uniss

UNIVERSITY of SASSARI

PhD school in

Biomolecular and Biotechnological Sciences

PhD Programme: Molecular and Clinical Microbiology

XXVI cycle

PhD School Director: Prof. Claudia Crosio

Development of new technologies to study gut microbiomes

Supervisor:
Prof. Sergio Uzzau

PhD candidate:
M. Sc. Antonio Palomba

PhD School Director:

Prof. Claudia Crosio

Table of contents

Abstract

Metaproteomics allows the qualitative and quantitative evaluation of the protein complement of an environment at a given time. Given the youth of this research field, significant efforts are needed to optimize sample preparation and data analysis workflows for metaproteome analysis.

A major task is aimed at developing novel, rapid and efficient workflows for shotgun metaproteomic analysis.

In the present PhD Thesis the investigation of a number of experimental methods have been developed to optimize sample preparation and its MS analysis. Methods were assessed on mock and real gut microbiome samples, combining bead-beating/freeze-thawing for protein extraction, FASP for clean-up and digestion, and single-run LC-MS/MS for peptide separation and identification. The impact of different sequence databases on data analysis was evaluated using mock microbial mixtures. Upon comparison of experimental metagenomic-derived and publicly deposited databases, complementary results suggested the use of iterative searches and suitable taxonomy filters to improve metaproteomic analysis. According to data obtained, the workflow enables protein identification also from fungi, showing high reproducibility (>99%), sensitivity ($<10^4$ bacterial CFUs) and dynamic range ($>10^4$).

Finally, this workflow was successfully applied to investigate the sheep fecal metaproteome, obtaining the identification of more than 35,000 proteins belonging to more than 700 microbial species (10 % of which fungi).

List of abbreviations

4MM: 4-organisms microbial mixture;

9MM: 9-organisms microbial mixture;

9MM-H: 9-organisms microbial mixture treated with method harsh;

9MM-M: 9-organisms microbial mixture treated with method mild;

aa: Amino acid;

ABC: Ammonium bicarbonate;

ABF: Archaea, Bacteria, Fungi;

ABFV: Archaea, Bacteria, Fungi, and Viruses;

ACN: Acetonitrile;

BFV: Bacteria, Fungi, and Viruses;

BLAST: Basic local alignment search tool;

Blat: *Brevibacillus laterosporus*;

bp: basepair;

CDS: Coding DNA sequence;

CFU: Colony-forming unit;

CID: Collision induced dissociation;

DB: Database;

DNA: Deoxyribonucleic acid ;

DTT: Dithiothreitol;

Ecol**:** *Escherichia coli*;

Efae**:** *Enterococcus faecalis*;

ESI: Electronspray ionization;

FASP: Filter-aided sample preparation;

FDR: False discovery rate;

GC: Gas chromatography;

GO: Gene ontology;

HCD: High-energy collision-induced dissociation;

HMP: Human microbiome project;

IAM: Iodoacetamide;

Laci**:** *Lactobacillus acidophilus*;

LC: Liquid chromatography;

LCA: Lowest common ancestor;

Lcas**:** *Lactobacillus casei* (group);

MEGAN: Metagenome analyzer;

Meta-6FT: Metagenome 6-frame translation;

Meta-PA: Metagenome predicted and annotated;

MFM: Murine faecal microbiome;

MMM: Mock microbial mixture;

MS/MS: Tandem mass spectrometry;

MS: Mass spectrometry;

NCBI: National center for biotechnology information;

NGS: Next generation sequencing;

NIH: National institutes of health;

NSAF: Normalized spectral abundance factor;

PBS: Phosphate-buffered saline;

PCR: Polymerase chain reaction;

***Pmul*:** *Pasteurella multocida*;

***Ppen*:** *Pediococcus pentosaceus*;

ppm: Parts per million;

PSM: Peptide-spectrum match;

RDP: Ribosomal database project;

***Rglu*:** *Rhodotorula glutinis*;

rpm: Revolutions per minute;

rRNA: Ribosomal ribonucleic acid;

***Scer*:** *Saccharomyces cerevisiae*;

SDS: Sodium dodecyl sulphate;

SDS-PAGE: Sodium dodecyl sulphate - polyacrylamide gel electrophoresis;

SGA-6FT: Single genomes assembly 6-frame translation;

SGA-PA: Single genomes assembly predicted and annotated;

SwissProt: UniProtKB SwissProt database;

TBP: Tributylphosphine;

TFA: Trifluoroacetic acid

TrEMBL: UniProtKB TrEMBL database;

WGS: Whole genome sequencing.

List of Tables

Chapter 1

Chapter 3

Chapter 4

Chapter 6

List of Figures

Chapter 1

Chapter 4

Chapter 5

Chapter 6

All the quoted figures are reproduced with the permission of their publishers.

Chapter 1

Introduction

1.1 Microbiome

1.1.1 What is it?

The word "microbiome" was coined by Joshua Lederberg and Alexa McCray in 2001 to indicate "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space and have been all but ignored as determinants of health and disease" (Lederberg and Mccray, 2001). Numerous studies have estimated that microbial cells in an animal body could exceed tenfold the number of host cells (approximately up to 100 trillion microbial cells against 10 trillion host cells), and that the total number of genes associated with the microbial organisms could be more than hundredfold superior than the total number of human genes (Bäckhed *et al.*, 2005; Ley *et al.*, 2006a). Some of these microorganisms can cause illnesses, and are thus to be considered as pathogens, but many other are not only harmless, but often absolutely necessary for host healthy.

For this reason, immediately after the conclusion of human genome sequencing, Relman and Falkow have highlighted the importance of the microbial component in the host life, asserting that "it is time to embark on a comprehensive genomic inventory of the large portion of cellular life within the human body that has been ignored so far, the endogenous microflora" (Relman and Falkow, 2001). In 2007, almost in response to this appeal, the National Institute of Health (NIH) of the United States of America launched the Human Microbiome Project (HMP), whose focal point was the description of the microbial diversity associated with health and disease (Peterson *et al.*, 2009; Turnbaugh *et al.*, 2007).

This project, beyond the shadow of a doubt, marked the beginning of the era of microbiome studies.

1.1.2 The Human Microbiome Project

The HMP has been immediately and universally considered as an extraordinarily ambitious project with a great deal of skepticism about the chances of success, but good results have immediately arrived. To give an example, one of the principal objectives of HMP was the production of reference genome sequences for at least 900 bacteria from several human body sites, and, just a couple of years later, 356 genomes were produced by the NIH HMP Jumpstart Consortium (formed by: the Human Genome Sequencing Center, Baylor College of Medicine, Houston; the Broad Sequencing Platform, Broad Institute of the Massachusetts Institute of Technology/Harvard, Cambridge, Massachusetts; The J. Craig Venter Institute, Rockville, Maryland; the Washington University Genome Sequencing Center, Washington University School of Medicine, St. Louis), including 178 genomes that have been completely annotated. These sequences, representing two kingdoms (Bacteria and Archaea), nine phyla, 18 classes, and 24 orders, were distributed among the gastrointestinal tract, the urogenital/vaginal tract, the skin, the oral cavity, and the respiratory tract (Nelson *et al.*, 2010).

One interesting finding described in this initial report was the distribution of the new species obtained by HMP researchers around the tree of life, as depicted in Figure 1-1. This picture shows the phylogenetic tree of 16S rDNA sequences with every specific phylum marked with a different color, and the organisms sequenced as part of the HMP project in blue. Despite a lack of detail in the branching structure (and several minor artifacts as acknowledged by the authors), it can be clearly seen an overall distribution of the HMP organisms around the whole tree of life, suggesting that there are microbial species still unknown in every phylum (Nelson *et al.*, 2010).

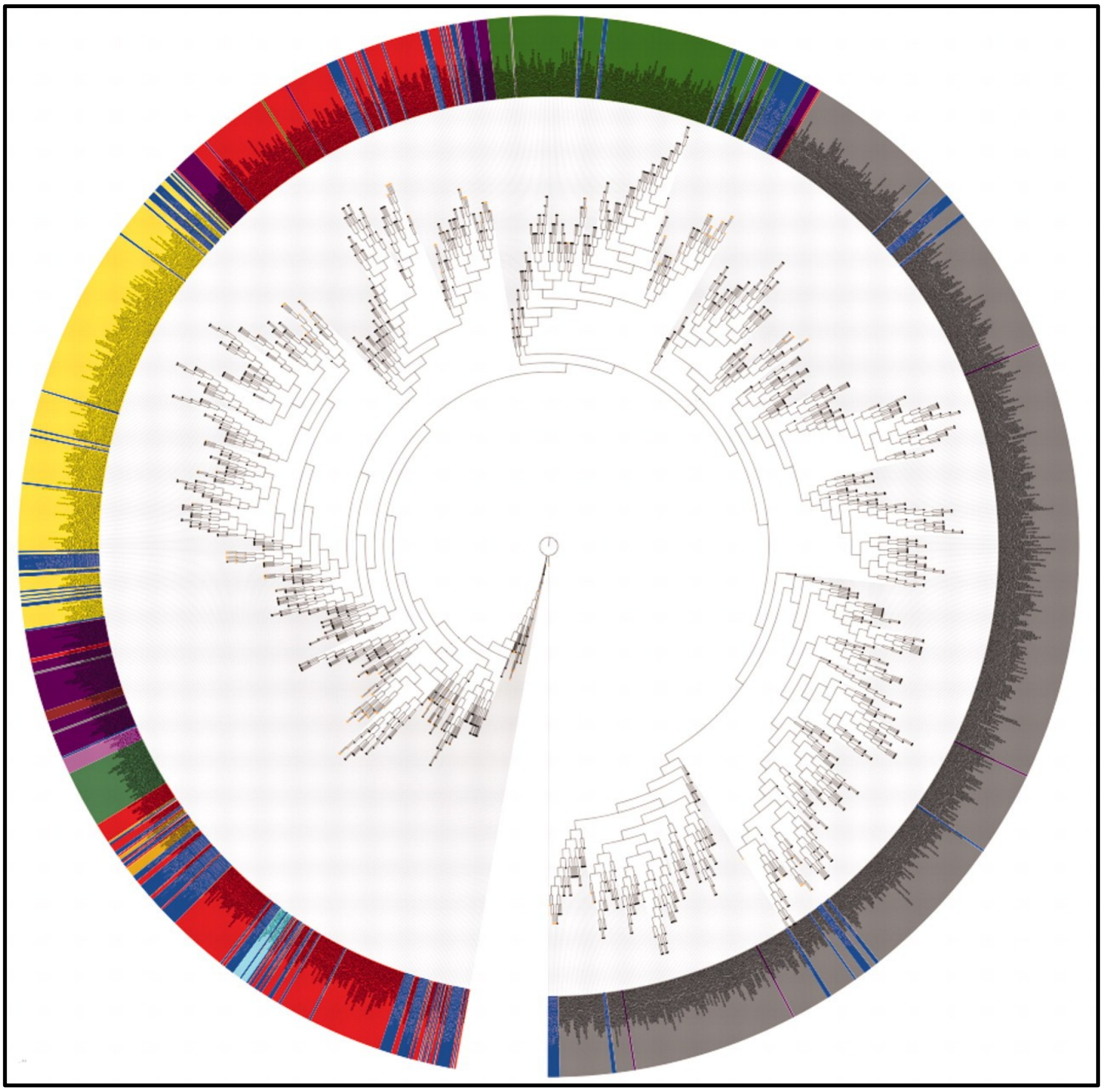


Figure 1 - 1. Overall distribution of the organism sequenced as part of the Human Microbiome Project (HMP) around the tree of life. The tree was created using 16S rDNAs representing single species. Organisms sequenced as part of the HMP are highlighted in blue. This image shows the phylogenetic tree of 16S rDNA sequences with any specific phylum marked with a different color: Actinobacteria in yellow, Bacteroidetes in dark green, Cyanobacteria in light green, Firmicutes in red, Fusobacteria in cyan, Planctomycetes in dark red, Proteobacteria in gray, Spirochaetes in magenta, TM7 in light pink, Tenericutes in tan.

In addition, the researchers compared 16.8 million microbial sequences found in public databases (DBs) to the genome sequences in the HMP reference collection, discovering that 62 genomes in the reference collection showed similarity with 11.3 million microbial sequences in public DBs, and 6.9 million of these (about 41%) corresponded with genome sequences in the reference collection. This analysis demonstrates that genomes sequenced as part of the reference collection add directly to an understanding of the human microbiome. Researchers also evaluated the microbial diversity present in the HMP reference collection, and found 29,693 previously undiscovered proteins, a number of protein superior to the estimated genes in the human genome (https://commonfund.nih.gov/hmp/) This excellent result has been followed by several others, allowing to achieve the objective initially fixed more rapidly than foreseen. For these reasons the original aim of 900 genomes (established in 2007) was changed, and the current (since 2012) objective of HMP is to sequence, or collect from publicly available sources, a total of at least 3,000 reference genomes isolated from human body sites (http://www.hmpdacc.org/).

1.1.3 Microbiome establishment and dynamics

The colonization of a specific biological niche might be more controlled by abiotic and biotic factors than microbial dispersal ability (Gonzalez *et al.*, 2011). Due to the generally unlimited dispersal capacity of microbes, the theory "everything is everywhere, but the environment selects" could be appropriate to explain the achievement dynamics of microbiome balance (Quispel, 1998; de Wit and Bouvier, 2006). As depicted in Figure 1-2, the host body can be seen as an ecosystem exposed to ecological processes including, for example, dispersal (horizontal transfer between two different individuals or between two different sites of the same individual), invasion (sudden appearance of a new "exotic" microorganism in a specific site), and succession (change in the species structure of an ecological community over time) (Gonzalez *et al.*, 2011).

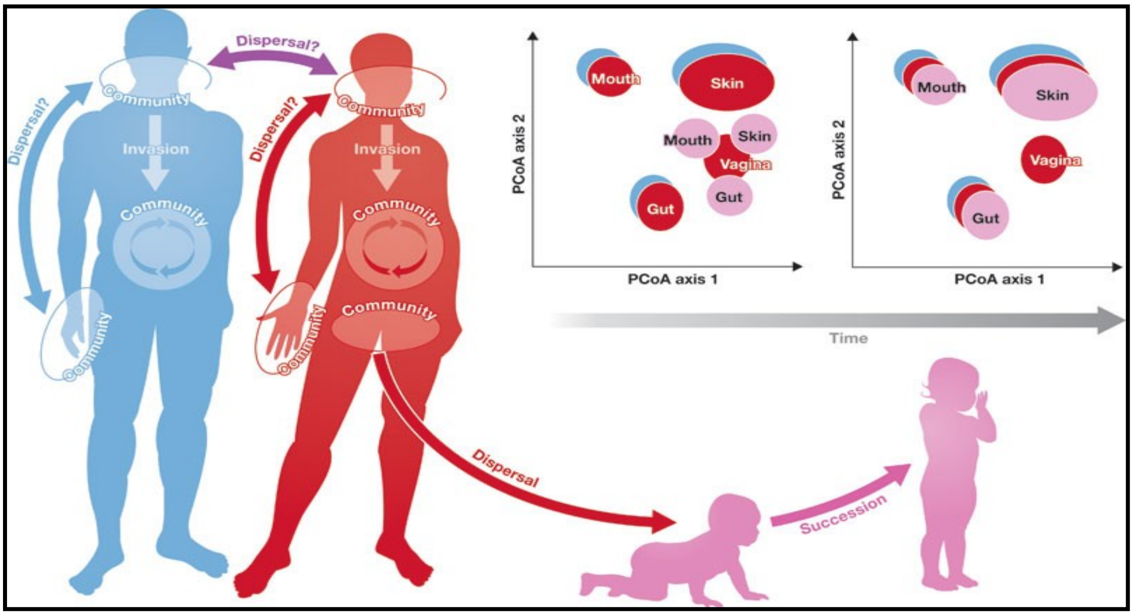


Figure 1 - 2. Ecology of human microbiome. The human body can be visualized as an ecosystem that is subject to the ecological processes that structure communities, including dispersal, invasion, succession, and meta-community dynamics.

Usually, in a first phase, the chemical and physical properties of the specific body site, namely pH, aerobic or anaerobic conditions, nourishment availability, etc., select against the microorganisms impaired to survive. Progressively such microorganisms contribute to modify the native environment in order to facilitate their self-survival, making the specific site always more selective. In this way, each site and its microbial composition is distinct, as it has been clearly shown by several studies describing the vagina, penis, intestinal tract, skin, and oral microbiome, as reported in Figure 1-3 (Jenkinson, 2011; Kong, 2011; Lamont *et al.*, 2011; Price *et al.*, 2010; Qin *et al.*, 2010; Ravel *et al.*, 2010; Turnbaugh *et al.*, 2009).

The microbial balance can be altered by drug application, especially antibiotic treatments that could eliminate, in addition to pathogens, also the commensal ones, resulting in a global remodeling of microbial hierarchy (Pérez-Cobas *et al.*, 2012). Furthermore, differences in life style that can condition food assumption, exposure to pets and livestock, and many other factors could influence how and where a gut microbiome is acquired (Yatsunenko *et al.*, 2012).

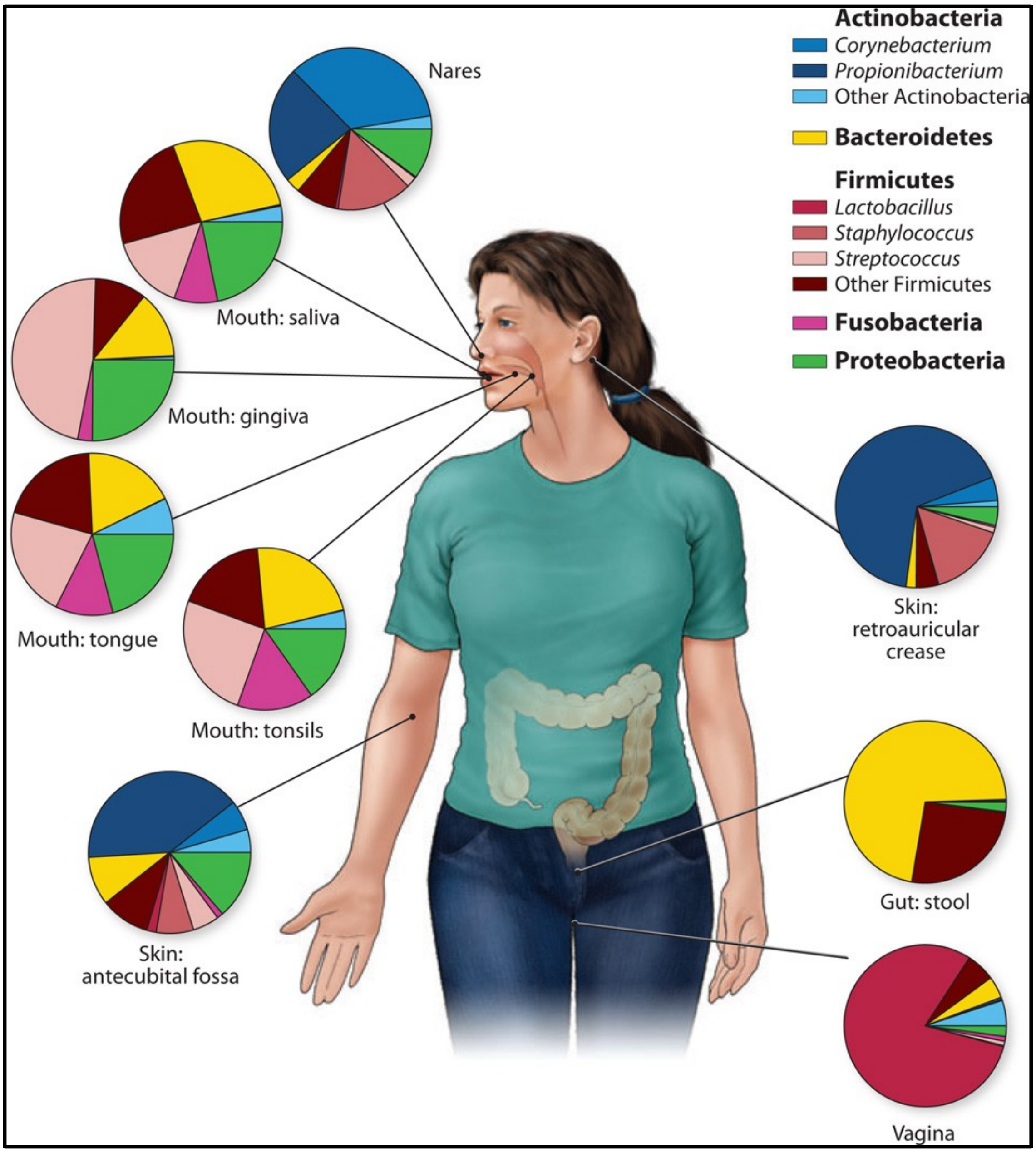


Figure 1 - 3. Genus- and phylum-level classification of Bacteria colonizing a human host. Each body site is characterized by specific bacterial taxonomy distribution. Districts with similar chemical and physical features share a greater similarity than others.

For this reason, the microbial composition among healthy individuals can be extremely different. In addition, the components of the human microbiome change over time, affected, for instance, by the patient disease state and medication. However, the microbiome eventually returns to a state of equilibrium, even if the composition of bacterial types has changed. Several studies have also shown that the microbiota of a specific site within the same individual is dynamic, varying naturally in correlation with age (Yatsunenko *et al.*, 2012).

Despite all these sources of variability, the presence of a “core microbiome”, defined as “those species-level phylotypes in a given body habitat that were observed across all sampling events”, has been demonstrated (Caporaso *et al.*, 2011). For example, as far as gut microbiome is concerned, in 2011 Arumugam *et al.* pointed out that human gastrointestinal microbiome can be clustered in three distinct groups, identifiable by the levels of one of the three bacterial genera *Bacteroides*, *Prevotella*, and *Ruminococcus*, and named enterotypes 1, 2, and 3, respectively. These enterotypes, each of which rich in genes involved in specific and alternative pathways exploited to generate energy from complex carbohydrates, seem to have correlation with none of the host properties evaluated, namely gender, age, or nationality (Arumugam *et al.*, 2011).

Nevertheless, just a year later another group begun to query whether this classification could simplify to an extreme degree the situation, suggesting, in turn, a continuum of species rather than discontinuous variation with segregated types (Jeffery *et al.*, 2012). Nonetheless, understanding which microbial taxa constitutes a “core microbiome” is of pivotal importance to enhance knowledge concerning microbial ecology, to determine their influence upon metabolic functions, as well as to use taxonomic profiles as possible diagnostic markers (Figure 1-4) (Li *et al.*, 2013).

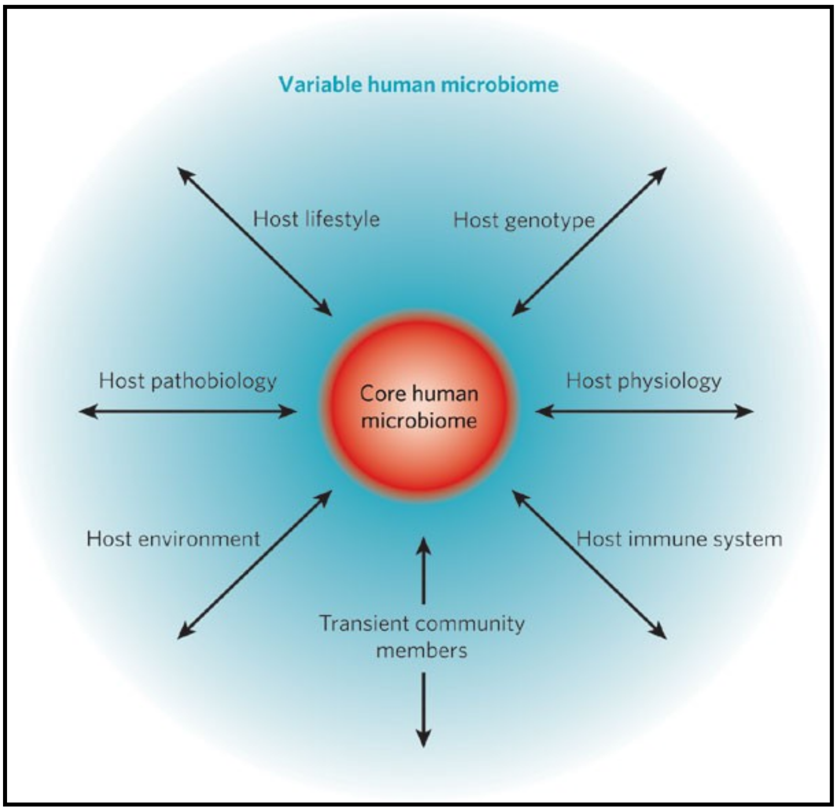


Figure 1 - 4. The concept of a core human microbiome. The core human microbiome (red) is the set of genes present in a given habitat in all or the vast majority of humans. The variable human microbiome (blue) is the set of genes present in a given habitat in a smaller subset of humans. This variation could result from a combination of factors.

Moreover, it is also important to note that Bacteria are the most abundant inhabitants inside microbiome, but not the only residents. Other organisms, as Fungi (forming "mycobiome"), Virus (forming "virome"), and Archaea (forming what can someone begins to call "archaeome"), although less abundant than Bacteria (also for this reason indicated as "rare biosphere"), are more variable between different individuals, and are deemed to play an increasingly pivotal role (Huffnagle and Noverr, 2013; Minton, 2012; Sogin *et al.*, 2006; Williams, 2013). In a complex microbial communities a relative small number of microbial species dominate, but hundreds to thousands of low abundance microorganisms also exist. It is also important to note that the word "rare" is used in correlation to specific environment analyzed. At mucosal sites, for instance, the most abundant bacterial specie can reach 10^{10} microbes per gram. Consequently, microorganisms present at 10^4 cells per gram can be considered “rare” since they count up for only the 0.0001% of the cellular content of the community. This “rare biosphere” may also harbor species that have an unbalanced effect (positive or negative) on the dominant members of the microbiome, a potential way by which they may support physiological or pathological effects (Huffnagle and Noverr, 2013). In support of this, accumulating evidence has delineated a correlation between species that are poorly represented within the microbiome and the host physiology, as depicted in Figure 1-5. For example, it has been shown that *Methanobrevibacter* (the most widespread Archaea genus in human gut) and *Candida* (the second most prevalent Fungi genus, after *Saccharomyces*, in the same human site) were positively associated with diets rich in carbohydrates, but negatively with diets high in amino acids, proteins, and fatty acids (Hoffmann *et al.*, 2013). Moreover, even though the *Prevotella/Bacteroides* ratio was not significantly correlated with the fungal types, the same study demonstrated that it was significantly correlated with relative proportions of Fungi present. As far as Archaea are concerned, there was a significant correlation between archaeal genera *Methanobrevibacter* and *Nitrososphaera* with the bacterial genus *Bacteroides*, specifically negative with the former and positive with the latter (Hoffmann *et al.*, 2013).

Several other studies concerning gut virome have suggested that bacteriophages (Virus infecting Bacteria) are the biggest regulators of bacterial abundance (Hofer, 2013; Williams, 2013). All together, these findings highlight that the microorganisms counterbalance in a microbiome is very complex, and numerous factors, both internal and external to the host, could have a crucial importance.

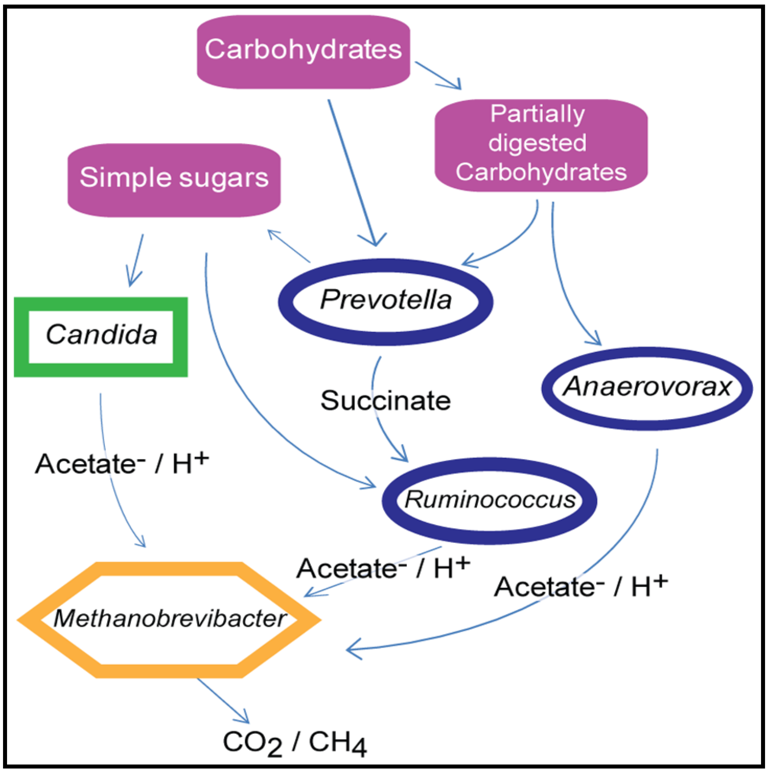


Figure 1 - 5. Possible syntrophic relationships in the human gut between fungal, archaeal, and bacterial microorganisms. Fungi are marked in green, Bacteria in blue, and Archaea in orange.

1.2 Techniques to study microbiomes

1.2.1 Metaculturomics

A microbiome can be studied for different purposes. On the one hand, some analyses, that could be roughed in as "descriptive", have the main objective to thrash out, both at qualitative and quantitative level, the microbial composition of a specific environment (for instance, an anatomic site of a particular animal). On the other hand, in studies that could be defined "associative", the chief aim is to identify a correlation between a specific physiologic or pathologic host status and some alteration in the microbial composition.

As represented in Figure 1-6, a microbiome can be characterized using different techniques. Traditionally, the best approach to study a microorganism has been culture-dependent. According to this method, two steps are crucial: plating the sample on a selective growth media, and recognizing the specific microorganism on the basis of several particular features, such as the morphological characteristics of colonies, the specific metabolic production, or the specific nutrient consumption. However, this kind of approach has huge limitations owing to the fact that, as it has been amply demonstrated by numerous evidences, more than 80% of the bacterial species present in the human body have not yet been cultured, or are even considered as unculturable (Bik *et al.*, 2006; Gevers *et al.*, 2012; Grice and Segre, 2012; Turnbaugh *et al.*, 2007). Another important limit is that an isolated microorganism in a pure culture obtained in laboratory is less representative of community interactions, due to the loss of precious information concerning the original ecological and molecular relationships between different microorganisms.

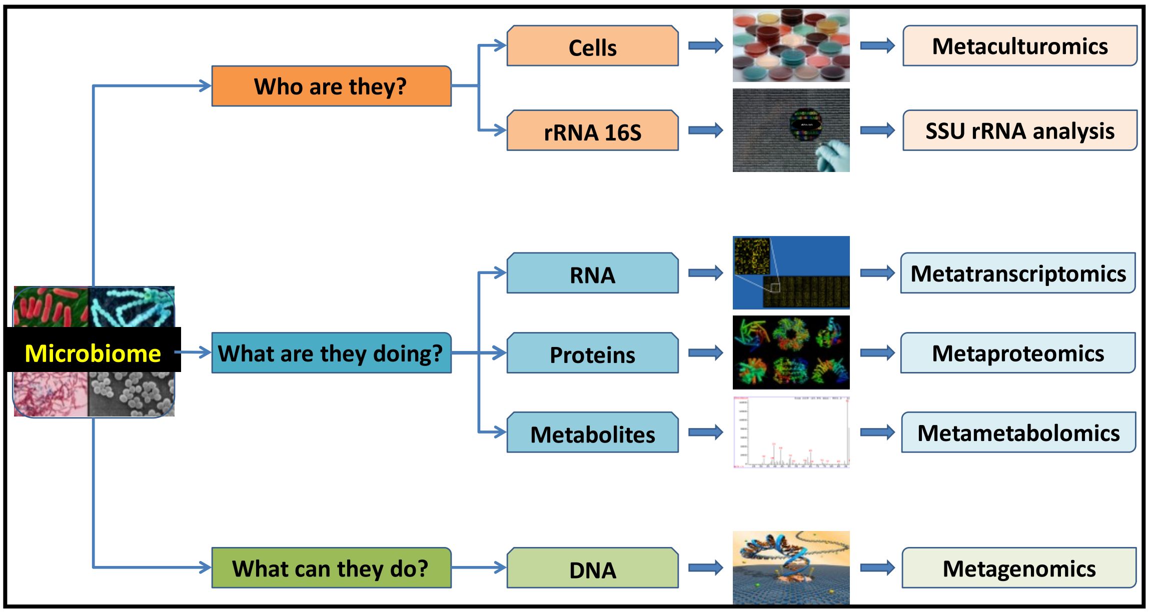


Figure 1 - 6. How to study a microbiome. Various approaches to answer important questions concerning a microbiome.

To overcome some of these limitations, in 2012 Lagier *et al.*, in a singular study, used more than 200 different culture conditions to identify microorganisms belonging to the human gut microbiome. With this analysis, researchers obtained 32,500 colonies belonging to 340 species (174 of which never described previously in the human gut) of Bacteria from seven phyla and 117 genera, including two species from rare phyla (Deinococcus-Thermus and Synergistetes), five Fungi, and a giant Virus (*Senegalvirus*, the largest Virus reported in the human gut). These results achieved with striking efforts were comparable, for the first (and, to date, unique) time, to those achievable with more sophisticated technologies. However, the extremely long time of this approach (also called "metaculturomics") and the extreme complexity of the experimental design indubitably reduce its routinely application (Lagier *et al.*, 2012).

1.2.2 Metagenomic

In the last years, along with advancements in molecular technologies, especially in sequencing and mass spectrometry (MS) instruments, alternative methods of microbial communities analysis have become available. This continuous improvement has led to the emergence of new branches of research, namely metagenomics, metaproteomics, metatranscriptomics, and metametabolomics, opening the door to a high-throughput analysis of microbial communities using culture-independent methods (Grice and Segre, 2012).

Metagenomics is based on extraction of DNA directly from a clinical or environmental sample. So far, two techniques have been the most adopted. The first one is the 16S rDNA tag sequencing, used to typify bacterial taxonomy according to information concerning 16S ribosomal RNA gene. The strong point of this method is the use of the *16S rDNA* gene, which contains both highly conserved sequences, that allow polymerase chain reaction (PCR) amplification using broad-range primers, and specie-specific hypervariable sequences, available for phylogenetic characterization (Hugenholtz and Pace, 1996). The *16S rDNA* gene gathers up all characteristics to be a perfect marker to identify the genome that contains it, without sequencing the entire genome. It is simple enough to be analyzed both for its reduced dimension, approximately 1500 base pairs (bps) in length, and for its high number of copies in some microbial genomes. Moreover, it is contained in every member of a population, differing only between distinct individuals with specific genomes and, in addition, it varies proportionally to the evolutionary distance between specific microbes, facilitating consequently taxonomic attribution (Morgan and Huttenhower, 2012). Actually, for all these reasons, it has reached an extremely high level of reliability, becoming the most popular technique to perform taxonomic classification (Almeida and Araujo, 2013; Carroll *et al.*, 2012; Han *et al.*, 2013; Hu *et al.*, 2013; Maughan *et al.*, 2012; Nava and Stappenbeck, 2011; Newton *et al.*, 2011; Santamaria *et al.*, 2012; Shahinas *et al.*, 2012; Tringe and Hugenholtz, 2008; Woo *et al.*, 2008). As reported by Grice and Segre in February 2012, the *16S rDNA* sequences deposited in Ribosomal Database Project (RDP) were more than 2 millions shared in 35 different phyla (Cole *et al.*, 2007; Grice and Segre, 2012).

The second genomic-based approach is founded on the Whole-Genome Sequencing (WGS) technology, that allows the identification of all genetic material from the different organisms making up a community in a specific ecosystem, by extracting and analyzing their DNA globally. The first studies have been focused on environmental and ecological communities, for example acid mine drainage (AMD), because of their lower complexity. The results of such analyses have been useful to pinpoint the presence of uncultivable microorganisms (Bacteria, Archaea, Fungi and Viruses), some of which, as mentioned above, can have pivotal importance to environment safeguard or host health, regardless of their abundance (Denef *et al.*, 2010). The complexity of a microbial community can range from very simple (extreme environments inhabited by few specialists, such as the above mentioned AMD) to very complex (environments inhabited by a variety of species, such as the gut microbiome). Obviously, the depth of knowledge which can be reached by the WGS approach can be dramatically different depending on the community under study. Simple communities, with only 76 Mb sequencing effort, can result in the assembly and analysis of almost complete genomes of the dominant species, including accurate metabolic reconstruction and detection of strain-specific genomic variants. More complex communities, with a much larger sequencing effort (almost 2 Gb), can result in very fragmented assemblies even for the most abundant species, with most of the dataset being represented by singleton sequencing reads (Chistoserdova, 2010). A huge technological advancement in sequencing instruments has made it possible to achieve these kind of results, allowing the investigation of several different genomes simultaneously. Taking into consideration, as described in Figure 1-7, that in 2002 millions of dollars were needed to obtain a complete genome sequence, it is truly amazing that, currently, the same information can be obtained with a few thousand dollars. In addition, it is also important to note that what was accessible in months or years of work just until a few years ago, now it can be obtained in few days, if not hours, with a higher reliability of the results.

A typical workflow of a WGS analysis (Figure 1-8) consists of a first fragmentation of genomic DNA, with the creation of a library of small segments, that in the next step are accurately sequenced in millions of parallel reactions.

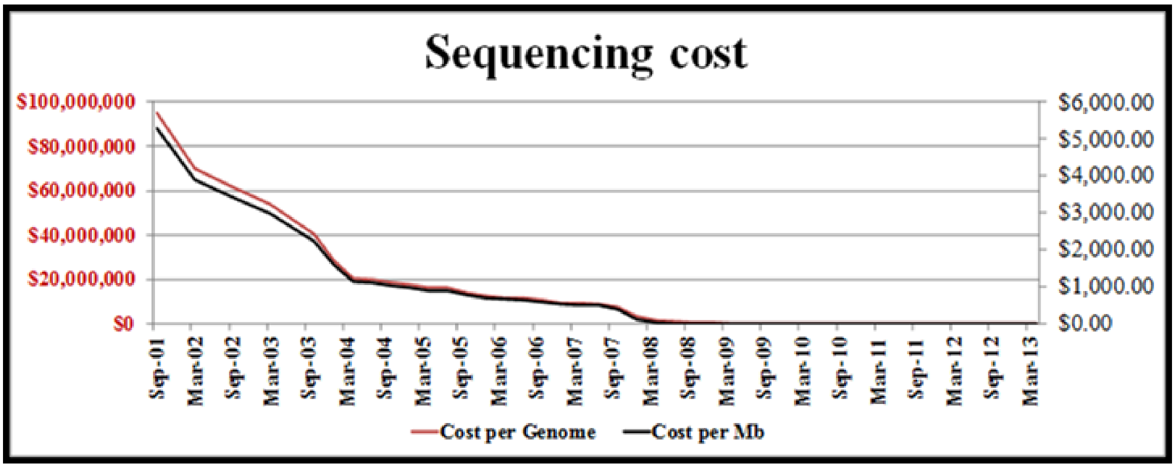


Figure 1 - 7. Trend of sequencing cost during the last years. Cost per genome (red line, values on the left) and per Mb (black line, values on the right) from 2001 to 2013 expressed in U.S. dollars. Data derived from the Genome Sequencing Program of the National Human Genome Research Institute (http://www.genome.gov/sequencingcosts/).

Then, these little nucleotide sequences, called reads, are reassembled either using a resequencing approach, that aligns the reads to a known reference genome employed as a scaffold, or using a *de novo* approach, where the alignment is achieved without reference information. This latter approach has the advantage to be useful to obtain genomic information about all microbial organisms (or, more in general, all organisms) independently from the preliminary level of information. Unfortunately, this strategy has the disadvantage to require very high quality data to achieve good results, i.e. an extremely high whole genome coverage, that is not always achievable, especially when several hundred different genomes are contained in the same sample. The resenquencing approach has reciprocal advantages and disadvantages compared to *de novo* sequencing; in fact, it is also possible to make use of data with poor coverage to identify a microbial organism, taking advantage of information concerning its reference genome. Obviously, this approach is not applicable to unknown, or not yet sequenced, microorganisms; in this regard, the lack of reference genome sequences represents the most important limitation to achieve trustworthy results with this technique. Therefore, further efforts to address these issues by generating new reference genomes, such as the Human Microbiome Project has done, are highly sought after (Grice and Segre, 2012).

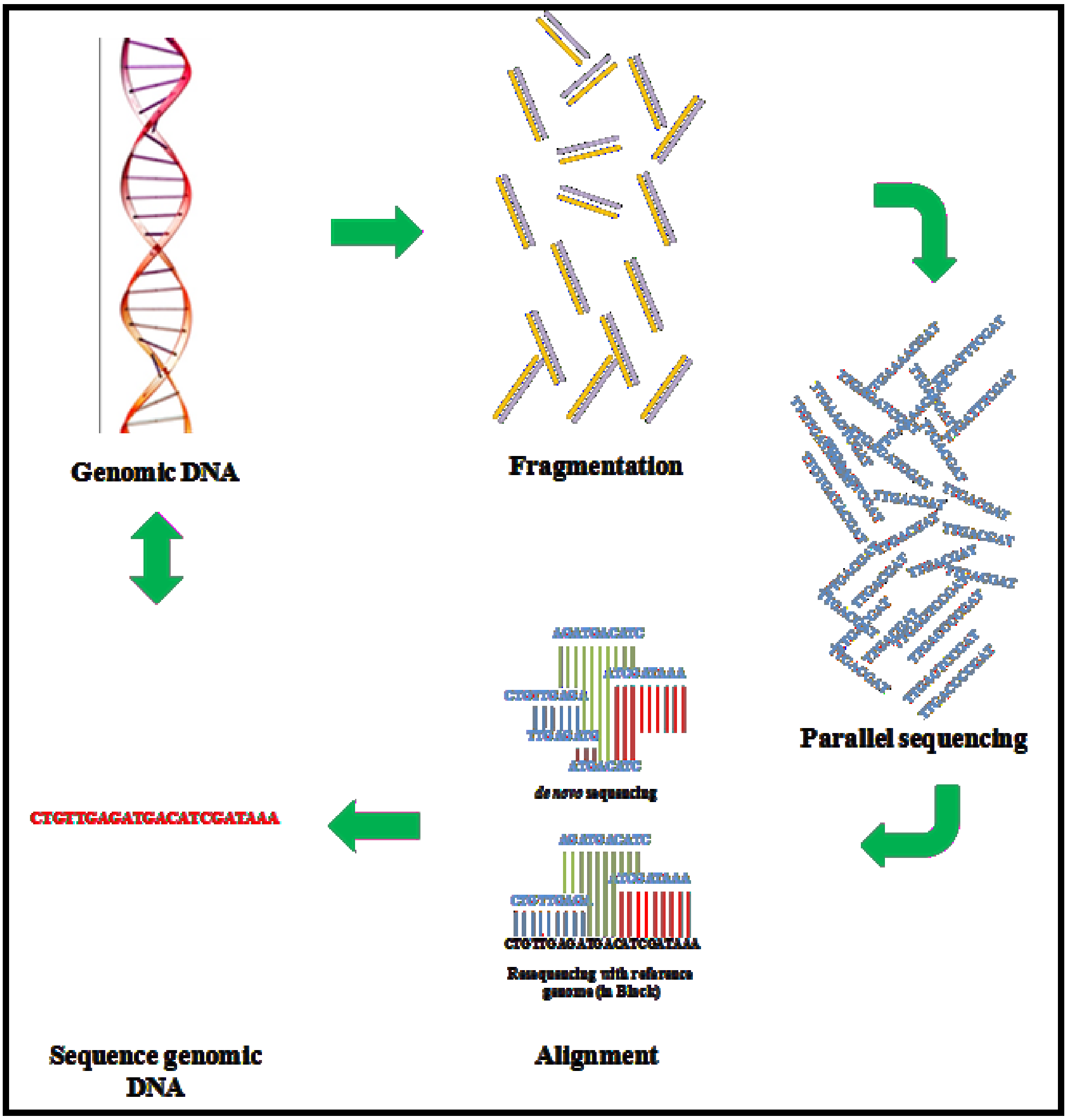


Figure 1 - 8. Whole genome sequencing (WGS) workflow.

To summarize, instruments with high performance in terms of both reliability and speed, combined with extremely reduced costs, have permitted to obtain in short time the sequencing of various genomes belonging to prokaryotic and eukaryotic organisms, and even of whole microbial communities with very high microbial complexity, such as human gut (Figure 1-9).

Another important limitation of this kind of techniques is that it is not possible to know if the extracted DNA originated from intact, viable cells, or not. For this reason, DNA is not the ideal system to evaluate functions carried out by the community at a specific point in time (Morgan and Huttenhower, 2012).

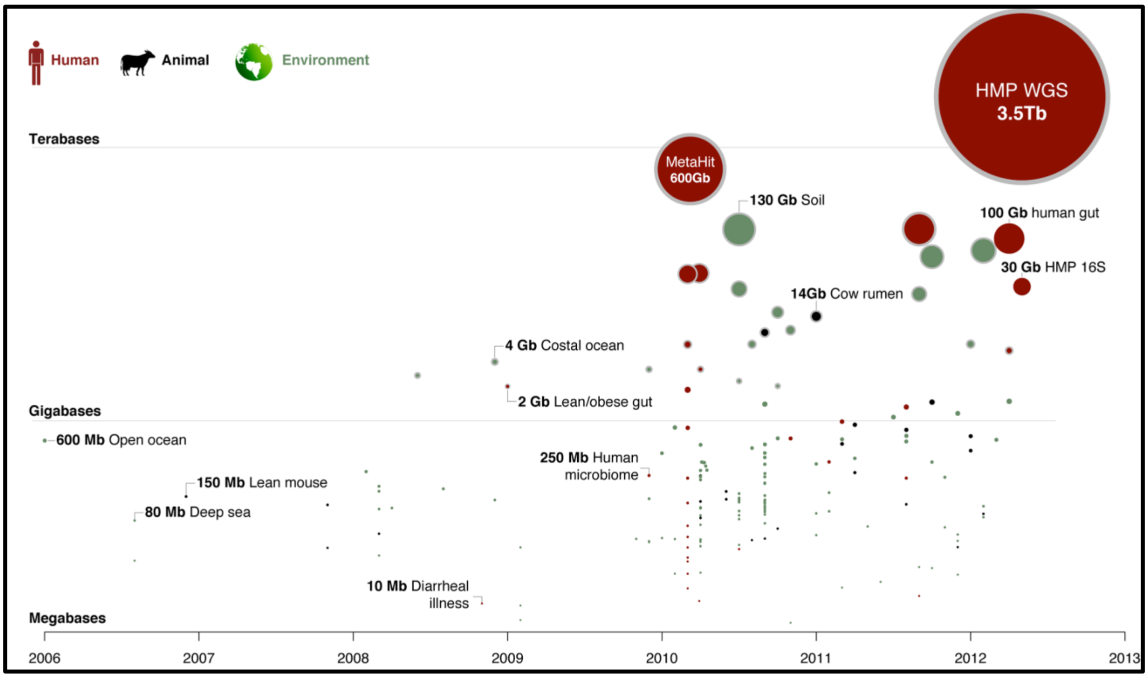


Figure 1 - 9. Timeline of microbial community studies using high-throughput sequencing. Each circle represents a high-throughput sequence-based 16S or shotgun metagenomic bioproject in NCBI (May 2012), indicating the amount of sequence data produced for each project (circle area and y-coordinate). Projects are grouped by human-associated (red), other animal (black), or environmental (green) communities, and shotgun metagenomic projects are marked with a grey band.

1.2.3 Other “omic” approaches: Metatranscriptomics

Through 16S rDNA tag sequencing and WGS it is possible to answer to two important questions, “Who is there?” and “What can they do?”, respectively, but different strategies are required to address another important question, namely “What are they doing?”. In order to obtain this kind of information, it is thus necessary to look at the expression profile (metatranscriptomics), and at metabolite (metametabolomics) or protein (metaproteomics) production. These kinds of analysis are still technically demanding and have only recently begun to be applied for studying microbial communities.

Metaproteomic techniques, probably the most interesting to assess the functions accomplished by microbes, are becoming increasingly popular. Since the main topic of this thesis is metaproteomics, this approach will be discussed in a chapter apart.

Metatranscriptomics, that is the qualitative and quantitative description of genes expressed at a given time by all organisms attending an ecological niche, can be seen as an interesting strategy to describe functionally a microbiome. Unfortunately this approach is very challenging due to various features concerning prokaryotic mRNA. For example, bacterial mRNA completely lacks the 3’-end poly(A) tail that instead marks mature molecules in eukaryoric mRNA, making their enrichment and analysis easier. In addition, this technology must deal with the intrinsic biases associated with the need for subtraction of ribosomal RNA (rRNA) that is normally the dominant RNA species extracted, usually comprising over 90% of the total RNA. As Figure 1-9 suggests, this problem could be overcome by deep sequencing of total RNA, including rRNA, since current depth of coverage would still be sufficient to obtain considerable mRNA transcripts (Lamendella *et al.*, 2012). In other words, whether we can generate about 50 Gb of sequences, 90% of which are rRNA (and so not usable to our purposes), the remaining 5 Gb of other RNA, including millions of employable transcript reads, can be sufficient to complete a transcriptomic analysis. Furthermore, RNA is more difficult to prepare and preserve compared to DNA, owing to its chemical nature. For these reasons, initially, this type of studies has been mainly applied to samples from water and soil environments; human samples, as those originated from the human gastrointestinal tract, have been successfully analyzed only in the last four years (Gosalbes *et al.*, 2011).

1.2.4 Other "omic" approaches: Metametabolomics

A further alternative is represented by metametabolomics (also called less awkwardly "community metabolomics"), that provides information concerning the complete spectrum of small-molecules, and their changes as a consequence of a particular stimulus. This approach includes various analytical technologies such as high-resolution nuclear magnetic resolution (NMR), GC-MS and LC-MS, in combination with chemometrics and bioinformatics tools (Turnbaugh and Gordon, 2008; Wikoff *et al.*, 2009; Xie *et al.*, 2013). Metabolites of microbial origin can be characterized by analyzing low molecular weight compounds in biofluids (blood and urine), intestinal contents, and tissues (especially feces), achieving a metabolic fingerprint profile, associable with individual phenotypes, in correlation with physiological and pathological statuses. Metabolomics can be applied to explain the molecular mechanisms of host-microorganism interactions during a disease, taking advantage of quantitative information about specific metabolite levels, such as bile acids and short-chain fatty acids (SCFAs) that are modulated during the pathologic process (Xie *et al.*, 2013).

1.3 Metaproteomics

1.3.1 Preliminary remarks

As de Hoog and Mann pointed out in their review on proteomics in 2004, “biological function is not carried out by the static genome but mainly by the dynamic population of proteins determined by an interplay of gene and protein regulation with extracellular influences”(de Hoog and Mann, 2004). In other words, the information obtainable by the global analysis of the proteins expressed in a given sample is key to carry out its full characterization. Proteomics offers the opportunity to identify the protein repertoire collectively expressed by an organism, making it possible to estimate protein abundance, either relatively or absolutely, and thus providing important insights into physiology, metabolism and cellular functionality, and/or confirming the real expression of proteins only inferred “*in silico”* from genome information (Hettich *et al.*, 2012; Siggins *et al.*, 2012; VerBerkmoes *et al.*, 2009a; Wilmes and Bond, 2006). This approach is able to provide details on the pathways that are actively functioning in a community, and on how the expression of specific proteins can change according to time, location, or environmental stimuli (Ottman *et al.*, 2012). In particular, when analyzing a particular microbiome, it can be more important to know which functions are carried out by the microbial components present in a biological district, than which specific microbial species are present within. Different microorganisms can in fact perform the same function, thus a divergence in microbial composition between two samples is not always correlated to an equivalent altered microbial functionality.

Unfortunately, several factors can seriously hamper a correct protein identification, and consequently a realistic proteome characterization. To give an example, one important issue in proteomics is the difficulty to access minor or under-represented proteins in a complex sample. The high dynamic range of proteins, that in some samples like blood can reach 12 orders of magnitude, still remains a challenging task despite the huge improvement in MS sensitivity (Zubarev, 2013).

In general, as illustrated in Figure 1-10, a typical shotgun proteomics workflow consists of few pivotal steps: protein extraction; protein digestion; MS analysis; computational analysis.

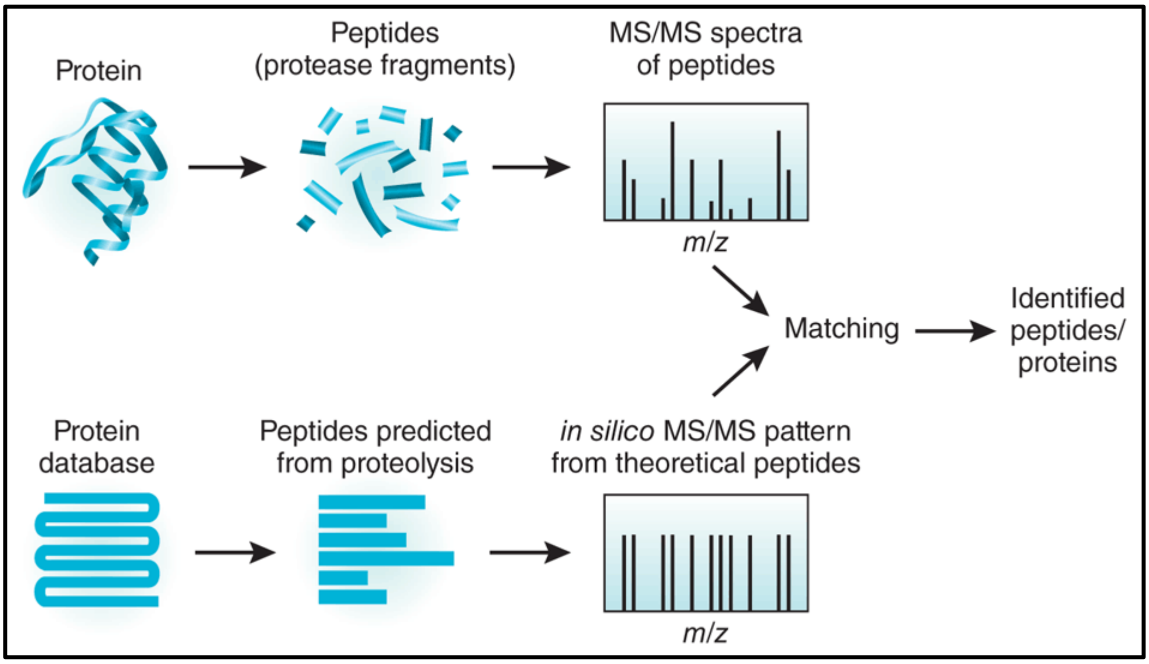

Protein
Peptides (protease fragments)
MS/MS spectra of peptides
m/z
Matching
Identified peptides/ proteins
Protein database
Peptides predicted from proteolysis
in silico MS/MS pattern from theoretical peptides
m/z

Figure 1 - 10. Typical proteomic workflow. General approach used by peptide-centric MS technologies for the identification of proteins in complex mixtures. After proteolysis of a protein or complex mixture of proteins, the spectra associated with protease fragments are matched with spectra generated "*in silico*" using information obtained from protein databases.

1.3.2 Protein extraction

The objective of the protein extraction step is to maximize the recovering of all proteins included in the sample and to minimize the presence of other molecules that can hamper the following analysis. To achieve this result, a wide assortment of methods are suitable: mechanical (French Press; bead-beating; grinding), physical (boiling, freeze-thawing; snap-freezing; sonication), chemical (using buffers containing one or more components among detergents, such as sodium dodecyl sulphate, SDS, CHAPS, and Triton X-100; chaotropic agents, such as urea and guanidine hydrochloride; reducing agents, such as dithiothreitol, DTT, and tributylphosphine, TBP; and other organic/inorganic compounds, such as phenol and sodium hydroxide) or enzymatic (deoxyribonuclease; ribonuclease) approaches can be used, depending on sample features (Abram *et al.*, 2009; Benndorf *et al.*, 2007; Chourey *et al.*, 2010; Fouts *et al.*, 2012; Kan *et al.*, 2005; Keiblinger *et al.*, 2012; Klaassens *et al.*, 2007; Kolmeder *et al.*, 2012; Leary *et al.*, 2012; Schneider *et al.*, 2012; Verberkmoes *et al.*, 2009a; Wilmes and Bond, 2004).

Since Gram-positive bacteria, Gram-negative bacteria and yeasts have important structural differences, and therefore a variable susceptibility to each protein extraction method, the choice of a specific approach may significantly bias the quality and the quantity of the proteomic results in the direction of a specific category of microorganisms. In same conditions the application of a single disruption method, among those mentioned above, can be sufficient for an efficient protein extraction, but in other circumstances, with more resistant samples, a combination of two or more of them might be necessary. In this regard, the combination of strong buffer components and harsh treatments may probably help maximize extraction yields and avoid selective depletion of species showing a higher resistance to lysis, such as yeasts and Gram positive bacteria, therefore enabling a more complete representation of the microbial community proteome.

1.3.3 Protein digestion

The subsequent step, protein digestion, is mainly achieved using specific enzymes that break peptide bonds in a process where all reaction conditions, such as duration, temperature, and pH, are carefully controlled. The most used enzyme is trypsin, that is a serine protease able to cleave peptide chains at the carboxyl side of the amino acids lysine or arginine, except when either ones are followed by proline. This process of proteolysis is also called trypsinization (Hustoft *et al.*, 2010). Prior to this step, it is also important to remove compounds that can limit enzymatic digestion and/or the following process steps, namely liquid chromatography (LC) separation and MS analysis. The typical way to achieve this result is using protein precipitation, which can be accomplished by adding, for instance, trichloroacetic acid, acetone, or ammonium acetate/methanol to the protein extract; then, the protein pellet is resuspended in a buffer compatible with the subsequent steps (Benndorf *et al.*, 2007; Chourey *et al.*, 2010; Leary *et al.*, 2012; Sharma *et al.*, 2012). However, significant sample losses due to protein aggregation may occur (Fic *et al.*, 2010; Jiang *et al.*, 2004). Another effective opportunity is to perform 1-dimension electrophoresis (1-DE) protein separation followed by in-gel digestion of the extracted proteins, which allows both the entrapment of interfering compounds within the gel matrix and the sample fractionation into gel slices (Ferrer *et al.*, 2012; Haange *et al.*, 2012; Kolmeder *et al.*, 2012). Unfortunately, although efficient, this method is labor-intensive and time-consuming, and reproducibility may not arrive to high values (Choksawangkarn *et al.*, 2012). A recent alternative is represented by the filter-aided sample preparation (FASP), in which sample clean-up and enzymatic cleavage take place in a molecular weight cut-off centrifuge filter (Wiśniewski *et al.*, 2009). This procedure has been applied with success in recent times to environmental microbiome samples, and was demonstrated to outperform several competing methods principally for low protein amounts (Sharma *et al.*, 2012).

1.3.4 Mass spectrometry analysis

The next step, MS analysis, is key because it allows the peptide identification to be achieved. Mass spectrometers are commonly interfaced upstream with separation devices, as gas chromatographs (GC) or LC. The separated components are then introduced into the mass spectrometer. Currently, the most performing instruments are based on a tandem mass spectrometry technology (MS/MS) that combines two mass spectrometers (Figure 1-11). Briefly, these instruments work by using magnetic and electric fields to exert forces on charge particles (ions). Therefore, the peptide mixture must be charged (ionization) to be analyzed. The choice of the ionization method depends on the nature of the sample. In proteomic analysis, one of the most performing ionization sources is the electrospray ionization (ESI). This technique, that can be classified as a “soft ionization” method, tends to produce mass spectra with little or no fragment-ion content. The sample solution is sprayed across a high potential difference (a few kilovolts) from a needle into an orifice in the interface. Heat and gas flows are used to desolvate the ions existing in the sample solution. This process often produces multiply charged ions with the number of charges tending to increase as the molecular weight increases. Later, the first mass analyzer detects a spectrum from which a single mass ion, also called precursor or parent ion, with a particular mass/charge (m/z) ratio, is selected. In turn, the precursor ion produces its fragment ions due to a harder ionization obtained by colliding the selected ions with a neutral gas. This process can be named collision-induced dissociation (CID) or higher-energy collision-induced dissociation (HCD), depending on the extent of collision energy used. Finally, such fragment ions are separated into the second mass analyzer according to their m/z ratio. The resulting MS/MS spectrum consists only of product ions generated from the selected precursor ion (Guthals and Bandeira, 2012; Rotilio *et al.*, 2012).

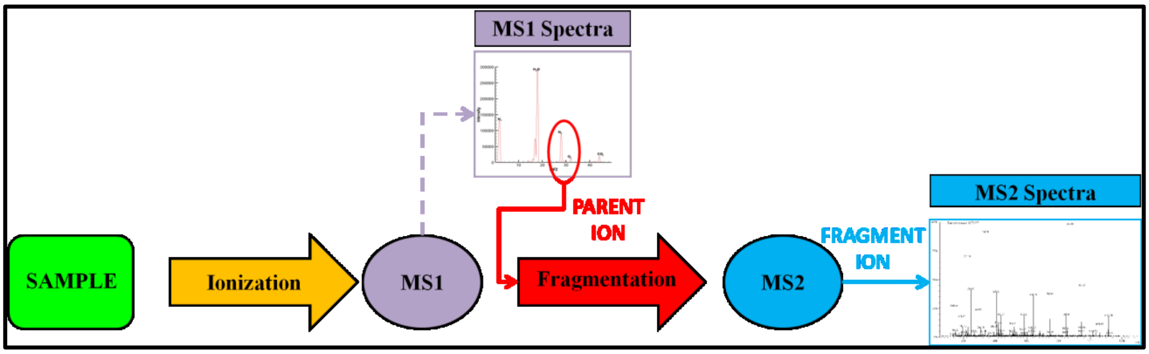


Figure 1 - 11. Diagram illustrating tandem mass spectrometry analyis worfklow (MS/MS). A sample is injected into the mass spectrometer, ionized and accelerated, and then analyzed by the first mass analyzer (MS1). Ions from the MS1 spectra are then selectively fragmented and analyzed by the second mass analyzer (MS2) to give the spectra for the ion fragments. While the diagram indicates separate mass analyzers (MS1 and MS2), some instruments can utilize a single mass analyzer for both rounds of MS.

Prior to the MS analysis step, sample complexity generally has to be reduced in order to improve the amount of information achievable by shotgun MS analysis. This has been attained in previous metaproteomic studies by carrying out a separation at the protein (mainly by 1-DE and GELFrEE approaches) and/or peptide level (most commonly by means of 2D-LC) (Kolmeder *et al.*, 2012; Pérez-Cobas *et al.*, 2012; Ram *et al.*, 2005; Schneider *et al.*, 2012; Sharma *et al.*, 2012; Verberkmoes *et al.*, 2009b). However, each additional fractionation step implies a corresponding increase in the quantity of starting material, laboratory effort, and/or MS measuring time required, as well as increasing challenges in analytical repeatability. In particular, 2D-LC-MS/MS, although reaching a very remarkable analysis depth, is technically challenging and, above all, requires extremely long times for a single sample to be analyzed (22 hours in a typical experimental setting) (Verberkmoes *et al.*, 2009a). Recently, a straightforward approach based on single-run nanoLC-MS/MS has been described, enabling the identification of several thousands of proteins per run from different kinds of sample (Köcher *et al.*, 2012; Nagaraj *et al.*, 2012; Pirmoradian *et al.*, 2013; Thakur *et al.*, 2011).

1.3.5 Peptide and protein identification

The last step concerns peptide identification, which is carried out starting from the m/z ratio signal obtained from MS analysis. To accomplish this result, a comparison is normally performed between the experimentally obtained signal (experimental spectrum) and the theoretical spectrum, which is obtained by digesting "*in silico*" all protein sequences contained in a sequence database (DB) through a dedicated software called search engine (such as SEQUEST, Mascot, OMSSA, and X!TANDEM) (Craig and Beavis, 2004; Geer *et al.*, 2004; Perkins *et al.*, 1999; Yates *et al.*, 1995). The resulting collection of identified peptide sequences is then assembled into an inventory of proteins that can account for the identified peptides. Since the likelihood of matching an MS/MS spectrum to a given peptide increases with the amount of that peptide in the sample under analysis, the number of MS/MS spectra that map to a given protein provides a quantitative estimation of the amount of that protein within the sample.

1.3.6 Quantification

A crucial issue in proteomic research concerns the accurate quantification of proteins contained in a sample (Bantscheff *et al.*, 2012). The most accurate quantitation methods consist of using heavy or light stable isotopes incorporated into the different proteomes to bring into comparison (Hinkson and Elias, 2011). On the one side these technologies provide very accurate measurements, on the other they require additional steps in sample preparation, reducing the experimental repeatability and increasing the laboratory effort.

For this reason, alternative approaches, especially the so-called "label-free quantification", characterized by more user-friendly procedure, are routinely preferred. The label-free based techniques, as the name suggests, do not involve any kind of marking of the samples. As stated before, they are based on correlation between several mass spectrometric signals (for example, peak intensity or spectral counting), linked to a specific peptide, and the original amount of the same peptide, and therefore of the protein, in the sample. This technique reduces substantially the number of steps in the procedure, resulting in a lower labor time and a higher reproducibility (Figure 1-12). Unfortunately, however, such techniques have an important limitation concerning sensitivity. In fact, label-free quantification requires the abundance variation to be at least 2-fold to be detected, whereas with metabolic labeling approach it could be possible to describe also protein variation of a few percent (Mann *et al.*, 2013).

1.3.7 Issues in metaproteomic analysis

Several studies have shown that the description of a protein expressed from a microbial community is very demanding, chiefly in data analysis and interpretation (Muth *et al.*, 2013; Seifert *et al.*, 2013). In this respect, two main issues do severely hamper the analysis of a metaproteome: first, genome sequence data might be unavailable for most of the species contained in the particular microbial community under study, thus considerably reducing the possibility of a correct matching between the experimental spectra and the theoretical spectra; second, a typical environmental sample contains thousands of proteins belonging to up to thousands of different microbial species, often having a high level of homology, making therefore both peptide-to-protein and peptide-to-taxon assignments a really tremendous task.

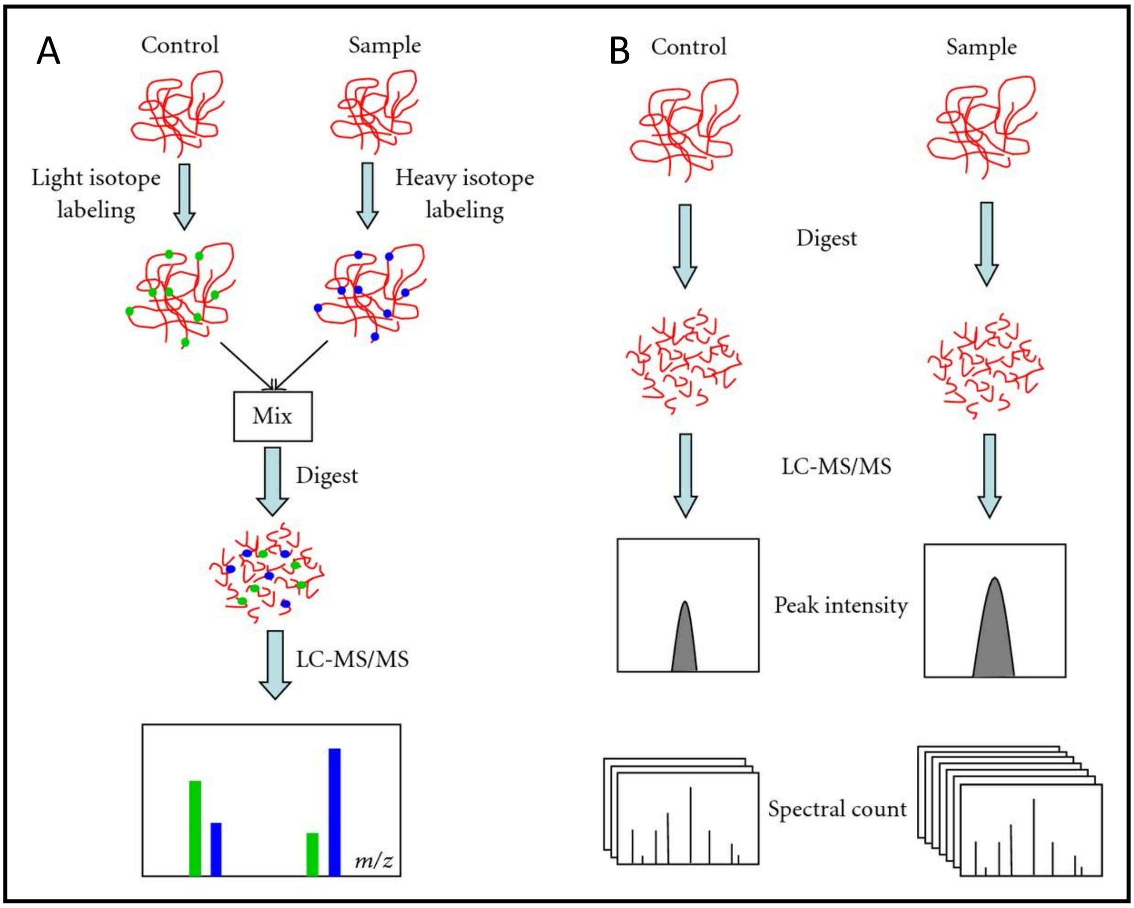


Figure 1 - 12. Quantitative proteomic approaches. (A) Shotgun isotope labeling method. After labeling by light and heavy stable isotope, control and sample are combined and analyzed by LC-MS/MS. The quantification is calculated based on the intensity ratio of isotope-labeled peptide pairs. (B) Label-free quantitative proteomics. Control and sample are subjected to individual LC-MS/MS analysis. Quantification is based on the comparison of peak intensity of the same peptide or the spectral count of the same protein.

1.3.8 Database impact

The selection of proper protein DBs represents an extremely critical step, especially when dealing with poorly characterized microbiomes. When a novel microbial community is subjected to metaproteome analysis, without further genomic investigation, publicly available DBs have to be used for peptide/protein identification, almost for a preliminary analysis. Protein DBs can be generally distinguished into non-manually annotated with plenty of information but huge dimensions, and thus very high computing times, such as NCBI and

UniProtKB/TrEMBL (TrEMBL), and manually curated sequences as UniProtKB/SwissProt (SwissProt), with inverse pros and cons in comparison with the first ones (Figure 1-13)(NCBI Resource Coordinators, 2013; The UniProt Consortium, 2012).

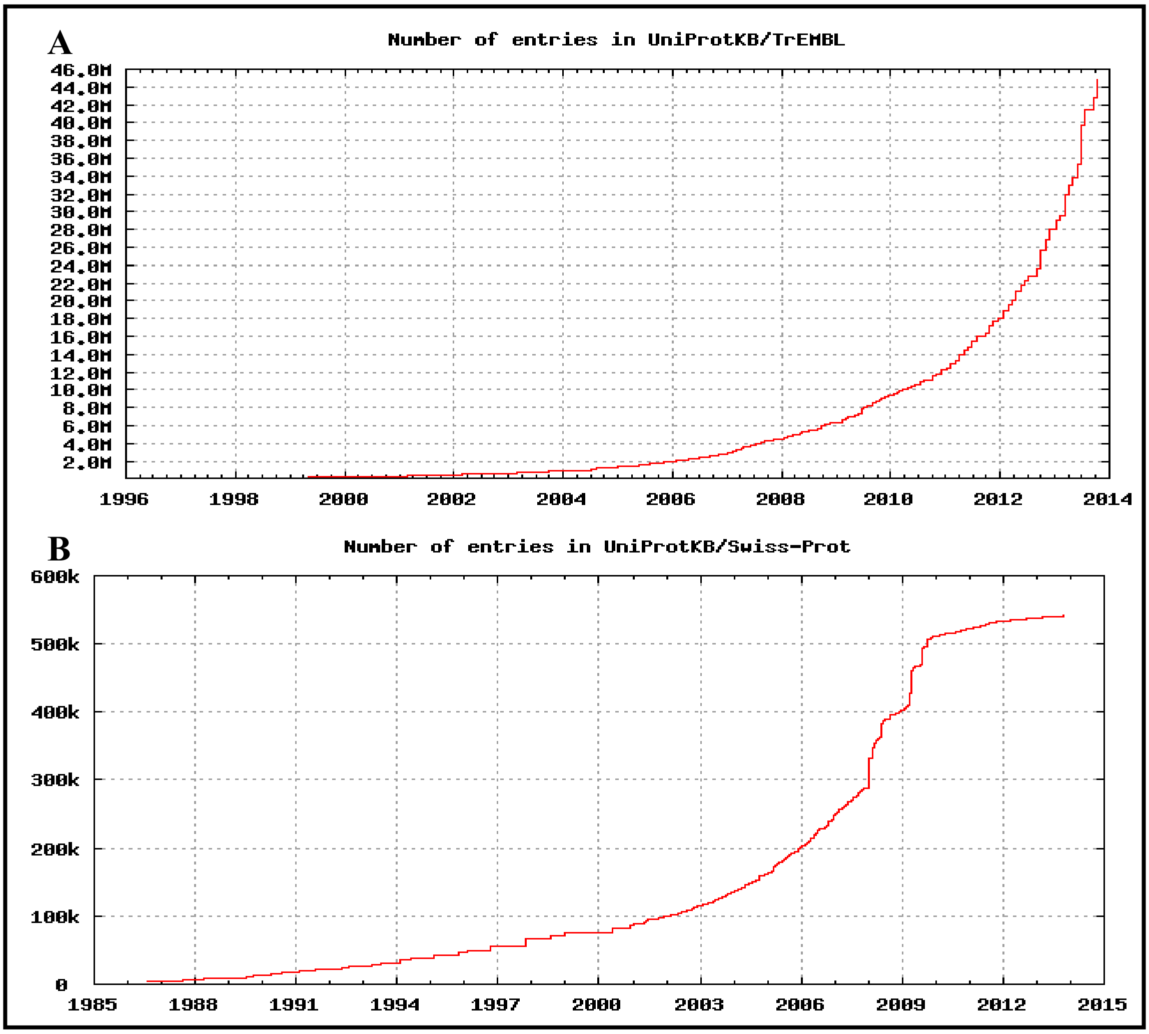


Figure 1 - 13. Number of protein sequences in UniProtKB/TrEMBL and UniProtKB/SwissProt databases.

Unfortunately, most uncultivable species have not yet been sequenced, in spite of the great efforts made in the last few years by genome scientists, and are therefore not present within the public resources. In this case, cross-species identification can occur when genome sequences of closely related species, with large sequence homology regions, are available (Wright *et al.*, 2010). Unlike "classical" DNA sequence homology search, in proteomics even slight differences in amino acid sequences lead to significant variations in peptide masses, thus making the proteomic characterization of unsequenced organisms extremely difficult. A possible alternative is using *de novo* sequencing, in which amino acid sequences are deduced directly from fragmentation spectra, without the need of a protein DB, followed by BLAST (basic local alignment search tool) search employed to identify candidate homology proteins (Seidler *et al.*, 2010; Shevchenko *et al.*, 2001). However, manual inspection of spectra is often required due to the error-prone nature of *de novo* sequencing, and very high quality data are necessary for reliable results to be achieved (Pevtsov *et al.*, 2006).

In the very recent past an increasing number of papers have described the integration of metagenomics and metaproteomics holds promising to address the above described issues (Cantarel *et al.*, 2011; Delmotte *et al.*, 2009; Denef *et al.*, 2009; Erickson *et al.*, 2012; Ferrer *et al.*, 2012; Rooijers *et al.*, 2011; Verberkmoes *et al.*, 2009b). Currently, such integration may occur at different levels:

1. using 16S (and/or 18S) rDNA gene sequencing information to assemble a customized DB (also named "pseudo-metagenome") restricted to the taxa which have been (or are expected to be) found within the microbiome under study, saving up analysis time and minimizing species misassignments (Verberkmoes *et al.*, 2009b);
2. using translated and annotated metagenome sequences as protein DB, ideally generated from the same sample being analyzed with metaproteomics (a so-called "matched" metagenome), but also retrieved from public metagenome archives, which are expected to impressively

grow in the years to come (Erickson *et al.*, 2012; Morris *et al.*, 2010; Ram *et al.*, 2005);

3. isolating further reference strains from the microbiome under study and performing individual genome sequencing, on the basis of a labor-intensive, in-depth approach, called "metaculturomics", as mentioned above (Lagier *et al.*, 2012).

Furthermore, according to a proteogenomic (*sensu stricto*) approach, metagenomic and genomic sequences can also be translated in all six reading frames (six-frame translation, 6FT), with the purpose of minimizing the inherent biases derived from gene prediction methods (Armengaud *et al.*, 2013; Renuse *et al.*, 2011). However, metagenome-derived DBs may suffer from technical biases in DNA extraction versus species having less abundance or particular cell wall features, as well as from bioinformatic issues in sequence assembly and annotation.

To date no reports have been described critically comparing the metaproteomic data which can be obtained using different types of publicly available and matched metagenome-derived DBs.

Interestingly, each of the above mentioned DB types exhibits specific features, mainly in terms of overall size and sequence redundancy, which might in turn considerably affect two of the main issues in proteome bioinformatics, namely false discovery rate (FDR) assessment and protein inference. FDR calculation applies a probabilistic method that inherently takes into account the effects of multiple testing, by estimating the proportion of peptide-spectrum matches (PSMs) that are incorrect among all significantly identified PSMs (Figure 1-14).

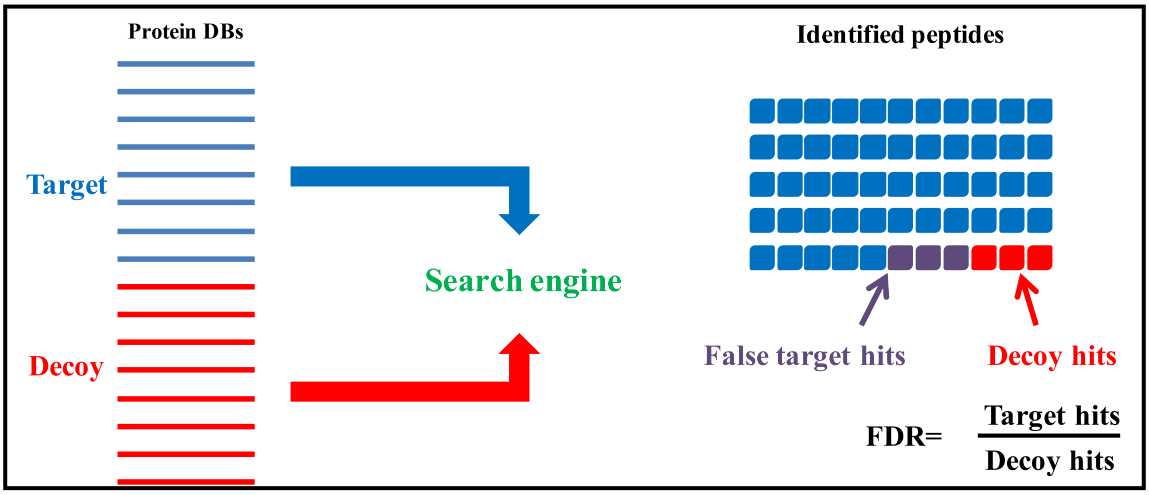


Figure 1 - 14. False discover rate (FDR) estimation. The decoy proteins are randomly generated, so any decoy hit is supposedly a false hit. The search engine is not able to distinguish target sequences from decoy sequences. Consequently, false positive identification can occur in both databases with equal probability. Thus, the total number of false target hits can be approximated by the number of decoy hits in the final result, and the FDR can be estimated by the ratio between the numbers of decoy hits and the number of target hits.

Several computational approaches have been developed to estimate the FDR at both peptide and protein level, usually exploiting the well-established target-decoy approach, although alternative statistical modeling approaches have been also developed (Elias and Gygi, 2007; Keller *et al.*, 2002; Renard *et al.*, 2010).

Advantages and limitations of the FDR approach in terms of quality, accuracy, and resolution have been critically discussed in various studies, but the simplicity and the global good efficiency of this approach make it the most often applied technique for the quality control of the data (Colaert *et al.*, 2011; Granholm and Käll, 2011; Vaudel *et al.*, 2011). The FDR applies globally to a set of PSMs, but single PSMs can also be associated with a *q*-value, defined as the minimal FDR of any PSM set that includes the given PSM (Granholm and Käll, 2011). Even though FDR estimation can be quite accurate and reproducible when a limited search space is concerned (e.g., a protein DB from a single organism), its resolution may significantly deteriorate when the search space complexity increases, as for proteogenomics and metaproteomics, with a consequent reduction in sensitivity (Blakeley *et al.*, 2012; Colaert *et al.*, 2011; Muth *et al.*, 2013). FDR accuracy and sensitivity are expected to be strongly influenced by the protein DB used, but this particular aspect has not been fully elucidated so far with regard to metaproteomic data (Blakeley *et al.*, 2012).

The second bioinformatic concern which may have a considerable impact on metaproteome analysis is represented by the protein inference problem, that is, how to assemble a list of peptides into a (reliable) list of proteins (Claassen, 2012; Nesvizhskii and Aebersold, 2005). When analyzing a single organism's proteome, ambiguities in peptide-to-protein assignment can be generally due to the presence of different splice variants or cleavage products. Unfortunately, when dealing with a metaproteome the scenario becomes terribly more complicated. In fact, many peptides (called degenerate peptides) can be shared among homologous proteins from different species, or even among recurring functional domains (Muth *et al.*, 2013). Under a DB perspective, a higher redundancy or homology in protein sequences corresponds to a higher degeneracy in peptide identification, and thus to harder issues in protein inference. Additionally, most of the widespread software suitable for protein/peptide identification usually display only a subset of all possible protein identifications; therefore, a tedious manual inspection for protein assignment is required in order not to over- or under-report important functional and taxonomic information (Kolmeder and de Vos, 2013).

1.3.9 Taxonomic Attribution

As far as taxonomic attribution is concerned, a simple but quite robust strategy to infer taxonomic information from (DNA or protein) sequence data is the so-called lowest common ancestor (LCA) approach (Huson *et al.*, 2007). According to this algorithm, a sequence is assigned to a given species only if it does not match with any other species contained in the sequence DB; conversely, if the sequence is shared among several species contained in the DB, all belonging to the same genus, the sequence is unambiguously assigned only to the genus level. Generally speaking, widely conserved sequences are always assigned to high-order taxa. When analyzing metaproteomics data, the LCA approach is clearly to be preferred over retrieving the taxonomic information using "classical" protein inference algorithms, as usually these systems select arbitrarily only one among the diverse taxonomic possibilities, with consequent loss of information (Kolmeder and de Vos, 2013). The LCA algorithm can be theoretically applied either at peptide or protein level: in the first case, LCA analysis should provide the most accurate results, in view of the peptide-centric nature of shotgun mass spectrometry; in the second case, as discussed before, the previous application of a protein inference algorithm not specifically suited for metaproteomics might introduce significant biases. The forerunner of LCA software, MEGAN, was originally developed for metagenomic data, but can be extended also to metaproteomics (Huson and Mitra, 2012; Huson *et al.*, 2011; Rudney *et al.*, 2010). Usually, in a preprocessing step protein/peptide sequences are compared against NCBI DB using BLAST, and MEGAN is then used to compute and explore the taxonomical content of the data set. A recent achievement of the metaproteomics community is the Unipept web application, which supports biodiversity analysis of metaproteome samples using tryptic peptide information obtained from shotgun MS/MS experiments, by retrieving all occurrences of the given peptides in UniProtKB records; taxon-specificity of the tryptic peptide is successively derived from these occurrences using a novel LCA approach (Mesuere *et al.*, 2012). To date, a critical evaluation of Unipept and/or MEGAN for the taxonomic profiling of metaproteomic data has not yet appeared in literature; furthermore, a possible influence of the DB choice on LCA results has not been investigated so far.

1.4 Gut microbiome

The human intestinal tract is colonized since birth by a large number of microorganisms that in adults can consist of around 10^{14} cells, with a complexity that is forecasted to include over 10^{3} microbial species, the vast majority of which has not yet been cultured (Zoetendal *et al.*, 2008). Several studies have suggested that most of the intestinal phylotypes belong to a limited set of phyla, including Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria and Verrucomicrobia (Figure 1-15 shows a preliminary overview of human gut microorganisms phyla) (Guarner and Malagelada, 2003; Lagier *et al.*, 2012). In healthy adults the intestinal microbiota fluctuates around a stable individual core of phylotypes that are affected by host genetics, environmental and stochastic factors (Rooijers *et al.*, 2011).

As Figure 1-16 shows, neonates develop in a bacteria-free environment until the delivery. Upon this moment they are exposed for the first time to a variety of different bacteria, in correlation with birth modality. During vaginal birth, the infant is exposed to microbes present in the birth canal of the mother, an environment that is mostly colonized by *Lactobacillus* (Ravel *et al.*, 2011). By contrast, babies delivered by caesarean section are not exposed to vaginal bacteria at birth and their bacterial communities resemble those found on skin (Dominguez-Bello *et al.*, 2010). Delivery mode thus leads to differences in the development of the microbiota, which may in turn contribute to variation in normal physiology or to disease predisposition. Some evidences have demonstrated the host-linked co-evolution of the immune responses and microbiota (Chung *et al.*, 2012; Hooper *et al.*, 2012; Olszak *et al.*, 2012). Moreover, it has been proved that early-life antibiotic exposure affects the long-term development of adipose tissue, lean muscle and bone (Cho *et al.*, 2012). Microbiota of genetically obese transgenic mice that show a particular taxonomy composition with a decrease in *Bacteroidetes* and an increase in *Firmicutes* species, found also in obese human volunteers, leads to a slight, but statistically significant, body fat increase when transferring in gut free mice (Ley *et al.*, 2006b; Turnbaugh *et al.*, 2006).

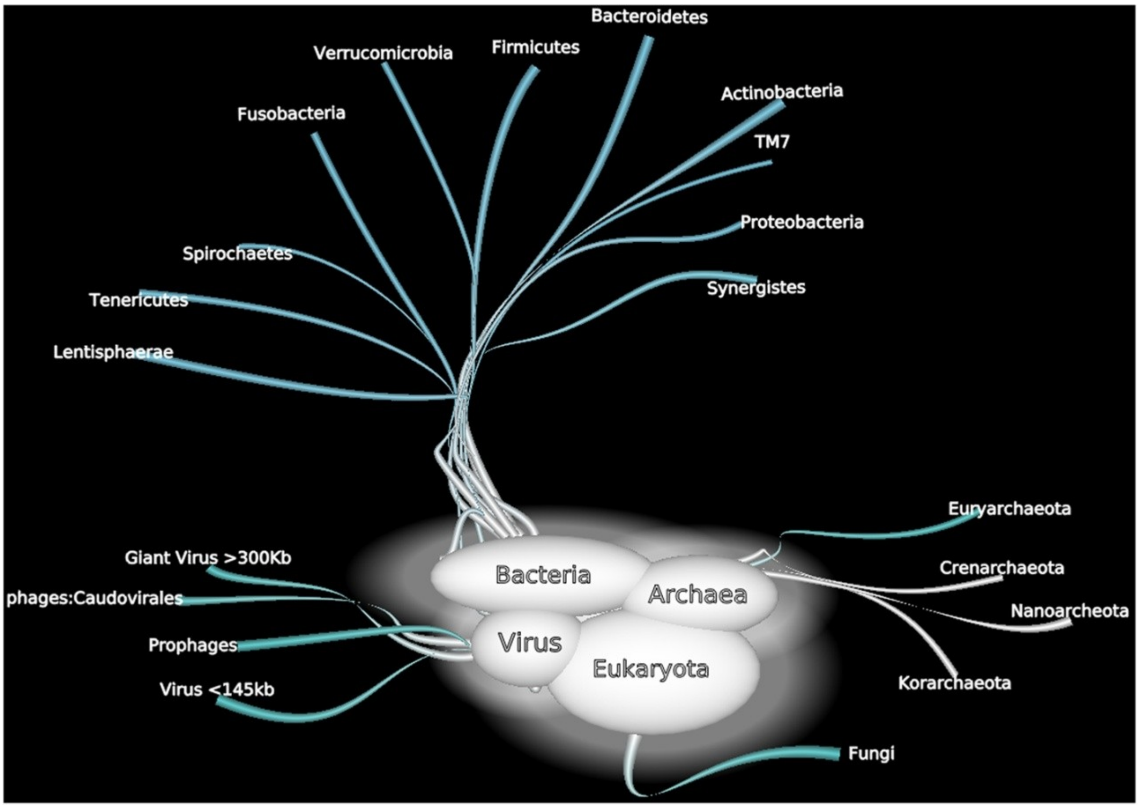


Figure 1 - 15. **A non exhaustive overview of human gut microorganisms among bacterial, Archaea, viral, and *Eukaryota* domains.**

In a study performed by Segata and coworkers in 2012, the important difference between composition, relative abundance, and metabolic potential of the bacteria population inhabitants of the adult digestive tract was clearly proved. In detail, a cluster of ten body habitats in four groups has been described, according to pattern of numerically dominant Bacteria taxa profiled using the 16S rDNA, as classified by the RDP (Cole *et al.*, 2009; Segata *et al.*, 2012).

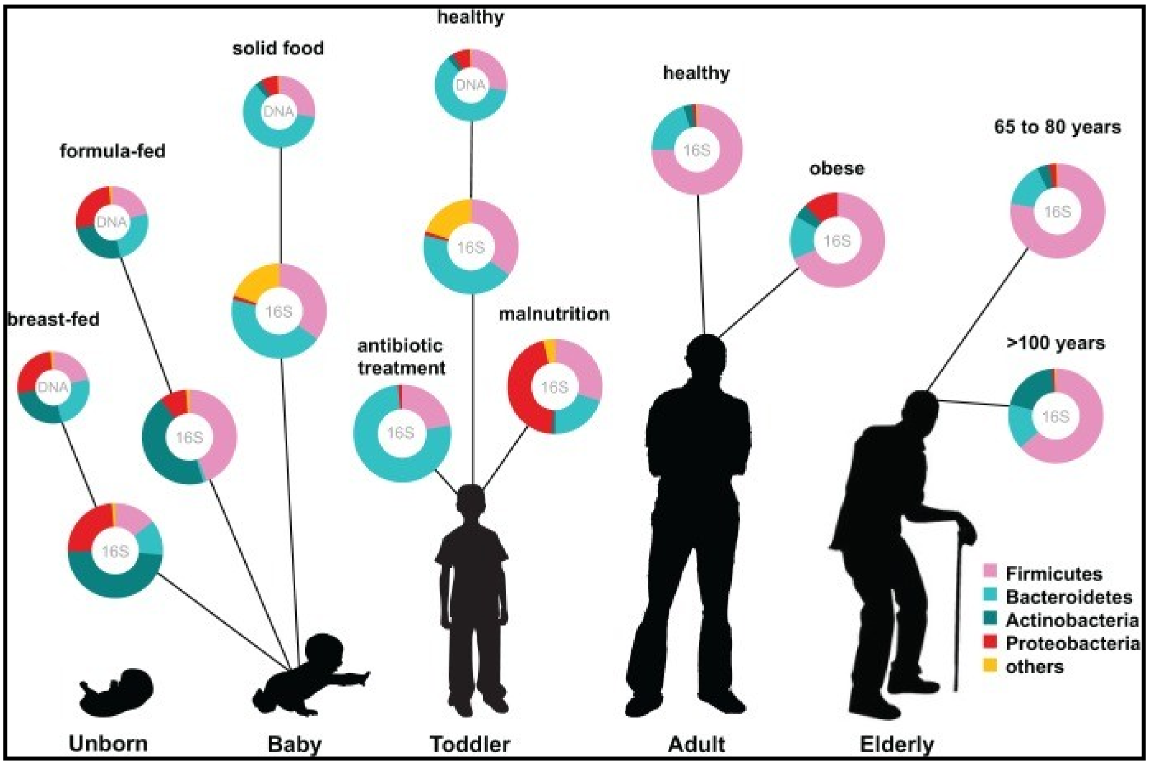


Figure 1 - 16. Human microbiota: onset and shaping through life stages and perturbations. The graph provides a global overview of the relative abundance of key phyla of the human microbiota composition in different stages of life, measured by either 16S RNA or metagenomic approaches.

The microbiota of Group 1 (buccal mucosa, keratinized gingiva, and hard palate) consisted mostly of Firmicutes followed in decreasing order of relative abundance by Proteobacteria, Bacteroidetes and either Actinobacteria or Fusobacteria. In comparison, Group 2 (saliva, tongue, tonsils, and throat or more exactly the back wall of oropharynx) had a decreased relative abundance of Firmicutes and increased levels of four phyla: Bacteroidetes, Fusobacteria, Actinobacteria and TM7. Group 3 (sub- and supra-gingival plaque) had a further decrease in Firmicutes compared to Groups 1 and 2, with a marked increase in the relative abundance of Actinobacteria (Figure 1-17).

The stool samples have been lonely gathered in a group (number 4) consisting mainly of Bacteroidetes (over 60%) followed by Firmicutes (including several genus-level biomarker for this group, as *Roseburia* and *Faecalibacterium* own to families Lachnospiraceae and Ruminococcaceae respectively), with very low relative abundances of Proteobacteria and Actinobacteria, and less than 0.01% of Fusobacteria (Segata *et al.*, 2012).

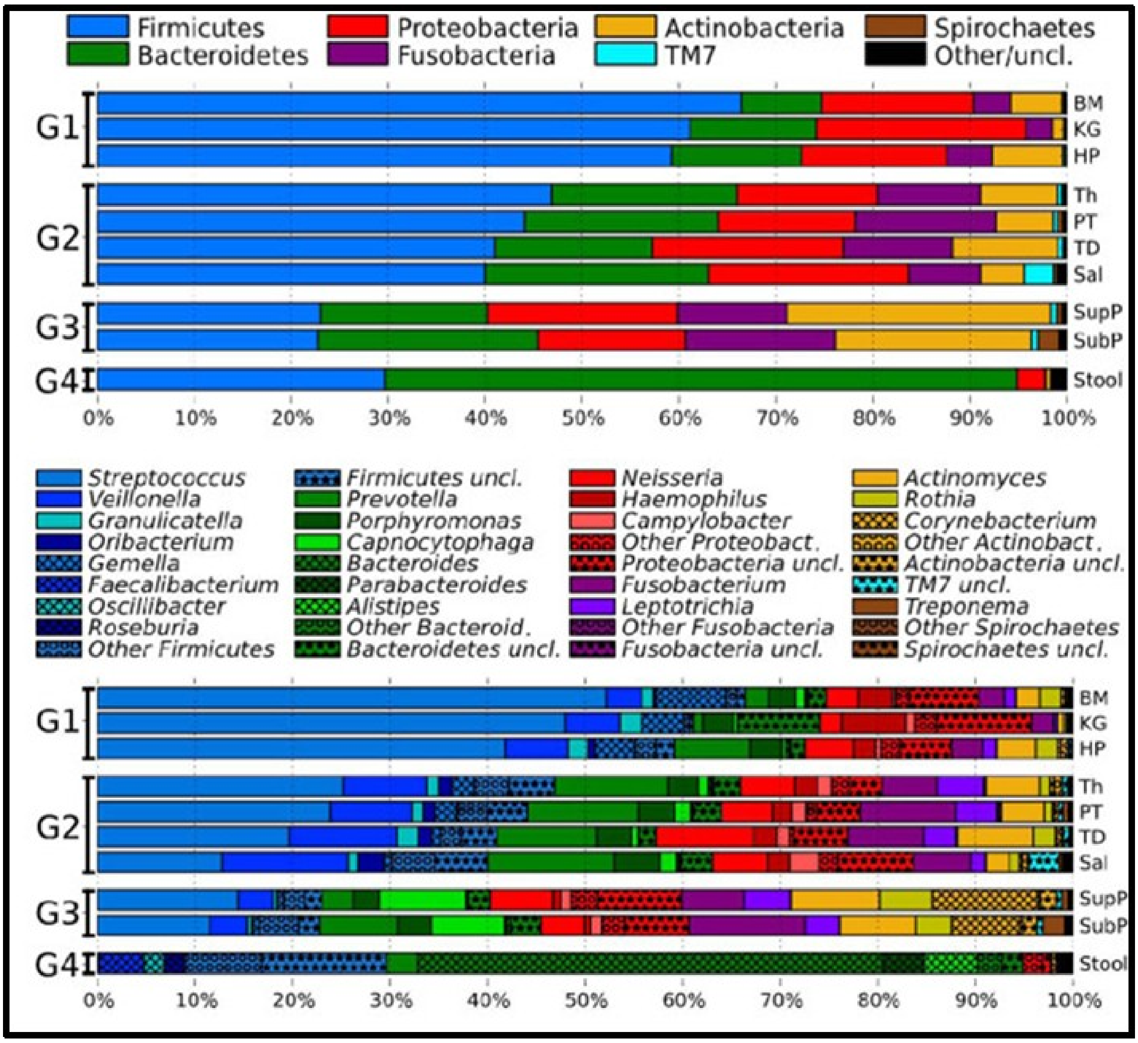


Figure 1 - 17. Groups detected in the sampled digestive tract microbiome sites based on similarities in microbial composition. Taxonomic composition of the microbiota in the ten digestive tract body habitats investigated based on average relative abundance of 16S rRNA pyrosequencing reads assigned to phylum (upper chart) and genus (lower chart). Microbiota from the ten habitats are grouped based on the ratio of Firmicutes to Bacteroidetes as follows: Group 1 (G1), buccal mucosa (BM), keratinized gingiva (KG) and hard palate (HP); Group 2 (G2), throat (Th), palatine tonsils (PT), tongue dorsum (TD) and saliva (Sal); Group 3 (G3), supraginval (SupP) and subgingival plaques (SubP); and Group 4 (G4), stool (Stool). Labels indicate genera at average relative abundance ≥2% in at least one body site. The remaining genera were binned together in each phylum as 'other' along with the fraction of reads that could not be assigned at the genus level as 'unclassified' (uncl.).

Several previous studies have observed different, and in some occurrences also opposite, results, with a preponderance of Firmicutes in respect to Bacteroidetes (Eckburg *et al.*, 2005; Turnbaugh *et al.*, 2009). These discrepancies could be explained by differences in geographic location, host genetics, or sample treatment protocols, but surely further studies are needed to throw light on. As mentioned above, this study confirms that the microbial composition of any specific site in the adult digestive tract shows extremely specific features even in absence of disease, and these differences reach to the genus level.

However, although microbial composition obtainable by analyses of stool samples may differ extensively from that achieved using colonic biopsies, the stool sample is preferable to the other because it is very easy to collect, due to the absolutely absence of risk to donor, and it is associated to a huge amount of information still obtainable.

1.5 How to exploit information on gut-microbiome?

The information about our microbial component, in which we can differ immensely, will be key in order to obtain truly personalized medicine, based not on the human genome, in which we are all 99.9% the same. Primarily the intestinal microbiome can be an optimal candidate for therapeutic microbial manipulation due to the wide possibility of interference, also with simple procedures such as adding probiotic microorganisms in food, or using specific antibiotic molecules, or modulating sources of nutrition according to different microbial response. Certainly, more information are necessary first about healthy microbiome and then about correlation between microbial changing and disease. These information can then be used to control and eventually prevent pathology or, for farm animals to improve their productivity. An example of this kind of intervention can be obtained by treatment of *Clostridium difficile* infection (CDI). CDI has emerged as a common complication of antibiotic usage. In the less severe cases of CDI, it can be sufficient the interruption of antibiotics assumption, but in more severe illness, with continuous diarrhea, fever, toxic megacolon, the final outgoing can be the death of patient (Kelly *et al.*, 2012). The recommended therapies for CDI include orally somministration of antibiotic metronidazole or vancomycin or other antibiotics, although less used, such as rifaximin, nitazoxanide or fidaxomicin. Although the use of antibiotics to treat a condition which was originally caused by the use of antibiotics seems counterintuitive, such therapies produce a clinical response for more than 80% of patients. Unfortunately a problem encountered in the treatment of CID is its recurrence, that can reach frequency between 20 and 30% of total affected patients unable to completely clear the infection. Recurrences of CDI (RCDI) are generally treated successfully changing antibiotic therapy. However some patients develop a chronic infection. In addition, a patient that shows recurrence once has more probabilities to show it other times. Owing to this significant failure rate of conventional antibiotic treatments and the lack of alternative therapies that have proven to be highly successful, clinicians have resorted to fecal microbiota transplantation (or fecal microbiome therapy, FMT). This kind of treatment, involving administration of fecal material from a healthy donor, usually a close relative of the patient, has as the mainly objective the recolonization with lost components of normal intestinal flora rather than specific eradication of the pathogen using a conventional antibiotic (Kelly *et al.*, 2012; Koenigsknecht and Young, 2013; Shahinas *et al.*, 2012). Even though the global experience with FMT is still limited, there is a growing number of scientific reports (about 200 cases reported in the world literature) on the efficacy of FMT for the treatment of RCDI with cure rates as high as 100% and a mean cure rate of 96% (Kelly *et al.*, 2012).

1.6 Non-human microbiomes

Although microbiomes have been studied mainly in association with human host, microbial communities correlated to other animal hosts have been also investigated. To date, a wide variety of animals have been characterized at various depth of analysis, such as vertebrates and invertebrates, domestic and wild animals. To give some examples, the microbiomes of carnivorous sponge, insect herbivore, ant, giant rodent, termites, and giant panda have been studied (Dupont *et al.*, 2013; Fang *et al.*, 2012; García-Amado *et al.*, 2012; Nakamura *et al.*, 2009; Poulsen and Sapountzis, 2012; Suen *et al.*, 2010; Warnecke *et al.*, 2007; Zhang *et al.*, 2012; Zhu *et al.*, 2011). Some of these studies were of pivotal importance to understand animal physiology, especially for those manifesting distinguishing features. For instance, the characterization of giant panda gut microbiome was key to explain its dietary oddities. This animal, indeed, consumes about 12.5 kg of bamboo each day, and, although possessing a gastrointestinal tract typical of carnivores, it can digest about 17% of dry matter consumed. Based on the panda's gut features, digestion of cellulose and hemicellulose contained in bamboo is impossible; consequently, this process must be dependent on microorganisms presents in panda gut. In support to this considerations, 16S rDNA analysis have shown that various microbial species present into the gut microbiota are able to digest cellulose, while metagenomic analysis has identified laccases, regarded as lignolysis-related enzymes, which may have positive roles in facilitating the breakdown of lignin and bamboo digestion (Fang *et al.*, 2012; Zhu *et al.*, 2011). Obviously, this kind of analyses was useful also to increase our knowledge about several physiological processes of domestic animals. In this regard, three independent research groups reviewed several canine microbiomes, respectively oral, gastrointestinal, and skin microbiome, at various depth of analysis, whereas Suchodoski and colleagues and Gnanandarajah and colleagues compared intestinal canine microbiome with acute diarrhea and idiopathic inflammatory bowel disease, and with calcium oxalate stones, respectively (Dewhirst *et al.*, 2012; Gnanandarajah *et al.*, 2012; Hooda *et al.*, 2012; Suchodolski *et al.*, 2012; Weese, 2013). Moreover, also farm animals were subjected to extensive investigations, principally to identify microbiome manipulation methods to increase productivity of each animal.

In Sardinia, sheep (and, at lower extent, goat) breeding has always been an activity of great economic importance, both for slaughtering purposes and for the production of milk and dairy products. In the first case, as shown in Table 1-1, in 2012 Sardinia was placed as second among Italian regions for ovine meat production, with more than 1.2 million units (about 20% of the overall Italian production) and over 8,500 tons of meat (about 15% of the overall Italian production). As far as the dairy industry is concerned, Sardinia plays a primary role in the Italian ovine and caprine sectors. As a matter of fact, it holds nearly half of sheep designed to milk production of the entire Italian territory (nearly 3 million ewes) and of about one-fifth of the nanny-goats (over 170,000 units), thus being by far the region with the highest production of sheep and goat milk, (64% and 34% of the overall Italian production, respectively) (Figure 1-18).

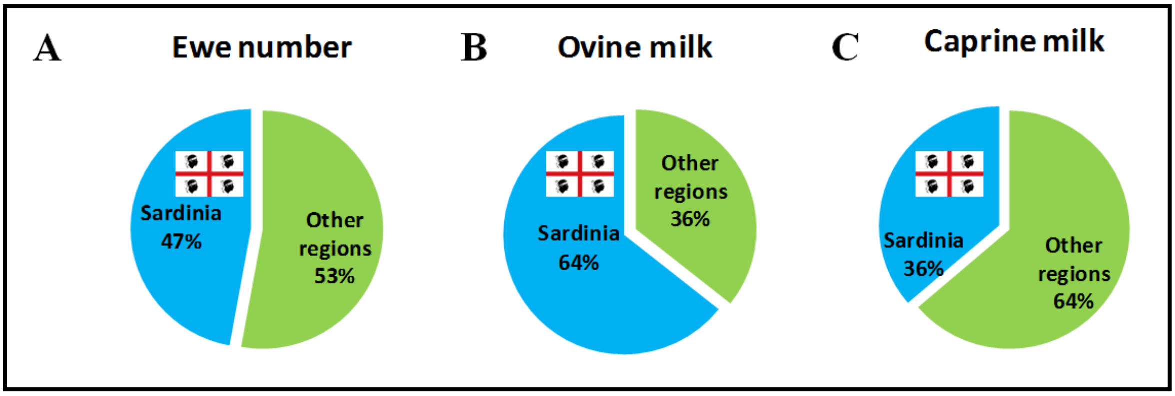


Figure 1 - 18. Comparison of ewe diffusion and milk production between Sardinia and the rest of Italy.

Table 1 - 1. Distribution of the dairy and slaughter areas among Italian regions.

Regions	Dairy area				Slaughter area	
	Units		Milk production (metric ton)		Units	Meat weight (metric ton)
	Ewe	Goat	Ovine	Caprine		
Abruzzi	190,472	17,399	36,09.3	1.4	524,284	6,443.2
Aosta Valley	1,951	3,574	-	370.6	1650	20
Apulia	213,797	30,700	3,912.4	583.8	790,065	7,294.8
Basilicata	229,939	48,362	91	423.7	294,864	2556
Calabria	229,274	105,864	4,389.5	91.5	152,368	1,179.5
Campania	165,847	21,999	1,783.5	101.7	374,043	3,244.8
Emilia-Romagna	58,819	13,368	1,351.1	127	11,559	164.5
Friuli-Venezia Giulia	8,458	2,040	-	91.3	4,450	54.2
Latium	605,873	33,119	39,763.2	2,561	1,477,977	1,3817.6
Liguria	16,311	13,346	-	236.9	4,112	33.1
Lombardy	57,267	56,853	305.6	5,097.2	39,464	526.5
Molise	65,497	6,659	-	-	122,799	1,225.8
Piedmont	77,335	58,598	1,884.1	4,308.9	45,649	404.5
Sardinia	**2,968,306**	**176,249**	**261,224.9**	**10,139**	**1,236,609**	**8,570.7**
Sicily	657,104	107,810	17,517.3	964.3	244,132	2,490.7
the Marches	119,535	4,400	2,675.4	-	146,313	1,354.5
Trentino-Alto Adige	60,404	12,161	-	1376.4	24,716	240
Tuscany	447,556	12,106	62,779.1	79.9	348,013	3,188.8
Umbria	92,868	4,752	4,518.8	41.2	129,575	1,382.1
Veneto	30,088	5,595	372.1	1,348	15,277	152.6
ITALY	**6,296,701**	**734,954**	**406,177.3**	**27,943.8**	**5,987,919**	**54,343.9**

Therefore, the achievement of further significant information concerning the physiology of these animals has a crucial economic relevance. In particular, the investigation on how the microbial component can correlate with variation in response to different farming methods could lead to the implementation of strategies ensuring higher productivity levels.

Except for studies aimed to find a cause-effect relationship with a specific disease, the microbial component of sheep has been poorly analyzed so far. The most investigated biological site has been the rumen, as in other ruminants, for two important reasons. The first one is its role in environmental pollution due to methane production, and the second one is the possibility of obtaining ideal candidates for industrial applications connected to the microbial ability to break down lignocelluloses.

Lignocellulosic biomass, the most abundant renewable polymer in nature, is made up of approximately 40% cellulose, and 20-30% both hemicelluloses and lignin (Khandeparker and Numan, 2008). The complex network formed by these two elements, hemicelluloses and lignin, allows digestibility of the cellulose only in ruminant mammalian animals. Rumen, in fact, represents a natural lignocelluloses-degrading system, where microorganisms inside have been found to produce enzymes able to digest lignocellulosic biomass (Facchini *et al.*, 2012; Yue *et al.*, 2013). In order to efficiently hydrolyze lignocelluloses, rumen microbes take advantage of the synergistic action of cellulase, hemicellulase and ligninolytic enzymes (Sun and Cheng, 2002). Lignocellulose can be converted to various energy products such as ethanol, butanol, iso-butanol, hydrogen, methane, and volatile fatty acids, through physicochemical approaches or by biological processes (Hendriks and Zeeman, 2009; Thanakoses *et al.*, 2003). Although this ecosystem is one of the most interesting environments to screen for novel biocatalysts, 85% of its inhabiting species remain uncultured, mainly due to its complexity and the anaerobic nature of the environment (Gong *et al.*, 2013; Krause *et al.*, 2003). In this case, meta-omic technologies are ideally suited for overcome the limit of pure cultivation methods, and can therefore be used to find out genes with functionally relevant properties owned by uncultured natural microorganisms.

At the time of this writing, this approach has been utilized to identify several novel enzymes, for example a new xylanase, named Xyn10N18, extracted from the rumen contents of a dairy cow, a bifunctional xylanase/endoglucanase, and two β-glucosidase/xylosidase enzymes from yak rumen metagenome, named RuCelA, RuBG3A and RuBG3B respectively (Bao *et al.*, 2012; Chang *et al.*, 2011; Gong *et al.*, 2013). All these enzymes have an important value due to their potential application in various industrial processes, such as textile, paper, food, animal feed, biobleaching and biofuel production (Collins *et al.*, 2005; Menon *et al.*, 2010; Polizeli *et al.*, 2005).

Ruminant animals have been deeply studied above all due to their involvements in greenhouse gas (GHG) production. In fact, with around 81-92 million tons of methane (end product of rumen fermentation during digestion) excreted per year, ruminants have been indicated as one of the greatest sources of anthropogenic emissions (GHG has been estimated to range between 9% and 18% of total). Taking into account this considerations, reducing methane emission has, primarily, crucial implications for global environmental protection. In addition, this process represents also an important loss of carbon and energy, accounting for 8 to 12% of the gross energy content of the animal diet; therefore, reducing methane formation can be useful to increase efficient animal production (Wang *et al.*, 2012). It has been amply demonstrated that methane production takes place in rumen with methanogenic Archaea as the master controller of the process, in collaboration with hydrolytic and fermentative microorganisms that, with degradation of organic matter, make hydrogen available. Nevertheless, the precise mechanisms underlying lignocellulose degradation are not yet fully understood. Some rumen cellulolytic bacterial species have been identified essentially using PCR assays. The most important are *Fibrobacter succinogenes*, *Ruminococcus albus*, *Ruminococcus flavefaciens*, (Koike and Kobayashi, 2001) and *Cellulosilyticum ruminicola* (Cai and Dong, 2010). Various strategies have been suggested in order to mitigate ruminant methane production, including the application of nutritional supplement, the manipulation of ruminal fermentation by changing feed composition, the addition of methane inhibitors as unsatured fatty acids (Sutton *et al.*, 1983) and the defaunation, i.e. the complete removal of hydrogen producers protozoa (Shibata and Terada, 2010). All these treatments modify rumen digestive features decreasing, or in some cases completely suppressing, methane production, and modulate rumen fermentation pattern frequently decreasing fiber digestion, that is the most commonly reported negative effect (Mosoni *et al.*, 2011; Wang *et al.*, 2012). A meta-omic approach (both metagenomics and metaproteomics) able to study globally rumen microbiome composition and response to external manipulation could enable to give clear insights about this issue.

To date, most of the known information has been obtained through methods classified as "classical", such as PCR and culturing, while “-omic” approaches in sheep have been less used. As was the case for other animals, mostly horses and cattle, the first step in this kind of investigation consists in studying the microbiomes features under physiological conditions (Costa and Weese, 2012; O’ Donnell *et al.*, 2013).

Considering the great invasiveness and complexity of the technical procedures needed to study rumen, different approaches are to be preferred. To give same examples, in 2012, Khianngam and colleagues isolated from buffalo faeces a new cellulose-degrading bacterium species belonging to genus *Cohnella*, for which was proposed the name *Cohnella cellulosilytica* sp. nov. (Khianngam *et al.*, 2012). In 2011 Calvo-Bado *et al.* characterized the ovine “pedomics”, i.e. the bacterial microbiome of ovine interdigital skin, using 16S rDNA by pyrosequencing and conventional cloning with Sanger-sequencing. The study reported Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria as the most abundant phyla and underlined an association between *Peptostreptococcus*, *Corynebacterium* and *Staphylococcus* genera and healthy interdigital skin, interdigital dermatitis, and virulent footrot, respectively (Calvo-Bado *et al.*, 2011).

Chapter 2

Aim of the Project

In keeping with the considerations outlined in the introduction, the main objective of this project was the development of a workflow for in-depth metaproteome characterization.

Accordingly, the following secondary aims were established:

- to critically evaluate the applicability of a multi-step workflow based on bead-beating/freeze-thawing, FASP, and single-run nanoLC-MS/MS for the metaproteomic analysis of microbial community samples;
- to test efficiency, reproducibility, sensitivity and dynamic range of the workflow by using lab-assembled, heterogeneous microbial mixtures, composed by bacterial and eukaryotic species in different proportions;
- to assess the impact of different protein databases on the metaproteomic investigation using mock microbial mixtures, specifically generating metagenomic- and genomic-derived DBs to compare them with publicly available DBs;
- to validate the workflow using a murine fecal sample, by testing its ability to provide reliable, in-depth taxonomic and functional information from complex microbiomes;
- to apply the optimized workflow to an animal sample of biotechnological interest, such as ovine stool, in order to characterize the gut metaproteome of healthy Sarda sheep.

Chapter 3

Materials and Methods

3.1 Samples

3.1.1 Microbial samples

Identity and features of the microbial strains used in this project are described in detail in Table 3-1. *Pasteurella multocida* was kindly provided by Dr. Gavino Marogna (Istituto Zooprofilattico Sperimentale della Sardegna), *Rhodotorula glutinis* by Prof. Ilaria Mannazzu (Department of Agricultural Sciences, University of Sassari), *Lactobacillus casei*, *Lactobacillus acidophilus*, *Pediococcus pentosaceus* and *Saccharomyces cerevisiae* by Dr. Pasquale Catzeddu and Dr. Manuela Sanna (Porto Conte Ricerche, Alghero), *Brevibacillus laterosporus* by Dr. Luca Ruiu (Bioecopest Srl, Alghero), whereas *Enterococcus faecalis* and *Escherichia coli* were available in the bacterial collection of the Department of Biomedical Sciences, University of Sassari. The microorganisms were seven bacterial strains and two yeasts, exhibiting wide differences both in terms of structural features and of reference sequences availability.

At the time when this study was performed, none of the specific microbial strains listed in Table 3-1 had its genome sequenced and deposited, except *B. laterosporus*.

Microbial cultures were grown to stationary phase using the appropriate standard medium and the most proper conditions for each microorganism. After an overnight culture, colony-forming unit (CFU) counting was performed to estimate the amount of viable microbial cells. The microbial cultures were then divided into aliquots, which were pelleted, washed three times in phosphate-buffered saline (PBS, pH 7.2) to eliminate medium residues, and stored as pellets at -80°C until use.

Table 3 - 1. Features of microorganism used in this project to assemble microbial mixtures.

Species	Cell type	Source	Genome size	Abbreviation
Brevibacillus laterosporus	Gram-variable bacillus	LMG 15441	5,180 Kb	*Blat*
Enterococcus faecalis	Gram-positive coccus	Field isolate	3,218 Kb	*Efae*
Escherichia coli	Gram-negative bacillus	Field isolate	4,600 Kb	*Ecol*
Lactobacillus acidophilus	Gram-positive bacillus	LMG 9433	1,993 Kb	*Laci*
Lactobacillus casei	Gram-positive bacillus	LMG 6904	2,900 Kb	*Lcas*
Pasteurella multocida	Gram-negative coccobacillus	Field isolate	2,250 Kb	*Pmul*
Pediococcus pentosaceus	Gram-positive coccus	Field isolate	1,832 Kb	*Ppen*
Rhodotorula glutinis	Yeast	Field isolate	20,300 Kb	*Rglu*
Saccharomyces cerevisiae	Yeast	CBS 1171	12,068 Kb	*Scer*

2.1.2 Fecal samples

The murine fecal sample (analyzed in Chapter 4) was kindly provided by Dr. Michael Silverman (Department of Microbiology and Immunobiology, Harvard Medical School, Boston, USA). The sample was collected from one female NOD mouse (38 weeks old, raised under standard condition), and stored at -80°C until use.

The ovine fecal samples, kindly provided by Dr. Gavino Marogna (Istituto Zooprofilattico Sperimentale della Sardegna), were collected from five Sarda sheep belonging to the same flock. The sheep were numbered from one to five. The animals were lactating females, free-grazing, and without evident clinical symptoms. Fecal samples were collected and stored at -80°C until use. The results obtained analyzing such samples are illustrated in Chapter 6.

3.2 Microbial mixture assembling

A nine-organism microbial mixture (9MM), composed of all microorganisms listed in Table 3-1, was assembled as follows. The first microbial pellet was resuspended in 500 µl of pre-heated (95°C) extraction buffer containing 2% sodium dodecyl sulphate (SDS), and 20 mM Tris-HCl pH 8.8. After careful pipetting, the microbial suspension was added to a second microbial pellet, and the procedure was sequentially repeated until the ninth pellet was resuspended and mixed. Each microbial pellet used contained 1 ml of overnight culture (approximately between 10^8 and 10^{10} cells).

Four bacteria (namely *P.multocida, E. coli, L. acidophilus, and E. faecalis*) were selected for further analyses. After overnight culture, each bacterial strain was subjected to accurate CFU counting, divided into 3 aliquots (corresponding to 10^{10}, 10^8 and 10^6 CFUs, respectively), pelleted, washed three times in PBS, dried, and stored at -80°C until use. A four-organism microbial mixture (4MM) was then assembled by merging a pellet corresponding to 10^{10} CFUs of *E. faecalis*, two pellets corresponding to 10^8 CFUs of *P.multocida* and *E. coli*, respectively, and a pellet corresponding to 10^6 CFUs of *L. acidophilus*. All four pellet were suspended using the same extraction buffer and the same procedure described above for 9MM.

3.3 Differential centrifugation

Two different approaches were used in this work to treat fecal samples (Figure 3-1). According to the first approach (called "homogenization"), used for protein extraction from ovine samples, feces were simply homogenized in extraction buffer (overall 1:2 w/v sample-to-buffer ratio) and directly subjected to protein extraction according to the method B detailed in the "Protein extraction" section in this chapter. According to the second approach, called "differential centrifugation" (Apajalahti *et al.*, 1998), feces were pre-processed before murine protein extraction. Briefly, after thawing at 4°C, fecal samples (approximately. 100 mg each) were resuspended in 10 ml of PBS, vortexed, shaken in a tube rotator for 45 minutes, and subjected to low-speed centrifugation at 500 x g for 5 minutes to eliminate gross particulate material; the supernatants were carefully transferred to clean polycarbonate centrifuge bottles (Beckman Coulter, Brea, CA, USA) and kept at 4°C, whereas the pellets were suspended again in PBS. The entire procedure was repeated for a total of three rounds. Then, the three supernatants obtained from each sample were centrifuged at 20,000 x g for 15 minutes, and the three derivative pellets were pooled after resuspension with the extraction buffer described below, and subjected to protein extraction as detailed in the "Protein extraction" section in this chapter.

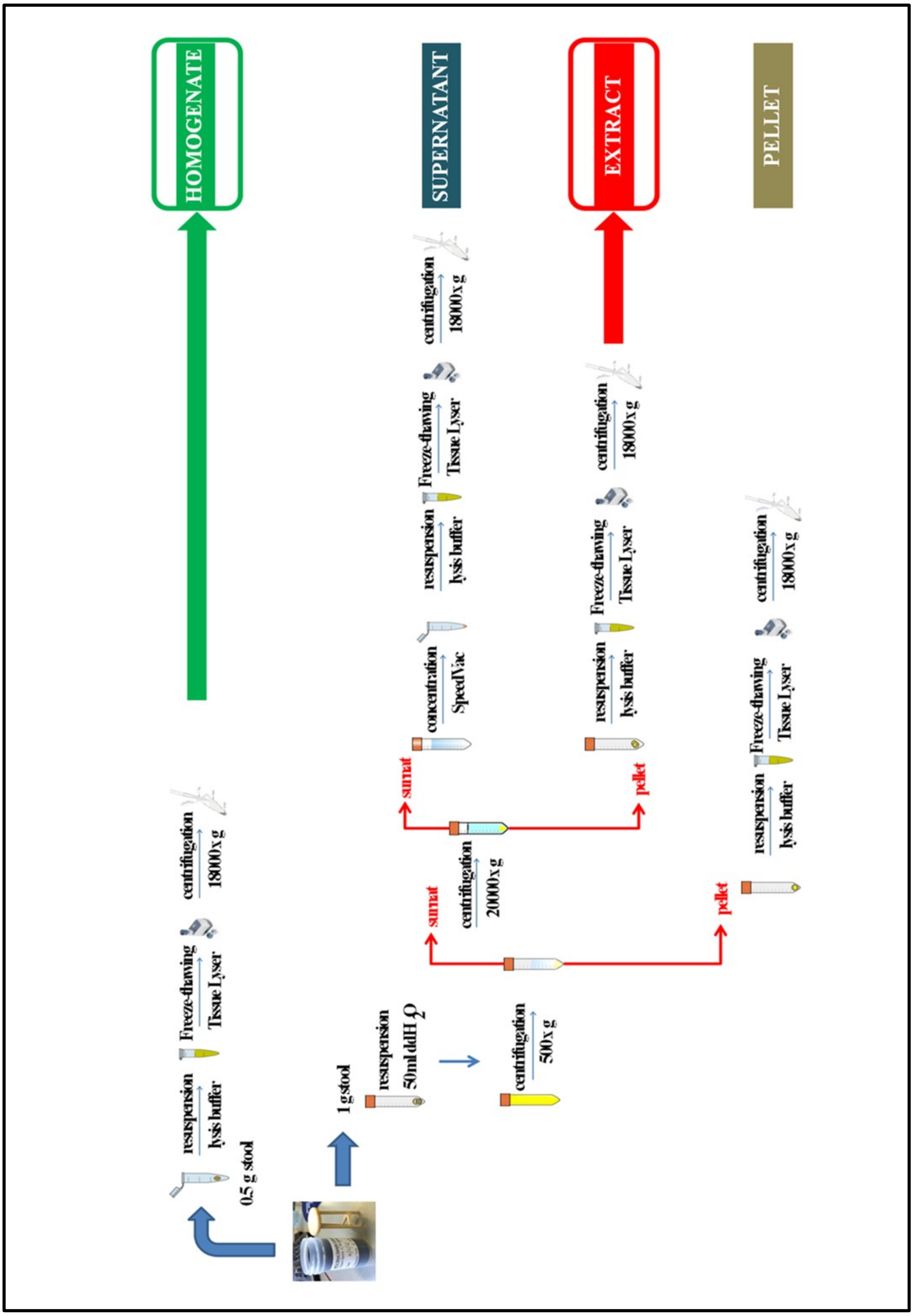


Figure 3 - 1. Fecal sample pre-processing workflow. After thawing at 4°C, the fecal samples were treated with different protocols: "homogenization" consisting of a direct resuspension of each sample with extraction buffer and differential centrifugation to enrich in microbial cells (Apajalahti *et al.* 1998).

3.4 Protein extraction

Proteins were extracted from single microbial pellets according to two different methods.

- A first pellet of each microorganism underwent high-temperature (Method A) extraction by incubation in 100 µl of extraction buffer at 95°C for 20 minutes in agitation (500 rpm) in a Thermomixer Comfort (Eppendorf, Hamburg, Germany), followed by centrifugation at 20,000 x g for 10 minutes at 4°C and collection of the protein containing supernatant.
- A second pellet of each microbe was subjected to high-temperature extraction followed by bead-beating (Method B). After the high temperature incubation described above as first step of Method A, a stainless steel bead (5 mm diameter, Qiagen, Hilden, Germany) was added to each sample. Samples were sequentially incubated at -80°C for 10 minutes, subjected to bead beating for 10 minutes (30 cycles/s in a TissueLyser LT mechanical homogenizer, Qiagen), incubated at -80°C for 10 minutes and then at 95°C for 10 minutes, and subjected to a further 10 minutes bead beating step. Finally, sample was centrifuged at 20,000 x g for 10 minutes at 4°C and the whole supernatant was collected.

The 9MM, 4MM, and fecal samples, both murine and ovine, were processed only according to the Method B.

3.5 Proteins and peptides quantification

Protein quantification was carried out by means of the 2-D Quant Kit (GE Healthcare, Little Chalfont, UK) following manufacturer's instructions.

In addition, 4 µl of each protein extract from single microbial organisms were separated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) using 10% polyacrilamide gels, which were stained with SimplyBlue SafeStain (Invitrogen, Carlsbad, CA, USA). Five microliters of Precision Plus Protein All Blue Standards (Bio-Rad, Life Science Research, Hercules, California, USA) for each gel was used as molecular weight marker.

Peptide mixture concentration was estimated by measuring absorbance at 280 nm with a NanoDrop 2000 spectrophotometer (Thermo Scientific, San Jose, CA, USA), using dilutions of the MassPREP *E. Coli* Digest Standard (Waters, Milford, MA, USA) to generate a calibration curve.

All measurements were carried out in three technical replicates.

3.6 Filter-aided sample preparation (FASP)

SDS protein extracts were diluted to 200 µl with UA solution (8M urea in 100 mM Tris-HCl, pH 8.8), loaded into the Microcon Ultracel YM-30 filtration devices (Millipore, now Merck Millipore, Billerica, MA, USA), and then processed according to filter-aided sample preparation (FASP) method, more in detail using the "FASP II" protocol (Liebler and Ham, 2009; Wiśniewski *et al.*, 2009; Tanca *et al.*, 2013), with minor modifications. Briefly, samples were centrifuged at 14,000 x *g* for 15 minutes, and the concentrates were diluted into the filter with 200 µl of UA solution and centrifuged again. After centrifugation, the concentrates were mixed with 100 µl of 10 mM dithiothreitol (DTT) in UA solution and incubated at 25°C for 30 minutes. After centrifugation, the concentrates were mixed with 100 µl of 50 mM iodacetamide (IAM) in UA solution and incubated at 20°C for 20 minutes. Following centrifugation, the concentrate was diluted with 100 µl of UA solution and concentrated again (this step was repeated twice). Next, the concentrate was diluted with 100 µl of 50 mM ammonium bicarbonate (ABC) and concentrated again. This step was repeated once. Subsequently, 40 µl of trypsin solution (150 ng in 50 mM ABC) were added to the filter, and the samples were incubated at 37°C overnight. Peptides were collected by centrifugation of the filter units, followed by an additional 50 µl wash with a solution containing 70% acetonitrile and 1% formic acid. Finally, the peptide mixture was brought to dryness and reconstituted in 0.2% formic acid to a final concentration of 1 mg/ml.

3.7 DNA extraction

DNA of single bacterial species was extracted according to a procedure hereafter called method M (Mild), based on detergent lysis and lysozyme treatment according to the DNeasy Blood & Tissue Kit protocol (Qiagen, Hilden, Germany), whereas yeast DNA was extracted according to a procedure hereafter called method H (Harsh), comprising a strong detergent pretreatment combined with freeze-thawing and bead beating steps (as previously described by Harju and coworkers (Harju *et al.*, 2004) followed by the Gentra Puregene kit protocol (Qiagen, Hilden, Germany). Furthermore, two identical replicates of the 9MM were assembled by merging 1 ml overnight culture cell pellets from the nine microorganisms mentioned above. Then, the first 9MM replicate was subjected to extraction according to method M, while the second according to method H, therefore producing two different 9MM extracts (called 9MM-M and 9MM-H respectively).

The extracted DNA was quantified using the Nanodrop 2000 (Thermo Scientific, Waltham, MA, USA), and quality was assessed by agarose gel electrophoresis.

All measurements were carried out in three technical replicates.

3.8 DNA sequencing

The 11 DNA extracts (9 individual microbes, 9MM-M and 9MM-H) were then subjected to next generation sequencing (NGS). Libraries were generated using the Illumina® TruSeq™ DNA Sample Preparation Kit (San Diego, CA, USA) according to the manufacturer's protocol with minor modifications. Briefly, genomic DNA was fragmented in an ultrasonic bath (Elmasonic S, Elma, Singen, Germany). After ligation to the adapters and gel purification of DNA ranging between 300 and 400 bps, the libraries were subjected to 15-20 polymerase chain reaction (PCR) cycles to enrich the DNA fragments with adapters ligated to both ends. The PCR products were purified and evaluated using the High Sensitivity DNA chip on an Agilent Technologies 2100 Bioanalyzer (Santa Clara, CA, USA). Normalized sample libraries were pooled and subjected to hybridization and cluster generation step on a v1 flow cell using the cBOT cluster generation station, according to the Illumina TruSeq PairedEnd Cluster Kit protocol. Libraries were sequenced (six samples per lane) with an expected coverage of at least 40X for each single microorganism except for *R. glutinis* (about 12X). The 9MM extracts were sequenced with a higher coverage (only two samples per lane) to achieve a better sequencing depth. DNA sequencing was performed with the Illumina HiScanSQ sequencer, using the paired-end method and 76 runs of sequencing.

After sequencing, all reads were subjected to a multiplexing step using Casava software version 1.8 implemented in the Illumina HiScanSQ sequencer.

3.9 Genome/metagenome gene finding, annotation and six-frame translation

Reads were assembled *de novo* into scaffolds using Velvet 1.2 (Zerbino and Birney, 2008), choosing the best K-mer values for each assembly to obtain nine genome drafts and two metagenome drafts. As detailed in Table 3-2, all the *de novo* drafts of the single microorganisms showed a N50 length >30 kbps and a coverage higher than 39X, except for *R. glutinis* (4910 bps and 12.6X, respectively, probably due to its wider genome); 9MM-M metagenome reads showed an assembly quality equivalent to the single genome sequences, whereas N50 length of the 9MM-H draft was significantly lower (< 1000).

Table 3 - 2. Characteristics of the drafts genomes upon individual sequencing of the nine microorganisms and metagenome sequencing of the 9MM. The median coverage depth was calculated by Velvet from the number of reads aligned against each contig.

Species	Number of contigs	N50 length (bp)	Total bps	Median coverage depth
B. laterosporus	409	119780	4862422	39.5
E. coli	324	131689	4670100	49.5
E. faecalis	241	178567	3006944	82.4
L. acidophilus	104	167344	1952696	111
L. casei	782	30833	2758628	73.4
P. multocida	350	55083	2286223	98.4
P. pentosaceus	168	282570	1826383	130.3
R. glutinis	8616	4910	17404538	12.6
S. cerevisiae	3521	59022	11464721	53.6
9MM-M	3029	49888	8543469	9.2
9MM-H	13811	724	5737764	7.8

The putative coding sequences (CDS) were identified with Prodigal 2.60 (Hyatt *et al.*, 2010). Each CDS was annotated evaluating the homology by BLAST search against TrEMBL Protein Database Release 2012_10 (E-value ≤ 10-8) (Altschul *et al.*, 1997).

Furthermore, each genome draft was translated in all six frames using the following perl script: translateWholeGenomeMultiChromosome.pl, freely available at http://proteomics.ucsd.edu/Downloads/.

3.10 LC-MS/MS analysis

MS analysis was carried out using an LTQ-Orbitrap Velos (Thermo Scientific, San Jose, CA, USA) interfaced with an UltiMate 3000 RSLCnano LC system (Dionex, Sunnyvale, CA, USA, now part of Thermo Scientific). After loading, 4 µg of each peptide mixtures were concentrated and desalted on a trapping pre-column (Acclaim PepMap C18, 75 µm × 2 cm nanoViper, 3 µm, 100 Å, Dionex), using 0.2% formic acid at a flow rate of 5 µl/min. The peptide separation was carried out at 35 °C using a C18 column (Acclaim PepMap RSLC C18, 75 µm x 15 cm nanoViper, 2 µm, 100 Å, Dionex) at a flow rate of 300 nL/min, subjecting the peptide mixtures to 305 or 510 minutes runs (280 or 485 minutes gradient from 1 to 50% eluent B in eluent A, where B is a solution 0.2% formic acid in 95% ACN, and A is 0.2% formic acid in 5% ACN). The mass spectrometer LTQ-Orbitrap Velos was set up in a data dependent MS/MS mode under direct control of the Xcalibur software (version 1.0.2.65 SP2), where a full-scan spectrum (from 300 to 1,700 m/z) was followed by tandem mass spectra (MS/MS). The instrument was operated in positive mode with a spray voltage of 1.2 kV, a capillary temperature of 275°C, and was calibrated before measurements. Full-scans were performed in the Orbitrap with resolution of 30,000 at 400 m/z, the automatic gain control was set to 1,000,000 ions and the lock mass option was enabled on a protonated polydimethylcyclosiloxane background ion ($(Si(CH_3)_2O)_6$; m/z = 445.120025) as internal recalibration for accurate mass measurements (Olsen *et al.*, 2010). Peptide ions were selected as the ten most intense peaks (Top 10) of the previous scan. The signal threshold for triggering an MS/MS event was set to 500 counts. Higher Energy Collisional Dissociation (HCD), performed at the far side of the C-trap, was chosen as the fragmentation method, by applying a 40% value for normalized collision energy, an isolation width of m/z 3.0, a Q-value of 0.25, and an activation time of 0.1 ms. Nitrogen was used as the collision gas.

3.11 Protein database construction

Thirteen protein databases (DBs) were used for protein/peptide identification from MS data to obtain results described in Chapter 5. The first nine DBs (Table 3-3) were assembled starting from publicly available sequences derived from NCBI, UniProtKB/SwissProt (hereafter simply called SwissProt), and UniProtKB/TrEMBL (hereafter simply called TrEMBL) records, using the Database Manager tool included in Mascot Server (version 2.4, Matrix Science, London, UK), and applying one of the three following taxonomy filters: Bacteria, Fungi, Viruses (BFV, corresponding to NCBI taxonomy IDs 2, 4751, and 10239), selected genera (*Brevibacillus*, *Escherichia*, *Enterococcus*, *Lactobacillus*, *Pasteurella*, *Pediococcus*, *Rhodotorula*, and *Saccharomyces*, corresponding to NCBI taxonomy IDs 55080, 561, 1350, 1578, 745, 1253, 5533, and 4930), or selected species (*B. laterosporus*, *E. coli*, *E. faecalis*, *L. acidophilus*, *L. casei* group, *P. multocida*, *P. pentosaceus*, *R. glutinis*, *S. cerevisiae*, corresponding to NCBI taxonomy IDs 1465, 562, 1351, 1579, 655183, 747, 1255, 5535, 4932).

Table 3 - 3. Public databases used for peptide identification from MS spectra. The databases were assembled starting from publicly available sequences derived from NCBI, UniProtKB/SwissProt, and UniProtKB/TrEMBL records, applying, one of the three following taxonomy filters: Bacteria, Fungi, Viruses (BFV), selected genera (*Brevibacillus*, *Escherichia*, *Enterococcus*, *Lactobacillus*, *Pasteurella*, *Pediococcus*, *Rhodotorula*, and *Saccharomyces*), or selected species (*B. laterosporus*, *E. coli*, *E. faecalis*, *L. acidophilus*, *L. casei* group, *P. multocida*, *P. pentosaceus*, *R. glutinis*, *S. cerevisiae*).

Database acronym	*Original database*	*Update*	*Taxonomy*	*Number of sequences*	*Average computing time per run (min)*
NCBI-BFV	NCBI	2012_12	BFV	16,175,389	817
TrEMBL-BFV	UniProtKB/ TrEMBL	2012_10	BFV	21,602,141	1002
SP-BFV	UniProtKB/ Swiss-Prot	2012_11	BFV	375,700	28
NCBI-G	NCBI	2012_12	8 selected genera	895,743	213
NCBI-S	NCBI	2012_12	9 selected species	554,718	219
TrEMBL-G	UniProtKB/ TrEMBL	2012_10	8 selected genera	2,622,251	269
TrEMBL-S	UniProtKB/ TrEMBL	2012_10	9 selected species	2,198,849	247
SP-G	UniProtKB/ Swiss-Prot	2012_11	8 selected genera	37,708	9
SP-S	UniProtKB/ Swiss-Prot	2012_11	9 selected species	33,130	8

The taxonomy *L. casei* group was preferred to *L. casei* (species) due to the very high level of sequence similarity and some ambiguity in taxonomic boundaries within the species comprised in this taxonomic group.

The remaining four DBs were constructed from genomic and metagenomic data experimentally obtained in our study. Specifically, the single predicted and annotated (PA) genomes assembly DB (SGA-PA) was obtained by concatenating in a single FASTA file the protein sequences obtained from each individual microbe upon CDS prediction and TrEMBL annotation, while the PA metagenome DB (Meta-PA) was obtained by concatenating in a single FASTA file the protein sequences obtained upon NGS of the two 9MM extracts, CDS prediction and TrEMBL annotation. Finally, the genome drafts of the nine sequenced microbes and the 9MM metagenome draft were also processed in an alternative way based on naïve six-frame translation, thus generating SGA-6FT and Meta-6FT DBs, respectively (Table 3-4).

Table 3 - 4. Custom databases used for peptide identification from MS spectra. Databases were constructed from genomic and metagenomic data experimentally obtained in this study.

Database acronym	*Original database*	*Processing*	*Number of sequences*	*Average computing time per run (min)*
Meta-PA	Matched metagenome	CDS prediction + TrEMBL annotation	24,673	10
Meta-6FT	Matched metagenome	six-frame translation	90,306	17
SGA-PA	Single genomes assembly	CDS prediction + TrEMBL annotation	52,455	10
SGA-6FT	Single genomes assembly	six-frame translation	54,948	28

As expected, the number of amino acid residues of the 6FT DBs was almost six time bigger than that of the corresponding PA DBs (specifically, 4.2 million residues for Meta-PA versus 26.1 for Meta-6FT, and 12.9 million residues for SGA-PA versus 84.2 for SGA-6FT). Features and composition of the in-house Meta-PA and SGA-PA DBs were as follows (as reported more in detail in Table 3-5). The percentage of annotated proteins were 71% and 54% of the overall protein sequences, and the number of non-redundant protein sequences within each DB amounted to 13270 and 27164 for Meta-PA and SGA-PA, respectively.

Table 3 - 5. Features of the in-house databases. Detail of the sequences (indicated as absolute and relative number) attributed to each species present in our microbial mixture. Specifically, SGA-PA DB was obtained by concatenating in a single FASTA file the protein sequences achieved from each individual microbe upon CDS prediction and TrEMBL annotation, while Meta-PA DB was obtained by concatenating in a single FASTA file the protein sequences obtained upon NGS of the two 9MM extracts, CDS prediction and TrEMBL annotation.

Species	Number of protein Sequenze		Percentage of protein sequences	
	Meta-PA	SGA-PA	Meta-PA	SGA-PA
B. laterosporus	4714	4286	35.52%	15.78%
E. faecalis	142	2847	1.07%	10.48%
E. coli	54	3337	0.41%	12.28%
L. acidophilus	1132	2184	8.53%	8.04%
L. casei group	2942	1827	22.17%	6.73%
P. multocida	3567	2109	26.88%	7.76%
P. pentosaceus	22	1527	0.17%	5.62%
R. glutinis	69	860	0.52%	3.17%
S. cerevisiae	1	5687	0.01%	20.94%
Other species	301	2411	4.72%	9.20%

Among these, 95% and 91% were correctly attributed to the species actually present in the 9MM, respectively. Concerning the species distribution of the protein sequences contained into the two DBs according to TrEMBL annotations, in the Meta-PA DB over 90% of protein sequences were from only 4 species (*B. laterosporus*, *P .multocida*, *L. casei* group, and *L. acidophilus*, representing 36%, 27%, 22% and 9% of the total, respectively), with a significant depletion in yeast sequences (for instance, only 1 from *S. cerevisiae*), whereas in the SGA-PA DB the abundance of the 9 actually present species ranged from 3 to 21% of the overall protein sequences.

Finally, a specific DB containing common contaminants (available at http://maxquant.org/contaminants.zip) was also used as a control for environmental and trypsin contamination.

In the case of murine and ovine fecal samples (results treated in Chapters 4 and 6, respectively), a slightly modified version of the "two-step" method, recently presented by Jagtap and coworkers, was applied (Jagtap *et al*., 2013). Briefly, MS spectra were first searched against a large database (UniProtKB) without establishing a FDR-based threshold. For each sample, the protein entries identified in the first search (in one or more replicates) were then used to create a smaller database (the DB composition is detailed in Table 3-6), which was employed to perform a second search with the FDR threshold set to 1%.

Table 3 - 6. Restricted databases used to analyze murine and ovine fecal samples. MS spectra were first searched against a large database (UniProtKB) without establishing a false discoverer rate (FDR)-based threshold. For each sample, the protein entries identified in the first search (in one or more replicates) were then used to create a smaller database according to the "two step" method presented by Jagtap and coworkers (Jagtap *et al*., 2013).

Database	Number of protein sequences
Sheep 1 (Obtained from one instrumental replicate)	65,345
Sheep 2 (Obtained from one instrumental replicate)	66,471
Sheep 3 (Obtained from one instrumental replicate)	62,274
Sheep 4 (Obtained from one instrumental replicate)	65,708
Sheep 5 (Obtained from one instrumental replicate)	57,887
MFM A (Obtained from two instrumental replicates)	107,842
MFM B (Obtained from two instrumental replicates)	109,979

3.12 Protein identification

Protein/peptide identification was performed using the Proteome Discoverer platform (Thermo Scientific, version 1.3.0.339 for results described in Chapter 5 and version 1.4 for results described both in Chapter 4 and 6), with a workflow consisting of the following nodes and respective parameters:

1) **Spectrum Selector:** Precursor Mass Range: 350-5,000 Da; S/N Threshold: 1.5;
2) **Search engine:**
   a) **Mascot** (in house server interfaced with Proteome Discoverer platform) for analyses reported in Chapter 5 according to the following criteria:
      (1) **Enzyme:** Trypsin;
      (2) **Maximum Missed Cleavage Sites:** 2;
      (3) **Precursor Mass Tolerance:** 10 ppm;
      (4) **Fragment Mass Tolerance:** 0.2 Da;
      (5) **Static modification:** Cysteine Carbamidomethylation;
      (6) **Dynamic Modifications:** N-terminal Glutamine conversion to Pyro-glutammic Acid, Methionine Oxidation and N-terminal Acetylation;
   b) **Sequest-HT** for analyses reported in Chapter 4 and 6 according to the following criteria:
      (1) **Protein Database:** UniProtKB, release 2013_07;
      (2) **Enzyme:** Trypsin;
      (3) **Maximum Missed Cleavage Sites:** 2;
      (4) **Peptide Length Range:** 5-50 amino acids;
      (5) **Maximum Delta Cn:** 0.05;
      (6) **Precursor Mass Tolerance:** 10 ppm;
      (7) **Fragment Mass Tolerance:** 0.02 Da;
      (8) **Static modification:** cysteine carbamidomethylation;
      (9) **Dynamic modification:** methionine oxidation;
3) **Percolator** for peptide validation (FDR<1%, based on peptide *q*-value, if not otherwise stated).

Results were filtered in order to keep only rank 1 peptides, and both peptide and protein grouping according to Proteome Discoverer's algorithms were allowed, applying strict maximum parsimony principle.

3.13 Proteomic data analysis

Peptides and proteins identified were subjected to the following further analyses:

- Peptide-spectrum matched (PSMs) sequences were imported on Unipept (http://unipept.ugent.be/), in order to infer taxonomic information about the identified peptides according to the Lower Common Ancestor (LCA) approach, and subjected to multi-peptide analysis setting the following parameters: "Equate I and L" and "Filter duplicate peptides" (Mesuere *et al.*, 2012).
- Peptide sequences were also subjected to standard protein BLAST search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) against the NCBI-nr DB using blastp with default parameters (included the automatic adjustment for short input sequences). BLAST output files (in xml format) were uploaded in MEGAN (MEtaGenome ANalyzer, version 4.70.4) to perform taxonomic analysis (Huson *et al.*, 2011). MEGAN parameters were left as default, except "Min support" which was set as needed (see Chapter 5 for details).
- Protein transmembrane helices were predicted using TMHMM server (v. 2.0; http://www.cbs.dtu.dk/services/TMHMM/).
- Protein annotation information, concerning subcellular localization, Gene Ontology (GO) categorization, and protein family assignment, were retrieved from UniProtKB (http://www.uniprot.org/).

- Proteins were mapped into metabolic pathways using the interactive Pathways Explorer (iPath v.2, http://pathways.embl.de/) (Yamada *et al.*, 2011).
- Venn diagrams were designed by means of Venny (http://bioinfogp.cnb.csic.es/tools/venny/index.html) or Venn Diagram Plotter (http://omics.pnl.gov/software/VennDiagramPlotter.php).
- Data elaboration was carried out using Microsoft Excel (Redmond, WA, USA) and in house scripts.

Chapter 4

Results and Discussion:

“Development of a rapid and efficient protocol based on bead-beating, FASP and single-run LC-MS/MS for in-depth metaproteome characterization”

4.1 Preliminary optimization of a protein extraction method for structurally different microbial species

We assessed the effect of adding bead-beating and freeze-thawing treatments (called method B in the “Materials and Methods” section) on protein extraction yield, compared to simple extraction by boiling in SDS-based buffer (called method A in the “Materials and Methods” section) in seven bacterial strains and two yeasts, exhibiting very different structural features as reported in Table 4-1 (and, in more detail, in Table 3-1 in the “Material and Methods” section). Six cell pellets (corresponding to 1 ml of overnight culture) per microorganism were subjected to protein extraction according to both methods (triplicate analysis per method).

Table 4 - 1. Features of microorganism used in this project to assemble microbial mixtures.

Species	Cell type	Abbreviation
Escherichia coli	Gram-negative bacillus	*Ecol*
Pasteurella multocida	Gram-negative coccobacillus	*Pmul*
Brevibacillus laterosporus	Gram-positive bacillus	*Blat*
Lactobacillus acidophilus	Gram-positive bacillus	*Laci*
Lactobacillus casei	Gram-positive bacillus	*Lcas*
Enterococcus faecalis	Gram-positive coccus	*Efae*
Pediococcus pentosaceus	Gram-positive coccus	*Ppen*
Rhodotorula glutinis	Yeast	*Rglu*
Saccharomyces cerevisiae	Yeast	*Scer*

The results shown in Figure 4-1 demonstrate that the combination of bead-beating and freeze-thawing dramatically increases protein extraction yields from yeasts (up to 14-fold) and Gram-positive bacteria (up to 10-fold), without detrimental effects on Gram-negative bacteria. Method B is therefore useful to maximize protein extraction from microbial species which are resistant to lysis using mild procedures. This is in line with previous data regarding DNA (and, in few cases, also protein) extraction from microbial cells and microbial communities (Kolmeder *et al.*, 2012; Salonen *et al.*, 2010).

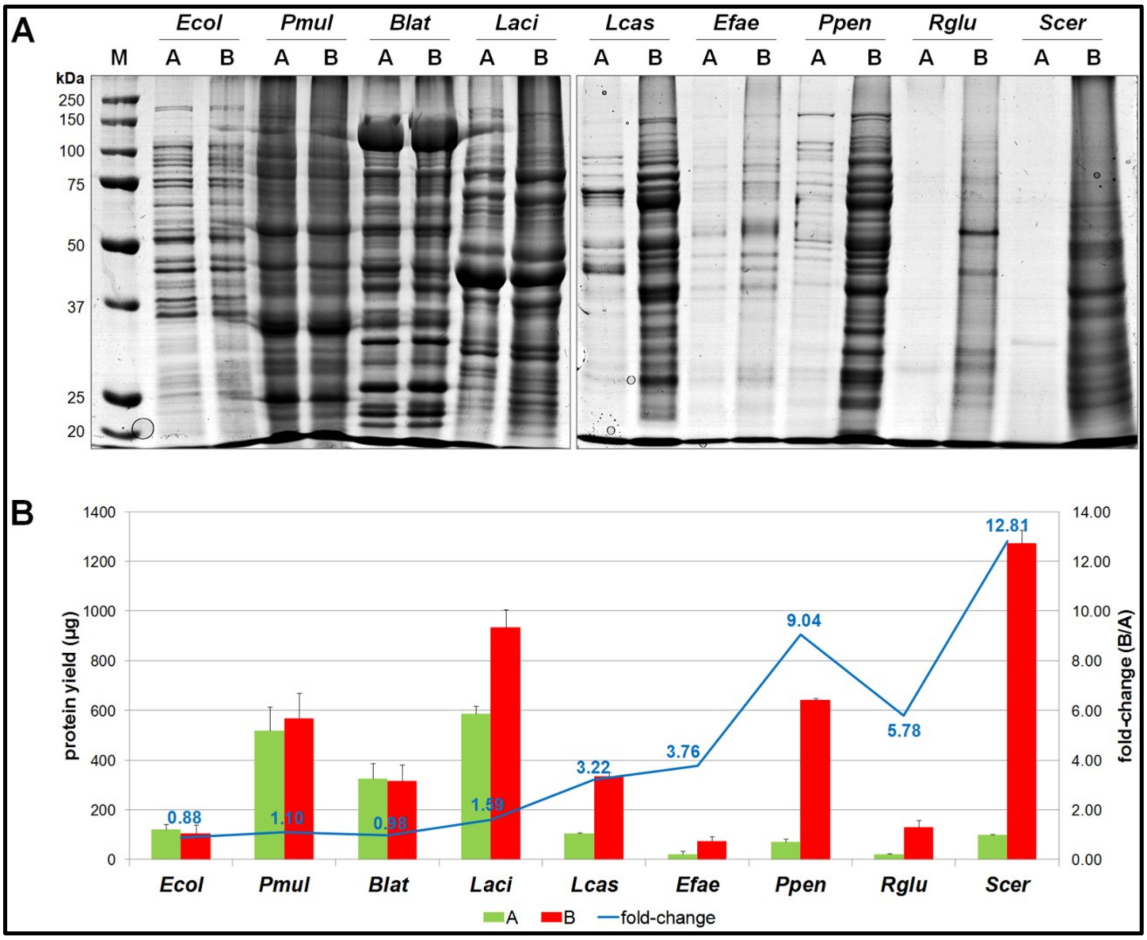


Figure 4 - 1. Protein extraction method optimization. Protein extraction yields using method A (boiling in SDS-based buffer) and method B (method A combined with bead-beating/freeze-thawing steps). For microbes abbreviation, see Table 4-1. A) SDS-PAGE pattern. Four microliters of each protein extract (obtained with method A or B, respectively) were loaded. M, molecular weight marker (Precision Plus Protein All Blue Standards, Bio-Rad). B) Histogram showing protein quantification results (mean of three replicates; error bars indicate standard deviation).

4.2 Overview of the study design for metaproteome analysis

The protocol for metaproteome analysis presented in this chapter (schematically illustrated in Figure 4-2, and detailed in the "Material and Methods" section) consists of the following steps:

1. **proteins extraction** from microbial community samples by boiling in SDS-based buffer combined with bead-beating/freeze-thawing steps (approximately 1.5 h);
2. **clean up and digestion** on-filter according to the FASP procedure of the protein extracts (minimum 8 h);
3. **single-run LC-MS/MS analysis** of the peptide mixtures using an 8 h gradient.

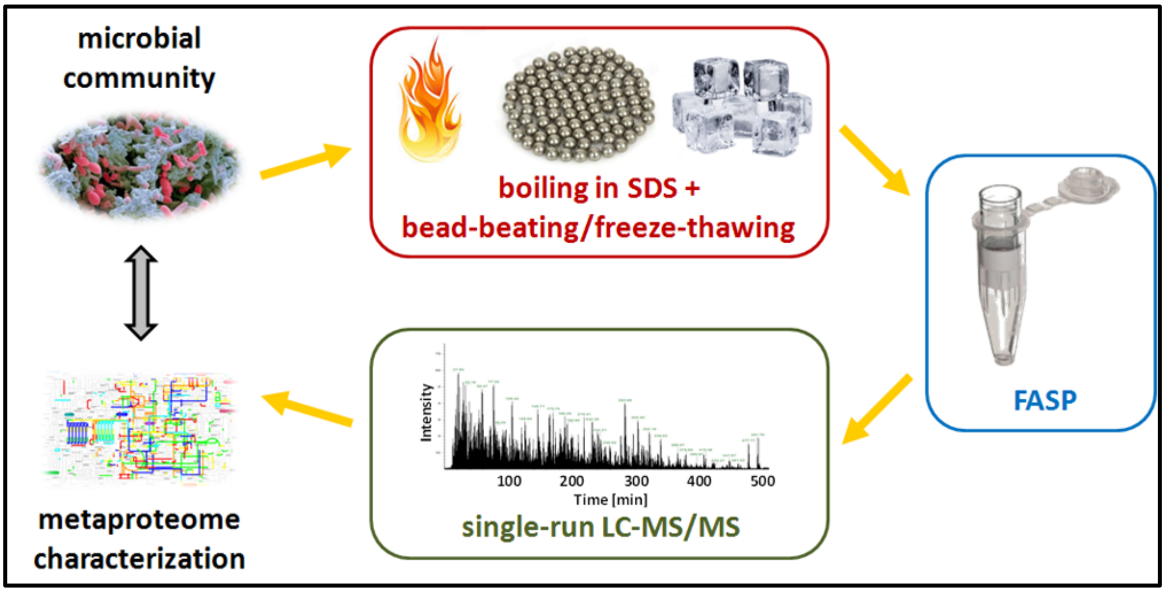


Figure 4 - 2. Schematic representation of the protocol workflow.

The protocol performance was first evaluated using lab-assembled microbial mixtures of known composition, in order to test its efficiency, reproducibility, sensitivity, dynamic range, and linearity. Then, its reliability and applicability to complex microbiome samples was validated by analyzing a mouse stool metaproteome.

4.3 Protocol evaluation on lab-assembled microbial mixtures

4.3.1 Nine-organism microbial mixture (9MM)

A first microbial mixture, named 9MM, was assembled by mixing all the nine above mentioned microbes (one pellet for each microbe, corresponding to 1 ml of overnight culture). The 9MM was subjected to the metaproteomic protocol described in the previous section, and the peptide mixtures obtained were analyzed in duplicate by single-run LC-MS/MS. Remarkably, the protocol showed a very high reproducibility among runs (over 99% based on PSMs, Figure 4-3A, left). Moreover, as detailed in Table 4-2, a mean of over 1,900 non-redundant proteins (of which about 1,250 identified with at least two non-redundant peptide sequences) and over 9,000 non-redundant peptides per run could be identified. Merged data from two replicate runs provided 2,186 non-redundant protein identifications. Proteins belonging to sequences from all nine microbial species contained within the 9MM (including Gram-positive bacteria and yeasts) were consistently detected (with FDR<1%), with mutual ratios comparable to those observed for protein extraction yields (PSM values were compared in Figure 4-3A, right). The percentage of transmembrane proteins identified was also estimated (6.2% on average per run), with 4.2% with a single transmembrane domain (TMD) and 1% with two or more TMDs (specifically over 40 multipass membrane proteins detected per run). On the other hand, 15%

of proteins having a localization annotation in UniProtKB were classified as belonging or associated to membrane. Finally, 10% of the identifications were low molecular weight proteins (MW<10 kDa), although a 30 kDa cut-off was chosen for on-filter sample preparation.

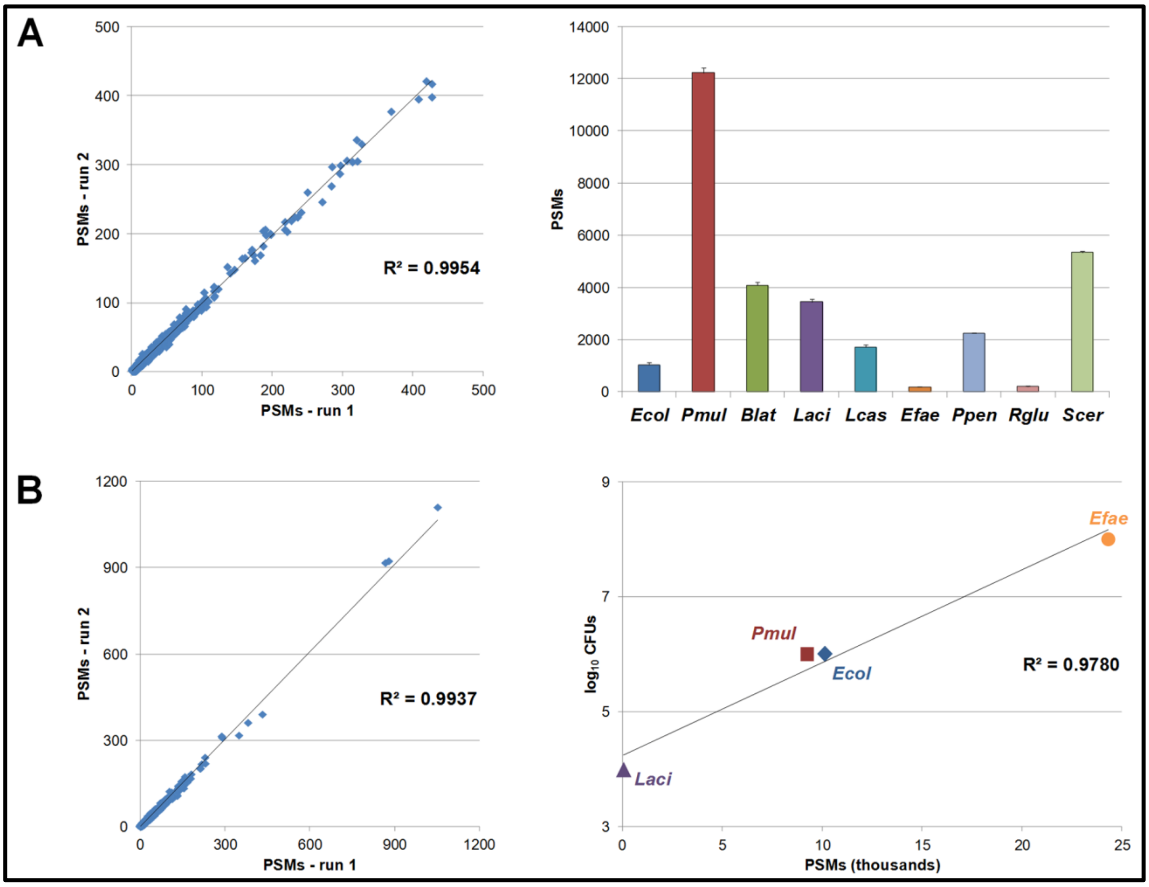


Figure 4 - 3. Results obtained with lab-assembled microbial mixtures. A) 9MM sample: correlation of PSM values among runs to estimate reproducibility (left), and distribution of the identified PSMs among microbial species (right). B) 4MM sample: correlation of PSM values among runs to estimate reproducibility (left), and correlation between the number of bacterial CFUs and the corresponding number of identified PSMs (right).

4.3.2 Four-organism unbalanced microbial mixture (4MM)

Four bacterial strains were then selected to assemble a simpler mock mixture, named 4MM. In this case, the amount of bacterial cells was accurately measured, and bacteria were mixed in unbalanced proportions (specifically, 10^{10} CFUs of *E. faecalis*, 10^{8} CFUs of *E. coli* and *P. multocida*, respectively, and 10^{6} CFUs of *L. acidophilus*), with the purpose of testing sensitivity and linearity of the method in relation to the bacterial cell amount. As a result, a mean of 11 proteins and 27 unique peptides per run were assigned to the less abundant species (*L. acidophilus*). Considering that the total amount of protein extracted from the 4MM was nearly 350 μg, and the amount of peptide mixture actually loaded per run was 4 μg (therefore about the ninetieth part of the initial 10^{6} *L. acidophilus* CFUs), the sensitivity of the protocol can be estimated as equal to (or lower than) 10^{4} CFUs. On the whole, the protocol exhibited a dynamic range of four orders of magnitude, slightly wider than previously observed (VerBerkmoes *et al.* 2009b). Interestingly, the number of PSMs assigned to each bacterial species was strictly correlated ($r^2 \geq 0.98$) to the bacterial cell amount (expressed as $\log_{10}$ CFUs; Figure 4-3B, right). Concerning standard identification statistics, reproducibility was confirmed as higher than 99% (based on PSMs; Figure 4-3B, left), whereas, as detailed in Table 4-2, a mean of 2,600 non-redundant proteins (of which 1,900 identified with at least two non-redundant peptide sequences) and over 14,000 non-redundant peptides were identified per run; cumulative identifications along two runs were even higher (nearly 3,000 and 17,000 non-redundant protein and peptide identifications, respectively). Furthermore, 11% of the identified proteins contained at least one TMD (of which 6.2% with two or more TMDs, corresponding to almost 170 multipass membrane proteins found per run). Based on UniProtKB localization annotation, 20% of proteins were classified as belonging or associated to membrane. Moreover, 9% of the identifications were low molecular weight proteins (MW<10 kDa), in spite of the putatively higher cut-off used for FASP.

On the whole, the data achieved with 9MM and 4MM samples are largely comparable, in qualitative and quantitative terms, to the results obtained using LC gradients of similar length in previous works, notably with less technically demanding samples (Nagaraj *et al.*, 2012; Pirmoradian *et al.*, 2013; Thakur *et al.*, 2011). Furthermore, the microbial mixtures used in this study were mainly composed by environmentally isolated (and not reference) strains; this implies that a slight, but not negligible, portion of the spectra may not have been correctly matched with the *in silico* peptide sequences due to differences between database reference sequences and experimental sequences (for instance, for the poorly characterized *R. glutinis*). This phenomenon, as well as the influence of protein sequence database selection on metaproteome characterization, will be specifically analyzed and discussed in the following chapter.

4.4 Protocol validation on murine fecal samples

The murine fecal microbiome (MFM) was chosen to validate the reliability and suitability of the protocol to complex metaproteome samples. Stool was preliminary subjected to differential centrifugation in order to enrich for microbial cells, and then subjected to the protocol previously evaluated on the lab-assembled microbial mixtures. Specifically, two technical replicates were processed in parallel (from differential centrifugation to FASP), and for each replicate two separate LC-MS/MS analyses were run (instrumental replicates, four replicates in total). An iterative strategy was adopted for protein identification, based on the recently published "two-step" database search method (Jagtap *et al.*, 2013).

Technical and instrumental reproducibility values were rather high (97% and 95% based on PSMs, respectively, as illustrated in Figure 4-4A). Table 4-2 shows the number of proteins, peptides and PSMs identified in each run (with FDR<1%), as well as the cumulative results (per technical replicate and overall). Up to 8,000 non-redundant proteins (of which 5,600 of microbial origin, and 2,400 identified with at least two non-redundant peptide sequences) and 11,500 non-redundant peptides could be identified per run, reaching 12,000 and over 18,000 protein identifications per technical replicate and in total, respectively. Concerning taxonomic distribution, 81% of proteins and 76% of PSMs were of microbial origin (specifically assigned to Bacteria, Archaea, Fungi or Viruses).

Table 4 - 2. Number of proteins, peptides and PSMs identified in each sample, replicate and run.

Sample type	Replicate	Run	Number of non-redundant proteins	Number of non-redundant proteins (≥ 2 pept)	Number of non-redundant peptides	Number of PSMs
4MM		1	2,598	1,903	14,746	38,903
		2	2,611	1,910	14,595	39,787
		1+2	*2,952*	*2,129*	*16,996*	*78,690*
9MM		1	1,973	1,266	9,173	27,015
		2	1,880	1,227	9,004	26,767
		1+2	*2,186*	*1,362*	*10,285*	*53,782*
MFM	a	1	7,429	1,664	10,381	19,033
		2	7,719	1,738	10,964	19,651
		1+2	*11,538*	*2,762*	*15,585*	*38,684*
	b	1	8,034	2,396	11,526	20,950
		2	6,994	2,049	10,097	19,146
		1+2	*11,999*	*3,000*	*16,153*	*40,096*
	a+b	***1+2***	***18,428***	***3,612***	***23,795***	***78,780***

Figure 4-4B depicts the taxonomic distribution according to lowest common ancestor (LCA) analysis carried out on PSM data using the Unipept web application. The distribution obtained by analyzing the two technical replicates was almost identical, as shown in Figure 4-4. According to LCA results, an outstanding microbial diversity could be observed in the MFM sample. In fact, peptide sequences were unambiguously classified as belonging to the following different microbial taxa: 41 phyla, 77 classes, 169 orders, 365 families, 1048 genera, and 1997 species.

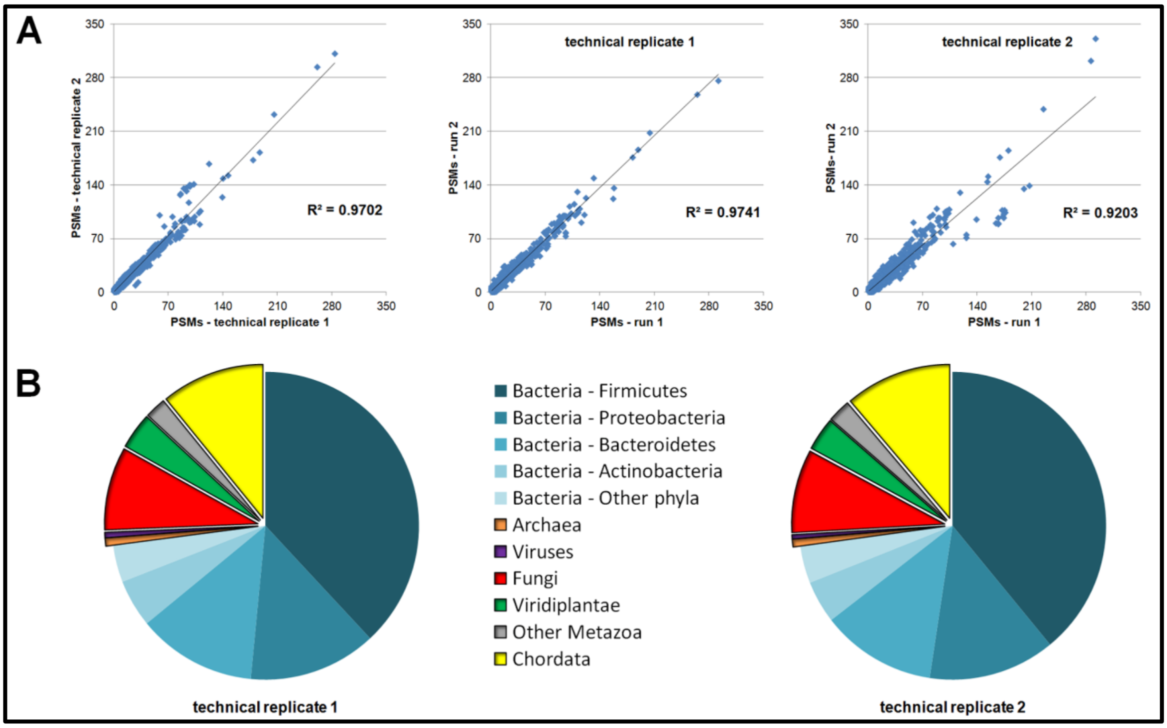


Figure 4 - 4. Evaluation of reproducibility in the analysis of the MFM sample. A) Correlation of PSM values among technical replicates (left) or among runs (center and right). B) Taxonomic distribution of the identified PSMs (first technical replicate, left; second technical replicate, right).

Notably, a considerable percentage of the identifications (specifically, 9%) was represented by fungal sequences, thus confirming the ability to properly extract and detect proteins from cells usually refractory to lysis. This is a significant result, because of the increasing importance recognized to the eukaryotic components of microbial communities, especially associated to human tissues and organs (Huffnagle and Noverr 2013; Minton 2012). Moreover, a significant part of the sequences were assigned to "exotic" fecal components, such as *Nematoda* (phylum, 0.6%) and *Insecta* (class, 1.4%), classified in the pie charts as “Other Metazoa”. Concerning protein topology and localization, 14% of protein identifications contained at least one TMD (of which 7.6% with two or more TMDs, corresponding to a total of 1,637 multipass membrane proteins detected in the MFM sample); 24% of proteins with UniProtKB localization annotation were classified as belonging or associated to membrane.

Mouse fecal metaproteome results were further analyzed to carry out a functional characterization of the microbiome. Specifically, according to UniProtKB protein family classification, 698 different protein families were identified in the MFM sample; among them, 603 were of microbial origin. The top 25 microbial protein families, based on the number of family member identified, are listed in Table 4-3; interestingly, they cover a wide range of enzymatic, transport and signaling functions. Furthermore, microbial protein identities were uploaded into iPATH web application with the aim of mapping proteins into metabolic pathways. As shown in Figure 4-5, the metaproteome data achieved using our protocol could be mapped to a high number of different metabolic pathways, and the specific contribution of the main bacterial phyla and of the fungal part (marked with different colors in the Figure 4-5) to the whole metabolic activity of the microbiome could be recognized.

The MFM data presented here represent the largest fecal metaproteome dataset published to date. Quantitatively comparable results per run (when considering proteins with at least two unique peptides) could be obtained only by applying 2D-LC approaches, but with higher FDR values compared to the stringent threshold used here (Erickson *et al.*, 2012; Verberkmoes *et al.*, 2009b).

More interestingly, the two main strategies so far successfully employed to unravel the stool metaproteome, the above mentioned 2D-LC-MS/MS preceded by in-solution digestion, and the extensive 1-DE fractionation before in-gel digestion and LC-MS/MS, are considerably more time-consuming and technically demanding compared to the protocol described here (Kolmeder *et al.*, 2012; Pérez-Cobas *et al.*, 2012). Moreover, gel cutting and destaining, as well as in-gel digestion, represent extremely labor-intensive steps for a laboratory operator; conversely, our method is more straightforward and also suitable for automation.

Table 4 - 3. Top 25 microbial protein families detected in the mouse stool sample. The table is ordered by the number of family members identified.

UniProt protein family	Number of family members identified	Total PSMs
ABC transporter superfamily	293	5739
GTP-binding elongation factor family	253	29255
TonB-dependent receptor family	108	357
ATPase alpha/beta chains family	104	4140
Glyceraldehyde-3-phosphate dehydrogenase family	94	4053
Chaperonin (HSP60) family	87	3521
Heat shock protein 70 family	85	1175
Glu/Leu/Phe/Val dehydrogenases family	84	2568
Class-II aminoacyl-tRNA synthetase family	78	383
NifJ family	77	4512
RNA polymerase beta chain family	74	2808
Phosphoglycerate kinase family	68	1941
Class-I aminoacyl-tRNA synthetase family	60	320
Acyl-CoA dehydrogenase family	55	939
Short-chain dehydrogenases/reductases (SDR) family	52	78
RNA polymerase beta' chain family	49	2201
Polyribonucleotide nucleotidyltransferase family	45	650
Binding-protein-dependent transport system permease family	43	92
Glycogen phosphorylase family	39	488
Aldehyde dehydrogenase family	38	57
Cation transport ATPase (P-type) family	36	223
ClpA/clpB family	33	418
Enolase family	33	378
Thiolase family	32	207
Phosphofructokinase family	26	304

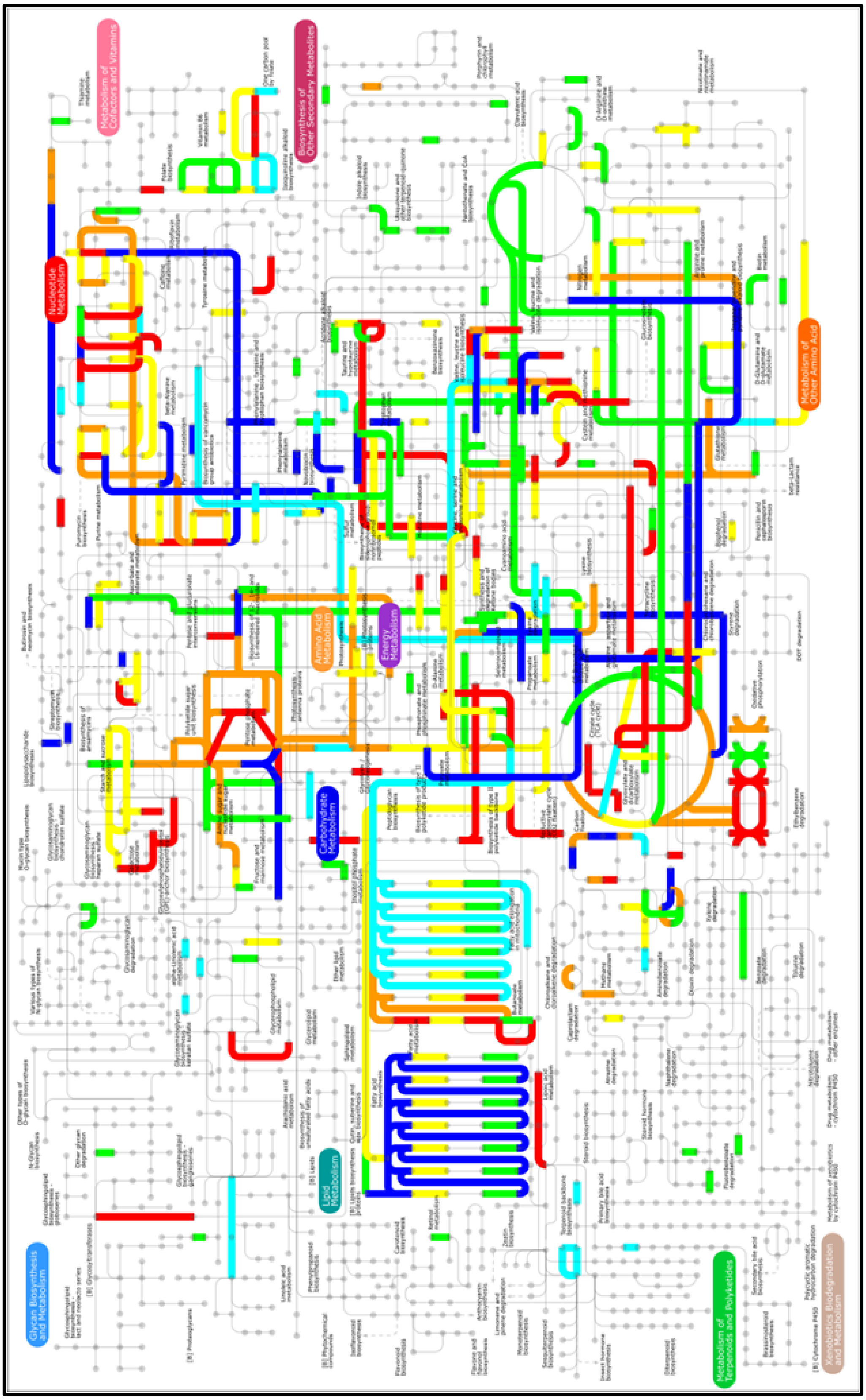

Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Nucleotide Metabolism
Metabolism of Other Amino Acid
Amino Acid Metabolism
Energy Metabolism
Carbohydrate Metabolism
Lipid Metabolism
Glycan Biosynthesis and Metabolism
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism

Figure 4 - 5. Distribution of the identified proteins belonging to the main microbial taxa into metabolic pathways. Colors assignments: green for Proteobacteria; red for Firmicutes; dark blue for Bacteroidetes; yellow for Actinobacteria; light blue for Fungi; orange for Archaea.

In conclusion, this chapter presents a rapid and efficient protocol for metaproteome analysis. The overall procedure can be accomplished in a minimum of ~18 h, compared to the best performing method developed to date (in-solution digestion coupled to 2D-LC-MS/MS) which requires at least 22 h solely for the 2D-LC separation. Our protocol, clearly, enables the identification of proteins from the diverse microorganisms (included Fungi) that might be part of a gut microbiome, showed a sensitivity down to 10^4 bacterial CFUs, a linear dynamic range of 4 orders of magnitude, and a reproducibility up to over 99%. When applied to fecal samples, it led to the identification of proteins belonging to nearly 2,000 different microbial species and mapping to over 600 functionally relevant protein families. In keeping with this, the protocol described here may be successfully used for the in-depth and time-effective characterization of complex microbiomes.

Chapter 5

Results and Discussion:

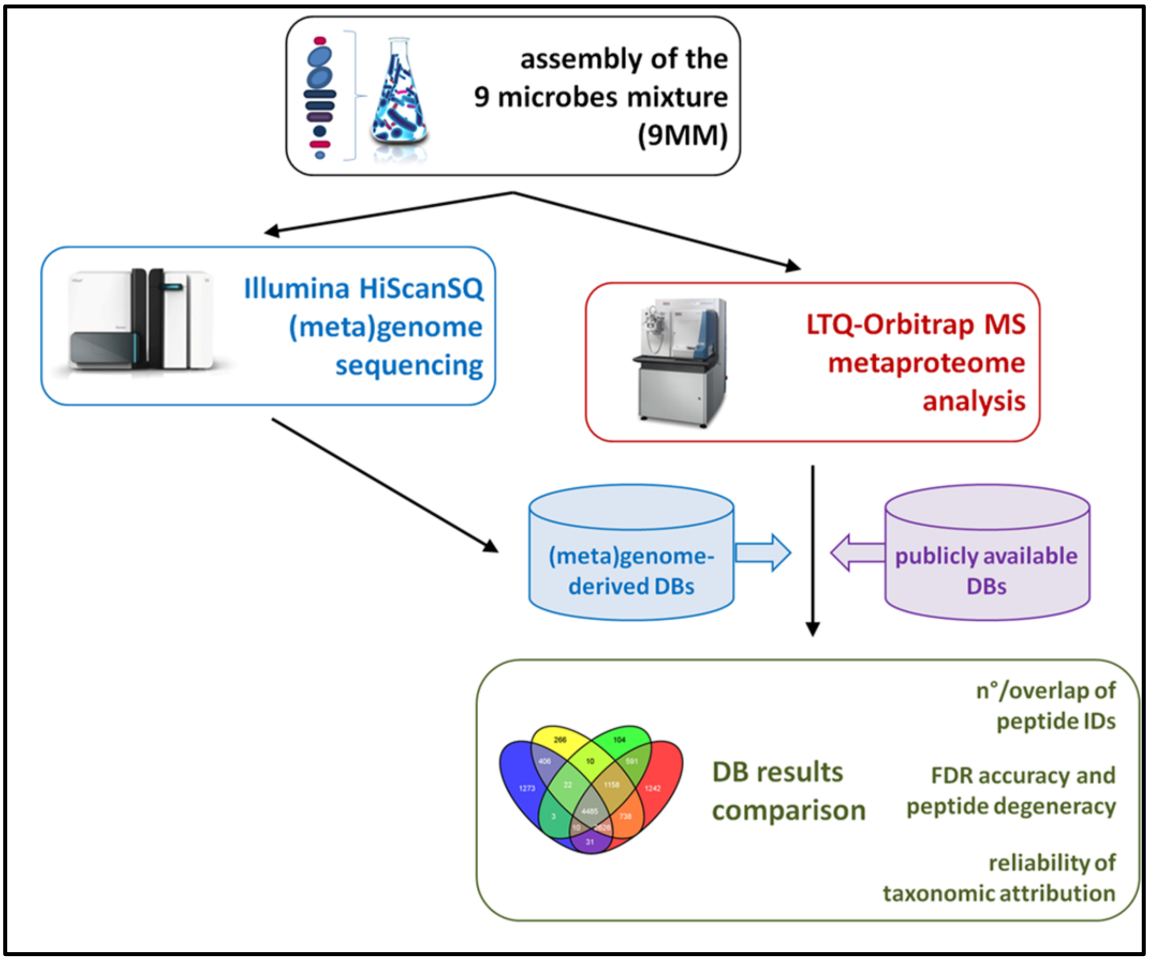
“Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture”

5.1 Global experimental design

This study was designed as schematized in Figure 5-1.

- As a first step, a nine organism microbial mixture (9MM) was assembled, by mixing seven prokaryotes and two eukaryotes microorganisms with heterogeneous structural characteristics, as summarized in Table 3-1 in the "Material and Methods" section. Differently from most of the studies published to date regarding the integration between metagenomics and metaproteomics, the 9MM used in this study contained, as mentioned, also eukaryotic microorganisms. This was in line with the recently growing awareness of the importance of Fungi within microbial communities and of their key functions for health and disease, which is opening the way to the study of the so-called "mycobiome" (Huffnagle and Noverr, 2013; Iliev *et al.*, 2012; Minton, 2012). Among all microorganisms used, species with lack of previous genomic characterization were included (such as *R. glutinis*), together with well-known species, both reference strains and environmental isolates (which may be expected to extensively differ from publicly deposited sequences), in order to take into account the high variability in the level of sequence information that might be encountered in environmental microbiomes.
- In the second step, the 9 microorganisms were subjected to Illumina NGS both as individual organism and as 9MM, in order to generate genome- and metagenome-derived protein DBs (see "Materials and Methods" for details).
- Then, as a third step, the 9MM metaproteome was analyzed by shotgun LTQ-Orbitrap MS, and MS data were searched against publicly available and matched experimental DBs to achieve peptide identification.

- Finally, as the last step, the information carried out using the different DBs was comparatively evaluated in relation to: number and overlap of peptide identifications; FDR behavior and peptide degeneracy; and reliability of taxonomic attribution using MEGAN and Unipept software.



assembly of the 9 microbes mixture (9MM)
Illumina HiScanSQ (meta)genome sequencing
LTQ-Orbitrap MS metaproteome analysis
(meta)genome-derived DBs
publicly available DBs
DB results comparison
n°/overlap of peptide IDs
FDR accuracy and peptide degeneracy
reliability of taxonomic attribution

Figure 5 - 1. Schematic illustration of the experimental design.

In particular, four main DB classes were considered for comparison, each one corresponding to a different experimental approach that might be used in a metaproteomics study (as represented in Figure 5-2):

1. public DBs (namely, NCBI, SwissProt and TrEMBL) with generic taxonomic indications (all microbial sequences, or rather those

belonging to Bacteria, Fungi, and Viruses, abbreviated as BFV), an approach needed when no precise taxonomic information and/or matched genome sequencing data are available for the microbiome under study;

2. protein sequences selected from the above mentioned public DBs, based on taxonomic information (referred to as "taxonomy-restricted" DBs, parsed at genus, G, or species, S, level) which may derive from previous *16S rDNA* gene sequencing or metaproteomic information;
3. matched metagenome sequence DBs (named "Meta" DBs), experimentally obtained from whole metagenome sequencing of the same microbiome subjected to metaproteomic analysis;
4. assembly of experimentally obtained individual genome sequences from the main species included in the microbiome (named "single genomes assembly", SGA), an approach that requires isolation of each strain of the culturable microbiome (Lagier *et al.*, 2012).

A further distinction must be made concerning genome data processing: both the metagenome and the single genomes were subjected either to coding sequence prediction and annotation (PA) or to naïve six-frame translation (6FT), thus generating four different experimental DBs. In addition, the information needed for generating a "taxonomy-restricted" DB can be easily (and usually) gathered by 16S-18S characterization, but a metaproteomic iterative approach can be also proposed, comprising a first search using a generic DB, sequentially followed by the identification of the main taxa of the microbiome of interest from metaproteomic data (using proper filters to improve reliability), the construction of a customized, smaller DB, and a second search with this latter DB to improve metaproteome coverage. This iterative metaproteomic strategy, which differs from the 'two-step method' proposed by Jagtap and coworkers in that the former is taxonomy-based, might be therefore successfully implemented without the need for additional genomic or metagenomic surveys (Jagtap *et al.*, 2013).

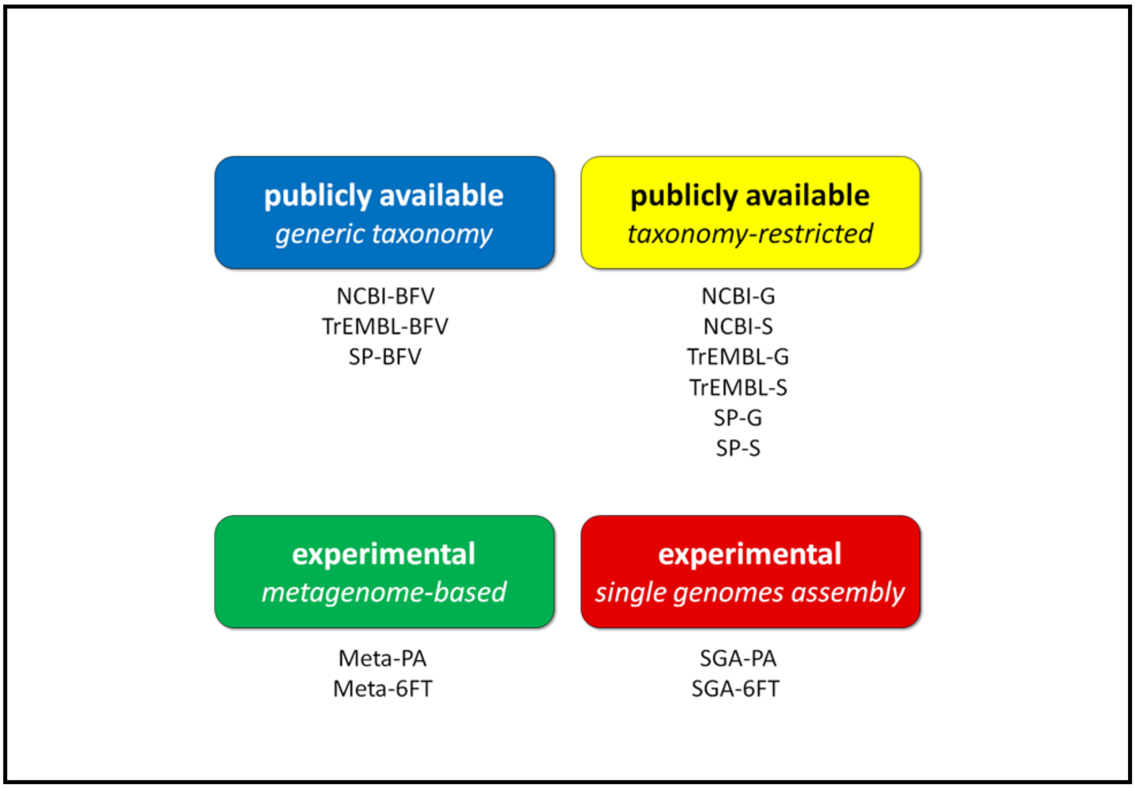


Figure 5 - 2. Schematic illustration of the database classes examined.

5.2 Comparison of metaproteomic data obtained using different protein databases

Figure 5-3 shows a comparison among the peptide identification data achieved by searching the MS spectra against the 13 DBs described above, using FDR<1% as a threshold. The use of SGA-PA led to the identification of the higher number of peptides (Figure 5-3A), while SwissProt-based DBs provided the least satisfactory results. Similar results were obtained according to the number of peptide-spectrum matches (PSMs; Figure 5-3B).

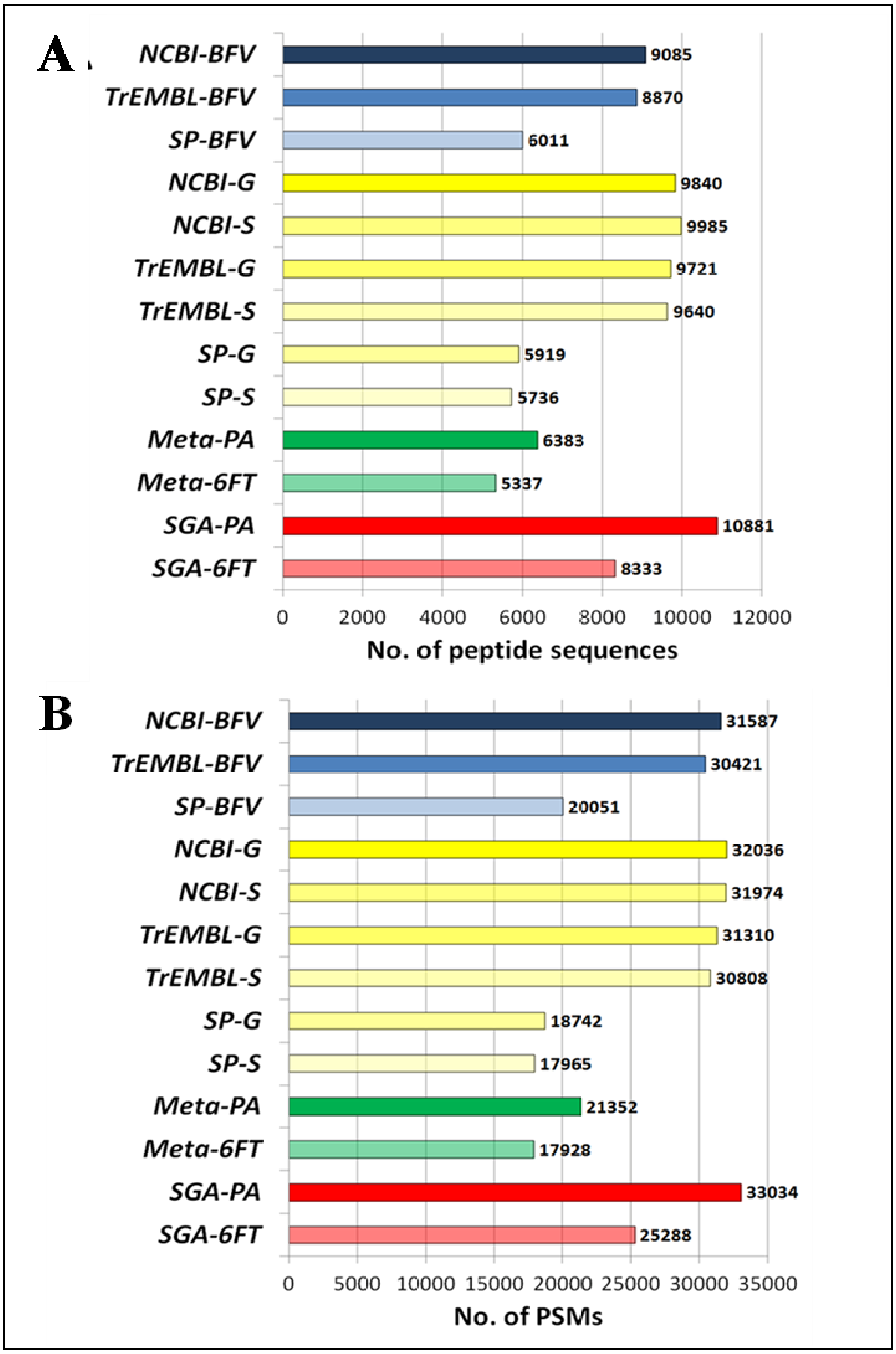


Figure 5 - 3. Comparison of metaproteomic data obtained with different databases. Number of peptide sequences (A) and peptide-spectrum matches (PSMs, B) identified in the 9MM using different sequence databases (FDR<1%).

The amount of peptide identifications achieved using the metagenome-derived DBs was slightly higher than those identified using SwissProt, but clearly lower than public non-manually annotated NCBI and TrEMBL DBs. Furthermore, "taxonomy-restricted" DBs from NCBI and TrEMBL performed better than the corresponding DBs with wider taxonomy. It has also to be noted that, as indicated in Table 3-3 and Table 3-4, the average computing time needed for the DB search differed dramatically among the DBs, proportionally to each DB size. On the whole, 12,911 different peptide sequences were identified by searching MS spectra against all DBs described above.

Four DBs were then selected as representative of the four main DB classes described above. Specifically, two were TrEMBL-based DBs, and two were (meta)genome-based DBs annotated against TrEMBL. The intersections among the peptide sequences identified with each DB were calculated and illustrated by means of a Venn diagram (Figure 5-4A). Surprisingly, only about one-third of the identified peptide sequences were common to all DBs, while 22% were unique to a single DB (of which nearly 90% were unique to TrEMBL-BFV or SGA-PA). Meta-PA identifications were common to SGA-PA at 98%, whereas the specific increment obtained with Meta-PA compared to the public DBs (given by the peptide sequences found only using Meta-PA and not detected using any publicly-available DB) could be estimated at 6%. Furthermore, 68% of peptide sequences were in common between TrEMBL-BFV and TrEMBL-G. When comparing DBs according to the public DB of origin (Figure 5-4B, left), approximately half of the peptides were common to NCBI, TrEMBL and SwissProt; NCBI and TrEMBL shared over 90% of the identified peptides, while about 8% of SwissProt peptide sequences (5% of the total) were not identified in the other DBs. As far as different taxonomy filters are concerned (Figure 5-4B, right), 70% of peptide identifications were common to all DBs, but the use of genus/species-specific DBs led to a 17% increase in identifications compared to search against a general microbial taxonomy (BFV).

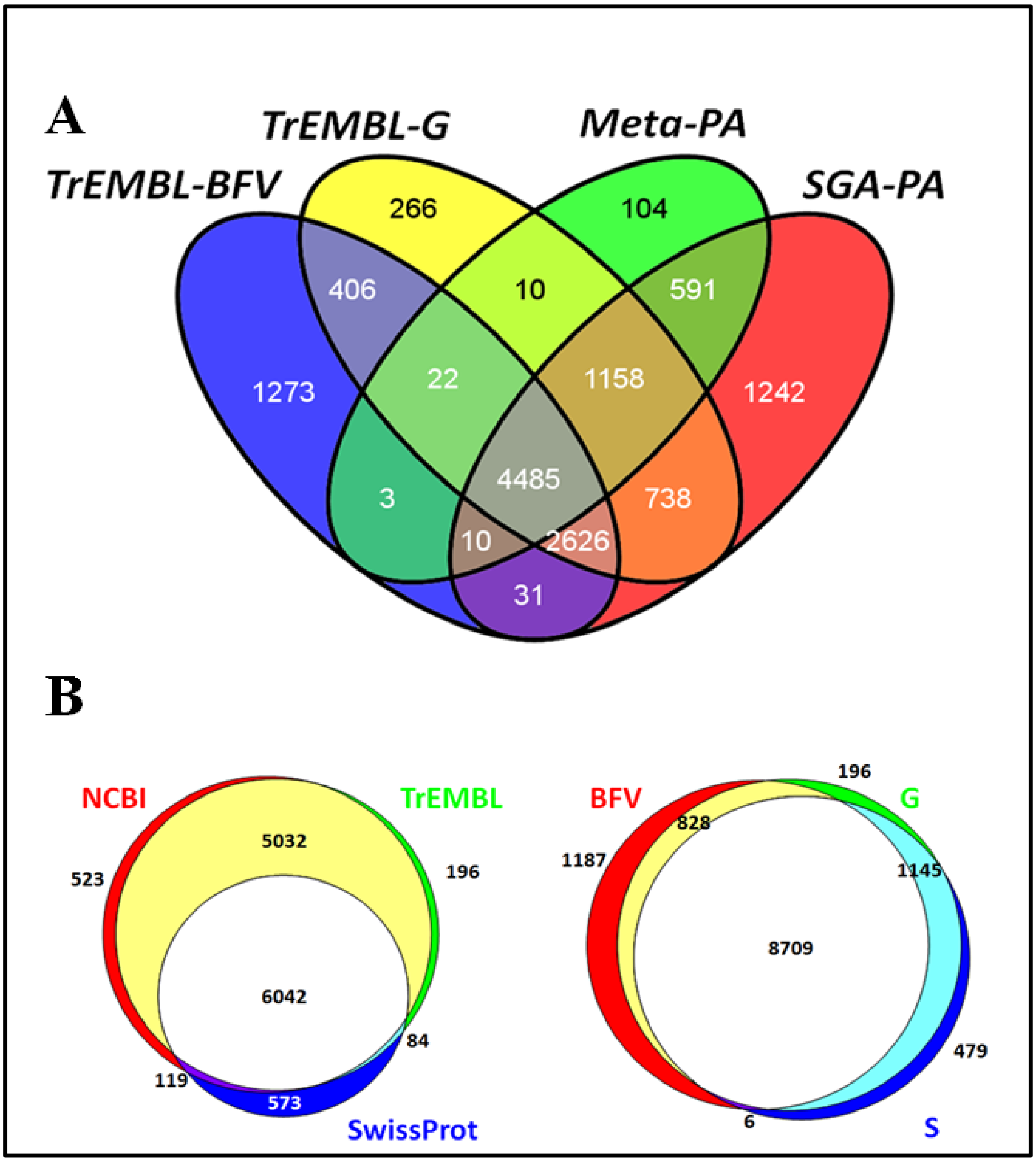


Figure 5 - 4. Comparison of metaproteomic data obtained with different databases. A) Venn diagram illustrating the peptide distribution among four different DB classes. B) Left, Venn diagram illustrating the peptide distribution among all NCBI-, TrEMBL- and SwissProt-based DBs used in this study. Right, Venn diagram illustrating the peptide distribution among all DBs with generic microbial taxonomy (BFV), genus-specific taxonomy (G), and species-specific taxonomy (S).

The performance of 6FT DBs were also evaluated. A total of 5,337 peptides were identified by searching MS spectra against Meta-6FT, of which 117 (2%) were unique when compared with the corresponding annotated DB (Meta-PA); SGA-6FT allowed the detection of 8,333 peptides, of which 757 (9%) had not been found using SGA-PA. On the whole, the employment of 6FT DBs enabled 783 additional identifications (6% increase).

It is worth noting that the number of peptides (for metaproteomics), as well as the number of reads (for metagenomics), attributed to each of the 9MM microbes was far from being equal, although they were theoretically present in comparable amounts based on CFU counting. This might be explained by the fact that the nine microbial species exhibited huge differences in size and cell structure, which may mean that similar numbers of cells contain different amounts of proteins, as well as by varying protein extraction yields among microorganisms.

5.3 Evaluation of FDR behavior and peptide degeneracy across different databases

Another aim of this study was to investigate how FDR behavior and peptide degeneracy are influenced by the particular DB used for metaproteome analysis. To evaluate FDR behavior, the number of peptides (Figure 5-5, left) and PSMs (Figure 5-5, right) identified with each DB were plotted as a function of FDR thresholds based on the Percolator *q*-values, as previously described by Spivak and colleagues (Spivak *et al.*, 2009). As a result, DBs could be distinguished into two groups based on the typical trend of their *q*-value curves: the first comprising all publicly available DBs with generic taxonomy (NCBI-BFV, TrEMBL-BFV and SP-BFV), whose curve kept on rising much longer compared to the remaining DBs, that tended considerably more rapidly to a *plateau*. Interestingly, the FDR evolution was quite different if either peptide sequences or PSMs were considered. For instance, SGA-PA achieved the higher number of peptide identified at any FDR, whereas in terms of PSMs the same DB passed from giving the best results at 1% FDR to being only the fourth best DB at 5% FDR.

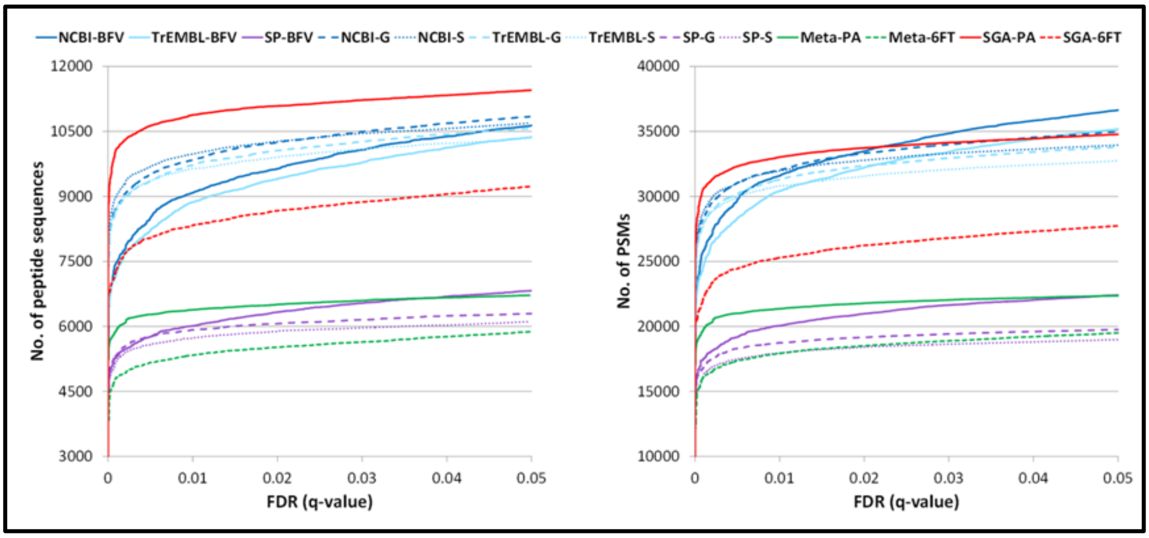


Figure 5 - 5. Evaluation of FDR behavior using different databases. Diagram plotting the number of peptides (left) and PSMs (right) identified with each database as a function of FDR thresholds based on the Percolator *q*-values.

The increment in peptide/PSM identifications when increasing the FDR threshold from 1 to 5% was also evaluated (Figure 5-6), and it was observed that the public DBs with generic taxonomy consistently yielded the highest percentage of additional hits when increasing the FDR threshold both for peptide sequences (Figure 5-6, left) and PSMs (Figure 5-6, right). Another significant observation could be made concerning 6FT DBs, which showed a two-fold percentage increase compared to the corresponding PA DBs when the FDR threshold was raised to 5%.

Furthermore, the degree of peptide degeneracy related to each DB was estimated by calculating the percentage of shared (or degenerate) peptides/PSMs. According to Proteome Discoverer’s algorithms, after protein identities are deduced from a set of identified peptides, proteins are grouped according to the peptide sequences identified for the proteins (in this case allowing the “Strict Maximum Parsimony Principle” option), and a master protein is reported for each protein group, which has been identified by a set of peptides that are not included (all together) in any other protein group. Each identified peptide can be therefore matched either with a single protein group (called “unique peptide”) or with multiple protein groups (called “shared peptide”).

In this context, the percentage of shared peptides out of the overall identifications gives an indication of the degeneracy associated to a particular DB. As shown in Figure 5-7, in general the percentage of shared PSMs (Figure 5-7, right) was higher compared to the percentage of shared peptides (Figure 5-7, left) measured for the same DB (with FDR<1%).

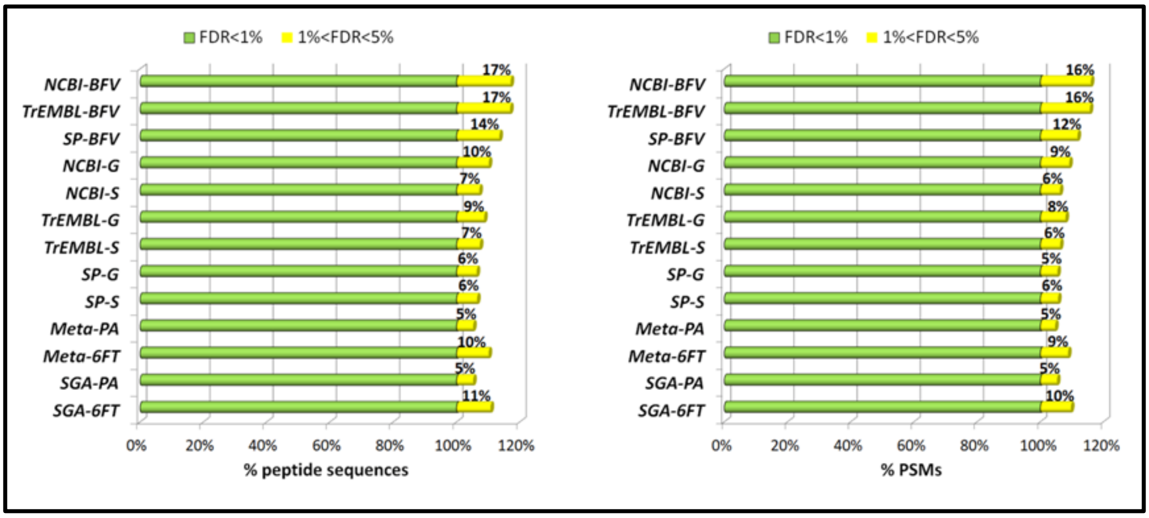


Figure 5 - 6. Evaluation of peptide degeneracy using different databases. Bar graph showing the percentage increment in peptide (left) and PSM (right) identifications achieved with each database when increasing the FDR threshold from 1 to 5%.

Moreover, experimental DBs exhibited significantly lower percentages of shared peptides (and even lower for PSMs) when compared with publicly available DBs, whose peptide degeneracy decreased, as expected, according to the following order: NCBI>TrEMBL>SwissProt and BFV>G>S.

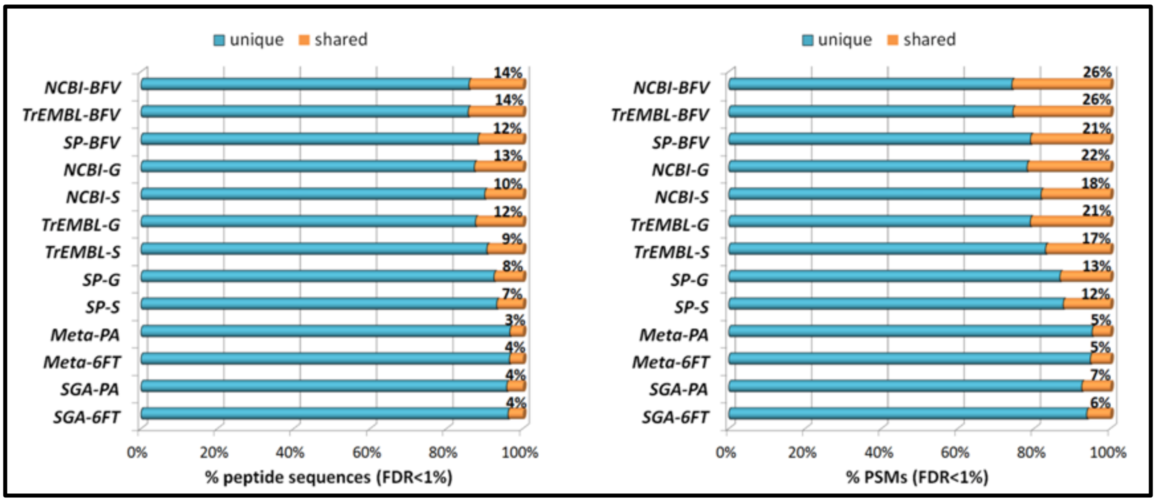


Figure 5 - 7. Evaluation of shared peptide and PSMs using different databases. Bar graph illustrating the percentage of shared peptides (left) and PSMs (right) identified with each database at FDR<1%.

The results presented here further highlight that the use of large and complex DBs required for multispecies samples (such as microbial communities) poses significant challenges in the implementation and optimization of search-decoy approaches for FDR calculation, and suggest that peptide/PSM identification significance thresholds are strongly influenced by DB size and redundancy, even when a post-search algorithm using semi-supervised machine learning (such as Percolator) is used. In fact, the use of "taxonomy-restricted" DBs led to a higher number of peptide identifications in comparison with those obtained with the same DBs with wider taxonomy (and thus larger size). This may seem quite surprising, given that "taxonomy-restricted" DBs were just a subset of the corresponding "general" DBs, containing no additional sequences when compared to the latter. Specifically, most of the peptide sequences uniquely detected with "taxonomy-restricted" DBs were not identified using the corresponding "general" DB, since those were discarded being below the 1% FDR threshold. Also the poorer performance of 6FT DBs when compared to the corresponding PA DBs may be explained in a similar way, since the former are almost six time bigger than the latter. In this respect, the use of alternative search-decoy strategies as those described by Blakeley *et al.* and, even more recently, by Jagtap *et al.* might partially address this problem and lead to an increase of peptide identifications, and may be the target of future studies (Blakeley *et al.*, 2012; Jagtap *et al.*, 2013). The same phenomenon could be observed for SwissProt when compared to TrEMBL (Figure 5-4): TrEMBL provided a much higher absolute number of identifications (mostly due to the lack of less characterized species within SwissProt), but the parallel use of SwissProt gave additional, unique information. Manually curated DBs offer also further advantages, including a higher level and quality of annotation concerning protein functions, processes, and localizations, which can be extremely useful in the functional perspective allowed by metaproteomics.

5.4 Reliability of taxonomic attribution by Unipept and MEGAN analysis of metaproteomic data

The metaproteomic data generated in this work were then used to evaluate the reliability of the taxonomic attribution of peptide identifications, with the aim of assessing the influence exerted by the DB choice in this type of investigations. Such evaluation was possible due to the *a priori* knowledge of the taxonomic composition of the lab-assembled 9MM. Specifically, the peptide sequences identified using the different DBs were parsed by means of two software enabling taxonomic analysis according to the LCA approach, namely Unipept and MEGAN (Huson *et al.*, 2011; Mesuere *et al.*, 2012). It is worth noting that MEGAN requires a preliminary BLAST search of the identified peptide (or protein) sequences to be performed, since a BLAST file is needed as input. Furthermore, MEGAN “Min Support” filter (that is, the number of reads/peptides that must be assigned to a taxon so that it appears in the results) was initially set to 1, according to Rudney *et al.* (Rudney *et al.*, 2010).

Figure 5-8 and Figure 5-9 illustrate the number of peptides detected as specific to family (top), genus (middle) or species (bottom) level upon, respectively, Unipept or MEGAN analysis, identified with five different DBs. Peptide distribution among the prokaryotic (blue) and eukaryotic (green) strains included in the 9MM was also taken into account, as well as the incorrect attributions (denominated "misassignments", in red). Genus/species-specific DBs were excluded from this comparison because it would have been superfluous to assess taxonomy attribution reliability when a specific "taxonomy filter" had been already set *a priori*, and therefore the number of misassignments had been "forced" to be zero. In general, the number of taxon-specific identifications decreased proportionally to the degree of taxonomic detail (for instance, nearly 4,500, 3,500, and 2,000 peptides could be found with family, genus, and species specificity with NCBI-BFV, respectively). Moreover, a higher amount of taxon-specific peptides could be yielded with Unipept analysis compared to MEGAN (for example up to over 4,500 family-specific peptides with Unipept versus less than 1,800 with MEGAN).

The impact of taxonomic "misassignments" was also evaluated. As a result, Unipept demonstrated a higher reliability, since the average percentage of incorrect attributions was 3%, 5% and 9% (at the family, genus and species level) compared to respective percentages of 7%, 17% and 32% with MEGAN. Among DBs, Meta-PA provided the most specific results, due to the lowest rate of "misassignments", whereas NCBI-BFV and TrEMBL-BFV performed worse in this respect. To the best of our knowledge, the data shown here represent the first comparative evaluation of tools enabling biodiversity analysis of metaproteome samples. In general, Unipept appeared to be more straightforward for this purpose (in terms of user-friendliness, analysis time, and reliability of the output), even though MEGAN can also provide functional and pathway information which are key for metaproteomic studies. In particular, two parallel MEGAN analyses were carried out as suggested by MEGAN developers (Huson *et al.*, 2011): the first using peptide sequences as BLASTP input, and the second using the inferred protein sequences to avoid issues due to the extreme shortness of peptide sequences. The second analysis produced a higher amount of information, but reliability of taxonomic attributions was rather poor, consistently to the protein inference issues which have to be expected in a metaproteomic experiment; therefore, we chose to use only the data obtained using peptide sequences for comparison with Unipept data, also taking into account the peptide-centric nature of shotgun proteomics. However, it has also to be mentioned that modifying the MEGAN parameter “Min Score” (which was not changed from the default settings in this study) may have led to different results, especially when dealing with peptide sequences.

With regard to the distribution of the taxon-specific peptides among the different microbial strains, no yeast-specific peptides could be identified using Meta-PA, because of the total lack of eukaryotic sequences in this DB. Bacterial family distribution was instead comparable among all DBs. Going down to the species level, the best coverage was achieved by SGA-PA, followed by NCBI-BFV and TrEMBL-BFV which provided similar results. Conversely, SP-BFV failed to detect peptides belonging to the species with lower level of genomic characterization (such as *B. laterosporus* and *R. glutinis*, since no protein sequences from these species were included within SwissProt records at the time of this study). *E. faecalis* and *E. coli* were significantly underrepresented with all DBs.

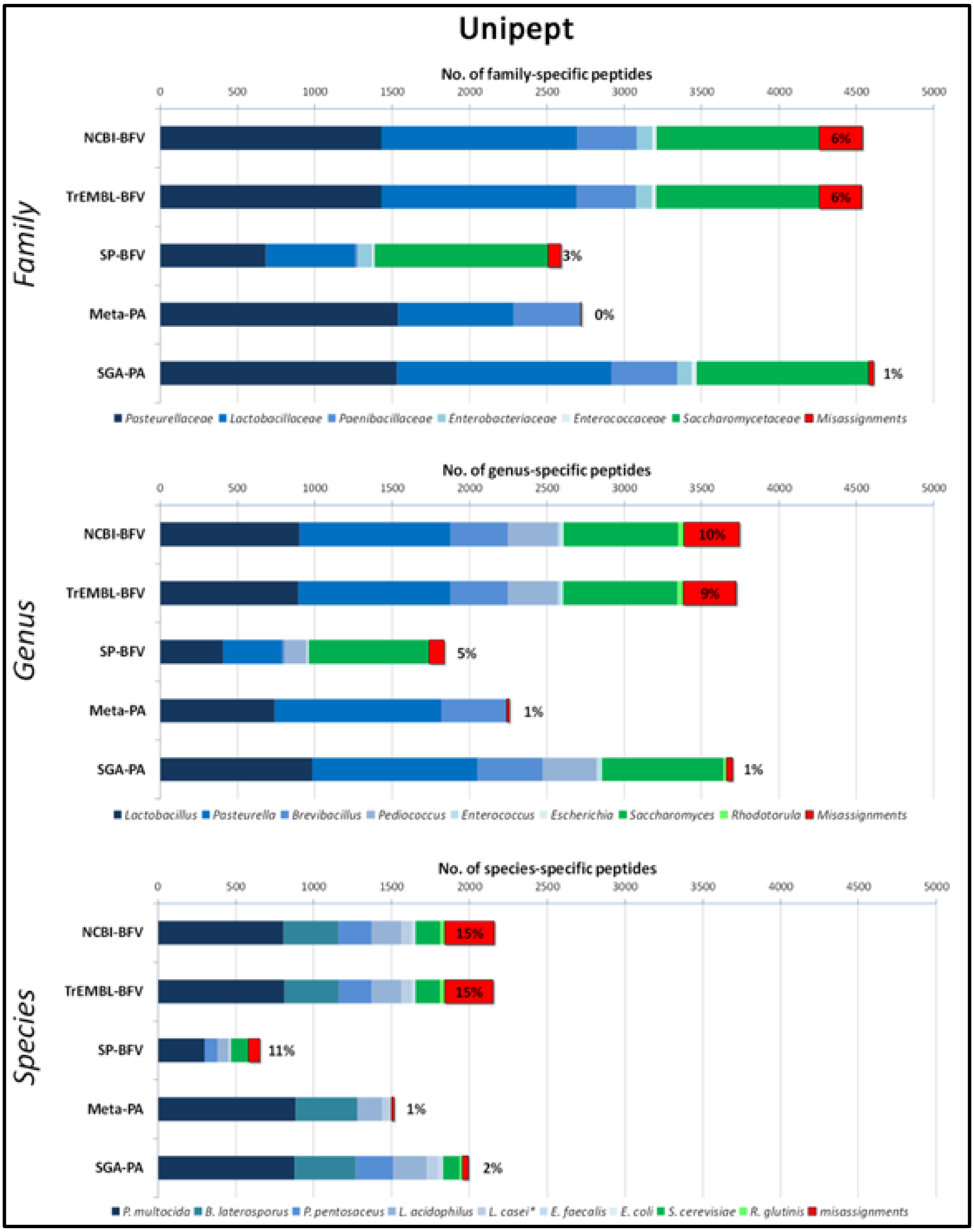


Figure 5 - 8. Reliability of taxonomic attribution using Unipept. Bar graphs showing taxonomic distribution of family (top), genus (middle) and species (bottom) specific peptides identified with different DBs, according to Unipept LCA analysis. Red rectangles illustrate misassignments (i.e. attributions to taxa not actually present in the 9MM), with indication of their percentage for each DB. Bacterial taxa are represented by various shades of blue, whereas yeast taxa by shades of green.

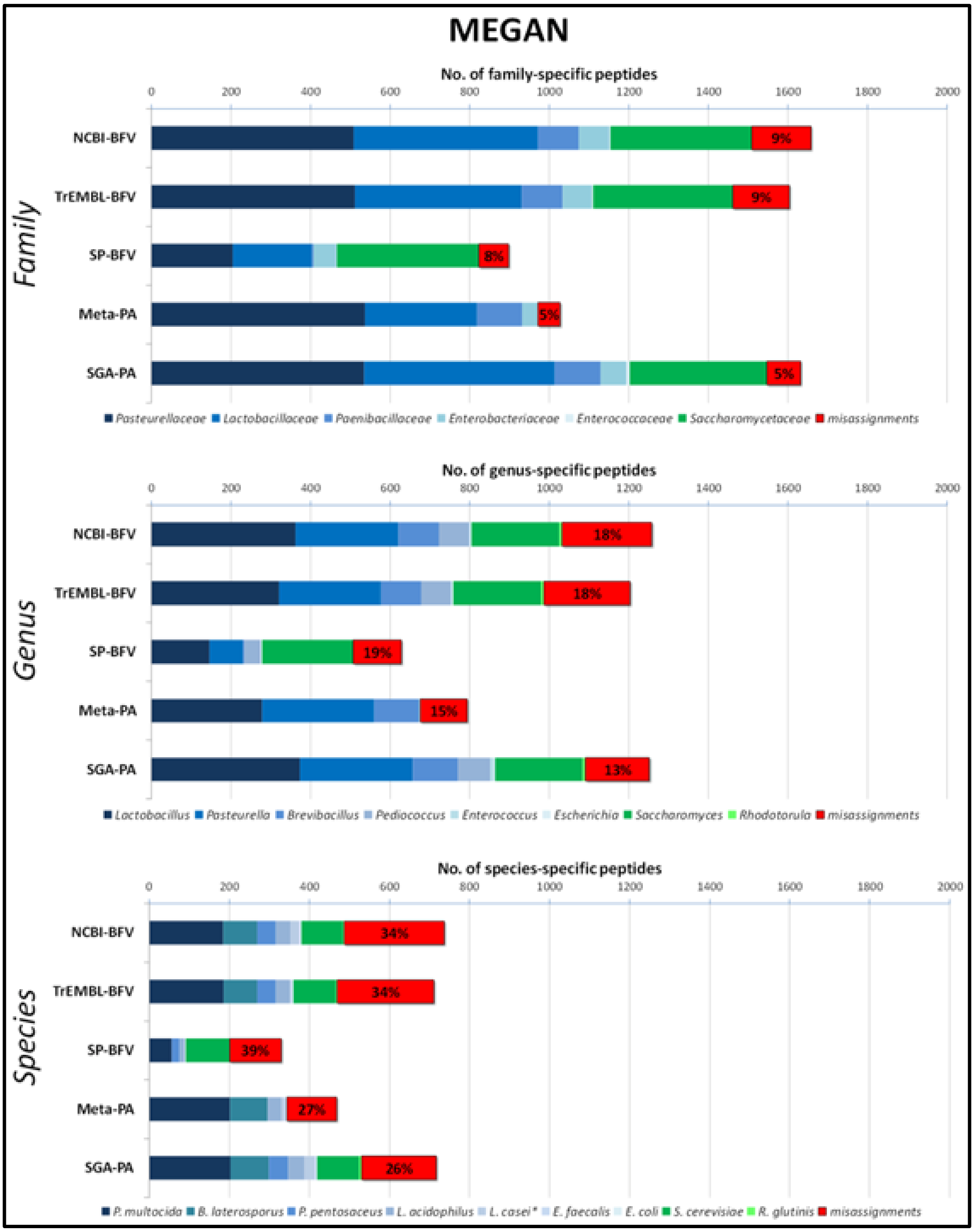


Figure 5 - 9. Reliability of taxonomic attribution using MEGAN. Bar graphs showing taxonomic distribution of family (top), genus (middle) and species (bottom) specific peptides identified with different DBs, according MEGAN LCA analysis. Red rectangles illustrate misassignments (i.e. attributions to taxa not actually present in the 9MM), with indication of their percentage for each DB. Bacterial taxa are represented by various shades of blue, whereas yeast taxa by shades of green.

In addition, when considering the overall number of families, genera, and species found with the different DBs, results were very far from the expected value. As an example, Unipept analysis of peptides identified using TrEMBL-BFV revealed the (purported) presence of 124 different families, 215 different genera, and 249 different species within the 9MM (in this case, MEGAN provided generally a lower number of false positives compared to Unipept). This, together with the non-negligible percentage of "misassignments" described above, demonstrates that taxonomic information gathered without adequate filtering can provide confounding information, dramatically decreasing the reliability of metaproteomic data. In keeping with this, an empirical filter was devised with the aim of eliminating false positive attributions and making the final result as similar as possible to the actual 9MM composition. Upon iterative analyses, a threshold corresponding to 0.5% of the total number of taxon-specific peptides was set, thus defining the taxa exhibiting a number of peptides below such value as false positives. As shown in Figure 5-10 ("u" indicates unfiltered data, whereas "f" indicates filtered data), in most cases the application of this filter allowed the elimination of all incorrect taxa (in red) without (or with only slight) loss of information about the actually present strains (in green).

The establishment of an empirical threshold to filter taxonomic classification, in order to discard false positive attributions, has been possible by analyzing a simple microbial community of known composition, and then searching for an optimized filter allowing the maximization of the real positive attributions and the minimization of the false positive ones. Specifically, the current version of Unipept does not allow the user to set a threshold (it should be done manually by parsing the csv output file); conversely, MEGAN includes a "Min Support" filter that can be easily modified according to the user's need. In particular, only two interesting reports (from the same research group) described the use of MEGAN for metaproteomic data analysis, and the first clearly stated that "because the number of reads in this proteomic dataset was considerably smaller than the thousands usual in a metagenomic dataset, the number of reads required for a taxon assignment was set to one" (Jagtap *et al.*, 2012; Rudney *et al.*, 2010). Here, we demonstrate that using such a low threshold can give rise to a significant percentage of misassignments. Clearly, the particular threshold adopted in this study might not be adequate for more complex environmental samples; however, our results underline that the raw taxonomic data may contain a significant share of false positives, and therefore strongly suggest a critical examination of the results. These incorrect species attributions might be generally due to the incompleteness of the genomic characterization of the species contained in a given microbial community. For instance, several strains of species "A" has been sequenced, and therefore different sequence variants are available in a DB. Conversely, species "B" has been less studied, and a single strain has been sequenced. As a consequence, an unknown sequence polymorphism (or even an inaccuracy in the deposited genome sequences) for species "B", which is shared with a species "A" strain, causes the erroneous attribution of its peptides to species "A", just for differences in the degree of information available for the two related species.

We also sought to investigate the taxonomic features of 6FT-unique peptide sequences. In fact, 783 peptides were identified only using 6FT DBs (Meta-6FT or SGA-6FT), since their sequence was absent from the corresponding predicted and annotated DBs. To this aim, the 6FT-unique sequences were classified based on the individual genome of origin (this information was available only for SGA-6FT), as well as subjected to BLAST sequence similarity analysis (both Meta-6FT and SGA-6FT). As a result, among 675 6FT-specific peptides detected using SGA-6FT, 77% matched with sequences belonging to *R. glutinis* genome, followed by 12% from *S. cerevisiae* (thus nearly 90% were from yeast sequences), about 4% each from *P. multocida* and *L. casei*, and an additional 4% from the remaining microbes.

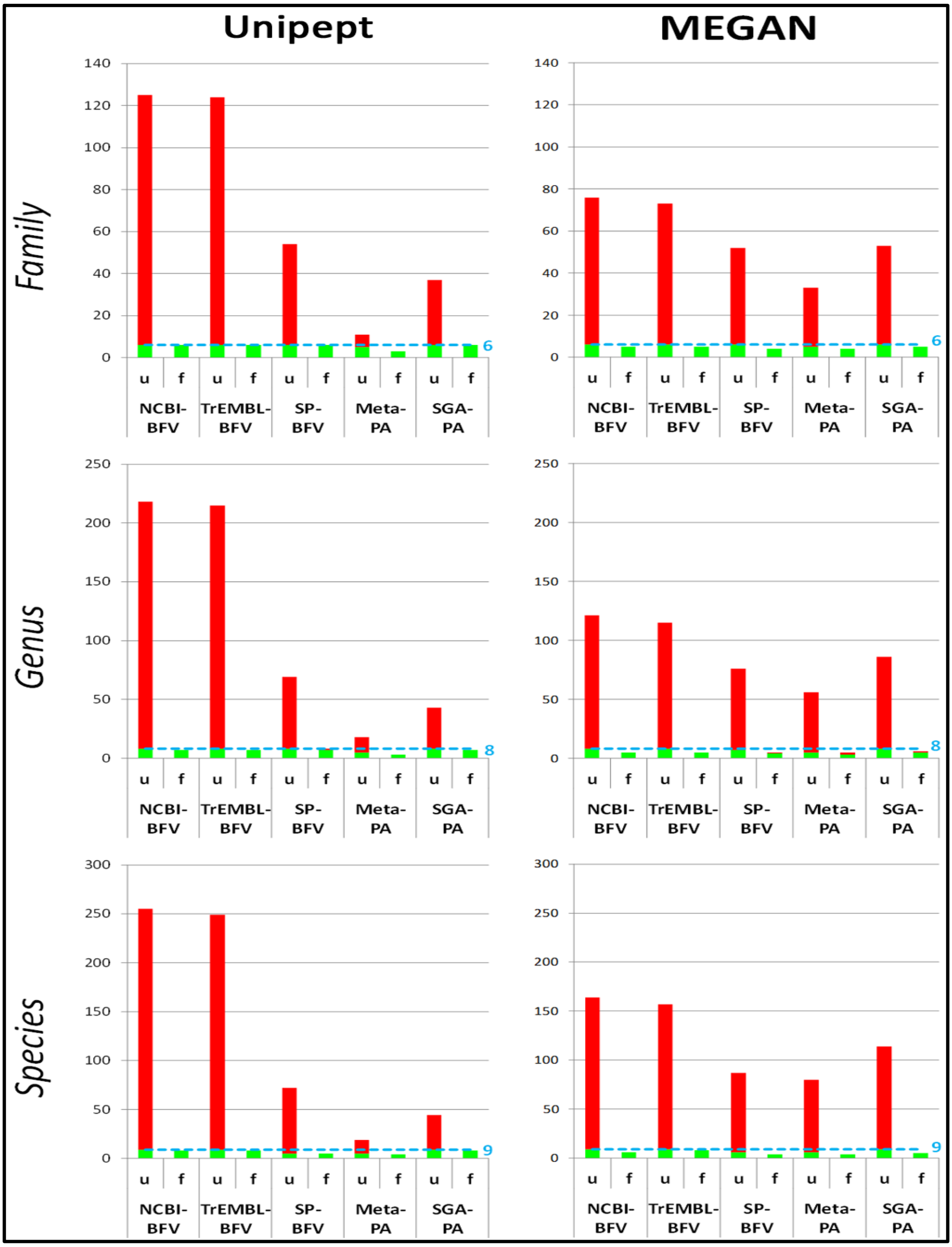


Figure 5 - 10. Improvement of the reliability of taxonomic attribution upon data filtering. Histograms showing the number of families (top), genera (middle) and species (bottom) detected upon Unipept (left) or MEGAN (right) LCA analysis using different DBs, before and after the application of a filter based on the number of taxon-specific peptides (u, unfiltered; f, filtered). The threshold was set to 0.5% of the overall number of peptides unambiguously assigned to a taxon at a particular taxonomic rank level (family, genus or species). Correct and incorrect attributions are represented in green and red, respectively. The light blue lines and numbers correspond to the number of families, genera or species actually present in the 9MM.

This result was confirmed by BLAST analysis, since 72% of the aligned sequences were found as significantly homologous to yeast sequences (21 to *R. glutinis*, 10 to *S. cerevisiae* and 34 to other Fungi).

The data presented here highlight that further efforts are needed to optimize characterization of fungal species, and in particular to enable an efficient extraction of yeast DNA together with the most accessible bacterial DNA. As above mentioned, the metagenome-derived DB was almost totally lacking eukaryotic sequences, thus impairing the identification of the corresponding peptides upon shotgun MS analysis. When considering only bacterial data, results attained using Meta-PA were comparable to those obtained with the remaining DBs (for example, 3,729 bacterial peptide identified with Meta-PA versus 4,698 with SGA-PA and 4,601 with TrEMBL-BFV). On the contrary, the exploitation of a proteogenomic approach can be useful mostly to increase yeast metaproteome coverage (rather than for the bacterial counterpart), most likely in view of the presence of alternative or non-conventional splicing forms in eukaryotes (Prasad *et al.*, 2012). It has also to be recognized that alternative bioinformatic strategies might have been used for genome sequence assembly, CDS finding, and gene annotation, especially to improve the quality of the 9MM-H metagenome draft which was not satisfactory (maybe due to the extreme harshness of the extraction conditions used to improve yeast DNA yield). Therefore, we cannot exclude that the application of data analysis approaches different from the ones chosen in this study might have led to a higher metagenome, and thus metaproteome, coverage.

In conclusion, a real effort is currently being made by proteome researchers to develop new bioinformatic strategies able to tackle data analysis issues typical of the metaproteomic field (Cantarel *et al.*, 2011; Jagtap *et al.*, 2013; Muth *et al.*, 2013; Rooijers *et al.*, 2011; Seifert *et al.*, 2013). In this context the results of this study confirm that DB selection is not a trivial issue in metaproteomics: data quality and quantity can in fact dramatically vary depending on this factor.

Based on our data, the following critical consideration and suggestion can be made:

1. when possible, the parallel use of multiple DBs has to be encouraged, as different DB types can lead to highly complementary results;
2. the use of iterative metaproteomic searches with DBs of decreasing size, based on protein identification data obtained with relaxed FDR thresholds or on taxonomic information obtained using generic DBs (as proposed in this study), can be key to achieve a wider metaproteome coverage (Jagtap *et al.*, 2013);
3. especially when dealing with poorly characterized microbial community samples, metagenomics (and, in some cases, sequencing of individual genomes) can help investigate less characterized species; however, special care needs to be taken in metagenomic data processing to ensure an adequate quality of the derived DBs (Cantarel *et al.*, 2011);
4. software enabling LCA analysis of metaproteome data (namely, Unipept and MEGAN) can provide reliable results even at the species level, but proper filters with specific thresholds (e.g. based on the total number of taxon-specific peptides, such as the one proposed above) have to be set to reduce false positive attributions.

On the whole, these data may be useful for all researchers dealing with microbiome characterization, and provide critical and concrete suggestions to improve reliability and analysis depth of metaproteomic results.

Obviously it has also to be noted that the results presented here were obtained using Percolator's and Proteome Discoverer's algorithms for FDR calculation and protein grouping, respectively. Several alternative, more sophisticated approaches are available to perform these post-processing operations (and metaproteomics-targeted software will be hopefully developed in the near future), which might deliver significantly different data (Claassen, 2012; Hoopmann and Moritz, 2013; Huang *et al.*, 2012; Nesvizhskii, 2010). Furthermore, the complexity of the lab-assembled microbial mixture used in this study was far from that of a typical “real-world” microbiome. This suggests that caution is required before extending the conclusions described here to the most heterogeneous environmental samples, and that further validation studies are needed to define an optimized pipeline for metaproteomic data analysis.

Chapter 6

Results and Discussion:

“Ovine gut microbiome characterization”

6.1 Proteomics analysis of ovine fecal samples

Fecal samples from five Sarda sheep (identified with numbers from 1 to 5) were analyzed with the proteomic pipeline developed within this project, and detailed in the "Material and Methods" section. Briefly, proteins were directly extracted from ovine stool samples using a combination of bead-beating/freeze-thawing, and subjected to FASP for clean-up and digestion, and single-run LC-MS/MS for peptide separation and identification. Then, an iterative strategy was adopted for protein identification, based on the recently published "two-step" database search method, consisting of a first unfiltered search against a large sequence database (DB), followed by a second step based on a search against a second, smaller DBs, comprising in turn all the sequences from the protein entries identified in the primary search (Jagtap *et al.*, 2013). The Table 6-1 shows the number of protein sequences, peptides, and peptide spectrum matches (PSMs) identified both overall and per sample using a FDR lower than 1%. In addition, the table also reports the number of proteins identified with at least 2 non-redundant peptide sequences.

Table 6 - 1. Number of proteins, peptides and PSMs identified in each sample, replicate and run.

Sample	Number of non-redundant proteins	Number of non-redundant Peptides	Number of PSMs	Number of non-redundant proteins (≥2 peptides)
Sheep 1	6,304	9,391	17,489	2,164
Sheep 2	9,591	13,380	22,436	2,887
Sheep 3	12,468	16,515	26,662	2,591
Sheep 4	8,537	12,435	21,780	2,764
Sheep 5	13,945	17,620	28,878	2,405
Total	35,971	44,204	117,245	5,852
Average	10,169	13,868.2	23,449	2,562
SD	3,061	3,295	4,448	287

The ovine fecal metaproteomes obtained were then analyzed to achieve a functional characterization of the microbiome. Specifically, according to UniProtKB protein family classification, 911 different protein families were identified among the five animals, 772 of which were of microbial origin (i.e. from Archaea, Bacteria, Fungi, or Viruses, ABFV; Table 6-2).

Interestingly, they cover a wide range of enzymatic, transport, and signaling functions.

The table shows in detail the microbial protein families reaching the cut-off value of 0.5% based on the number of family members or the number of PSMs assigned specifically to the particular protein family identified.

Table 6 - 2. Microbial protein families detected in the mouse stool sample above 0.5% threshold. The table is ordered by the number of family members identified.

Microbial Protein Families (Archaea-Bacteria-Fungi-Viruses)	Number of family members identified	Total PSMs	Percentage of family members identified Members	Percentage of PSMs
GTP-binding elongation factor	359	24,566	5.56%	41.81%
TonB-dependent receptor	341	1,309	5.28%	2.23%
ABC transporter	315	838	4.88%	1.43%
ELFV dehydrogenases	202	5,082	3.13%	8.65%
GAPDH	117	5,258	1.81%	8.95%
HSP70	114	1,078	1.77%	1.83%
NifJ	105	2,621	1.63%	4.46%
Class-II aa-tRNA synthetase	99	193	1.53%	0.33%
Class-I aa-tRNA synthetase	96	130	1.49%	0.22%
SDR	82	138	1.27%	0.23%
HSP60	77	657	1.19%	1.12%
Aldehyde dehydrogenase	73	92	1.13%	0.16%
Phosphoglycerate kinase	71	631	1.10%	1.07%
Ribosomal protein L7/L12P	70	589	1.08%	1.00%
Acyl-CoA dehydrogenase	66	182	1.02%	0.31%
ATPase α/β chains	64	694	0.99%	1.18%
Cation transport ATPase (P-type)	63	114	0.98%	0.19%

Microbial Protein Families (Archaea-Bacteria-Fungi-Viruses)	Number of family members identified	Total PSMs	Percentage of family members identified Members	Percentage of PSMs
RNA polymerase β chain	59	594	0.91%	1.01%
Binding-protein-dependent transport system permease	57	80	0.88%	0.14%
RNA polymerase β' chain	49	615	0.76%	1.05%
PEPCK [ATP]	46	1,294	0.71%	2.20%
Ribosomal protein L5P	45	697	0.70%	1.19%
Ribosomal protein S2P	39	798	0.60%	1.36%
Ribosomal protein S7P	39	305	0.60%	0.52%
Phosphohexose mutase	35	206	0.54%	0.35%
Ribosomal protein S4P	35	231	0.54%	0.39%
Thiolase	33	48	0.51%	0.08%
Actin	10	586	0.15%	1.00%
Other ABF families (744)	3,694	9,136	57.23%	15.55%
Total	**6,455**	**58,762**	**100%**	**100%**

6.2 Taxonomic distribution

It is also important to highlight that through the use of a metaproteomic approach we can achieve significant information about all organisms within the sample. According to this, Figure 6-1 illustrates the taxonomic distribution based to lowest common ancestor (LCA) analysis carried out on PSMs data using the Unipept web application (as described in the “Material and Methods” section) for all the five animals. The pie-chart reports all phyla reaching an average cut-off value of 0.5% of total PSMs identified among all the animals analyzed,. This threshold was selected according to the indications suggested by the analysis of a mock microbial mixture (this specific cut-off value allowed, in fact, the lowest number of taxonomic misassignments, for more details see Chapter 5), which might require adjustments depending on the particular microbial community under analysis; further studies are thus still necessary to evaluate this aspect. In total, 15 phyla overcoming this threshold were identified in the ovine fecal samples. All these phyla could have a key value according to the different kind of information researched on the basis of our experimental plan:

- the phylum Chordata (on average more than 32% of PSMs identified among all five animals), mainly represented from *Bos Bovis* species, corresponds to the “host” proteome;
- the phyla Streptophyta (11.6%) and Chlorophyta (0.75%), belonging to the Viridiplantae kingdom, represent the “nourishment” of the animals analyzed;

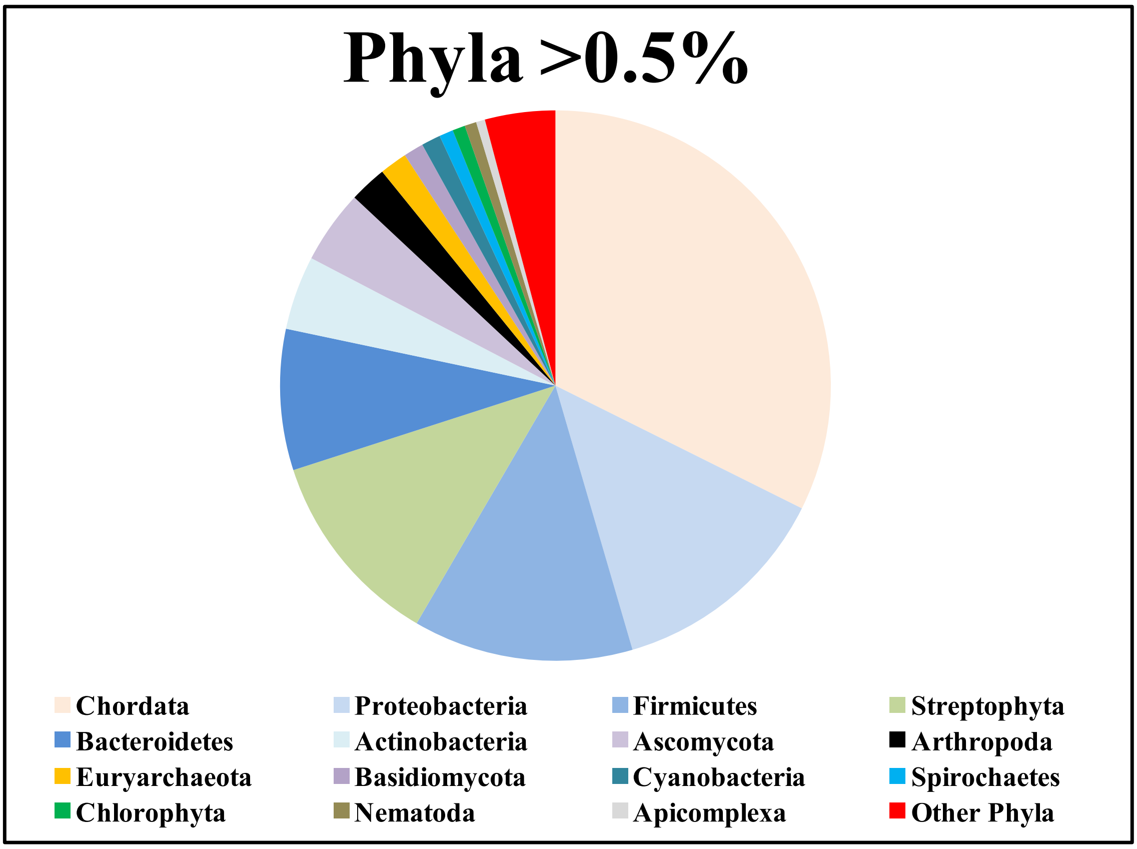


Figure 6 - 1. Distribution of phyla identified in the ovine gut reaching the 0.5% threshold. Taxonomy attribution was perfomed according to lowest common ancestor (LCA) analysis carried out on PSMs data using the Unipept web application.

- the various phyla forming the “true” ovine microbiome, namely:

    - Actinobacteria (4.4%), Bacteroidetes (8.2%), Cyanobacteria (1.1%), Firmicutes (13%), Proteobacteria (13%), and Spirochetes (0.8%), belonging to the kingdom Bacteria, that are generally the most abundant microorganisms in the gut microbiome, extensively investigated in numerous studies;
    - Ascomycota (4.3%), belonging to the kingdom Fungi, that can be important decomposers, breaking down organic materials, such as dead leaves. The components of this phylum, along with other Fungi, are able to degrade large molecules such as cellulose or lignin, and thus have important roles in host nutrient cycling and in industrial process;
    - Euryarchaeota (1.6%), belonging to the kingdom Archaea, the most important greenhouse gas (GHG) producers in the ruminant animals (Wang *et al.*, 2012);
- "exotic" phyla such as:
    - Arthropoda (2.2%), including the class *Insecta*, that can be a contaminant of the nourishment;
    - Apicomplexa (0.5%), a large group of parasitic protists belonging to the kingdom Chromalveolata, some of which indicated, for example, as abort agents in sheep such as *Toxoplasma gondii*, *Sarcocystis ovicanis* and *Sarcocystis arieticanis* that seem to be widespread in sheep flocks of Sardinia (Natale *et al.*, 2007; Porqueddu *et al.*, 2006);
    - Nematoda (0.7%), an heterogeneous group of organisms often involved in different pathologies of veterinary interest such as, for example, the *Haemonchus contortus* that is one of the most abundant infectious agents in sheep around the world, causing great economic damage to ovine breeding (Akkari *et al.*, 2013; Santos *et al.*, 2012);

- the "Other Phyla" group, comprising all phyla identified with an average abundance lower than 0.5%, include 49 different phyla, 37 of which found in all five animals.

Consequently, according to the specific aim of the project it is possible to focus the analysis on proteins belonging to one (or more than one) of the groups mentioned above.

As illustrated in Table 6-3, the host gut proteome characterization host is easily achievable thanks to more than three thousand proteins belonging to phylum Chordata. Specifically, according to UniProtKB protein family classification, 180 different protein families attributable to the host were identified among the five animals (43 of which reaching the 0.5% abundance cut-off).

Table 6 - 3. Host protein families. Protein families attributable to the host (Phylum chordata) identified according to UniProtKB classification, reaching the cut-off value of 0.5% based on the number of family members or the number of PSMs assigned.

Host Protein Families	Number of family members identified	Total PSMs	Percentage of family members identified Members	Percentage of PSMs
Intermediate filament	193	18,518	22.13%	57.96%
Peptidase S1	53	2,164	6.08%	6.77%
Serpin	26	707	2.98%	2.21%
G-protein coupled receptor 1	25	34	2.87%	0.11%
MHC class I	25	245	2.87%	0.77%
Actin	24	1,943	2.75%	6.08%
HSP70	18	422	2.06%	1.32%
Annexin	15	915	1.72%	2.86%
SDR	13	40	1.49%	0.13%
Mitochondrial carrier	12	200	1.38%	0.63%
Small GTPase	12	64	1.38%	0.20%
AB hydrolase	10	760	1.15%	2.38%
Cation transport ATPase (P-type)	10	219	1.15%	0.69%

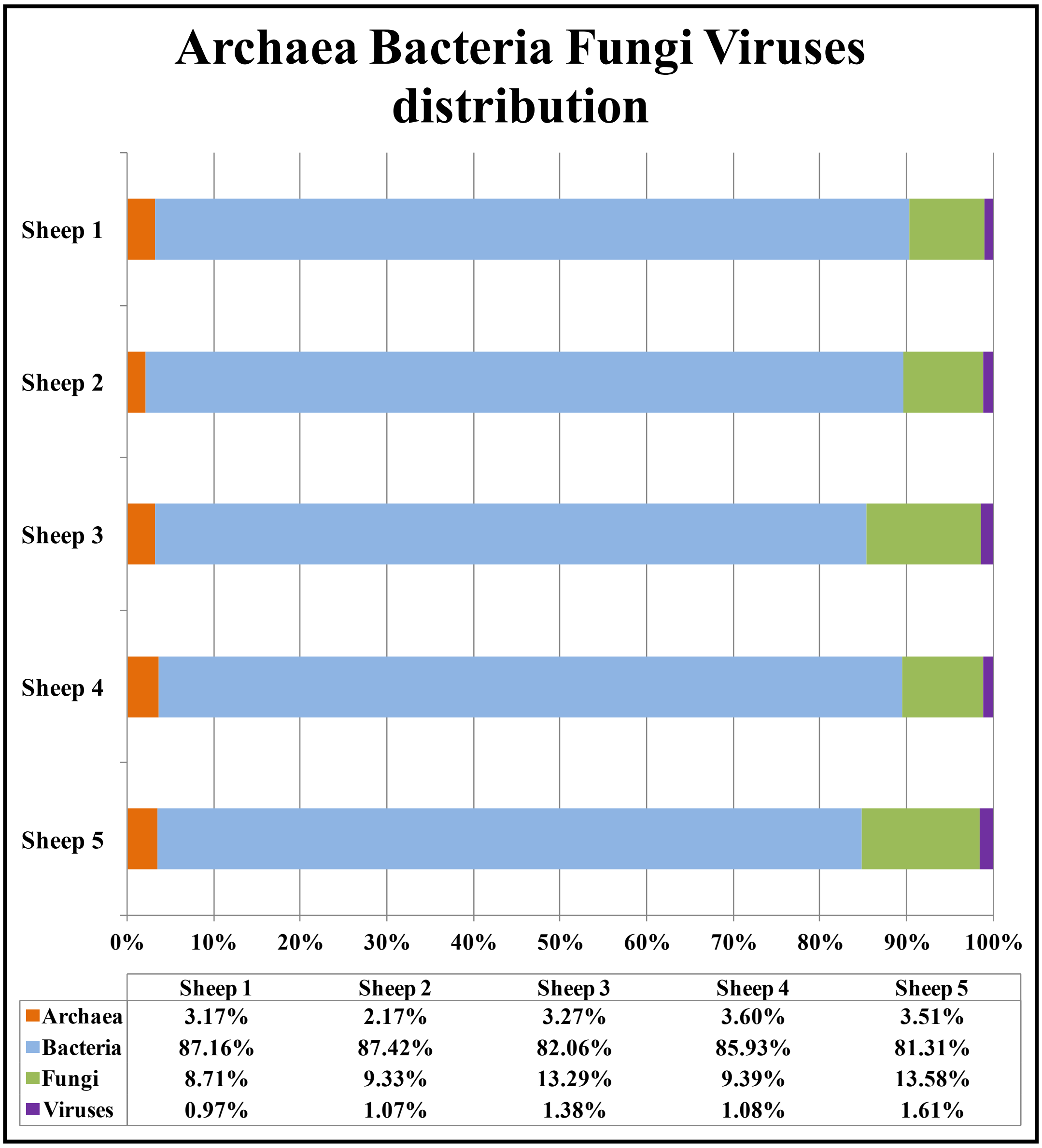
Host Protein Families	Number of family members identified	Total PSMs	Percentage of family members identified Members	Percentage of PSMs
Aldehyde dehydrogenase	9	93	1.03%	0.29%
ATPase α/β chains	9	874	1.03%	2.74%
GTP-binding elongation factor	9	260	1.03%	0.81%
Kinesin-like protein	9	23	1.03%	0.07%
Peptidase M16	9	178	1.03%	0.56%
Protein kinase	9	13	1.03%	0.04%
Acyl-CoA dehydrogenase	8	37	0.92%	0.12%
LDH/MDH	8	64	0.92%	0.20%
Peptidase T1B	8	18	0.92%	0.06%
14-3-3	7	110	0.80%	0.34%
GAPDH	7	268	0.80%	0.84%
Glycosyl hydrolase 13	7	164	0.80%	0.51%
Glycosyl hydrolase 22	7	229	0.80%	0.72%
Peptidase C14A	7	37	0.80%	0.12%
Alkaline phosphatase	6	110	0.69%	0.34%
Calycin	6	31	0.69%	0.10%
Cytochrome P450	6	9	0.69%	0.03%
Peptidase C19	6	13	0.69%	0.04%
Peptidase T1A	6	34	0.69%	0.11%
Protein disulfide isomerase	6	98	0.69%	0.31%
Tubulin	6	237	0.69%	0.74%
UDP-glycosyltransferase	6	46	0.69%	0.14%
WD repeat coronin	6	11	0.69%	0.03%
ATP:guanido phosphotransferase	5	499	0.57%	1.56%
Class-I PNDR	5	17	0.57%	0.05%
Complex I subunit 5	5	5	0.57%	0.02%
Integrin α chain	5	20	0.57%	0.06%
Phospholipase A2	5	103	0.57%	0.32%
Globin	4	362	0.46%	1.13%
Other Host Protein Families (137)	235	1,755	27.52%	5.49%
Total	**872**	**31,949**	**100%**	**100%**

It is also interesting to note that the abundance of members belonging to a specific protein family and the abundance of the specific protein family in the sample (based on PSMs) are not always correlated. For this reason, the protein families with the highest abundance account for around the 95% of total PSMs, but only for the 75% of the family members. These results suggest that the less abundant protein families can be represented by a significant number of members, although each identified with a very low number of PSMs, thus significantly limiting the frequency of their identification.

Focusing on the microbial organisms, the taxonomic distribution of PSMs assigned to Bacteria, Archaea, Fungi or Viruses was found to be almost identical among the five animals; specifically, Archaea, Bacteria, Fungi, and Viruses accounted for about 3%, 85%, 10%, and 1% of total PSMs of microbial origin, respectively. According to LCA results, the following number of different microbial taxa were unambiguously identified: 41 phyla, 80 classes, 176 orders, 387 families, 1,123 genera, and 2,304 species among all five animals (Figure 6-2).

6.3 Investigation of the ovine "core" microbiome

The "core" microbiome of these animals was also defined, by examining the taxa reaching the cut-off value of 0.5% in all five animals. This core microbiome consisted of the following different taxa: 11 phyla, 26 classes, 43 orders, 50 families, 23 genera, and (only) 3 species (Table 6-5, at the end of this chapter, reports the complete list of microbial taxa reaching the selected cut-off value of 0.5% in at least one sheep). Figure 6-3 gives an overall representation of the huge microbial biodiversity identified in these animals.



	Sheep 1	Sheep 2	Sheep 3	Sheep 4	Sheep 5
🟧 Archaea	3.17%	2.17%	3.27%	3.60%	3.51%
🟦 Bacteria	87.16%	87.42%	82.06%	85.93%	81.31%
🟩 Fungi	8.71%	9.33%	13.29%	9.39%	13.58%
🟪 Viruses	0.97%	1.07%	1.38%	1.08%	1.61%

Figure 6 - 2. Archaea, Bacteria, Fungi, and Viruses distribution among each analyzed animals. Taxonomy attribution was perfomed according to lowest common ancestor (LCA) analysis carried out on PSMs data using the Unipept web application.

Concerning bacterial phyla, the following average distribution was observed: Firmicutes 25.8%, Proteobacteria 25.8%, Bacteroidetes 16.4%, Actinobacteria 8.6%, Cyanobacteria 2.2%, Spirochaetes 1.6%, Verrucomicrobia 0.5%. Among fungal phyla, Ascomycota and Basidiomycota amounted to 8.3% and 2.3%, respectively. The Euryarchaeota phylum (from Archaea) was 3.2%. Microbes belonging to Firmicutes and Proteobacteria were the most represented in all animals, amounting to more than 50% of microbial peptide sequences.

It is worth noting that, on the whole, about 90% of total microbial peptide sequences belonged to only six phyla (namely Proteobacteria, Firmicutes, Bacteroidetes, Actinobacteria, Ascomycota, and Euryarchaeota).

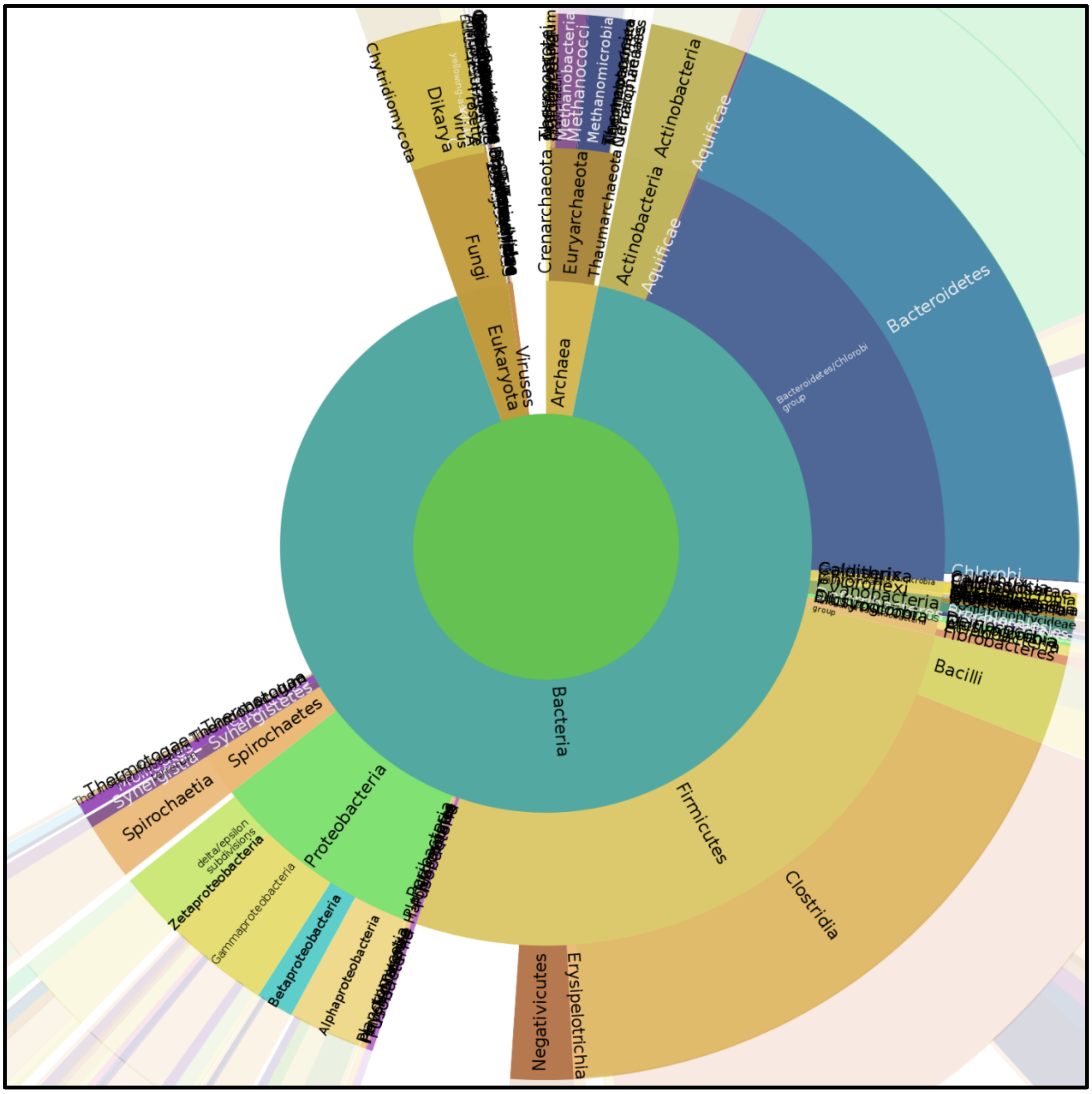


Figure 6 - 3. Overall representation of the huge microbial biodiversity identified in sheep. Taxonomy attribution was perfomed according to lowest common ancestor (LCA) analysis carried out on PSMs data using the Unipept web application.

This data give a further confirmation that fungal and archaeal organisms are significantly represented within the gut community. In this regard, interesting information is provided by archaeal proteins according to their Gene Ontology (GO) categories. In fact, as Table 6-4 shows, the fourth most represented GO category was "Methanogenesis", supporting the crucial role of these microorganisms in such processes.

Table 6 - 4. Archaeal protein gene ontology. Gene ontology categories attributable to the Archaea microorganisms identified according to UniProtKB classification, reaching the cut-off value of 0.5% based on the number of family members or the number of PSMs assigned.

Archaeal Gene ontology categories	Number of family members identified	Percentage of family members identified
ATP binding	115	6.253%
Cytoplasm	60	3.263%
DNA binding	41	2.229%
Methanogenesis	**35**	**1.903%**
Metal ion binding	35	1.903%
Catalytic activity	32	1.740%
Oxidoreductase activity	31	1.686%
Metabolic process	30	1.631%
Iron-sulfur cluster binding	28	1.523%
Integral to membrane	26	1.414%
Flavin adenine dinucleotide binding	24	1.305%
Membrane	21	1.142%
One-carbon metabolic process	19	1.033%
4 iron, 4 sulfur cluster binding	19	1.033%
Hydrolase activity	18	0.979%
Structural constituent of ribosoma	17	0.924%
Zinc ion binding	17	0.924%
DNA replication	17	0.924%
Translation	16	0.870%
Magnesium ion binding	16	0.870%
nucleic acid binding	15	0.816%
sequence-specific DNA binding transcription factor activity	14	0.761%
Ribosoma	14	0.761%

Archaeal Gene ontology categories	**Number of family members identified**	**Percentage of family members identified**
coenzyme-B sulfoethylthiotransferase activity	14	0.761%
electron carrier activity	14	0.761%
regulation of transcription, DNA-dependent	13	0.707%
signal transduction by phosphorylation	12	0.653%
Proteolysis	12	0.653%
phosphorelay sensor kinase activity	12	0.653%
DNA repair	12	0.653%

Along with the above mentioned phyla, also Acidobacteria, Actinobacteria, Aquificae, Chlamydiae, Chlorobi, Chloroflexi, Chytridiomycota, Crenarchaeota, Deferribacteres, Deinococcus-Thermus, Elusimicrobia, Fibrobacters, Fusobacteria, Gemmatimonadetes, Ignavibacteriae, Lentisphaerae, Microsporidia, Nitrospirae, Poribacteria, Synergistetes, Tenericutes, Thaumarchaeota, Thermodesulfobacteria, Thermotogae, and Verrucomicrobia were identified in all animals, even though at lower abundance, and may therefore be considered as the main components of the ovine core microbiome.

6.4 Conclusion

Further data are clearly needed to validate the results presented here, mainly because of the low number of samples analyzed; furthermore, considerable differences in the core microbiome composition may be expected when varying, for instance, diet, breed or farming conditions. Anyhow, these results represent the first description of the ovine fecal metaproteome, and demonstrate its outstanding biological diversity. Based on these data, large-scale studies could be carried out to correlate changes in sheep gut microbiota to zootechnical and production variables, with the final aim of optimizing livestock animals productivity or of protecting them from several disease types.

Table 6-5. Complete list of microbial taxa reaching the cut-off value of 0.5% in at least one sheep.

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Archaea	3.211%	2.208%	3.330%	3.653%	3.568%	3.194%	0.579%
Archaea|Crenarchaeota|Thermoprotei	**0.494%**	**0.482%**	**0.338%**	**0.501%**	**0.444%**	**0.452%**	**0.068%**
Archaea|Euryarchaeota	3.500%	2.117%	3.309%	3.782%	3.464%	3.235%	0.648%
Archaea|Euryarchaeota|Halobacteria	**1.011%**	**0.680%**	**1.080%**	**1.092%**	**1.186%**	**1.010%**	**0.195%**
Archaea|Euryarchaeota|Halobacteria|Halobacteriales	1.003%	0.673%	1.061%	1.089%	1.169%	0.999%	0.192%
Archaea|Euryarchaeota|Halobacteria|Halobacteriales|Halobacteriaceae	**1.182%**	**0.821%**	**1.196%**	**1.295%**	**1.274%**	**1.154%**	**0.192%**
Archaea|Euryarchaeota|Methanobacteria	0.989%	0.705%	0.714%	0.713%	0.605%	0.745%	0.144%
Archaea|Euryarchaeota|Methanobacteria|Methanobacteriales	**0.981%**	**0.698%**	**0.701%**	**0.711%**	**0.596%**	**0.738%**	**0.144%**
Archaea|Euryarchaeota|Methanobacteria|Methanobacteriales|Methanobacteriaceae	1.104%	0.791%	0.779%	0.774%	0.624%	0.814%	0.176%
Archaea|Euryarchaeota|Methanobacteria|Methanobacteriales|Methanobacteriaceae|*Methanobrevibacter*	**0.822%**	**0.427%**	**0.322%**	**0.572%**	**0.272%**	**0.483%**	**0.222%**
Archaea|Euryarchaeota|Methanobacteria|Methanobacteriales|Methanobacteriaceae|*Methanobrevibacter*|*Methanobrevibacter smithii*	0.529%	0.263%	0.198%	0.195%	0.160%	0.269%	0.150%
Archaea|Euryarchaeota|Methanomicrobia	**1.258%**	**0.507%**	**1.244%**	**1.775%**	**1.348%**	**1.226%**	**0.457%**
Archaea|Euryarchaeota|Methanomicrobia|Methanomicrobiales	0.847%	0.282%	0.777%	1.362%	0.851%	0.824%	0.383%
Archaea|Euryarchaeota|Methanomicrobia|Methanomicrobiales|Methanocorpusculaceae	**0.447%**	**0.045%**	**0.342%**	**0.954%**	**0.451%**	**0.447%**	**0.328%**
Archaea|Euryarchaeota|Methanomicrobia|Methanomicrobiales|Methanocorpusculaceae|*Methanocorpusculum*	0.451%	0.046%	0.344%	0.978%	0.457%	0.455%	0.337%

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Archaea|Euryarchaeota|Methanomicrobia|Methanomicrobiales|Methanocorpusculaceae|*Methanocorpusculum*|*Methanocorpusculum labreanum*	0.600%	0.061%	0.452%	1.291%	0.595%	0.600%	0.445%
Bacteria	**88.418%**	**88.776%**	**83.508%**	**87.183%**	**82.738%**	**86.124%**	**2.816%**
Bacteria|Acidobacteria	0.590%	0.380%	0.428%	0.363%	0.510%	0.454%	0.095%
Bacteria|Actinobacteria	**8.417%**	**7.176%**	**9.439%**	**7.779%**	**9.977%**	**8.558%**	**1.154%**
Bacteria|Actinobacteria|Actinobacteria	9.618%	8.174%	10.425%	8.768%	10.901%	9.577%	1.129%
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales	**8.473%**	**6.965%**	**9.164%**	**7.671%**	**9.646%**	**8.384%**	**1.087%**
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Actinomycetaceae	0.315%	0.463%	0.288%	0.558%	0.433%	0.411%	0.111%
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Corynebacteriaceae	**0.394%**	**0.299%**	**0.576%**	**0.414%**	**0.520%**	**0.441%**	**0.109%**
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Corynebacteriaceae|*Corynebacterium*	0.398%	0.305%	0.580%	0.424%	0.527%	0.447%	0.109%
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Frankiaceae	**0.552%**	**0.254%**	**0.480%**	**0.504%**	**0.485%**	**0.455%**	**0.116%**
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Frankiaceae|Frankia	0.557%	0.259%	0.483%	0.517%	0.492%	0.462%	0.117%
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Gordoniaceae	**0.447%**	**0.373%**	**0.555%**	**0.306%**	**0.607%**	**0.458%**	**0.124%**
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Gordoniaceae|*Gordonia*	0.451%	0.381%	0.558%	0.314%	0.615%	0.464%	0.124%
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Micrococcaceae	**0.499%**	**0.388%**	**0.502%**	**0.288%**	**0.381%**	**0.412%**	**0.090%**
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Micromonosporaceae	0.289%	0.314%	0.609%	0.288%	0.399%	0.380%	0.136%
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Mycobacteriaceae	**0.893%**	**0.732%**	**1.014%**	**1.098%**	**1.205%**	**0.988%**	**0.183%**

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Mycobacteriaceae|*Mycobacterium*	0.902%	0.732%	0.988%	1.126%	1.213%	0.992%	0.189%
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Nocardiaceae	**0.762%**	**0.762%**	**0.779%**	**0.738%**	**0.780%**	**0.764%**	**0.017%**
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Nocardiaceae|*Rhodococcus*	0.504%	0.503%	0.590%	0.535%	0.563%	0.539%	0.038%
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Pseudonocardiaceae	**0.893%**	**0.627%**	**0.673%**	**0.702%**	**0.832%**	**0.745%**	**0.112%**
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Streptomycetaceae	2.864%	2.120%	2.616%	2.321%	2.687%	2.521%	0.298%
Bacteria|Actinobacteria|Actinobacteria|Actinomycetales|Streptomycetaceae|*Streptomyces*	**2.865%**	**2.151%**	**2.544%**	**2.362%**	**2.646%**	**2.514%**	**0.272%**
Bacteria|Actinobacteria|Actinobacteria|Bifidobacteriales|Bifidobacteriaceae	0.420%	0.373%	0.395%	0.288%	0.520%	0.399%	0.084%
Bacteria|Actinobacteria|Actinobacteria|Coriobacteriales	**0.580%**	**0.612%**	**0.493%**	**0.590%**	**0.453%**	**0.546%**	**0.069%**
Bacteria|Actinobacteria|Actinobacteria|Coriobacteriales|Coriobacteriaceae	0.683%	0.747%	0.555%	0.702%	0.494%	0.636%	0.107%
Bacteria|Bacteroidetes	**16.755%**	**20.780%**	**14.024%**	**18.075%**	**12.591%**	**16.445%**	**3.250%**
Bacteria|Bacteroidetes|Bacteroidia	10.876%	14.369%	8.477%	12.242%	7.617%	10.716%	2.753%
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales	**10.792%**	**14.225%**	**8.330%**	**12.211%**	**7.507%**	**10.613%**	**2.761%**
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Bacteroidaceae	3.048%	3.823%	2.616%	3.544%	2.141%	3.034%	0.681%
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Bacteroidaceae|*Bacteroides*	**3.078%**	**3.905%**	**2.630%**	**3.635%**	**2.171%**	**3.084%**	**0.710%**
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Porphyromonadaceae	1.524%	2.210%	1.025%	1.511%	0.797%	1.413%	0.545%
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Porphyromonadaceae|*Porphyromonas*	**0.531%**	**0.412%**	**0.204%**	**0.332%**	**0.132%**	**0.322%**	**0.160%**
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Prevotellaceae	3.022%	4.211%	2.413%	3.742%	2.895%	3.256%	0.715%

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Prevotellaceae|*Prevotella*	2.733%	3.707%	2.222%	3.451%	2.681%	2.959%	0.607%
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Prevotellaceae|*Prevotella|Prevotella ruminicola*	**0.917%**	**1.415%**	**0.551%**	**1.462%**	**0.995%**	**1.068%**	**0.378%**
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Rikenellaceae	0.946%	1.463%	0.811%	1.403%	0.451%	1.015%	0.423%
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Rikenellaceae|*Alistipes*	**0.955%**	**1.495%**	**0.816%**	**1.439%**	**0.457%**	**1.033%**	**0.437%**
Bacteria|Bacteroidetes|Bacteroidia|Bacteroidales|Rikenellaceae|*Alistipes|Alistipes putredinis*	0.459%	0.546%	0.282%	0.463%	0.103%	0.371%	0.178%
Bacteria|Bacteroidetes|Cytophagia	**1.101%**	**0.779%**	**1.032%**	**0.880%**	**1.025%**	**0.963%**	**0.131%**
Bacteria|Bacteroidetes|Cytophagia|Cytophagales	1.093%	0.771%	1.014%	0.878%	1.010%	0.953%	0.128%
Bacteria|Bacteroidetes|Cytophagia|Cytophagales|Cytophagaceae	**0.709%**	**0.478%**	**0.619%**	**0.450%**	**0.511%**	**0.554%**	**0.108%**
Bacteria|Bacteroidetes|Flavobacteriia	1.820%	1.694%	1.919%	1.623%	1.719%	1.755%	0.116%
Bacteria|Bacteroidetes|Flavobacteriia|Flavobacteriales	**1.739%**	**1.640%**	**1.867%**	**1.589%**	**1.638%**	**1.695%**	**0.111%**
Bacteria|Bacteroidetes|Flavobacteriia|Flavobacteriales|Flavobacteriaceae	1.813%	1.762%	1.847%	1.745%	1.664%	1.766%	0.070%
Bacteria|Bacteroidetes|Sphingobacteriia	**0.337%**	**0.804%**	**0.810%**	**0.956%**	**0.742%**	**0.730%**	**0.233%**
Bacteria|Bacteroidetes|Sphingobacteriia|Sphingobacteriales	0.334%	0.783%	0.796%	0.953%	0.732%	0.720%	0.231%
Bacteria|Bacteroidetes|Sphingobacteriia|Sphingobacteriales|Sphingobacteriaceae	**0.210%**	**0.538%**	**0.406%**	**0.612%**	**0.329%**	**0.419%**	**0.160%**
Bacteria|Chloroflexi	0.570%	0.380%	0.358%	0.336%	0.458%	0.421%	0.095%
Bacteria|Cyanobacteria	**2.262%**	**2.171%**	**2.628%**	**1.925%**	**2.119%**	**2.221%**	**0.259%**
Bacteria|Cyanobacteria|Unclassified|Chroococcales	0.914%	0.857%	1.042%	0.802%	0.978%	0.919%	0.095%
Bacteria|Cyanobacteria|Unclassified|Nostocales	**0.624%**	**0.355%**	**0.569%**	**0.409%**	**0.366%**	**0.464%**	**0.124%**
Bacteria|Cyanobacteria|Unclassified|Nostocales|Nostocaceae	0.631%	0.329%	0.448%	0.342%	0.269%	0.404%	0.142%

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Bacteria|Cyanobacteria|Unclassified|Oscillatoriales	0.580%	0.637%	0.682%	0.605%	0.469%	0.595%	0.080%
Bacteria|Deinococcus-Thermus	0.492%	0.304%	0.375%	0.525%	0.369%	0.413%	0.092%
Bacteria|Deinococcus-Thermus|Deinococci	0.562%	0.346%	0.415%	0.592%	0.403%	0.464%	0.107%
Bacteria|Firmicutes	27.807%	28.694%	22.494%	28.937%	20.936%	25.774%	3.769%
Bacteria|Firmicutes|Bacilli	6.157%	5.775%	6.394%	6.766%	6.625%	6.343%	0.393%
Bacteria|Firmicutes|Bacilli|Bacillales	3.233%	3.024%	3.450%	3.435%	3.515%	3.331%	0.202%
Bacteria|Firmicutes|Bacilli|Bacillales|Bacillaceae	2.049%	1.702%	2.007%	2.177%	1.863%	1.960%	0.182%
Bacteria|Firmicutes|Bacilli|Bacillales|Bacillaceae|*Bacillus*	1.698%	1.312%	1.696%	1.569%	1.459%	1.547%	0.165%
Bacteria|Firmicutes|Bacilli|Bacillales|Bacillaceae|*Bacillus*|*Bacillus cereus*	0.565%	0.384%	0.508%	0.414%	0.412%	0.457%	0.076%
Bacteria|Firmicutes|Bacilli|Bacillales|Paenibacillaceae	0.867%	0.926%	0.961%	0.936%	0.979%	0.934%	0.043%
Bacteria|Firmicutes|Bacilli|Bacillales|Paenibacillaceae|*Paenibacillus*	0.663%	0.686%	0.655%	0.775%	0.765%	0.709%	0.057%
Bacteria|Firmicutes|Bacilli|Lactobacillales	2.809%	2.595%	2.796%	3.253%	2.990%	2.889%	0.247%
Bacteria|Firmicutes|Bacilli|Lactobacillales|Enterococcaceae	0.473%	0.373%	0.512%	0.558%	0.451%	0.473%	0.069%
Bacteria|Firmicutes|Bacilli|Lactobacillales|Enterococcaceae|*Enterococcus*	0.425%	0.305%	0.451%	0.535%	0.352%	0.413%	0.089%
Bacteria|Firmicutes|Bacilli|Lactobacillales|Lactobacillaceae	0.893%	0.821%	1.036%	1.116%	1.040%	0.981%	0.120%
Bacteria|Firmicutes|Bacilli|Lactobacillales|Lactobacillaceae|*Lactobacillus*	0.902%	0.824%	0.955%	1.052%	0.984%	0.944%	0.086%
Bacteria|Firmicutes|Bacilli|Lactobacillales|Streptococcaceae	1.130%	1.120%	1.036%	1.367%	1.075%	1.145%	0.130%
Bacteria|Firmicutes|Bacilli|Lactobacillales|Streptococcaceae|*Streptococcus*	1.114%	1.098%	0.913%	1.236%	1.002%	1.073%	0.122%
Bacteria|Firmicutes|Clostridia	19.236%	21.405%	14.264%	19.296%	12.023%	17.245%	3.925%

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Bacteria|Firmicutes|Clostridia|Clostridiales	17.770%	20.088%	13.069%	18.097%	10.903%	15.985%	3.835%
Bacteria|Firmicutes|Clostridia|Clostridiales|Clostridiaceae	**3.915%**	**3.718%**	**2.765%**	**3.293%**	**2.739%**	**3.286%**	**0.537%**
Bacteria|Firmicutes|Clostridia|Clostridiales|Clostridiaceae|*Clostridium*	3.794%	3.570%	2.673%	3.211%	2.575%	3.165%	0.537%
Bacteria|Firmicutes|Clostridia|Clostridiales|Eubacteriaceae	**0.972%**	**0.866%**	**0.758%**	**0.990%**	**0.754%**	**0.868%**	**0.113%**
Bacteria|Firmicutes|Clostridia|Clostridiales|Eubacteriaceae|*Eubacterium*	0.584%	0.610%	0.451%	0.591%	0.571%	0.561%	0.063%
Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae	**2.181%**	**3.151%**	**2.028%**	**2.699%**	**1.733%**	**2.358%**	**0.565%**
Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|*Blautia*	0.345%	0.549%	0.290%	0.517%	0.299%	0.400%	0.124%
Bacteria|Firmicutes|Clostridia|Clostridiales|Peptococcaceae	**0.604%**	**0.672%**	**0.801%**	**0.576%**	**0.754%**	**0.681%**	**0.096%**
Bacteria|Firmicutes|Clostridia|Clostridiales|Peptostreptococcaceae	0.394%	0.597%	0.342%	0.468%	0.373%	0.435%	0.102%
Bacteria|Firmicutes|Clostridia|Clostridiales|Ruminococcaceae	**2.049%**	**2.150%**	**1.260%**	**2.375%**	**0.962%**	**1.759%**	**0.613%**
Bacteria|Firmicutes|Clostridia|Clostridiales|Ruminococcaceae|*Faecalibacterium|Faecalibacterium prausnitzii*	0.353%	0.566%	0.325%	0.341%	0.160%	0.349%	0.144%
Bacteria|Firmicutes|Clostridia|Clostridiales|Ruminococcaceae|*Ruminococcus*	**0.982%**	**1.022%**	**0.612%**	**1.218%**	**0.501%**	**0.867%**	**0.300%**
Bacteria|Firmicutes|Clostridia|Clostridiales|Ruminococcaceae|*Ruminococcus|Ruminococcus champanellensis*	0.318%	0.344%	0.297%	0.560%	0.149%	0.333%	0.148%
Bacteria|Firmicutes|Clostridia|Clostridiales|Unclassified|*Pseudoflavonifractor*	**0.902%**	**0.427%**	**0.258%**	**0.351%**	**0.176%**	**0.423%**	**0.284%**
Bacteria|Firmicutes|Clostridia|Clostridiales|Unclassified|*Pseudoflavonifractor|Pseudoflavonifractor capillosus*	1.200%	0.566%	0.339%	0.463%	0.229%	0.559%	0.380%
Bacteria|Firmicutes|Clostridia|Thermoanaerobacterales	**0.803%**	**0.600%**	**0.588%**	**0.469%**	**0.644%**	**0.621%**	**0.121%**
Bacteria|Firmicutes|Erysipelotrichia	0.562%	0.507%	0.482%	0.683%	0.476%	0.542%	0.086%
Bacteria|Firmicutes|Erysipelotrichia|Erysipelotrichales	**0.557%**	**0.502%**	**0.474%**	**0.681%**	**0.469%**	**0.537%**	**0.088%**

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Bacteria|Firmicutes|Erysipelotrichia|Erysipelotrichales|Erysipelotrichaceae	0.657%	0.612%	0.534%	0.810%	0.511%	0.625%	0.119%
Bacteria|Firmicutes|Negativicutes	1.708%	1.608%	1.553%	2.063%	1.937%	1.774%	0.219%
Bacteria|Firmicutes|Negativicutes|Selenomonadales	1.695%	1.591%	1.526%	2.058%	1.909%	1.756%	0.223%
Bacteria|Firmicutes|Negativicutes|Selenomonadales|Veillonellaceae	1.629%	1.478%	1.463%	1.871%	1.716%	1.631%	0.171%
Bacteria|Firmicutes|Negativicutes|Selenomonadales|Veillonellaceae|*Selenomonas*	0.584%	0.473%	0.397%	0.664%	0.519%	0.527%	0.102%
Bacteria|Planctomycetes	0.433%	0.586%	0.585%	0.431%	0.716%	0.550%	0.121%
Bacteria|Planctomycetes|Planctomycetia	0.494%	0.631%	0.608%	0.470%	0.718%	0.584%	0.102%
Bacteria|Planctomycetes|Planctomycetia|Planctomycetales	0.468%	0.588%	0.512%	0.439%	0.684%	0.538%	0.099%
Bacteria|Planctomycetes|Planctomycetia|Planctomycetales|Planctomycetaceae	0.552%	0.717%	0.566%	0.522%	0.737%	0.619%	0.100%
Bacteria|Proteobacteria	25.192%	22.712%	28.414%	23.028%	29.680%	25.805%	3.142%
Bacteria|Proteobacteria|Alphaproteobacteria	8.854%	7.283%	8.738%	6.978%	8.973%	8.165%	0.954%
Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales	3.946%	3.415%	3.886%	3.177%	3.499%	3.585%	0.325%
Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|Bradyrhizobiaceae	0.815%	1.015%	1.142%	1.098%	0.893%	0.992%	0.138%
Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|Bradyrhizobiaceae|*Bradyrhizobium*	0.265%	0.549%	0.676%	0.535%	0.439%	0.493%	0.153%
Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|Methylobacteriaceae	0.552%	0.388%	0.352%	0.252%	0.399%	0.389%	0.108%
Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|Phyllobacteriaceae	0.578%	0.523%	0.737%	0.702%	0.468%	0.601%	0.115%
Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|Rhizobiaceae	1.419%	1.120%	1.110%	0.900%	1.127%	1.135%	0.185%
Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|Rhizobiaceae|*Rhizobium*	0.796%	0.641%	0.741%	0.424%	0.571%	0.635%	0.146%

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Bacteria|Proteobacteria|Alphaproteobacteria|Rhodobacterales	2.274%	1.799%	2.056%	1.710%	2.004%	1.969%	0.222%
Bacteria|Proteobacteria|Alphaproteobacteria|Rhodobacterales|Rhodobacteraceae	**2.365%**	**1.941%**	**2.071%**	**1.871%**	**1.941%**	**2.038%**	**0.196%**
Bacteria|Proteobacteria|Alphaproteobacteria|Rhodospirillales	1.271%	0.796%	0.929%	0.772%	1.129%	0.979%	0.216%
Bacteria|Proteobacteria|Alphaproteobacteria|Rhodospirillales|Acetobacteraceae	**0.631%**	**0.448%**	**0.491%**	**0.468%**	**0.511%**	**0.510%**	**0.072%**
Bacteria|Proteobacteria|Alphaproteobacteria|Rhodospirillales|Rhodospirillaceae	0.867%	0.523%	0.555%	0.450%	0.719%	0.623%	0.168%
Bacteria|Proteobacteria|Alphaproteobacteria|Sphingomonadales	**0.713%**	**0.392%**	**0.777%**	**0.363%**	**1.026%**	**0.654%**	**0.279%**
Bacteria|Proteobacteria|Alphaproteobacteria|Sphingomonadales|Sphingomonadaceae	0.736%	0.448%	0.833%	0.396%	0.997%	0.682%	0.256%
Bacteria|Proteobacteria|Betaproteobacteria	**4.427%**	**4.031%**	**5.796%**	**3.701%**	**5.810%**	**4.753%**	**0.992%**
Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales	3.367%	3.183%	4.653%	2.981%	4.445%	3.726%	0.768%
Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Alcaligenaceae	**0.525%**	**0.388%**	**0.662%**	**0.360%**	**0.503%**	**0.488%**	**0.121%**
Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Burkholderiaceae	1.498%	1.956%	2.466%	1.709%	2.323%	1.990%	0.406%
Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Burkholderiaceae|*Burkholderia*	**0.929%**	**1.495%**	**1.342%**	**1.107%**	**1.222%**	**1.219%**	**0.217%**
Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Burkholderiaceae|*Cupriavidus*	0.265%	0.244%	0.805%	0.314%	0.747%	0.475%	0.277%
Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Burkholderiaceae|*Cupriavidus*|*Cupriavidus metallidurans*	**0.247%**	**0.040%**	**0.650%**	**0.171%**	**0.641%**	**0.350%**	**0.280%**
Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Comamonadaceae	0.920%	0.732%	1.110%	0.702%	0.971%	0.887%	0.171%
Bacteria|Proteobacteria|Betaproteobacteria|Neisseriales|Neisseriaceae	**0.525%**	**0.463%**	**0.395%**	**0.252%**	**0.537%**	**0.435%**	**0.117%**
Bacteria|Proteobacteria|Deltaproteobacteria	3.416%	2.374%	3.125%	2.746%	3.324%	2.997%	0.433%

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Bacteria|Proteobacteria|Deltaproteobacteria|Desulfovibrionales	1.271%	0.661%	0.806%	1.014%	1.066%	0.963%	0.237%
Bacteria|Proteobacteria|Deltaproteobacteria|Desulfovibrionales|Desulfovibrionaceae	**1.445%**	**0.791%**	**0.875%**	**1.169%**	**0.988%**	**1.054%**	**0.261%**
Bacteria|Proteobacteria|Deltaproteobacteria|Desulfovibrionales|Desulfovibrionaceae|*Desulfovibrio*	1.300%	0.656%	0.762%	0.960%	0.782%	0.892%	0.253%
Bacteria|Proteobacteria|Deltaproteobacteria|Desulfuromonadales	**0.334%**	**0.269%**	**0.370%**	**0.348%**	**0.525%**	**0.369%**	**0.095%**
Bacteria|Proteobacteria|Deltaproteobacteria|Myxococcales	0.847%	0.710%	1.232%	0.575%	0.978%	0.868%	0.253%
Bacteria|Proteobacteria|Deltaproteobacteria|Myxococcales|Myxococcaceae	**0.525%**	**0.478%**	**0.673%**	**0.324%**	**0.529%**	**0.506%**	**0.125%**
Bacteria|Proteobacteria|Epsilonproteobacteria	1.326%	0.940%	0.984%	1.032%	0.928%	1.042%	0.164%
Bacteria|Proteobacteria|Epsilonproteobacteria|Campylobacterales	**1.226%**	**0.894%**	**0.872%**	**0.984%**	**0.891%**	**0.973%**	**0.148%**
Bacteria|Proteobacteria|Epsilonproteobacteria|Campylobacterales|Helicobacteraceae	0.867%	0.582%	0.534%	0.522%	0.485%	0.598%	0.154%
Bacteria|Proteobacteria|Epsilonproteobacteria|Campylobacterales|Helicobacteraceae|*Helicobacter*	**0.743%**	**0.366%**	**0.344%**	**0.369%**	**0.422%**	**0.449%**	**0.167%**
Bacteria|Proteobacteria|Epsilonproteobacteria|Campylobacterales|Helicobacteraceae|*Helicobacter*|*Helicobacter pylori*	0.565%	0.303%	0.169%	0.244%	0.172%	0.291%	0.163%
Bacteria|Proteobacteria|Gammaproteobacteria	**9.910%**	**10.610%**	**12.171%**	**10.938%**	**12.741%**	**11.274%**	**1.159%**
Bacteria|Proteobacteria|Gammaproteobacteria|Alteromonadales	1.516%	1.310%	1.422%	1.407%	1.622%	1.455%	0.119%
Bacteria|Proteobacteria|Gammaproteobacteria|Alteromonadales|Alteromonadaceae	**0.552%**	**0.672%**	**0.576%**	**0.774%**	**0.667%**	**0.648%**	**0.088%**
Bacteria|Proteobacteria|Gammaproteobacteria|Alteromonadales|Shewanellaceae	0.447%	0.343%	0.395%	0.360%	0.546%	0.418%	0.082%
Bacteria|Proteobacteria|Gammaproteobacteria|Alteromonadales|Shewanellaceae|*Shewanella*	**0.451%**	**0.351%**	**0.397%**	**0.369%**	**0.554%**	**0.424%**	**0.082%**
Bacteria|Proteobacteria|Gammaproteobacteria|Chromatiales	0.535%	0.539%	0.531%	0.439%	0.541%	0.517%	0.044%

<table>
<tr><th>TAXONOMY</th><th>S 1</th><th>S 2</th><th>S 3</th><th>S 4</th><th>S 5</th><th>AVG</th><th>SD</th></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Enterobacteriales</td><td>2.297%</td><td>2.485%</td><td>3.099%</td><td>2.845%</td><td>3.062%</td><td>2.757%</td><td>0.355%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Enterobacteriales |Enterobacteriaceae</td><td>2.706%</td><td>3.031%</td><td>3.491%</td><td>3.383%</td><td>3.337%</td><td>3.189%</td><td>0.320%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Oceanospirillales</td><td>0.357%</td><td>0.600%</td><td>0.654%</td><td>0.545%</td><td>0.461%</td><td>0.523%</td><td>0.117%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Pasteurellales| Pasteurellaceae</td><td>0.473%</td><td>0.478%</td><td>0.459%</td><td>0.540%</td><td>0.468%</td><td>0.484%</td><td>0.032%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Pseudomonadales</td><td>1.784%</td><td>2.093%</td><td>2.417%</td><td>1.982%</td><td>2.831%</td><td>2.221%</td><td>0.411%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Pseudomonadales |Moraxellaceae</td><td>0.315%</td><td>0.508%</td><td>0.470%</td><td>0.594%</td><td>0.789%</td><td>0.535%</td><td>0.174%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Pseudomonadales |Moraxellaceae|<i>Acinetobacter</i></td><td>0.239%</td><td>0.381%</td><td>0.322%</td><td>0.517%</td><td>0.536%</td><td>0.399%</td><td>0.127%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Pseudomonadales |Pseudomonadaceae</td><td>1.787%</td><td>2.046%</td><td>2.231%</td><td>1.763%</td><td>2.297%</td><td>2.025%</td><td>0.246%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Pseudomonadales |Pseudomonadaceae|<i>Pseudomonas</i></td><td>1.751%</td><td>1.983%</td><td>2.147%</td><td>1.753%</td><td>2.312%</td><td>1.989%</td><td>0.246%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Vibrionales</td><td>0.669%</td><td>0.820%</td><td>1.042%</td><td>0.908%</td><td>1.050%</td><td>0.898%</td><td>0.160%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Vibrionales| Vibrionaceae</td><td>0.788%</td><td>0.986%</td><td>1.174%</td><td>1.080%</td><td>1.144%</td><td>1.034%</td><td>0.155%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria|Vibrionales| Vibrionaceae|<i>Vibrio</i></td><td>0.557%</td><td>0.717%</td><td>0.794%</td><td>0.794%</td><td>0.747%</td><td>0.722%</td><td>0.098%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria| Xanthomonadales</td><td>0.602%</td><td>0.526%</td><td>0.578%</td><td>0.454%</td><td>0.668%</td><td>0.566%</td><td>0.081%</td></tr>
<tr><td>Bacteria|Proteobacteria|Gammaproteobacteria| Xanthomonadales|Xanthomonadaceae</td><td>0.657%</td><td>0.642%</td><td>0.619%</td><td>0.522%</td><td>0.719%</td><td>0.632%</td><td>0.072%</td></tr>
<tr><td>Bacteria|Spirochaetes</td><td>1.514%</td><td>1.835%</td><td>1.476%</td><td>1.575%</td><td>1.669%</td><td>1.614%</td><td>0.144%</td></tr>
<tr><td>Bacteria|Spirochaetes|Spirochaetia</td><td>1.730%</td><td>2.090%</td><td>1.630%</td><td>1.775%</td><td>1.824%</td><td>1.810%</td><td>0.172%</td></tr>
<tr><td>Bacteria|Spirochaetes|Spirochaetia|Spirochaetales</td><td>1.717%</td><td>2.069%</td><td>1.602%</td><td>1.770%</td><td>1.797%</td><td>1.791%</td><td>0.172%</td></tr>
</table>

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Bacteria|Spirochaetes|Spirochaetia|Spirochaetales|Spirochaetaceae	1.576%	2.016%	1.302%	1.583%	1.456%	1.587%	0.266%
Bacteria|Spirochaetes|Spirochaetia|Spirochaetales|Spirochaetaceae|*Treponema*	**1.300%**	**1.648%**	**1.031%**	**1.236%**	**1.169%**	**1.277%**	**0.230%**
Bacteria|Synergistetes	0.511%	0.369%	0.515%	0.458%	0.325%	0.436%	0.085%
Bacteria|Synergistetes|Synergistia	**0.472%**	**0.420%**	**0.434%**	**0.516%**	**0.315%**	**0.431%**	**0.075%**
Bacteria|Synergistetes|Synergistia|Synergistales	0.468%	0.416%	0.426%	0.514%	0.310%	0.427%	0.076%
Bacteria|Synergistetes|Synergistia|Synergistales|Synergistaceae	**0.552%**	**0.508%**	**0.480%**	**0.612%**	**0.338%**	**0.498%**	**0.102%**
Bacteria|Tenericutes	0.315%	0.467%	0.384%	0.511%	0.318%	0.399%	0.088%
Bacteria|Tenericutes|Mollicutes	**0.360%**	**0.532%**	**0.424%**	**0.576%**	**0.347%**	**0.448%**	**0.103%**
Bacteria|Verrucomicrobia	0.590%	0.358%	0.375%	0.458%	0.694%	0.495%	0.144%
Eukaryota (Fungi)	**7.389%**	**7.926%**	**11.759%**	**8.068%**	**12.056%**	**9.440%**	**2.270%**
Eukaryota (Fungi)|Ascomycota	6.529%	7.448%	10.548%	7.376%	10.080%	8.396%	1.796%
Eukaryota (Fungi)|Ascomycota|Dothideomycetes	**1.011%**	**1.088%**	**1.283%**	**0.986%**	**1.251%**	**1.124%**	**0.136%**
Eukaryota (Fungi)|Ascomycota|Dothideomycetes|Pleosporales	0.557%	0.563%	0.663%	0.484%	0.549%	0.563%	0.064%
Eukaryota (Fungi)|Ascomycota|Eurotiomycetes	**1.685%**	**2.226%**	**3.009%**	**2.033%**	**2.542%**	**2.299%**	**0.504%**
Eukaryota (Fungi)|Ascomycota|Eurotiomycetes|Eurotiales	0.803%	1.322%	1.734%	1.241%	1.519%	1.324%	0.348%
Eukaryota (Fungi)|Ascomycota|Eurotiomycetes|Eurotiales|Trichocomaceae	**0.946%**	**1.613%**	**1.954%**	**1.475%**	**1.655%**	**1.529%**	**0.370%**
Eukaryota (Fungi)|Ascomycota|Eurotiomycetes|Eurotiales|Trichocomaceae|*Aspergillus*	0.292%	0.549%	0.784%	0.461%	0.932%	0.604%	0.255%
Eukaryota (Fungi)|Ascomycota|Eurotiomycetes|Eurotiales|Trichocomaceae|*Talaromyces*	**0.212%**	**0.458%**	**0.515%**	**0.498%**	**0.325%**	**0.402%**	**0.130%**
Eukaryota (Fungi)|Ascomycota|Eurotiomycetes|Onygenales	0.780%	0.771%	1.099%	0.726%	0.795%	0.834%	0.150%

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Eukaryota (Fungi)|Ascomycota|Eurotiomycetes|Onygenales|Arthrodermataceae	0.368%	0.358%	0.555%	0.288%	0.373%	0.388%	0.099%
Eukaryota (Fungi)|Ascomycota|Leotiomycetes	**0.404%**	**0.519%**	**0.733%**	**0.425%**	**0.581%**	**0.533%**	**0.133%**
Eukaryota (Fungi)|Ascomycota|Leotiomycetes|Helotiales	0.379%	0.331%	0.559%	0.333%	0.509%	0.422%	0.106%
Eukaryota (Fungi)|Ascomycota|Saccharomycetes	**1.573%**	**1.422%**	**2.054%**	**1.790%**	**2.009%**	**1.770%**	**0.273%**
Eukaryota (Fungi)|Ascomycota|Saccharomycetes|Saccharomycetales	1.561%	1.408%	2.019%	1.785%	1.972%	1.749%	0.263%
Eukaryota (Fungi)|Ascomycota|Saccharomycetes|Saccharomycetales|Saccharomycetaceae	**0.893%**	**0.896%**	**1.121%**	**1.169%**	**1.092%**	**1.034%**	**0.131%**
Eukaryota (Fungi)|Ascomycota|Sordariomycetes	2.360%	2.758%	3.935%	2.609%	4.051%	3.142%	0.790%
Eukaryota (Fungi)|Ascomycota|Sordariomycetes|Glomerellales	**0.379%**	**0.588%**	**0.787%**	**0.409%**	**0.795%**	**0.591%**	**0.199%**
Eukaryota (Fungi)|Ascomycota|Sordariomycetes|Glomerellales|Glomerellaceae	0.315%	0.523%	0.737%	0.378%	0.615%	0.514%	0.172%
Eukaryota (Fungi)|Ascomycota|Sordariomycetes|Glomerellales|Glomerellaceae|Colletotrichum	**0.133%**	**0.320%**	**0.515%**	**0.295%**	**0.360%**	**0.325%**	**0.137%**
Eukaryota (Fungi)|Ascomycota|Sordariomycetes|Hypocreales	1.070%	1.151%	1.592%	1.332%	1.511%	1.331%	0.224%
Eukaryota (Fungi)|Ascomycota|Sordariomycetes|Hypocreales|Nectriaceae	**0.473%**	**0.597%**	**0.779%**	**0.846%**	**0.572%**	**0.653%**	**0.154%**
Eukaryota (Fungi)|Ascomycota|Sordariomycetes|Hypocreales|Nectriaceae|*Fusarium*	0.292%	0.320%	0.515%	0.627%	0.343%	0.420%	0.145%
Eukaryota (Fungi)|Ascomycota|Sordariomycetes|Magnaporthales	**0.491%**	**0.331%**	**0.531%**	**0.393%**	**0.525%**	**0.454%**	**0.088%**
Eukaryota (Fungi)|Ascomycota|Sordariomycetes|Magnaporthales|Magnaporthaceae	0.578%	0.403%	0.598%	0.468%	0.572%	0.524%	0.084%
Eukaryota (Fungi)|Ascomycota|Sordariomycetes|Sordariales	**0.312%**	**0.441%**	**0.739%**	**0.363%**	**0.907%**	**0.552%**	**0.258%**
Eukaryota (Fungi)|Basidiomycota	2.203%	1.748%	2.462%	2.032%	3.109%	2.311%	0.516%
Eukaryota (Fungi)|Basidiomycota|Agaricomycetes	**1.258%**	**0.878%**	**1.485%**	**1.289%**	**1.614%**	**1.305%**	**0.280%**

TAXONOMY	S 1	S 2	S 3	S 4	S 5	AVG	SD
Eukaryota (Fungi)|Basidiomycota|Agaricomycetes|Agaricales	0.535%	0.379%	0.692%	0.560%	0.700%	0.573%	0.132%
Eukaryota (Fungi)|Basidiomycota|Tremellomycetes	**0.472%**	**0.482%**	**0.395%**	**0.228%**	**0.637%**	**0.443%**	**0.149%**
Eukaryota (Fungi)|Basidiomycota|Tremellomycetes|Tremellales	0.468%	0.477%	0.389%	0.227%	0.628%	0.438%	0.146%
Eukaryota (Fungi)|Basidiomycota|Tremellomycetes|Tremellales|Tremellaceae	**0.525%**	**0.358%**	**0.320%**	**0.144%**	**0.563%**	**0.382%**	**0.169%**
Eukaryota (Fungi)|Basidiomycota|Tremellomycetes|Tremellales|Tremellaceae|*Filobasidiella*	0.531%	0.366%	0.322%	0.148%	0.571%	0.388%	0.171%
Viruses	**0.982%**	**1.091%**	**1.404%**	**1.096%**	**1.638%**	**1.242%**	**0.272%**

Chapter 7

Conclusions and Future Perspectives

To summarize, during my PhD I have worked at the development of a new, rapid and efficient protocol combining bead-beating/freeze-thawing, FASP, and single-run LC-MS/MS (respectively for protein extraction, for clean-up and digestion, and for peptide separation and identification) to enhance metaproteome characterization.

The first important advantage of this pipeline is that the whole procedure can be accomplished in a minimum of ~18 h, that is 4 h less than the best performing method developed to date, which required at least 22 h solely for the 2D-LC separation. Moreover, the results obtained analyzing mock microbial mixture highlighted that the workflow can be successfully applied to the identification of proteins belonging to different microorganisms, included Fungi, usually extremely resistant to lysis. Importantly, the method showed a sensitivity down to 10^4 bacterial CFUs, a linear dynamic range of 4 orders of magnitude, and a reproducibility up to over 99%. Furthermore, using the same lab-assembled microbial mixture, the impact of different sequence DBs on metaproteome analysis was investigated. These results confirmed that DB selection can dramatically modify the quality and quantity of achievable data, and that, consequently, the choice of the protein DB must be carefully evaluated. Our data suggest that, when possible, the parallel use of multiple DBs has to be encouraged, because different DB types can lead to highly complementary results. An alternative/complementary method involves the use of iterative metaproteomic searches with DBs of decreasing size, based on protein identification data obtained with relaxed FDR thresholds or on taxonomic information obtained using generic DBs, as proposed in this study, allowing to achieve a wider metaproteome coverage. Metagenomics and, in some cases, sequencing of individual genomes can help investigate less characterized species. This study also demonstrated that software enabling LCA analysis of metaproteome data can provide reliable results even at the species level, but proper filters with specific thresholds have to be set to increase coverage and trustworthiness of metaproteomic data. As a final point, the complete metaproteomic analysis workflow was successfully applied to investigate the metaproteome of Sarda sheep, an animal of significant veterinary interest although little studied, obtaining the identification of more than 35,000 proteins belonging to more than 700 different microbial species (10 % of which of fungal origin). In keeping with these results, the workflow described here may be successfully used for the in-depth characterization of complex microbiomes.

In the next future, the results described in Chapter 5 might be further validated, evaluating the impact of different sequence databases on metaproteome analysis using "real-world" and more complex microbial community samples. This is in order to find an optimized bioinformatic pipeline that permits to maximize the information achievable from each sample. In this regard, it will be interesting to carefully investigate advantages and drawbacks, above all in terms of information increase/sequencing effort ratio, of the complementary use of several metagenomics approaches.

This information will be of pivotal importance to reliably describe the microbial community of various animals of biotechnological, veterinary, and sanitary interest. To achieve these results, the first mandatory step is the improvement of our knowledge concerning animal microbiomes in physiological condition (Costa and Weese, 2012; Dewhirst *et al.*, 2012; Hooda *et al.*, 2012). The next step builds on the need to correlate the modification of a microbial community with a given pathological status, as already illustrated in numerous studies on humans or other animals (Erickson *et al.*, 2012; Gnanandarajah *et al.*, 2012; Hwang *et al.*, 2012; Suchodolski *et al.*, 2012). The findings achieved through these steps could be applied with the aim of modulating the microbiota members to improve host life conditions; for instance, this can be useful to increase livestock animal productivity or to limit human (or more in general host) susceptibility to disease (Cani and Delzenne, 2011; O' Donnell *et al.*, 2013; Zimmer *et al.*, 2012).

Surely, microbiomes represent an amazing treasure of information concerning physiological and pathological processes, that only now we begin to investigate and understand. Without doubts, a more in-depth understanding of such microbial communities can lead to get responses to important biological questions still unanswered.

References

Abram, F., Gunnigle, E., and O'Flaherty, V. (2009). Optimisation of protein extraction and 2-DE for metaproteomics of microbial communities from anaerobic wastewater treatment biofilms. Electrophoresis *30*, 4149–4151.

Akkari, H., Jebali, J., Gharbi, M., Mhadhbi, M., Awadi, S., and Darghouth, M.A. (2013). Epidemiological study of sympatric Haemonchus species and genetic characterization of Haemonchus contortus in domestic ruminants in Tunisia. Vet. Parasitol. *193*, 118–125.

Almeida, L. a, and Araujo, R. (2013). Highlights on molecular identification of closely related species. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. *13*, 67–75.

Altschul, S.F., Madden, T.L., Schäffer, a a, Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Apajalahti, J.H.A., Särkilahti, L.K., Mäki, B.R.E., Heikkinen, J.P., Nurminen, P.H., William, E., and Ma, B.R.E. (1998). Effective Recovery of Bacterial DNA and Analysis of Community Structure in the Gastrointestinal Tract of Broiler Chickens Effective Recovery of Bacterial DNA and Percent-Guanine-Plus-Cytosine-Based Analysis of Community Structure in the Gastrointestinal .

Armengaud, J., Hartmann, E.M., and Bland, C. (2013). Proteogenomics for environmental microbiology. Proteomics *13*, 2731–2742.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., *et al.* (2011). Enterotypes of the human gut microbiome. Nature *473*, 174–180.

Bäckhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D. a, and Gordon, J.I. (2005). Host-bacterial mutualism in the human intestine. Science *307*, 1915–1920.

Bantscheff, M., Lemeer, S., Savitski, M.M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal. Bioanal. Chem. *404*, 939–965.

Bao, L., Huang, Q., Chang, L., Sun, Q., Zhou, J., and Lu, H. (2012). Cloning and characterization of two β-glucosidase/xylosidase enzymes from yak rumen metagenome. Appl. Biochem. Biotechnol. *166*, 72–86.

Benndorf, D., Balcke, G.U., Harms, H., and von Bergen, M. (2007). Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. ISME J. *1*, 224–234.

Bik, E.M., Eckburg, P.B., Gill, S.R., Nelson, K.E., Purdom, E. a, Francois, F., Perez-Perez, G., Blaser, M.J., and Relman, D. a (2006). Molecular analysis of the bacterial microbiota in the human stomach. Proc. Natl. Acad. Sci. U. S. A. *103*, 732–737.

Blakeley, P., Overton, I.M., and Hubbard, S.J. (2012). Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. J. Proteome Res. *11*, 5221–5234.

Cai, S., and Dong, X. (2010). Cellulosilyticum ruminicola gen. nov., sp. nov., isolated from the rumen of yak, and reclassification of Clostridium lentocellum as Cellulosilyticum lentocellum comb. nov. Int. J. Syst. Evol. Microbiol. *60*, 845–849.

Calvo-Bado, L. a, Oakley, B.B., Dowd, S.E., Green, L.E., Medley, G.F., Ul-Hassan, A., Bateman, V., Gaze, W., Witcomb, L., Grogono-Thomas, R., *et al.* (2011). Ovine pedomics: the first study of the ovine foot 16S rRNA-based microbiome. ISME J. *5*, 1426–1437.

Cani, P.D., and Delzenne, N.M. (2011). The gut microbiome as therapeutic target. Pharmacol. Ther. *130*, 202–212.

Cantarel, B.L., Erickson, A.R., VerBerkmoes, N.C., Erickson, B.K., Carey, P. a, Pan, C., Shah, M., Mongodin, E.F., Jansson, J.K., Fraser-Liggett, C.M., *et al.* (2011). Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. PloS One *6*, e27173.

Caporaso, J.G., Lauber, C.L., Costello, E.K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., *et al.* (2011). Moving pictures of the human microbiome. Genome Biol. *12*, R50.

Carroll, I.M., Ringel-Kulka, T., Siddle, J.P., Klaenhammer, T.R., and Ringel, Y. (2012). Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. PloS One *7*, e46953.

Chang, L., Ding, M., Bao, L., Chen, Y., Zhou, J., and Lu, H. (2011). Characterization of a bifunctional xylanase/endoglucanase from yak rumen microorganisms. Appl. Microbiol. Biotechnol. *90*, 1933–1942.

Chistoserdova, L. (2010). Recent progress and new challenges in metagenomics for biotechnology. Biotechnol. Lett. *32*, 1351–1359.

Cho, Y.S., Lee, J.-Y., Park, K.S., and Nho, C.W. (2012). Genetics of type 2 diabetes in East Asian populations. Curr. Diab. Rep. *12*, 686–696.

Choksawangkarn, W., Edwards, N., Wang, Y., Gutierrez, P., and Fenselau, C. (2012). Comparative Study of Workflows Optimized for In-gel, In-solution, and On-filter Proteolysis in the Analysis of Plasma Membrane Proteins. 8–12.

Chourey, K., Jansson, J., Verberkmoes, N., Shah, M., Chavarria, K.L., Tom, L.M., Brodie, E.L., and Hettich, R.L. (2010). Direct Cellular Lysis / Protein Extraction Protocol for Soil Metaproteomics research articles. J. Proteome Res. *9*, 6615–6622.

Chung, H., Pamp, S.J., Hill, J. a, Surana, N.K., Edelman, S.M., Troy, E.B., Reading, N.C., Villablanca, E.J., Wang, S., Mora, J.R., *et al.* (2012). Gut immune maturation depends on colonization with a host-specific microbiota. Cell *149*, 1578–1593.

Claassen, M. (2012). Inference and validation of protein identifications. Mol. Cell. Proteomics MCP *11*, 1097–1104.

Colaert, N., Degroeve, S., Helsens, K., and Martens, L. (2011). Analysis of the resolution limitations of peptide identification algorithms. J. Proteome Res. *10*, 5555–5561.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, a S., McGarrell, D.M., Bandela, a M., Cardenas, E., Garrity, G.M., and Tiedje, J.M. (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res. *35*, D169–72.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, a S., McGarrell, D.M., Marsh, T., Garrity, G.M., *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. *37*, D141–5.

Collins, T., Gerday, C., and Feller, G. (2005). Xylanases, xylanase families and extremophilic xylanases. FEMS Microbiol. Rev. *29*, 3–23.

Costa, M.C., and Weese, J.S. (2012). The equine intestinal microbiome. Anim. Health Res. Rev. Conf. Res. Work. Anim. Dis. *13*, 121–128.

Craig, R., and Beavis, R.C. (2004). TANDEM: matching proteins with tandem mass spectra. Bioinforma. Oxf. Engl. *20*, 1466–1467.

Delmotte, N., Knief, C., Chaffron, S., Innerebner, G., Roschitzki, B., Schlapbach, R., von Mering, C., and Vorholt, J.A. (2009). Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc. Natl. Acad. Sci. U. S. A. *106*, 16428–16433.

Denef, V.J., VerBerkmoes, N.C., Shah, M.B., Abraham, P., Lefsrud, M., Hettich, R.L., and Banfield, J.F. (2009). Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. Environ. Microbiol. *11*, 313–325.

Denef, V.J., Mueller, R.S., and Banfield, J.F. (2010). AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. ISME J. *4*, 599–610.

Dewhirst, F.E., Klein, E. a, Thompson, E.C., Blanton, J.M., Chen, T., Milella, L., Buckley, C.M.F., Davis, I.J., Bennett, M.-L., and Marshall-Jones, Z. V (2012). The canine oral microbiome. PloS One *7*, e36067.

Dominguez-Bello, M.G., Costello, E.K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc. Natl. Acad. Sci. U. S. A. *107*, 11971–11975.

Duncan, M.W., Aebersold, R., and Caprioli, R.M. (2010) The pros and cons of peptide-centric proteomics. Nat. Biotechnol. 28, 659-664.

Dupont, S., Corre, E., Li, Y., Vacelet, J., and Bourguet-Kondracki, M.-L. (2013). First insights into the microbiome of a carnivorous sponge. FEMS Microbiol. Ecol. 1–12.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D. a (2005). Diversity of the human intestinal microbial flora. Science *308*, 1635–1638.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods *4*, 207–214.

Erickson, A.R., Cantarel, B.L., Lamendella, R., Darzi, Y., Mongodin, E.F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., *et al.* (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn’s disease. PloS One *7*, e49138.

Facchini, F.D. a, Reis, V.R. a, Roth, A.P., Magalhães, K. a, Peixoto-Nogueira, S.C., Casagrande, D.R., Reis, R. a, and Polizeli, M.D.L.T.M. (2012). Effects of Aspergillus spp. exogenous fibrolytic enzymes on in vitro fermentation of tropical forages. J. Sci. Food Agric. *92*, 2569–2573.

Fang, W., Fang, Z., Zhou, P., Chang, F., Hong, Y., Zhang, X., Peng, H., and Xiao, Y. (2012). Evidence for lignin oxidation by the giant panda fecal microbiome. PloS One *7*, e50312.

Ferrer, M., Ruiz, A., Lanza, F., Haange, S.-B., Oberbach, A., Till, H., Bargiela, R., Campoy, C., Segura, M.T., Richter, M., *et al.* (2012). Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. Environ. Microbiol.

Fic, E., Kedracka-Krok, S., Jankowska, U., Pirog, A., and Dziedzicka-Wasylewska, M. (2010). Comparison of protein precipitation methods for various rat brain structures prior to proteomic analysis. Electrophoresis *31*, 3573–3579.

Fouts, D.E., Pieper, R., Szpakowski, S., Pohl, H., Knoblach, S., Suh, M.-J., Huang, S.-T., Ljungberg, I., Sprague, B.M., Lucas, S.K., *et al.* (2012). Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. J. Transl. Med. *10*, 174.

García-Amado, M.A., Godoy-Vitorino, F., Piceno, Y.M., Tom, L.M., Andersen, G.L., Herrera, E. a, and Domínguez-Bello, M.G. (2012). Bacterial diversity in the cecum of the world’s largest living rodent (Hydrochoerus hydrochaeris). Microb. Ecol. *63*, 719–725.

Geer, L.Y., Markey, S.P., Kowalak, J. a, Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., and Bryant, S.H. (2004). Open mass spectrometry search algorithm. J. Proteome Res. *3*, 958–964.

Gevers, D., Knight, R., Petrosino, J.F., Huang, K., Mcguire, A.L., Birren, B.W., Nelson, K.E., White, O., and Methe, B.A. (2012). The Human Microbiome Project : A Community Resource for the Healthy Human Microbiome. PloS Biol. *10*, 6–10.

Gnanandarajah, J.S., Johnson, T.J., Kim, H.B., Abrahante, J.E., Lulich, J.P., and Murtaugh, M.P. (2012). Comparative faecal microbiota of dogs with and without calcium oxalate stones. J. Appl. Microbiol. *113*, 745–756.

Gong, X., Gruniniger, R.J., Forster, R.J., Teather, R.M., and McAllister, T. a (2013). Biochemical analysis of a highly specific, pH stable xylanase gene identified from a bovine rumen-derived metagenomic library. Appl. Microbiol. Biotechnol. *97*, 2423–2431.

Gonzalez, A., Clemente, J.C., Shade, A., Metcalf, J.L., Song, S., Prithiviraj, B., Palmer, B.E., and Knight, R. (2011). Our microbial selves: what ecology can teach us. EMBO Rep. *12*, 775–784.

Gosalbes, M.J., Durbán, A., Pignatelli, M., Abellan, J.J., Jiménez-Hernández, N., Pérez-Cobas, A.E., Latorre, A., and Moya, A. (2011). Metatranscriptomic approach to analyze the functional human gut microbiota. PloS One *6*, e17447.

Granholm, V., and Käll, L. (2011). Quality assessments of peptide-spectrum matches in shotgun proteomics. Proteomics *11*, 1086–1093.

Grice, E. a, and Segre, J. a (2012). The human microbiome: our second genome. Annu. Rev. Genomics Hum. Genet. *13*, 151–170.

Guarner, F., and Malagelada, J. (2003). Gut flora in health and disease. Lancet *360*, 512–519.

Guthals, A., and Bandeira, N. (2012). Peptide identification by tandem mass spectrometry with alternate fragmentation modes. Mol. Cell. Proteomics MCP *11*, 550–557.

Haange, S.-B., Oberbach, A., Schlichting, N., Hugenholtz, F., Smidt, H., von Bergen, M., Till, H., and Seifert, J. (2012). Metaproteome analysis and molecular genetics of rat intestinal microbiota reveals section and localization resolved species distribution and enzymatic functionalities. J. Proteome Res. *11*, 5406–5417.

Han, Y., Li, L., and Liu, J. (2013). Characterization of the airborne bacteria community at different distances from the rotating brushes in a wastewater treatment plant by 16S rRNA gene clone libraries. J. Environ. Sci. *25*, 5–15.

Harju, S., Fedosyuk, H., and Peterson, K.R. (2004). Rapid isolation of yeast genomic DNA: Bust n' Grab. BMC Biotechnol. *4*, 8.

Hendriks, a T.W.M., and Zeeman, G. (2009). Pretreatments to enhance the digestibility of lignocellulosic biomass. Bioresour. Technol. *100*, 10–18.

Hettich, R.L., Sharma, R., Chourey, K., and Giannone, R.J. (2012). Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. Curr. Opin. Microbiol. *15*, 373–380.

Hinkson, I. V, and Elias, J.E. (2011). The dynamic state of protein turnover: It's about time. Trends Cell Biol. *21*, 293–303.

Hofer, U. (2013). Viral evolution: Variation in the gut virome. Nat. Rev. Microbiol. *11*, 596–597.

Hoffmann, C., Dollive, S., Grunberg, S., Chen, J., Li, H., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. PloS One *8*, e66019.

Hooda, S., Minamoto, Y., Suchodolski, J.S., and Swanson, K.S. (2012). Current state of knowledge: the canine gastrointestinal microbiome. Anim. Health Res. Rev. Conf. Res. Work. Anim. Dis. *13*, 78–88.

De Hoog, C.L., and Mann, M. (2004). Proteomics. Annu. Rev. Genomics Hum. Genet. *5*, 267–293.

Hooper, L. V, Littman, D.R., and Macpherson, A.J. (2012). Interactions between the microbiota and the immune system. Science *336*, 1268–1273.

Hoopmann, M.R., and Moritz, R.L. (2013). Current algorithmic solutions for peptide-based proteomics data generation and identification. Curr. Opin. Biotechnol. *24*, 31–38.

Hu, Y.-J., Wang, Q., Jiang, Y.-T., Ma, R., Xia, W.-W., Tang, Z.-S., Liu, Z., Liang, J.-P., and Huang, Z.-W. (2013). Characterization of oral bacterial diversity of irradiated patients by high-throughput sequencing. Int. J. Oral Sci. *5*, 21–25.

Huang, T., Wang, J., Yu, W., and He, Z. (2012). Protein inference: a review. Brief. Bioinform. *13*, 586–614.

Huffnagle, G.B., and Noverr, M.C. (2013). The emerging world of the fungal microbiome. Trends Microbiol. *21*, 334–341.

Hugenholtz, P., and Pace, N.R. (1996). Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. Trends Biotechnol. *14*, 190–197.

Huson, D.H., and Mitra, S. (2012). Introduction to the analysis of environmental sequences: metagenomics with MEGAN. Methods Mol. Biol. Clifton NJ *856*, 415–429.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. *17*, 377–386.

Huson, D.H., Mitra, S., Ruscheweyh, H., Weber, N., and Schuster, S.C. (2011). Integrative analysis of environmental sequences using MEGAN4. 1552–1560.

Hustoft, H.K., Malerod, H., Wilson, S.R., Reubsaet, L., Lundanes, E., and Greibrokk, T. (2010). A Critical Review of Trypsin Digestion for LC-MS Based Proteomics. Integr. Proteomics.

Hwang, J.-S., Im, C.-R., and Im, S.-H. (2012). Immune disorders and its correlation with gut microbiome. Immune Netw. *12*, 129–138.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

Iliev, I.D., Funari, V. a, Taylor, K.D., Nguyen, Q., Reyes, C.N., Strom, S.P., Brown, J., Becker, C. a, Fleshner, P.R., Dubinsky, M., *et al.* (2012). Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. Science *336*, 1314–1317.

Jagtap, P., McGowan, T., and Bandhakavi, S. (2012). Deep metaproteomic analysis of human salivary supernatant. Proteomics 12, 992–1001.

Jagtap, P., Goslinga, J., Kooren, J.A., McGowan, T., Wroblewski, M.S., Seymour, S.L., and Griffin, T.J. (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics *13*, 1352–1357.

Jeffery, I.B., Claesson, M.J., O'Toole, P.W., and Shanahan, F. (2012). Categorization of the gut microbiota: enterotypes or gradients? Nat. Rev. Microbiol. *10*, 591–592.

Jenkinson, H.F. (2011). Beyond the oral microbiome. Environ. Microbiol. *13*, 3077–3087.

Jiang, L., He, L., and Fountoulakis, M. (2004). Comparison of protein precipitation methods for sample preparation prior to proteomic analysis. J. Chromatogr. A *1023*, 317–320.

Kan, J., Hanson, T.E., Ginter, J.M., Wang, K., and Chen, F. (2005). Metaproteomic analysis of Chesapeake Bay microbial communities. Saline Syst. *1*, 7.

Keiblinger, K.M., Wilhartitz, I.C., Schneider, T., Roschitzki, B., Schmid, E., Eberl, L., Riedel, K., and Zechmeister-Boltenstern, S. (2012). Soil metaproteomics - Comparative evaluation of protein extraction protocols. Soil Biol. Biochem. *54*, 14–24.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. *74*, 5383–5392.

Kelly, C.R., de Leon, L., and Jasutkar, N. (2012). Fecal microbiota transplantation for relapsing Clostridium difficile infection in 26 patients: methodology and results. J. Clin. Gastroenterol. *46*, 145–149.

Khandeparker, R., and Numan, M.T. (2008). Bifunctional xylanases and their potential use in biotechnology. J. Ind. Microbiol. Biotechnol. *35*, 635–644.

Khianngam, S., Tanasupawat, S., Akaracharanya, A., Kim, K., Lee, K., and Lee, J. (2012). Cohnella cellulosilytica sp. nov., isolated from buffalo faeces. Int. J. Syst. Evol. Microbiol. *62*, 1921–1925.

Klaassens, E.S., de Vos, W.M., and Vaughan, E.E. (2007). Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. Appl. Environ. Microbiol. *73*, 1388–1392.

Köcher, T., Pichler, P., Swart, R., and Mechtler, K. (2012). Analysis of protein mixtures from whole-cell extracts by single-run nanoLC-MS/MS using ultralong gradients. Nat. Protoc. *7*, 882–890.

Koenigsknecht, M.J., and Young, V.B. (2013). Faecal microbiota transplantation for the treatment of recurrent Clostridium difficile infection: current promise and future needs. Curr. Opin. Gastroenterol. *29*, 628–632.

Koike, S., and Kobayashi, Y. (2001). Development and use of competitive PCR assays for the rumen cellulolytic bacteria: Fibrobacter succinogenes, Ruminococcus albus and Ruminococcus flavefaciens. FEMS Microbiol. Lett. *204*, 361–366.

Kolmeder, C. a, and de Vos, W.M. (2013). Metaproteomics of our microbiome - Developing insight in function and activity in man and model systems. J. Proteomics.

Kolmeder, C. a, de Been, M., Nikkilä, J., Ritamo, I., Mättö, J., Valmu, L., Salojärvi, J., Palva, A., Salonen, A., and de Vos, W.M. (2012). Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. PloS One *7*, e29913.

Kong, H.H. (2011). Skin microbiome: genomics-based insights into the diversity and role of skin microbes. Trends Mol. Med. *17*, 320–328.

Krause, D.O., Denman, S.E., Mackie, R.I., Morrison, M., Rae, A.L., Attwood, G.T., and McSweeney, C.S. (2003). Opportunities to improve fiber degradation in the rumen: microbiology, ecology, and genomics. FEMS Microbiol. Rev. *27*, 663–693.

Lagier, J.-C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., Bittar, F., Fournous, G., Gimenez, G., Maraninchi, M., *et al.* (2012a). Microbial culturomics: paradigm shift in the human gut microbiome study. Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. *18*, 1185–1193.

Lagier, J.-C., Million, M., Hugon, P., Armougom, F. and Raoult, D. (2012b). Human gut microbiota: repertoire and variations. Front. Cell. Inf. Microbio. 2:136.

Lamendella, R., VerBerkmoes, N., and Jansson, J.K. (2012). “Omics” of the mammalian gut--new insights into function. Curr. Opin. Biotechnol. *23*, 491–500.

Lamont, R.F., Sobel, J.D., Akins, R. a, Hassan, S.S., Chaiworapongsa, T., Kusanovic, J.P., and Romero, R. (2011). The vaginal microbiome: new information about genital tract flora using molecular based techniques. BJOG Int. J. Obstet. Gynaecol. *118*, 533–549.

Leary, D.H., Hervey, W.J., Li, R.W., Deschamps, J.R., Kusterbeck, A.W., and Vora, G.J. (2012). Method development for metaproteomic analyses of marine biofilms. Anal. Chem. *84*, 4006–4013.

Lederberg, B.J., and Mccray, A.T. (2001). Ome Sweet ' Omics-- A Genealogical Treasury of Words. Scientist *15*.

Ley, R.E., Peterson, D. a, and Gordon, J.I. (2006a). Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell *124*, 837–848.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006b). Microbial ecology: human gut microbes associated with obesity. Nature *444*, 1022–1023.

Li, K., Bihan, M., and Methé, B. a (2013). Analyses of the stability and core taxonomic memberships of the human microbiome. PloS One *8*, e63139.

Liebler, D.C., and Ham, A.-J.L. (2009). Spin filter-based sample preparation for shotgun proteomics. Nat. Methods *6*, 785; author reply 785–786.

Mann, M., Kulak, N. a, Nagaraj, N., and Cox, J. (2013). The coming age of complete, accurate, and ubiquitous proteomes. Mol. Cell *49*, 583–590.

Maughan, H., Wang, P.W., Diaz Caballero, J., Fung, P., Gong, Y., Donaldson, S.L., Yuan, L., Keshavjee, S., Zhang, Y., Yau, Y.C.W., *et al.* (2012). Analysis of the cystic fibrosis lung microbiota via serial Illumina sequencing of bacterial 16S rRNA hypervariable regions. PloS One *7*, e45791.

Menon, V., Prakash, G., Prabhune, A., and Rao, M. (2010). Biocatalytic approach for the utilization of hemicellulose for ethanol production from agricultural residue using thermostable xylanase and thermotolerant yeast. Bioresour. Technol. *101*, 5366–5373.

Mesuere, B., Devreese, B., Debyser, G., Aerts, M., Vandamme, P., and Dawyndt, P. (2012). Unipept: Tryptic Peptide-Based Biodiversity Analysis of Metaproteome Samples. J. Proteome Res. *11*, 5773–5780.

Minton, K. (2012). Mucosal immunology: Don't forget our fungal friends. Nat. Rev. Immunol. *12*, 476.

Morgan, X.C., and Huttenhower, C. (2012). Chapter 12: Human microbiome analysis. PLoS Comput. Biol. *8*, e1002808.

Morris, R.M., Nunn, B.L., Frazar, C., Goodlett, D.R., Ting, Y.S., and Rocap, G. (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. ISME J. *4*, 673–685.

Mosoni, P., Martin, C., Forano, E., and Morgavi, D.P. (2011). Long-term defaunation increases the abundance of cellulolytic ruminococci and methanogens but does not affect the bacterial and methanogen diversity in the rumen of sheep. J. Anim. Sci. *89*, 783–791.

Muth, T., Benndorf, D., Reichl, U., Rapp, E., and Martens, L. (2013). Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. Mol. Biosyst. *9*, 578–585.

Nagaraj, N., Kulak, N.A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Mol. Cell. Proteomics MCP *11*, M111.013722.

Nakamura, N., Gaskins, H.R., Collier, C.T., Nava, G.M., Rai, D., Petschow, B., Russell, W.M., Harris, C., Mackie, R.I., Wampler, J.L., *et al.* (2009). Molecular ecological analysis of fecal bacterial populations from term infants fed formula supplemented with selected blends of prebiotics. Appl. Environ. Microbiol. *75*, 1121–1128.

Natale, A., Porqueddu, M., Capelli, G., Mocci, G., Marras, A., Sanna Coccone, G.N., Garippa, G., and Scala, A. (2007). Sero-epidemiological update on sheep toxoplasmosis in Sardinia, Italy. Parassitologia *49*, 235–238.

Nava, G.M., and Stappenbeck, T.S. (2011). Diversity of the autochthonous colonic microbiota. Gut Microbes *2*, 99–104.

NCBI Resource Coordinators (2013). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *41*, D8–D20.

Nelson, K.E., Weinstock, G.M., Highlander, S.K., Worley, K.C., Creasy, H.H., Wortman, J.R., Rusch, D.B., Mitreva, M., Sodergren, E., Chinwalla, A.T., *et al.* (2010). A catalog of reference genomes from the human microbiome. Science *328*, 994–999.

Nesvizhskii, A.I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteomics *73*, 2092–2123.

Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. Mol. Cell. Proteomics MCP *4*, 1419–1440.

Newton, R.J., Jones, S.E., Eiler, A., McMahon, K.D., and Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria.

O' Donnell, M.M., Harris, H.M.B., Jeffery, I.B., Claesson, M.J., Younge, B., O' Toole, P.W., and Ross, R.P. (2013). The core faecal bacterial microbiome of Irish Thoroughbred racehorses. Lett. Appl. Microbiol. *57*, 492–501.

Olsen, J. V, Godoy, L.M.F. De, Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2010). Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap. Mol. Cell. Proteomics MCP *4*, 2010–2021.

Olszak, T., An, D., Zeissig, S., Vera, M.P., Richter, J., Franke, A., Glickman, J.N., Siebert, R., Baron, R.M., Kasper, D.L., *et al.* (2012). Microbial exposure during early life has persistent effects on natural killer T cell function. Science *336*, 489–493.

Ottman, N., Smidt, H., de Vos, W.M., and Belzer, C. (2012). The function of our microbiota: who is out there and what do they do? Front. Cell. Infect. Microbiol. *2*, 104.

Pérez-Cobas, A.E., Gosalbes, M.J., Friedrichs, A., Knecht, H., Artacho, A., Eismann, K., Otto, W., Rojo, D., Bargiela, R., von Bergen, M., *et al.* (2012). Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. Gut 1–11.

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data Proteomics and 2-DE. Electrophoresis *20*, 3551–3567.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. a, Bonazzi, V., McEwen, J.E., Wetterstrand, K. a, Deal, C., *et al.* (2009). The NIH Human Microbiome Project. Genome Res. *19*, 2317–2323.

Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C., and Zhang, X. (2006). Performance evaluation of existing de novo sequencing algorithms. J. Proteome Res. *5*, 3018–3028.

Pirmoradian, M., Budamgunta, H., Chingin, K., Zhang, B., and Zubarev, R.A. (2013). Rapid and deep human proteome analysis by single- dimension shotgun proteomics Rapid and deep single-dimension shotgun proteomics. 1–32.

Polizeli, M.L.T.M., Rizzatti, a C.S., Monti, R., Terenzi, H.F., Jorge, J. a, and Amorim, D.S. (2005). Xylanases from fungi: properties and industrial applications. Appl. Microbiol. Biotechnol. *67*, 577–591.

Porqueddu, M., Marras, A., Tanda, B., Pipia, A.P., Varcasia, A., Garippa, G., and Scala, A. (2006). Apicomplexa diffusion in tissue samples from slaughtered sheep in Sardinia (Italy). Parassitologia *48*, 273–273.

Poulsen, M., and Sapountzis, P. (2012). Behind every great ant, there is a great gut. Mol. Ecol. *21*, 2054–2057.

Prasad, T.S.K., Harsha, H.C., Keerthikumar, S., Sekhar, N.R., Selvan, L.D.N., Kumar, P., Pinto, S.M., Muthusamy, B., Subbannayya, Y., Renuse, S., *et al.* (2012). Proteogenomic analysis of Candida glabrata using high resolution mass spectrometry. J. Proteome Res. *11*, 247–260.

Price, L.B., Liu, C.M., Johnson, K.E., Aziz, M., Lau, M.K., Bowers, J., Ravel, J., Keim, P.S., Serwadda, D., Wawer, M.J., *et al.* (2010). The effects of circumcision on the penis microbiome. PloS One *5*, e8422.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Quispel, a (1998). Lourens G. M. Baas Becking (1895-1963). Inspirator for many (micro)biologists. Int. Microbiol. Off. J. Span. Soc. Microbiol. *1*, 69–72.

Ram, R.J., Verberkmoes, N.C., Thelen, M.P., Tyson, G.W., Baker, B.J., Ii, R.C.B., Shah, M., Hettich, R.L., and Banfield, J.F. (2005). Community Proteomics of a Natural Microbial Biofilm. *1915*.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S.K., Mcculle, S.L., Ault, K., Peralta, L., and Forney, L.J. (2010). Vaginal microbiome of reproductive-age women.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S.K., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., *et al.* (2011). Vaginal microbiome of reproductive-age women. Proc. Natl. Acad. Sci. U. S. A. *108 Suppl 1*, 4680–4687.

Relman, D. a, and Falkow, S. (2001). The meaning and impact of the human genome sequence for microbiology. Trends Microbiol. *9*, 206–208.

Renard, B.Y., Timm, W., Kirchner, M., Steen, J.A.J., Hamprecht, F.A., and Steen, H. (2010). Estimating the Confidence of Peptide Identifications without Decoy Databases. Anal. Chem. *82*, 4314–4318.

Renuse, S., Chaerkady, R., and Pandey, A. (2011). Proteogenomics. Proteomics *11*, 620–630.

Rooijers, K., Kolmeder, C., Juste, C., Doré, J., de Been, M., Boeren, S., Galan, P., Beauvallet, C., de Vos, W.M., and Schaap, P.J. (2011). An iterative workflow for mining the human intestinal metaproteome. BMC Genomics *12*, 6.

Rotilio, D., Della Corte, A., D'Imperio, M., Coletta, W., Marcone, S., Silvestri, C., Giordano, L., Di Michele, M., and Donati, M.B. (2012). Proteomics: bases for protein complexity understanding. Thromb. Res. *129*, 257–262.

Rudney, J.D., Xie, H., Rhodus, N.L., Ondrey, F.G., and Griffin, T.J. (2010). A metaproteomic analysis of the human salivary microbiota by three-dimensional peptide fractionation and tandem mass spectrometry. Mol. Oral Microbiol. *25*, 38–49.

Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović, M., Kekkonen, R. a, Palva, A., and de Vos, W.M. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J. Microbiol. Methods *81*, 127–134.

Santamaria, M., Fosso, B., Consiglio, A., De Caro, G., Grillo, G., Licciulli, F., Liuni, S., Marzano, M., Alonso-Alemany, D., Valiente, G., *et al.* (2012). Reference databases for taxonomic assignment in metagenomics. Brief. Bioinform. *13*, 682–695.

Santos, M.C., Silva, B.F., and Amarante, A.F.T. (2012). Environmental factors influencing the transmission of Haemonchus contortus. Vet. Parasitol. *188*, 277–284.

Schneider, T., Keiblinger, K.M., Schmid, E., Sterflinger-Gleixner, K., Ellersdorfer, G., Roschitzki, B., Richter, A., Eberl, L., Zechmeister-Boltenstern, S., and Riedel, K. (2012). Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. ISME J. *6*, 1749–1762.

Segata, N., Haake, S.K., Mannon, P., Lemon, K.P., Waldron, L., Gevers, D., Huttenhower, C., and Izard, J. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol. *13*, R42.

Seidler, J., Zinn, N., Boehm, M.E., and Lehmann, W.D. (2010). De novo sequencing of peptides by MS/MS. Proteomics *10*, 634–649.

Seifert, J., Herbst, F.-A., Halkjaer Nielsen, P., Planes, F.J., Jehmlich, N., Ferrer, M., and von Bergen, M. (2013). Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. Proteomics 2786–2804.

Shahinas, D., Silverman, M., Sittler, T., and Chiu, C. (2012). Toward an Understanding of Changes in Diversity Associated with Fecal Microbiome Transplantation Based on 16S rRNA Gene Deep Sequencing. mBio *3*, 1–10.

Sharma, R., Dill, B.D., Chourey, K., Shah, M., VerBerkmoes, N.C., and Hettich, R.L. (2012). Coupling a detergent lysis/cleanup methodology with intact protein fractionation for enhanced proteome characterization. J. Proteome Res. *11*, 6008–6018.

Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., and Standing, K.G. (2001). Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal. Chem. *73*, 1917–1926.

Shibata, M., and Terada, F. (2010). Factors affecting methane production and mitigation in ruminants. Anim. Sci. J. Nihon Chikusan Gakkaihō *81*, 2–10.

Siggins, A., Gunnigle, E., and Abram, F. (2012). Exploring Mixed Microbial Community Functioning: Recent Advances in Metaproteomics. FEMS Microbiol. Ecol. *80*, 265–280.

Sogin, M.L., Morrison, H.G., Huber, J. a, Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M., and Herndl, G.J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. U. S. A. *103*, 12115–12120.

Spivak, M., Weston, J., Bottou, L., Käll, L., and Noble, W.S. (2009). Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. J. Proteome Res. *8*, 3737–3745.

Suchodolski, J.S., Markel, M.E., Garcia-Mazcorro, J.F., Unterer, S., Heilmann, R.M., Dowd, S.E., Kachroo, P., Ivanov, I., Minamoto, Y., Dillman, E.M., *et al.* (2012). The fecal microbiome in dogs with acute diarrhea and idiopathic inflammatory bowel disease. PloS One *7*, e51907.

Suen, G., Scott, J.J., Aylward, F.O., Adams, S.M., Tringe, S.G., Pinto-Tomás, A. a, Foster, C.E., Pauly, M., Weimer, P.J., Barry, K.W., *et al.* (2010). An insect herbivore microbiome with high plant biomass-degrading capacity. PLoS Genet. *6*, e1001129.

Sun, Y., and Cheng, J. (2002). Hydrolysis of lignocellulosic materials for ethanol production: a review. Bioresour. Technol. *83*, 1–11.

Sutton, J.D., Knight, R., McAllan, a B., and Smith, R.H. (1983). Digestion and synthesis in the rumen of sheep given diets supplemented with free and protected oils. Br. J. Nutr. *49*, 419–432.

Tanca, A., Biosa, G., Pagnozzi, D., Addis, M.F., and Uzzau, S. (2013). Comparison of detergent-based sample preparation workflows for LTQ-Orbitrap analysis of the *Escherichia coli* proteome. Proteomics 13, 2597-2607.

Thakur, S.S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J., and Mann, M. (2011). Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. Mol. Cell. Proteomics MCP *10*, M110.003699.

Thanakoses, P., Black, A.S., and Holtzapple, M.T. (2003). Fermentation of corn stover to carboxylic acids. Biotechnol. Bioeng. *83*, 191–200.

The UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. *40*, D71–5.

Tringe, S.G., and Hugenholtz, P. (2008). A renaissance for the pioneering 16S rRNA gene. Curr. Opin. Microbiol. *11*, 442–446.

Turnbaugh, P.J., and Gordon, J.I. (2008). An invitation to the marriage of metagenomics and metabolomics. Cell *134*, 708–713.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature *444*, 1027–1031.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature *449*, 804–810.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B. a, Affourtit, J.P., *et al.* (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480–484.

Vaudel, M., Burkhart, J.M., Sickmann, A., Martens, L., and Zahedi, R.P. (2011). Peptide identification quality control. Proteomics *11*, 2105–2114.

Verberkmoes, N.C., Russell, A.L., Shah, M., Godzik, A., Rosenquist, M., Halfvarson, J., Lefsrud, M.G., Apajalahti, J., Tysk, C., Hettich, R.L., *et al.* (2009a). Shotgun metaproteomics of the human distal gut microbiota. ISME J. *3*, 179–189.

VerBerkmoes, N.C., Denef, V.J., Hettich, R.L., and Banfield, J.F. (2009b). Functional analysis of natural microbial consortia using community proteomics. Nat. Rev. Microbiol. *7*, 196–205.

Wang, J.-K., Ye, J.-A., and Liu, J.-X. (2012). Effects of tea saponins on rumen microbiota, rumen fermentation, methane production and growth performance--a review. Trop. Anim. Health Prod. *44*, 697–706.

Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T.H., Stege, J.T., Cayouette, M., McHardy, A.C., Djordjevic, G., Aboushadi, N., *et al.* (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature *450*, 560–565.

Weese, J.S. (2013). The canine and feline skin microbiome in health and disease. Vet. Dermatol. *24*, 137–45.e31.

Wikoff, W.R., Anfora, A.T., Liu, J., Schultz, P.G., Lesley, S. a, Peters, E.C., and Siuzdak, G. (2009). Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. Proc. Natl. Acad. Sci. U. S. A. *106*, 3698–3703.

Williams, S.C.P. (2013). The other microbiome. Proc. Natl. Acad. Sci. U. S. A. *110*, 2682–2684.

Wilmes, P., and Bond, P.L. (2004). The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ. Microbiol. *6*, 911–920.

Wilmes, P., and Bond, P.L. (2006). Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol. *14*, 92–97.

Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.

De Wit, R., and Bouvier, T. (2006). "Everything is everywhere, but, the environment selects"; what did Baas Becking and Beijerinck really say? Environ. Microbiol. *8*, 755–758.

Woo, P.C.Y., Lau, S.K.P., Teng, J.L.L., Tse, H., and Yuen, K.-Y. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. *14*, 908–934.

Wright, J.C., Beynon, R.J., and Hubbard, S.J. (2010). Cross species proteomics. Methods Mol. Biol. Clifton NJ *604*, 123–135.

Xie, G., Zhang, S., Zheng, X., and Jia, W. (2013). Metabolomics approaches for characterizing metabolic interactions between host and its commensal microbes. Electrophoresis 1–12.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., and Bork, P. (2011). iPath2.0: interactive pathway explorer. Nucleic Acids Res. *39*, W412–5.

Yates, J.R., Eng, J.K., McCormack, a L., and Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal. Chem. *67*, 1426–1436.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., *et al.* (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227.

Yue, Z.-B., Li, W.-W., and Yu, H.-Q. (2013). Application of rumen microorganisms for anaerobic bioconversion of lignocellulosic biomass. Bioresour. Technol. *128*, 738–744.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhang, T., Desimone, R. a, Jiao, X., Rohlf, F.J., Zhu, W., Gong, Q.Q., Hunt, S.R., Dassopoulos, T., Newberry, R.D., Sodergren, E., *et al.* (2012). Host genes related to paneth cells and xenobiotic metabolism are associated with shifts in human ileum-associated microbial composition. PloS One *7*, e30044.

Zhu, L., Wu, Q., Dai, J., Zhang, S., and Wei, F. (2011). Evidence of cellulose metabolism by the giant panda gut microbiome. Proc. Natl. Acad. Sci. U. S. A. *108*, 17714–17719.

Zimmer, J., Lange, B., Frick, J.-S., Sauer, H., Zimmermann, K., Schwiertz, A., Rusch, K., Klosterhalfen, S., and Enck, P. (2012). A vegan or vegetarian diet substantially alters the human colonic faecal microbiota. Eur. J. Clin. Nutr. *66*, 53–60.

Zoetendal, E.G., Rajilic-Stojanovic, M., and de Vos, W.M. (2008). High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. Gut *57*, 1605–1615.

Zubarev, R. a (2013). The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics *13*, 723–726.