UNIWERSYTET ŚLĄSKI
WYDZIAŁ MATEMATYKI, FIZYKI I CHEMII
INSTYTUT FIZYKI

Robert Kwapich
276941

Badanie mikrobiomu pacjentów z toczniem rumieniowatym układowym poprzez wykorzystanie metod sekwencjonowania nowej generacji

PRACA MAGISTERSKA

Promotorzy:
dr hab. prof. UŚ Anna Michnik
Uniwersytet Śląski w Katowicach

dr Patrick Gaffney
Oklahoma Medical Research Foundation

KATOWICE 2017

Słowa kluczowe: mikrobiom, mikrobiota, toczeń rumieniowaty układowy, sekwencjonowanie nowej generacji, bioinformatyka

Oświadczenie autora pracy

Ja, niżej podpisany:

Robert Sebastian Kwapich

autor pracy dyplomowej pt. „*Badanie mikrobiomu pacjentów z toczniem rumieniowatym systemowym poprzez wykorzystanie metod sekwencjonowania nowej generacji*”

Numer albumu: **276941**

Student Wydziału Matematyki, Fizyki i Chemii Uniwersytetu Śląskiego w Katowicach kierunku studiów **Fizyka Medyczna** specjalności **promieniowanie jonizujące** oświadczam, że ww. praca dyplomowa:

- została przygotowana przeze mnie samodzielnie[1],
- nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (tekst jednolity Dz. U. z 2006 r. Nr 90, poz. 631, z późn. zm.) oraz dóbr osobistych chronionych prawem cywilnym,
- nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,
- nie była podstawą nadania dyplomu uczelni wyższej lub tytułu zawodowego ani mnie, ani innej osobie.

Oświadczam również, że treść pracy dyplomowej zamieszczonej przeze mnie w Archiwum Prac Dyplomowych jest identyczna z treścią zawartą w wydrukowanej wersji pracy.

Jestem świadomy/-a odpowiedzialności karnej za złożenie fałszywego oświadczenia.

…… …….

Data Podpis autora pracy

[1] uwzględniając merytoryczny wkład promotora (w ramach prowadzonego seminarium dyplomowego)

UNIVERSITY OF SILESIA
FACULTY OF MATHEMATICS, PHYSICS AND CHEMISTRY
AUGUST CHEŁKOWSKI INSTITUTE OF PHYSICS

Robert Kwapich
276941

Microbiome study of Systemic Lupus Erythematosus patients through Next Generation Sequencing Methods

MASTER THESIS

Thesis supervisors:
Anna Michnik Ph.D.
University of Silesia in Katowice

Patrick M. Gaffney, M.D.
Oklahoma Medical Research Foundation

KATOWICE 2017

Contents

Streszczenie

Abstract in Polish

Mikrobiom jest złożonym czynnikiem środowiskowym związanym bezpośrednio ze swym gospodarzem. W ostatnich latach stał się on obiektem wielu badań naukowych mających na celu zbadanie i zrozumienie jego kompozycji i funkcji zarówno u zdrowych jednostek, jak i w przypadku stanów patologicznych. Zmiany wywołane warunkami, czy nawykami żywieniowymi (tj. otyłość, niedożywienie), leczenie antybiotykowe, alergie, stany nowotworowe, choroby neurologiczne, stany zapalne, a także choroby autoimmunologiczne są obecnie intensywnie badane pod kątem związku przyczynowo-skutkowego ze stanem mikroflory bakteryjnej. W przypadku chorób autoimmunologicznych takie schorzenia jak stwardnienie rozsiane, czy reumatoidalne zapalenie stawów zostały bardziej szczegółowo zbadane, w porównaniu do tocznia rumieniowatego układowego.

Niniejsza praca bada zależność między toczniem rumieniowatym układowym (ang. *Systemic Lupus Erythematosus (SLE)*) a stanem mikroflory bakteryjnej. W tym celu zebrano 93 próbki ludzkiego kału (61 pacjentów z toczniem oraz 32 zdrowych osobników), w większości od kobiet pochodzących z różnych grup etnicznych. Próbki zebrano w dwóch punktach czasowych z dwóch rejonów Stanów Zjednoczonych: Stanu Oklahoma oraz Stanu Kalifornii. Wyekstrahowane DNA z zebranych próbek sekwencjonowano przy wykorzystaniu regionów zmiennych mikroorganizmów (gen 16S rRNA) z wykorzystaniem par starterów „Caporaso" dla zmiennego regionu V4. Zastosowana analiza danych poza standardowymi procedurami obejmowała nowatorskie podejście analityczne dotyczące klastrowania sekwencji porzucające standardową metodę opartą na 97% podobieństwie sekwencji, tworząc tzw. operacyjne jednostki taksonomiczne o zerowym promieniu (Zero-radius Operational Taxonomic Units) poprzez wykorzystanie algorytmu `unoise2`. Zastosowano metodę heurystyczną detekcji i eliminacji tzw. przesłuchu (`uncross`) pomiędzy odczytami sekwencji dla różnych próbek, co skutkowało zmniejszeniem błędów typu I. Częścią niniejszej pracy jest także dokumentacja oraz otwarty kod źródłowy zastosowanej analizy danych dostępnej w sieciowym repozytorium.

Wynikiem pracy jest stwierdzenie statystycznie istotnego zmniejszenia liczby obserwowanych bakterii (p=0.001, metryka alfa) w przypadku grupy z toczniem rumieniowatym układowym w porównaniu do zdrowej grupy kontrolnej. Obserwacja ta jest zgodna z badaniami przeprowadzonymi na myszach jako organizmach modelowych tocznia rumieniowatego układowego. W przypadku analizy porównawczej mikrobiomu między grupami próbek, przy zastosowaniu metryk beta (Unifrac) nie zauważono wyraźnej separacji próbek między grupą kontrolną a pacjentami z toczniem. Oznacza to, iż nie zaobserwowano istotnych statystycznie różnic mogących opisywać duże zmiany w drzewie filogenetycznym pacjentów z toczniem.

Zaobserwowano jednakże zmiany w abundancji poszczególnych bakterii, pośród których bakterie z rodzaju Akkermansia (rodziny Verrucomicrobiaceae), Shewanella oraz Atopobium, charakteryzują się większa (istotną statystycznie, p=0.048 po zastosowaniu korekcji dla

wielu porównań metodą Storey'a dla testu 't' Welch'a) abundancją u pacjentów z toczniem rumieniowatym układowym. Dodatkowo na podstawie otrzymanych danych dokonano predykcji pełnego profilu metagenomicznego przy wykorzystaniu oprogramowania PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States). Dzięki tej metodzie możliwym okazało się oszacowanie funkcjonalnego repertuaru mikrobioty i dalsze porównanie grup kontrolnych z pacjentami z SLE. Statystycznie przewidziane ortologi z bazy danych KEGG (Kyoto Encyclopedia of Genes and Genomes) pozwoliły na dalsze ich mapowanie na większe grupy funkcyjne, tj. moduły i ścieżki. Zidentyfikowano zwiększoną abundancję ortologu K11089, genu TROVE2 zwanego także Ro60 w mikrobiomie pacjentów z toczniem rumieniowatym układowym. Zwiększoną ilość przeciwciał przeciwko białkom kodowanym przez ten gen obserwuje się u pacjentów z toczniem rumieniowatym układowym, a także w innych chorobach autoimmunologicznych, np. Syndrom Sjögrena. Otrzymane wyniki sugerują, że mikrobiota pacjentów z toczniem rumieniowatym układowym różni się od tego zaobserwowanego w grupach kontrolnych. Jednocześnie badanie to nie jest całkowicie wyczerpujące, pozostawia miejsce dla bardziej szczegółowych badań.

Abstract

Microbiome is a complex host-environmental factor, that is rapidly becoming elucidated in terms of its composition and functionality in various healthy and pathogenic conditions. Among many, changes in the microbiome caused by nutrition (obesity/malnourishment), antibiotic treatments, allergies, cancers, neurological dysfunctions, inflammation of the bowels, and even autoimmune diseases are widely studied. While some autoimmune diseases such as multiple sclerosis and rheumatoid arthritis have been investigated, the contributory effects of the microbiome on SLE are largely unexplored.

This study has investigated SLE in human subjects. Fecal samples of 93 mostly female human subjects (61 SLE patients, 32 control groups) from different ethnic groups, have been collected at two time periods. Extracted fecal DNA has been sequenced using 16S ribosomal RNA amplicon sequencing method with Caporaso primers for V4 variable gene region. Data analysis pipeline used novel sequence clustering method that created Zero-radius Operational Taxonomic Units through `unoise2` algorithm. Combined with cross-talk detection among multiplexed samples it reduced many false positive OTUs. As a part of this thesis the assembled analysis pipeline has accompanying documentation and open source code available in accompanying online repository.

Study discovered statistically significant overall decrease (p=0.001) in observed number of species (alpha diversity) in SLE patients compared with healthy controls. This observation is consistent with other studies investigating SLE, for example in murine lupus models. Across-samples diversity (beta diversity) observed no distinct separation of grouped clusters (healthy vs diseased), meaning that disease state is not characterized by broad changes of whole communities, rather than by single taxonomic ranks and overall number of observed bacteria. Analyzing microbial abundances, study identified microbial taxa belonging to Verrucomicrobiaceae family (Akkermansia genus), Shewanella and Atopobium genus, that are statistically more abundant in lupus patients (p=0.048), where p-values were corrected for multiple comparisons using Storey FDR parameter using Welch's t-test (unequal variance and sample size t-test).

Moreover, from obtained datasets a predicted metagenomic profile was reconstructed using Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) software package. Predicted orthologs from Kyoto Encyclopedia of Genes and Genomes (KEGG) database allowed to further infer metabolic traits and functions in order to deepen analytical approach. Study identified ortholog K11089 known as `TROVE2` gene or `Ro60` to be more abundant in SLE patients. Autoantibodies against `Ro60` proteins are often a hallmark of SLE as well as Sjögren's syndrome. Obtained results demonstrate that SLE associated microbiome differs from healthy control groups. At the same time study leaves an open space for further, more detailed investigations.

Acknowledgements

I'd like to thank my thesis supervisors: **Patrick Gaffney M.D**. for entrusting me with this fascinating project, giving me a unique experience while working on it from sample collection, through processing, sequencing and data analysis, making it a meaningful and rich experience. **Anna Michnik Ph.D.**, for excellent guidance spanning throughout my studies from early university years. I would like also to thank **Richard Pelikan Ph.D.** for bioinformatics and personal guidance, **Melissa Bebak** and **Mandi Wiley Ph.D.** for wet-lab and organisational instructions, **Kandice Tessner Ph.D.** for invaluable help and advice during preparing this thesis, and all the people from **Gaffney Lab** at **Oklahoma Medical Research Foundation**. I'd like to thank **Polish-American Fulbright Commission** for granting me the possibility to embark on this scientific journey. Last but not least, I'd like to thank my **parents** for their constant support and encouragement, without you I wouldn't be able to do it.

Part I

Theory

Chapter 1

Human microbiome

1.1 Basic definitions, characteristics and essential aspects

In 1676 a Dutch scientist Anthony Van Leeuwenhoek discovered *'animalcules'* ("*little animals*"), both protists and bacteria living in a raindrop through his lensed microscopes. Most researchers didn't have the technical capabilities to match Leeuwenhoek's resolutions, (less then $1\mu m$), and for this reason many of his discoveries were doubted and refused [149]. Anthony Van Leeuwenhoek is officially acknowledged as the father of microbiology. Microbiome is a collective genome of microbiota. Microbiota is the variety of microbial communities found in all multicellular organisms - from plants to animals. Human microbiome would therefore refer to the multiplicity of microorganisms found in a human at a particular site (gastrointestinal, oral, vaginal, skin, nose etc.). Human gut microbiome, the main interest of this thesis, is the multiplicity of microorganisms in human intestine. It is present across three dimensions of lumen and mucus layers. Some papers regard human gut microbiota as a multicellular organism, that consists of nearly two hundred prevalent bacterial species, and approximately one thousand uncommon species [175]. Although many papers (ex. [216](August 2010), [184](May 2014), [21] (December 2014),) still contain information that the number of microbial cells in human gut microbiome tenfolds that of human cells, it is being questioned [2](January 2016). Most recent ratio of microbial cells to human cells is 1.3 to 1.0, however it is a subject to a significant variability and uncertainty [190].

The estimated number of microbial cells is reported to be approximately 40 trillion [189]. Human gut is not only a host for trillions of microorganisms. Human microbiota plays an important role in maintaining human health, and has been proven to contribute to the pathogenesis of various diseases [78], [33], [183], [184]. Current level of understanding of the human gut microbiome cannot specify what constitutes a "healthy" microbiota. Studies now only begin to address whether there might be a particular composition that is shared among healthy or diseased individuals [56] - so called "core microbiome". Technological advancements, namely high-throughput DNA-based pyrosequencing technology helped to classify bacteria and archaea to individual 16S rRNA sequences without the need for culturing (culture independent

techniques). These technologies have became less costly and more rapid, making detailed means of profiling complex communities of microorganisms possible.

Human gut microbiota plays a major role in an organism's health by providing necessary nutrients (vitamins, short chain fatty acids), extracting energy, metabolizing drugs and environmental toxins, and also digesting complex polysaccharides [175]. Many comparative studies state that gut microbiota among higher vertebrates is host-specific. Numerous factors (see section 2.3 on page 30) seem to affect gut microbiome composition. It has been discovered [12] that the richness and abundance of the gut phylum varies along the length of the gut in a proximal to distal gradient of abundance - that is from small intestine, through cecum and colon. During studies of human subjects it has been established that acquisition of the gut microbiota occurs from the first years of life, from the maternal bacteria obtained during vaginal delivery or cesarean section, and later breastfeeding [32], [53]. Microbiota changes are considered to be adaptations to surrounding environment, or associated with antibiotic treatments, dietary changes [30], and other factors discussed later in this text.

Next-Generation Sequencing (NGS) technologies indisputably added a new dimension to the studies of microbiota, as these technologies do not relay on cultivable bacteria, resulting in new observed species, that are a challenge for taxonomist (see "*Current efforts in microbiome research*" section 2.5 on page 33). In addition the falling costs of sequencing (see figure 4.1 on page 43) created increasing of publicly available sequences of human microbiome, allowing big data approaches with novel computational tools and methods. In fact, the volume of sequence data acquired by environmental sequencing is several orders of magnitude greater than that acquired by sequencing of a single genome [175].

1.2 Healthy microbiome

There is a plethora of factors that have been identified to differentiate microbiome composition and functional content - age, ethnicity, BMI, short-term diet, gender, blood pressure, pH and others. However, those phenotypic factors do not well explain the sources of those variations in the healthy human microbiota [78]. Human Microbiome Consortium [78] published in 2012 the results from a wide-range of microbiota studies of healthy human subjects (males and females). Samples from various sites like mouth, gastrointestinal (GI) tract, skin and vagina were collected. Study group consisted of 300 samples from subjects gathered from multiple body sites (18 for women, 15 for men - excluding vaginal samples) at different time points. Researchers found that microbial diversity and abundance varies greatly between each habitat, and that human microbiome could be unique to each individual. Each healthy person is subject to microbiome changes that are often regarded as a "microbiome cloud" in Principal Coordinates Analysis (PCoA) plots (see section 6.2: *Visualizing Microbiome Diversity*, page 80). Despite those variations, healthy subjects exhibited stable (i.e. statistically uniform, not significantly different) metagenomic carriage of metabolic pathways.

Until Next-Generation Sequencing (NGS) technologies were developed, the major challenge in the study of microbiota in various human sites was the inability

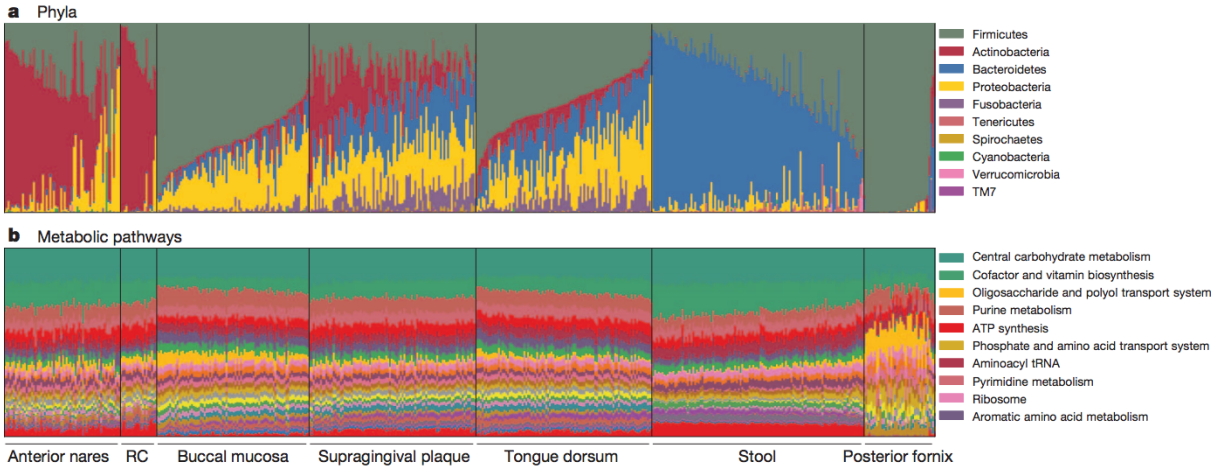


Figure 1.1: Results from wide-range human microbiota studies from various sites - microbial taxa composition varies from subject to subject while metabolic pathways remain stable within a healthy population and body site. Directly reproduced from: [78].

to culture most of the bacteria - as identification methods relied on the culturable bacteria [175]. For the purpose of this thesis the descriptions of microbiome compositions will be focused on the **gut microbiota**. Many studies now show that some phyla proportions, even among healthy individuals, may vary and change over time. Since the composition between individuals has been shown to vary greatly, it is somewhat inadequate to discuss healthy gut microbiota composition. Undisputedly *Bacteroidetes* and *Firmicutes* phylums are the most abundant taxa, ranging around $80\% - 90\%$ of the total gut microbes [164]. However, the trillions of microorganisms are likely composed of several bacterial phyla [36]: *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, *Proteobacteria*, *Fusobacteria* and *Verrucomicrobia*.

Most of the bacterial phyla are considered to be non-pathogenic [45], [88]. Microbiota is responsible for performing essential metabolic functions: it is a source of essential nutrients and vitamins that help in energy and nutrient extraction (such as Short-Chain Fatty Acids (SCFA) and amino-acids), it is crucial for the breakdown of indigestible complex plant polysaccharides, has significant role in induction of `IgA`, and maintenance of homeostasis in various `T-cell` populations in the gut [164] (including Regulatory T cells (Tregs) and Th17, Th1 cells). It plays a protective role against pathogen colonization, working in tandem with host defenses and the immune system. Products of microbiota aren't unique for digestion, production of nutrients, detoxification, defense against pathogens, but also in the development of competent immune system, although this is intensely being studied now [67]. Some efforts focus on establishing a so called "healthy core microbiome" [127], i.e. a microbiome composition that is shared among healthy individuals. Ideally this approach would allow for predictive measures in development and/or progression of various diseases, and perhaps lead to microbiome-targeted therapies, simply due to the fact that gut microbiota is responsible for several critical metabolic and immunological functions that maintain homeostasis [175]. At the same time considering low-level

taxonomic hierarchy (like genus, species or strain level) tge idea of microbiota fingerprint, i.e. unique microbiome composition for each individual emerges, has emerged. This was shown by researchers [49] studying microbiome composition of monozygotic twins, that revealed only approximately 40% of species is shared between them.

Human microbiome consortium discovered that "healthy human microbiota" might in fact contain some pathogenic bacteria. High-risk pathogens are absent in healthy individuals, yet an analogous model to genetic traits, where "recessive alleles of modest risk are maintained in a population" [78] is proposed to microbiota composition. In this model healthy human microbiota could consist of potentially pathogenic bacteria that in homeostasis do not exhibit disease state on its host. Thus, considering aforementioned factors, current microbiome studies try to resolve many unknowns that still accompany microbial studies and its effect on its host.

1.3 Host-microbes interactions

This section briefly characterizes some of the host-microbiome interactions. Those interactions could be assigned into several categories [175]:

1. mutualism (both organisms benefit),
2. amensalism (one organism is inhibited or destroyed and the other is unaffected),
3. commensalism (one organism gets the advantage without any help or harm to the other),
4. competition (both organisms harm each other),
5. parasitism (one organism benefits while harming the other).

Microbial interaction networks, i.e. networks conveying interactions of microbes and host cells, are currently being investigated by scientific community [175]. Major function of intestinal microbiome is nutrient extraction, providing vitamins and short chain fatty acids. Microbiome also processes complex carbohydrates and makes substances such as butyrates (source of nutrients for cells lining the mammalian colon) [33]. There is evidence that gut microbiome shapes the immune system by inducing T helper 17 cell (Th17) cells or TNFa-secreting cells. However, autoimmunity and the gut microbiota is discussed in detail in section 2.6 on page 35. In order to elucidate mechanisms of complex interplay between microbiota, epithelium, immune system and even enteric nervous system, researchers often use germ-free mice - mice raised without any resident microorganisms [123]. This provides for controlled environment in which comparisons between germ-free and normally raised mice could be made; so called Microflora Associated Characteristics (MACs) are drawn.

Already mentioned role of microbiome in nutrient extraction has been shown on germ-free mice. Research shows [61] that conventionally raised mice require around 30% less calorie intake compared with germ-free mice in order to maintain its body weight. Microflora thus might help in conversion of many food products into nutrients for the host, but it is also hypothesized that it affects metabolic capacities of host

cells, which results in increased efficiency of host metabolism [123]. Another example is the carbohydrate metabolism. Without the microflora hydrolyzing carbohydrates, an essential part of energy would remain unextracted for the host. Mammals have evolved certain mechanisms for absorption and utilization of products of bacterial fermentation - like Short-Chain Fatty Acids (SCFA). Gut microflora is also responsible for vitamin synthesis, where members from different taxonomic ranks, such as *Bacteroides*, *Eubacterium*, *Fusobacterium* and *Propionibacterium*, have been shown to synthesize vitamins [63].

1.4 Gastrointestinal dysbiosis

Dysbiosis (also called dysbacteriosis) refers to perturbations of microbiota composition which favours harmul to protective species. Microbiome of an individual tends to stay relatively stable in adult subject. Potential changes in abundances of individual phylotypes may occur due to several reasons: obesity (and resulting diabetes), asthma (and allergies), crohn's disease and ulcerative colitis (inflammatory bowel diseases), irritable bowel syndrome, cancer, autism and HIV [74]. Dysbiosis of the microbiota first was correlated with GI related diseases (like Inflammatory Bowel Disease (IBD) or Irritable Bowel Syndrome (IBS)); however, studies are now exploring microbiome role in functioning of Central Nervous System (CNS) [56]. The "gut-brain axis" was explored in mice models, under which anxiety0induced behavioral changes were significantly smaller in germ-free mice [167]. In fact, pathogenic infections (for example from *Campylobacter jejuni*) have been shown to induce anxiety-like symptoms and subsequent brain-stem activations [157]. On the other hand, human trials with *Lactobacillus casei* studying depression showed improvement in mood regulation after probiotic intake compared with placebo groups [68]. Also the development of Autism Spectrum Disorder (ASD) is associated with the microbiota changes in mice models of Maternal Immune Activation (MIA) [124]. Those mice were administered poly(I:C), a viral mimetic that caused behavioural changes resembling those of ASD.

It has been observed that generally different disease conditions [56] are positively correlated with loss of microbial diversity. Thus gut microbiota in healthy human subject tends to be more "diverse". Such observations were possible only after the introduction of NGS technologies in 2005. The major challenge when attempting to characterize disease-altered microbiota is that there is no established "core healthy microbiota" (see previous section 1.2 on page 19). In many disease conditions, a mutualistic, inter-dependent relation between host immune system, microbial community and its metabolic products are observed. Early environmental exposures (like maternal vaginal bacteria) shape the microbiota, where effects of this exposure span into the adulthood [6]. Studies conducted on mice models of autoimmune diseases (like IBD), housed under germ-free conditions, reveal that the microbiota can modulate the severity and incidence of a particular disease, such that in most cases [56] germ-free conditions prevent development or mitigate the severity of symptoms.

The analysis of microbiome dysbiosis aims at achieving the theoretical grounds for next-generation therapeutics that could target intestinal microbiota [191]. As various disease units show changes induced on microbiome composition, the potential

of artificially manipulating microbiota to mitigate the symptoms or even reverse disease pathogenesis arises. Currently however, the approaches to exactly identify pathobionts that could reproduce the effects of certain diseases have failed.

Chapter 2

Systemic Lupus Erythematosus

2.1 SLE Pathogenesis

In an autoimmune disease, immune cells start attacking the very tissues that they are supposed to protect. Systemic Lupus Erythematosus (SLE) results in severe inflammations that lead to tissue damages in lungs, kidneys, heart, joint and even brain. The onset of this particular disease is characterized when abnormally functioning B lymphocytes produce auto-antibodies to DNA and nuclear proteins, which results in immune complexes that cause damages to aforementioned tissues. SLE is thus autoimmune systemic disease that is characterized by formation of a variety of auto-antibodies (mainly Immunoglobulin G (IgG) and Immunoglobulin M (IgM)), often *glomerular nephritis* (group of kidney diseases) and recruitment of auto-reactive or inflammatory T cells and abnormal production of proinflammatory cytokines, that cause cell infiltration and glomerular necrosis. The word *"systemic"* refers to the fact that this disease affects multiple organs, i.e. brain, liver, skin, kidneys and many others. *"Erythematosus"* refers to the reddening of the skin where *erythema* comes directly from the domain of pathology meaning abnormal redness of the skin due to local inflammation. The term *"lupus"* is a Latin word for a *"wolf"*, however the modern and medical usage of the word refers to a variety of diseases that affect the skin [50].

According to Lupus Foundation of America, around 1.5 million Americans have lupus, and it is estimated that around 5 million people around the world may have a form of lupus [64]. It is more common among people of African, Hispanic and Asian ethnicity. It is observed in 20 to 200 cases for each 100,000 people from general population. Women of childbearing age are said to be around nine to ten times more likely to develop SLE than men. Moreover African-American women suffer from more severe symptoms and a higher mortality rate. Causes of this disease remain unclear, and thus there is currently no available cure. At this time long-term usage of immunosuppressants remains the main treatment, despite their side-effects like susceptibility to infections [27]. Triggering factors still remain unknown and are now being studied (among them age, sex and dietary practices (see section 2.3 on page 30) mainly in murine lupus models. The current hypothesis, however, is that SLE results from interactions between dietary, environmental and genetic factors [27]. As dietary factors are major regulators of immune function, and taking into

1	Malar rash	
2	Discoid rash	
3	Photosensitivity	
4	Oral ulcers	
5	Arthritis	
6	Serositis:	Pleuritis or pericarditis
7	Renal disorder:	Persistent proteinuria of >0.5 g/24 h or cellular casts
8	Neurological disorder (having excluded other causes):	Seizures or psychosis
9	Hematological disorders:	Hemolytic anemia or
		Leukopenia $<4.0 \times 10^9$/liter on 2 or more occasions
		Lymphopenia $<1.5 \times 10^9$/liter on 2 or more occasions
		Thrombocytopenia $<100 \times 10^9$/liter
10	Immunological disorders:	Anti-ds DNA antibody
		Anti-Sm antibody
		Positive antiphospholipid antibodies
11	Antinuclear antibody in raised titer	

Figure 2.1: American College of Rheumatology (ACR) classification criteria for SLE from 1997. SLE may be diagnosed if 4 or more of the 11 criteria are present. Source: [50] (table 10.1).

account the incidence rise in recent years in the developed and developing regions, the western diet may hold at least partial responsibility for the development of this autoimmune disease [70].

One known environmental factor for lupus is UV radiation, combined with susceptibility genes this could initiate certain cell apoptosis. Apoptosis in turn produces apoptotic bodies that include part of the nucleus (DNA, histones, other proteints). Susceptibility genes hypothesis [80] assumes that they may have an effect on the immune system, in which immune cells recognize apoptotic bodies as foreign (nuclear antigens), therefore initiating autoimmune response [66]. Susceptibility genes might be also responsible for less effective clearance, leading to buildup of nuclear antigens. In turn B-cells initiate antinuclear antibodies, products that bind to nuclear antigens and form antigen-antibody complexes, present in majority of cases of SLE. When those complexes penetrate to bloodstream they are able to circulate freely in the body. Eventually those complexes adhere to vessel walls in various organs and tissues leading to local inflammatory responses [135]. Local inflammations activate complement system, a part of the immune system that complements antibodies and phagocytic cells to clear off damaged cells and pathogens from an organism by attacking plasma membrane. In effect this system promotes inflammation [75]. Complement systems, through cascade of various enzymes causes cell membrane to loose its continuity, causing more cell death. However, as it was stated before, UV radiation is not the only factor implicated in influencing the onset of SLE. Other factors might include (but are not restricted to) certain medications (like hydralazine or procainamide), smoking, and sex hormones, like estrogens, which may explain why SLE is more prevalent among women. And finally viruses and bacteria - i.e. microbiome [135], which is the focus of this study. This mechanism is oftentimes referred to in literature as type III hypersensitivity reaction - accumulation of immune complexes (antigen-antibody) and inadequate clearance by innate immune cells - see figure 2.2 for illustration of this process. Type II hypersensitivity mechanism is also observed among SLE patients. Type II hypersensitive mechanism is sometimes referred to as tissue-specific hypersensitivity [75]. Many SLE patients develop auto-antibodies against various phospholipids or red and white blood cells that lead inevitably to their destruction. This mechanism, which is still under scientific investigations, leads to additional SLE symptoms.

Certain studies [125] view SLE as Th1 and Th2 imbalance. It has been noted that cytokine imbalances play a role in acceleration of lupus-like autoimmune disease. The shifting of Th1 to Th2 immune responses results in B cells hyperactivity, and the resulting production of pathogenic auto-antibodies and further inflammation. Recent study [40] suggests an important role of toll-like receptor (TLR)7/9, which contribute to the development of SLE. For non-autoimmune mice, the researchers have modified T cell equilibrium by TLR9 engagement [24]. This suggests that nucleic acids from commensal bacteria might affect immunoregulatory pathways that lead to systemic autoimmunity.

SLE is characterized by periods of "*flare-ups*" - sudden manifestation of symptoms discussed below - and periods of "*remittence*", when symptoms are not manifested, or their intensities are greatly reduced. The diagnosis of SLE is not a straightforward procedure (see figure 2.2), as many symptoms are not SLE-specific, and

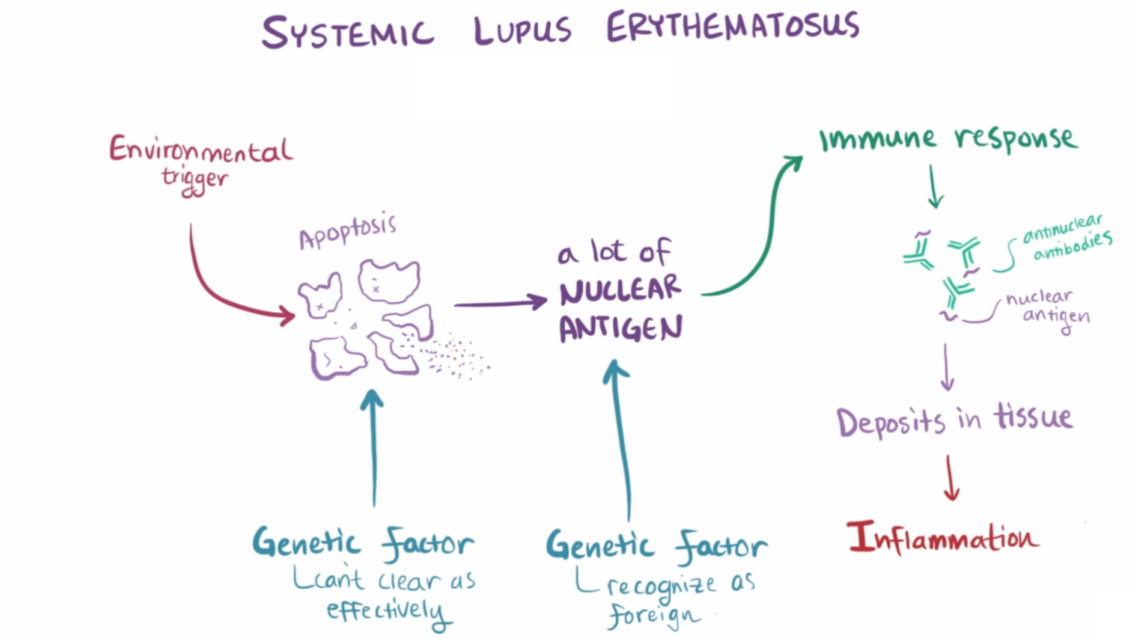


Figure 2.2: Overview of SLE onset caused by environmental factor. Adapted from source: [135].

could be equally attributed to other diseases. Generally they include fever and weight loss, while specific symptoms are directly attributed to the organs being affected. For this reason American College of Rheumatology (ACR) classification for SLE require the presence of minimum four symptoms out of eleven (see figure 2.1 on page 25) in order to diagnose this autoimmune disease. There are symptoms that affect the skin (malar rash, discoid rash) associated with general photo-sensitivity that causes rash, like characteristic "butterfly rash" over cheeks of affected individual. The second group of symptoms include inner membrane of various tissues - *mucosa*. Manifestation in this subcategory include ulcers in oral area. *Serosa* - the outer membrane of various tissues could also be affected, which is manifested by *serositis*: *pleuritis* - an inflammation of the lining around the lungs or *pericarditis* - inflammation of the membrane (pericardium) surrounding the heart. Criteria for SLE diagnosis also include the detection of several antibodies in blood. Already discussed antinuclear antibodies are not a good hallmark of SLE, as this test is not very specific. Antinuclear antibodies are present in several other diseases. More SLE specific antibodies include:

- Anti-Sm antibody (anti-smith antibodies are produced against small ribonucleoproteints),
- Anti-Ds DNA antibody (targeting double-stranded DNA)

Other antibodies and symptoms associated with SLE are listed on figure 2.1. For the plethora of symptoms that might be manifesting itself in SLE there are equal variety of drugs (see figure 2.3 on page 28). SLE patients are advised to avoid sunlight in order to prevent lupus flare-ups. Corticosteroids are often used for limiting immune

Symptom	Drug	Regimen
Arthralgias/fever	NSAIDs (caution with renal disease) Hydroxychloroquine	No special recommendation Hydroxychloroquine 400 mg daily; ophthalmic examination yearly, although the risk of untoward events is low.
Malar/discoid rash	Prednisone, hydroxychloroquine, sunscreen	
Arthritis/serositis/myositis	Prednisone, methotrexate, azathioprine, leflunomide	20–40 mg daily for 2–4 weeks, then reducing dose 5 mg steps each week. Require bone prophylaxis against osteoporosis if dose remains at 7.5 mg or above for more than 3 months.
Autoimmune anemia or thrombocytopenia (ITP)	Prednisone, azathioprine, IVIG	60–80 mg prednisolone daily for 2 weeks, reducing in 10 mg steps per week after depending on response. 2.5 mg/kg azathioprine.ITP might also require immunoglobulin or splenectomy
Renal	Prednisone, mycophenolate, azathioprine, cyclophosphamide	Severe disease may require monthly IV steroid and cyclophosphamide for 6 months then 2–3 monthly for 2 years
Central nervous system	Prednisone, azathioprine	Up to 80 mg daily of prednisone (1 mg/kg/day)

Figure 2.3: Recommendations for drug use in Systemic Lupus Erythematosus (SLE). Source: [50], *table: 10.5*.

responses, immunosuppresants are used when SLE manifestation are severe and potentially life-threatening.

2.2 Gut Microbiota in SLE

2.2.1 Lupus-associated microbiome

Studies involving lupus often involve murine models of lupus-like diseases. Classical example is lupus-prone MRL/Mp-Fas^{lpr} (MRL/lpr) mice (homozygous for the mutation Fas^{lpr}). This particular mouse manifests systemic autoimmunity, lymphadenopathy, and glomerulonephritis, all of which are similar to human lupus-associated symptoms. Concerning commensal bacteria, a classical study [10] briefly characterized a number of changes with regard to lupus-prone mice: depletion of *Lactobacillaceae*, increases of *Lachnospiraceae* and *Clostridiaceae* phyla. Dietary interventions using Retinoic Acid (RA) restored *Lactobacilli* phylum, at the same time improving or even reversing SLE symptoms, and partially restoring changes in microbial functions. Hence, it has been suggested to use probiotic *Lactobacilli* and Retinoic Acid as dietary supplements for lupus patients in order to relieve inflammatory flares. Specific factors are discussed in section 2.3 on page 30, but it is worth mentioning here that an overrepresentation of *Lachnospiraceae* in females was associated with an earlier onset of and more severe lupus symptoms [10].

PCoA plots from study [27] showed a clear distinction between MRL vs MRL/lpr - lupus-prone (see figure 2.4 on page 29). In this study fecal samples were gathered from 5-week and 14-week old females. Through diversity analysis using `UniFrac` distance metric, researchers observed significant phylogenetic differences. As in the previously mentioned study [10], this study [27] also observed a reduction of family *Lactobacillaceae*, and increases in *Lachnospiraceae*, *Ruminococcaceae*, *Rikenellaceae (genus Alistipes)*, *Clostridiales* family XIII, and the *Streptococcacea* family (see figure 2.5). It is worth noting that lupus prone mice were characterised as having higher microbial diversity compared to healthy ones. Neither RA or VARA (vitamin A-retinoic acid) treatments changed this phenomenon, which implies that this diversity may depend on genetic predispositions of the mouse strains, but not their phenotype

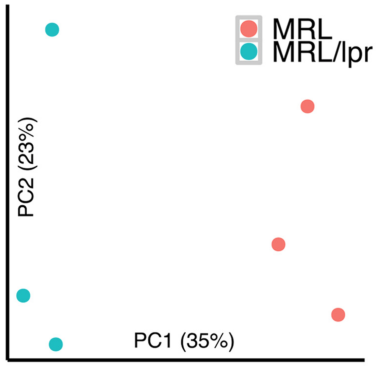


Figure 2.4: PCoA (first two principal coordinates, unweighted UniFrac distances) of MRL and MRL/lpr mice gut microbiota. Fecal samples from 5-week old female mice. Source [27], fig. 1A.

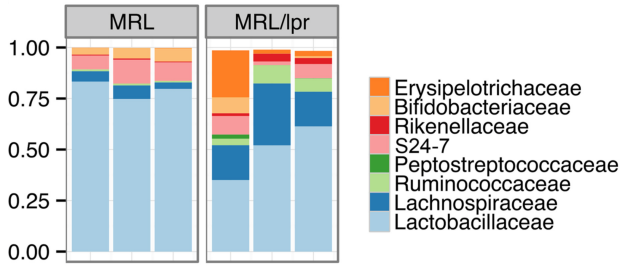


Figure 2.5: Taxonomy analysis at the family level for abundant bacterial OTU (> 0.1%) (y-axis relative abundance). Rare OTUs accounted for 3.5% of individual microbiota. Source: [27], fig. 1C.

[27].

2.2.2 Improving lupus symptoms

It was already mentioned (see section 2.2.1), that Retinoic Acid (RA) restores certain bacterial phyla like *Lactobacilli*. Discussed study [27] investigated the effects of vitamin A and RA on MRL/lpr female mice. Study involved two therapeutic groups, one with vitamin A with RA $6mg/kg$ (in canola oil), second with retinyl palmitate $11.2mg/kg$ and $0.6mg/kg$ RA. Daily oral distribution continued for a period of six to fourteen weeks. Control mice were treated only with canola oil (*"a vehicle"*), as it contains polyunsaturated fatty acids. Researchers observed reduced turnover of bacterial DNA, RNA and protein synthesis comparing lupus prone mice to control groups [27]. On the other hand, increases in cell motility and sporulation genes occurred, altogether with increased membrane transport, amino-acid metabolism and signal transduction. Vitamin A treatment with RA attenuated lupus-like changes and symptoms, reversing changes in carbohydrate and amino-acid metabolism (e.g., histidine metabolism, and phenylalanine, tyrosine and tryptophan biosynthesis) [27]. This study [27] confirmed the previously cited study [10], that RA restored bacteria

from *Lactobacillaceae* phylum. Abundance of mentioned phyla in untreated mice was reduced, while the treatment restored *Lactobacillaceae* pyhulm, improving at the same time lupus symptoms. It is worth noting that another therapeutic group treated with retinyl palmitate + 10% of RA not only didn't show positive effects, but even worsened the lupus symptoms. Still, it remains unclear whether retinoid-mediated changes cause subsequent alterations in gut microbiota, or rather those alterations are the by-products of vitamin A.

2.3 Factors shaping SLE microbiota

2.3.1 Dietary factors

Human and mice microbiome exhibit similarity, that allows for performing studies on mice and inferring results for human subjects. As it was stated in the introduction, mostly two bacterial phyla: *Bacteroidetes* and *Firmicutes*, and one archea appear to dominate the microbiome. Altogether they make up 98% of the 16S rRNA sequences obtained from GI tract [11]. Simple dietary modifications like caloric restriction have shown to be effective in preventing progression of lupus-like disease in NZB mice [13], as well as the SLE-associated Antiphospholipid Syndrome (APS) [195]. NZB (New Zealand Black) mice displays several characteristic autoimmune-like abnormalities: hemolytic anemia, elevated levels of immunoglobulin, anti-DNA antibodies, anti-thymocyte antibodies, and circulating immune complexes causing glomerulonephritis [138]. This mice are widely used as a model for SLE. Calorie restriction improved SLE survival by eliminating immunoglobulin (Ig)A and IgG2 auto-antibodies as well as increasing secretion of interleukin (IL)-12 and interferon-γ (INFγ)[125]. Also vitamins A, D and E with polyunsaturated fatty acids (n-6/n-3 fatty acids) and phytoestrogens, when applied orally to mice models of lupus, have shown to lead to improved outcomes - reduction of proteinuria and glomerulonephritis [125]. For example, vitamin D inhibits B cell activation and differentiation into plasmablast and subsequent immunoglobulin production [125]. Enriching food that was served to mice with *n-3 polyunsaturated fatty acids* proved to prevent many clinical symptoms of SLE, like fetal loss or APS which are considered lupus-associated manifestations [178].

Another study [9] considered simple dietary deviation, namely changing *pH* of drinking water, to investigate the effects on SLE incidence, and composition of gut microbiota. Study was conducted on mice that were administered acidic pH water (AW) or neutral *pH* water (NW). Mice that were given AW developed nephritis at a slower pace compared to the other group. 16S rRNA gene-targeted sequencing revealed that the composition of gut microbiome was significantly different between NW and AW groups of mice. At the same time this study revealed that Segmented Filamentous Bacteria (SFB) colonization was unaffected by *pH* of drinking water, and it didn't cause a profound increase in Th17 response. Since SFB promotes Th17 response and autoimmunity in several mouse models (of arthritis, multiple sclerosis etc.), researchers stated that *pH* of water had no significant effect on lupus incidence. There are numerous dietary factors that could, or are reported to modify SLE symptoms and pathogenesis [9]. Some of them are presented in table 2.1 on

page 32. In general, dietary factors that act on one of the ways listed below, are beneficial for delayed onset of proteinuria, and for longer life span of murine lupus [125]:

- suppress Th2 cytokines, such as IL-4 at early phase,
- lower PGE2 production,
  Prostaglandin E2 (PGE2), also known as dinoprostone.
- inhibit inflammatory cytokines: IL-1β, IL- 6, TNFα and IFNγ,
- enhance production of TGF-β or IL-10 at late phase

Such dietary functions have been shown to affect autoantibody production, cytokine secretions, inflammatory mediators and subsequently alter life span of lupus mice [125]. Still, many dietary interactions between SLE and gut microbiota remain unknown. Moreover not enough clinical studies have involved dietary interventions for SLE patients to determine the effects of dietary factors on human SLE.

2.3.2 Sex

Sexual dimorphism is shown to be responsible for hormone dependent autoimmunity in non-obese diabetic (NOD) mice [41], which is a model of female Sjörgen syndrome and autoimmune polyglandular syndrome type II manifesting with thyroiditis and adrenalitis [195]. Moreover, female mice exhibited accelerated disease progression compared with male counterparts. Males on the other hand, do not exhibit significant differences in the gut, when lupus prone and healthy individuals were compared [27].

Comparing sexes among healthy mice in that study has shown that certain phyla differ in terms of abundance. Namely females had higher abundance of *Lactobacillaceae* and *Streptococcaceae*, a lower abundance of *Lachnospiraceae* and *Clostridiaceae* compared with male age-matched counterparts. Comparing differences among lupus prone and healthy mice, it has been noted that males and females had comparable levels of relative abundance of *Lactobacillaceae* phylum. Where they differed was in *Lachnospiraceae* and *Bacteroidetes* phylas, for which females had much higher abundance compared to males. Males on the other hand had much higher abundance of *Bifidobacterium* and *Erysipelotrichaceae* [27]. *Bifidobacterium* like *lactobacilli* has been suggested to exert anti-inflammatory functions [35], which suggests the hypothesis that the greater the abundance of the *bifidobacterium*, the greater the attenuation of lupus symptoms in males. It is at the same time worth noting that Principal Coordinates Analysis (PCoA) showed no distinct microbiota in control (healthy) male vs. female groups, only in lupus prone mice. Study [27] concludes with observation that higher abundance of *Lachnospiraceae* may be associated with an earlier onset or more severe lupus symptoms in female MRL/lpr (lupus prone) mice.

2.3.3 Virome

Since 16S rRNA gene sequencing survey is unable to access the virome, its role and potential will not be discussed extensively; only more accurate (and more costly)

Dietary manipulation	Cytokine and autoantibody regulation	Effect on animal model
1. Calorie restriction	↑ CD8+ T lymphocytes ↓ IL-2 and IFN-γ in splenocyte ↓ IFN-γ, IL-10 and IL-12 secretions, and mRNA in kidney and submandibular gland ↓ IgA and IgG2 levels and polymeric immunoglobulin receptor mRNA in submandibular gland ↓ NF-κB activation in kidney	↓ Glomerulonephritis and deposits, proteinuria ↑ Life span in NZB/W mice
2. Dietary oil		
n−6 PUFA	↑ IL-6, TNFα and PGE_2 in macrophage ↓ TGFβ mRNA in splenocyte, lymphocyte proliferation ↓ Anti-dsDNA IgG autoantibody in serum	Counteracted effect in NZB/W mice
n−3 PUFA	↓Th2/Th1 cytokine and ↑TGFβ in splenocyte ↓ IL-6, TNF-α, PGE_2 and IL-1β productions, and NF-κB activation in macrophage ↓ Anti-ssDNA and dsDNA IgG autoantibody in serum	↑ Antioxidant enzymes, ↓ adhesion molecules ↓ Glomerulonephritis and proteinuria ↑ Life span in NZB/W mice
3. Vitamin A	↓ IFN-γ, IL-2 and IL-10 secretions in serum ↓ IFN-γ, IL-2 and IL-12 secretions in splenic CD4+ T cells ↓ Monocyte chemoattractant protein1 mRNA in kidney ↓ Anti-DNA autoantibody in serum ↓ IL-1α, IL-1β, IFN-γ-inducible factor, IL-12 and ↑ TGF-β mRNA in kidney	↓ Proteinuria, glomerular IgG deposits ↑ Life span in NZB/W mice ↓ T cell and macrophage infiltrates and proteinuria in the kidney of MRL/l mice
4. Vitamin D	↑ Th2 (IL4, IL5, IL-10, IL-13)/Th1 (IFN-γ, IL-2, IL-12) cytokine in splenocyte ↑ Treg activation and natural killer T cells ↑ IL-4 and TGFβ in splenocyte ↓ anti-ssDNA autoantibody ↑ Macrophage chemoattractant osteopontin	↓ Proteinuria ↑ Life span in MRL/l mice ↑ Lupus nephritis in NZB/W and MRL/l mice
5. Vitamin E		
Oxidation	↓ IL-6 and IFN-γ in splenocyte ↓ Anti-dsDNA IgG autoantibody in serum	↓ Oxidative stress in spleen and kidney ↓ Proteinuria, ↑ life span in NZB/W F1 mice
High dose	↑ Anti-dsDNA and cardiolipin IgM autoantibody in serum ↓ IL-2, ↑ IL-4 and IL-10 in splenocyte	↓ Life span in high vitamin E in MRL/l mice Opposite effects of low and high vitamin E
6. Phytoestrogens		
Isoflavones	↓ IFN-γ in splenocytes ↓ anti-dsDNA and cardiolipin IgG autoantibody in serum	↓ renal damage, proteinuria, ↑ life span in MRL/l mice
Alfalfa sprout	↓ IFN-γ and IL-4 in splenocytes ↓ TNFα and IL-1β in peritoneal cells	↓ glomerulonephritis, ↑ life span in MRL/l mice
Coumestrol	↓ anti-dsDNA IgG and anti-chromatin autoantibody in serum	↓ proteinuria in NZB/W mice
Indole-3-carbinol	↓ Anti-dsDNA autoantibody in serum	↓ Glomerulonephritis, IgG deposits in kidney ↑ life span in NZB/W mice
7. Traditional medicine		
Ganoderma	↓ Anti-dsDNA autoantibody in serum	↓ Perivascular and parenchyma mononuclear cell infiltration in vital organs ↓ Proteinuria, ↑ life span in NZB/W mice
Hachimi-jio-gan	↓ IFNγ production and IL-12 mRNA expression, tended to promote IL-4 production in splenocyte ↓ Anti-dsDNA IgG2a autoantibody in serum	↓ Immune complex deposition in kidney and proteinuria, and nephritis in MRL/l mice

Table 2.1: Dietary modifications in lupus animal models. Directly reproduced from [125].

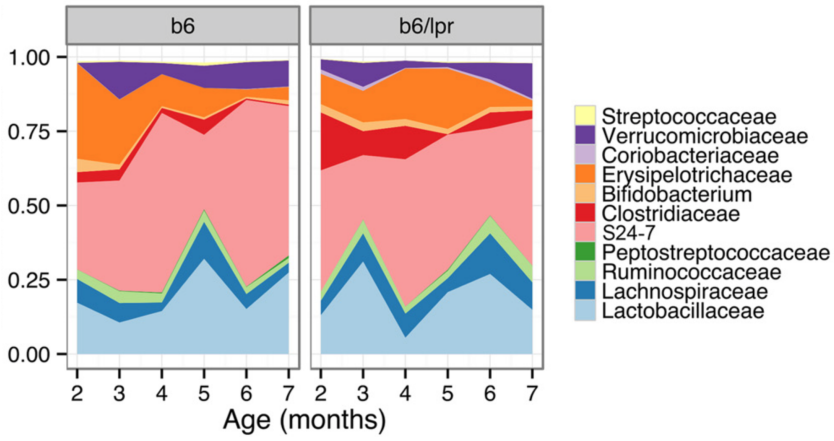


Figure 2.6: Time-dependent (with relation to age) microbiota changes (y-axis relative abundance) in healthy and lupus-prone mice. Source [27], fig. 2C.

study employing Whole Genome Sequencing (WGS) would explore the gut virome. The Virome - viral microbiome - is much less studied than the microbiome. It is a diverse community consisting of eukaryotic RNA and DNA viruses and bacteriophages. Preliminary studies state that is might also play a role in human health [19]. It has been observed that gut microbiota interacts with host viruses, therefore influencing the virome. The RNA and DNA viruses could reside within gut commensal bacteria. The interferon-α (INFα) signature in SLE patients suggest a potential viral factor to consider for future therapies and studies [18]. Study conducted on monozygotic twins and their mothers [5] showed that, although bacterial abundance in the gut tend to change over time, virotypes remain constant.

2.4 Time-dependent microbiota changes

A time-dependent microbiota study [27] involving B6.MRL-Fas^{lpr} mice, due to the fact that their lifespan is 7 months, compared to previously mentioned mice MRL/Mp-Fas^{lpr} (MRL/lpr), which have a lifespan is approximately 4.5 months. Even before manifesting lupus-like symptoms, lupus prone mice (B6.MRL-Fas^{lpr}) showed higher levels of *Clostridiaceae* abundance, which occurred from two to five months. On the other hand, bacteria from *Lachnospiraceae* phylum were more abundant after lupus-like symptoms occurred in *lpr* mice - see figure 2.6. Study concluded that microbiome changes occur throughout the entire lifetime in both (healthy and lupus prone) mice types. Bacterial phyla, such as *Clostridiaceae* and *Lachnospiraceae* are more abundant in *lpr* mice at specific time points.

2.5 Current microbiome efforts

Microbiome studies are vast, and concerned with many different topics like allergies, cancer progression, autoimmune diseases, and establishing a microbial "core" for healthy and diseased patients. This "functional core" presumably exists, meaning

that almost 40% of the microbiome is shared among each human individual [51], although these are just preliminary results. Researchers are just begining to tackle the difficulties in establishing relations between human health and relative abundance of specific microbial communities. However, due to intra- and inter-individual differences that are influenced by (among others) sex, dietary and age factors, this connection is complex. It is even questionable whether occurring changes in phylogenetic composition are causative or consecutive for a studied disease [56]. Studies often identify bacteria with potential to cause disease, and yet it remains also possible that these microorganisms colonize the gut after the onset of a certain disease [164]. One of many tasks in microbiome studies is to identify the commensal bacteria in the GI tract, that promote or prevent certain diseases - i.e. search for *pathobionts* and *symbionts* [195]. However, determining the modulating factors that change microbiota is the primary concern. There are numerous gaps in our knowledge in understanding the interplay of diet and gut microbiota in human immune-mediated diseases. The molecular and cellular mechanisms that drive gut commensals to influence autoimmune responses are just now being studied and understood.

For studying microbiota in pathogenesis of autoimmune diseases, research often involves Germ-free (GF) mice models of corresponding human-autoimmune disease. Studies show that GF mice poorly develop lymphoid tissues, have spleens with fewer germinal centers and poorly formed T and B cell zones, lower numbers of *lamina propria* $CD4^+$ cells and IgA-producing plasma cells [65]. Therefore protective immune responses could be induced by "healthy" and balanced microbiota, could suppress inflammation in organs distant from a gut by induction of tolerogenic DC and Regulatory T cells (Tregs). However, how these cells traffic from the gut to the distal lymph nodes and possibly to the peripheral tissues like Central Nervous System (CNS), joints and pancreas, is still unknown and needs to be determined [164]. Microbiota is currently being viewed as a trigger in GF mice models. Still, attempting to identify those pathological members that reproduce the effects on the host proved fruitless [56]. Germ-free animals are often used in pair with wild-type ones to compare the effects the microbiota has on its hosts in health and disease. One major problem is that microbiome of animals under investigations often change dramatically from one to another facility, which in turn hinders identification of potential genetic biomarkers of disease [175]. Often correct and proper statistical analysis is difficult to achieve, and many researchers use different algorithmic and statistical approach to infer their conclusions. For this reason establishing a so called Standard Operating Procedure (SOP) is proposed as one of the most important aspects of microbiome study. SOP would enforce greater reproducibility in microbiome studies, and eliminate often contradictory results [39]. The long-term goal of microbiome studies is however to establish microbial networks and understand their functionality. This understanding would further drive potential to manipulate individual microbiome, or even individual bacteria for targeted ("personal") therapy through (for example) designer probiotics. Also the ability to predict perturbation effects of the gut microflora with relation to certain conditions (like obesity) or diseases, would allow for immediate counteracting procedures against potential pathogenic changes [175].

2.6 Gut microbiota and the immune system

Autoimmune diseases are said to develop in genetically susceptible individuals [122], yet environmental factors are being studied, due to the fact that the environment seems to modify immunological responses. Since studies on homozygotic twins didn't record 100% concordance rate for autoimmune and inflammatory diseases (the onset, symptoms, and severity) further suggesting additional environmental factors play a significant role. In recent years the impact of microbiota on the immune system responses has begun to be studied. For genetically susceptible individuals, there are at least two general factors that would induce the onset of a disease (see figure 2.7 on page 36):

1. reduction of commensal bacteria favoring regulatory cells,
2. enrichment in commensal bacteria that favors the induction of potential pathogenic cells.

Finding correlations between microbial organisms in the human intestine, the immunological responses and the onset of a disease is a nontrivial endeavour. Preliminary studies indicate that the same commensal bacteria could possibly induce a protective response or a pathogenic one, which seem to depend on the susceptibility of the individual. Spore-forming Segmented Filamentous Bacteria (SFB) are a perfect example - protective in type 1 diabetes [42], causing at the same time Experimental Autoimmune Encephalitis (EAE) in mouse models [29] or Autoimmune Arthritis [26]. In fact SFB were shown to promote the development of intestinal Th17 cells, which are crucial for fighting bacterial infections. At the same time SFB are involved in the pathogenesis of a number of inflammatory and autoimmune diseases [59]. Microbiota that favors SFB could thus have an impact on the immune response, and consequently on the development of Th17-mediated inflammatory or autoimmune diseases in the gut at distant sites in predisposed individuals. Microbiota is said to have a profound impact on the maturation and development of immune system, and therefore modulations in commensal bacteria composition could possibly influence onset and progression of certain autoimmune or inflammatory diseases [164].

2.6.1 Inflammation

Generally, commensal bacteria control the inflammation in the gut via two mechanisms:

1. controlling bacterial phyla that directly cause intestinal inflammation,
2. indirectly affecting the composition of the microbiota - reducing disease-mediating commensals - to balance immune responses in favor of the regulation.

Bacteroidetes and *Firmicutes* phyla are said to induce Tregs that control Th17 cells, which in turn are responsible for intestinal inflammation. During a certain study [14], human bacteria *Lactobacillus* from the *Firmicutes* phylum were orally distributed to mice models of colitis (inflammatory disease of the colon), and have

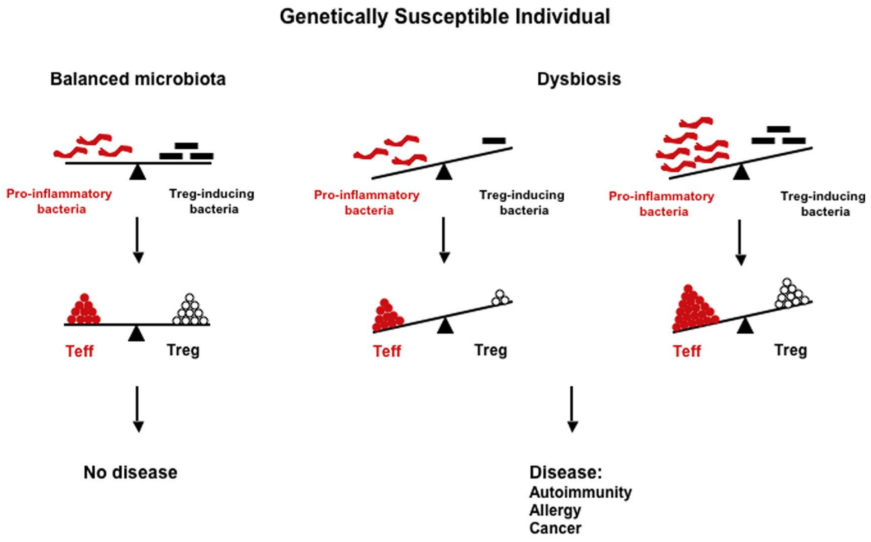


Figure 2.7: Microbiota and its impact on immune response and disease development. Directly reproduced from [164], figure 1.

been shown to be protective against inflammation. *Lactobacillus* bacteria attenuated the symptoms and effect via induction of Regulatory T cells (Tregs). Certain commensal bacteria exist (like those from *Actinobacteria* phylum) that appear to control the levels of other commensal bacteria, which in turn cause intestinal inflammation [164].

2.6.2 Autoimmunity and commensal bacteria

Compared with inflammation, our current understanding of the role of commensal bacteria in autoimmune disease (induction or exacerbation) is more limited. Spore-forming Segmented Filamentous Bacteria (SFB) has been shown to induce Interleukin 17 (IL-17) production, and subsequent induction of Rheumatoid Arthritis (RA) in arthritis-prone K/BxN mice via activation of Th17 cells [26]. It is worth noting that Experimental Autoimmune Encephalitis (EAE), Rheumatoid Arthritis (RA) and colitis are mediated by IL-17 producing cells. In fact, a single commensal bacteria is capable of producing an immune response that are extra-testinal, i.e. outside of intestine in distant sites, extending to peripheral lymphoid organs and tissues. Evidence shows that SFB-induced Th17 cells could be found in spinal cord, spleen and joints [26]. Human species could be colonized with SFB [23] for protective role (like in type 1 diabetes [42]), however, there is not enough evidence suggesting that SFB are directly responsible for autoimmune or inflammatory diseases in humans. One of the current focuses in the microbiome study is in determining the causative or consecutive role of the microbiota in a particular disease - whether certain bacteria are present, or gut is colonized after the disease onset. However, there is substantial evidence that IBD causes abnormal immune responses to bacterial components of the gut microbiota [73].

2.6.3 Protective role of commensal bacteria

As it was stated in the previous section, commensal bacteria serve a protective role at sites distant from the intestine. Early study from 1993 [12] stated that microbiota plays a significant role in a model of collagen-induced RA, where rats maintained in GF environment developed more severe RA symptoms, than those kept in a conventional environment. More recent studies from 2009 [29] mention microbiota's role in protecting CNS. In this study mice were treated orally with an antibiotic with Polysacharide A (PSA) that was isolated from *B. fragilis* (*Bacteroidetes* phylum). It was discovered that *B. fragilis* prevents EAE development, and attenuated its symptoms. At the same time oral administration of probiotics proved to protect against autoimmune diseases like type 1 diabetes [43], EAE [55], and even SLE [48]. Generally, it is said that abundance of microorganisms from *Bifidobacterium*, *Bacteroides*, *Clostridium*, and *Lactobacillus* phyla are positively correlated with protection against inflammation and autoimmune diseases [164]. It is for this reason, that there are numerous efforts in establishing a balanced microbiota that favors these protective commensal bacteria. Determining the factors that shape a "balanced" microbiota would therefore stand as a strategy for the prevention and perhaps treatment of inflammatory and autoimmune diseases in susceptible individuals.

2.6.4 Summary

Genetics and environment shape the gut microbiota composition, which shapes the immune response at both levels: intestinal and extra-intestinal. Gut microbiota also shapes the development of some types of autoimmune and allergic diseases, including SLE. Many gut microbiota changes are sex-specific, depend on the environment and dietary practices, and finally vary over time during lupus progression. Current treatments propose certain probiotics with anti-inflammatory functions [57], for the discussed reduction in members of *Lactobacillaceae* family. Currently no direct correlation between intestinal *Lactobacillaceae* and autoimmune lupus has been established, although colonization of these bacteria has been shown to reduce inflammation both in mice and humans [37]. Strong positive correlations exist between the abundance of *Lachnospiraceae* and lupus disease parameters, including lymphadenopathy and renal pathology. In the future researchers would like to design patient-specific therapies, that would be able to prevent development of disease in high-risk (susceptible) individuals, while at the same time treat already diseased patients. There are hopes to establish microbiota species' profiles in healthy and diseased individuals that would predict possible outcomes of drug treatments, suggest dietary practices and even help in diagnostics. It would be another step into the era of personalized therapies.

Chapter 3

Genomic techniques

3.1 16S rRNA microbial survey

16S rRNA Gene Amplicon Sequencing, sometimes referred as 16S survey, is a common and relatively affordable method to study bacterial phylogeny and taxonomy. It uses a 16S rRNA marker gene usually restricted to certain regions, among which V4 is the most widely used. The 'S' in 16S represents Svedberg units, which is a non-metric unit describing sedimentation rate, that is affected by its shape and mass. Svedberg is a time unit, where $1Sv = 1S = 10^{-13}[s]$. 16S ribosomal RNA (16S rRNA) is a part of the 30S small subunit of a prokaryotic ribosome. 16s rRNA gene is the 1542 base pair (bp) long gene that codes for ribosomal subunit. This gene is present in almost the entire known spectrum of bacteria [139]. In addition, this region is highly conserved, meaning that function of this gene over evolutionary time has not changed much [139]. Despite evolutionary conservation (i.e. ubiquity) the gene also has highly variable regions that can be used to differentiate between various bacteria (see figure 3.1).

16S rRNA gene is too long to be directly sequenced to completion from a single sequencer read, therefore it has been divided into several sequencing portions that are several hundred base pairs long. Choosing highly conserved regions would not be useful for identification of bacterial taxonomy. On the other hand, choosing highly variable regions, such a with many changes occurring from species to species would unnecessarily complicate possible analyses, for example: identification of bacteria belonging to the same family would be hindered, as no common sub-sequences are present.

It has been shown [215] that no variable region is able to exactly classify all bacteria from Domain to Species levels, however, some (like V4 region) can reliably predict specific taxonomic levels. V4 region of 16s rRNA gene is able to provide resolution comparable to entire 16S gene. Underestimation of the total diversity of the sample occurs for certain families, like *Enterobacteriaceae* or *Clostridiaceae*, from which it is reported [34] that bacterial species could share up to 99% sequence similarity. The standard threshold of 97% similarity produce OTUs that underestimate total diversity. One of the main disadvantages to 16S rRNA analysis is that produced results have low phylogenetic power at the genus and species level [139], in fact some researchers even argue that species should not be predicted [89]. Two

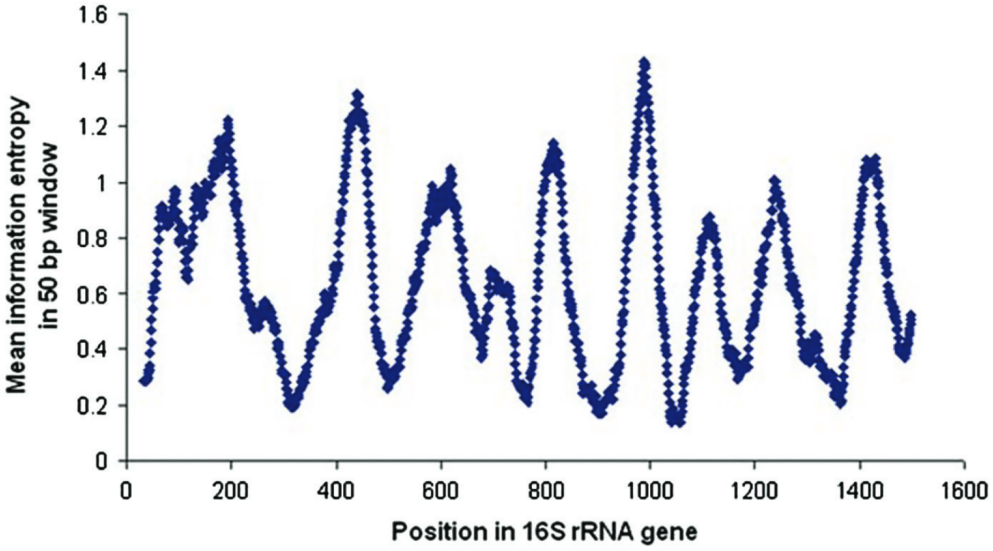


Figure 3.1: Variability within 16S rRNA gene from pre-aligned, sequenced data. Variability measured as Shannon information entropy, calculated at each sequence position, using only positions without a gap in *E. coli*. Y-axis showing the Shannon Entropy over 50-bp windows, centered at each position in the gene (X-axis). Shannon entropy calculated as $\sum p(x_i)log2(p(x_i))$, where $p(x_i)$ is the frequency of nucleotide i. Figure directly reproduced from [145].

or more species could have identical tag sequences, making it impossible to correctly identify species [199].

3.2 Metagenomics

The term "metagenome" was first used in a publication from 1998 [31], which is regarded as the first use of this term to describe a "collective genome of soil microflora" [212]. For the first time the collective genome was regarded as a "genomic unit". The prefix *meta* comes from the greek, and means *after*, *beyond* or even *transcendent*. "*Meta*" in the context of metagenomics stands for the fact that there is a need to develop methods that go *beyond* exact and complete characterization of the sample (due to its inherent complexity), and instead try to maximize understanding by developing advanced sampling methods. There is however another point to metagenomics, where *meta* stands for the analysis not on individual, but rather whole communities, and how they might influence each other.

Metagenomics is regarded as a set of research techniques, but also as a valid research field [163] that studies genetic material of environmental samples, instead of cultured-based ones; for this reason it is sometimes referred as *eco-genomics* or *community genomics*. Broad definition of metagenomics states that it involves cultivation-independent genome-level characterization of microbial communities. It aims to understand transorganismal behaviors and biosphere at the genomic level [163]. Metagenomic (shotgun sequencing) approach is different from 16S rRNA gene

survey approach. It involves the study of many microorganisms by randomly shearing its DNA and sequencing many short sequences. This approach does not amplify (i.e. clone) any specific regions of microbial DNA (like 16S rRNA gene), and very often does not perform Polymerase Chain Reaction (PCR) amplification of sheered microbial DNA [39], thus allowing it to bypass PCR errors and study, at the same time, bacteria, microbial eukaryotes and viruses. Short sequences could be bioinformatically reconstructed into a consensus sequence, or marker-based approaches involving many specific tags, like 16S rRNA genes and others could be used [58]. Due to the nature of shotgun sequencing, study of environmental samples ensures that low abundant organisms will be represented by at least small sequence segments - if enough coverage (sequencing depth) is provided [22]. It still remains impossible to fully capture all DNA sequenced from every microorganism in the studied environment. For this reason sequencing depth estimates are necessary to compromise between required depth, potential sequencing costs and achieving the study objectives [140]. Metagenomics allows for functional profiling (i.e. what metabolic processes are possible) from gene composition of microbial communities, that is not available from 16S rRNA gene surveys allowing for phylogenetic and diversity analyses [188]. For this reason, and for its superior accuracy compared with 16S analyses it is used for generating novel hypotheses of microbial function. [58].

3.3 Comparing 16S rRNA and Whole Genome Sequencing

Previous chapters describe the approach used for 16S surveys (16S rRNA gene sequencing) and metagenomic studies. This section briefly summarizes the main differences between those two approaches.

It is worth noting that 16S surveys sometimes are wrongly to referred as “metagenomic studies”. This is a common mistake, due to the fact that this survey targets only a single microbial gene, and most often a single region from that gene. While metagenomic study should be identified with Whole Genome Sequencing (WGS) (shotgun sequencing) where no specific region is targeted, and produced reads could come from any microbial gene regions. 16S survey is a relatively affordable approach for identification of known bacteria based on reference databases with known genomes. It does not provide functional characterization of the analysed community, except for possible predictions using `PICRUSt` approach (see section 7.3.1 on page 90). While metagenomic approach provides functional output about analysed community and, if the sequencing provides deep coverage, it is more accurate than the `PICRUSt` predictive metagenomic approach. 16S survey is unable to reach acceptable accuracy for strain-level microbial identification. In fact, even for higher level of species and genus, the accuracy could be questioned (see figure 6.12 on page 77). Since gene contents may differ between bacterial strains despite having identical 16S gene sequence, this survey may not be able to characterize microbial genes responsible for investigated disease states [175]. Exploring pathogenesis and potential toxicity of certain bacteria may be hindered by 16S survey, while metagenomic studies are capable of this type of analyses.

Deep metagenomic studies yield much more data. While data generated from 16S surveys could potentially be computed on a regular computing machine (PC or Mac), most metagenomic datasets would need a computing cluster to process it. Raw sequence data gathered from 2016 study cohort (see part II of this thesis, *Study group* chapter on page 102) in this experiment weighted approximately 10GB (uncompressed data) for 66 samples. While shotgun sequencing experiment being a follow up to this study (while not being part of this thesis) calculated for the same amount of samples (66) generated approximately 600GB of raw uncompressed reads. Thus sixty times more (60000%) data is generated. 16S rRNA sequencing introduces primer bias, a bias that is driven toward certain organisms. If a primer doesn't match a particular region of a microbial genome it is not sequenced [145]. Whole genome sequencing, especially after adapting recommended PCR-free library preparation methodologies, is not only free from this bias, but also avoids selective amplifications and could potentially aim at measuring absolute abundances of microbial communities (in tandem with qPCR approaches) [39]. Note that microbiota refers not only to bacteria, but also viruses and microbial eukaryotes. 16S rRNA gene analysis aims only at bacteria, while 18S ribosomal RNA surveys specifically target small eukaryotic ribosomal subunits [208]. Metagenomic studies are able to reveal full microbiome composition including bacteria, eukaryotes and viruses (virome).

Metagenomic shotgun sequencing however has some drawbacks. While in 16S surveys, a certain bacterial gene (or its fragment) is amplified, shotgun sequencing is not targeting microbial genes in general. It may therefore suffer from significant contamination from host or site genome. This hinders the metagenomic analyses. For example microbiota from human biopsies will include many sequenced copies of the human genome. This would in turn dramatically decrease depth of microbiome coverage, making it impossible to assemble microbial genes, and prevent downstream analyses. Human Microbiome Consortium [78] recommends full metagenomic sequencing in detailed studies that aim at deepening the understanding of the disease pathogenesis and identification of potential new targets for therapy, as these types of studies are able to reveal minor genomic variations within species that could cause altered phenotypes.

In summary: deep metagenomic sequencing offers better resolution and accuracy in taxonomy assignments, identification of viruses and microbial eukaryotes and functional profiling, but remains an expensive approach. At present time the costs of 16S survey is around $50 per sample, while metagenomic shotgun sequencing costs around $350 per sample. Considering the fact that researchers need to build a relatively big sample size for downstream analyses, it still remains prohibitively expensive for many researchers.

Chapter 4

Sequencing technologies

4.1 Introduction

General term *sequencing* refers to the process of determining the primary structure of an unbranched biopolymer. Restricting the definition of sequencing to DNA, sequencing refers then to determining the nucleotide order (`A` - *adenine*, `G` - *guanine*, `C` - *cytosine* and `T` - *thymine*) of a given DNA fragment. The genome of a particular organisms refers thus to the order of bases in all the DNA of that particular organism. Costs of sequencing rapidly dropped in the year 2008 (see figure 4.1) after introducing third generation sequencing - Next-Generation Sequencing (NGS). This reflects the time period of transition from second-generation Sanger-based dideoxy chain termination sequencing to modern approaches like Sequencing by Synthesis (SBS). NGS is a technology that allows sequencing of millions of molecules in a parallel manner, including samples from multiple individuals at the same time. Affordable prices increased the availability of sequencing throughout the world, this is reflected by the fact that sequencing costs exceeded Moore's Law. NGS sequencing workflow adapted for Illumina machines is reproduced on figure 4.2; it consists of several steps described in subsequent sections.

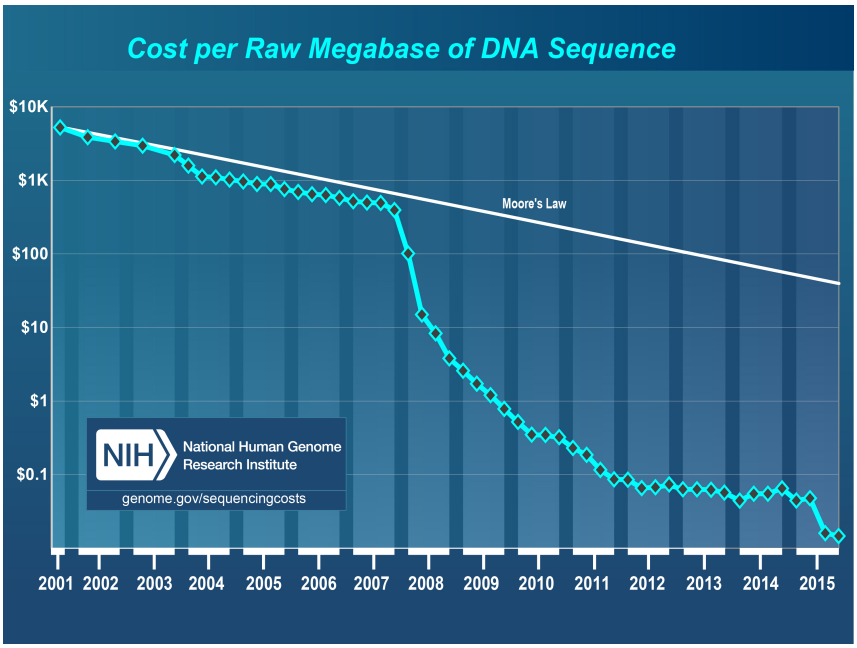


Figure 4.1: Falling costs of sequencing, and Moore's Law that describes long-term observed trends in computer hardware reflecting doubling of computing power every two years. Affordability of Next-Generation Sequencing (NGS) reportedly exceeded Moore's Law around year 2007. Source [136].

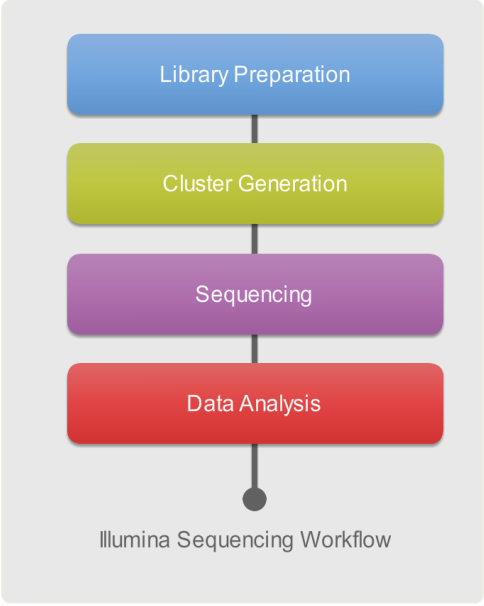


Figure 4.2: Standard sequencing workflow for NGS Illumina machines [131].

4.2 NGS Sample preparation

In order to sequence a desired sample a *library* has to be prepared either from genomic DNA or total RNA, that would become a valid input for NGS machines. A library is a collection of randomly sized DNA fragments representing the sample [3]. The aim of library preparation is to obtain nucleic acid fragments with adapter sequences attached on both ends. **Tagmentation** is a process of obtaining short, randomly cut DNA fragments ("tags"). Enzymes called transposomes are responsible for random cuts of DNA. Adapters are then added on both sides of the DNA fragments ("ligation"). The use of different adapters allow for multiple libraries to be pooled together and sequenced at the same time. It is worth noting that DNA tags that don't incorporate primers on both ends are subsequently washed away during the sequencing process - see section 4.3.2.1 on page 47. Sample preparation is therefore divided into several steps listed below (see also figure 4.3):

1. Fragmentation of DNA / target selection:
   Extracted DNA is fragmented using multiple existing methods belonging to two basic approaches: physical or enzymatic. If the sequence of a specific target region is known, such as 16s rRNA SSU, PCR amplification of those targets is performed resulting in amplicons of the desired length.

2. Addition of adapter sequences:
   Addition (annealing) of specific adapters to both 3' and 5' ends takes place. These double-stranded adapter sequences, usually ranging from 20 to 40 bp long, are complementary to two types of oligos present on a sequencing flow cell. For detailed discussion see Sequencing by Synthesis (SBS) section on page 47.

3. Size selection:
   If DNA fragmentation was preformed, fragment size selection is performed. This step is performed using various methods including gel electrophoresis or bead-based size selection method. If target selection was performed, the sample library has DNA fragments of known size that need to be verified; hence no size selection is performed.

4. Library quantification and quality control:
   There are several library quantification methods. 1) Bioanalyzer® System provides the user with information about concentration and resulting fragment sizes. 2) qPCR measures amplifiable library fragments, but without measurement of library sizes. Quantification step is important as high concentrations of DNA produce low-quality data due to flow cell saturation; low concentrations would greatly decrease coverage - the number of times a nucleotide is sequenced. Low coverage may yield reads that are producing shorter contigs, thus reducing the downstream analysis capabilities.

Sample preparation greatly effects the outcome of NGS. Significant differences in taxonomy are observed among the four different next-generation sequencing library preparations using a DNA mock community and a cell control of known concentration [39]. There exits four types of NGS applications (different sample preparations):

1. Whole Genome Sequencing (WGS),

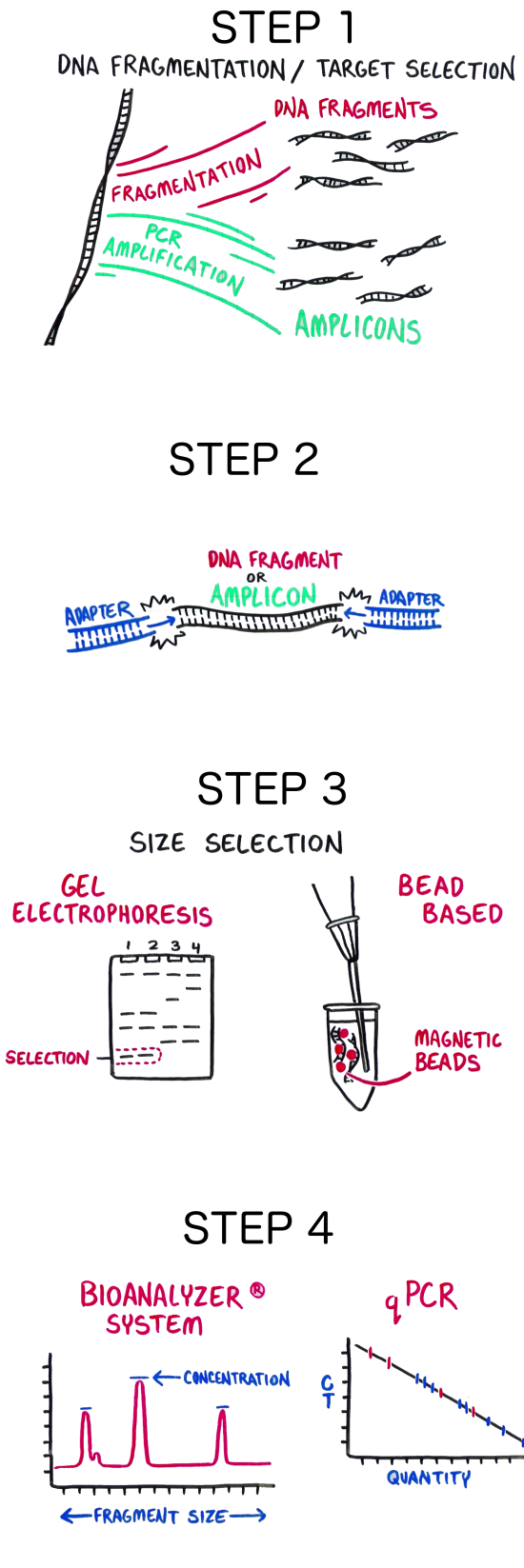


Figure 4.3: Next Generation Sequencing Library Preparation steps: **1.** Fragmentation of DNA or target selection; **2.** Addition of sequencing adapters. **3.** Library size selection. **4.** Library quantification and quality control. Adapted from [187].

2. Exome Sequencing (Exome-Seq),

3. RNA Sequencing (RNA-Seq),

4. Methylation Sequencing (Methyl-Seq)

For the purposes of brevity and for relatively small significance to this thesis they would not be discussed here. For more information see [39]. Estimating sequencing errors and problems are discussed on section 4.7 - page 57.

4.3 DNA sequencing methods

4.3.1 Sanger sequencing

4.3.1.1 Overview

Sanger sequencing, developed by Frederick Sanger in 1975, is a gold standard for sequencing. F. Sanger was awarded the Nobel Prize in chemistry in 1980 as a result. The method is used currently for routine sequencing application or for validation purposes in NGS analysis. Sanger-sequencing belongs to the so called "first-generation sequencing" technologies. Since Sanger sequencing methodology is not a part of this thesis it is only briefly described below.

4.3.1.2 Methodology

The first step involves producing many copies of single strand of DNA. There are a variety of methods used for preparation of single-stranded DNA (ssDNA); among many: asymmetric PCR, size separation on denaturing-urea PAGE, biotin-streptavidin separation or lambda exonuclease digestion [158]. Many copies of ssDNA are added to a mixture containing an enzyme called DNA polymerase, whose main function is to add complementary bases to a ssDNA. This mixture also needs the "building blocks", a single dNTPs (deoxynucleotide triphosphate) of `A`,`T`,`G` and `C`, and ddNTPs (**di**deoxynucleotide triphosphate) - i.e. "terminated" nucleotides that can't be extended. Each ddNTP carries a different fluorescent molecule (fluorophore).

dNTPs and ddNTPs are attached to complementary bases by DNA polymerase. The chain is terminated when a ddNTP is attached. At this point the mixture contains a mixture of copies of ssDNA having different lengths. The copy strands are chemically separated from each other (denatured), separated by their size using gel electrophoresis, and used to read the base sequence. As the DNA is negatively charged, smaller DNA molecules migrate further toward the positive anode, and accordingly bigger DNA molecules would stay closer to the negative cathode. Therefore the first sequence to pass the fluorescent detector would be the first base in the sequence, and further according to sequence length from the shortest to the longest [132].

4.3.2 Sequencing by synthesis (SBS)

Sequencing by Synthesis (SBS) is a technology developed by Illumina company. Rather than relaying on measurements of the gel-transported labeled fragments (Sanger sequencing, section 4.3.1 page 46), it relays on a known position of each template, attached physically to a surface of acrylamide-coated glass flow cell. The flow cell, comprised of several lanes (usually 8 - see figure 4.4), has a coating of oligonucleotides at the bottom of it, which functions to attach the short DNA strands tagged with adapters by reverse complement matching.

SBS technology also uses before-mentioned fluorophore-labelled ddNTP nucleotides. Sequencing is performed in a number of cycles, during each the addition and reading of a particular base takes place. The length of the read is hence dependant on the number of cycles. Colored dots as registered on a CCD camera are actually clusters of small regions of amplified DNA. An image is captured during each sequencing cycle, allowing it to capture the nucleotides of each cluster [132].

4.3.2.1 Cluster generation

A cluster refers to a small region of amplified DNA; every cluster contains a different fragment of amplified DNA. Clusters are generated on the flow cell described above. Flow cell oligos (primers) are complimentary to adapters added to DNA templates during DNA preparation. The purpose of clustering is to reach satisfactory signal strength for optimal detection coming from the oligos ddNTP fluorophores. Briefly, a clustering process comprises of several major steps. Some steps are illustrated with accompanying figures: 4.5 - 4.10.

1. Sample hybridization to a flow cell
   A multitude of primers are attached to the surface of a flow cell. Primers are short sequences of DNA, and since DNA polymerase can only add new nucleotides to an existing strand of DNA, those primers are complementary to the adapters on the sample fragments. As the sample passes across the surface of the flow cell it hybridizes to its complimentary oligo. DNA molecules are thus bound to the surface in a random pattern (figure 4.5).

2. Reduced cycle amplification
   This step involves the addition of motifs: sequencing binding site, indices (unique, short, usually six base-pair long sequences used to identify samples sequenced), and regions compliment to the flowcell oligos (see figure 4.6).

3. Sample amplification in a process called *bridge amplification*,
   a) A complimentary copy of the original strand is made. As sample's short fragment of DNA is bound to the surface of the flow cell, then the primer is extended by DNA polymerases, making a complimentary copy of DNA. This double-stranded molecule is then denatured, and the original template is washed away. However, newly synthesized DNA remains covalently attached to the flow cell surface. On the top of the newly synthetized DNA strand there is an adapter sequence (figure 4.7).
   b) *Bridge amplification* is a process in which a single, newly synthesized DNA strand

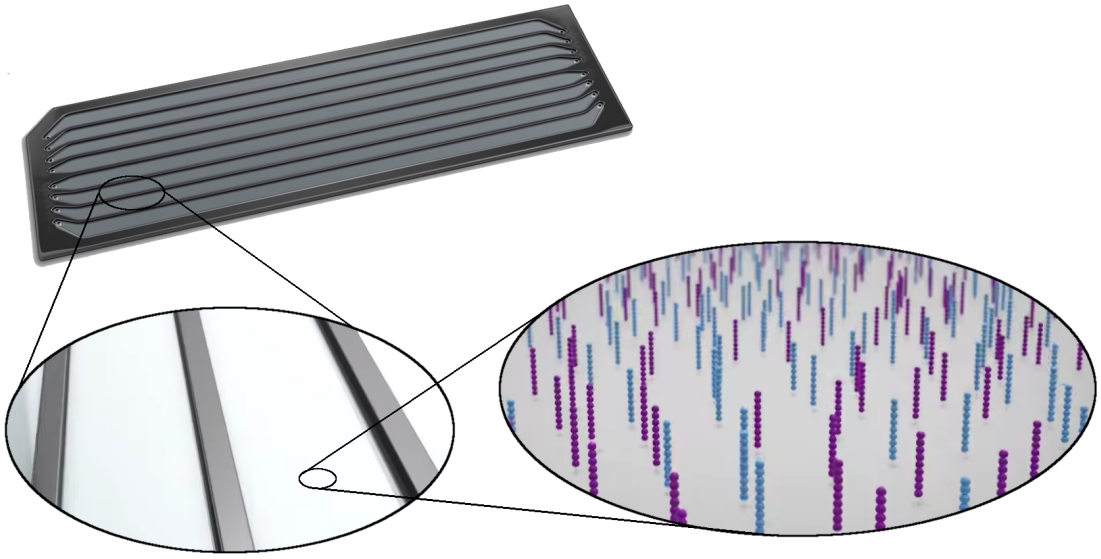


Figure 4.4: Illustration of a glass flow cell containing 8 lanes, and its surface containing two types of oligos (blue and violet). Combined materials from source [134].

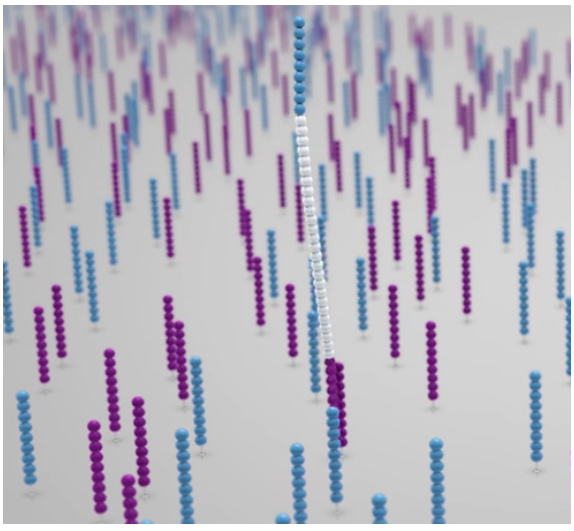


Figure 4.5: STEP 1: Sample hybridization to a complimentary oligo on a flow cell. Source [134].

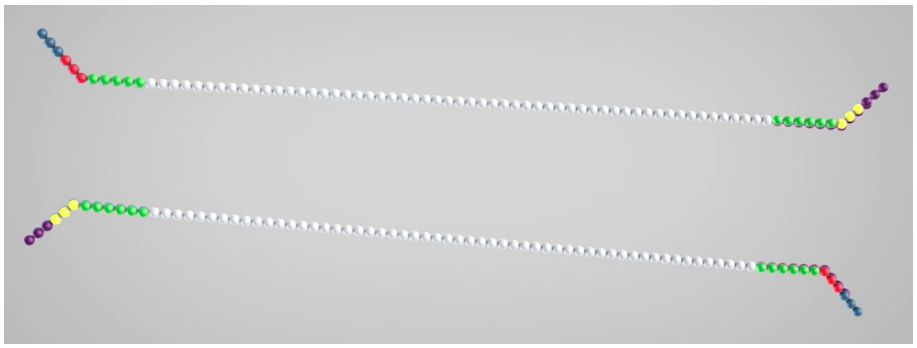


Figure 4.6: STEP 2: Illustration representing short fragments of DNA with additional motifs added: sequencing binding site, index #1, index #2, region #1 and region #2 complimentary to the flow cell oligos. Source: [134].

bends and hybridizes to an adjacent oligo on the flow cell. dNTPs and DNA polymerases are then added, to create another extension (double-stranded bridge) of the strand. This double strand bridge is denatured, resulting in two copies of covalently bound single-stranded templates (figure 4.9).

4. Clonal Amplification
   Cycles are repeated until multiple clusters are formed, resulting in multiple copies of DNA strand from a single source. Clonal amplification is essential for quality control. A single error, such as attachment of a wrong base (*"base substitution errors"*), becomes statistically insignificant, and reduces signal to noise ratio. Also forward and reverse strands that should be complimentary to one another allow for further quality checks (figure 4.9).

5. Linearization of fragments, in a state optimal for sequencing,

6. Blocking of fragments by ddNTP attachment, preventing further attachment of additional nucleotides

7. Hybridization of sequencing primers to the fragments

4.3.2.2 Sequencing process

Sequencing is performed on linearized cluster of amplified sequences - those not forming bridges. After all the reverse strands are washed off the flow cell, only forward strands remain, 3' ends are blocked to prevent priming. Primers attached to forward strands add ddNTPs with fluorescent dyes (fig. 4.10 **a**), and block addition of further nucleotides through Reversible Terminator Chemistry (RTC - see box below). Only one nucleotide gets implemented based on the sequence of the template. Each of the four bases (A,C,G,T) emits electromagnetic waves of different wavelength, that are detected (fig. 4.10 **b**).

After unincorporated ddNTPs are removed the signal for each cluster is detected. Depending on the fluorescent signal it is possible to determine which base has been incorporated. The key concept to sequencing relies on the fact that ddNTPs can be

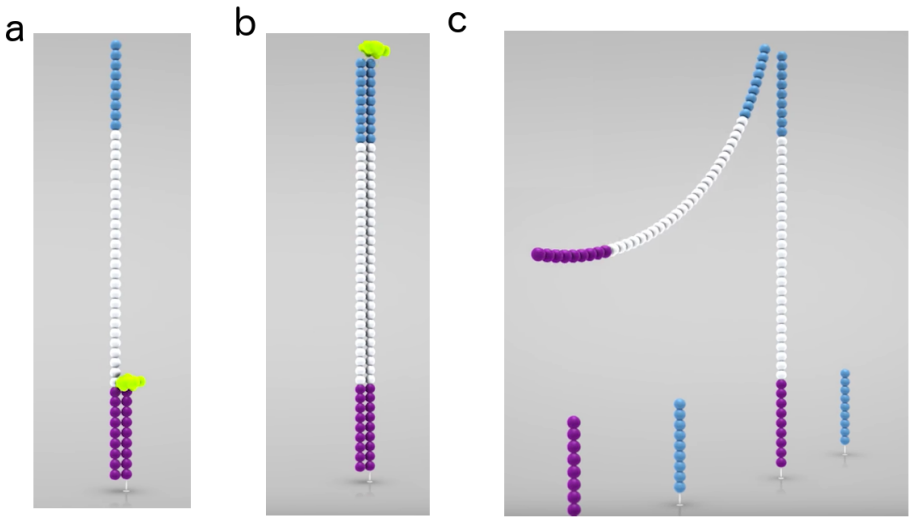


Figure 4.7: STEP 3a: The DNA fragment attaches through its oligo to a complimentary adapter on a flow cell surface **(a)**, then DNA polymerase creates a complementary sequence of the original fragment **(b)**. Finally the double-strand molecule is denatured, and the original strand is washed away. Combined materials from source: [134].

unblocked (see *Reversible Terminator Chemistry* box), and another cycle of added reagents, washing and measurement is possible.

When the first read is sequenced, the newly created strand is washed away (fig. 4.10 **c**), then index 1 read primer is introduced by hybridization to the template (fig. 4.10 **d**). Another index read is generated in a similar manner to the first read, but from a "different direction". Read product is washed off after the reading index is read. Now the 3' ends are not protected and the template folds again (fig. 4.10 **e**), but binding to the second oligo. Index 2 is read (fig. 4.10 **f**), and DNA polymerase extends the second flow cell oligo (fig. 4.10 **g**). Again, a double-stranded bridge is constructed, then denatured in order to linearize it (fig. 4.10 **h**), and 3' ends are blocked again. Original forward stand is cleaved off (fig. 4.10 **i**), and only reverse strand remains. Read 2 sequencing primer is introduced (fig. 4.10 **j**), and sequencing steps are repeated: ddNTPs are introduced, fluorescent signal is measured (fig. 4.10 **k**), and finally read 2 product is washed away (fig. 4.10 **l**).

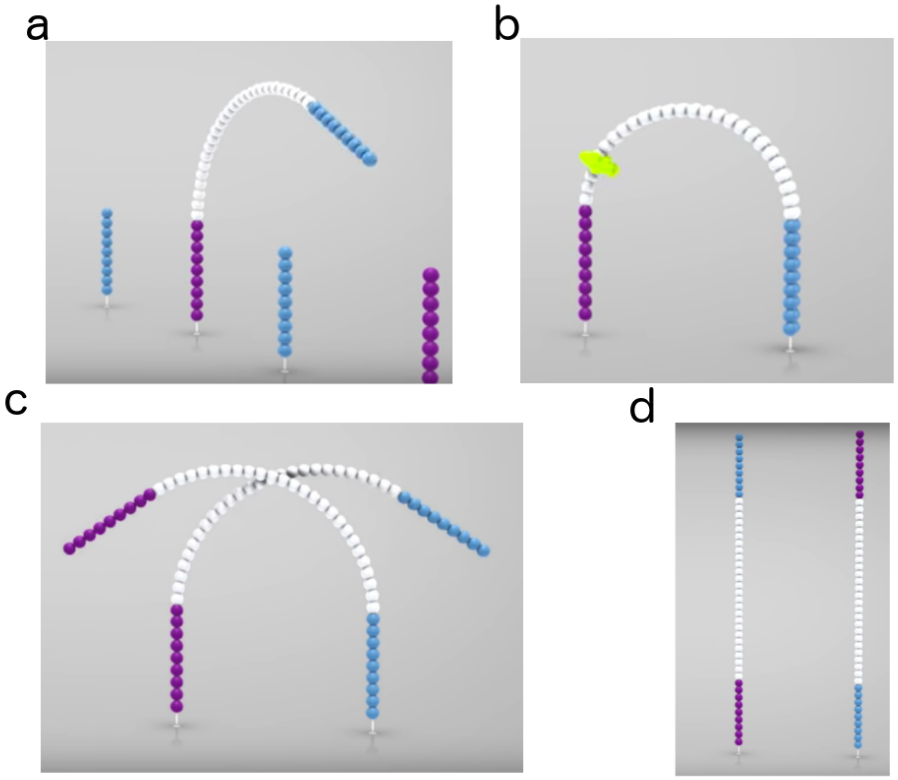


Figure 4.8: STEP 3b: Illustration of bridge amplification. **(a)** Hybridized strand folds over to an adjacent, second-type oligo on a flow cell. **(b)** DNA polymerase reconstructs a complimentary strand - a double stranded "bridge" is created. **(c)** Denaturation of the bridge. **(d)** Two single-stranded copies of the molecule arise that are attached to a flow cell. Combined materials from source: [134].

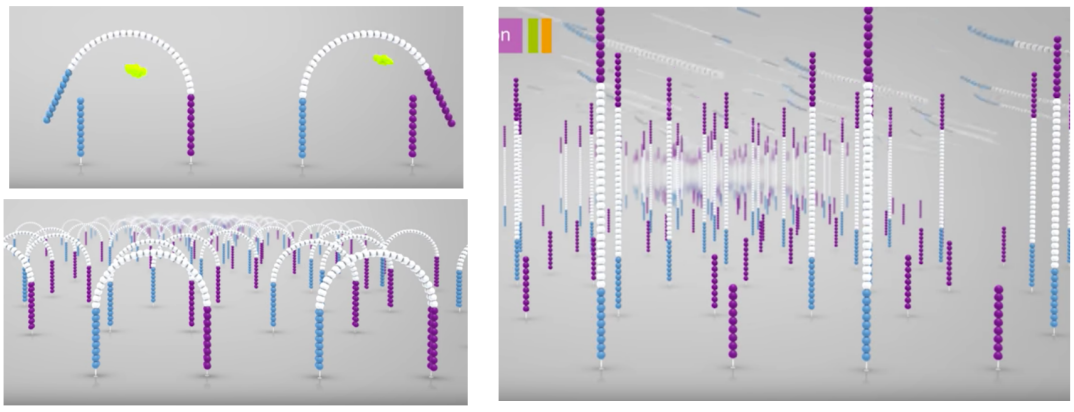


Figure 4.9: STEP 4: Illustration of clonal amplification - repeating the process of bridge parallelization. Combined materials from source: [134].

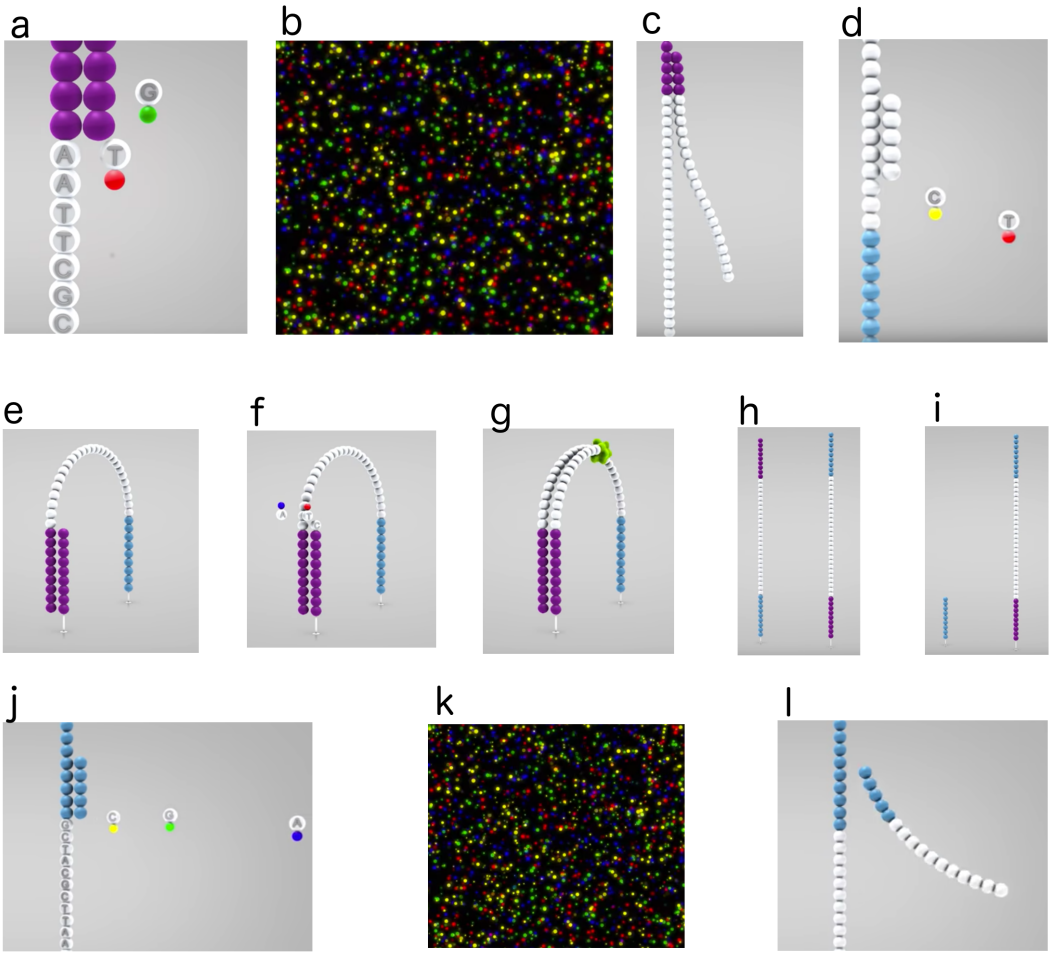


Figure 4.10: Illustration of paired-end sequencing steps. **(a)** SBS process - extension of first sequencing primer with ddNTPs. **(b)** Simultaneous registration of fluorescent signal from forward strand from all clusters. Steps (a) and (b) are repeated up to allowed/desired number of cycles. **(c)** Denaturation and washing away of read product. **(d)** Hybridization of the first index read primer to template. Index is then washed off, and 3' end read becomes unprotected. **(e)** DNA template folds over and binds to second oligo. **(f)** Second index is read. **(g)** DNA polymerase adds dNTPs making a double-stranded bridge. **(h)** Denaturation of double-stranded bridge linearizes forward and reverse strand. **(i)** Original forward strand is cleaved (washed away). **(j)** Introduction of second reading primer, and further extension of this primer with ddNTPs. **(k)** Simultaneous registration of fluorescent signal from reverse strand from all clusters. Steps (j) and (k) are repeated up to allowed/desired number of cycles. **(l)** Read product is washed away. Combined materials from source: [134].

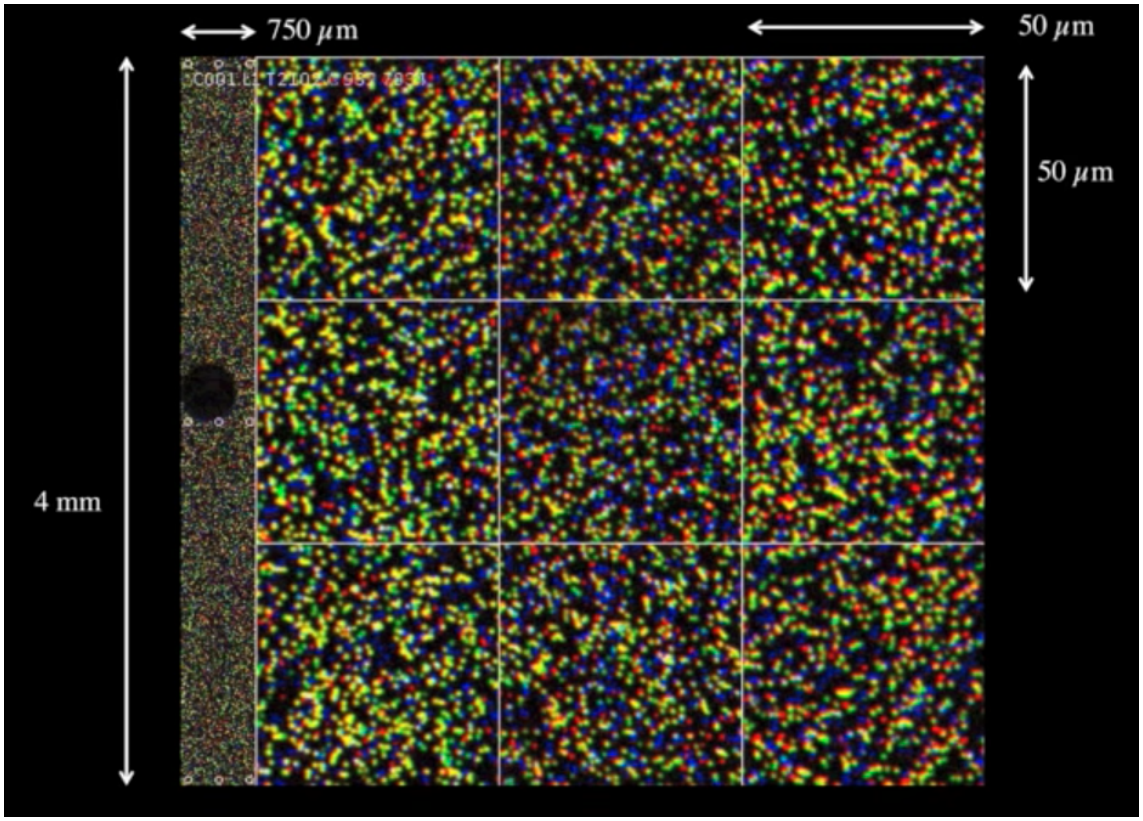


Figure 4.11: Image from a CCD camera showing a small fraction of a flow cell during sequencing. Each dot represents a cluster of sequences. Source: [179].

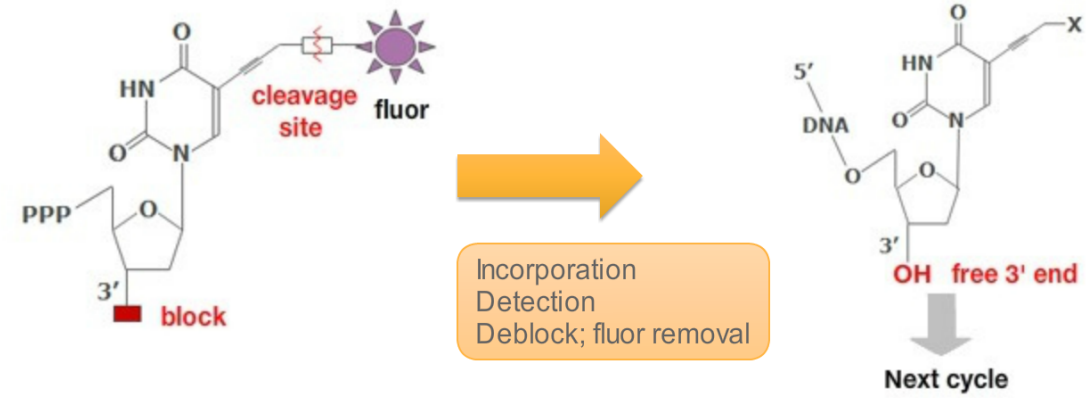


Figure 4.12: Reversible Terminator Chemistry. Source: [131].

Reversible Terminator Chemistry (RTC)

In order to ensure base-by-base incorporation of nucleotides in a stepwise manner, dideoxynucleotide triphosphate (ddNTP) is used. These four reversible terminators, 3'-O-azidomethyl 2'-deoxynucleoside triphosphates (`A`, `C`, `G` and `T`) are labelled with a different removable fluorophores. Terminating the addition of another base is essential for registration of fluorescent signal. The fluorescence is removed by cleaving fluorophore from the added base. The 3' fragment is removed allowing the addition of the next base.

4.3.2.3 Data gathering

Millions of generated reads corresponding to fragmented DNA templates are assigned to a particular sample ID based on unique indices (barcodes) used in sample preparation step. Usually barcodes are automatically stripped before producing output sequences for further bioinformatics processing pipelines. Raw data output files are in `bcl` format (see section 5.1 on page 60), which are then de-multiplexed to `fastq` files (see section 5.3 on page 61).

4.3.3 Comparison of Sanger and SBS approaches

A variety of factors differentiating Sanger Sequencing and NGS by Illumina SBS technology are summarized in a table 4.1.

Feature	Sanger Sequencing	Sequencing By Synthesis (SBS)
Sequencing	combination of color and physical movement of fluorophore-tagged fragments through a gel	combination of color, and fixed (known) position of each cluster attached to the flow cell surface
Read lengths	typically in the range of 300-800 bases	maximum of 150 bases with paired-end reads determined prior to sequencing, during run setup
Quality of data	phred scores > 20 is a golden standard quality if always visualized on electropherograms	phred scores > 30 is a golden standard quality of reads needs to be computed before manual inspection (high throughput)
Accuracy	sequencing both (forward and reverse) strands increases accuracy	generating sufficient number of reads increases genome coverage
Template validation	not required - templates are not fragmented, and map to a specific region	required - fragmented template is routinely sequenced across multiple regions
Primer design	pior knowledge of genome is required in order to design PCR and sequencing primers	pcr primers are universal

Table 4.1: Major differences between Sanger Sequencing and NGS by Illumina SBS technology. Based on: [132].

4.4 Other sequencing methods

There are several other sequencing methods in addition to first- (Sanger) and SBS methods. This section briefly characterizes some of them:

1. Pyrosequencing

Pyrosequencing belongs to SBS methodology family. It is considered as second-generation sequencing method. Pyrosequencing relies on detecting the signal of pyrophosphate release on nucleotide incorporation. There is no chain termination as described in Illumina RTC [151].

2. Sequencing by ligation

Sequencing by ligation relies on short fragments of DNA called oligonucleotides, rather than single bases in order to sequence DNA (see figure 4.13). DNA *ligase* incorporates oligonucleotides, whereas in SBS approach DNA polymerase attaches ddNTPs and dNTPs. DNA ligase ensures that only oligonucleotides with bases matching a template are incorporated. Sequencing mechanism also relies on measurement of fluorescent signals released from labeled oligonucleotides. This method is easy to implement, but is time-consuming and limited to short reads [153].

3. Ion semiconductor sequencing

Ion semiconductor sequencing belong to SBS methodology family. This approach is based on detection of H^+ ions released during the polymerization of DNA [152], and does not require modified nucleotides tagged with fluorescent fluorophore. Amount of hydrogen ions (H^+) released is measured (i.e. pH) by the ISFET sensors placed on a number of micro-wells, and then translated into corresponding base calls. This method is rapid and affordable in terms of costs, but generates shorter reads compared with Sanger or pyrosequencing. At the same time it has problems with registration of homopolymers (like. AAAAAAA) which results in greater pH change, and loss of resolution.

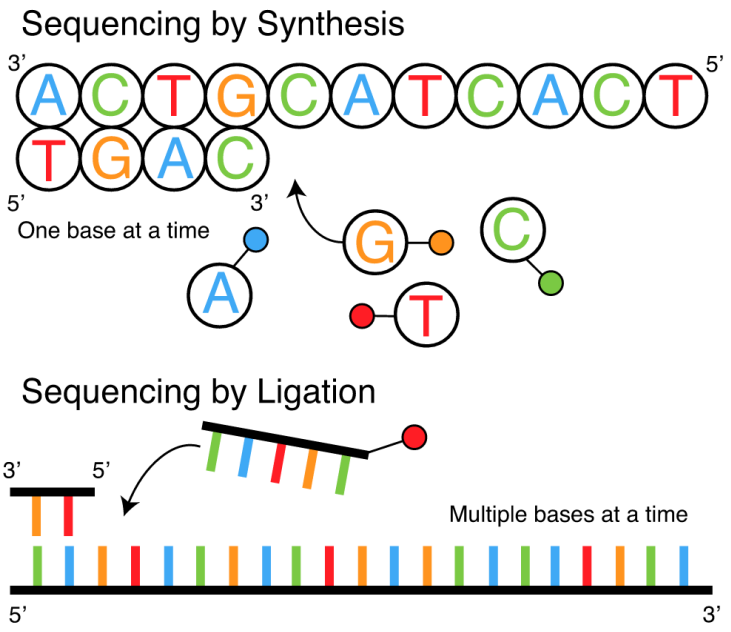


Figure 4.13: Illustration comparing SBS and sequencing by ligation. Source [153].

4.5 Sequencing error estimation

Sequencing machines provide not only sequencing reads, but also accompanying quality for each base in a read. Quality scores (Q-score) are expressed in a logarithmic relation of base-calling probabilities. This Q-score is referred to as Phred score, which is then later converted to a single `ASCII` character relating to each base - see section 5.3.2 on page 62 for more details.

Phred score reflects predictive measure of quality. Schirmer et al. [44] observed that predictive quality scores do not accurately reflect errors of amplicon sequences, and are underestimated. They conclude their work stating that perhaps newer imaging strategies, such as 2-channel SBS chemistry, will yield measured error rates that correspond to actual error rates.

Another approach for estimating quality/error rate is to map produced reads to a reference genome. Number of mismatches allow to produce a observed measure of empirical quality or raw read accuracy. However, this empirical approach counts valid biological variants as mismatches to a reference genome. Ths is one of the challenges in human genomics or metagenomics.

4.6 phiX control

The phiX 174 (or $\phi X174$) bacteriophage is a virus and was the first DNA-based genome to be sequenced [20]. In the NGS sequencing domain it now functions as a control for prepared libraries. There is one major reason why phiX is widely used as a "spike-in" control [133] - it has a relatively small, known genome that could be aligned to a reference database for error estimation. Generating PhiX libraries allows

for quality control in cluster generation, sequencing, further alignment or calibration for potential cross-talk (see chapter 4.7.1 on page 57).

During the sequencing process an algorithm operating in "real time" (i.e. during the sequencing process) aligns complete reads to the PhiX reference. Sample sequencing success is dependent on the potential alignment to phiX genome, which usually takes place during the 25th cycle [133]. PhiX is especially recommended for non-mammalian samples, as those are characterized by unbalanced nucleotide composition, i.e. unequal proportion of nucleotides `A,C,G` and `T`. PhiX is not tagged with indexes during the sequencing run. Low diversity libraries are libraries where a significant number of the reads have the same sequence. Illumina company recommends for 16S amplicon sequencing using a minimum 5% of PhiX to serve as control, as this types of sample produces low-diversity libraries.

4.7 Sequencing errors and problems

There are a variety of sequencing errors and problems that influence sequencing results. Those errors span from library preparation techniques, associated for example with PCR-related errors, or contaminating bacteria (during DNA extraction), up to sequencing limitations. Produced reads lose overall quality with increasing read length. Another type of errors are inherent to the nature of tagmentation-based sequencing - cluster mixing.

A very thorough and detailed description for Illumina platform used for sequencing DNA extracted for this experiment is found in publication: [44]; other errors are discussed in [7] . One of the major techniques to investigate any potential biases in obtained sequences is to use mock communitys - an artificially lab-cultured group of bacteria with known abundances (see section 6.5 on page 87). Those communities oftentimes serve as positive controls in microbial studies, such as metagenomics and 16S surveys. Establishing analysis pipeline over expected microbial (mock) communities at multiple parameter thresholds for each step used in constructed pipeline (ex. sequence filtering, OTU picking, read-pair merging, etc.) can yield sufficient analytical setup for conservative and accurate representation of microbial communities. This chapter only briefly discusses some potential errors that have the most profound impact on the relevance of this study. For more information please refer to the cited publications.

4.7.1 Cross-talk

Cross-talk (also known as cluster mixing) is an error occurring among multiplexed samples within a single sequencing run. Samples are multiplexed in order to identify the sample of origin. When a read from one sample is wrongly assigned to another sample, a cross-talk error occurs [91]. Cross-talk error significance grows with the number of multiplexed samples, where each sample represents a small fraction of the total number of reads. Illumina has reportedly around 2% cross-talk error rate, where for a given OTU the number of reads assigned to a signle sample could be inflated up to 0.5% [96].

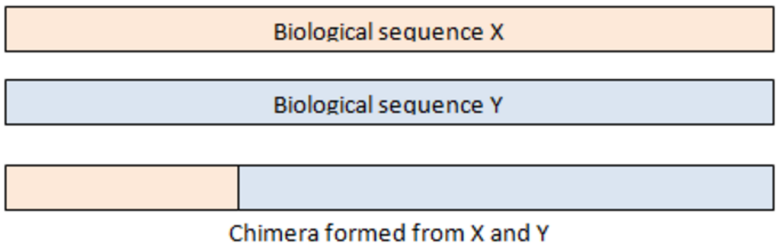


Figure 4.14: Illustration of chimeric sequence formation. Source [90].

Researchers attribute cross-talk to multiple phenomenons, such as: experimental mistakes, cross-contamination during primer synthesis, multiple misread bases within index sequences, or remnants from previous sequencing runs on the same machine [214]. Cross-talk error directly influences false-positive identification of the species; if computed OTU table shows that around 0.5% of the reads are assigned to a particular sample, then the correct count could be zero [96]. Cross-talk influences downstream analyses. It has been shown that it has the greatest impact on alpha diversity metrics, while still influencing beta-diversity calculation [7]. One way to reduce this type of errors, and its influence on downstream analyses, is to use phylogenetically-aware distance metrics for alpha and beta diversity, *phylogenetic diversity (PD)* and *UniFrac* respectively, or novel cross-talk detection algorithms, like `uncross` [96].

4.7.2 Chimeras

Chimera sequence is a sequence that is formed from at least two biological sequences joined together. They primarily form during sequencing when incomplete products from previous cycles start acting as primers during the extension step [8]. Chimeric sequences are very rare in shotgun sequencing, but common in amplicon sequencing where amplification of closely-related sequences take place. Chimeric template is created usually in one sequencing cycle, and then is amplified during subsequent cycles producing chimeric amplicons. Chimeras produced during early cycles will be highly perpetuated, and the effect is amplified by high-cycling conditions. Other approaches involve computer science methods [90]. Some researchers [60] reduce chimera formation by using low temperature extension step (4 minutes at 60 $^{\circ}C$), or utilizing low-cycling conditions [16].

4.7.3 Library preparation

Differences in library preparation methodologies have been shown [39] to affect metagenomic studies in terms of inferred taxonomic composition and functional predictions. In this paper researchers analyzed mock community - artificially construed microbial communities - as well as stool samples of patients treated with amoxicillin for different time-points. In both set of samples they observed differences in taxonomic abundances, however the overall diversity measured through Shannon index

didn't manifest statistically significant differences.

PCR-free library preparation kits produce libraries that are characterized by low duplication rates, low numbers of low-quality reads and longer contig length compared to PCR-based systems (for metagenomic studies). However, both methods, PCR-free and PCR-based, generate long contigs on mock community datasets (150–178 kb and ≈142 kb accordingly), and relay on the same preparation steps, i.e., DNA fragmentation, library amplification, and size selection. Researchers pose a strong suggestion to adapt calibration controls in metagenomic research by inclusion of mock communities (see section 6.5 on page 87). Is is of particular importance in metagenomic research, where standardization to one system in order to reduce errors is not feasible due to constant changes in the technologies, character of microbiome samples and limited scientific budgets. Developing a standardized practices would reduce the risk of over-interpreting generated datasets that often generate conflicting results, and hinder further progress in the field.

Amplicon sequencing (like 16S rRNA gene surveys) relay on PCR. Methodological tests performed on mock communities revealed that among falsely-positive OTUs identified in their dataset [7] one third were contaminant arising from cluster mixing (i.e. cross-talk), and two thirds were attributed to sequence variants of the constituent OTUs. Those variants arise during PCR amplification, and probably the sequencing process itself. Although cross-talk could be reduced by double-indexing [144] or certain algorithmic approaches (like the one used in this study through `uncross` algorithm [96]), PCR-related error remain inherent to amplicon sequencing. One solution it to adapt aforementioned PCR-free methods for Whole Genome Sequencing (WGS) protocols, but that is not possible for 16S rRNA gene surveys.

Chapter 5

Data formats

5.1 BCL format

BCL stands for "basecall", since Illumina sequencing machines generate per-cycle files as primary sequencing output. Downstream analyses use FASTQ files that correspond to per-read structure. BCL file format is binary, while FASTQ is encoded by simple ASCII characters (see FASTQ file format on page 61). Conversion from BCL to FASTQ files is automatized in Illumina machines, and starts as soon as the first read has been completely sequenced. It is worth noting that it is during this conversion that the demultiplexing process takes place. The samples are identified by index sequences that were attached to the template during sample preparation [129]. During the conversion of BCL to FASTQ, detection of Illumina specific adapters may take place, in which case matching nucleotide sequences are changed to "N" character.

5.2 FASTA file format

FASTA file format is very common in the domain of bioinformatics. Files often consists of multiple entries, each entry constructed of two sections:

1. **Sequence identifier** - starting with ">" character, indicating a FASTA record beginning; it is often followed by an arbitrary number of unique sequence identification characters, provided they reside on the same line,

2. **Read sequence** - string of characters that often correspond to biological entity (like nucleic acid or amino acid); there is no formal restriction as to how many lines a particular sequence has, nor of what letters the alphabet is comprised.

Characters in a sequence may contain the standard four nucleotides ACGT, and may sometimes be extended by N character that corresponds to unknown nucleotide. FASTA format is also used for amino acid sequences. Possible characters are standardized by International Union of Pure and Applied Chemistry (IUPAC), and can be found on bioinformatics.org/sms/iupac.html.

Although it is not formally required, many algorithms and software packages assume that sequence lines always wrap at the same width (except for the last sequence

line). Another "good practice" is to use upper-case letters, as some software packages assign a different meaning to corresponding lower-case letters [62]. Sometimes lower-case letters are used to indicate a repetitive region of the genome, ex.:

```
ATGACGAGCatgagcACGTGAC
```

here `atgagc` indicates that this part of a sequence is present in multiple locations of the larger genome. In other applications lower-case letters may be omitted.

Due to the fact that `FASTA` format definition is not very strict, it is within the user responsibility to draw attention to produced outputs, and required inputs for each particular software application. `FASTA` format has a file extension "`.fasta`", however, there are some variations that retain the same file structure:

- `.fna` - `fasta` for nucleic acid sequences,
- `.faa` - `fasta` amino acid sequences

5.3 FASTQ

5.3.1 FASTQ format overview

`FASTQ` format is a golden standard in the field of bioinformatics. All main sequencing instruments represent data in this format. It is often regarded as a variant of `FASTA` format (see paragraph 5.2 on page 60), that includes a quality score for each nucleotide in a sequence read.

`FASTQ` format is a multi-lane format where each read consists of four sections:

1. **Sequence identifier** - starts with `@` character followed by sequencer identification name,
2. **Read sequence** - the actual string of nucleotides,
3. **Spacer** character - "`+`", that may be optionally followed by the same sequence identification found in section one,
4. **Quality score** - given for each nucleotide in line two; must be the same length as the read sequence in section two.

Sequence identifier (@) can contain additional information that are standardized for particular applications. Illumina machines include information like: unique instrument name, run and flowcell identification number, indexing sequence and many others. Sequences from NCBI Sequence Read Archive include NCBI-assigned identifier [211]; however, it can be (and often is) changed during downstream analyses (like OTU picking). When pooling (i.e. concatenating) entries from different samples into one file, it is crucial to be able to distinguish from which sample a particular sequence comes from.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Figure 5.1: An example of a single `FASTQ` record [211].

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|    |    |    |    |    |    |    |    |
0....5...10...15...20...25...30...35...40
|    |    |    |    |    |    |    |    |
worst.................................best
```

Figure 5.2: Characters encoding Phred33 quality scores (in ascending order), and some of their corresponding phred scores. Source [62], *FASTQ format*.

Quality scores are calculated by the machines from fluorescence peak shapes, and any potential overlaps at every base. Scores are computed from a logarithmic equation (see sub-paragraph *Phred score*), and then encoded via single letter encoding. The aim of an encoded quality value is to represent multi-digit number by a single character in order to ensure that quality and sequence reads have the same length.

The current standard is often called *Phred33* (as shown on figure 5.2), where "33" denominates the shift by which ASCII letters are moved. *Phred64* also exists, shifting by a corresponding number of letters.

5.3.2 Phred score

Phred quality scores characterize the quality of DNA sequences, and can be used to compare the efficacy of different sequencing methods [213]. Phred score is used to represent error probabilities. It is directly computed through the logarithmic equation below [101]:

$$Q = -10 \cdot log_{10}(P) \tag{5.1}$$

- Q - phred quality score,
- P - base-call error probability.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Figure 5.3: Phred quality scores are logarithmically linked to error probabilities. Directly reproduced from [211].

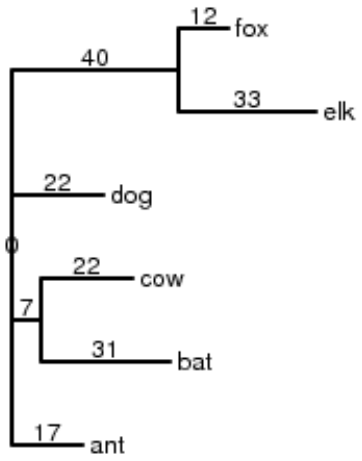


Figure 5.4: An example of Newick tree corresponding to a structure on equation 5.2, page 63. Source [121].

5.4 Newick Tree format

Newick tree format stores spanning-trees with weighted edges and node names in a minimal file format. In 16S rRNA gene surveys or Whole Genome Sequencing (WGS) downstream analyses usually store phylogenetic trees and taxonomies in this file format [85]. Rooted trees are represented as labeled nodes specifying length between parent and child. The tree ends with a semicolon "`;`". Trees are represented as an annotated nested list of node names. Node names are a letter optionally followed by more letters, digits or an underline. [105].

An example tree reproduced from [121] is showed below:

```
(ant:17, (bat:31, cow:22):7, dog:22, (elk:33, fox:12):40);
```
(5.2)

5.5 Biological Observation Matrix format

The Biological Observation Matrix (BIOM) file format is a general-use format created for feature count representation per sample basis. It is recognized by Genomics Standards Consortium. BIOM can represent features like OTUs, KEGG Orthologs (KOs), KEGG Pathways and any other feature. It has a broad application in comparative analyses, like marker-gene surveys.

BIOM format holds a matrix of features and counts of a particular feature observed per sample (or category). In addition BIOM has implemented structures to hold various metadata [15]. The latest BIOM format uses `HDF5` data model for storage and management, while the previous versions used `JSON` (JavaScript Object Notation) or plain text files. `HDF5` data model is designed for time efficient input and output operations for high volume complex data, as encountered in genomic research. More technical information can be found on the HDF Group website [120] or BIOM consortium website [15].

Chapter 6

Bioinformatics

Bioinformatics is a relatively new field of interdisciplinary science. It is not only a field merging branches of computer science and biology, but also statistics (biostatistics), mathematics and big data-science approaches, like deep- and machine-learning.

This chapter only briefly characterizes some aspects of bioinformatics relevant to the field of microbial ecology thorough Next-Generation Sequencing (NGS) technologies. It discusses standard operating procedures (SOPs) for data processing in 16S rRNA survey, most of which are applied in a pipeline constructed for this study (discussed in part II: materials and methods, chapter 15.2, page 118). It covers such topics as: filtering of the sequences, merging read-pairs, OTUs picking, clustering, assigning taxonomy, diversity analyses with visualization techniques, statistical testing, mock communitys role in enhancing data analyses, and more general nuances about biomarker discovery, PCoA and `UniFrac` statistical approaches.

6.1 16S rRNA survey analysis pipeline

This section discusses several steps necessary to process data from microbial amplicon sequences.

6.1.1 Sequence filtering and adapter removal

Obtained sequences have accompanying quality scores (see for example `fastq` format overview and `phred` quality scores: section 5.3 page 61). There are a variety of approaches to filter raw sequences produced by sequencing machines. The most simple and straightforward approach is to specify a *cutoff point*: when nucleotide quality in a read falls below specified level, the read is trimmed, and subsequent bases are deleted. Other approaches implemented, including `timmomatic` software package [69] used in the analysis pipeline for this study, use a sliding-window technique. This technique used performs sequence quality trimming based on average quality values within specified window width. Window width refers to the number of nucleotides considered in a single step. The first filtering step might ensure that each read has a specified minimum and maximum length, which can be beneficial to apply to 16S amplicon survey, as reads already have expected lengths that shouldn't be exceeded nor fall too short of this value.

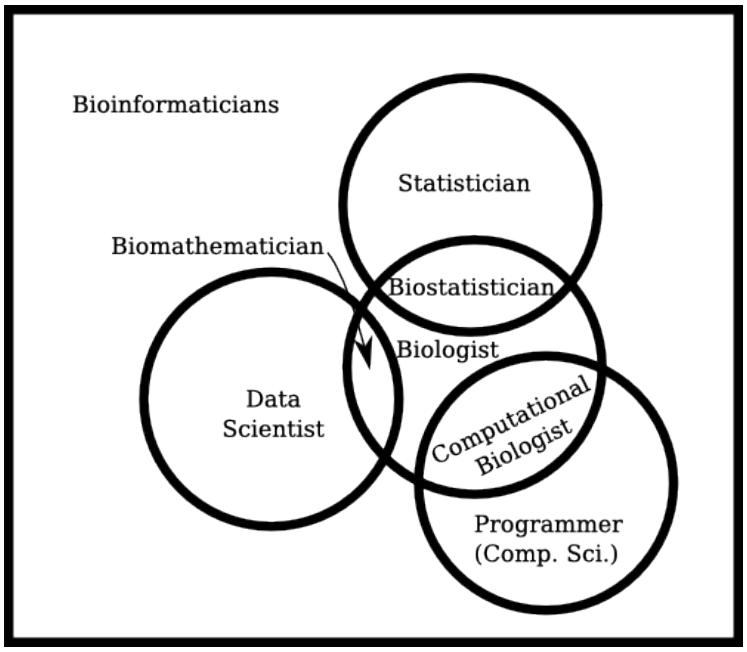


Figure 6.1: Bioinformatics as an interdisciplinary field combining statistics, data science, biology and computer science. Source: [104].

The role of adapters for sequencing is discussed in section 4.3.2 (page 47). Before sequences are subject to further processing steps, adapter sequences at both ends have to be removed. Adapter sequences are usually several nucleotides long, and are established during study design. `Trimmomatic` is a software package able to remove Illumina-specific adapters. Resulting sequences (still in `fastq` format) after initial filtering are subject to read-pair merging.

6.1.2 Merging read-pairs, additional filtering

Merging read-pairs produced during paired-end sequencing creates longer consensus sequences, and consensus quality scores. As read quality at the end of sequences decrease due to deteriorating fluorescent signal, one corrective approach is to merge sequences obtained from forward and reverse reads. This approach is well established for relatively short reads, like amplicon reads having lengths approximately 250bp, while each read produced has length approximately 150bp. This approach is then able to assembly consensus sequences corresponding to amplicon of desired region, like V4 region of 16S rRNA gene for microbial identification [95]. Often the maximum number of differences in nucleotides is the parameter which users can adjust for specific needs (like in `usearch` software package [98]).

Consensus sequences (in `fastq` format) can be then further filtered. This additional filtering step is performed after truncation methods described in previous subsection, and considers total expected errors allowed - E. Expected number of errors is the sum of the error probabilities, and those probabilities are reflected by aforementioned `phred` quality scores. It is most common to establish $E < 1$; then the most probable number of errors is zero.

6.1.3 Decontaminating sequences

During DNA extraction there is high probability of microbial contamination. When extracting microbial DNA, it is a standard procedure to prepare controls. Control samples, also referred to as *"blanks"*, are samples that underwent the same extraction procedures, except for sample DNA addition. "Blanks" are then sequenced on the sequencing machine after "tagmentation" (see section 4.2 on page 44) and PCR amplification steps. Usually those samples produce reads corresponding to contaminating sequences. Sample sequences are searched against contaminating sequences through algorithms such as `blast`. Resulting sequences that do not match control sequences are then considered as decontaminated sequences, and are subject to subsequent processing steps. Matching is set according to user-specified similarity threshold, where 98% or 99% are the most often used parameters.

6.1.4 Dereplication of the reads

Dereplication is a process of finding unique sequences from the input file. As input files might have many identical copies of particular reads, for OTU picking (discussed in the next section), it is important to provide a set of unique (i.e. dereplicated) reads. This process is computationally intensive as sequences are compared nucleotide by nucleotide. Dereplication is performed on one file consisting of sequences coming from all samples sequenced in the experiment. Sequences are compared over full lengths, thus sub-strings would not yield a match. Often the minimum abundance is set above one, as singletons (read with a sequence that is present once) might come from erroneous sequences with several incorrect base-pairs. Some researchers recommend discarding singletons, as they may be attributed to independently and randomly distributed errors [94]. Therefore, one of the subsequent steps (OTU picking) often doesn't consider singletons for potential bias it may introduce.

6.1.5 Sequence clustering

After finding unique sequences through dereplication process (see section 6.1.4), all sequences from samples in the study have to be clustered into OTUs [145].

6.1.5.1 Clustering approaches

Centroid-based clustering (see figure 6.2) is a very common method used to cluster sequences. This approach relies on picking a single sequence - a *"centroid"* - to which other sequences are compared. If a sequence displays a certain amount of similarity above a certain threshold (like 97%), then this new sequence becomes a part of that centroid. However, if similarity does not exceed the required threshold for any cluster, then this new sequence becomes a new centroid onto which all other sequences are additionally compared. This process is iterated until all sequences are analyzed. Size of the cluster (i.e. similarity threshold) is irrespective of the approach, i.e. can be user-defined to match study requirements and/or prior knowledge. Similarity threshold greatly increase the expected number of clusters - see figure 6.6 on page 72.

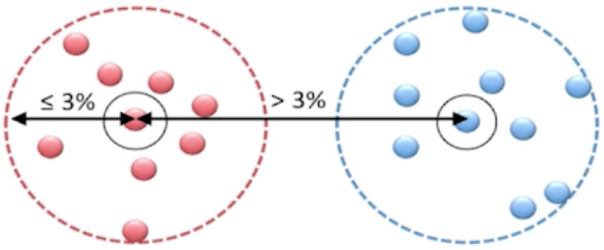


Figure 6.2: Centroid-based clustering. Source: [145].

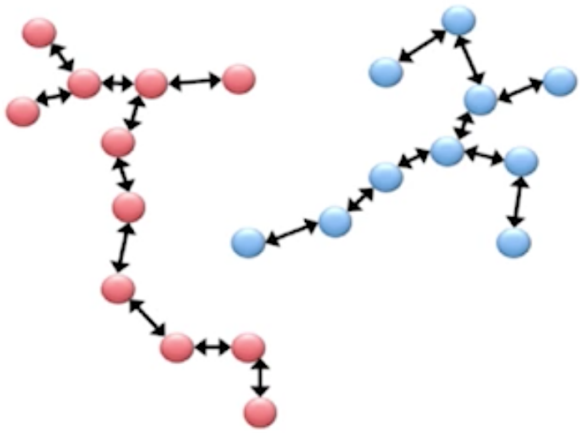


Figure 6.3: Single-linkage based clustering. Source: [145].

Single-linkage-clustering, often referred to as "*friend-of-a-friend* clustering", is a type of agglomerative clustering that is also iterative. Each iteration contains a step that assigns a sequence within a particular reach distance to another sequence already belonging to a cluster, or initiates a new cluster. This method doesn't specify cluster boundary, like centroid-based clustering in which similarity threshold is specified for a centroid sequence. For this reason it can produce long clusters in which sequences at the "opposite ends" may not share much similarity, which leads to additional problems. It is not often used for microbial studies.

Complete-linkage clustering is a combination of two aforementioned approaches: *centroid-based* and *single-linkage* clustering. Sequence considered has to remain within a particular *reach distance* of closest sequence belonging to a cluster, while at the same time this particular sequence has to be within a certain threshold range to a centroid sequence. *Complete-linkage* clusters tend to have a spherical structure (see figure 6.4 on page 69).

Average-linkage is another approach that is a compromise between single- and complete-linkage clustering. In this approach, a sequence has to lay within a particular *reach distance* (from *single-linkage* clustering), and then the average distance

Figure 6.4: Complete-linkage-based clustering. Source: [145].

between sequences is computed. If the average distance between sequences exceeds the threshold (ex. 3%), then a particular sequence is either tested against another cluster of sequences, or becomes a centroid for a new cluster if no matches are successful.

***Zero-Radius* OTU picking** approach is a relatively new method introduced by independent researcher Robert C. Edgar [97]. This method does not relay on similarity thresholds for reasons discussed in section 6.1.6 (page 70). Picking Zero-Radius Operational Taxonomic Units (zOTUs) is performed through `UNOISE2` algorithm: whether a sequence `M` is a valid member of the cluster defined by a centroid `C` is dependent on several parameters:

- a_c, a_m - abundances of the centroid and member sequence respectively,
- d - Levenshtein distance (number of differences including both substitutions and gaps),
- α - user-settable parameter [default 2]

Abundance skew is defined as skew(M,C) $= \frac{a_m}{a_c}$. If a particular sequence (`M`) has relatively small abundance (small skew) and small Levenshtein distance, then it is probably a read of centroid sequence (`C`) with small number of `d` point-errors. This algorithm is represented by a phenomenological model as described by inequality below:

$$\beta(d) = (\frac{1}{2})^{\alpha d+1} \geqslant \text{skew(M,C)} \tag{6.1}$$

When this inequality is met, then a particular sequence `M` is part of centroid `C`. β function was based on results obtained from several mock communities. [97].

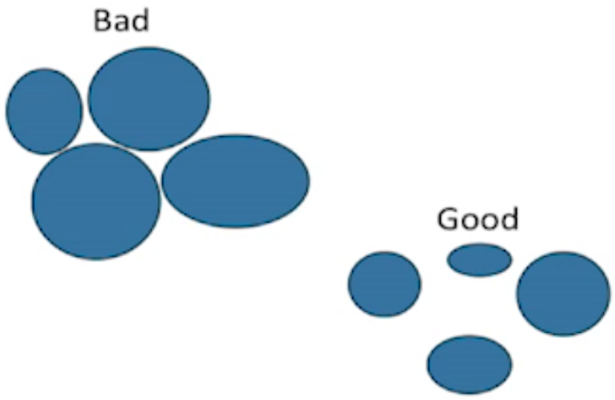


Figure 6.5: Schematic representation of quality of clusters. Bad quality of clusters are characterized by overlapping regions - i.e. relatively small between-cluster distance. Good quality clusters are separated between each other - big between-cluster distance. Source: [145].

6.1.5.2 Quality of the clusters

In order to quantify the quality of the clusters produced, within-cluster distance and between-cluster distance are measured. Good quality of the cluster is represented by large between-cluster distances and small within-cluster distances (see figure 6.5, page 70). Quality of clusters is oftentimes described in *Silhoutette index*, an absolute measure of cluster quality giving a score in the range of $(-1, +1)$ (see table 6.1, page 71). *Silhoutette index* is defined by an equation:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} - 1 \leqslant s(i) \leqslant 1 \tag{6.2}$$

where:
$a(i)$ - average distance to nearest other cluster,
$b(i)$ - average distance to assigned cluster

There are many models to quantify the quality of clusters, such as: Calinski-Harabasz relative measure, and prediction strength or mixture models, that have not been discussed here.

6.1.6 Picking OTUs

Operational Taxonomic Unit (OTU) term comes from numerical taxonomy [186] and is vaguely defined as the object of study, thus this term can be attributed to individual organisms, or certain taxonomic ranks, such as phylum, genus, etc. NGS technology does not observe individual organisms directly. The most common practice for OTU picking - identifying sequences coming from different bacteria (at given accuracy) is to use dereplicated (see subsection 6.1.4, page 67) and then clustered reads (see subsection 6.1.5, page 67), and finally select (*"pick"*) OTUs.

Oftentimes sequences that share 97% or more similarity (comparative sequence alignment) are considered as one OTU. This approach, although backed-up by thor-

Range of SI	Interpretation
0.71-1.00	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial. Try additional methods of data analysis.
< 0.25	No substantial structure has been found

Table 6.1: Ranges for silhouette index values, and their interpretation for quality of the clusters. Source: [145].

ough scientific investigations [147], has several drawbacks. First, there are different species having much more similarity at the nucleotide level, which would result in an OTU representing multiple species. Secondly, the opposite situation is possible: certain species might have paralogs (genes related by duplication within a genome) that share less than 97% of similarity (see figure 6.8 page 73), which would result in two observed OTUs associated with two species, where in fact one is present. Finally, the third possible error is the presence of spurious clusters that arise due to sequencing errors (see section 4.7, page 57) like chimeras and read-errors [93]. As considered, similarity threshold greatly impacts observed number of OTUs (see figure 6.6, page 72), OTU picking with sequence clustering plays one of the most important roles on the effects of later, downstream analyses.

6.1.6.1 General OTU picking approaches

This subsection briefly characterizes several OTU picking approaches with relation to prior knowledge stored in databases.

***De novo* OTU picking** is an approach where reads are clustered based on similarity threshold or others like zOTUs (see section 6.1.5.1 page 69). This method does not relay on any prior knowledge, i.e. databases of known sequences. Instead it utilizes all reads and allows for study of unknown species, however from computational perspective is relatively slow, as it cannot be parallelized [145], [173]. Another issue with this approach is that, since there is no standard reference (i.e. database), results from one study using *de novo* picking cannot be compared directly to results from another study. In order to compare studies, data from them needs to be combined and recomputed. This approach is used for uncharacterized environments, like soil and water samples from different regions.

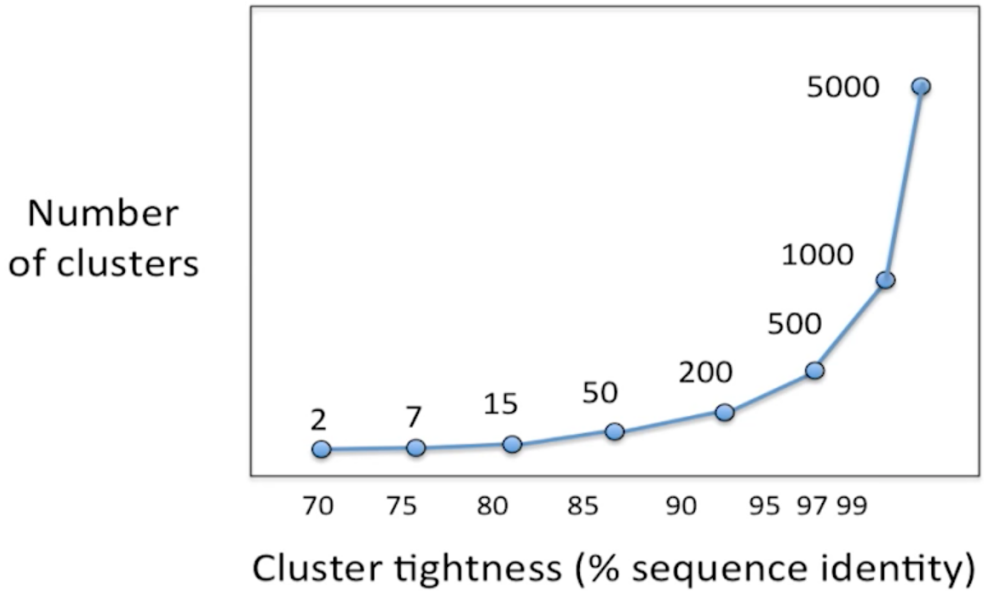


Figure 6.6: Number of created clusters as a function of cluster tightness. Source: [145].

***Closed-reference* OTU picking** method uses predefined database of OTU clusters, to whom closest matching sequences are assigned. Sequences that fail to match to a database of cluster sequences are not considered, i.e. discarded (see figure 6.7, page 73). This approach is easily parallelizable, OTUs are defined by high-quality trusted sequences where error rates are strictly controlled [145], [173]. This method serves as another quality filter for errors (PCR and sequencing errors). This method however would never yield new species, and the observed results are highly-depended on the database used. As reference databases are growing in rapid succession it is necessary to update and re-compute analyses for further comparative analyses. One potential drawback is that reference databases might still have erroneous representative OTU sequences. This approach is used for well-characterized environments like human/mouse gut, oral or skin samples.

***Open-reference* OTU picking** is a hybrid of *de novo* and *closed-reference* approaches. First OTUs are picked in a *closed-reference* manner, then sequences that failed to match are subjected to *de novo* OTU picking. In this approach all reads are clustered, however, the method in only partially parallelizable. It is used for relatively uncharacterized environments that have been previously identified (to some extent), like soil and water samples [145], [173].

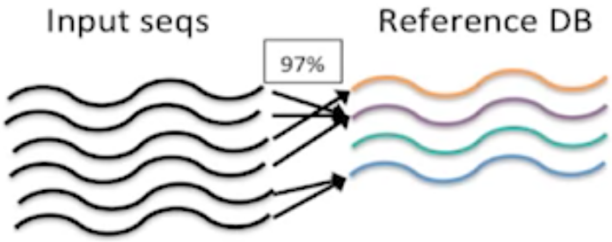


Figure 6.7: Closed-reference OTU picking. Mapping sequences to a reference database= that have up to 3% of variation (97% sequence similarity). Source: [145].

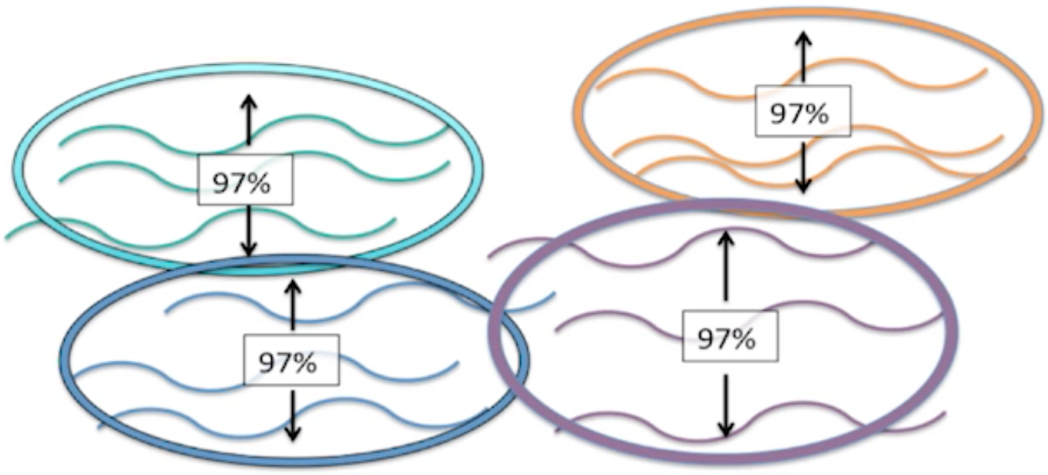


Figure 6.8: Clustering sequences that have up to 3% variation (97% sequence similarity). Source: [145].

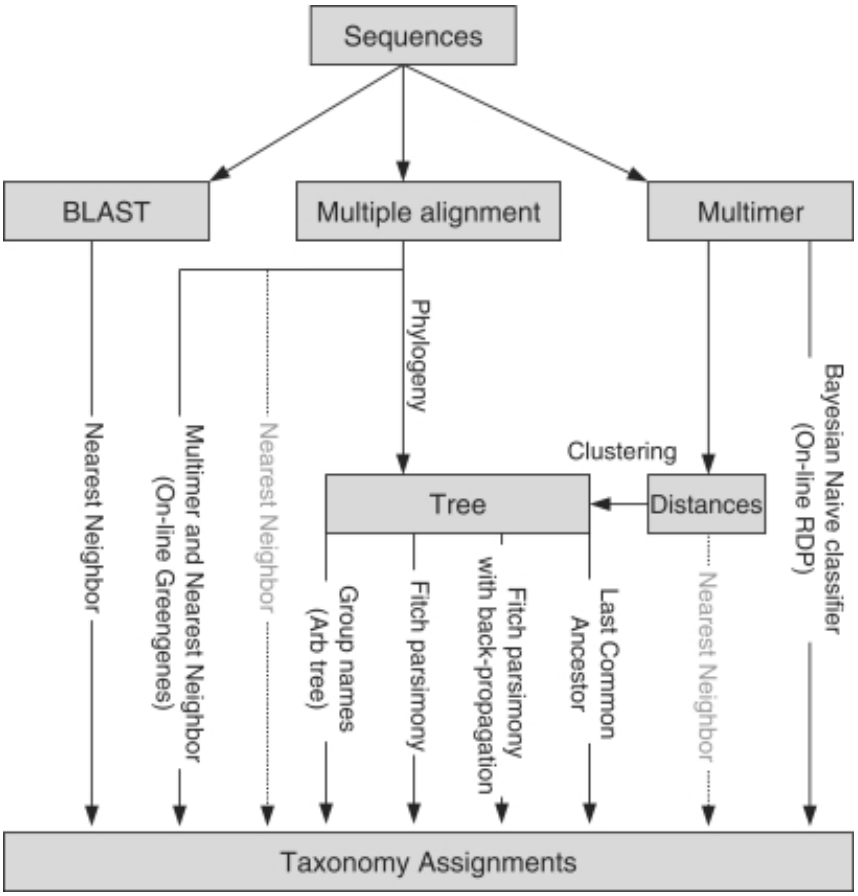


Figure 6.9: Several methods for taxonomy assignment. Source: [154].

6.1.7 Assigning Taxonomy

Taxonomy assignment is viewed as a standard classification problem by machine learning field. The training dataset is comprised of sequences with corresponding taxonomic classification to a particular phylogenetic branch. In machine learning terminology, the input sequences need to be labeled properly - i.e. annotated with taxonomy, and accompanied by a confidence level.

There are many approaches for taxonomy assignment (see figure 6.9 on page 74), but generally they are divided into two general groups: alignment-based and k-mer based [145]. The choice of a particular method has been shown [154] to greatly impact the outcome, i.e. computed expected taxonomic composition. This chapter, in order to be concise, will not thoroughly review all taxonomy assignment methods, rather it will focus on general mechanisms and accompanying caveats.

Taxonomy assignment is not unequivocal for different sets of primers, and often depends on assignment algorithm [154] (see figure 6.10 on page 75). Choosing among many variable regions to amplify thus has a significant effect on downstream analyses. For this particular reason it is often advised not to directly compare results from different studies, as there are plethora of non-biological factors influencing (i.e. shifting) the final outcome. The sole effect of primer selection is shown on figure 6.11 (page 76). This figure compares the number of OTUs per genus between 16S

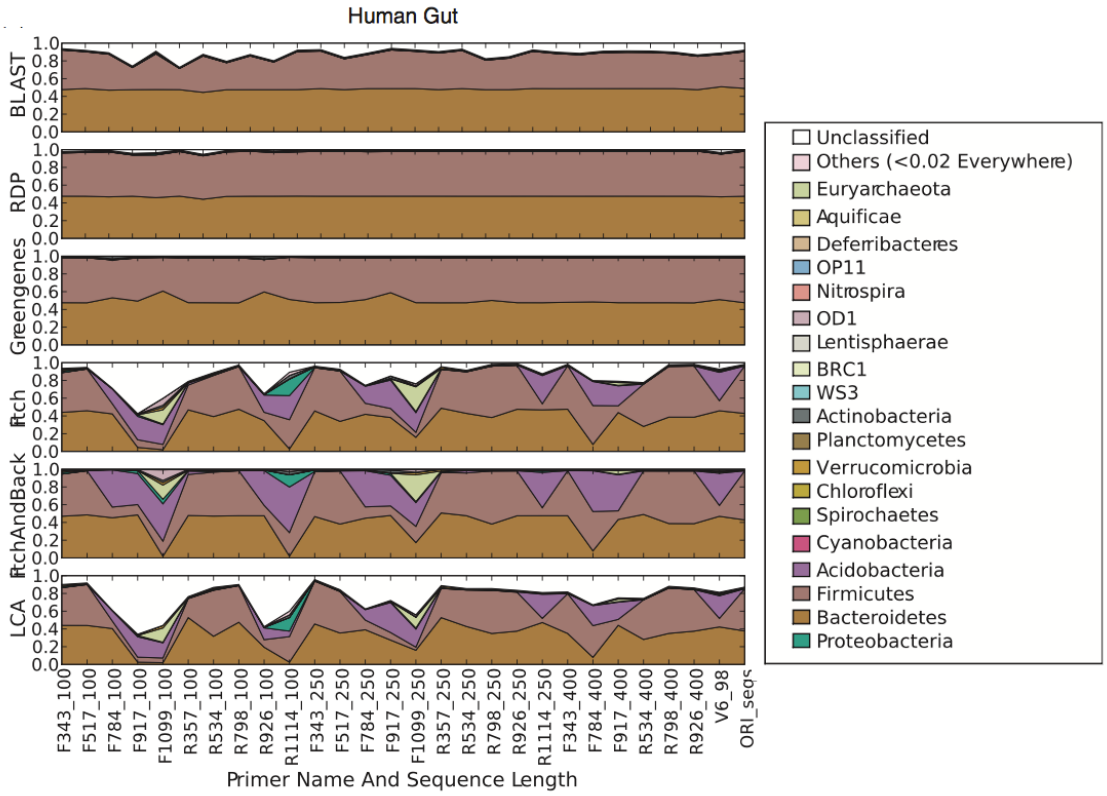


Figure 6.10: Compositions at the phylum level for human gut microbiome using different set of primers and taxonomy assignment methods. Values are reported in relative abundances. Source: [154].

windows: V13 and V35. Majority of the data falls at low values of OTU per genus; however, there are certain bacteria that are overrepresented in a set of primers, as compared to others (ex. overrepresentation of *coprococcus* in V35 primers, and underrepresentation in V13).

One of the most significant classifiers uses Naive-Bayes classification method [52] from 2007 and is called an `RDP` (*Ribosomal Database Project*) classifier. Naive-Bayes approach relies on Bayes's rule briefly reviewed below for genus assignment to a particular sequence. It is worth mentioning that `RDP` classifier works on a set of k-mers - a set of subsequences of a particular sequence.

$$P(genus|sequence) = \frac{\{P(kmer_1|genus) \cdot ... \cdot P(kmer_n|genus)\} \cdot P(genus)}{P(sequence)} \quad (6.3)$$

- $P(genus|sequence)$ - a desired *posterior* conditional probability of obtaining a particular taxonomic rank (*genus* in this example) given the *sequence* as an input,
- $\{P(kmer_1|genus) \cdot ... \cdot P(kmer_n|genus)\}$ - a likelihood of a *sequence* coming from a particular taxonomic rank (*genus* in this example); easily computed from a database,

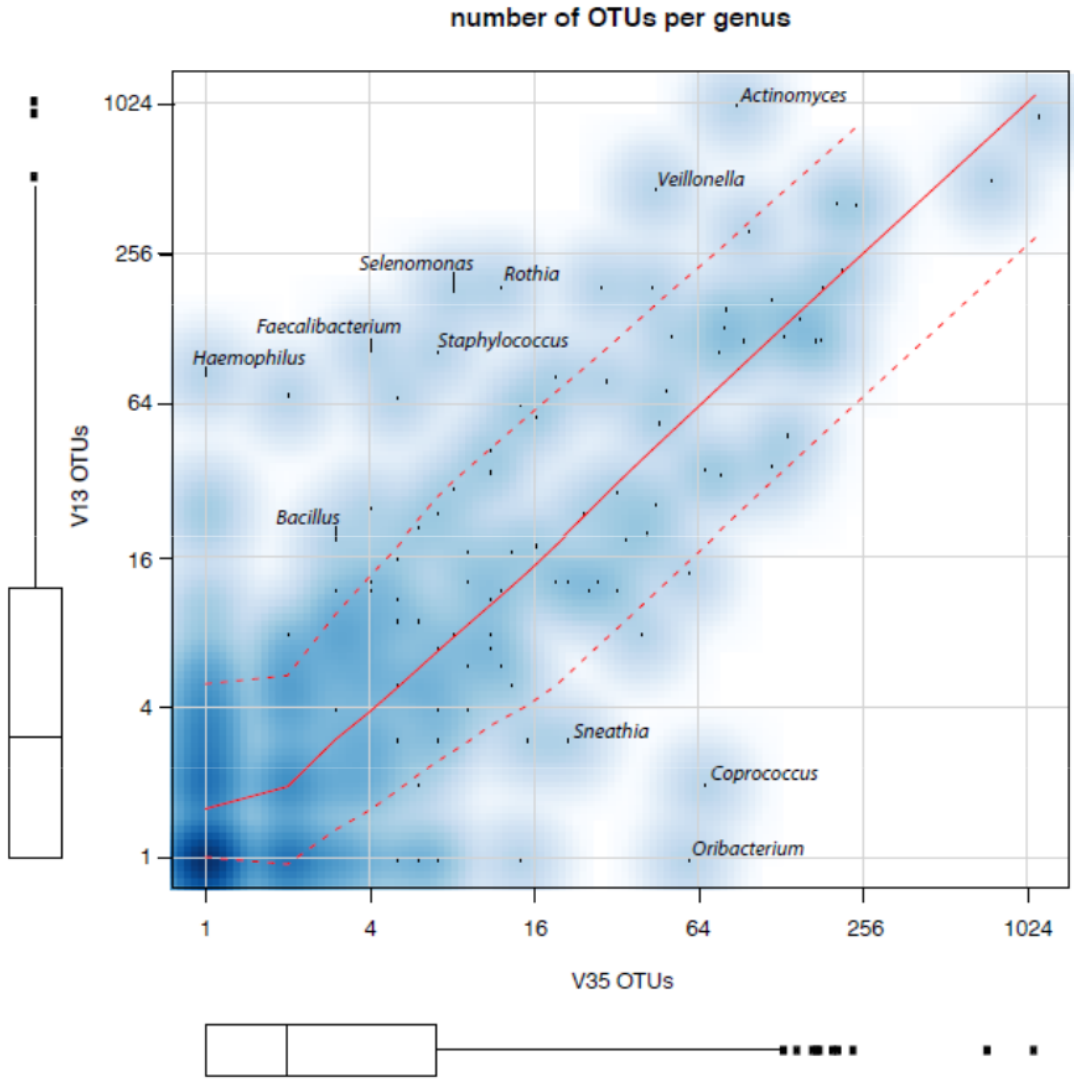


Figure 6.11: Comparison of the number of OTUs per genus between 16S windows. The density of the data is represented as blue gradient. Box plots represent the distribution within each of the windows individually. Red lines are LOESS curves fitted to the data, where dotted lines represent +/- root-mean-square values. Source [77], (*fig. 3*).

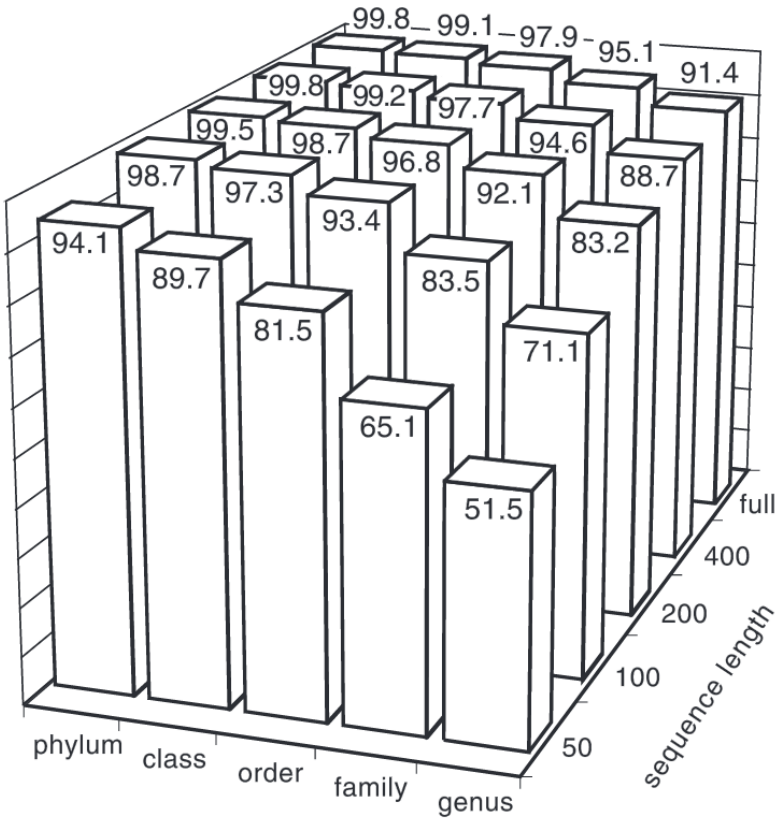


Figure 6.12: Quality of classification depending on query size. Numbers correspond to percentages of tests correctly classified. Source: [52], fig. 1.

- $P(genus)$ - prior probability reflecting how probable is a particular rank (*genus* in this example) to be observed in a particular dataset,

- $P(sequence)$ - total probability (i.e. *evidence*); the probability of observing this *sequence* from the total input data.

The problem put forward is to establish what genus a particular sequence represents, where database files serve as training datasets. Often several simplifications are established. First, researchers often assume that all genera have the same prior probability. This assumption is not always true, as genera often do not form uniform distributions. Second, obtaining the total probability is not necessary to find the best match, thus operating on relative probabilities for all different genera is acceptable. This assumption results in simplified probabilities $P(\text{sequence}) = 1$, and $P(genus) = 1$. One possible way of improving the rate of false positives in obtained datasets is to improve classification methods that do not oversimply abovementioned assumptions. Performance of Naive-Baye's approach depends on the required accuracy level and also on the sequence length (see figure 6.12, page 77).

6.1.8 Diversity analysis

Diversity analysis is concerned with quantification of different types of features in a dataset. For microbial ecology, those features would include different taxonomic ranks or functions. Diversity analysis is done after OTU picking and taxonomy assignment.

6.1.8.1 Rarefaction

Rarefaction in the domain of numerical microbial ecology is the process that assesses OTU richness or other microbe-relevant metrics from each sample to ensure, that sufficient observations have been made in order to regard the samples as reliable representations of the whole community. Datasets obtained from sequencing experiments, like WGS or 16S rRNA gene amplicons are often characterized by unequal number of reads per sample. Comparing downstream results obtained from samples having different depths means that each sample can have different numbers of distinct OTUs that are not of biological origin, but of different sequencing efforts characterized by sequencing depth. Rarefaction allows for comparison of observed number of OTUs, and ensures normalization for the purposes of later downstream analyses. Rarefaction is often represented in a rarefaction plot, where number of distinct features like OTUs are plotted against different sequencing depths obtained from randomly sub-sampling sample sequences. The most common procedure is then to establish the maximum sequencing depth for all samples, for possible exclusion of particularly shallow samples. From rarefaction plots, it is also possible to estimate whether a particular microbial community has been sequenced deeply enough, i.e. obtained sample sequences are representative of the community. This is usually visible directly from rarefaction plots, as the curve representing number of distinct OTUs over sequencing depth reaches a *plateau* (see figure 6.13 on page 79). If a particular sample (or group of samples) on a rarefaction plot is non-converging then more sequences (or samples, for group of samples) are necessary to ensure that all taxa are sequenced. However, read errors have also been shown to drive not converging rarefaction plots, making it increase indefinitely. Rarefaction curve sensitivity is limited and should be regarded as suggestive only [92]. Although the curve may be converging, it is false to assume that deeper sequencing would not yield more species, as certain environments (like gut microbiota) are characterized by the presence of a "long-tail" of low abundance taxa.

6.1.8.2 Alpha diversity

Alpha diversity is used to assess mean richness of number of taxa within a microbial environment [128]. There are many methods to measure alpha diversity through various metrics. The most straightforward method counts observed OTUs, however it ignores features like close-relatedness. Another metric is *phylogenetic diversity* (PD), where from phylogenetic tree the sum of branch lengths are considered [102] to quantify the value of this metric. *Chao1 estimator* is attempting to predict how many species are in sample (or a particular community), given that the input consists of finite number of samples of that community [126]. *Chao1* metric takes into

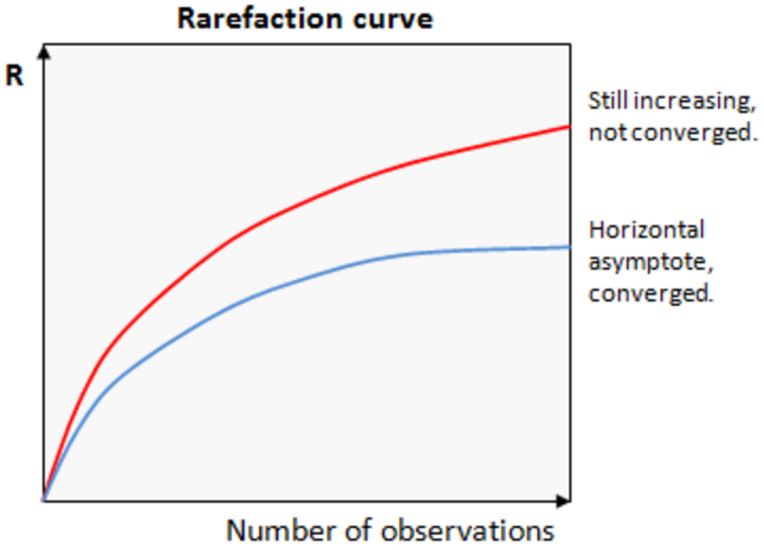


Figure 6.13: Example of rarefaction plots. Red line represents a sample in which increasing number of observations (sequencing depth) would yield more unique features R (like OTUs), while blue line represents a converging line, where community would probably not exhibit more unique features (most common OTUs) when sequencing depth are increased. Source: [92].

consideration so called *singletons* and *doubletons* in the below equation:

$$S_1 = S_{obs} + \frac{\overbrace{F_1^2}^{\text{number of singletons}}}{\underbrace{2F_2}_{\text{number of doubletons}}} \tag{6.4}$$

S_{obs} - the number of observed species,

F_1 - number of singletons (species that are observed once, one observation of OTU of that community),

F_2 - number of doubletons (species that are observed twice in that community).

However, *Chao1* estimator is rarely considered for microbiome data as singletons are often recommended to be discarded. Most singletons are statistically very likely to contain errors [94], and can be attributed to erroneous amplicons (like chimeras - see section 4.7.2 on page 58). However, most of alpha diversity metrics, like phylogenetic diversity, OTU count, Shannon diversity (and many others) tend to be highly correlated with each other [145].

6.1.8.3 Beta diversity

Beta diversity is used often to assess and express diversity between habitats or groups of samples, i.e. healthy vs. disease [128]. Beta diversity is thus used to quantify different communities in a region of interest, and also to measure the degree of differentiation among communities in biological samples [196].

UniFrac (*Unique Fraction*) metric is the most popular metric for beta diversity. In contrast to other used metrics, like Bray-Curtis, chi-square or euclidean, UniFrac uses phylogenetic (close-relatedness) information in order to provide more sensitive comparisons of communities. UniFrac describes the percentage of observed branch length that is unique to either sample (see figure 6.14 on page 81). For the purpose of brevity, the whole methodology would not be discussed here. Exact implementation and specificity is discussed in author's paper [156]. UniFrac uses analogical phylogenetic approach that is performed for phylogenetic diversity (see previous alpha diversity subsection).

UniFrac performs calculations between every pair of samples, thus resulting in distance metrics that can be further analyzed and visually explored through Principal Coordinates Analysis (PCoA) plots or hierarchical clustering methods (see figure 6.15 on page 81, and PCoA section 6.2 on page 80). UniFrac metric is comprised of two sub-metrics: weighted and unweighted. Weighted UniFrac takes into account relative abundance of microbes across compared samples; phylogenetic branch lengths describing phylogenetic distance (close-relatedness) are weighted additionally by its abundance observed across considered samples. This approach emphasizes the dominant microbes. This method is more likely to reveal community differences that are caused by relative abundances of observed taxons. Unweighted UniFrac is a more qualitative method that doesn't weight branch lengths by observed microbial abundances. This method is more appropriate when communities are differentiated by the capability to host certain microbes (ex. environments different in nutrients or different temperatures), where abundance can hinder to reveal significant patterns of variation. In research, both metrics are computed and compared to infer significant observations. UniFrac metric is described by a number between $< 0, 1 >$ (see figure 6.14 on page 81), where 0 represents identical communities, and 1 represents unrelated communities.

6.1.8.4 Gamma diversity

Gamma diversity describes total microbe diversity within a particular landscape. Gamma diversity is determined by alpha and beta diversities. Alpha diversity is a "local" component, while beta diversity accounts for differences between habitats. It is not commonly used in microbiome studies, but rather environmental ecology, like amazon jungle diversity studies.

6.2 Visualizing Microbiome Diversity

Microbial diversity is often visualized using ordination techniques on multidimensional matrices, such as those computed for beta diversity (UniFrac). There are variety of techniques: Principal Coordinates Analysis (PCoA), Principal Components Analysis (PCA) or Non-metric Multidimensional Scaling (NMDS). For the purposes of brevity only PCoA method is discussed in more detail, while others are briefly summarized in subsection 6.2.2 (page 83) "*comparing ordination techniques*". All ordination techniques produce a set of uncorrelated (orthogonal) axes in order to visualize the variability in the dataset [72]. Each axis is accompanied

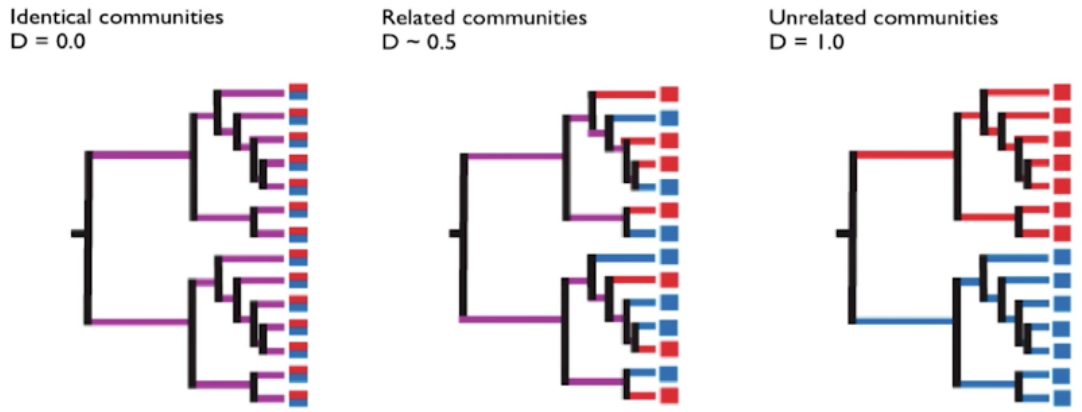


Figure 6.14: Illustration of UniFrac metric in assessment of beta diversity for three microbial communities and their appropriate UniFrac score. Identical communities are represented by UniFrac value of 0, related communities are represented by numbers within range $(0, 1)$, and unrelated communities are described by UniFrac value of 1. Source: [145].

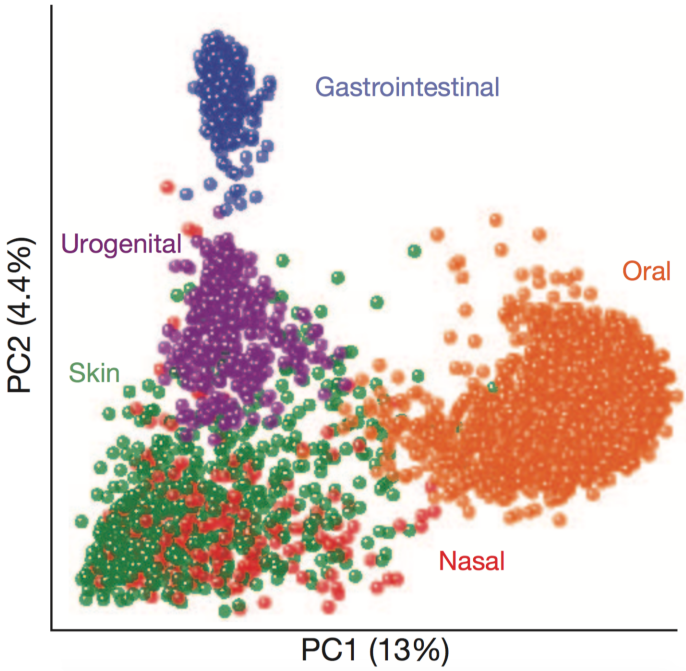


Figure 6.15: Example beta-diversity representation through ordination plot. PCoA plot showing variations among samples from Human Microbiome Project. Clustering is visible by body area - oral, gastrointestinal, skin and urogenital clusters are clearly visible. Source: [78], fig 1.c.

by a value ("*eigenvalue*"), that reveals the amount of variation captured from the dataset. Successful ordination methods will capture all or most variability in two or three axes that could be easily explored on a two- or three-dimensional plot. Each object explored, after applying one of the ordination methods, is accompanied by a *score* that provides coordinates in the ordination plot [72].

6.2.1 Principal Coordinates Analysis (PCoA)

Principal Coordinates Analysis (PCoA), sometimes referred to as *metric multidimensional scaling* is a technique for presentation of inter-object (dis)similarity that forms multidimensional array defined by the number of samples and features observed (like OTUs), in a low dimensional Euclidean space [71]. In other words it helps to extract and visualize a set of few highly-informative components (axes) of variation from complex and often multidimensional data [174]. PCoA thus allows to position studied objects (like samples) in a space of reduced dimensionality, at the same time preserving their distance in a chosen (by the user) metric [171]. Below is a brief mathematical description of PCoA method.

The initial input is a distance matrix $\mathbf{D} = [D_{hi}]$, that is then transformed into a new matrix $\mathbf{A}$ [171] with elements a_{hi}:

$$a_{hi} = -\frac{1}{2}{D_{hi}}^2 \tag{6.5}$$

Next, the $\mathbf{A}$ matrix is centered with row $\bar{a}_h$ and column $\bar{a}_i$ means of the matrix element a_{hi}; $\bar{a}$ is the mean of all a_{hi} values:

$$\delta_{hi} = a_{hi} - \bar{a}_h - \bar{a}_i + \bar{a} \tag{6.6}$$

Centering positions at the origin of the new axes at the center position (centroid) of the scattered objects, without affecting distances between studied objects. This results in a new matrix $\mathbf{\Delta_1} = [\delta_{hi}]$. Finally eigenvalues λ_k and eigenvectors $\mathbf{u}'_k, \mathbf{u}_k$ are computed (through standard algebraic calculations). Eigenvectors are scaled to lengths equal to square roots of their respective eigenvalues:

$$\sqrt{\mathbf{u}'_k \mathbf{u}_k} = \sqrt{\lambda_k} \tag{6.7}$$

Generally in order to represent n points in Euclidean space $(n - 1)$ axes are required. Due to centering of the matrix (equation 6.6) it always has at least one zero eigenvalue, although there may be more than one zero eigenvalue. In that case the matrix is *degenerate*, and studied objects (ex. samples) can be represented by fewer number of axes. Finally, after the eigenvectors are computed, they can be stored in a matrix (see figure 6.17 on page 84). If the eigenvectors are columns filled with eigenvalues for that eigenvector, then the rows of the constructed table (each attributed to a particular object/sample) are the coordinates in the space of principal coordinates [171]. Thus plotting two or three principal coordinates allows for visualization in space-reduced ordination.

It has been shown [119] that relationships of distance are conserved in the full-dimensional principal coordinate space. This proof will not be discussed here, but can be found in sources [119], [171] (and others).

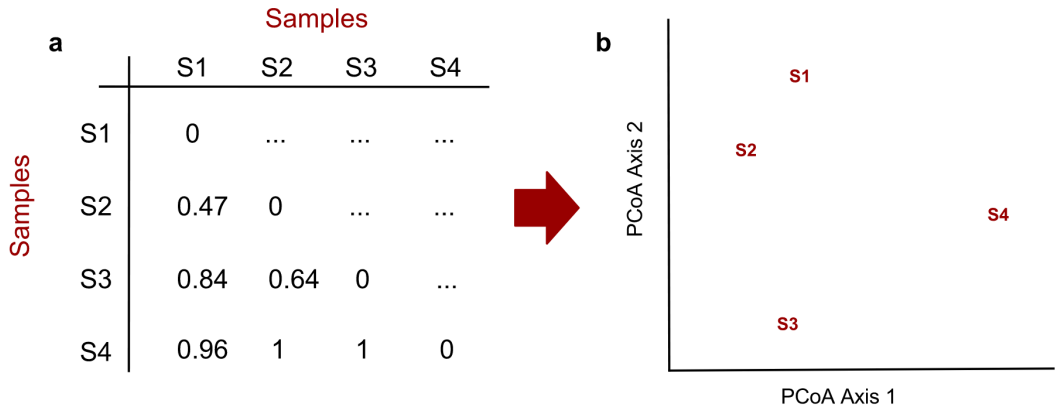


Figure 6.16: Principal Coordinates Analysis (PCoA) ordination for example distance (dissimilarity) matrix. Dissimilarity is expresses as a number from a range of $< 0, 1 >$, where 0 describes identical samples, 1 complete dissimilarity between compared samples. After PCoA analysis samples are plotted on PCoA axes. Samples ordinated closer together have relatively small dissimilarity, compared to samples further apart. Source: [72].

6.2.2 Comparison of ordination techniques

Principal Components Analysis (PCA) is used very often outside the field of ecology, however very rarely in microbiome data. It is a method equivalent to calculation of euclidean distances on obtained distance metric matrix. The main limit of this methodology is that studied matrices need to be characterized by greater number of samples than features (i.e. OTUs), and that euclidean metric is rather unsuitable for microbial ecology data represented by close-relatedness of bacteria.

Non-metric Multidimensional Scaling (NMDS) is a method where the primary importance is to represent objects (samples) in a required number of dimensions, and preserving distances among objects is not a primary concern. It shares similarity with PCoA as both methods are not restricted to a particular distance metric (like Euclidean), however, it can also handle missing (empty) distances in matrices. For this reason it is a primary application in laboratory assays [171] where direct observations of some results are often missing pairwise distances. Non-metric Multidimensional Scaling (NMDS) axes, as opposed to Principal Components Analysis (PCA) and PCoA, do not maximize variability of the axis - axes in NMDS are thus arbitrary.

	Eigenvalues			
	λ_1	λ_2	$\cdots$	λ_c
Objects	**Eigenvectors**			
$\mathbf{x}_1$	u_{11}	u_{12}	$\cdots$	u_{1c}
$\mathbf{x}_2$	u_{21}	u_{22}	$\cdots$	u_{2c}
$\mathbf{x}_h$	u_{h1}	u_{h2}	$\cdots$	u_{hc}
$\mathbf{x}_i$	u_{i1}	u_{i2}	$\cdots$	u_{ic}
$\mathbf{x}_n$	u_{n1}	u_{n2}	$\cdots$	u_{nc}
Length: $\sqrt{\sum_i u_{ik}^2}$ =	$\sqrt{\lambda_1}$	$\sqrt{\lambda_2}$	$\cdots$	$\sqrt{\lambda_c}$
Centroid: $\bar{u}_k$ =	0	0	$\cdots$	0

Figure 6.17: Example PCoA table with scaled eigenvectors. Objects studied (ex. samples), represented in rows after computations described in equations 6.5 - 6.7, have accompanying principal coordinates attributed to subsequent eigenvectors. Source: [171] (tab. 9.7).

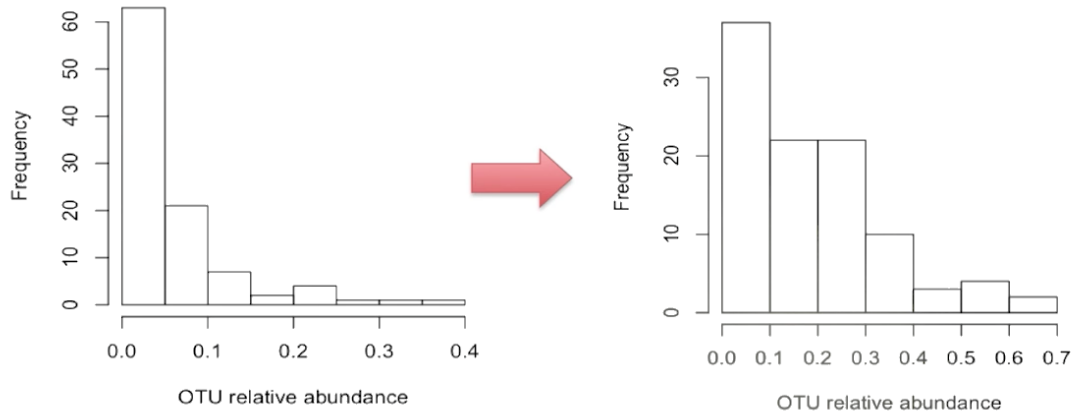


Figure 6.18: Example of typical OTU distribution within hypothetical population. The trend follows negative binomial distribution (left plot), in which there are relatively few individuals with big relative abundance of a particular bacteria, and the majority of the population has small abundance or no observance of this bacteria. After applying square root transformation the distribution is more similar to normal, however this transformation is not very efficient. Source [145].

6.3 Statistical testing of microbiome data

This section briefly discusses some statistical caveats and methods when dealing with microbiome data. Species distribution in the microbiota are often zero-inflated, which means there are certain subjects that might lack certain bacteria present in other subjects. Thus, assuming normal distribution for microbiome data might not be an appropriate approximation. Generally species (or more broad taxonomic ranks like genera) distribution follows negative binomial distribution, also referred to as Pascal distribution (see figure 6.18).

Oftentimes researchers transform the structure of the input data with *arcsin-square root transform* or *logit transform* for further downstream analyses. The aim of *Arcsin* transform, sometimes referred to as angular transformation, "pulls out" and stretches the ends of the distribution over a broader range of values (see figure 6.19, page 86). This transform spreads smaller values more than the middle-range, thus putting more emphasis on low abundance taxa. Unlike logit transform, it is able to process zero values [207]. Due to the fact that *arcsin* transform puts more emphasis on low and high abundance taxa, it is often argued that *square-root transform* might have more biological rationale for microbial studies [145], as it only emphasizes small values (small-abundance taxa) (see figure 6.18, page 85).

Non-parametric tests in the domain in microbial ecology are sometimes preferred, as they do not relay on null distributions, like normal distribution, but on ranks of values. However, using non-parametric tests doesn't allow for controlling for confounding factors that are often valid characteristics for microbiome data. Most importantly non-parametric tests have generally lower statistical power.

Parametric testing with linear models is sometimes considered for microbial ecology data, as not the original data, but the residuals of the data need to be normally

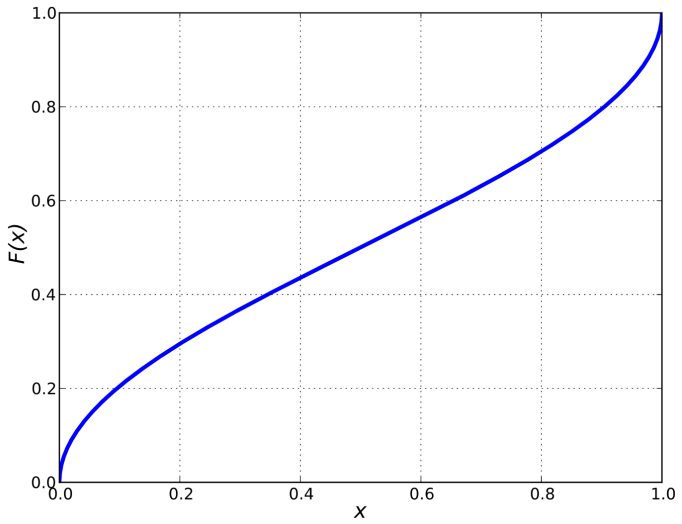


Figure 6.19: Cumulative distribution of `Arcsine` function. Source: [209].

distributed. Linear models allow to account for multiple confounding factors. Utilizing parametric tests should always come with caution, as assumptions like normality or normality of the residuals are often violated. One possible approach to account for that is first to test whether residuals are normally distributed, which is often tested with Kolmogorov–Smirnov test. Oftentimes after performing this test one must reject the null hypothesis - that residuals come from normal distribution. For this reason generalized linear models are used. Those models don't relay on normal distribution, and previously discussed *negative binomial distribution* are often utilized. *Generalized linear models* don't operate on relative values, like relative abundances, thus operating on absolute values (like absolute counts of microbes) is required.

Microbiome also exhibits *heteroscedasticity* - the problem of unequal variances. With increasing number of observations (i.e. reads of a particular sample), the uncertainty (i.e. variability) increases. For microbial studies this means that bacteria with higher abundances tend to have greater variances with relation to microbes with smaller abundances [145], [161]. For this reason oftentimes linear model assumption of uniform distribution is violated, as microbiome data often exhibits lack of *homoscedasticity*, thus generalized linear models are used.

Multiple test corrections are necessary when performing multiple comparisons. Having performed 100 test with $\alpha = 0.05$ one can expect that 5 results might appear significant as *p-values* are normally distributed. For genetic testing the *Bonferroni correction* is often performed. For n multiple tests, Bonferroni correction introduces adjusted alpha threshold value $\alpha_{adj} = \frac{\alpha}{n}$ for a result to be considered as statistically significant. Other methods like False Discovery Rate (FDR) are also used [145].

Finally, it is often recommended to omit (i.e. exclude) very rare taxonomic ranks, i.e. small abundances, as there may be a mere lack of data to ensure satisfactory statistical power in order to infer meaningful results.

6.4 Biomarker discovery

Biomarker is generally a measurable biological property that can be indicative of some phenomena, such as infection, disease or environmentally-caused disruption. For metagenomics, there are two types of potential biomarkers: those relating to taxonomic composition and functional biomarkers; 16S rRNA gene survey only involves the former. Taxonomic composition includes specific genera or species, although this requires deep metagenomic sequencing to provide necessary accuracy for taxonomic classifiers. For 16S rRNA gene surveys, the quality of classification depends of the query size (see figure 6.12 on page 77), thus a study design might limit the potential of bio-marker discovery.

Second biomarker includes potential functional characteristics, like genes, proteins or metabolites specific to a single organism or particular for a specific community of organisms. While deep metagenomic study can provide those types of information, 16S rRNA survey is generally considered only for taxonomic and phylogenetic investigations; there are however machine learning approaches for conducting general inferences about community functions (see `PICRUSt` software on page 90).

Biomarker discovery can be described as a multivariate approach to removing non-informative or redundant gene sequences, and focusing on features differentiating between two conditions. A good biomarker has low variance, for example consistent taxa abundance among each group. Ideally its abundance across samples follows a normal distribution, where means between groups of interest are significantly apart from each other. Biomarker can be summarized as a set of genes/species that are predictive of some clinical manifestation on host or particular state of the analysed environment.

6.5 Mock communities

Mock community is essentially a DNA-free water that is spiked with DNA from multiple known taxa of lab-cultured bacteria. Knowing compositions of these bacteria allows then to use *mock communities* as a positive control in microbial analyses. It has been shown [7] that using different parameters in downstream analyses (filtering of sequences, OTU picking, etc.) can lead potentially to large overestimation of microbial diversity. Thus, backing created analysis pipelines with appropriate positive controls is a basic scientific approach that has become a part of Standard Operating Procedure (SOP).

Mock communities are very often offered without charge for laboratories and public institutions and can be ordered from the Human Microbiome Project (HMP) website. The best practice involves sequencing samples of interest with mock community. Ideally a study would include two identical samples coming from the same mock community, so as it would be possible to compare various statistics among them (alpha- and beta-diversity, microbial compositions) in order to observe any potential false positive differences arising due to various sequencing errors (see section 4.7 on page 57).

Sequences of known mock communities are also publicly available through repository called *"mockrobiota"* [168]. This repository currently contains 26 sets of different

mock communities mostly obtained through 16S rRNA survey, but `ITS` or `18S` gene sequences are also available. Although it allows for bench-marking for downstream analyses, this approach would not help to asses sequencing-specific errors and biases involved with a particular run. Mock communities might also be associated with communities created *"in silico"*, by extracting raw sequences from microbial databases.

Mock communities revealed that certain algorithmic approaches may be context-dependent, thus there may be no one most-appropriate computational method for a specific aim [7]. Finally, it is worth stating that mock communities might not reflect the complexities found in environmental samples, like gut microbiome that is characterized by a "long-tail" of low abundance microbes, thus mock communities may not well generalize to environmental samples [39]. Despite those problems, inclusion of mock communities into a NGS-based microbial study remains a recommended practice in order to ensure greater accuracy for taxonomic assignment.

Chapter 7

Software Packages

This chapter only briefly describes some of the software packages used in this study, there is no necessity for detailed discussions. Each software package has rich and detailed documentation: manuals, protocols, accompanying scientific papers and usually a user community forum.

7.1 QIIME

Quantitative Insights Into Microbial Ecology (QIIME) is an open-source bioinformatic platform for microbiome analysis from raw DNA sequencing data, through demultiplexing (if necessary), quality filtering, various OTU picking methods, taxonomy assignment, phylogenetic reconstructions to diversity analyses and visualizations [174]. Current version 1.9.1 will be officially supported until the end of 2017, as version 2.0 is still in development stage. QIIME has received a grant from National Science foundation [107] to prepare a platform that comprises of various techniques and statistical analyses, often performed for microbiome data, in one package that would ensure reproducibility of conducted research.

7.2 Mothur

Mothur is an open-source software package for microbiome data processing, analysis and visualization [54]. Mothur software package is constantly being upgraded and updated (at the time of this thesis, the latest version was released in March 2017). Mothur offers similar functionality as QIIME; it offers algorithms to process raw reads and later downstream statistical analyses. While QIIME is a more scientific community-driven project, where each algorithm is usually developed by different researchers, mothur is developed by a relatively small scientific group at the Department of Microbiology & Immunology at the University of Michigan. Mothur also allows users to create additional analysis plugins that might comprise of statistical tests or novel algorithmic approaches.

7.3 PICRUSt

7.3.1 PICRUSt overview

Marker gene studies, such as 16S rRNA gene studies (see paragraph: 3.1 on page 38), focus particularly on universal genes, and it is impossible to directly identify metabolic capabilities of the microorganisms studied. Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) is a tool that provides 80-85% accuracy for predicting full metagenomic repertuar in the microbiome. Note that because of this range the authors themselves [47] (Dan Knights, Ph.D.) suggest to treat obtained results as "suggestive only" [145]. The main advantage of this approach is that relatively little sequencing depth is needed to characterize the diversity and phylogeny of a sample. Deep Whole Genome Sequencing (WGS) still remains costly and at times prohibitively expensive, yet it is able to provide characterization of rare organisms and genes. PICRUSt functionality, described below in more detail, takes advantage of the fact that phylogeny and biomolecular functions are strongly correlated [47]. PICRUSt creators state that phylogenetic trees based on 16S survey closely resemble clusters obtained based on shared gene content [47].

PICRUSt is a software technique that uses evolutionary modeling to predict metagenomes from 16S survey data and a reference genome database. Its processing pipeline comprises of two main workflows: 1) gene content inference, and 2) metagenome inference, that are described in paragraphs below. The user provies an as input OTU table with associated GreenGenes database identifiers. The current PICRUSt uses 13.5 GreenGenes microbial database from May 2013.

It is worth noticing that computed outputs are directly comparable to metagenome profiles generated through deep, whole-genome sequencing studies, through pipelines utilizing software packages such as HMP Unified Metabolic Analysis Network (HUMAnN) or MG-RAST. Additionally PICRUSt is able to perform estimates of the contributions made by each OTU to a given gene function, which is not a straightforward procedure for data generated through WGS. PICRUSt key features: [47]:

- recaptures key findings from the Human Microbiome Project (HMP) and predicts metagenomes across a broad range of host-associated and environmental samples,
- tested on human gut, soil, sea (hyper-diverse and underexplored Guerrero Negro microbial mat), and mammalian gut samples,
- in the best case, correlations between inferred and metagenomically measured gene content approach 90%; average is approximately 80%,
- outperforms the metagenomes measured at particularly shallow sampling depths,

7.3.2 Gene Content Inference

Although phylogenetics is a rapidly developing field of science, many organisms still have not been studied. PICRUSt aims at estimating the properties of ancestral

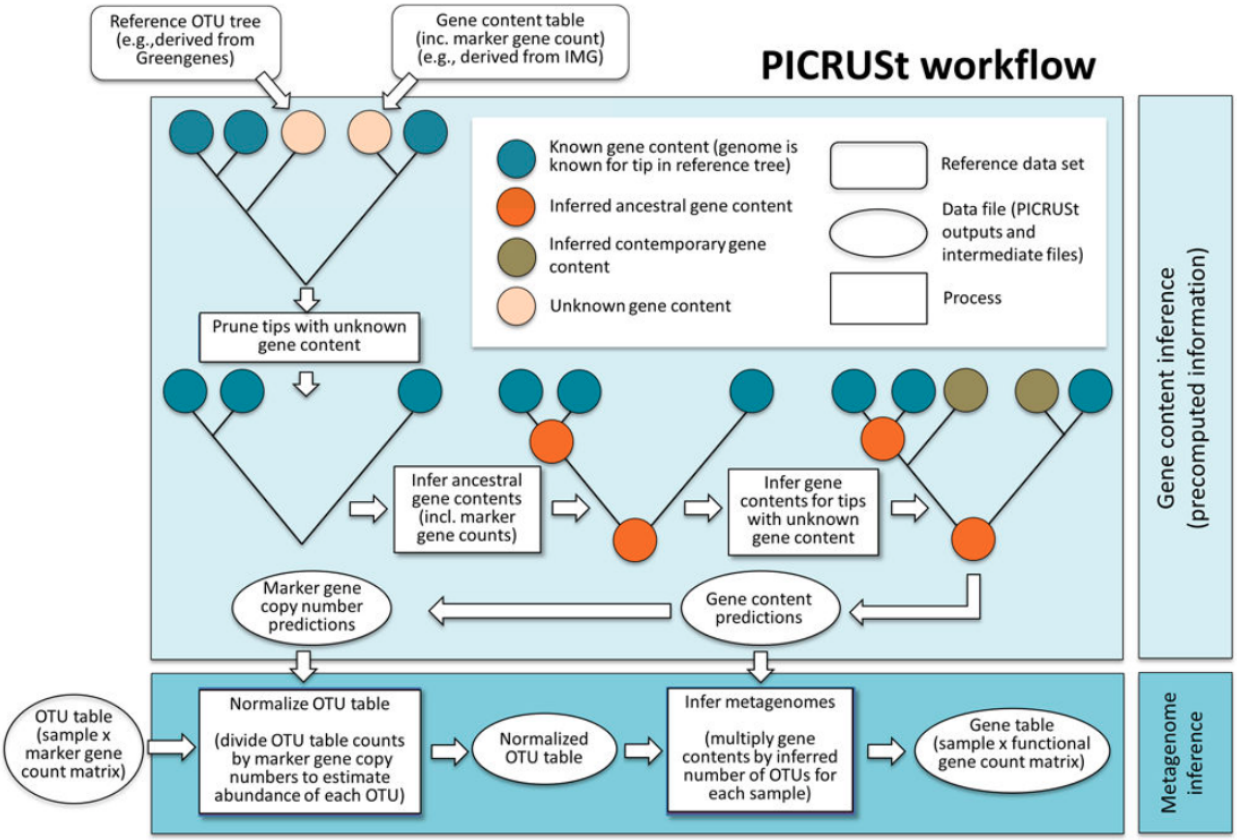


Figure 7.1: PICRUSt workflow comprises of two pipelines described in subsequent subparagraphs (7.3.2 and 7.3.3). One is gene content inference (top of the figure), and the other metagenome inference. Figure is directly reproduced from [47], fig.1. p.17.

organisms from living relatives. The main reason for developing Ancestral State Reconstruction (ASR) algorithmic approach is that microbial genomes of related bacteria tend to be more conserved compared with those more distantly related on the tree of life. From obtained phylogenetic tree, it is possible to infer other gene families present, i.e. those that have been studies in more detail.

The main source of uncertainty is the dynamics at which particular bacterial genes can change over evolutionary time. For this reason PICRUSt developers report that each prediction is accompanied by a 95% confidence interval that reflects this uncertainty [46]. Gene content is precomputed for each organism identified in a studied sample from a reference phylogenetic tree. It is important that this step is independent of community sample type. Gene Content Inference is comprised of several steps: tree pruning and formatting, Ancestral State Reconstruction (ASR), and predicting traits. Corresponding first step is required for ASR algorithms, where identification numbers must match between matrix of character states and the phylogeny. In other words, it is required to reduce the trait table, such that it only contains organisms shared between tree and trait table. ASR algorithm is explained in section 7.3.4 (page 93). This algorithm uses a common ancestor to predict gene content of a particular taxon that has not been thoroughly studied, yet whose close relative gene content is present in databases. The last step - predicting traits - is dependent on the position in the phylogenetic tree of a particular organism. Therefore the estimates are weighted based on phylogenetic distance. Gene Content Inference is a workflow for predicting unknown gene content directly from OTU tables with known content (taxa) and a phylogenetic tree (in Newick format - see paragraph 5.4 at page 63) relating OTUs with known or unknown gene content.

7.3.3 Metagenome inference

The second step is the metagenome inference that takes into account relative abundances of 16S rRNA genes in one or more microbial communities. Those predictions are corrected for expected 16S rRNA gene copy number in order to generate the expected abundances of gene families in the entire community [47]. A process called "*copy number normalization*" is the first step in metagenome inference (see figure 7.1 at page 91). Since 16S rRNA gene operons can vary from 1 to 15 copies in bacteria, the observed relative abundances are likely to vary from true organismal abundances. 16S rRNA copy number varies greatly among different bacteria and archea. Normalization processes divide OTU abundance level for each sample by its predicted 16S rRNA copy number. The goal is to obtain OTU table that corresponds to the relative abundances of organisms, rather than relative abundances of the 16S rRNA gene. Normalization process accomplishes just that, utilizing precomputed gene copies for each microbial family. Finally, simple multiplication of vector containing gene counts for each OTU by the abundance of that OTU in each of the sample produces a predicted metagenome - an annotated table of predicted gene family counts for each samples. Gene families can be orthologous groups (like KEGG orthologs - KOs), as well as Clusters of Orthologous Groups (COGs) or Pfams (database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models). It is important that

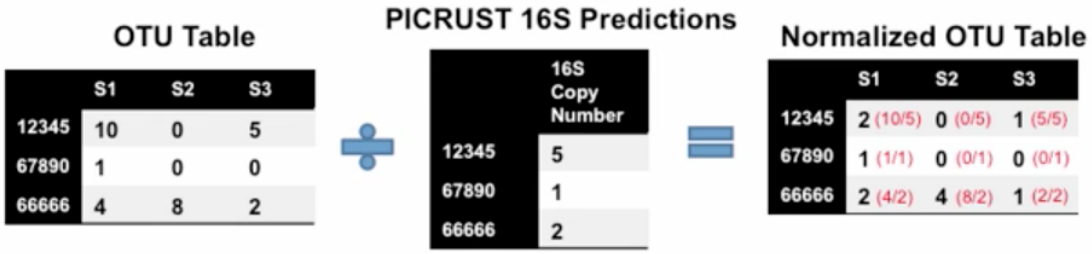


Figure 7.2: OTU table normalization. Source [150].

OTUs were picked by *closed-reference* method (see: paragraph 6.1.6 *OTU picking approaches*, page 70), discarding novel and uncharacterized (in the databases) bacterial taxonomic rank.

7.3.4 Ancestral State Reconstruction Algorithm (ASR)

The rationale of ASR algorithms is straightforward: since the gene content is mostly conserved among highly related microorganisms it is thus possible to infer gene content present in the common ancestor of a particular organism. Hence, it is possible to predict gene content in an observed, but not yet thoroughly studied organisms.

ASR algorithm fits evolutionary models to the distribution of traits observed in a living organism using probabilistic criteria, such as Bayesian posterior probability or maximum likelihood. ASR also extends its prediction to extant (in addition to ancestral) organisms. By default the count of each gene family is regarded as a continuous evolutionary character evolving under Brownian Motion model [47]. This is a simpler model of evolution that is more easily computable, compared to discrete character model. For more discussion and information refer to [46], [47].

Model produces estimates of uncertainty for each ancestral state that reflect how fast a particular gene family is being changed in terms of copy number over evolutionary time. Model therefore summarizes uncertainties coming from mechanisms of microbial gene plasticity, ex. horizontal gene transfer.

7.3.5 PICRUSt limitations

Since PICRUSt output is of a predictive nature and there are requirements with relation to the nature of the input - several limitations arise. Firstly, the input OTU table has to be picked in a closed-reference manner (see OTU picking methods, section 6.1.6 page 70) with relation to a particular database and its version - GreenGenes 13_5. Since *de novo* OTUs have no accompanying database identification number, they can't be analyzed in terms of predictive metabolic potential. Another limitation is that gene contents often differ between bacterial strains, for which 16S survey has no possible access. Thus microbes from different strains, exhibiting different gene contents are presented by identical 16S rRNA gene sequences.

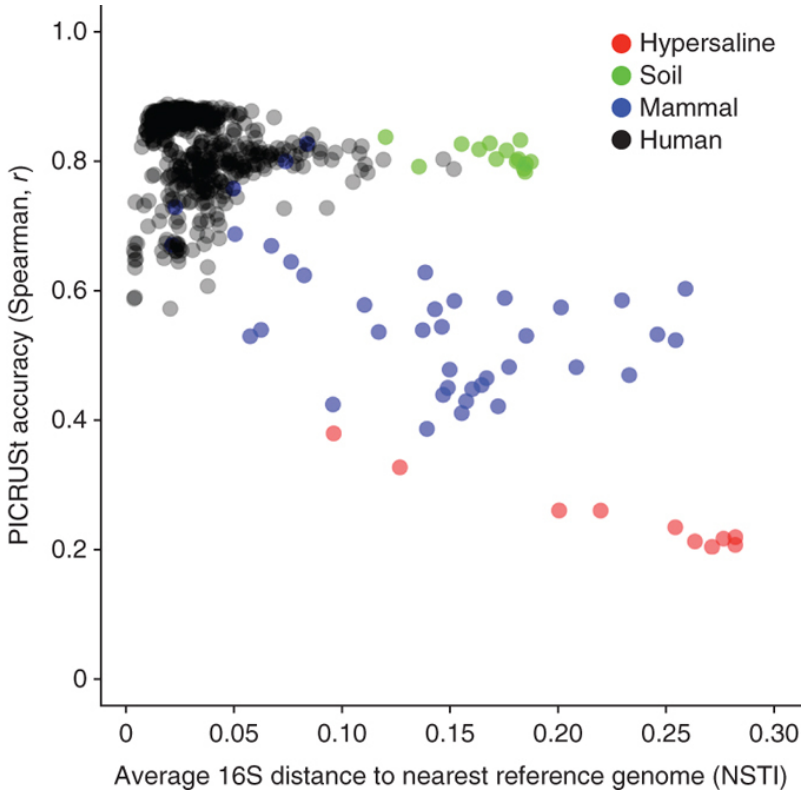


Figure 7.3: PICRUSt accuracy for different environments. Spearman correlation was computed for reference data from functional profiling from WGS. NSTI scores represent distance to a nearest reference genome. Source [47].

This underlines to some degree the ability to recognize and categorize their functional potential that might be responsible for toxicity and pathogenesis [47]. Because of the homology of certain genes (genes coming from different species, like human and bacteria), PICRUSt output might correlate certain microbial functions to human functions. They should be either discarded or properly examined in order to infer proper biological meaning. Last, but not least, PICRUSt accuracy is limited for particular environments (see figure 7.3 on page 94); human gut samples are relatively well characterized (expected accuracy lay between 60% to 90%), while other environments should not even be considered for meaningful metagenome predictions (like hypersaline samples). For this reason the authors themselves suggest to treat obtained results as *"suggestive only"*, even for well-characterized human gut microbiome samples.

7.4 STAMP

Statistical Analysis of Metagenomic Profiles (STAMP) is a software package for standardizing and ensuring the reproducibility of statistical analyses in microbial profiles [169]. It allows for performing two-sample, two-group or multiple group statistical tests with inclusion of study-specific metadata. For taxonomic profiles, it allows comparing bacteria at different taxonomic ranks separately, i.e. performing

separate statistical tests for genera, family, order, etc. This open-source software aims at promoting *best practices* in selection of statistical techniques appropriate for microbial ecology.

STAMP has a rich documentation in which all features are discussed. For this reason only a brief overview of statistical methods available is shown. STAMP performs various parametric and non-parametric statistical tests for:

- Multiple-group tests: `ANOVA` (parametric), Kurskal-Wallis H-test (non-parametric)
- Two-group tests: t-test of equal variance, Welch's t-test for unequal variances and White's non-parametric t-test
- Two-sample tests: Fisher's exact test, chi-square, bootstrap and many others.

Additionally, it allows for controlling of confidence intervals, such as: Games-Howell, Tukey-Kramer, Scheffe (multiple-groups), Welch's and t-test inverted confidence intervals (two-groups), and Newcombe-Wilson, Asymptotic and others (two-samples). More importantly, STAMP promotes correction for multiple comparisons that are key feature for many genomic studies: Storey's False Discovery Rate (FDR), Bonferroni, Sidak and Benjamini-Hochberg FDR.

7.5 UPARSE

`UPARSE` is a software package and an algorithm created by an independent researcher, Robert C. Edgar [98], [99]. `UPARSE` is available in a broader package called `USEARCH` that comprises of more algorithms like *unoise2* [97] (for Zero-Radius Operational Taxonomic Unit (zOTU) picking - see section 6.1.5.1 on page 69), or *uncross* [96] algorithm implementation for removing cross-talk. `UPARSE` allows for many steps discussed in section 6.1 "*16S rRNA survey analysis pipeline*" (page 65). `UPARSE` allows for most steps in read preparation: performing paired-end merging, filtering reads by quality scores, discarding singletons, relabeling and identifying sequences, and trimming sequence lengths to obtain a fixed value (this is important for unpaired sequences). `UPARSE` allows for further steps like read dereplication, clustering sequences through `UPARSE-OTU` algorithm (see figure 7.4, page 96) [99], chimera filtering and, finally, preparing OTU tables.

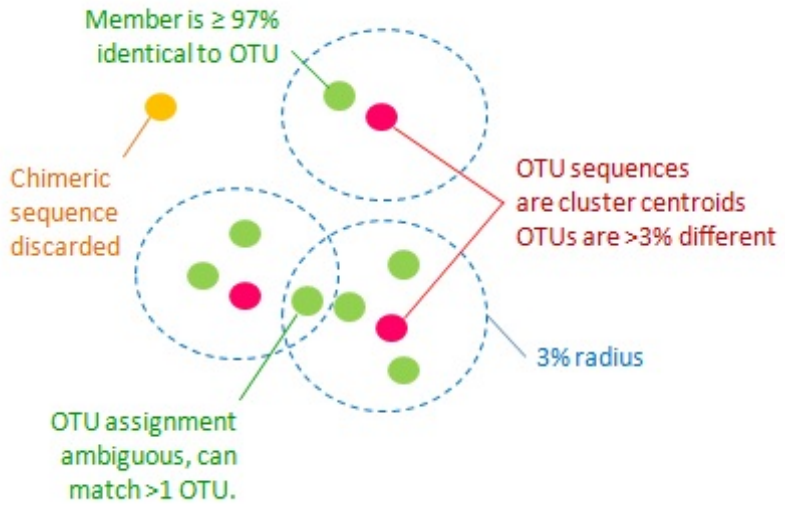


Figure 7.4: UPARSE-OTU algorithm: principles of clustering sequences together. Source: [99].

Part II

Materials and Methods

Chapter 8

Online repository

This study is accompanied by a corresponding online repository containing all analysis scripts, anonymised subjects' metadata, protocols and manuals in order to ensure the reproducibility of the research. As this and later work (shotgun sequencing of fecal microbiota of SLE patients and control groups) will later become part of scientific publication, the raw sequences will be thus be published and deposited in National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) with appropriate annotations. Online repository will be updated with appropriate SRA sequence accession numbers. For the purposes of brevity and clarity, all analysis scripts can not be reproduced here without greatly increasing the number of pages. Instead, analysis pipeline containing several analysis scripts has been depicted as an algorithm block in chapter 15.2 *Analysis pipeline for 16S rRNAamplicon sequencing* - (page 118) and all analysis scripts are available to explore in online repository. Some of patients' metadata has been shown on histograms in chapter 11 (page 102), however full anonymised study subjects' metadata is also available in the online repository.

The repository is available on GitHub platform under link:

`https://github.com/vaxherra/MicrobiomeSLE`

GitHub platform is a web-based version control repository and internet hosting service that is widely used for open-source software projects, sharing scientific code and computational workbooks. Its financial model allows for free public hosting, where materials are available for any interested party.

Chapter 9

Computing cluster

This chapter only briefly characterizes the computing cluster used for bioinformatic analyses in this study, and the necessities of computing cluster for bioinformatic analyses.

Bioinformatics processes, like assembling genomes, clustering or matching sequences, analyzing diversity and performing statistical analysis on large datasets requires computers with relatively large amount of memory and computational power. Most of bioinformatics tasks are computationally intensive, and most processes, especially *upstream* computation (handling raw sequences) can not be run on standard personal computers. The supercomputer (a highly efficient computing unit) used in this study consisted of a coordinate server that could handle basic processing - computationally not demanding scripts. It was used primarily as a coordination server, from where jobs were submitted and queued through job scheduler. "Sun Grid Engine" used for this study is a computing cluster consisting of a master host and 59 execution host. Each execution hosts had an average of 24 computing cores, a single execution host with 40 cores and two hosts with 16 cores. Available memory ranged from $94.4GB$ to $377.9GB$. Open Grid Scheduler, often referred to as Grid Engine, is a commercially supported open-source batch-queuing system for distributed resource management [165]. The coordinating server schedules particular jobs, that were submitted to be run on more memory and computationally efficient clusters. The scheduling takes into account resources availability for a given request. Each analysis script is written in **`.pbs` script**. User specifies the amount of cores and memory required for a particular analysis. If submitted job has insufficient memory, the master node terminates further execution of that job. Analysis scripts in `.pbs` format are available in an online repository accompanying this thesis (see chapter 8, page 98).

Chapter 10

Collection protocol

Sample collection was performed at two separate time points by two different facilities. In the first study cohort gathered in year 2014 (referred hereinafter as 2014 dataset), sample collection was conducted by clinicians working at Oklahoma Medical Research Foundation (OMRF). In 2016 *Sanguine*, an external company in California was employed to perform a similar collection. Detailed protocols used for sample collection are included in an online repository accompanying this thesis at: `https://github.com/vaxherra/MicrobiomeSLE` (see chapter 8, page 98). This chapter contains more detailed descriptions of required criteria for patients to be included in this study.

Each patient was given a basic health questionnaire and informed as to the nature and importance of the study, after which he or she signed a written consent. Lupus patients had to have documented *prednisone* (most commonly prescribed steroid for lupus) dosage intake less than 20mg/day. One month prior to sample collection, patients and control groups were requested not to take antibiotics. None of the participants from both groups had previous GI surgery or GI conditions.

Although this study explores fecal microbiome, collected materials included fecal, saliva and blood samples:

1. **Saliva**: $5 - 10[ml]$ of saliva was collected for each patient using *OMNIGINE ORAL* collection kit (see supplementary materials in an online repository for a full description). The collection tube included a buffer released after closing the lid, allowing the sample to be stored at room temperature ($15°C - 30°C$) for up to two months prior to DNA extraction.

2. **Feces**: A commode style fecal collection unit and zip-lock bag was provided for each patient (see supplementary materials in an online repository for a full description). Participant transferred approximately 50mg of fecal sample to collection tube containing stabilizing liquid, allowing the sample to be stored at room temperature ($15°C - 30°C$) for up to two months prior to DNA extraction.

3. **Blood**: Several blood samples were collected. Samples were stored in different solutions, and needed to be processed within 24 hours of collection:

(a) 4× blood tubes containing Acid-citrate-dextrose (ACD) solution (an anticoagulant to preserve blood specimens) for **plasma** and **Peripheral Blood Mononuclear Cells (PBMCs)** (most specifically lymphocytes and monocytes) extraction. Later in the laboratory cells were separated from plasma by centrifugation ($1,200g$ for 20 minutes).

(b) 2× BD Vacutainer **Serum** Blood Collection Tubes Plastic (8.5 ml). After collecting the blood it is left undisturbed at room temperature to allow the blood to clot ($15-30min$). Later the samples were sent for processing at OMRF laboratory. Clot was removed by centrifuging at $1,000-2,600\times g$ for 10 minutes in a centrifuge (see entire blood processing protocol in supplementary materials).

(c) 1× Ethylenediamine Tetraacetic Acid (EDTA) (anticoagulant) treated collection tube for **plasma** and **PBMCs** extraction. Later in the laboratory cells were separated from plasma by centrifugation ($1,200g$ for 20 minutes).

(d) 2× PAXGene Blood RNA System collection tube was intended for stabilization of intracellular **RNA** from whole blood. PAXGene collection tubes were stored without processing in $-80°C$.

Study group characteristics are described in chapter 11 (page 102), while subsequent microbial DNA extraction from fecal samples are described in chapter 12 (page 107).

Chapter 11

Study group

Chapter 10 (page 100) indicated that sample collection was performed in two study cohorts: one conducted in 2014 in Oklahoma, and other in 2016 in California. However, both sample cohorts were processed at OMRF laboratory. For complete metadata files characterizing each sample, refer to online repository accompanying this study (see paragraph 8 on page 98). The basic characteristics of study cohorts is presented in terms of BMI, gender, ethnicity and other factors. Briefly, first cohort (2014) consisted of 31 study subjects, among which 21 were SLE patients, and 13 Healthy Controls (HCs). Second cohort (2016) used 66 patient samples from which 45 were SLE patients and 18 HCs. Subsequent figures explore factors like BMI, age, SLE presence or absence and ethnicity of studied subjects.

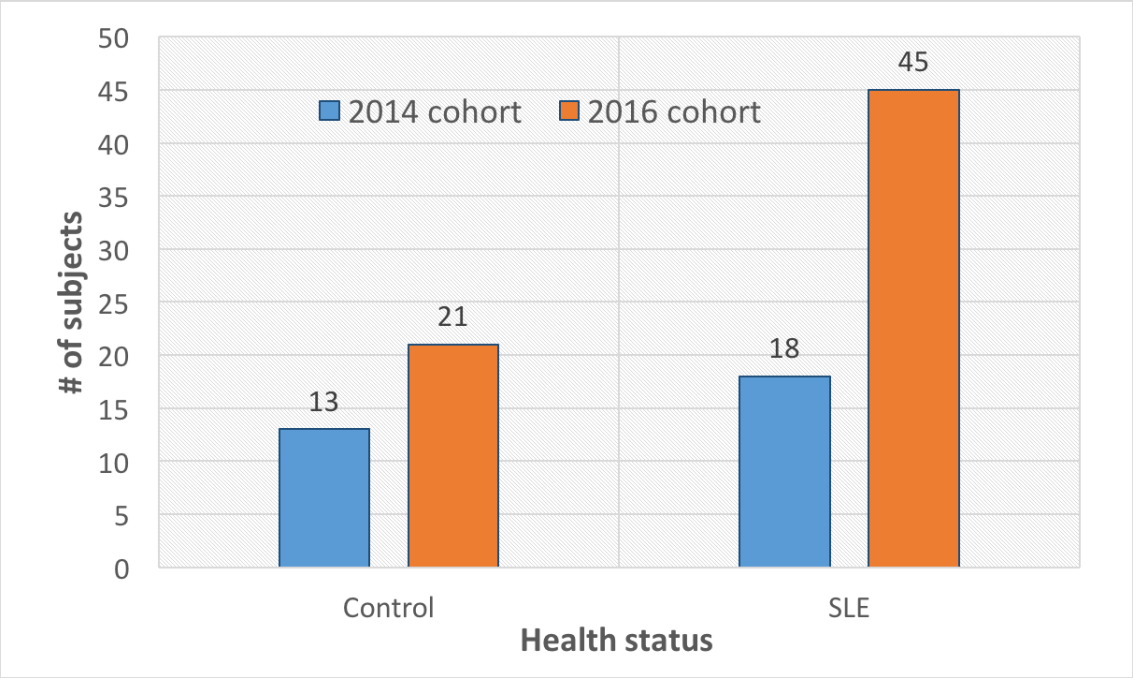


Figure 11.1: Histogram showing number of subjects in control groups (*control*) and SLE patients (*SLE*) in corresponding study cohorts.

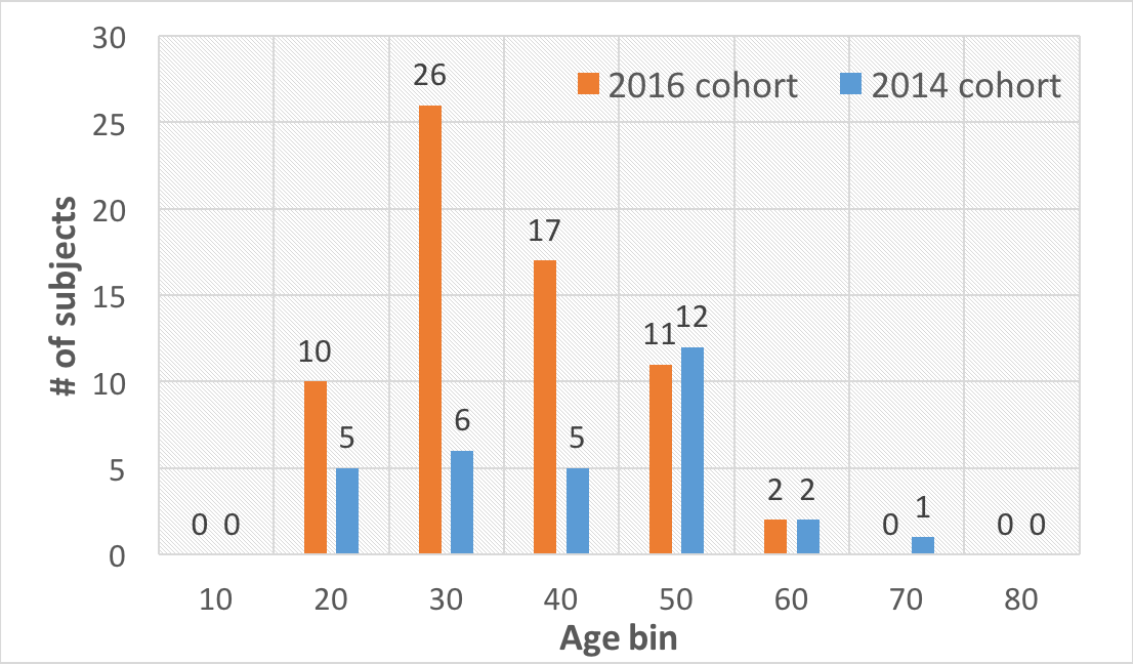


Figure 11.2: Histogram showing number of subjects in corresponding age bin according to study cohorts.

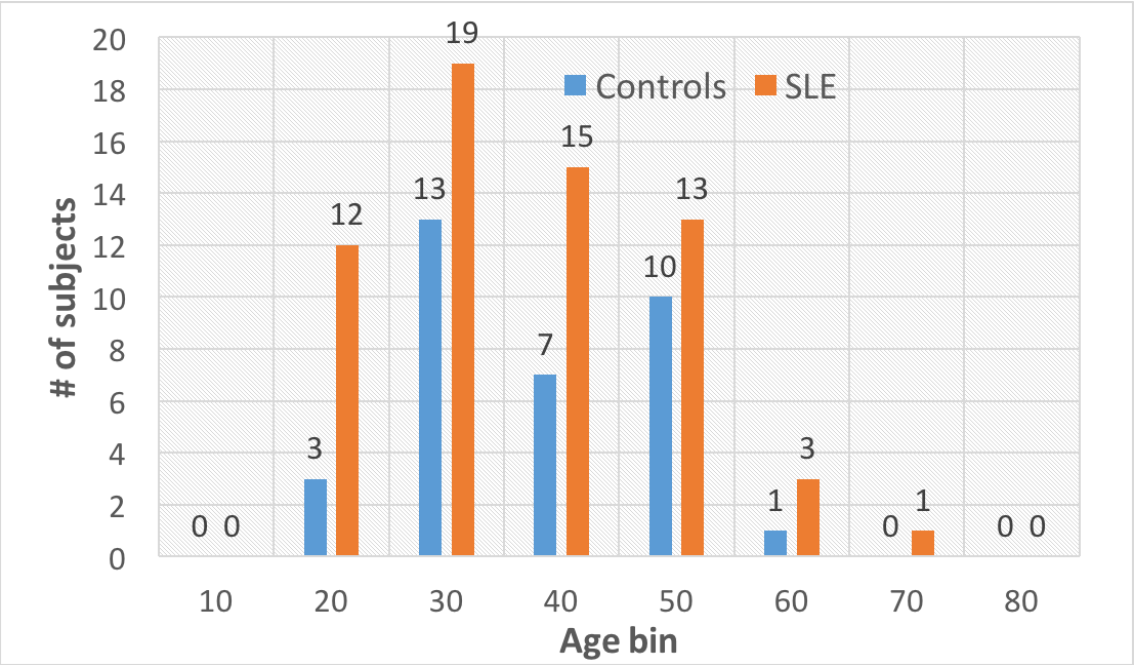


Figure 11.3: Histogram showing number of subjects in corresponding age bin according to health status (SLE/HC).

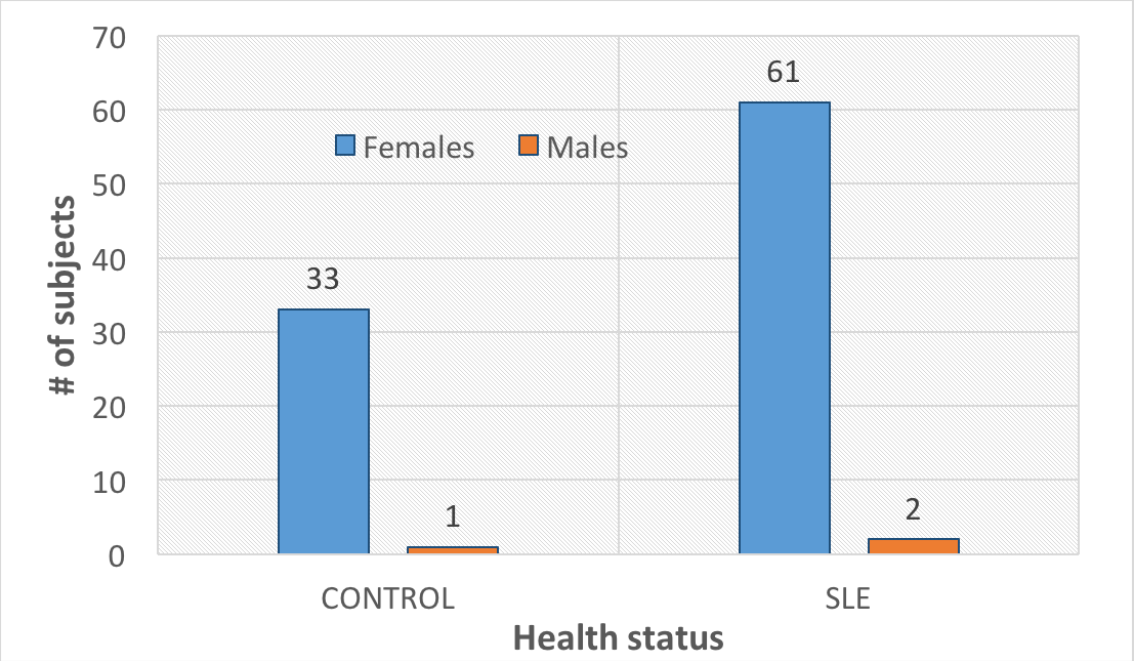


Figure 11.4: Histogram showing number of subjects: females and males in SLE patients and HCs.

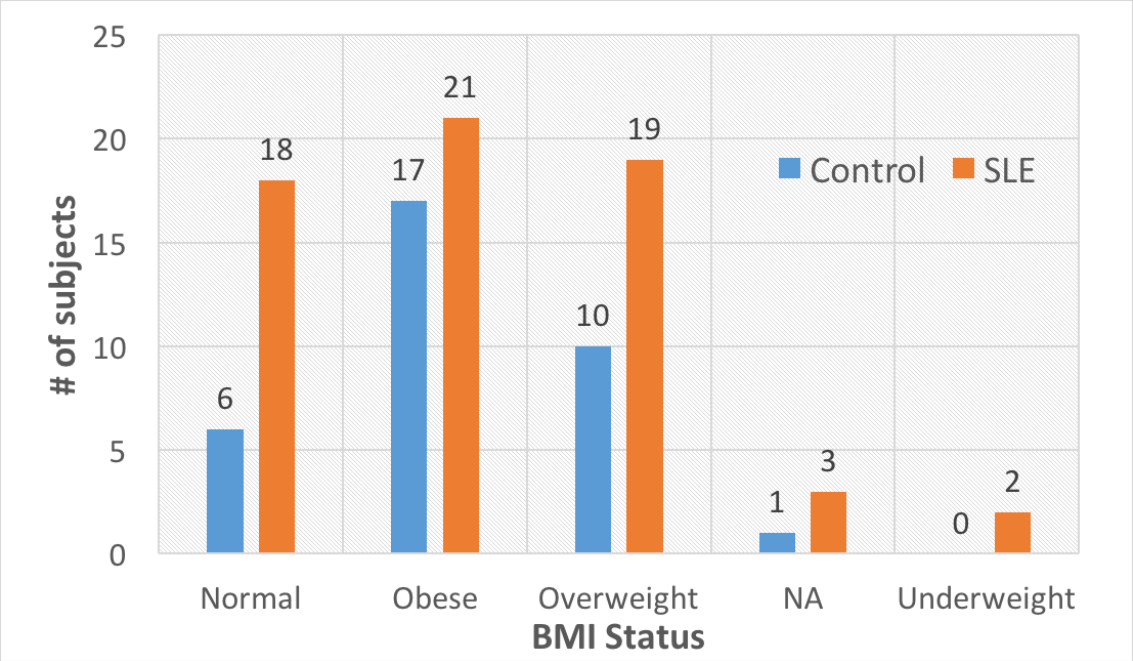


Figure 11.5: Histogram showing number of subjects according to their BMI status (underweight/normal/overweight/obese) in SLE patients and HCs.

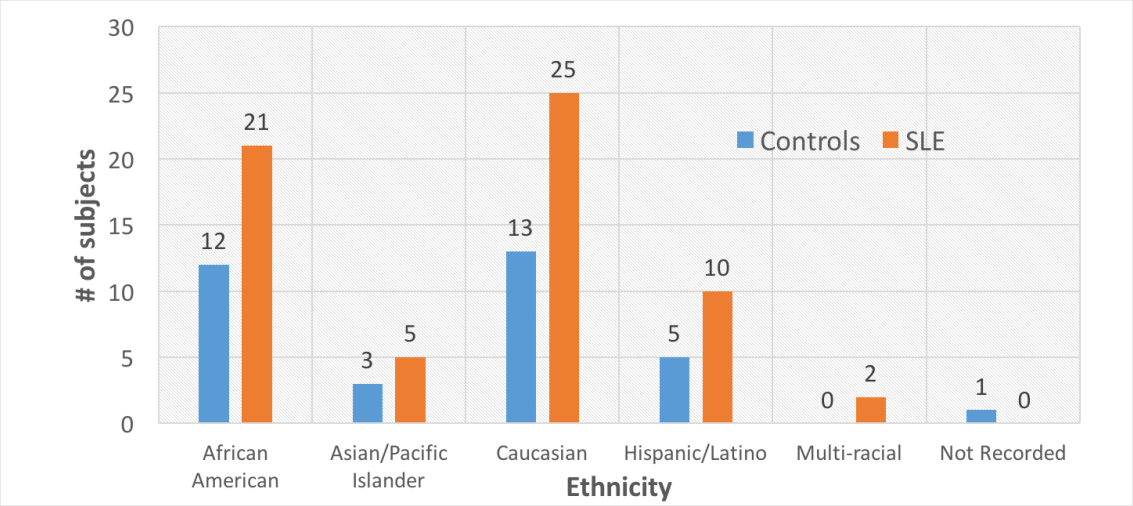


Figure 11.6: Histogram showing number of subjects belonging to a particular ethnic group, distinguishing between HCs and SLE patient.

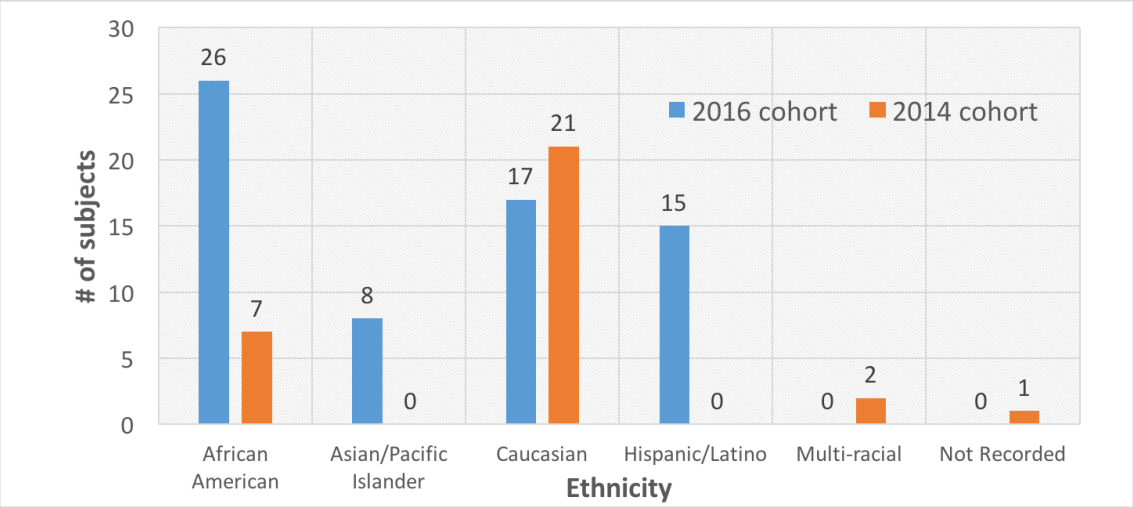


Figure 11.7: Histogram showing number of subjects belonging to a particular ethnic group, distinguishing between study cohort.

Chapter 12

DNA extraction

This experiment analyzed only fecal samples. For extracting microbial DNA from fecal samples, *PowerFecal® DNA Isolation Kit* by *Mo Bio Laboratories Inc.* was used. For full, detailed protocol description refer to an online repository accompanying this thesis - a supplementary materials sub-folder at `https://github.com/vaxherra/MicrobiomeSLE` or refer to appropriate product manual. Below is a brief description of fecal DNA extraction protocol using PowerFecal® DNA Isolation Kit. Each fecal sample from stored aliquotes (approximately 500 mg) was transferred to separate tubes that contain beads. Next the bead solution is added and the sample is vortexed in order to homogenize it and perform cell lysis with the help of other solutions (like Sodium dodecyl sulfate (SDS) - breaks down fatty acids and lipids associated with the cell membrane of several organisms). Cell lysis breaks the cell membrane, opens and exposes the DNA along with the cytoplasm. Sample was heated to $65°C$ in order to increase reaction rate, and aid cell lysis. Then, a "*lysate*" was removed after centrifugation.

From the lysate, non-DNA organic and inorganic materials were removed adding appropriate solutions, vortexing and incubating for several minutes. The process, called Inhibitor Removal Technology®, is patented. Since only DNA (excluding non-DNA organic and inorganic material) tends to bind silica at high salt concentrations, the next solution was highly concentrated with salts. It was put in a tube containing a spin filter made from silica onto which DNA binds. After several loads of prepared solution and subsequent centrifugation all DNA was bound to spin filter. Contaminants pass through the filter membrane, leaving the DNA bound to the membrane. Final steps included washing spin filter with ethanol-based solution to further clean the DNA from contaminants. In the end, elution buffer containing tris(hydroxymethyl)aminomethane (TRIS) released the DNA from spin filter. Obtained DNA was contained in $100\mu l$ volume, and was ready for any downstream applications, like PCR amplification. However, before any further processing, DNA was stored in $-20°C$ as all samples cannot be extracted simultaneously. DNA was extracted in batches of samples; each batch was accompanied by an extraction "*blank*".

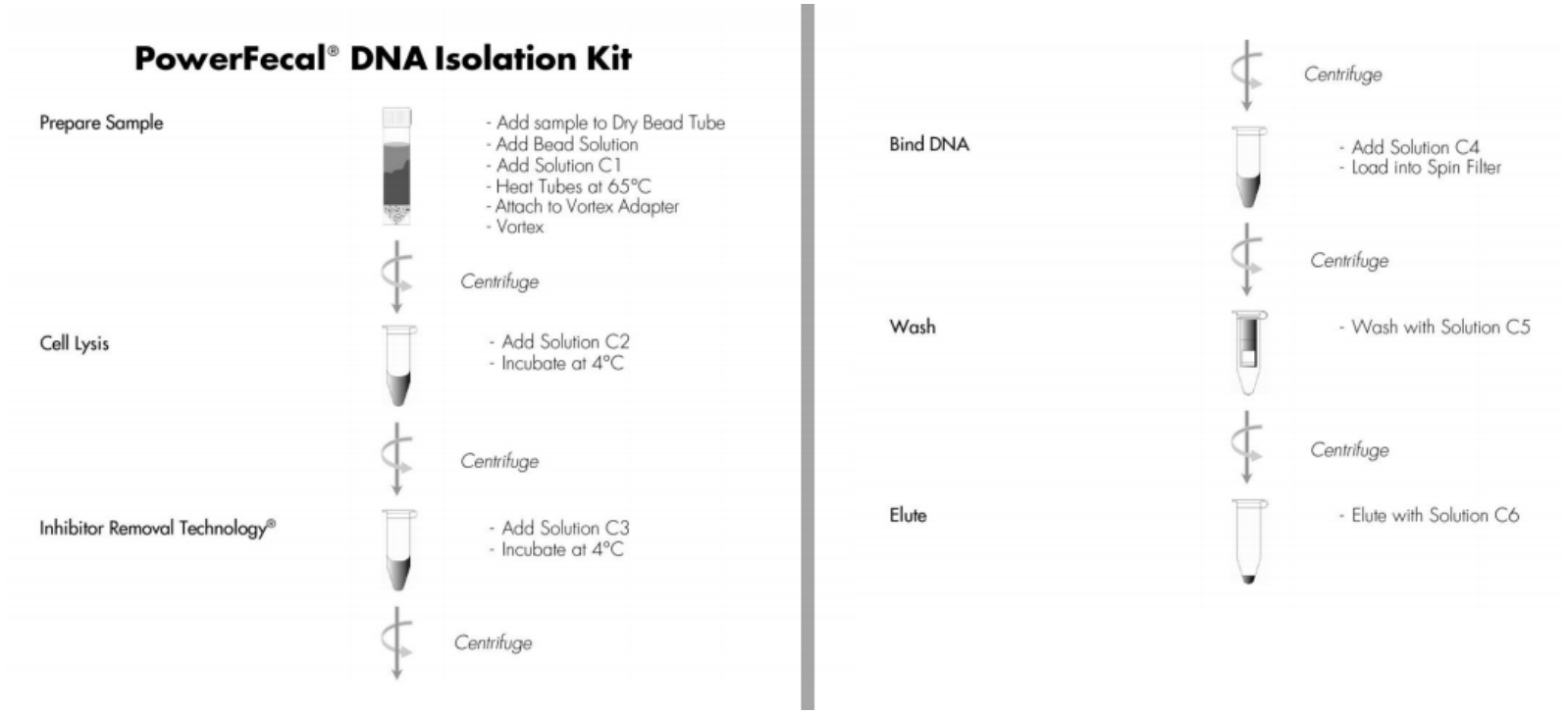


Figure 12.1: PowerFecal® DNA Isolation Kit workflow for DNA extraction. Directly reproduced from producer materials [148].

Chapter 13

Sequencing

13.1 Measuring DNA concentration

Extracted DNA was quantified in terms of concentration using Qubit® 1.0 Fluorometer with Qubit® dsDNA HS Assay Kit. Manuals provided by the manufacturer *Invitrogen* are shown in supplementary materials (see online repository at `https://github.com/vaxherra/MicrobiomeSLE`). Here a brief description is provided for steps undertaken to quantify the concentration of extracted microbial DNA from fecal samples.

Assay was performed under room temperature conditions. The first step involved a preparation of working station buffer (*Qubit® working solution*). For each sample $1\mu l$ of dye (*Qubit® dsDNA HS Reagent*) was mixed with 199 μl of read buffer (*Qubit® dsDNA HS Buffer*). Each sample requires 180-199μl of working station buffer, and for this experiment $199\mu l$ of working station buffer was used, and then mixed with $1\mu l$ of DNA. The two standards require $190\mu l$ of working station buffer mixed with $10\mu l$ of control, known concentration. The final volume must be $200\mu l$ for each sample. Prepared samples were mixed by vortexing, and allowed to incubate at room temperature for recommended 2 minutes.

For DNA concentration measurements Qubit® 1.0 Fluorometer was used, for which option *dsDNA High Sensitivity* was selected as the assay type. The first step was calibration using the two prepared calibration samples. Later, DNA concentration in samples were computed according to prepared standards (controls). The output is a `csv` format table, where concentrations are displayed in $\frac{ng}{\mu l}$. Results can be computed with relation to user-specified dilution ratio, which for our experiment was 1:200, based on the simple equation showed below:

$$\text{Sample concentration} = \text{QF value} \cdot \frac{200}{x} \tag{13.1}$$

where:
`x` - number of micro-liters of sample added to the assay tube,
`QF value` - the value given by fluorometer.

13.2 PCR amplification

16S rRNA sequencing survey preparation utilized *"Caporaso Primers"* and Miseq Illumina sequencing machine that generated 250bp paired-end reads. Selected primers were designed to amplify bacteria and archaea (prokaryotes). For more general information about 16S analysis see section 3.1 on page 38. Primers used for sequencing were 515F-806R primers (so called *"Caporaso primers"*). These primers target the V4 region of the 16S SSU rRNA (see chapter 3.1 on page 38). This is a standardized protocol as described on "Earth Microbiome Project" [172] with slight modifications. Protocol comprised of several steps listed below:

1. Normalized DNA concentrations (obtained from section 13.1) by dilluting DNA samples to obtain final concentration of $10\frac{ng}{\mu l}$.

2. Prepared a triplicate PCR reaction mixture as shown on table 13.1. Triplicate means each sample is amplified in 3 replicates, $25\mu l$ each.

3. Pooled triplicates for each sample into single volume of $75\mu l$.

4. Verified presence of PCR product through 4200 TapeStation System (see section 13.3). Expected band size for Caporaso Primers (515f-806r) are around 300-400bp. The target V4 region (515F-806R) is 291bp long but other factors have to be considered, like the added length of priming and indexing bases; hence 300-400bp is the expected region.

5. Quantified amplicon with Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher/Invitrogen cat. no. P11496; follow manufacturer's instructions; see supplementary materials in an online repository accompanying this thesis),

6. Pooled equal amounts of amplicon from each sample into a single sterile tube. It is recommended to take approximately 240 ng from each sample.

7. Cleaned amplicon pool with UltraClean® PCR Clean-Up Kit (for more information see supplementary materials in an online repository accompanying this thesis) designed to purify PCR products from primers, dNTPs (i.e. nucleotides) and reaction components in the size range $60bp - 10kbp$.

   Prior and after this step electropherograms were generated with Agilent 2200 TapeStation. For brief overview as well as electropherograms generated for thesis experiment refer to section 13.3 on page 111.

8. Measured the final concentration of final, cleaned pool.

9. Submitted an aliquot for sequencing to Clinical Genomics Center at OMRF

Table 13.1: Composition of triplicate PCR reaction mixture.

Reagent	**Volume**
PCR-grade water	*13.0 µl*
PCR master mix (2×)	*10.0 µl*
Forward primer (10 µM)	*0.5 µl*
Reverse primer (10 µM)	*0.5 µl*
Template DNA	*1.0 µl*
Total reaction volume	**25.0** *µl*

13.3 Verification of PCR products

The 2200 TapeStation system performs electrophoretic separation of nucleic acids and proteins, therefore is widely used for quality control in NGS, micro-array and qPCR workflows. It can also be used in protein purification and antibody production steps [202]. This system detects fluorescently stained double stranded DNA including genomic DNA, stained total RNA and labelled proteins. The disposable screen-tape contains a series of lanes allowing for individual sample separation (up to 96 samples) and further quality control.

Results can be displayed as:

- electropherograms,
- gel images,
- tabular data format.

Agilent 2200 TapeStation was used to verify the presence of PCR products from each sample pool corresponding to study cohort (see study group chapter on page 102). Next page presents two electropherograms, one before and one after cleaning steps as described in previous section 13.2 on page 110 for 2016 study cohort.

13.4 DNA sequencing

13.4.1 Cluster generation and sequencing

In order to perform a run on the MiSeq platform, MiSeq Reagent Kit v2 was used (Catalog # MS-102-2003). For a detailed protocol refer to appropriate manufacturer Preparation Guide [130]. This documents are reproduced in supplementary materials accompanying this thesis in an online repository.

Kit reagent of the 500-cycle size was used. It allows up to 525 cycles of sequencing, which is satisfactory for up to a 251-cycle paired-end run plus two eight-cycle index reads. Kit reagents included in MiSeq product need to be stored in appropriate temperature indicated in the manual. Prior to sequencing they need to be thawed using a room temperature water bath. Sample libraries are then loaded onto the cartridge in appropriate position. MiSeq reagent cartridge is a single-use consumable that has pre-filled and numbered reservoirs with clustering and sequencing reagents

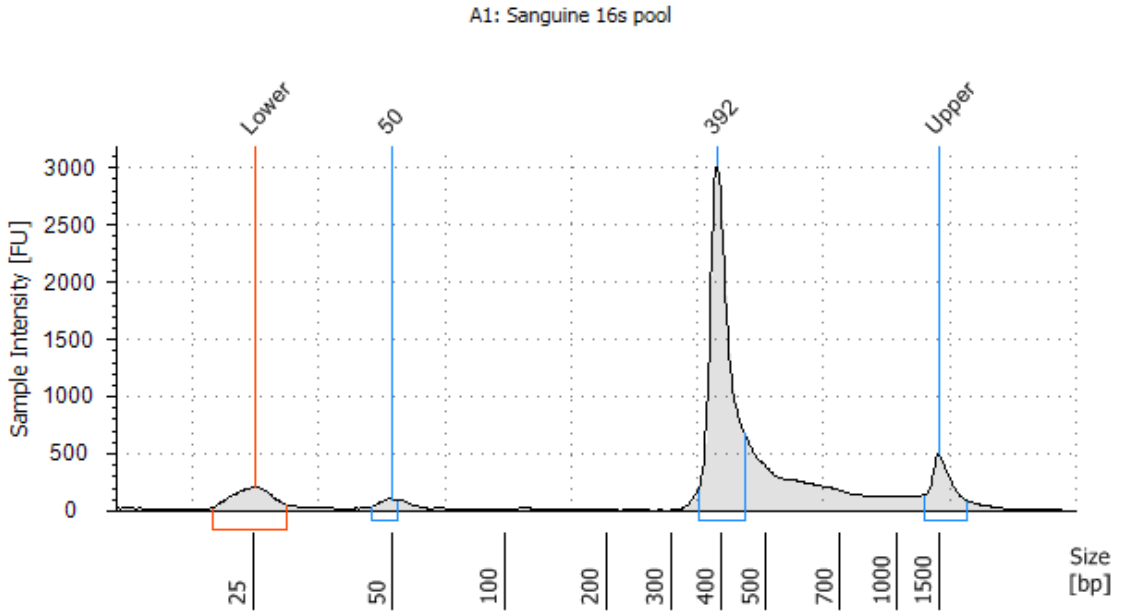


Figure 13.1: Electropherogram representing pooled amplicons from 2016 sample cohort before cleanup process. Visible contaminant PCR products at 50bp region, and amplified 16s rRNA V4 region of around 390bp length are shown. The "upper" and "lower" regions correspond to the TapeStation internal markers used for calibration.

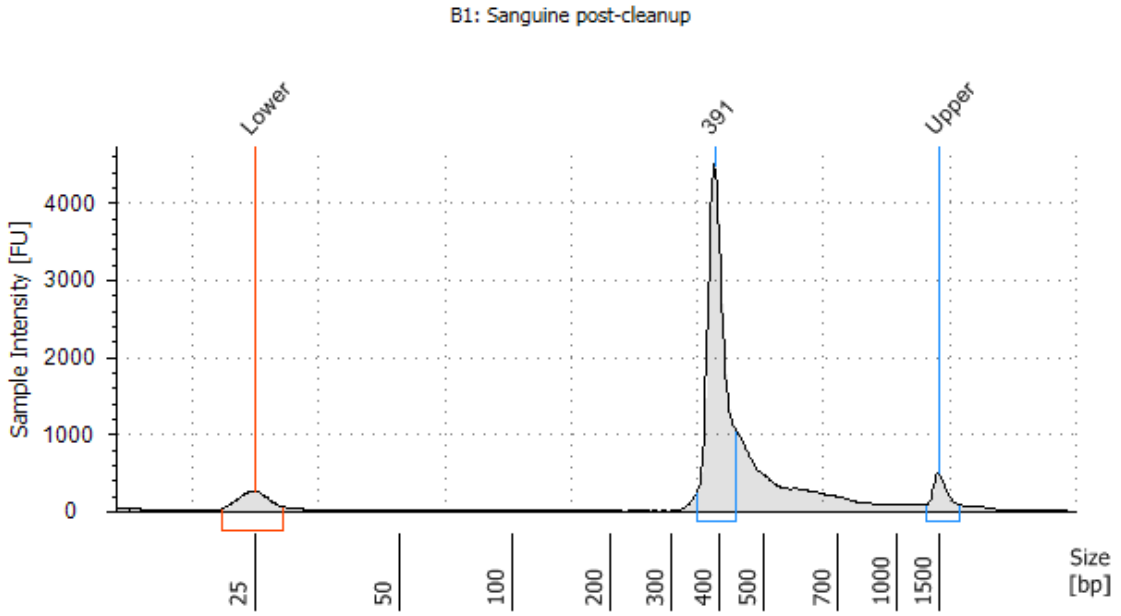


Figure 13.2: Electropherogram representing pooled amplicons from 2016 sample cohort after cleanup process. No visible contaminant PCR products at 50bp region, but amplified 16s rRNA V4 region of around 390bp length is visible. The "upper" and "lower" regions correspond to the TapeStation internal markers used for calibration.

sufficient for sequencing only one flow cell. Through the MiSeq Control Software (MCS) interface, a step-by-step procedure of loading a flow cell and reagents is performed.

Sequencing technology, Sequencing by Synthesis (SBS), is described in the theoretical part of this thesis (see section 4.3.2 on page 47). Cluster generation is described in section 4.2 on page 44. Cluster generation and sequencing were preformed by NGS Clinical Genomics Center at Oklahoma Medical Research Foundation (OMRF).

13.4.2 Sequencing results statistics

This chapter briefly characterizes the quality of obtained reads from each sequencing run (cohort 2014 and 2016 - see study group description at page 102 or collection protocol description at page 100).

Q-scores represent the overall technical quality of obtained sequences. The greater the number, the lesser the probability of incorrect base calling. See more detailed description of `phred` quality scores at page 62. Q-Score distributions from 2014 and 2016 cohort are shown on figures 13.3 and 13.4 respectively. For 2014 cohort, the percentage of bases with quality score above `30` is 77.3%, whereas for 2016 cohort it is 80.8%. Incomplete removal of fluorescent ddNTPs occurs for increasing read lengths. This leads directly to decreasing quality scores - see discussion of *sequencing errors and problems* at page 57.

Figures 13.5 and 13.6 represent the same q-scores for two studied cohorts but as a heat-map. Quality scores are plotted for each cycle. It is important to remember that the sequencing was conducted in paired-end mode, where one read was obtained through 250 cycles. Figures 13.5 and 13.6 clearly show the loss of quality at the end of sequencing cycles (approximately from 200 to 250 and 450 to 500), where sequences with lower qualities are more prevalent.

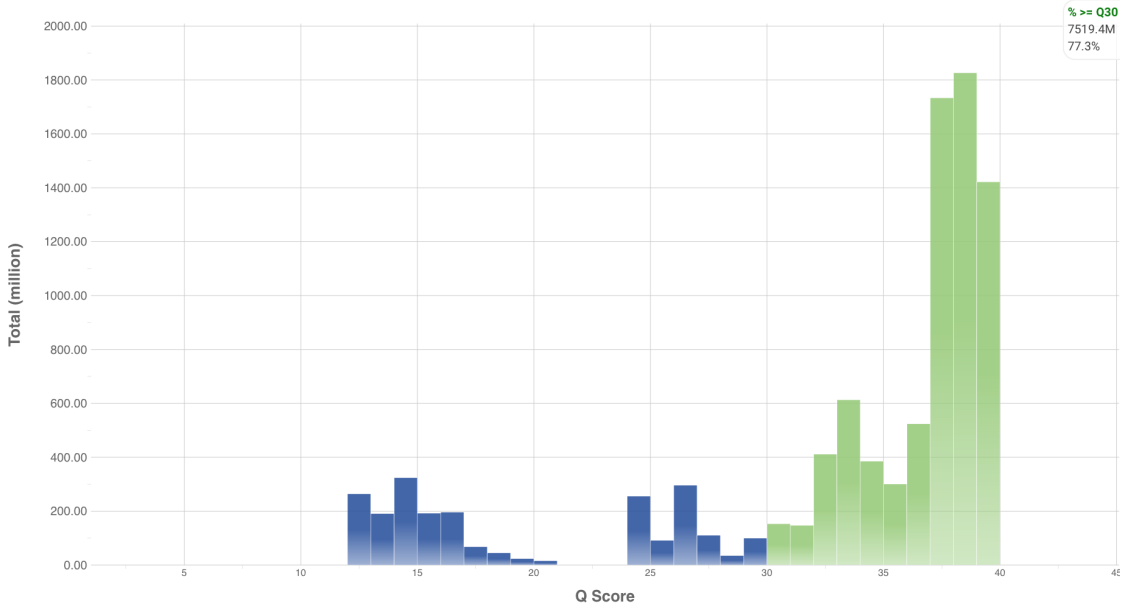


Figure 13.3: Q-score distribution for **2014** cohort.

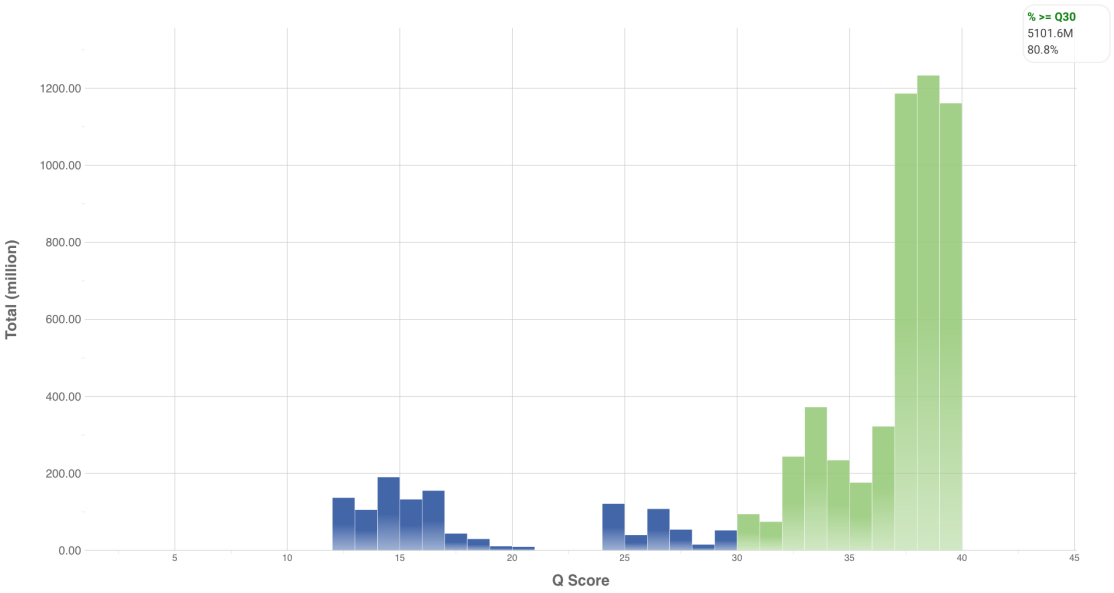


Figure 13.4: Q-score distribution for **2016** cohort.

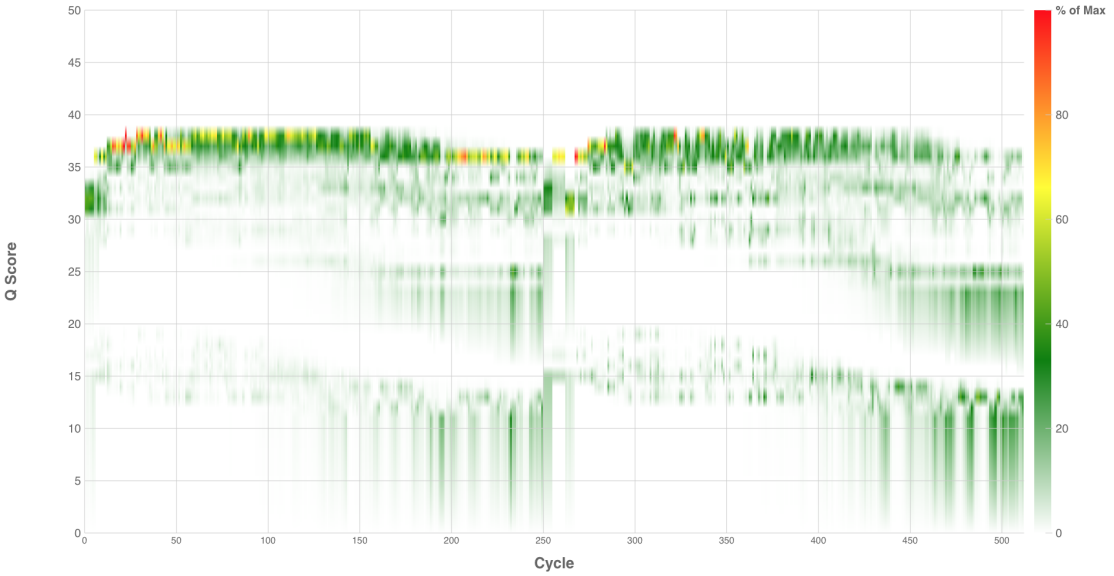


Figure 13.5: Q-score heat-map for **2014** cohort. Scores are plotted against number of sequencing cycles.

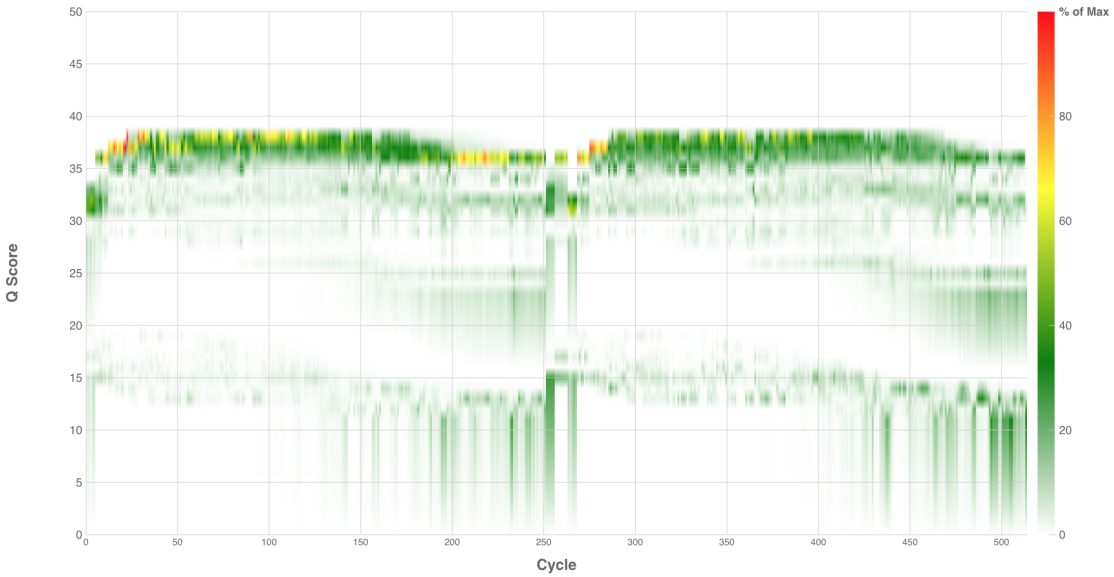


Figure 13.6: Q-score heat-map for **2016** cohort. Scores are plotted against number of sequencing cycles.

Chapter 14

Software versions and dependencies

This chapter lists all major software packages used for data analysis. Figure 14.1 on the next page specifies software package names and versions. As reproducibility in the microbiome field has been reported to be a major problem [176], this thesis aims at facilitating reproducibility of research conducted by several means, among which listing all software versions and their dependencies is one of the measures taken.

Software package	Version
usearch	v9.2.64_i86linux32
PICRUSt	1.1.0
STAMP	2.13
trimmomatic	0.35
ncbi_blast	2.2.22
QIIME library	1.9.1
QIIME script	1.9.1
qiime-default-reference	0.1.3
NumPy	1.11.2
SciPy	0.18.1
pandas	0.13.0
matplotlib	1.1.1
biom-format	2.1.5
qcli	0.1.1
pyqi	0.3.2
scikit-bio	0.2.3
PyNAST	1.2.2
Emperor	0.9.51
burrito	0.9.1
burrito-fillings	0.1.1
sortmerna	2.0, 29/11/2014
sumaclust	Version 1.0.00
swarm version	1.2.19 [Oct 31 2016 11:19:15]
GreeGenes	13.5 (PICRUSt predictions)

Figure 14.1: Table listing software packages and their appropriate version that were used directly or indirectly for data analysis in this experiment.

Chapter 15

Analysis pipeline for 16S rRNA amplicon sequencing

15.1 Rationale

Microbiome research is a complex study that includes many variables spanning from clinical metadata, DNA (RNA or metabolite) treatment and extraction, to various analytical and computational approaches in upstream and downstream analyses. The word *"upstream"* refers to processing of raw sequences - sequences obtained from sequencing machines. The word *"downstream"* in this context implies various statistical tests on obtained features, like OTUs, taxonomic groups, their relative abundances, orthologous groups or certain signaling pathways of microbial communities in each sample. Since microbiome research is a relatively new field, where new tools for *upstream* and *downstream* analyses are either created or continuously updated, certain researchers (see for example: [180] or [177]) have urged the scientific community to put enormous significance on the reproducibility of the research. Listing software packages used for the study is often not sufficient; the corresponding versions need to be provided, as well as the way in which output from one processing step is used as the input for another. Specifying various parameters of many statistical tests or algorithms used, ex. quality filtering, is necessary to achieve reproducibility in microbiome research.

15.2 Overview

Taking into account the above-mentioned reasons, an analysis pipeline including both upstream and downstream analysis steps was created. Source code is available at online repository accompanying this thesis (see chapter 8 at page 98). The overview of the pipeline is presented as a block structure on figures 15.1 and 15.2. Previous chapter 14 (page 116) lists all software and databases used with corresponding versions. As shown on figure 15.1 (page 121), this pipeline takes as an input raw amplicon read files (in `fastq` format). Blue background of this figure corresponds to a *preprocessing* step. During this step technical adapter sequences are removed through the help of `trimmomatic` software. The reasons for presence

of adapter sequences are described in *sequencing by synthesis* section on page 47. Each sample has two corresponding reads: forward and reverse that are then merged allowing only for base error (`-fastq_maxee 1.0`). Samples are separated based on a sequencing run (a cohort) and then divided into desired sequences coming from samples, and negative controls, i.e. contaminating sequences coming from DNA extraction controls that were sequenced. After this step, sample sequences are aligned to contaminating sequences. Only sequences unmapped to contaminants are considered for further downstream analyses. Sequences are then pooled, i.e. concatenated into one file, where each read has an identifying header name referring to sample ID - see `FASTQ` file format structure on page 61.

Green background (figure 15.2, page 122) contains steps undertaken for composition and diversity analysis, while the brown background refers to steps for inferring a metagenome content. Finally, the purple background represents downstream analyses, obtained feature profiles or compositions, and methods for analyzing them. Composition and diversity processing steps begin with a so called "*dereplication*" process, also referred to as unique sequences search. This step scans the entire pooled sequence file and outputs sorted `fastq` file, where first sequences are most abundant (decreasing order). Each head of the read contains additional field separated by semicolon describing its abundance. This step is crucial for OTU picking approaches. Sequences are then clustered into OTUs, or rather ZOTUs [97] (see section 6.1.5.1, page 69). ZOTUs (Zero-radius OTUs) are valid OTUs not based on 97% similarity threshold but rather on a phenomenological model constructed for positive controls obtained from several mock communities and Illumina in-vivo Illumina datasets. Produced OTU tables are then split into two (or several) parts depending on the number of sequencing runs; this experiment consisted of two study cohorts. Since reportedly up to 2% of reads are assigned to incorrect samples it is a source of possible bias, especially when low-abundance taxa are considered. `Uncross` algorithm aims at removing unexpected counts of taxa by simple heuristics described in detail in [96]. OTU tables need to be split into tables containing samples sequenced on the same run, as this is the characteristic of cross-talk phenomenon. Parallel steps to generate and filte OTU tables are: predicting taxonomy and preparing phylogenetic tree. Taxonomy in this step is predicted with RDP classifier containing 13 thousand sequences providing classification up to genus level [206]. This is publicly released version 16 a collection of 12,681 bacterial and 531 archaeal 16S rRNA gene sequences with an improved taxonomy. Taxonomic annotation is performed for each OTU, and added to the filtered OTU table. Third step prepares a phylogenetic tree, a tree relating each OTU in terms of sequence similarity and phylogenetic relationship on the tree of life. First the multiple sequence alignment in performed with PyNAST algorithm, alignment is filtered for gap signs and then phylogenetic tree is constructed. Phylogenetic tree, OTU table with taxonomic annotations, a mapping file with clinical metadata and a `QIIME`-specific configuration file stand as an input for downstream analyses. Qiime-specific configuration file can contain many parameters; for this pipeline it only manually specifies that all alpha metrics should be computed or paths to databases located in custom directory on a remote computing server (see `qiime_config.txt` in online repository). The results obtained are rarefaction plots, alpha diversity metrics, beta diversity metrics, PCoA plots, and

taxonomic plots. Some of those results are further analyzed, for example to infer statistical differences between control and lupus patients. Brown background on the figure 15.2 corresponds to `PICRUSt` metagenomic predictions. `PICRUSt` overview and mechanism is discussed in section 7.3.1 on page 90. `PICRUSt` requires first that the OTUs be picked against *GreenGenes* database version 13.5 specifically. Each OTU needs to have a *GreenGenes* identifier, thus closed-reference OTU picking is required. For this step "`usearch_ref`" method was used through `QIIME` functionality. Other methods are possible, like "`uclust_ref`" however this is not recommended for OTU picking [95], as it was not intended for this purpose. Obtained OTU table is normalized for the expected 16S gene copy number for each taxon, then. through ASR algorithm orthologous groups are predicted. `PICRUSt` is able to collapse predicted KEGG Orthologs (KOs) into modules (level 2) or pathways (level 3). The output is a matrix in which columns correspond to samples, and rows to features (orthologs, modules or pathway abundance). Using Statistical Analysis of Metagenomic Profiles (STAMP) software package, samples can be compared according to selected clinical metadata features (like healthy control and disease state). `PICRUSt` functionality allows for backtracking of metagenome contributions in order to identify certain taxa responsible for enrichment/depletion of orthologs, modules or pathways in one of compared the groups.

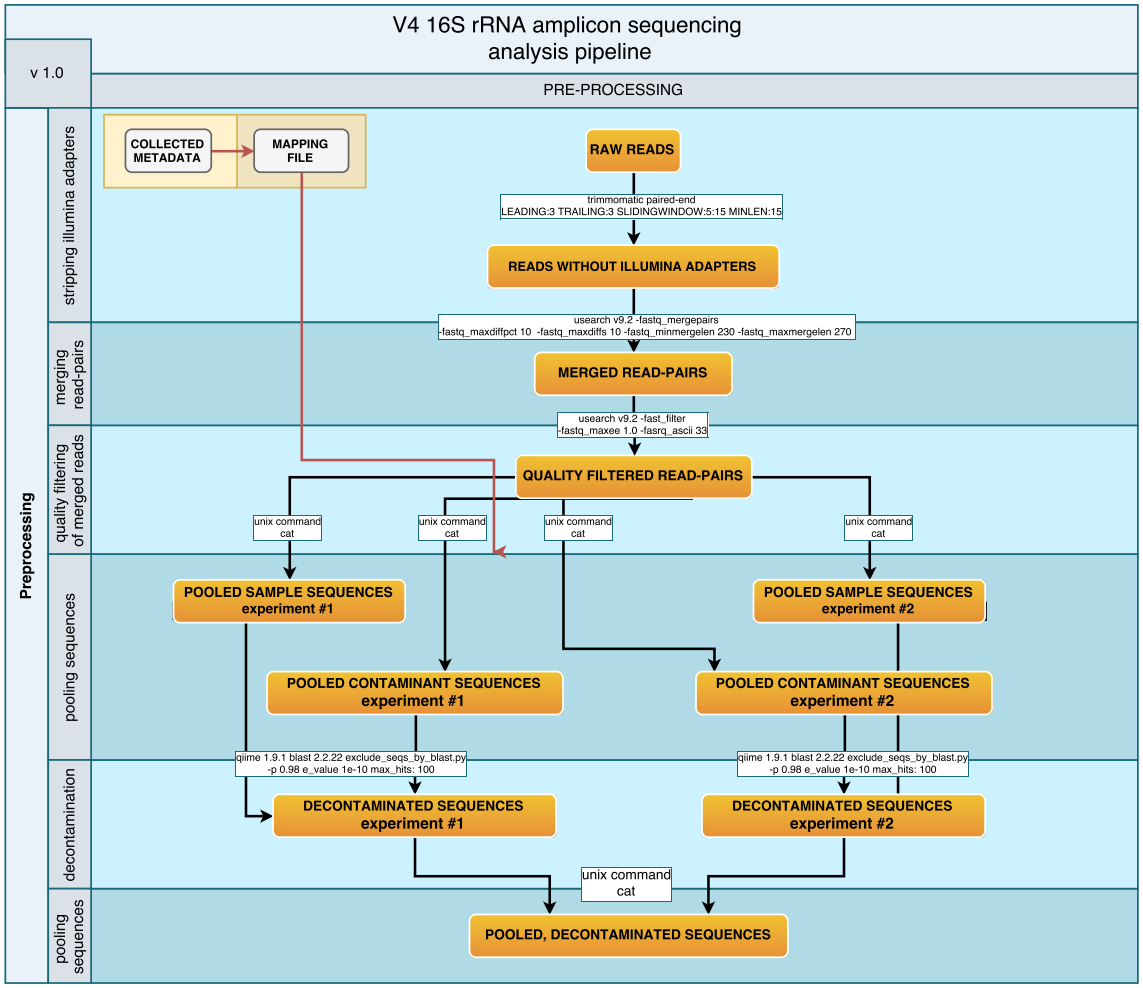


Figure 15.1: Analysis pipeline - preprocessing. A block diagram representing preprocessing steps in 16S rRNA amplicon sequencing survey. Yellow boxes correspond to the state of the data, white rectangular boxes on arrows represent software packages with particular scripts and parameters used. Blue-gray boxes with vertical text explain analysis steps performed.

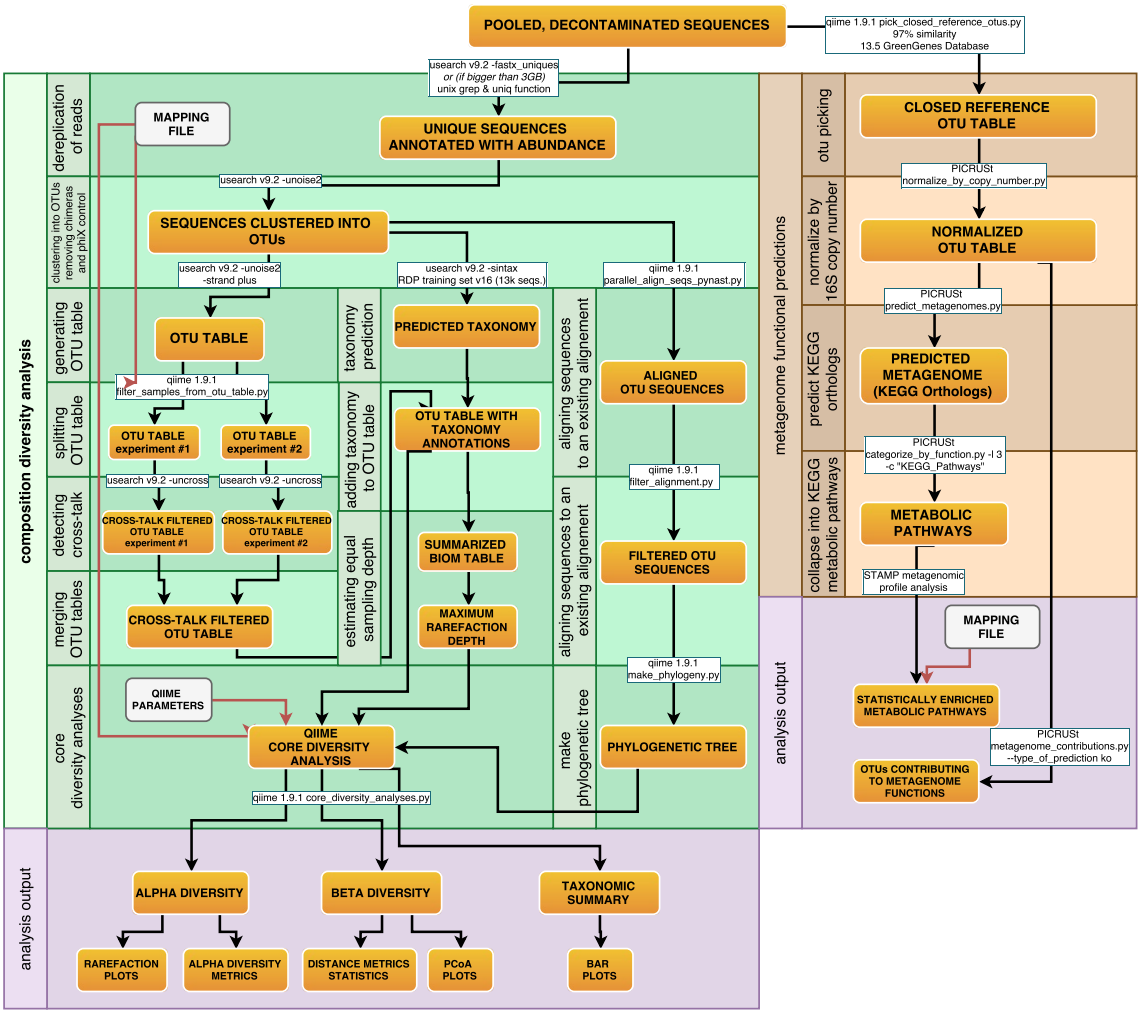


Figure 15.2: Analysis pipeline - composition diversity analysis and metagenomic predictions. A block diagram representing downstream analysis steps in 16S rRNA amplicon sequencing survey. Yellow boxes correspond to the state of the data, white rectangular boxes on arrows represent software packages with particular scripts and parameters used. Vertical text on left side of each block explains analysis steps performed. Light purple box contains outputs from analysis pipeline.

Part III

Results & Discussion

Chapter 16

Diversity analyses

16.1 Rarefaction plots

The basic definition of a rarefaction step and the method used for obtaining rarefaction plots are described on page 78 (section 6.1.8.1). Rarefaction is a technique used for measuring "saturation", i.e. a state when increasing sequencing depth (obtaining more reads) would not yield additional, detailed insight into a studied community. In other words, when curves presented on a rarefaction plot tend to *plateau* after reaching certain sequencing depths, we could ascertain that we've observed a representative sample, or samples of certain subjects. Figure 16.1 and 16.2 show rarefaction plots for all samples but different metrics. Figure 16.1 describes the number of observed species, and 16.1 phylogenetic diversity over increasing sequencing depth. Observed number of species is a straightforward measure, where species actually refers to OTU and each OTU is regarded as being observed or not observed; frequencies are not considered in this metric. Thus observed number of distinct taxa are reported through this metric. Phylogenetic diversity metric uses a phylogenetic tree and is expressed as the sum of the phylogenetic branch lengths in a phylogenetic tree that is covered (represented) in a given sample [103]. Subsequent figures (16.3 and 16.4) present rarefaction plots where individual samples are collated to compare Healthy Controls (HCs) and SLE patients. Presented figures suggest that this experiment presented enough sequencing depth for the samples to be considered representative of studied communities. Rarefaction plots grouped according to status (healthy vs SLE patient) suggest that SLE patients exhibit a smaller number of distinct taxa. Indeed this difference is statistically significant and is further explored in next section - *alpha diversity* 16.2, page 127.

Rarefaction is an important measure that ensures the correctness of made comparisons between samples or group of samples. Amplicon sequencing results in sample reads with varying number of reads. Comparing samples with unequal number of reads (unrarefied samples) is an error to be avoided, as more sequences would/could yield more OTUs, i.e. observed bacteria. Therefore one approach is to normalize the number of reads to the smallest number of reads among all samples sequenced. This approach has a drawback of discarding valid sequences, however amplicon sequencing results in relatively uniformly distributed read lengths, therefore small percentage of data is actually discarded. This is a more serious concern in WGS.

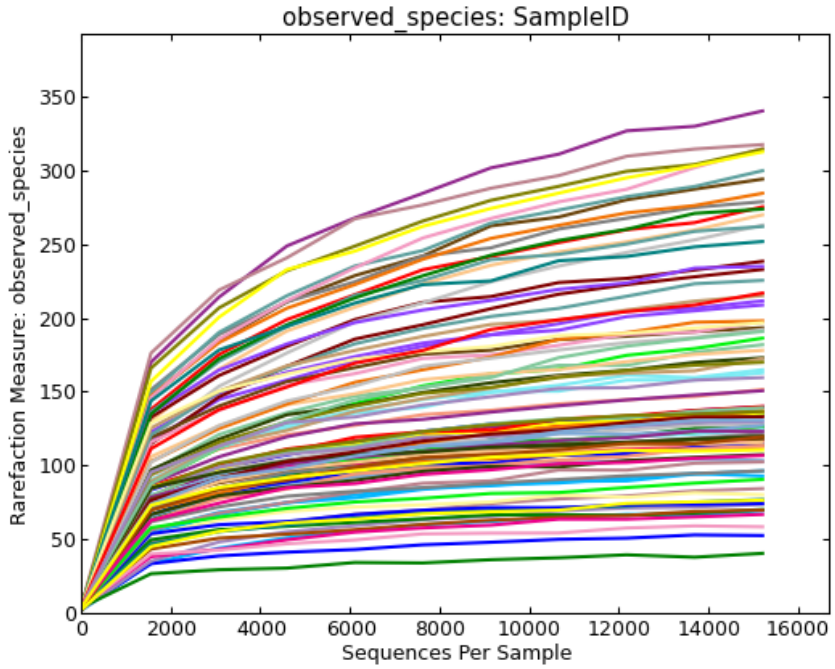


Figure 16.1: Rarefaction plot. Each colored line represents a sample studied. Used metric: **observed number of species**.

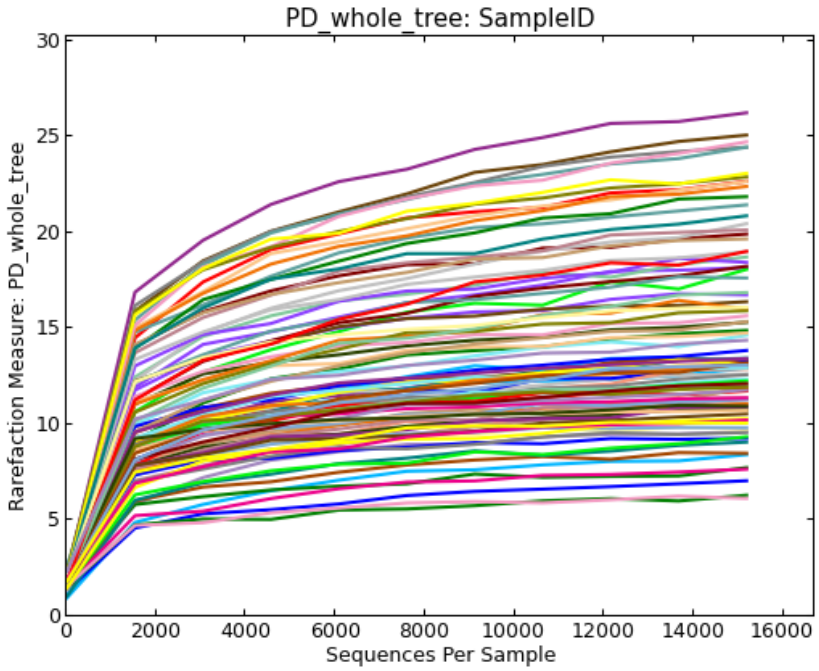


Figure 16.2: Rarefaction plot. Each colored line represents a sample studied. Used metric: **phylogenetic diversity (PD)**.

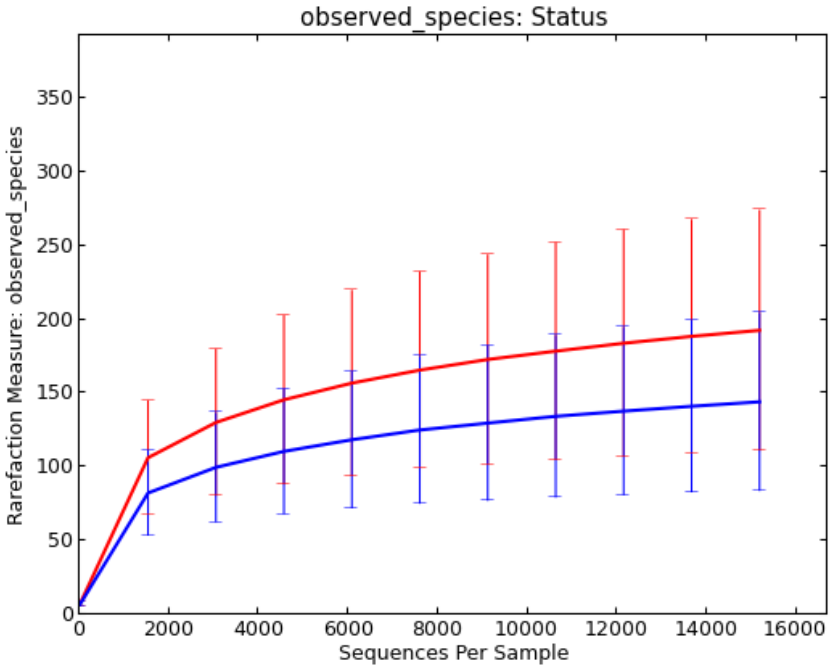


Figure 16.3: Rarefaction plot collating Healthy Controls (HCs) and SLE patients. Used metric: **observed number of species**.

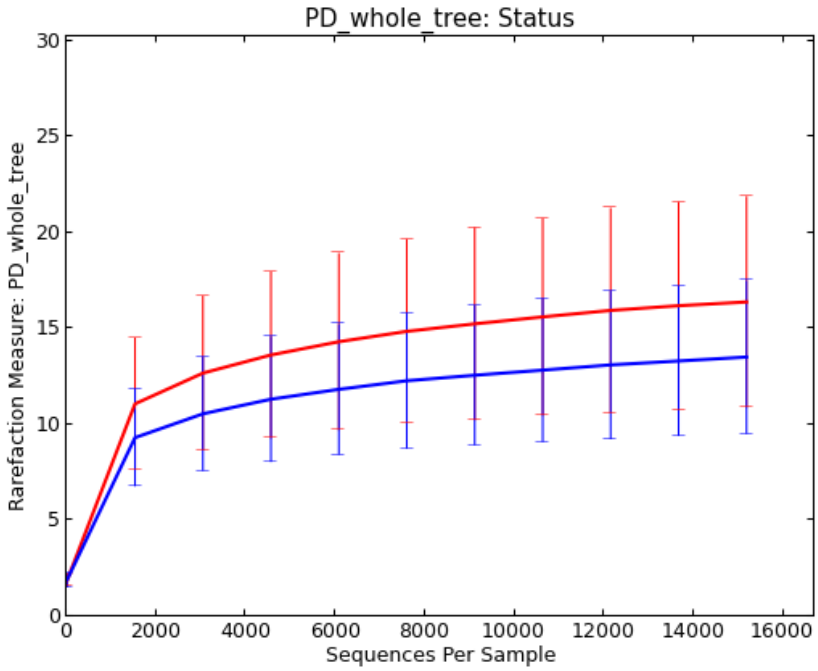


Figure 16.4: Rarefaction plot collating Healthy Controls (HCs) and SLE patients. Used metric: **phylogenetic diversity (PD)**.

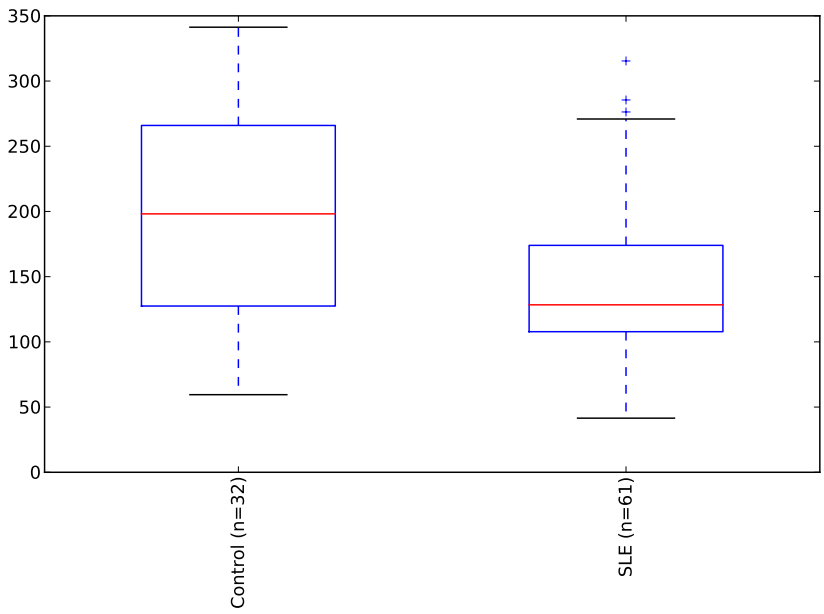


Figure 16.5: Alpha diversity with the number of **observed species** (Y-axis) for Healthy Controls (HCs) group and SLE patients.

16.2 Alpha diversity

Alpha diversity, a diversity also referred to as within-sample diversity or richness, conveys in many available metrics how many different types of organisms are present (identified) in a particular sample or group of samples. Alpha diversity and its metrics are briefly described in section 6.1.8.2, page 78. Figure 16.5 and 16.6 show alpha diversity for observed number of species and phylogenetic diversity respectively. For both metrics there is a clear decrease in alpha diversity for SLE patients. Comparison groups of samples coming from either healthy controls or SLE patients was done through non-parametric t-test, using Monte Carlo permutations to calculate p-value. The case for most of genomic and metagenomic research is that a normal distribution can not be assumed, therefore a non-parametric test was used.

Table 16.1 on page 128 presents results of non-parametric t-test with Monte Carlo permutations. Both metrics used the observed number of species (i.e. OTUs) and phylogenetic diversity and found 2 a statistically significant decrease in alpha diversity: $p = 0.001$ and $p = 0.008$ respectively. Decreased alpha diversity is a hallmark of many diseases including Inflammatory Bowel Disease (IBD) or Irritable Bowel Syndrome (IBS). For gut dysbiosis discussion in general see chapter 1.4 at page 22; for significance discussion of these results see discussion (section 18) at page 169.

16.3 Beta diversity

Beta diversity (β-*diversity*), sometimes referred as *across sample diversity* or *between sample diversity*, is a statistical method for comparing different communities

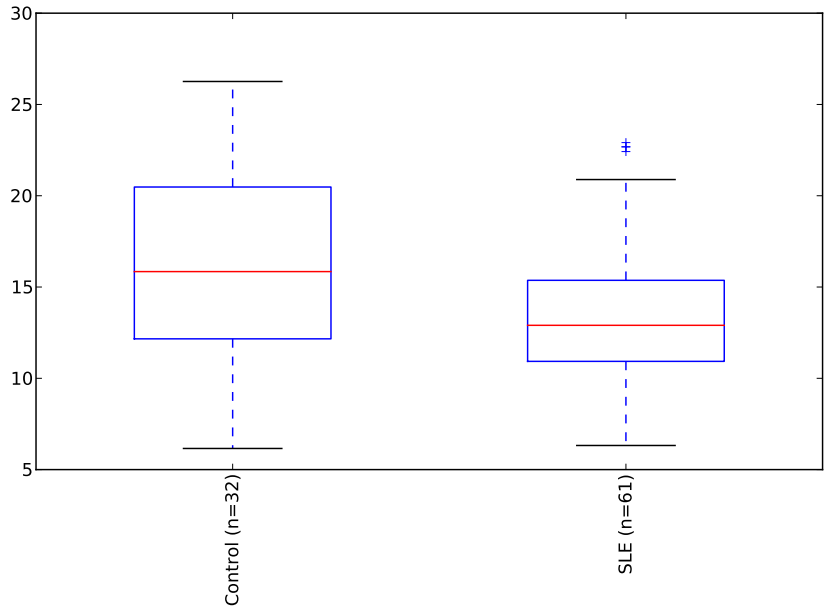


Figure 16.6: Alpha diversity with **phylogenetic diversity** metric values (Y-axis) for Healthy Controls (HCs) and SLE patients.

Table 16.1: Alpha diversity statistical test results showed for two metrics used: observed species and phylogenetic diversity. After non-parametric t-test with Monte Carlo permutations, both report statistically significant metric value decrease in SLE patients compared to Healthy Controls (HCs).

Metric	Control mean	Control std	SLE mean	SLE std	t stat	p-value
Observed species	192.62	81.64	144.17	60.45	3.21	0.001
Phylogenetic diversity (PD)	16.40	5.51	13.52	4.00	2.84	0.008

and estimating, through a quantitative measure, how much similar or dissimilar one community (or group of communities) is in comparison with another. For general beta diversity theory and overview see section 6.1.8.3 on page 79. In a manner analogous to alpha diversity, beta diversity is calculated through various distance metrics. `UniFrac` is a metric incorporated in this study. `UniFrac` is a phylogenetically aware distance metric that can take into account relative microbial abundance. `UniFrac`, being phylogenetically aware, allows for more sensitive comparisons; its value representing community dissimilarity (i.e. distance) is a single number in the range of $< 0, 1 >$, where 0 describes complete similarity between communities (zero distance), and 1 represents full dissimilarity between compared communities.

16.3.1 PCoA plots

Principal Coordinates Analysis (PCoA) is an ordination method that is the most widely used method to study biological diversity. The main purpose is to visualize distance matrix (like `UniFrac` distance) in order to interpret the results. PCoA method is described in section 6.2.1 on page 82. Figure 16.7 and 16.8 show PCoA plots for **unweighted** `UniFrac` metric from different angles. Figures 16.9 and 16.10 show the same results but for **weighted** `UniFrac` metric. The percentages of variability explained by three main (highest contributing) axes on PCoA plots add up to 28.15% for unweighted and 61.33% for weighted `UniFrac`, the latter capturing more variability. PCoA plots do not exhibit strong and visible "clustering" of samples coming from HCs and SLE subjects. It is worth mentioning here that the Human Microbiome Project (HMP) observed clear, yet not perfect separation in PCoA plots of distance matrices for different healthy body sites (see figure 6.15 on page 81). Contrary to that study [78] investigating microbiome in healthy individuals among different body sites, this study investigated within-site variability, i.e. microbiome from the same body site, incorporating different study subjects: healthy and diseased. HMP concludes that the separation is significant, yet not definitive ([78]). This is an important aspect to consider before interpreting results being part of this thesis. PCoA plots are used for visualization of microbiome diversity, further statistical tests are explored in subsequent sections.

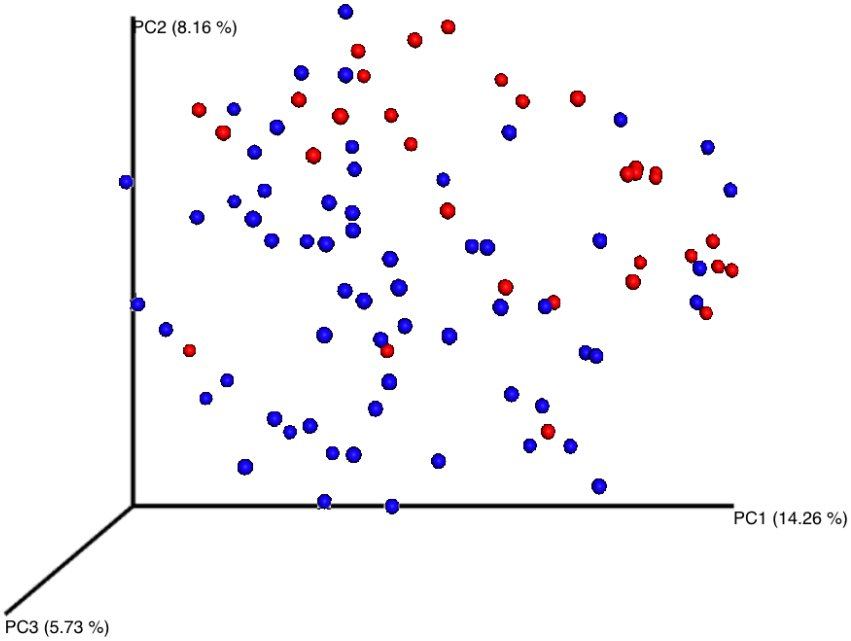


Figure 16.7: Three dimensional PCoA plot of **unweighted** `UniFrac` for **Healthy Controls (HCs)** and **SLE** subjects.

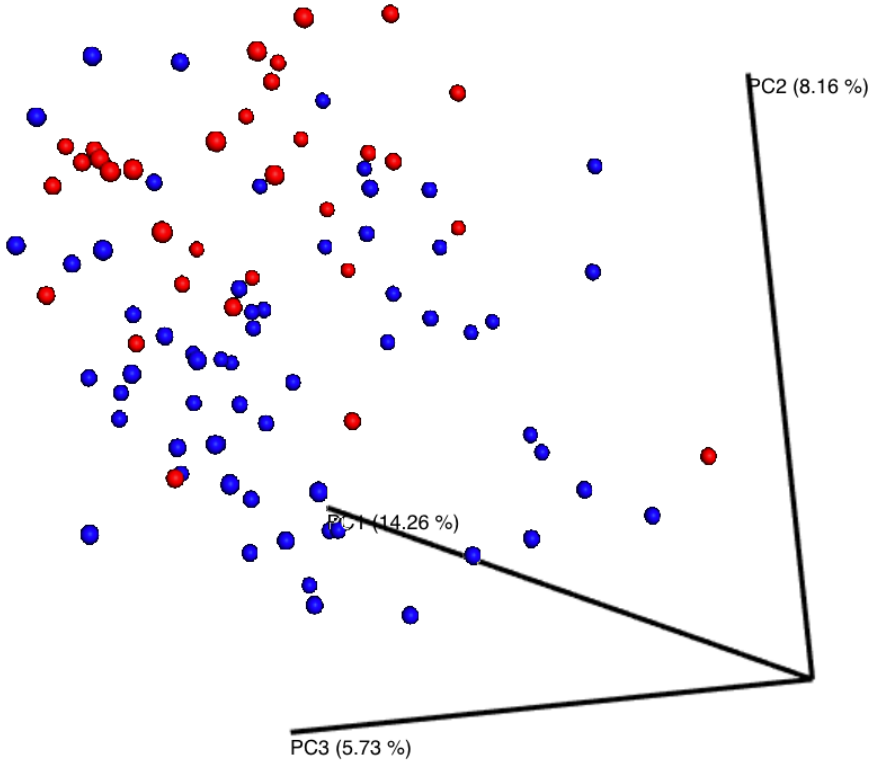


Figure 16.8: Three dimensional PCoA plot of **unweighted** `UniFrac` for **Healthy Controls (HCs)** and **SLE** subjects. Different angle compared to figure 16.7.

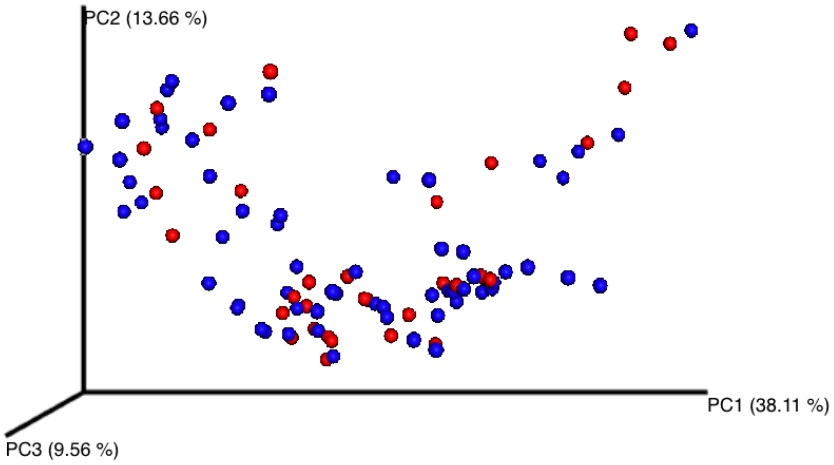


Figure 16.9: Three dimensional PCoA plot of **weighted** `UniFrac` for **Healthy Controls (HCs)** and **SLE** subjects.

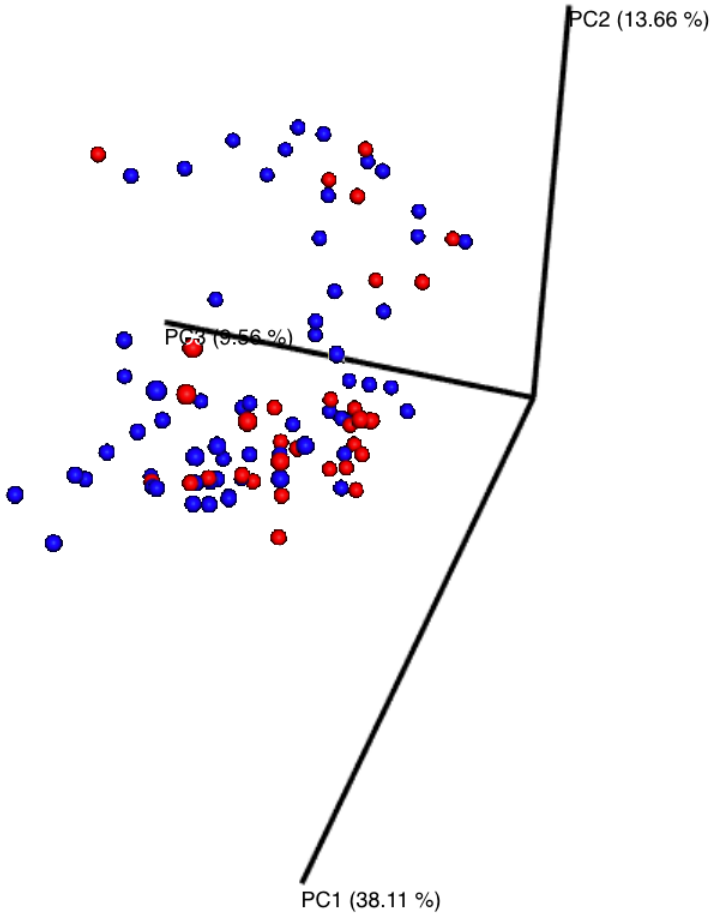


Figure 16.10: Three dimensional PCoA plot of **weighted** `UniFrac` for **Healthy Controls (HCs)** and **SLE** subjects. Different angle compared to figure 16.9.

16.3.2 Distance boxplots

Distance comparison plots, plots representing mean UniFrac (weighted and unweighted) values for all samples, are shown on figures 16.11 (page 135) and 16.12 (page 136). The first boxplot on each figure contains the distances within HC samples and SLE samples. "*Status*" is a metadata field describing whether a sample comes from Healthy Control (HC) or SLE subjects. Correspondingly, the second boxplot contains the distances between HC and SLE samples. The next two boxplots represent the individual within-distances separately for HCs and SLE subjects. The final boxplot on each mentioned figure represents the individual between-distances. Due to the nature of this study there are only two possible states for "*Status*" field (i.e. HC or SLE); for this reason the *all between status* is the same as the last boxplot - *Control vs. SLE*.

Multiple Student's two-sample t-test for every pair of boxplots was calculated, to determine if they were significantly different from each other. The non-parametric p-values were calculated using 999 Monte Carlo permutations. Since multiple comparisons are performed a Bonferroni correction for multiple comparisons was used. Results from above-mentioned comparisons for unweighted and weighted `UniFrac` are shown on Tables 16.2 (page 133) 16.3 (page 134) and respectively .

Considering comparisons of unweighted `UniFrac` between *all-within* and *all-between* samples the q value (p value after Bonferroni FDR correction) is $q = 0.01$, which is statistically significant. However, taking into account the relative abundances of microbial taxa, the same comparison looses its statistical significance $q = 1.00$. The unweighted `UniFrac` metric doesn't take into account relative abundance, thus increasing the importance of small-abundance lineages. For data in this experiment, this conveys that not accounting for abundance, those two compositions have statistically different phylogenetic compositions. However, if weighted UniFrac is taken into account, placing importance on more abundant lineages, HCs and SLE patients have similar broad phylogenetic compositions.

Table 16.2: The results of multiple Student's two-sample t-tests, comparing every pair of boxplots for **unweighted** `UniFrac` metric. First yellow row corresponds to comparison of interest: two distributions of distances are significantly different even after FDR correction for nonparametric test.

Group 1	Group 2	t statistic	Parametric p-value	Parametric p-value (Bonferroni corrected)	Nonparametric p-value	Nonparametric p-value (Bonferroni corrected)
All within Status	*All between Status*	-7.327142122	2.80E-13	2.80E-12	0.001	0.01
All within Status	*Control vs. Control*	1.207912825	0.227182085	1	0.218	1
All within Status	*SLE vs. SLE*	-0.545904298	0.585161038	1	0.587	1
All within Status	*Control vs. SLE*	-7.327142122	2.80E-13	2.80E-12	0.001	0.01
All between Status	*Control vs. Control*	5.408467781	6.97E-08	6.97E-07	0.001	0.01
All between Status	*SLE vs. SLE*	6.5407274	6.94E-11	6.94E-10	0.001	0.01
All between Status	*Control vs. SLE*	0	1	1	1	1
Control vs. Control	*SLE vs. SLE*	-1.546556778	0.122106313	1	0.127	1
Control vs. Control	*Control vs. SLE*	-5.408467781	6.97E-08	6.97E-07	0.001	0.01
SLE vs. SLE	*Control vs. SLE*	-6.5407274	6.94E-11	6.94E-10	0.001	0.01

Table 16.3: The results of multiple Student's two-sample t-tests, comparing every pair of boxplots for **weighted** `UniFrac` metric. First yellow row corresponds to comparison of interest: two distributions of distances are **not** significantly different after FDR correction for non-parametric test, and even before this correction.

Group 1	Group 2	t statistic	Parametric p-value	Parametric p-value (Bonferroni corrected)	Nonparametric p-value	Nonparametric p-value (Bonferroni corrected)
All within Status	All between Status	1.308864728	0.190650513	1	0.202	1
All within Status	Control vs. Control	4.013421586	6.14E-05	0.000614033	0.001	0.01
All within Status	SLE vs. SLE	-1.791343956	0.073310888	0.733108884	0.078	0.78
All within Status	Control vs. SLE	1.308864728	0.190650513	1	0.172	1
All between Status	Control vs. Control	3.097159598	0.001976017	0.019760173	0.001	0.01
All between Status	SLE vs. SLE	-2.961259549	0.003082902	0.030829016	0.003	0.03
All between Status	Control vs. SLE	0	1	1	1	1
Control vs. Control	SLE vs. SLE	-5.119372812	3.32E-07	3.32E-06	0.001	0.01
Control vs. Control	Control vs. SLE	-3.097159598	0.001976017	0.019760173	0.002	0.02
SLE vs. SLE	Control vs. SLE	2.961259549	0.003082902	0.030829016	0.003	0.03

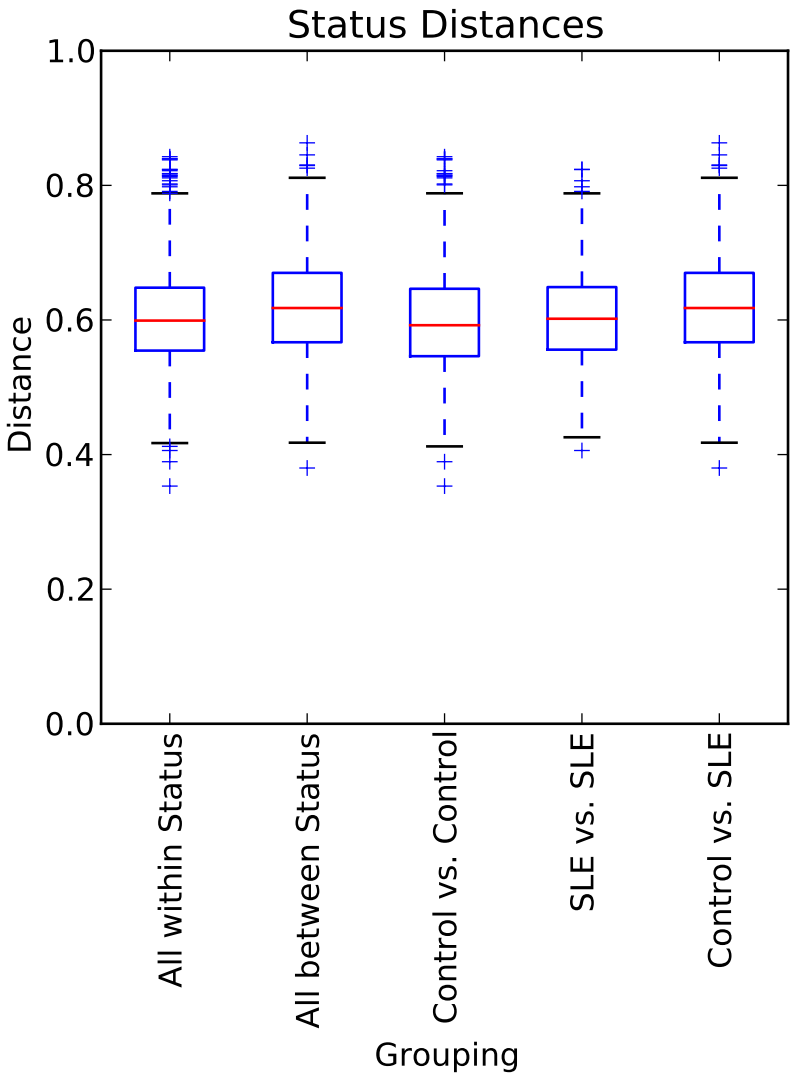


Figure 16.11: Distance comparison plots for **unweighted** `UniFrac` metric.

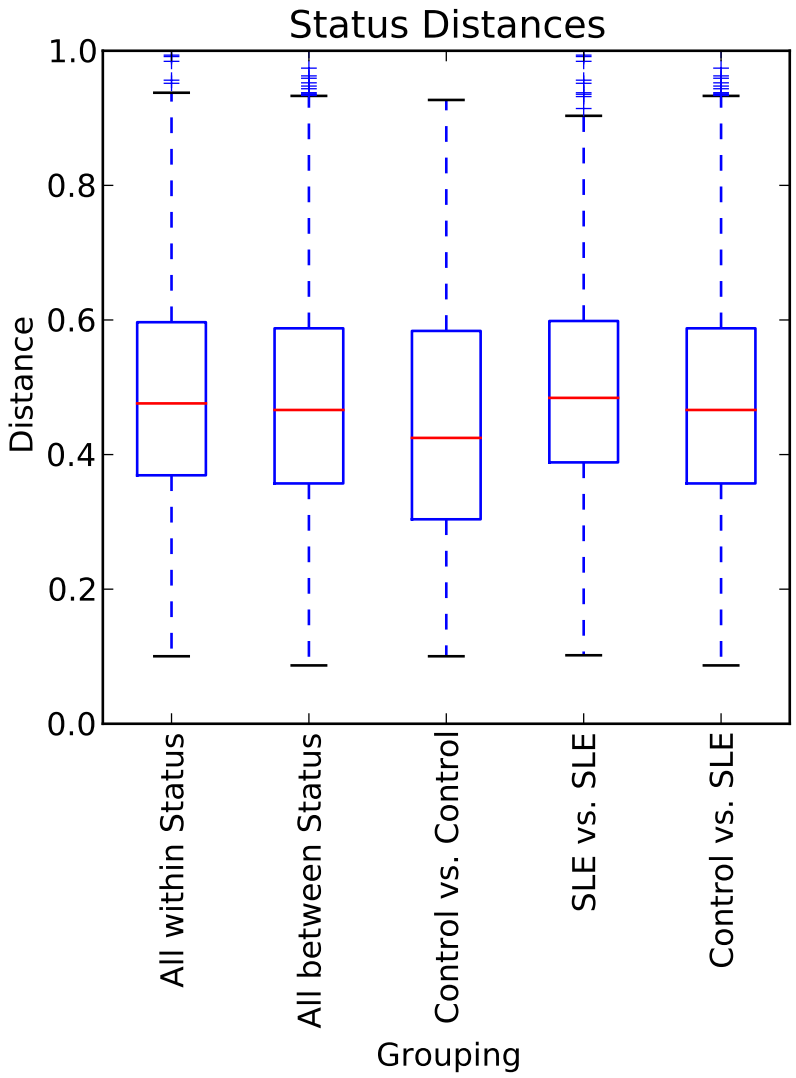


Figure 16.12: Distance comparison plots for **weighted** `UniFrac` metric.

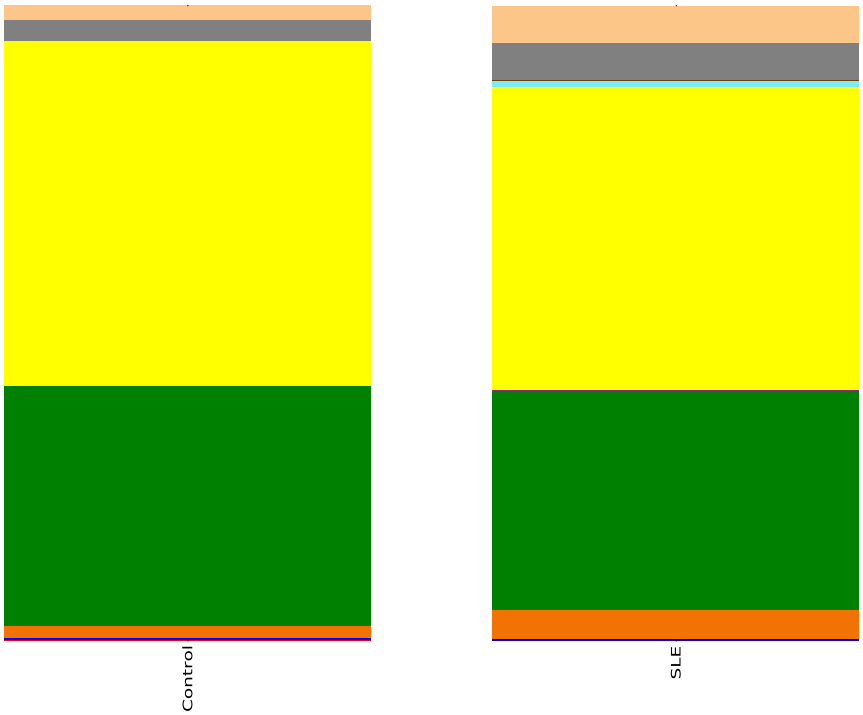
16.4 Taxonomic compositions

This section presents selected taxonomic compositions for metadata of interest - Healthy Control (HC) and Systemic Lupus Erythematosus (SLE) subjects. All of the taxonomic plots are placed on an online repository accompanying this thesis - see section 8, page 98. As detailed taxonomic compositions are hard to reproduce on a single or double page, they become uninformative; for example, detailed taxonomic compositions at genus level for each sample individually represent a matrix of 242 rows and 94 columns. Figures 16.13 and 16.14 show taxonomic composition for two groups (SLE patients and HCs) at phylum and class level respectively. The most abundant phylum included *Firmicutes* and *Bacteroidetes* (see figure 16.13). Altogether they account for up to 92% composition in healthy subjects and 82% composition for SLE patients. The most abundant classes (figure 16.14) included *Clostridia* from *Firmicutes* phylum and *Bacteroidia* from *Bacteroidetes* phylum. For healthy controls, they represent approximately 86% of total taxa observed (including bacteria and low abundance archaea), and 75% for SLE patients respectively. All other bacterial or archeal classes fall into abundances below 2%. This is consistent with Human Microbiome Project (HMP) findings describing human microbiota as comprised of a relatively small number of high-abundance taxa, followed by a long-tail of low abundance taxa [78].

Plotting taxonomic composition at deeper levels of phylogenetic tree would yield impractical, and uninterpretable results. This approach would also not yield any biological relevance considering the aim of this study. One valid approach is to statistically compare mean relative abundances for HCs and SLE patients in order to determine whether certain taxonomic ranks (phyla, classes, orders, families, genuses or species) are enriched or depleted in SLE patients.

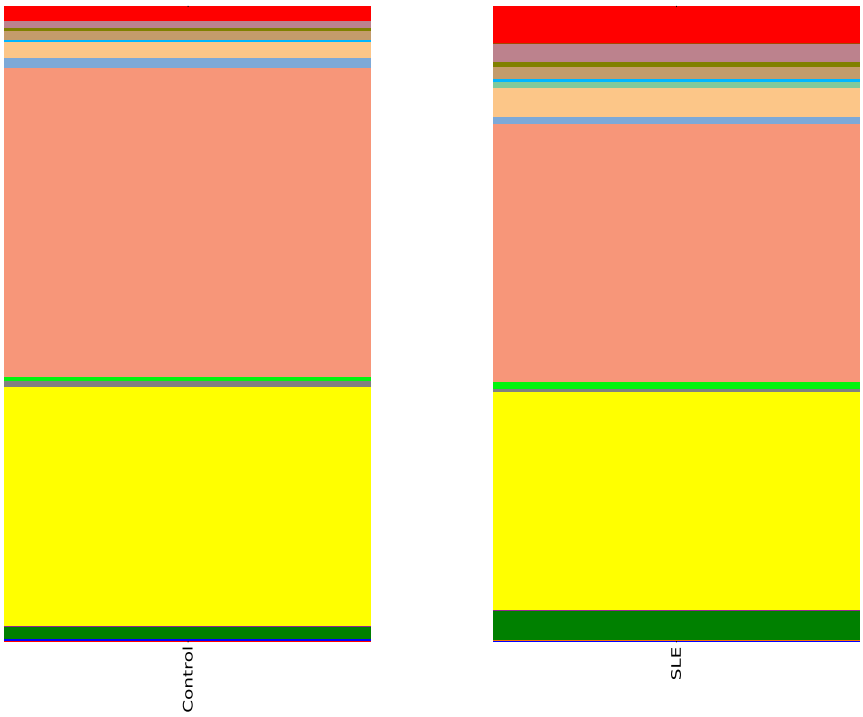
In order to analyze abundance profiles statistical, analysis software dedicated for metagenomic profiles called `STAMP` was used [169]. For a brief description of the statistical package please refer to section 7.4 on page 94. OTU table with taxonomy annotations (refer to pipeline diagram 15.2 on page 122) was converted to `STAMP` specific format through auxiliary scripts available as part of the package published in Janurary 2017 [76]. This taxonomic profile of relative abundances expressed for each sample, together with metadata mapping file constituted the input on which inferences were drawn. Two group Welch's t-test was used to compare relative abundances. This test is a variation of Student's t-test, that does not assume two groups having equal variances [185]. Two-sided confidence intervals at 0.95 are calculated by inverting Welch's t-test, and finally multiple comparison correction is performed with Storey FDR method. This FDR method is a relatively new method for controlling false discoveries. It is reportedly more robust than Benjamini-Hochberg approach [142], [198].

Figure 16.15 (page 141) identifies statistically significant more abundant taxa in SLE patients. This bacteria taxa belongs to one phylogenetic branch: *Verrucomicrobia* phylum (p) [$p = 0.011$], *Verrucomicrobiae* class (c) [$p = 0.040$], *Verrucomicrobiales* order (o) [$p = 0.032$] and *Verrucomicrobiaceae* family (f) [$p = 0.048$]. Under further inspection the only identified taxa from analyzed samples belonging to this particular branch comes from *Akkermansia* genus, and further *muciniphila* species



Legend	Taxonomy	Total %	Control %	SLE %
	d:Archaea;p:Euryarchaeota	0.1%	0.1%	0.1%
	d:Bacteria;Other	0.3%	0.4%	0.2%
	d:Bacteria;p:Actinobacteria	3.3%	1.9%	4.7%
	d:Bacteria;p:Bacteroidetes	36.2%	37.8%	34.5%
	d:Bacteria;p:Cyanobacteria/Chloroplast	0.0%	0.0%	0.0%
	d:Bacteria;p:Firmicutes	51.0%	54.2%	47.8%
	d:Bacteria;p:Fusobacteria	0.6%	0.2%	1.0%
	d:Bacteria;p:Gemmatimonadetes	0.0%	0.0%	0.0%
	d:Bacteria;p:Lentisphaerae	0.0%	0.0%	0.0%
	d:Bacteria;p:Parcubacteria	0.0%	0.0%	0.0%
	d:Bacteria;p:Proteobacteria	4.6%	3.3%	6.0%
	d:Bacteria;p:Spirochaetes	0.0%	0.0%	0.0%
	d:Bacteria;p:Synergistetes	0.0%	0.0%	0.0%
	d:Bacteria;p:Tenericutes	0.0%	0.0%	0.0%
	d:Bacteria;p:Verrucomicrobia	4.0%	2.2%	5.8%

Figure 16.13: Taxonomic compositions. Bar charts qualitatively describe microbiota composition at **phylum** level grouped for HCs and SLE patients; its height corresponds to relative abundance of a particular phylum. Legend describes mean relative abundance of each taxa for a particular group with appropriate name and corresponding color code.



Legend	Taxonomy	Total %	Control %	SLE %
	d:Archaea;p:Euryarchaeota;c:Methanobacteria	0.1%	0.1%	0.1%
	d:Bacteria;Other;Other	0.3%	0.4%	0.2%
	d:Bacteria;p:Actinobacteria;Other	0.0%	0.0%	0.0%
	d:Bacteria;p:Actinobacteria;c:Actinobacteria	3.3%	1.9%	4.7%
	d:Bacteria;p:Bacteroidetes;Other	0.1%	0.1%	0.1%
	d:Bacteria;p:Bacteroidetes;c:Bacteroidia	36.1%	37.7%	34.4%
	d:Bacteria;p:Bacteroidetes;c:Cytophagia	0.0%	0.0%	0.0%
	d:Bacteria;p:Bacteroidetes;c:Flavobacteriia	0.0%	0.0%	0.0%
	d:Bacteria;p:Bacteroidetes;c:Sphingobacteriia	0.0%	0.0%	0.0%
	d:Bacteria;p:Cyanobacteria/Chloroplast;c:Chloroplast	0.0%	0.0%	0.0%
	d:Bacteria;p:Firmicutes;Other	0.6%	0.9%	0.3%
	d:Bacteria;p:Firmicutes;c:Bacilli	0.9%	0.6%	1.1%
	d:Bacteria;p:Firmicutes;c:Clostridia	44.6%	48.6%	40.6%
	d:Bacteria;p:Firmicutes;c:Erysipelotrichia	1.4%	1.7%	1.1%
	d:Bacteria;p:Firmicutes;c:Negativicutes	3.6%	2.4%	4.7%
	d:Bacteria;p:Fusobacteria;c:Fusobacteriia	0.6%	0.2%	1.0%
	d:Bacteria;p:Gemmatimonadetes;c:Gemmatimonadetes	0.0%	0.0%	0.0%
	d:Bacteria;p:Lentisphaerae;c:Lentisphaeria	0.0%	0.0%	0.0%
	d:Bacteria;p:Parcubacteria;Other	0.0%	0.0%	0.0%
	d:Bacteria;p:Proteobacteria;Other	0.0%	0.0%	0.0%
	d:Bacteria;p:Proteobacteria;c:Alphaproteobacteria	0.3%	0.2%	0.4%
	d:Bacteria;p:Proteobacteria;c:Betaproteobacteria	1.7%	1.4%	2.0%
	d:Bacteria;p:Proteobacteria;c:Deltaproteobacteria	0.6%	0.5%	0.7%
	d:Bacteria;p:Proteobacteria;c:Epsilonproteobacteria	0.0%	0.0%	0.0%
	d:Bacteria;p:Proteobacteria;c:Gammaproteobacteria	2.0%	1.1%	2.9%
	d:Bacteria;p:Spirochaetes;c:Spirochaetia	0.0%	0.0%	0.0%
	d:Bacteria;p:Synergistetes;c:Synergistia	0.0%	0.0%	0.0%
	d:Bacteria;p:Tenericutes;c:Mollicutes	0.0%	0.0%	0.0%
	d:Bacteria;p:Verrucomicrobia;c:Opitutae	0.0%	0.0%	0.0%
	d:Bacteria;p:Verrucomicrobia;c:Verrucomicrobiae	4.0%	2.2%	5.8%

Figure 16.14: Taxonomic compositions. Bar charts qualitatively describe microbiota composition at **class** level grouped for HCs and SLE patients; its height corresponds to relative abundance of a particular class. Legend describes mean relative abundance of each taxa for a particular group with appropriate name and corresponding color code.

and this explains the fact that throughout the phylogenetic branch, from phylum up until family taxonomic rank, this bacteria exhibits the same proportion of sequences, i.e. compositions presented on histograms on Figure 16.15 (page 141).

However, conducted test doesn't reflect a significance between HCs and SLE patients for *Akkermansia* genus ($p = 0.106$). However, the specificity and sensitivity for this type of analyses (16S rRNA gene survey) deteriorates at deeper taxonomic ranks, such as *genus* or *species* - see figure 6.12 on page 77. According to figure 6.12, the nature of the experiment could reach approximately 83% accuracy at taxonomic assignment for genus level when using 515f-806r primers (515f-806r primers - see section 13.2 on page 110). Consequently species level would have smaller accuracy. Some researchers even argue that species should not be considered for 16S rRNA gene surveys - for more discussion refer to 16S methodology overview, section 3.1 on page 38.

Figure 16.16 (page 142) also identifies bacterial taxa belonging to related order, family and genus. *Alteromonadales* order and *Shewanellaceae* family have been identified to be statistically more abundant within SLE patients. This is reflected with $p = 0.015$ (for order) and $p = 0.022$ (for family). P-values reported are considered after FDR correction. For *Shewanella* genus reported p-value is $p = 0.048$ (FDR corrected). The latter result should be treated with caution for the reasons discussed above - genus predictions have a certain, questionable confidence threshold that might influence downstream comparative analyses. The association at order and phylum level are stronger and more confident.

Finally figure 16.17 (page 143) plots proportion of sequences assigned to *Atopobium* genus. Considering even sampling for each sample in the study, this result suggests that bacteria coming from this genus are significantly more abundant in SLE patients. However, the same restrictions apply here for genus-level predictions.

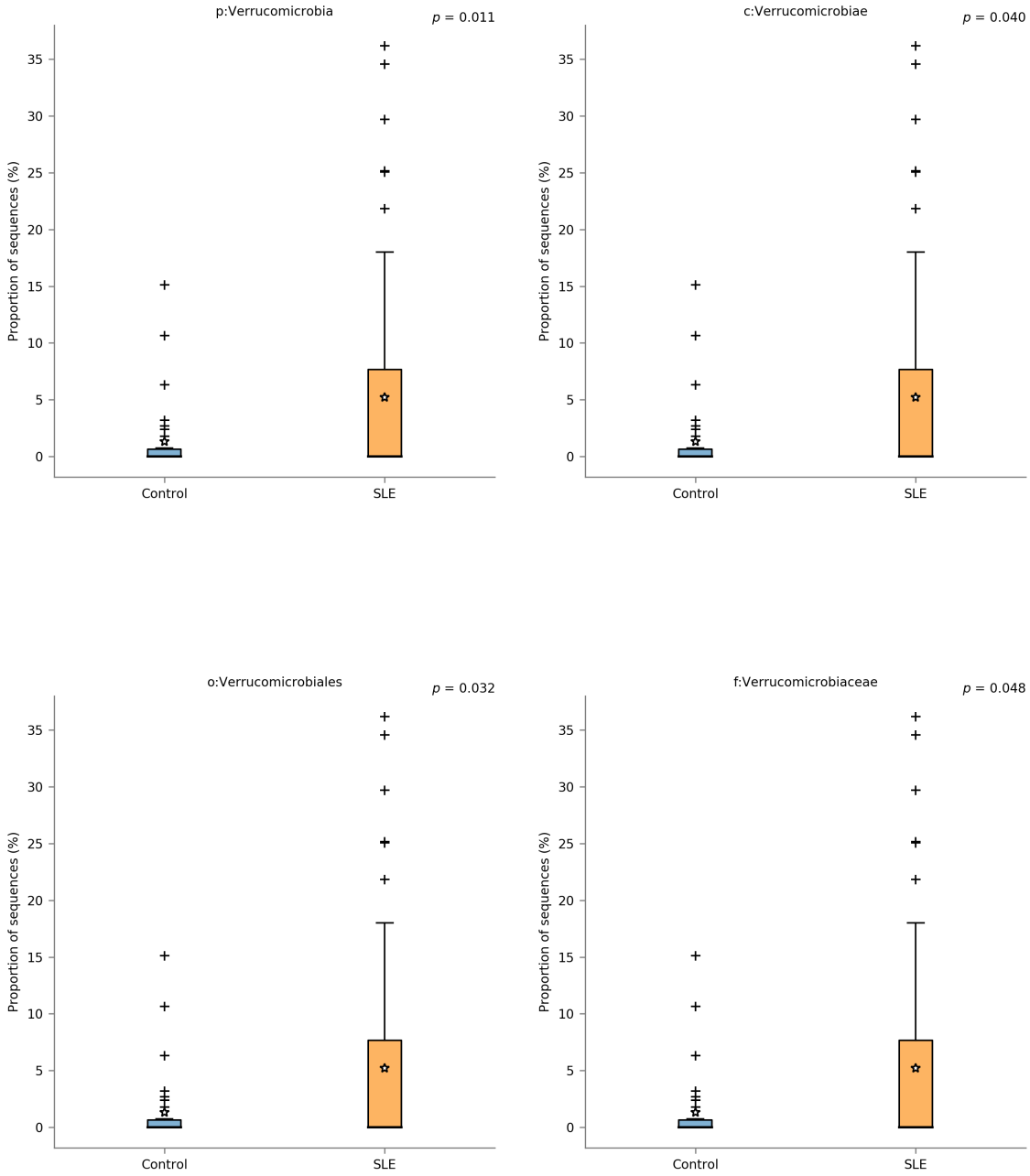


Figure 16.15: Statistically more abundant bacterial taxa belonging to one phylogenetic branch: *Verrucomicrobia* phylum (p), *Verrucomicrobiae* class (c), *Verrucomicrobiales* order (o) and *Verrucomicrobiaceae* family (f). Proportion of sequences on the y-axis are expressed in percentages, reflecting how many sequences were assigned to a particular taxonomic rank.

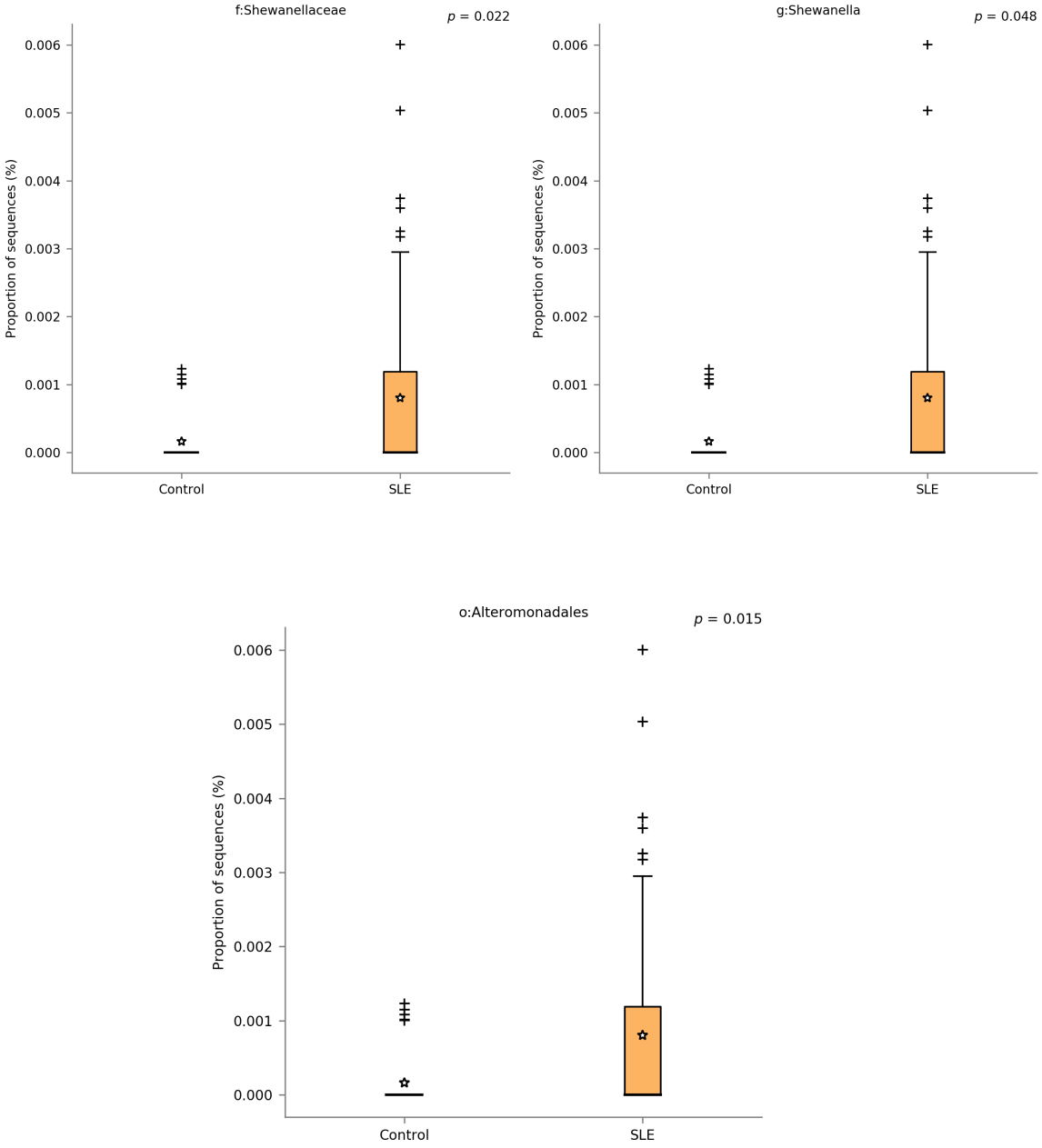


Figure 16.16: Statistically more abundant bacterial taxa belonging to one fragment of phylogenetic branch: *Alteromonadales* order (o) *Shewanellaceae* family (f) and *Shewanella* genus (g). Proportion of sequences on the y-axis are expressed in percentages, reflecting how many sequences were assigned to a particular taxonomic rank.

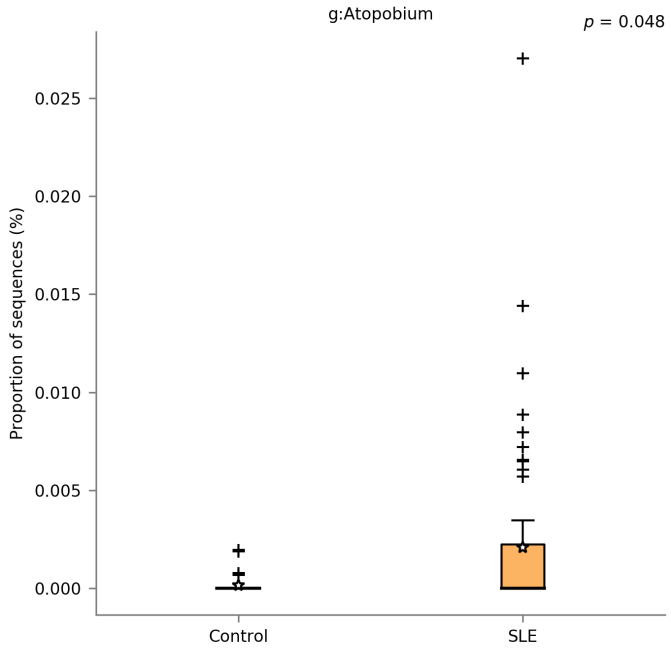


Figure 16.17: Microbes from *Atopobium* genus are statistically more abundant ($p = 0.048$ after FDR correciton) among SLE patients.

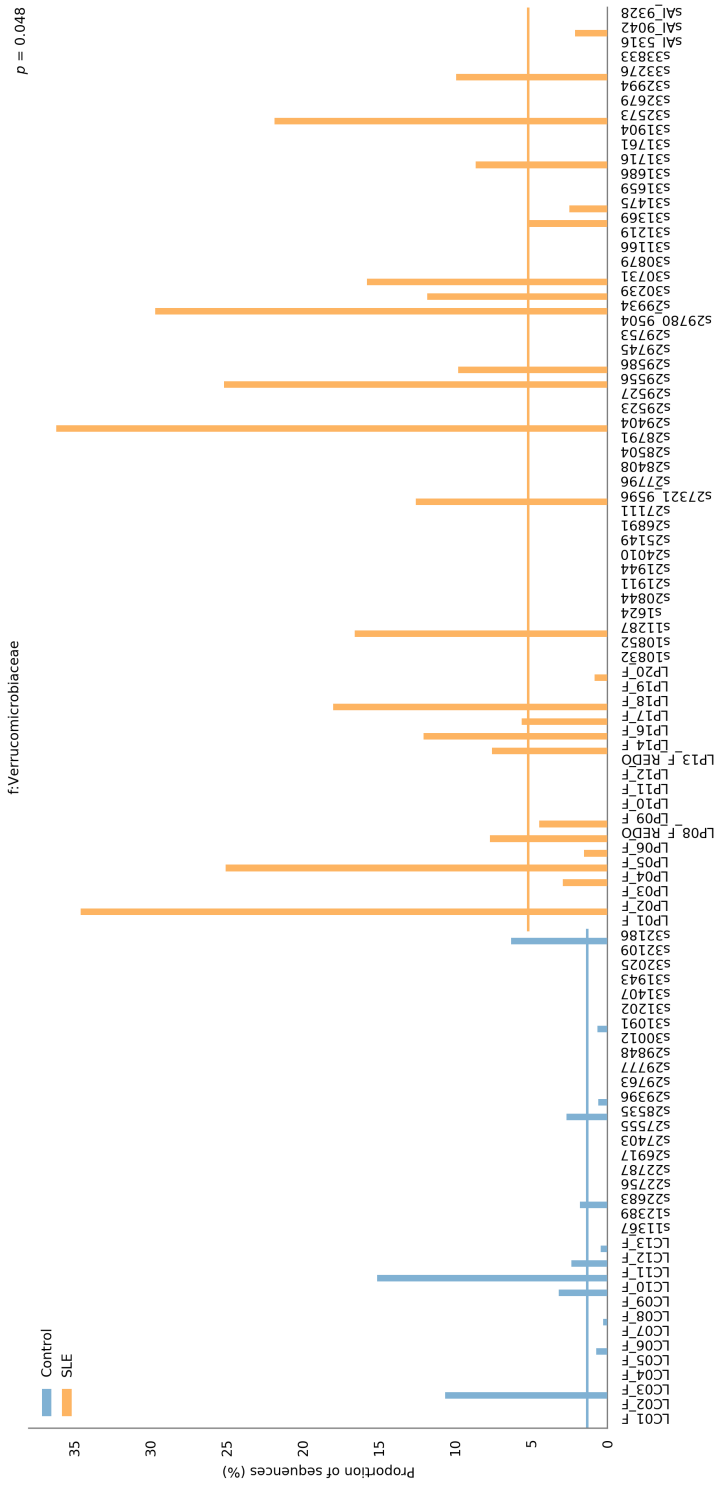


Figure 16.18: Bar plot representing abundance of *Verrucomicrobiaceae* family in each sample from this study. Horizontal lines represents average for each group.

Chapter 17

Metagenome predictions

17.1 Orthologs, modules and pathways

PICRUSt software package was used to infer a putative metagenome. For this purpose, the decontaminated and pooled sequences served as an input for metagenome predictions. Since PICRUSt relies heavily on a training dataset constructed solely (although not limited to) on GreenGenes database version 13_5, OTUs were picked in a closed-reference manner - see pipeline overview section 15.2, page 118, and figure 15.2 on page 122.

Final output consisted of predicted matrix, where each column represented a sample, and each row represented KEGG Ortholog (KO) (Kyoto Encyclopedia of Genes and Genomes (KEGG)). Through PICRUSt functionality KOs were collapsed into modules and pathways. Output BIOM table was converted to *tsv* format through BIOM package functionality. This *tsv* file together with mapping metadata file consisted an input for downstream analyses with STAMP software package.

PICRUSt predicted a total 6909 KOs, that were collapsed into 41 modules and 328 KEGG pathways. No module was identified to be significantly enriched or depleted in SLE patients - see figure 17.1. The feature "*metabolism*" having $p = 0.089$ had the lowest, yet still not significant value. It is worth mentioning that modules are very broad, manually drawn categories. For example, the mentioned *metabolism* category is comprised of many different features: carbohydrate, energy, lipid, nucleotide, amino-acid metabolisms, glycan biosynthesis and metabolism, or cofactors and vitamins metabolism, and even biosythesis of secondary metabolites or xenobiotics biodegradation and metabolism. For this reason the metagenome at this level is too general for informative interpretations.

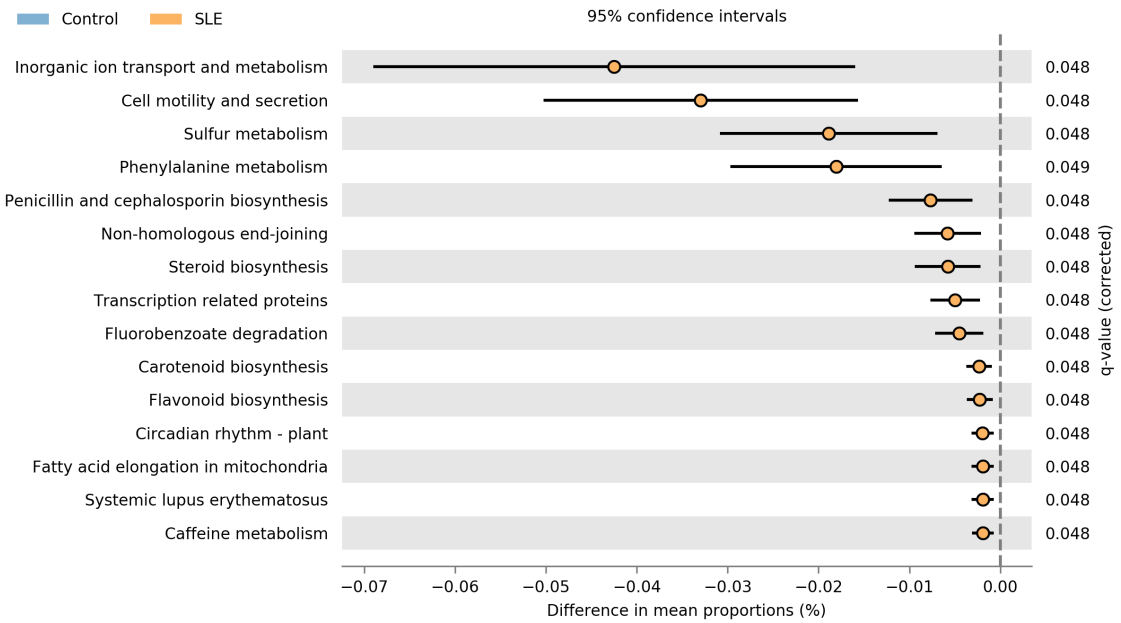
Collapsing the metagenome at more detailed levels allowed comparison of pathways available in the KEGG database. Figure 17.2 shows significantly enriched pathways in SLE patients compared to HCs. Among 15 significantly enriched pathways that have been identified, there is a pathway of a particular interest - *Systemic Lupus Erythematosus*. Significance of other pathways, and potential explanation is discussed in section 17.5 on page 159. Figure 17.3 (page 149) shows a box-plot for the SLE-pathway abundance for HCs and SLE subjects, while figure 17.4 on page 150 is a bar-plot that shows pathway abundance for each sample separately.

Although the SLE-pathway is significantly more abundant in SLE patients, from

above-mentioned bar-plot it is clear that this distinction is ambiguous, since this pathway is not present for all samples from SLE patients. It is worth mentioning here that PICRUSt creators themselves advise to treat metagenome predictions as "*suggestive only*" [145], since the accuracy spans from the range 60% to 90% for human gut microbiome, and 80% is approximately the expected accuracy value (see figure 7.3 page 94).

Feature	Diff. between means	p-value	Corrected p-value
Metabolism	-0.101	3.33e-3	0.089
Xenobiotics Biodegradation and Metabolism	-0.075	0.019	0.130
Lipid Metabolism	-0.094	0.015	0.130
Excretory System	-0.006	0.020	0.130
Environmental Adaptation	0.008	0.039	0.210
Translation	0.231	0.056	0.236
Replication and Repair	0.300	0.071	0.236
Poorly Characterized	-0.068	0.068	0.236
Signaling Molecules and Interaction	-0.011	0.131	0.247
Signal Transduction	-0.089	0.137	0.247
Nucleotide Metabolism	0.127	0.139	0.247
Neurodegenerative Diseases	-0.010	0.135	0.247
Immune System	0.005	0.134	0.247
Cellular Processes and Signaling	-0.097	0.137	0.247
Cell Motility	0.210	0.088	0.247
Enzyme Families	0.036	0.169	0.279
Carbohydrate Metabolism	-0.192	0.178	0.279
Metabolic Diseases	0.003	0.229	0.339
Transport and Catabolism	-0.022	0.264	0.365
Cell Growth and Death	0.013	0.274	0.365
Transcription	0.042	0.349	0.397
Metabolism of Cofactors and Vitamins	0.046	0.372	0.397
Infectious Diseases	-0.011	0.319	0.397
Energy Metabolism	0.058	0.348	0.397
Digestive System	0.008	0.370	0.397
Glycan Biosynthesis and Metabolism	-0.104	0.405	0.416
Amino Acid Metabolism	-0.052	0.444	0.439
Nervous System	0.002	0.544	0.518
Metabolism of Terpenoids and Polyketides	0.017	0.569	0.520
Metabolism of Other Amino Acids	-0.011	0.605	0.520
Genetic Information Processing	-0.012	0.598	0.520
Biosynthesis of Other Secondary Metabolites	-0.009	0.688	0.573
Membrane Transport	-0.148	0.726	0.586
Sensory System	0.000	1.000	0.650
Immune System Diseases	-0.000	0.849	0.650
Folding, Sorting and Degradation	0.006	0.872	0.650
Endocrine System	0.001	0.945	0.650
Circulatory System	-0.000	0.928	0.650
Cell Communication	0.000	1.000	0.650
Cardiovascular Diseases	-0.000	0.988	0.650
Cancers	-0.000	0.991	0.650

Figure 17.1: Comparative analysis of metagenome collapsed into 41 **modules**. No modules are statistically more abundant or depleted in SLE patients compared with Healthy Controls (HCs). Modules are sorted in ascending order for corrected p-values (Storey FDR). Welch's t-test was used to compare module abundances between study groups.



Feature	Diff. between means	p-value	Corrected p-value
Cell motility and secretion	-0.033	2.75e-4	0.048
Transcription related proteins	-0.005	4.72e-4	0.048
Fluorobenzoate degradation	-0.005	9.99e-4	0.048
Carotenoid biosynthesis	-0.002	1.26e-3	0.048
Penicillin and cephalosporin biosynthesis	-0.008	1.37e-3	0.048
Fatty acid elongation in mitochondria	-0.002	1.92e-3	0.048
Systemic lupus erythematosus	-0.002	1.94e-3	0.048
Circadian rhythm - plant	-0.002	1.95e-3	0.048
Inorganic ion transport and metabolism	-0.042	1.97e-3	0.048
Caffeine metabolism	-0.002	1.99e-3	0.048
Steroid biosynthesis	-0.006	2.07e-3	0.048
Non-homologous end-joining	-0.006	2.26e-3	0.048
Sulfur metabolism	-0.019	2.32e-3	0.048
Flavonoid biosynthesis	-0.002	2.47e-3	0.048
Phenylalanine metabolism	-0.018	2.73e-3	0.049

Figure 17.2: Significantly enriched **pathways** in SLE patients compared to Healthy Controls (HCs). First (top) sub-figure shows extended error bar plot for 95% confidence intervals and q-values (Story FDR corrected p-values). Second (bottom) sub-figure shows values for difference between means, p- and q-values.

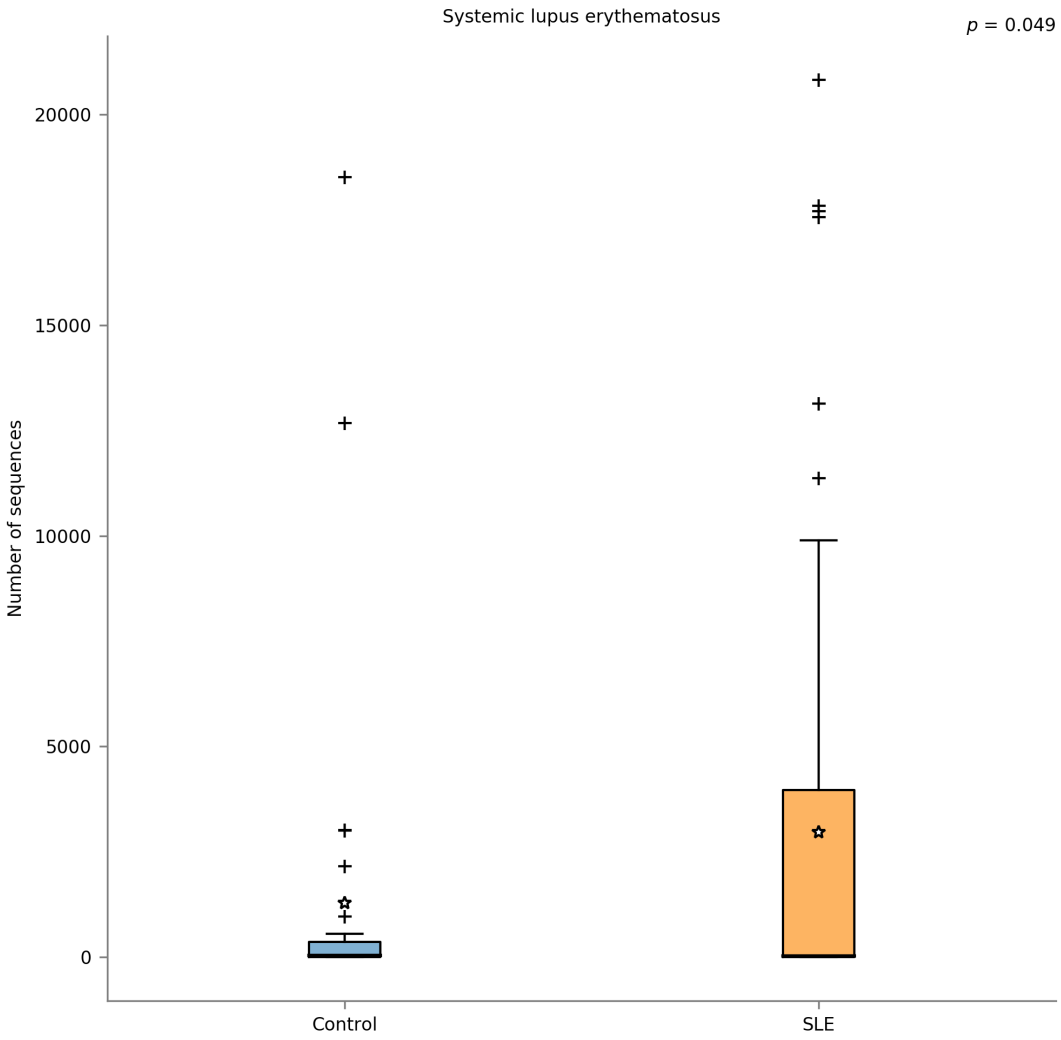


Figure 17.3: Box-plot of Systemic Lupus Erythematosus (SLE) pathway abundance for Healthy Controls (HCs) and SLE patients.

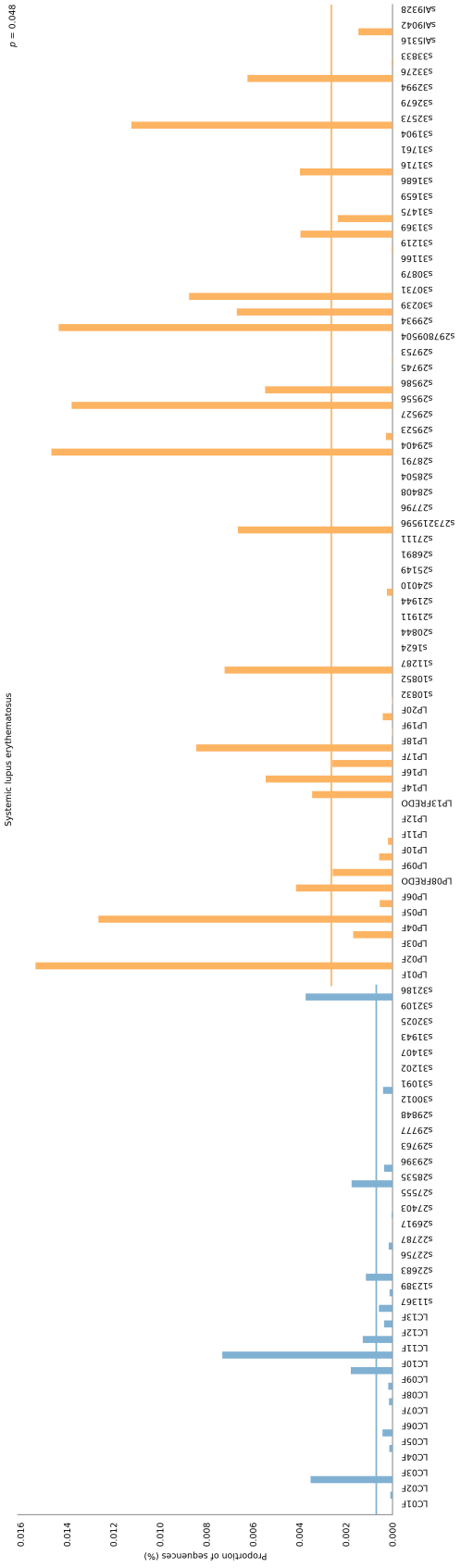


Figure 17.4: Bar-plot of Systemic Lupus Erythematosus (SLE) pathway abundance for Healthy Controls (HCs) and SLE patients.

17.2 NSTI values

One of the major factors contributing to PICRUSt accuracy for inferring putative metagenome is the availability of sequenced genomes for a particular bacteria. PICRUSt creators developed a metric called weighted Nearest Sequenced Taxon Index (NSTI). This metric summarizes in a quantitative measure the extent to which bacteria or archea in a particular sample are related to sequenced genomes. Computationally NSTI values are the "average branch length that separates each OTU in a sample from a reference bacterial genome, weighted by the abundance of that OTU in the sample" [46]. NSTI scores have been pre-calculated for common databases, including GreenGenes used for this study. Values of NSTI below 0.03 mean that microbes (its functional gene content) from a particular sample can be predicted using a related bacteria from the same species (assuming 97% similarity). For gathered datasets (see figure 17.5, page 151) NSTI values fall in the approximate range $0.0260 - 0.1718$, with average 0.090 and median 0.089. Thus the majority of microbial functional content was inferred through Ancestral State Reconstruction (ASR) algorithm through more distantly related species. NSTI values are not directly tied to PICRUSt predictions, but are used as simple estimators regarding how well a sample is represented by the reference genomes available to PICRUSt.

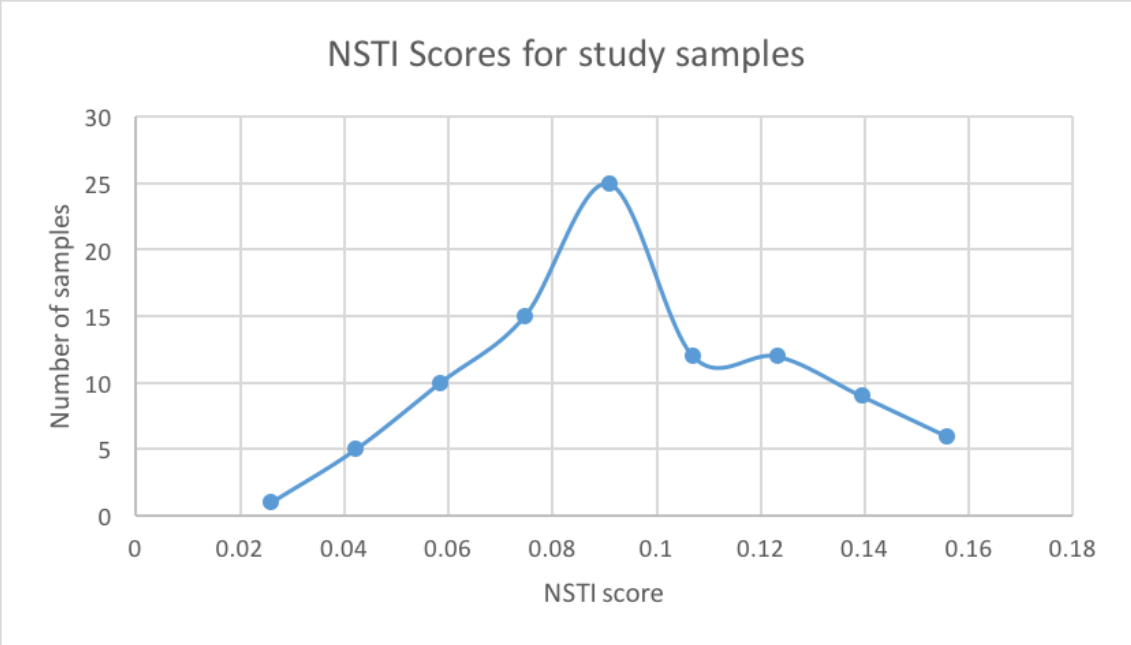


Figure 17.5: A histogram presenting NSTI values for inferred metagenomic content from input study dataset.

Table 17.1: Table of KOs constituting to SLE pathway as represented in KEGG database. Source: [116]. Bold text corresponds to three present (observed) KOs in the predicted metagenome for this study attributed to SLE pathway. Yellow background corresponds to a KO that is significantly more abundant in SLE patients. Numeric identifiers come directly from internal KEGG annotations for SLE pathway.

ID	Definition
K01319	**CTSG; cathepsin G [EC:3.4.21.20]**
K01327	ELANE; leukocyte elastase [EC:3.4.21.37]
K01330	C1R; complement component 1, r subcomponent [EC:3.4.21.41]
K01331	C1S; complement component 1, s subcomponent [EC:3.4.21.42]
K01332	C2; complement component 2 [EC:3.4.21.43]
K03156	TNF, TNFA; tumor necrosis factor superfamily, member 2
K03160	TNFRSF5, CD40; tumor necrosis factor receptor superfamily member 5
K03161	TNFSF5, CD40L, CD154; tumor necrosis factor ligand superfamily member 5
K03986	C1QA; complement C1q subcomponent subunit A
K03987	C1QB; complement C1q subcomponent subunit B
K03988	C1QG; complement C1q subcomponent subunit C
K03989	C4; complement component 4
K03990	C3; complement component 3
K03994	C5; complement component 5
K03995	C6; complement component 6
K03996	C7; complement component 7
K03997	C8A; complement component 8 subunit alpha
K03998	C8B; complement component 8 subunit beta
K03999	C8G; complement component 8 subunit gamma
K04000	C9; complement component 9
K04687	IFNG; interferon gamma
K05209	GRIN2A; glutamate receptor ionotropic, NMDA 2A
K05210	GRIN2B; glutamate receptor ionotropic, NMDA 2B
K05412	CD80; CD80 antigen
K05413	CD86; CD86 antigen
K05443	IL10, CSIF; interleukin 10
K05699	ACTN1_4; actinin alpha 1/4
K06463	FCGR3, CD16; low affinity immunoglobulin gamma Fc receptor III
K06470	CD28; CD28 antigen
K06472	FCGR2A, CD32; low affinity immunoglobulin gamma Fc receptor II-a
K06498	FCGR1A, CD64; high affinity immunoglobulin gamma Fc receptor I
K06752	MHC2; major histocompatibility complex, class II
K06856	IGH; immunoglobulin heavy chain
K10651	TRIM21, SSA1; tripartite motif-containing protein 21 [EC:2.3.2.27]
K10784	TRAV; T cell receptor alpha chain V region
K10785	TRBV; T-cell receptor beta chain V region
K11086	SNRPB, SMB; small nuclear ribonucleoprotein B and B'
K11087	SNRPD1, SMD1; small nuclear ribonucleoprotein D1
K11088	SNRPD3, SMD3; small nuclear ribonucleoprotein D3
K11089	**TROVE2, SSA2; 60 kDa SS-A/Ro ribonucleoprotein**
K11090	LA, SSB; lupus La protein
K11251	H2A; histone H2A
K11252	H2B; histone H2B
K11253	**H3; histone H3**
K11254	H4; histone H4

17.3 Metagenome contributions

SLE pathway in KEGG database [116] is comprised of 45 KOs - see table 17.1 on page 152. Estimating which KOs are contributing to the enrichment of SLE pathway, and identifying microbial taxons at specific taxonomic ranks contributing to the abundance of these orthologs is be crucial for biological significance of this study.

Firstly, predicted orthologs were scanned according to orthologs present in SLE pathway. It turns out that only 3 out of 45 orthologs were present. It is worth bearing in mind that this KEGG pathway is a human disease pathway, and many orthologs might be absent simply because the subject of the study is a microbiome. Table 17.1 (page 152) shows with bold text the three observed orthologs: `K01319`,`K11089` and `K11253`. Upon further inspection relative abundances of KOs: `K01319` and `K11253` are rounded to 0.0% for HCs and SLE patients. This leaves one KO, `K11089`, that is indeed more abundant in SLE patients - see figure 17.6 on page 154. This KO (see table 17.1, page 152) is a `TROVE2` gene, also referred to as `60 kDa SS-A Ro` ribonucleoprotein, or simply *Ro*60. Potential significance of this discovery is discussed in next section 17.4 on page 157. Comparing figures 17.3 and 17.6, there is similarity of abundance, because `K11089` ortholog is the main contributor to observed SLE pathway.

Having identified the major ortholog contributing to enriched SLE pathway in diseased subjects, it is of crucial importance to estimate certain microbial taxonomic ranks directly contributing to this ortholog abundance. Through PICRUSt functionality it is possible to pin-point all the taxa from the initial input dataset (closed-reference picked OTUs) that contribute to abundance of desired orthologs. Contributions for genus level are visualized on figure 17.7 (page 155) through auxiliary scripts in microbiome helper package [76] (`plot_metagenome_contributions.R` in *R-cran* statistic environment). Identifying microbes at lower, i.e. broader, less specific taxonomic rank, such as family level (figure 17.8, page 156), reveals the same contributory pattern. This is important, as it has been previously stated that identifying microbes at genus level yields approximately 82% accuracy, while family level yields minimum 92% expected accuracy.

Described analyses point to bacteria belonging to *Akkermania* genus from *Verrucomicrobiaceae* family. Although species identification for 16S survey is often not recommended, as taxonomic assignment might lack sensitivity and specificity, upon further inspection all the contributions to KO `K11089` were assigned to species *muciniphila*. This bacterium will be further referred to as *Akkermania muciniphila*.

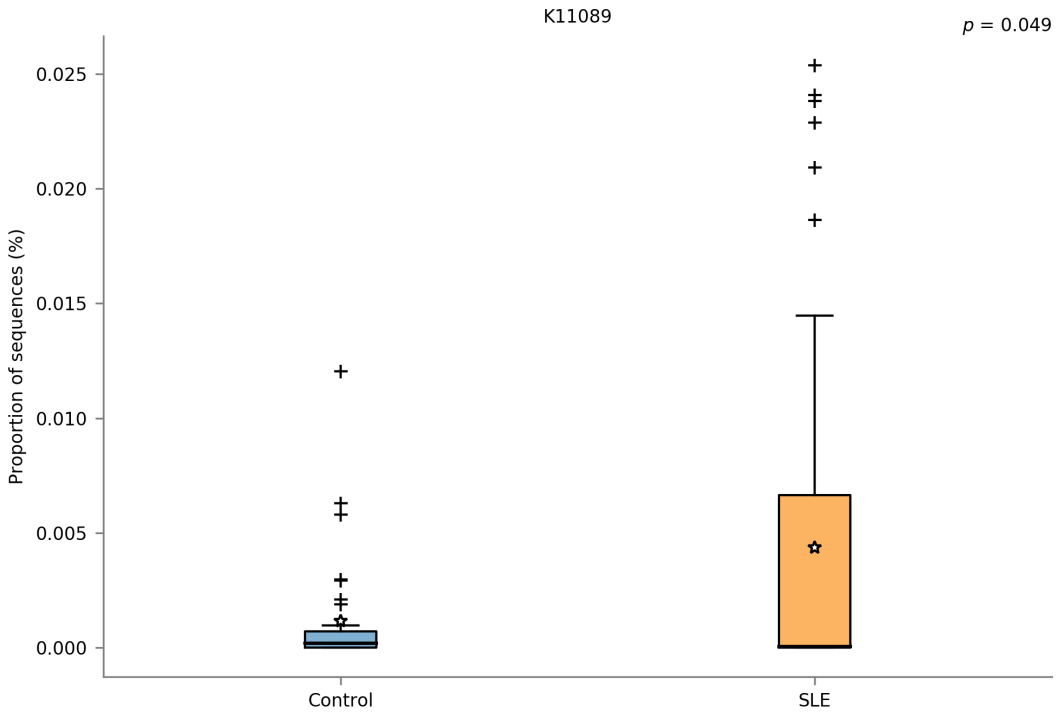


Figure 17.6: Significantly more abundant KO `K11089` in SLE patients. This orthologous group is `TROVE2` gene, also referred to as `60 kDa SS-A Ro` ribonucleoprotein, or simply `Ro60`. Significance level of $p = 0.049$ is a value after multiple correction test (Storey `FDR`) for all the orthologs present in the dataset (6909).

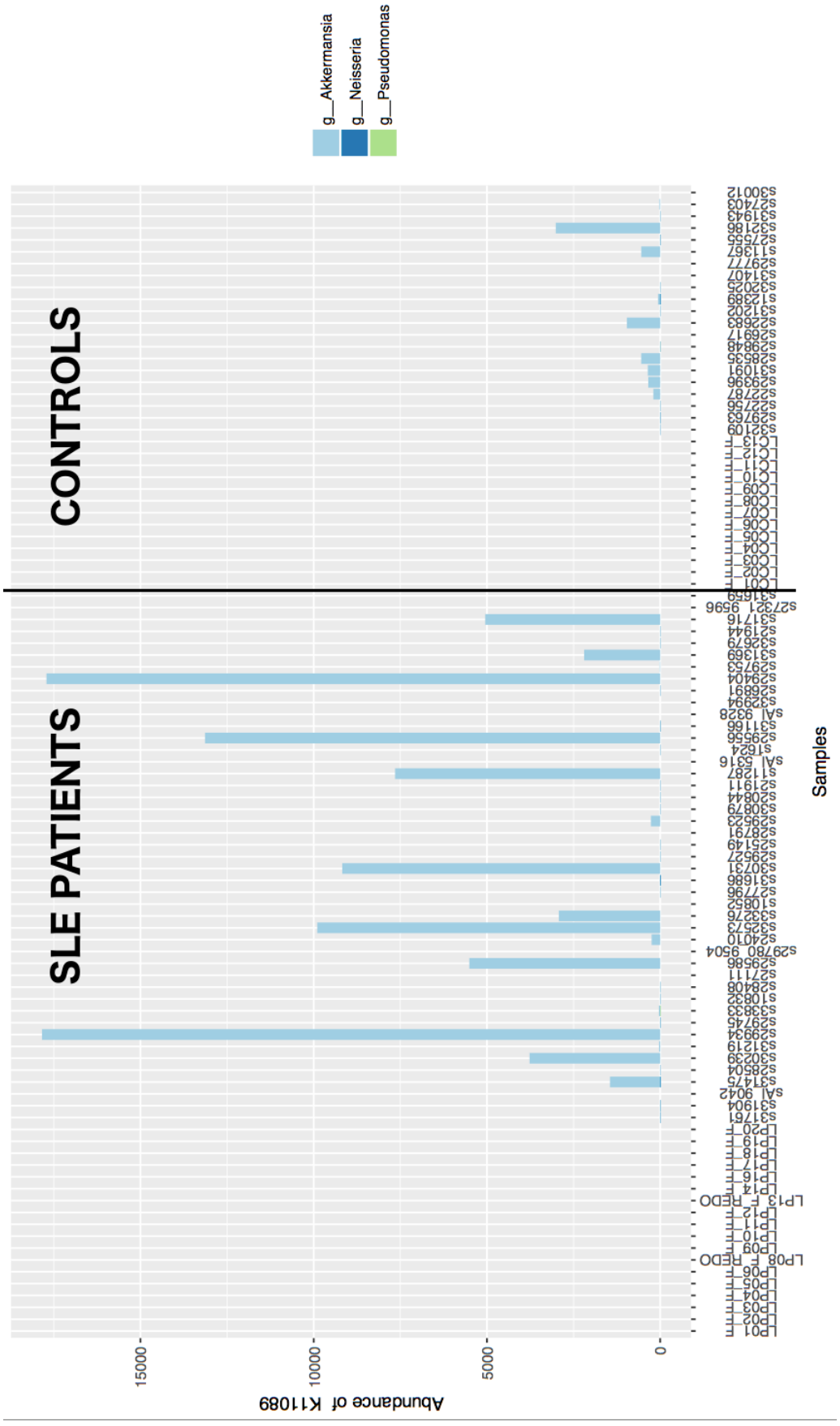


Figure 17.7: Metagenome contributions for K11089 ortholog at genus rank. Plot representing all of the samples gathered for experiment, ordered by SLE and healthy status. Plot points to three microbial genuses: *Akkermania*, *Neisseria* and *Pseudomonas*, among which *Akkermansia* has the strongest contributions; others are not visible on the plot.

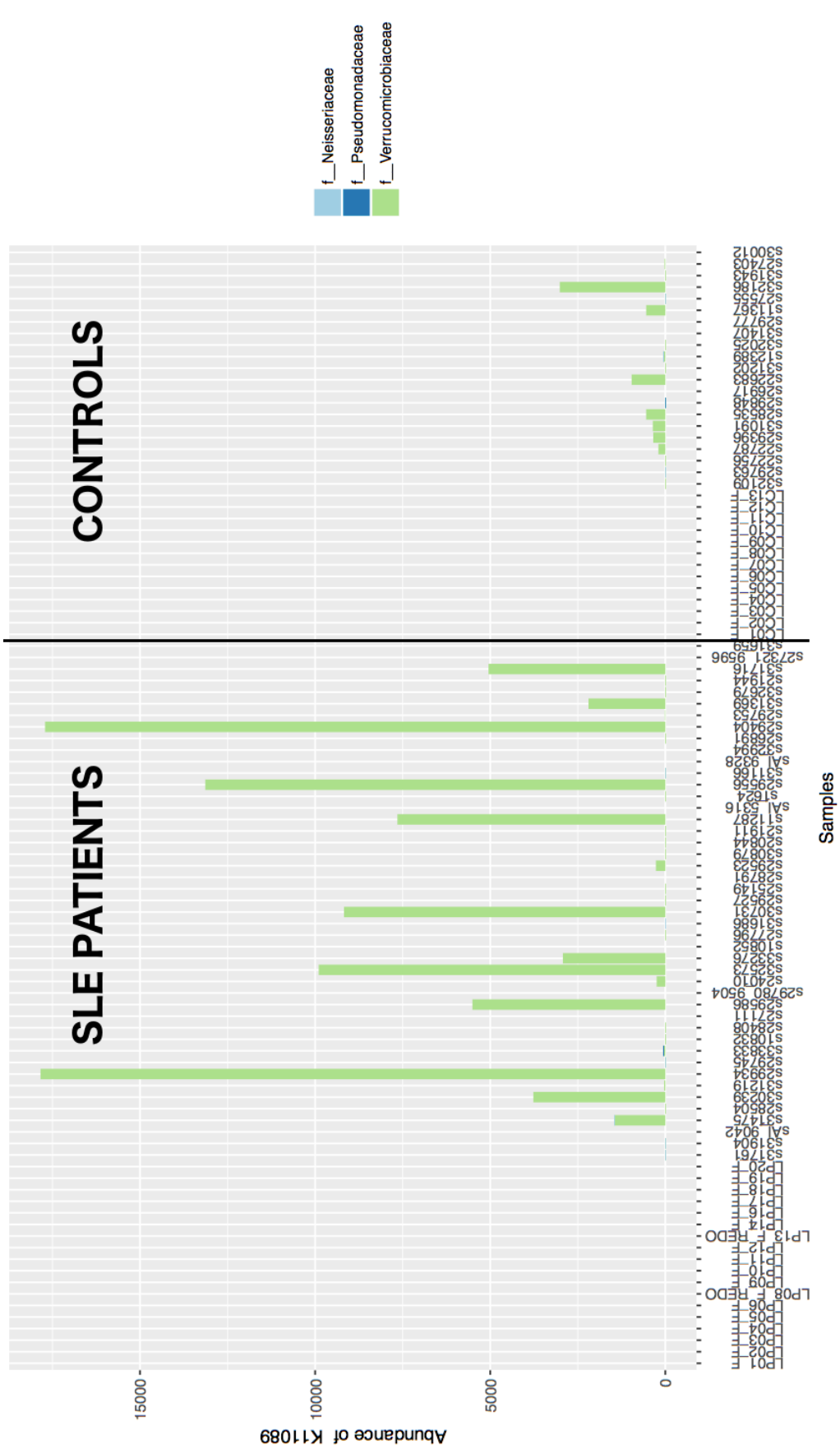


Figure 17.8: Metagenome contributions for K11089 ortholog at family rank. Plot representing all of the samples gathered for experiment, ordered by -SLE and healthy status. Plot points to three microbial families: *Verrucomicrobiaceae*, *Neisseriaceae* and *Pseudomonadaceae*, among which *Verrucomicrobiaceae* has the strongest contributions; others are not visible on the plot.

17.4 *Ro*60 and *Akkermansia muciniphila*

17.4.1 Rationale

It is important to notice that taxonomy composition analysis (section 16.4, page 137) pointed to a phylogenetic branch of *Verrucomicrobia* phylum, *Verrucomicrobiae* class, *Verrucomicrobiales* order and *Verrucomicrobiaceae* family as statistically more abundant in SLE patients. Metagenomic predictions also pointed to this microbial branch, but as having profound (statistically significant) contribution to SLE pathway through `TROVE2` ortholog (also referred to as *Ro*60) in SLE patients compared to HCs. It is crucial to assign a biological relevance to a statistical significance from previous sections. What might be the role of *Ro*60 ortholog present is certain bacteria, highly associated in our study with *Akkermansia muciniphila* from *Verrucomicrobiaceae* family? This section provides a brief literature overview for obtained results, and proposes a few possible explainations.

17.4.2 *Akkermansia muciniphila*

Akkermansia muciniphila is a relatively recently discovered bacterium (2004) [82], and is the only cultivated intestinal representative of the *Verrucomicrobia* phylum [83]. Its role is now extensively being studied in obesity and diabetes [100], but also in inflammation [83]. It is a gram-negative mucin-degrading bacterium [181]. Mucins are proteins produced by epithelial tissues in most organisms from *animalia* kingdom. They have several functions, among which the ability to form gels is one of the major ones. Mucins are involved in lubrication that creates efficient barriers against chemical factors [159].

Intestinal barrier serves a crucial role in protecting the intestinal cells from bacteria residing in lumen. This protection is accomplished by the production and shedding of mucus, secretion of anti-microbial peptides, proteins and production of secretory Immunoglobulin A (IgA) by immune exclusion [83]. Mucins are composed of several amino-acids and oligosaccharides and are often of nutritional value to some commensal bacteria. Not only *A. muciniphila* is capable of mucin degradation; several species from *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria* are also capable of mucin degradation [200]. Only the outer, thicker mucus layer is in contact with commensal bacteria, while the inner layer does not come into contact with bacteria in normal, non-pathological conditions [141].

Akkermansia muciniphila is reportedly inversely correlated with inflammations in the GI tracts (mainly Ulcerative Colitis (UC)), thus suggesting potential *anti-inflammatory* properties [106]. It has been observed that this species abundance is increased in humans when calorie restriction diet is practised [83]. Mucin degradation releases fucose, galactose, N-acetylglucosamine, N-acetylgalactosamine, sialic acid, sulfate, and also di-sacharides and small oligosaccharides which can be further used by the remaining bacteria not able to degrade mucin [83].

Treatment of mice with *A. muciniphila* revealed that it increased the number of Regulatory T cells (Tregs) and goblet cells in the gut, increased mucus production and immune signaling. In fact the indirect effect of taking anti-inflammatory drugs such as *predisone* most commonly used for SLE symptoms alleviation or current

clinical tests with *metformin* for reducing lupus flares might stimulate growth of *A. muciniphila* [205]. *Metformin* - the mitochondrial metabolism inhibitor - has been already shown to stimulate growth of *A. muciniphila* [193].

Compromised barrier function is associated with many diseases, like Inflammatory Bowel Disease (IBD) or metabolic syndrome [83]. Potential pathogenic role of *A. muciniphila* remain unclear, as it both reportedly stimulates mucin production, and degrades mucin [193]. It is therefore possible that overabundance of this bacteria is correlated with ongoing drug treatments for SLE patients, rather than as a hallmark of the disease itself. This can lead to the conclusion that changes induced by SLE progression could reside at higher functional level in the microbiota - i.e. metabolite level [17]. For assessing metabolic functions of a community of bacteria residing in the gut, a metabolomic study would be necessary. Also Whole Genome Sequencing (WGS) might yield helpful results in order to overcome many limitations associated with 16s rRNA gene surveys, such as limited accuracy, potential biases connected to PCR errors resulting in distorted community abundances, and, most importantly, to gain an deeper insight into community's physiology by assessing the true collective functional capability that are encoded in the microbial genes constituting a community, instead of relying on metagenome predictions through PICRUSt Ancestral State Reconstruction (ASR) algorithm.

17.4.3 Infection-induced autoimmunity hypothesis

Systemic Lupus Erythematosus (SLE) onset is still scientifically debated and investigated. As mentioned in the theory part of this thesis (chapter 2.1 page 24) its origins involve the interplay of environmental and host-genetic factors. The onset of many autoimmune diseases is unknown. Infection-induced autoimmunity hypothesis takes into consideration infection-induced activation of self-reactive lymphyocytes through *molecular mimicry* process (see *Molecular mimicry* frame).

The presence of `Ro` antibodies in SLE patients is common, although they are not always present. They are present in pre-clinical periods, before the disease manifests and during the disease progression in a subset of patients, i.e. they are not SLE specific, ubiquitous markers. `Ro` antibodies are usually the earliest auto-antibodies detected in SLE individuals. Some studies [160] try to identify environmental agents of potential etiologic role, like Epstein-Barr virus (EBV), and further characterize and explore the epitope of the auto-antibody response to `60 kDa Ro`. It is important to understand that auto-antibodies to `Ro` and other major immunogen `Sm` are present in around 40% of patients [160]. EBV might not be the only foreign factor initiating SLE disease onset. Infection-induced autoimmunity could be caused by other viral or bacterial agents. This statistic is also observed in gathered dataset, clearly visible for "metagenomic contributions" shown on figures 17.7 and 17.8 (pages: 155 and 156). In these figures, contributions to `Ro60` ortholog are not present for all SLE samples.

This thesis serves as a survey that identified a bacterial genus of *Verrucomicrobiaceae* that might play a role in SLE onset and progression through presence of a gen - orthologs mimicking `Ro60` (TROVE2) genes. Further detailed studies are required to isolate this genus and conduct detailed research. It is worth noting that initiation

of autoimmune disease by infection might not be involved in all SLE patients [160]. Other factors, like viral infections for which 16S rRNA survey is not able to assess, like EBV, may be possible.

Molecular mimicry

Molecular mimicry originally was associated with sequence similarity between foreign and host-produced peptides. Similarities between self-derived and potentially pathogenic viral or bacterial peptides could explain the underlying mechanism resulting in the cross-activation of autoreactive T cells leading to autoimmune responses. Foreign antigen can share sequence or structural similarities with self-antigents. Infection thus serves as a stimulus for initiating autoimmune disease - self-directed immune responses. Tissue pathology is caused by a chain of reactions starting from initial expansion of auto-reactive T-cells. This expansion is caused by a cross-linking of T-cell receptors (TCRs) by either *viral* or *bacterial* peptides. Those T-cells then cross-react with self-epitopes initiating immune responses [146].

A More recent hypothesis considers dual TCRs on a single T-cell. Those TCRs are characterized by reactivity to foreign and self-antigens, thus making individuals with dual TCR susceptible of an autoimmune disease [81].

17.5 Significance of other pathways

So far the main focus was put on SLE pathway identified through metagenomic predictions. However, there are other significantly enriched pathways identified in SLE patients as compared to HCs (see figure 17.2 on page 148). This section is concerned with assigning biological meaning to statistical significance for these identified pathways.

17.5.1 Cell motility and secretion

A microbiome study of murine lupus model [27], previously discussed in theoretical part of this thesis (chapter 2.1 page 24), identified a significant increase in cell motility genes for SLE mice. The authors of this paper concluded that this finding may suggest that bacteria adjust their location in order to actively access greater amounts of substrates.

17.5.2 Sulfur metabolism

The role of sulfur metabolism still remains to be fully understood in SLE patients. Sulfate-reducing bacteria (SRB) reportedly colonize the guts of around 50% of humans [182], moreover SRB are positively associated with inflammation [155], [84], i.e. one of the characteristic symptoms in SLE disease progression.

Sulfur is metabolized by all organisms, from bacteria and archaea to plants and animals. Many studies report that SLE patients exhibit impaired sulfur metabolism:

Table 17.2: Significantly enriched KEGG Orthologs (KOs) for sulfur metabolism in SLE patients compared to Healthy Controls (HCs).

KO ID	DEFINITION
K00299	ssuE; FMN reductase
K00380	cysJ; sulfite reductase (NADPH) flavoprotein alpha-component
K00381	cysI; sulfite reductase (NADPH) hemoprotein beta-component
K01011	TST, MPST, sseA; thiosulfate/3-mercaptopyruvate sulfurtransferase

either impaired sulphoxidation or no sulphoxidation is observed [118]. However an alternative pathway of cysteine oxygenase (plasma cysteine/sulphate) is reported in the same study to be significantly enriched in SLE patients ($p = 0.00001$). In fact *sulfur metabolism* is only partially impaired in SLE patients: only S-oxidation is impaired, while S-methylation is reportedly unimpaired in SLE patients [118].

Performing statistical test on all predicted orthologs among SLE patients and HCs with aforementioned Welch's t-test (does not assume that two groups having equal variances) and applying multiple test correction (Storey FDR), revealed four significantly enriched KOs in SLE patients (see table 17.2, page 160). Sulfur metabolism in KEGG database is comprised of 99 KOs, [117], 45 KOs were identified in the gathered dataset. Enriched pathways correspond to sulfite reductases and FMN (also known as riboflavine mononucleotide reductase), and thiosulfate (3-mercaptopyruvate sulfurtransferase). Reductases are the enzymes that catalyze a reduction reaction. This is an interesting observation, although should be treated with caution as KEGG Orthologs (KOs) are obtained indirectly through PICRUSt predictions, and should be regarded as suggestive only.

17.5.3 Phenylalanine metabolism

SLE patients often exhibit increased food sensitivity [162], leading to abnormal metabolism of various amino-acids, such as phenylalanine. Also, increased *phenylalanine metabolism*, as well as *tyrosine metabolism* (unobserved for this dataset) are reported [79] to aggravate the symptoms of SLE. SLE patients are therefore often recommended by doctors to avoid phenylalanine and tyrosine amino-acids in their diets. In fact SLE patients are often recommended a vegan diet, as beef and diary products are rich in aforementioned amino-acids. Increased food sensitivity or diet restrictions can therefore explain enrichment of this pathway.

17.5.4 Penicillin and cephalosporin biosynthesis

Penicillin is not only synthesized by filamentous fungi, such as *penicillium chrysogenum*; penicillins and cephalosporins can be produced by variety of microorganisms including many gram-positive streptomycetes, and a few gram-negative unicellular bacteria [4]. The bacteria are able to synthesize a plethora of cephalosporins and cephamycins but do not produce penicillins as end products [4].

Study design prohibited subjects included in this experiment to participate if they've been subjected to antibiotic treatment less than a month prior to sample

collection. It is reported that SLE patients exhibit antibiotic allergy to penicillin, cephalosporins, sulfonamides, tetracyclines, and erythromycin [170]. The symptoms include most commonly a skin rash at various body sites.

Since *penicillin and cephalosporin biosynthesis* KEGG pathway is comprised of several orthologs, identifying major contributors to this pathway significance is important. Among several orthologs present in KEGG database [114], `K01060` is the only major contributor. Through PICRUSt functionality, it is possible to identify taxonomic ranks driving overabundance of this ortholog. Figure 17.9 (page 162) presents metagenome contributions for `K01060` ortholog - *ephalosporin-C deacetylase*. The major contributor is *Akkermansia* genus, while other minor contributors include *Collinsella* genus, and unclassified genus coming from *Ruminococcaceae* family and others. *Cephalosporin-C deacetylase* is an enzyme that catalyzes the reversible chemical reaction shown below [210]:

$$\text{cephalosporin C} + H_2O \rightleftharpoons \text{deacetylcephalosporin C} + \text{acetate}$$

Is is therefore possible that *Akkermansia muciniphila* has genes that could catalyze cephlalosporin biosynthesis. As was mentioned earlier, SLE patients exhibit allergic reactions to cephalosporins, (such as skin rush) that are also a hallmark of SLE. However these are only assumptions based on putative metagenome inferred through PICRUSt Ancestral State Reconstruction (ASR) algorithm, and should be treated as suggestive only.

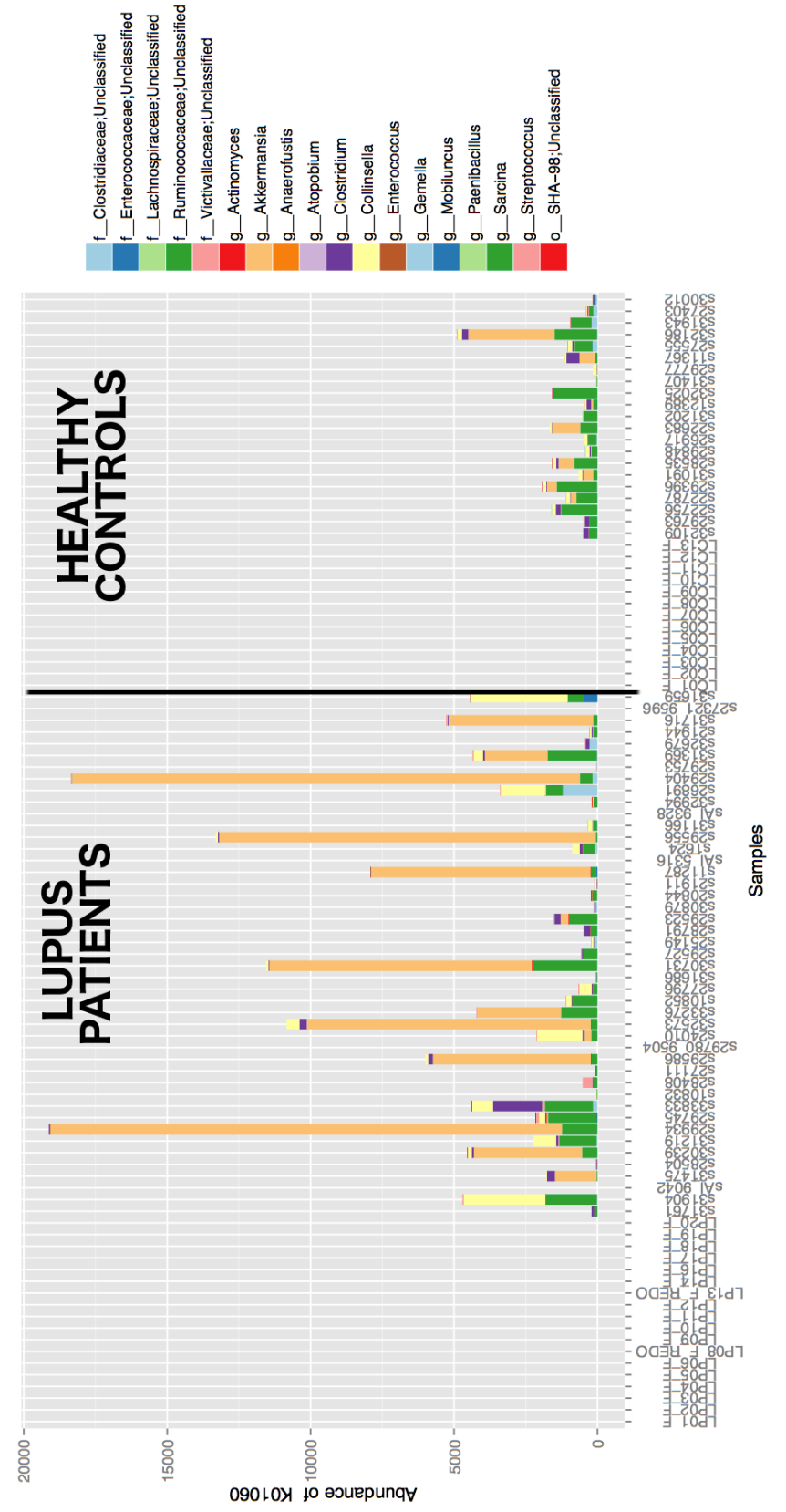


Figure 17.9: Metagenome contributions for K01060 ortholog at genus rank. Plot representing all of the samples gathered for experiment ordered by SLE and HC status. Plot points to several microbial genuses: *Akkermania*, *Neisseria* and *Pseudomonas*, among which *Akkermansia* has the strongest contributions; others are not visible on the plot due to small contributions.

17.5.5 Inorganic ion transport and metabolism

The role of increased potential for inorganic ion transport and metabolism is not clear. Inorganic ions play an essential part in the growth and survival of the bacterial cells [87]. They are essential for providing nutrition and sustaining metabolism of bacterial cells. They also prevent lysis of bacterial cells. It is therefore possible that for some yet to be explored reason, the conditions in GI tract of SLE patients require more effort on the part of bacterial metabolism to prevent cell lysis and sustain bacterial existence.

17.5.6 Non-homologous end-joining

Double-strand DNA breaks may occur for many exogenous factors, such as ionizing radiation or therapeutic drugs. Two main mechanisms occur for repairing those breaks: homologous and non-homologous end-joining. The first is a complete, error-free repair pathway relying on DNA homology. Non-homologous repair is a simple religation of broken DNA that doesn't require a template. It has been shown that double strand breaks of DNA in bacteria are not only repaired by homologous recombination but also by non-homologous end-joining [194]. Non-homologous end-joining in bacteria is a relatively new field of study, where many branches still remain under-explored. The same scientific study [194] hypothesizes that perhaps non-homologous end-joining might have potential role in the insertion of bacterial DNA into the host genome. Occurrences of this type of DNA repairs could be also correlated with SLE patients' drug intake [197], [28].

17.5.7 Steroid and carotenoid biosynthesis

Current knowledge about steroid biosynthesis in bacteria is very limited. In fact, some papers state that "almost nothing is known" about its mechanisms and functions [25]. Steroids have anti-inflammatory and immunosuppresive effects, and for this reason they are administered to SLE patients as a part of treatments.

Steroid biosynthesis pathway in KEGG database is characterized by 31 KOs [115], however only 3 orthologs were detected, with one significantly more abundant in SLE patients - `K00801`: FDFT1; farnesyl-diphosphate farnesyltransferase, also known as squalene synthase (SQS). SQS is found in plants, yeast and animals, however, in terms of structure (sequence similarity) and mechanics, it closely resembles phytoene synthase (PHS). PHS serves a similar role as SQS in plants and bacteria - it catalyzes the synthesis of phytoene, a precursor of carotenoid compound. [201]

Carotenoids are synthesized by all photosynthetic organisms (including plants) and some non-photosynthetic bacteria and fungi. KEGG pathway of *carotenoid biosynthesis* is represented by 46 orthologs (KOs) [109], where 22 orthologs were identified in the study dataset. Among those 22 orthologs, only one `K02291`: 15-cis-phytoene/all-trans-phytoene synthase were statistically more abundant in SLE patients (see figure 17.11). Again the major contributor to this ortholog comes from *Verrucomicrobiaceae* family containing *Akkermansia muciniphila* bacteria (see figure 17.12 on page 166).

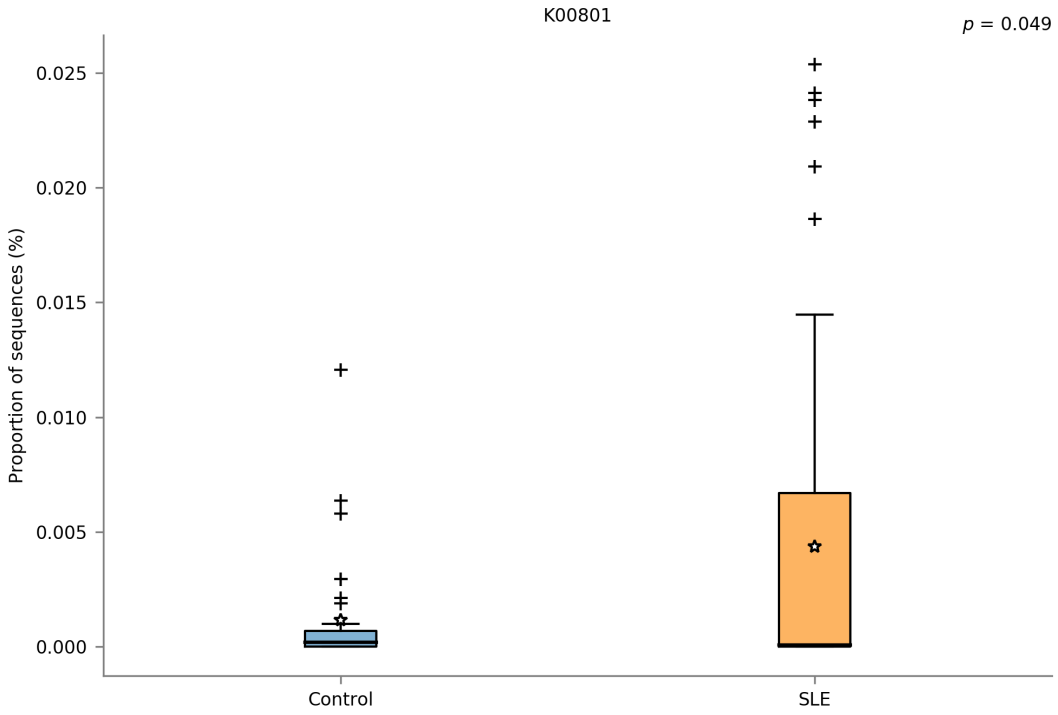


Figure 17.10: Statistically abundant KEGG ortholog `K00801`: FDFT1; farnesyl-diphosphate farnesyltransferase in SLE patients. p-value = 0.049 corrected for multiple comparison with Storey `FDR` correction.

A potential anti-inflammatory role of carotenoids is proposed by researchers. The concentration of carotenoids is the highest in the GI tract, and they have been found to positively modulate markers of inflammation and oxidative stress [143]. It is thus possible that diet rich in carotenoids, such as certain fruits and vegetables combined with drug intake of SLE patients is the reason behind the overabundance of carotenoid-related orthologs. Current research studies demonstrate that there are several pathways possibly related to carotenoids, inflammation and oxidative-stress, where many aspects still remain to be elucidated [143].

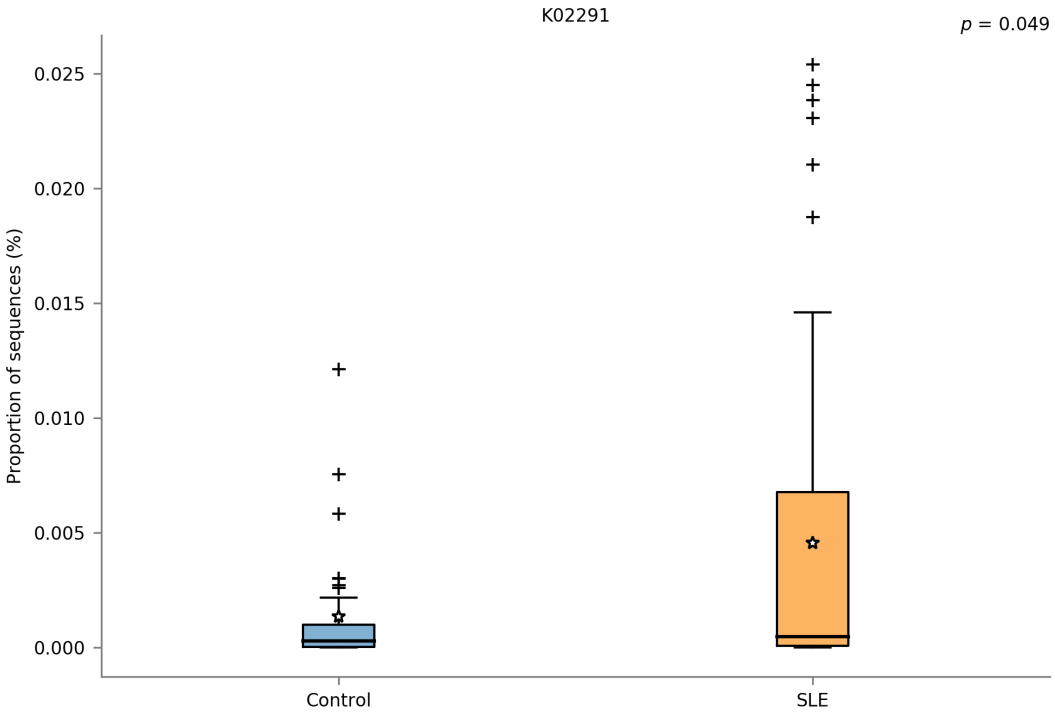


Figure 17.11: Statistically abundant KEGG ortholog `K02291`: 15-cis-phytoene/all-trans-phytoene synthase in SLE patients. p-value = 0.049 corrected for multiple comparison with Storey `FDR` correction.

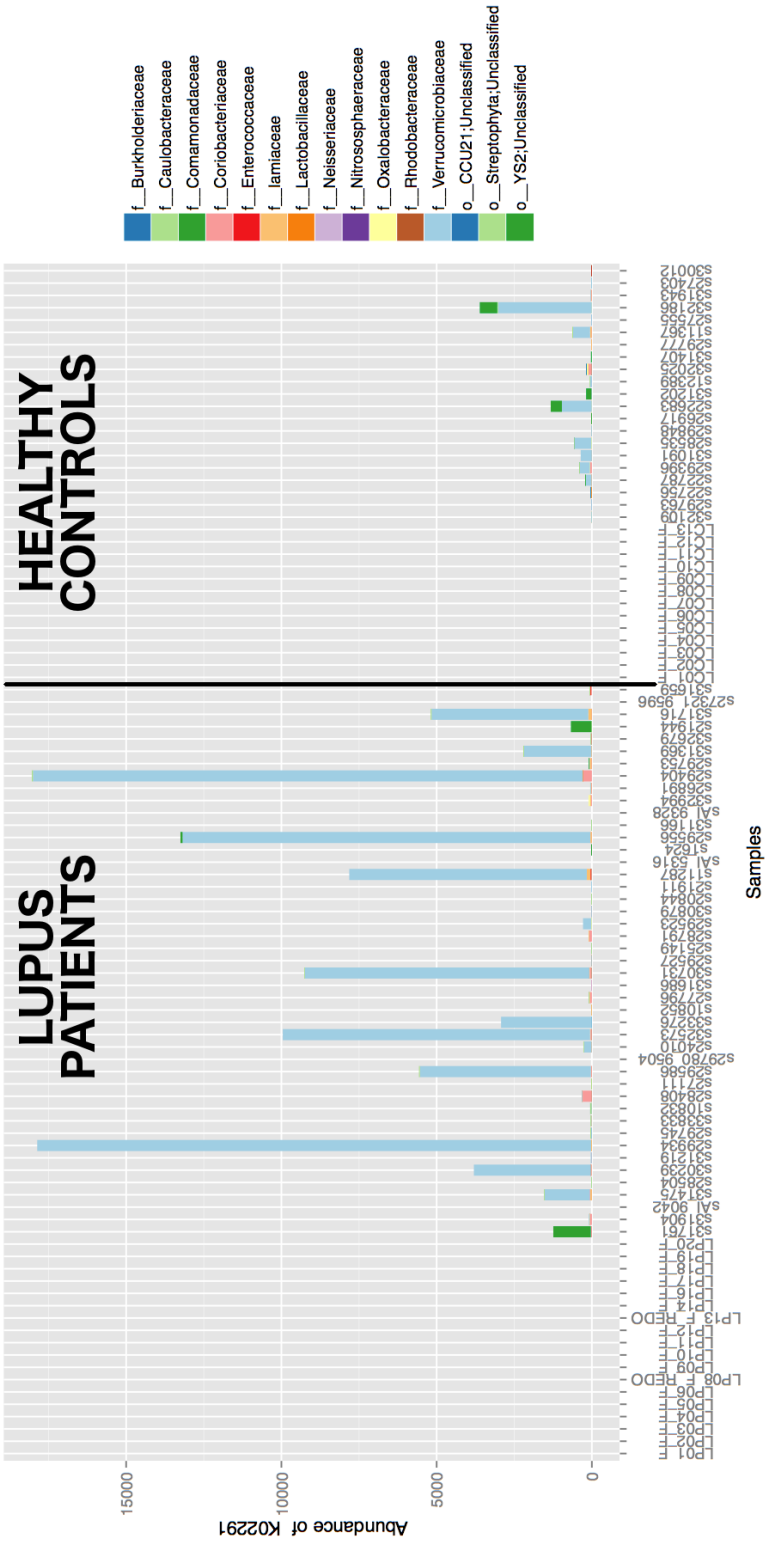


Figure 17.12: Metagenome contributions for KEGG ortholog K02291: 15-cis-phytoene/all-trans-phytoene synthase. The major contributor is *Verrucomicrobiaceae* family containing *Akkermansia muciniphila* bacteria.

17.5.8 Fluorobenzoate degradation

Fluorobenzoate degradation KEGG pathway is comprised of 18 orthologs [113], among which 11 are present in the gathered dataset. Only one ortholog was significantly more abundant in microbiome of SLE patients: `K01061` - carboxymethylenebutenolidase. This ortholog is homologous to dienelactone hydrolase involved in the bacterial halocatechol degradation pathway. It is is thought to predominantly contribute to the quick onset of drug action [137].

17.5.9 Flavonoid biosynthesis

Flavonoid biosynthesis KEGG pathway is comprised of 21 orthologs [112], however only two are present in the gathered dataset: `K00588` and `K00660`. Only `K00660` ortholog representing chalcone synthase (CHS) was significantly more abundant in SLE microbiome. This enzyme is ubiquitous in higher plants, it is therefore possible that there is a bacterial ortholog similar to CHS. In fact some research [166] describes plant-like biosynthetic pathways in bacteria that include chalcone pathways. CHS catalysis serves as the initial step for flavonoid biosynthesis. Flavonoids are secondary metabolites in plants that serve numerous functions, however the role of its orthologous group in bacteria is unknown.

17.5.10 Circadian rythm (plant)

Circadian rhythm pathway is represented in KEGG database by 27 orthologs [110]. Among them only one was identified through PICRUSt metagenome predictions. `K00660` ortholog representing chalcone synthase (CHS) is significantly more abundant in SLE microbiome. The same ortholog contributed to *flavonoid biosynthesis* pathway. Again, bacteria may exhibit plant-like biosynthetic pathways including chalcone pathway [166], however its role remains to be elucidated.

17.5.11 Fatty acid elongation in mitochondria

Fatty acid elongation pathway in KEGG database [111] is represented by 23 orthologs. Among them only one ortholog was present and statistically more abundant (p=0.049 after Storey FDR correction) in SLE fecal microbiome. This ortholog is `K07512` - mitochondrial trans-2-enoyl-CoA reductase (**TER**, MECR, NRBF1). The literature on TER remains scarce, it is reported that the prokaryotic TER homolog has not been investigated so far [204].

17.5.12 Caffeine metabolism

Caffeine metabolism in KEGG database is represented by 9 orthologs [108]. Among them only three (`K00106`, `K00622` and `K00365`) orthologs were identified in the gathered dataset. However, only one ortholog was significantly more abundant in SLE patients: `K00365` - urate oxidase (Uox) (p=0.049 after Storey FDR correction) .

Uox is ubiquitously found in most organisms: from bacteria to mammals (absent however in humans and other higher apes) where it plays a different metabolic role.

Bacterial Uox presence is said to be important for rapid elimination of the metastable intermediary products of urate oxidation [86].

Uox prevents the uric acid levels from rising to dangerously high levels in bacteria. More orthologous groups identified in SLE patients combined with the fact that uric acid is produced during the breakdown of purines found in certain food products, like meats (bacon, beef, pork, and lamb), beer, and seafood could be linked with differences in dietary habits. It has been already mentioned that SLE patients are recommended a plant-based diet, limiting meat (especially beef) and dietary products. Dietary habits however were not recorded as a part of this study. Therefore significance of this pathway remains unclear, and probably not relevant to the aim of this study.

Chapter 18

Results discussion

16S rRNA gene sequencing survey approach was used to study fecal microbiota of SLE patients and appropriate Healthy Controls (HCs). This analytical approach is the preferred way of broad community characterization, although it exhibits limited sensitivity and specificity for higher taxonomic ranks. Its drawbacks and limitations are discussed in section 3.3 (page 40). Although this study established a pipeline limiting the effects of some of the errors discussed in section 4.7 (page 57), the upper limits are due to the specificity of amplicon sequencing. This may in turn fail to resolve a substantial fraction of the actual diversity in a studied community [192]. Additionally, PICRUSt functionality for infering putative metagenome is limited (see figure 7.3 on page 94); it is reasonably well characterized for human samples, as opposed to other mammalian samples, and hyper-saline samples. Despite the fact that expected accuracy for human samples is in the range of 75%-90% ,the creators themselves insist that these predictions should be treated as "suggestive only", thus requiring cautionary measures of interpretation.

16S rRNA analysis showed statistically significant broad changes in the SLE microbiome. Namely the within-sample diversity (*alpha diversity*) is significantly decreased in SLE patients. This means that overall number of observed bacteria is smaller in SLE patients, exhibiting features of gut microbial dysbiosis. Across-sample diversity analyses reveal however that phylogenetic compositions, tend to be comparable, as shown in beta-diversity plots (figures 16.7 - 16.10, pages 130 - 131). Two "clusters" of healthy controls (HCs) and SLE patients are not separated on PCoA plots; they tend to be intermingled with each other. Figure 16.9 (page 131) shows that even samples among one group, regardless whether SLE patients or HCs, exhibit greater diversity shifts within the group, compared to between diseased and healthy state. One of the interesting finding of the Human Microbiome Project (HMP) is that human gut exhibits great variations among healthy subjects in terms of taxonomic composition, while at the same time maintaining the functional repertoire [78] - see figure 1.1 on page 20. However, after performing multiple Student's two-sample t-test for every pair of boxplots for `UniFrac` distances, it revealed that differences in unweighted `UniFrac` were statistically significant, whereas for weighted `UniFrac` those differences were insignificant. The former method, not accounting for bacterial abundances, reveals that SLE and Healthy Control (HC) microbiomes are differentiated by the capability to host certain microbes, likely of small relative

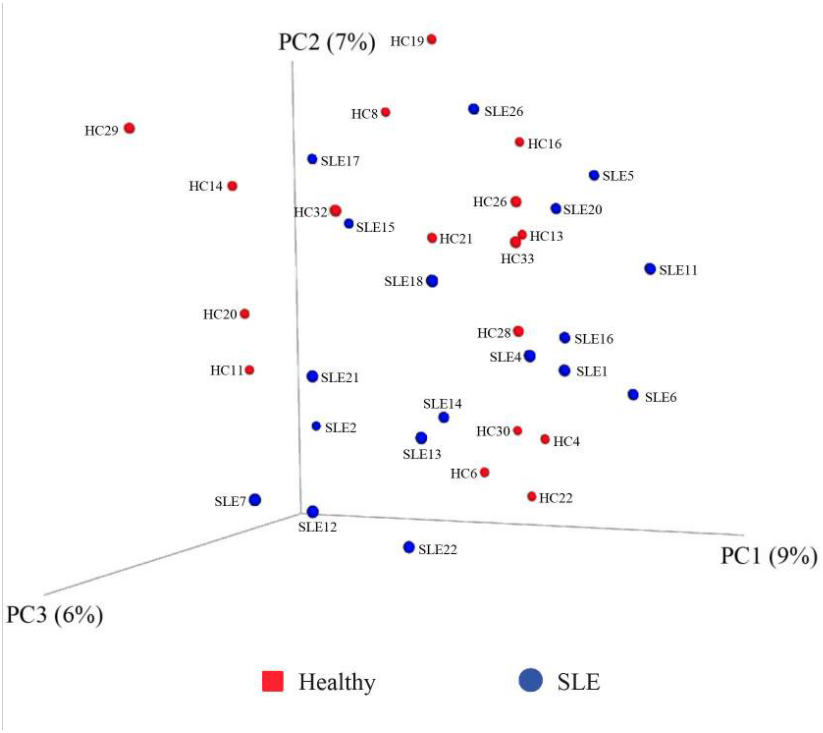


Figure 18.1: A similar study result showing PCoA plot of Healthy Control (HC) and Systemic Lupus Erythematosus (SLE) microbiome samples, showing no subcategorized and distinct clusters. Source: [17].

abundances. The latter, weighted `UniFrac` method, didn't reveal statistically significant differences in this metric. Thus, accounting for microbial abundances `UniFrac` metric did not reveal shifts in more dominant bacteria - i.e. there were no significant shifts in the dominant, broad microbiota compositions. Interesting results were observed in a similar study of microbiome in SLE patients and Healthy Control (HC) subjects [17]. Researchers in this study have also not observed overall significant differences in bacterial compositions explored through beta diversity (see figure 18.1 reproduced from this study). Researchers argue that changes between those two compared groups are marked at higher level of functional hierarchy, such as metabolite level. Aforementioned study has not however explored the putative metagenome through PICRUSt prediction by Ancestral State Reconstruction (ASR) algorithms, or any other method.

This study identified members of *Verrucomicrobiaceae* family to be particularly more abundant in SLE patients, however, this is not a general rule. Although there was statistical difference, where SLE patients exhibited increased abundance of *Verrucomicrobiaceae* family, figure 16.18 (page 144) shows that this bacterial family was not always more abundant in SLE patients. In other words, there are certain healthy study control subjects that exhibited greater abundance of this family compared to particular SLE patients. For this particular reason members of *Verrucomicrobiaceae* family can't be regarded as a potential biomarker of the disease.

Inferred putative metagenome exhibited several differences between compared healthy and diseased subjects. One major difference with a particular relevance to the

study was significant presence of Ro60 orthologs assigned to the *Verrucomicrobiaceae* family that was more abundant in SLE patients. As was discussed in section 17.4 (page 157), the presence of this ortholog is important as antibodies against Ro60 are the most frequently found antibodies in SLE and Sjogren's syndrome.

Establishing causality is a major challenge in microbiome studies. Although study group was not subject to antibiotics, SLE patients were undergoing treatment that reduces SLE manifestations and symptoms. SLE is a systemic autoimmune disease that affects multiple organs, therefore for each disease manifestation there is a different drug recommendation (see figure 2.3, page 28). In fact gastrointestinal symptoms do not always accompany SLE onset and/or progression, and every organ could be affected [203]. GI conditions in SLE patients are potentially life-threatening, thus all of the patients (including this study) were likely to be undergoing treatments (although not necessarily, if GI symptoms were not observed). Researchers have found no specific auto-antibodies associated with SLE gastroenteropathy, however it has been observed that most complications are responsive to corticosteroids and immunosuppresive agents [203]. Therefore identified changes in the microbiome could be associated with drug intake or inflammations in the GI tract, discussed in *"significance of other pathways"* section 17.5 (page 159).

The role of *Akkermansia muciniphila* remains to be elucidated. The nature of this study was unable to access the full genome of this bacteria, and orthologs associated with it should be considered with caution. This bacterium however has been shown to be inversely correlated with inflammation [38] [83] (May 2017) which puts into question infection-induced hypothesis on the part of *Akkermansia muciniphila*. It is however possible that there may exist certain, unidentified strains of this species having parts of their genome mimicking Ro60 peptide, and thus having a role in SLE pathogenesis. However, this scientific question requires further detailed investigation.

Summary

Microbiome elucidation in terms of composition and functional repertoire in various healthy and disease conditions is one of decennial challenges for this decade, as published in "*Nature Visions*" [1]. This study showed that microbial composition between healthy controls and Systemic Lupus Erythematosus (SLE) patients exists. Namely SLE patients exhibit smaller alpha (within-sample) diversity, meaning that observed number of bacteria is smaller as compared to healthy controls (HCs). However, overall community structure tends to be diverse both in SLE patients and HCs, sharing many similarities. Both microbiota differ only in the capacity to host certain microbes. This reveals that SLE state might not be driven by broad changes in phylogenetic compositions, as *beta diversity* plots didn't exhibit clear separation. One bacterium *Akkermansia muciniphila* of *Verrucomicrobiaceae* family was particularly abundant among SLE patients. Literature overview of recently published scientific papers revealed that it might have mucin degrading and anti-inflammatory properties. Potential pathogenic role of this bacterium through *molecular mimicry* mechanisms was also discussed due to the fact that it has been associated with abundance Ro60 orthologs in SLE patients. However, methodology used for this study imposes certain constraints in terms of limited accuracy and predictive nature of in-

ferred metagenome. For this reason further exploration of this intricate and complex interdependent relationship between microbiota and host in Systemic Lupus Erythematosus (SLE) patients by other methods, like Whole Genome Sequencing (WGS), meta-transcriptomics or metabolomics, might reveal additional relationships, and may help to fill a gap in knowledge that currently exists in understanding SLE onset and progression.

Glossary

core microbiome is a working hypothesis, that states that there may be conserved microbial composition among healthy population, that is disrupted in diseased states.. 18, 20

ddNTP **di**deoxynucleotide triphosphate, i.e. nucleotide, single units of the bases A, T, G, and C, which are essentially "building blocks" for new DNA strands, but lack the 3'-OH group from standard deoxynucleotide triphosphate. . 46, 47, 49, 50, 52, 54, 55, 113

dNTP deoxynucleotide triphosphate, i.e. nucleotide, single units of the bases A, T, G, and C, which are essentially "building blocks" for new DNA strands.. 46, 49, 52, 55, 110

k-mer A `k-mer` is the term applied for a sub-string of length `k` in a genetic string. The term is frequently applied in reference to genome sequencing, in which relatively short, overlapping reads are used to reconstruct the genome. 74, 75

mock community refers to a positive control in microbiome study, it is either: 1) DNA-free water that is spiked with DNA from mul- tiple known taxa of lab-cultured bacteria. 2) Sequences of microbial communities with known members and their abundance. 3) Artificially constructed sequences of microbial members pulled from microbial databases, pooled (i.e. concatenated) together so as to obtain an expected profile for later comparisons. . 57–59, 65

OTU clusters of sequences, intended to represent some degree of taxonomic relatedness. When sequences are clustered at 97% sequence similarity, each resulting cluster is typically thought of as representing a species. This model and the current techniques for picking OTUs are known to be flawed, however, in that 97% OTUs do not match what humans have called species for many microbes. Determining exactly how OTUs should be defined, and what they represent, is an active area of research. 29, 38, 57–59, 61, 64, 65, 67, 70–76, 78, 79, 82, 83, 85, 87, 89, 90, 92, 93, 95, 118–120, 124, 127, 137, 145, 151, 153

RTC Reversible Terminator Chemistry, refers to the process of adding one nucleotide base at a time in order to measure the nucleotide specific fluorescent signal, while preventing other complementary nucleotides to be added to a strand. . 49, 54, 55

Acronyms

ACD Acid-citrate-dextrose. 101

ACR American College of Rheumatology. 25, 27

APS Antiphospholipid Syndrome. 30

ASD Autism Spectrum Disorder. 22

ASR Ancestral State Reconstruction. 92, 93, 120, 151, 158, 161, 170

BIOM Biological Observation Matrix. 64, 145

bp base pair. 38, 39, 44

CNS Central Nervous System. 22, 34, 37

COGs Clusters of Orthologous Groups. 92

EAE Experimental Autoimmune Encephalitis. 35–37

EBV Epstein-Barr virus. 158, 159

EDTA Ethylenediamine Tetraacetic Acid. 101

FDR False Discovery Rate. 86, 95, 132–134, 137, 140, 143, 147, 148, 160

GF Germ-free. 34, 37

GI gastrointestinal. 19, 22, 30, 34, 100, 157, 163, 164, 171

HC Healthy Control. 102, 104, 105, 124, 126–132, 137–140, 145, 147–150, 153, 157, 159, 160, 162, 169–171

HMP Human Microbiome Project. 87, 90, 129, 137, 169

HUMAnN HMP Unified Metabolic Analysis Network. 90

IBD Inflammatory Bowel Disease. 22, 36, 127, 158

IBS Irritable Bowel Syndrome. 22, 127

IgA Immunoglobulin A. 157

IgG Immunoglobulin G. 24

IgM Immunoglobulin M. 24

IL-17 Interleukin 17. 36

INFα interferon-α. 33

INFγ interferon-γ. 30

IUPAC International Union of Pure and Applied Chemistry. 60

KEGG Kyoto Encyclopedia of Genes and Genomes. 14, 145, 152, 153, 160, 161, 163–167

KO KEGG Ortholog. 64, 92, 120, 145, 152–154, 160, 163

MAC Microflora Associated Characteristic. 21

MIA Maternal Immune Activation. 22

NCBI National Center for Biotechnology Information. 98

NGS Next-Generation Sequencing. 19, 22, 42–44, 46, 54–56, 65, 70, 88, 111, 113

NMDS Non-metric Multidimensional Scaling. 80, 83

NOD non-obese diabetic. 31

NSTI Nearest Sequenced Taxon Index. 151

OMRF Oklahoma Medical Research Foundation. 100–102, 110, 113

PBMC Peripheral Blood Mononuclear Cell. 101

PCA Principal Components Analysis. 80, 83

PCoA Principal Coordinates Analysis. 19, 28, 29, 31, 65, 80–84, 119, 129–131, 169, 170

PCR Polymerase Chain Reaction. 40, 41, 44, 46, 57, 59, 67, 72, 110, 111, 158

PICRUSt Phylogenetic Investigation of Communities by Reconstruction of Unobserved States. 14, 90–94, 145, 146, 151, 153, 158, 160, 161, 167, 169, 170

PSA Polysacharide A. 37

QIIME Quantitative Insights Into Microbial Ecology. 89

RA Rheumatoid Arthritis. 36, 37

RA Retinoic Acid. 28–30

SBS Sequencing by Synthesis. 42, 44, 47, 52, 54–56, 113

SCFA Short-Chain Fatty Acids. 20, 22

SDS Sodium dodecyl sulfate. 107

SFB Segmented Filamentous Bacteria. 30

SFB Spore-forming Segmented Filamentous Bacteria. 35, 36

SLE Systemic Lupus Erythematosus. 12–14, 24–28, 30, 31, 33, 37, 98, 102–105, 124, 126–132, 137–140, 143, 145–150, 152–165, 167–172

SOP Standard Operating Procedure. 34, 65, 87

SRB Sulfate-reducing bacteria. 159

ssDNA single-stranded DNA. 46

STAMP Statistical Analysis of Metagenomic Profiles. 94, 95, 120, 145

TCR T-cell receptor. 159

Th17 T helper 17 cell. 21, 30

TLR toll-like receptor. 26

Treg Regulatory T cell. 157

Tregs Regulatory T cells. 20, 34–36

TRIS tris(hydroxymethyl)aminomethane. 107

UC Ulcerative Colitis. 157

WGS Whole Genome Sequencing. 33, 40, 44, 59, 63, 78, 90, 94, 124, 158, 172

zOTU Zero-Radius Operational Taxonomic Unit. 69, 71, 95

Bibliography

[1] "2020 visions". In: *Nature* 463.7277 (Jan. 2010), pp. 26–32. DOI: 10.1038/463026a. URL: https://doi.org/10.1038/463026a.

[2] Alison Abbott. "Scientists bust myth that our bodies have more bacteria than human cells". In: *Nature* (Jan. 2016). DOI: 10.1038/nature.2016.19136. URL: https://doi.org/10.10382Fnature.2016.19136.

[3] Applied Biological Materials - abm. *Next Generation Sequencing (NGS) - An Introduction*. video. June 2015.

[4] Y. Aharonowitz, G Cohen, and J F Martin. "Penicillin and Cephalosporin Biosynthetic Genes: Structure, Organization, Regulation, and Evolution". In: *Annual Review of Microbiology* 46.1 (Oct. 1992), pp. 461–495. DOI: 10.1146/annurev.mi.46.100192.002333. URL: https://doi.org/10.1146%5C%2Fannurev.mi.46.100192.002333.

[5] Alejandro Reyes et al. "Viruses in the faecal microbiota of monozygotic twins and their mothers". In: *Nature* 466 (July 2010), pp. 334–338.

[6] Andrew K. Benson et al. "Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors". In: *PNAS, Proceedings of the National Academy of Sciences* 107.44 (Nov. 2010), pp. 18933–18938.

[7] Andrew Krohn et al. "Optimization of 16S amplicon analysis using mock communities: implications for estimating community diversity". In: *PeerJ* (Oct. 2016).

[8] B. J. Haas et al. "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons". In: *Genome Research* 21.3 (Jan. 2011), pp. 494–504. DOI: 10.1101/gr.112730.110. URL: https://doi.org/10.1101%2Fgr.112730.110.

[9] B. M. Johnson et al. "Impact of dietary deviation on disease progression and gut microbiome composition in lupus-prone SNF1 mice." In: *Clinical and Experimental Immunology: The Journal of Translational Immunology* 181.2 (Aug. 2015), pp. 323–337.

[10] B. S. Andrews et al. "Spontaneous murine lupus-like syndromes: clinical and immunopathological manifestations in several strains". In: *Journal of Experimental Medicine* 148.5 (Nov. 1978), pp. 1198–1215.

[11] Bäckhed F et al. "Host-Bacterial Mutualism in the Human Intestine". In: *Science* (Mar. 2005), pp. 1915–1920.

[12] Breban MA et al. “Influence of the bacterial flora on collagen-induced arthritis in susceptible and resistant strains of rats.” In: *Clinical and Experimental Rheumatology* 11.1 (Feb. 1993), pp. 61–64.

[13] Chiharu Kubo et al. “Effects of calorie restriction on immunologic functions and development of auto- immune disease in NZB mice”. In: *Society for Experimental Biology and Medicine* (Nov. 1992), pp. 192–199.

[14] Claudia Di Giacinto et al. “Probiotics ameliorate recurrent Th1-mediated murine colitis by inducing IL-10 and IL-10-dependent TGF-beta-bearing regulatory cells.” In: *Journal of Immunology* 174.6 (Mar. 2005).

[15] Daniel McDonald et al. “The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome”. In: *GigaScience* 1.1 (July 2012). DOI: 10.1186/2047-217x-1-7. URL: https://doi.org/10.1186%5C%2F2047-217x-1-7.

[16] Daryl M. Gohl et al. “Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies”. In: *Nature Biotechnology* 34.9 (Sept. 2016), pp. 942–949. DOI: 10.1038/nbt.3601.

[17] David Rojo et al. “Ranking the impact of human health disorders on gut metabolism: Systemic lupus erythematosus and obesity as study cases”. In: *Scientific Reports* 5.1 (Feb. 2015). DOI: 10.1038/srep08310. URL: https://doi.org/10.1038%5C%2Fsrep08310.

[18] Eben I. Lichtman et al. “Emerging therapies for systemic lupus erythematosus–focus on targeting interferon-alpha”. In: *Clinical Immunology* 143.3 (June 2012), pp. 210–221.

[19] Efrem S Lim et al. “Early life dynamics of the human gut virome and bacterial microbiome in infants.” In: *Nature Medicine* 21.10 (Oct. 2015), pp. 1228–1234.

[20] F. Sanger et al. “Nucleotide sequence of bacteriophage phiX174 DNA”. In: *Nature* 265.5596 (Feb. 1977), pp. 687–695. DOI: 10.1038/265687a0. URL: https://doi.org/10.1038%5C%2F265687a0.

[21] Filip Scheperjans et al. “Gut microbiota are related to Parkinson’s disease and clinical phenotype”. In: *Movement Disorders* 30.3 (Dec. 2014), pp. 350–358. DOI: 10.1002/mds.26069. URL: https://doi.org/10.1002%5C%2Fmds.26069.

[22] Gene W. Tyson et al. “Community structure and metabolism through reconstruction of microbial genomes from the environment”. In: *Nature* 428 (Mar. 2004).

[23] H. L. B. M. Klaasen et al. “Intestinal, segmented, filamentous bacteria in a wide range of vertebrate species”. In: *Laboratory Animals* 27 (Apr. 1993), pp. 141–150.

[24] Hall Jason A. et al. “Commensal DNA limits regulatory T cell conversion and is a natural adjuvant of intestinal immune responses”. In: *Immunity* (Oct. 2008), pp. 637–649.

[25] Helge Björn Bode et al. “Steroid biosynthesis in prokaryotes: identification of myxobacterial steroids and cloning of the first bacterial 2, 3(S)-oxidosqualene cyclase from the myxobacterium Stigmatella aurantiaca”. In: *Molecular Microbiology* 47.2 (Jan. 2003), pp. 471–481. DOI: 10.1046/j.1365-2958.2003.03309.x. URL: https://doi.org/10.1046%5C%2Fj.1365-2958.2003.03309.x.

[26] Hsin-Jung Wu et al. “Gut-Residing Segmented Filamentous Bacteria Drive Autoimmune Arthritis via T Helper 17 Cells”. In: *Immunity* (June 2010), pp. 815–827.

[27] Husen Zhang et al. “Dynamics of Gut Microbiota in Autoimmune Lupus”. In: *Applied and Environmental Microbiology* 80.24 (Dec. 2014), pp. 7551–7560.

[28] Husen Zhang et al. “Dynamics of Gut Microbiota in Autoimmune Lupus”. In: *Applied and Environmental Microbiology* 80.24 (Dec. 2014), pp. 7551–7560.

[29] Javier Ochoa-Repáraz et al. “Role of Gut Commensal Microflora in the Development of Experimental Autoimmune Encephalomyelitis.” In: *Journal of Immunology* 183.10 (Nov. 2009), pp. 6041–6050.

[30] Jeremy E. Koenig et al. “Succession of microbial consortia in the developing infant gut microbiome”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108 (Mar. 2010).

[31] Jo Handelsman et al. “Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products”. In: *Cell Chemical Biology* 5.10 (Oct. 1998), pp. 245–249.

[32] John Penders et al. “Factors influencing the composition of the intestinal microbiota in early infancy.” In: *Pediatrics Journal* 118.2 (Aug. 2006).

[33] Jose C. Clemente et al. “The Impact of the Gut Microbiota on Human Health: An Integrative View”. In: *Cell* 148.6 (Mar. 2012), pp. 1258–1270.

[34] Juan Jovel et al. “Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics”. In: *Frontier in microbiology* 7.459 (Apr. 2016).

[35] Julia Preising et al. “Selection of bifidobacteria based on adhesion and anti-inflammatory capacity in vitro for amelioration of murine colitis.” In: *Applied and Environmental Microbiology* 76.9 (May 2010), pp. 3048–3051.

[36] Ley et al. “Ecological and evolutionary forces shaping microbial diversity in the human intestine”. In: *Cell* 124 (2006), pp. 837–848.

[37] Liam O’Mahony et al. “Lactobacillus and bifidobacterium in irritable bowel syndrome: symptom responses and relationship to cytokine profiles.” In: *Gastroenterology* 128.3 (Mar. 2005), pp. 541–551.

[38] Marc Schneeberger et al. “Akkermansia muciniphila inversely correlates with the onset of inflammation, altered adipose tissue metabolism and metabolic disorders during obesity in mice”. In: *Scientific Reports* 5.1 (Nov. 2015). DOI: 10.1038/srep16643. URL: https://doi.org/10.1038/srep16643.

[39] Marcus B. Jones et al. “Library preparation methodology can influence genomic and functional predictions in human microbiome research”. In: *PNAS: Proceedings of the National Academy of Sciences* 112.45 (Nov. 2015), pp. 14024–14029.

[40] Marie-Laure Santiago-Raber et al. “Emerging roles of TLR7 and TLR9 in murine SLE”. In: *Journal of Autoimmunity* (2009), pp. 231–238.

[41] Markle JG et al. “Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity”. In: *Science* (2013), pp. 1084–1088.

[42] Martin A. Kriegel et al. “Naturally transmitted segmented filamentous bacteria segregate with diabetes protection in nonobese diabetic mice.” In: *Proceedings of the National Academy of Sciences of the United States of America* 108.28 (July 2011), pp. 11548–11553.

[43] Matsuzaki T et al. “Prevention of onset in an insulin-dependent diabetes mellitus model, NOD mice, by oral feeding of Lactobacillus casei.” In: *APMIS* 105.8 (Aug. 1997), pp. 643–649.

[44] Melanie Schirmer et al. “Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform”. In: *Nucleic Acids Research* 43.6 (Mar. 2015).

[45] Mirjana Rajilič-Stojanovič et al. “Diversity of the human gastrointestinal tract microbiota revisited.” In: *Environmental Microbiology* 9.9 (Sept. 2007), pp. 2125–2136.

[46] Morgan G.I. Langille et al. *PICRUSt: Phylogenetic Investigation of Communities by Reconstruction of Unobserved States*. webpage. 2013. URL: `http://picrust.github.io/picrust/`.

[47] Morgan G.I. Langille et al. “Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.” In: *Nature Biotechnology* 31.9 (Sept. 2013), pp. 814–821.

[48] Pascale Alard et al. “Probiotics control lupus progression via induction of regulatory cells and IL-10 production.” In: *Journal of Immunology* 182.1 (Apr. 2009).

[49] Peter J. Turnbaugh et al. “Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins”. In: *PNAS Proceedings of the National Academy of Sciences* 108.16 (Apr. 2010), pp. 7503–7508.

[50] Philip Seo et al. *Oxford American Handbook of Rheumatology*. second. Oxford University Press, 2013.

[51] Qij J et al. “A human gut microbial gene catalogue established by metagenomic sequencing.” In: *Nature* 464.7285 (2010), pp. 59–65.

[52] Qiong Wang et al. “Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.” In: *Applied and Environmental Microbiology* 73.16 (Aug. 2007), pp. 5261–5267.

[53] Raul Cabrera-Rubio et al. "The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery." In: *American Journal of Clinical Nutrition* 96.3 (Sept. 2012), pp. 544–551.

[54] Schloss P. D. et al. "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities". In: *Applied and environmental microbiology* 75.23 (Dec. 2009), pp. 7537–751.

[55] Shahram Lavasani et al. "A novel probiotic mixture exerts a therapeutic effect on experimental autoimmune encephalomyelitis mediated by IL-10 producing regulatory T cells." In: *Plos One* 5.2 (Feb. 2010).

[56] Simon Carding et al. "Dysbiosis of the gut microbiota in disease." In: *Microbial Ecologhy* 26.2 (Feb. 2015).

[57] Todd R. Klaenhammer et al. "The impact of probiotics and prebiotics on the immune system." In: *Nature Reviews Immunology* 12 (Oct. 2012), pp. 728–734.

[58] Torsten Thomas et al. "Metagenomics - a guide from sampling to data analysis". In: *Microbial Informatics and Experimentation* 2.3 (Feb. 2012).

[59] Valérie Gaboriau-Routhiau et al. "The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses." In: *Immunity Journal* 31.4 (Oct. 2009), pp. 677–689.

[60] Victor Kunin et al. "A Bioinformatician's Guide to Metagenomics". In: *Microbiology and Molecular Biology Reviews* 72.4 (Dec. 2008), pp. 557–578.

[61] Wostmann BS et al. "Dietary intake, energy metabolism, and excretory losses of adult male germfree Wistar rats". In: *Journal of the American Association for Laboratory Animal Science* 33.1 (Feb. 1983), pp. 46–50.

[62] Dr. Istvan Albert. *Bioinformatics Handbook Guide*. Jan. 2017. URL: https://read.biostarhandbook.com/.

[63] M. J. Albert, V. I. Mathan, and S. J. Baker. "Vitamin B12 synthesis by human small intestinal bacteria". In: *Nature* 283.5749 (Feb. 1980), pp. 781–782. DOI: 10.1038/283781a0. URL: https://doi.org/10.1038/283781a0.

[64] Lupus foundation of America. *What is Lupus?* 2016. URL: http://www.lupus.org/answers/entry/what-is-lupus.

[65] Nicola L. Harris Andrew J. Macpherson. "Interactions between commensal intestinal bacteria and the immune system". In: *Nature Reviews: Immunology* 4 (June 2004), pp. 478–485.

[66] M Barbhaiya and KH Costenbader. "Ultraviolet radiation and systemic lupus erythematosus". In: *Lupus* 23.6 (May 2014), pp. 588–595. DOI: 10.1177/0961203314530488. URL: https://doi.org/10.1177/0961203314530488.

[67] Yasmine Belkaid and Timothy W. Hand. "Role of the Microbiota in Immunity and Inflammation". In: *Cell* 157.1 (Mar. 2014), pp. 121–141. DOI: 10.1016/j.cell.2014.03.011. URL: https://doi.org/10.1016%5C%2Fj.cell.2014.03.011.

[68] D Benton, C Williams, and A Brown. "Impact of consuming a milk drink containing a probiotic on mood and cognition". In: *European Journal of Clinical Nutrition* 61.3 (Dec. 2006), pp. 355–361. DOI: 10.1038/sj.ejcn.1602546. URL: https://doi.org/10.1038%5C%2Fsj.ejcn.1602546.

[69] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (Apr. 2014), pp. 2114–2120. DOI: 10.1093/bioinformatics/btu170. URL: https://doi.org/10.1093/bioinformatics/btu170.

[70] N. Braun et al. "Accelerated atherosclerosis is independent of feeding high fat diet in systemic lupus erythematosus-susceptible LDLr-/- mice". In: *Lupus* 17.12 (Dec. 2008), pp. 1070–1078. DOI: 10.1177/0961203308093551. URL: https://doi.org/10.1177/0961203308093551.

[71] Pier Luigi Buttigieg and Alban Ramette. "A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses". In: *FEMS Microbiology Ecology* 90.3 (Nov. 2014), pp. 543–550. DOI: 10.1111/1574-6941.12437. URL: https://doi.org/10.1111/1574-6941.12437.

[72] Pier Luigi Buttigieg and Alban Ramette. *A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses: Principal coordinates analysis*. Nov. 2014. URL: https://mb3is.megx.net/gustame/dissimilarity-based-methods/principal-coordinates-analysis.

[73] R. B. Sartor C. D. Packey. "Interplay of commensal and pathogenic bacteria, genetic mutations, and immunoregulatory defects in the pathogenesis of inflammatory bowel diseases." In: *Journal of Internal Medicine* 263.6 (June 2008), pp. 597–606.

[74] P. Desreumaux C. P. Tamboli C Neut and J. F. Colombel. "Dysbiosis in inflammatory bowel disease". In: *Gut* 53.1 (Jan. 2004), pp. 1–4.

[75] Mark Walport Charles A Janeway Jr Paul Travers. *Imunobiology: The Immune System in Health and Disease*. fifth. Garland Science, 2001.

[76] André M. Comeau, Gavin M. Douglas, and Morgan G. I. Langille. "Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research". In: *mSystems* 2.1 (Jan. 2017). Ed. by Jonathan Eisen, e00127–16. DOI: 10.1128/msystems.00127-16. URL: https://doi.org/10.1128%5C%2Fmsystems.00127-16.

[77] Human Microbiome Project Consortium. "A framework for human microbiome research." In: *Nature* 13.486 (June 2012), pp. 215–221. URL: http://www.nature.com/nature/journal/v486/n7402/extref/nature11209-s1.pdf.

[78] The Human Microbiome Project Consortium. "Structure, Function and Diversity of the Healthy Human Microbiome". In: *Nature* 486.7402 (June 2012), pp. 207–214.

[79] Lourdes C. Corman. "The role of diet in animal models of systemic lupus erythematosus: Possible implications for human lupus". In: *Seminars in Arthritis and Rheumatism* 15.1 (Aug. 1985), pp. 61–69. DOI: 10.1016/0049-0172(85)90010-1. URL: https://doi.org/10.1016%5C%2F0049-0172%5C%2885%2990010-1.

[80] S. P. Crampton, P. A. Morawski, and S. Bolland. "Linking susceptibility genes and pathogenesis mechanisms using mouse models of systemic lupus erythematosus". In: *Disease Models & Mechanisms* 7.9 (Aug. 2014), pp. 1033–1046. DOI: 10.1242/dmm.016451. URL: https://doi.org/10.1242/dmm.016451.

[81] Matthew F. Cusick, Jane E. Libbey, and Robert S. Fujinami. "Molecular Mimicry as a Mechanism of Autoimmune Disease". In: *Clinical Reviews in Allergy & Immunology* 42.1 (Nov. 2011), pp. 102–111. DOI: 10.1007/s12016-011-8294-7. URL: https://doi.org/10.1007%5C%2Fs12016-011-8294-7.

[82] M. Derrien. "Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium". In: *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY* 54.5 (Sept. 2004), pp. 1469–1476. DOI: 10.1099/ijs.0.02873-0. URL: https://doi.org/10.1099%5C%2Fijs.0.02873-0.

[83] Muriel Derrien, Clara Belzer, and Willem M. de Vos. "Akkermansia muciniphila and its role in regulating host functions". In: *Microbial Pathogenesis* 106 (May 2017), pp. 171–181. DOI: 10.1016/j.micpath.2016.02.005. URL: https://doi.org/10.1016%5C%2Fj.micpath.2016.02.005.

[84] Suzanne Devkota et al. "Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in Il10-/- mice". In: *Nature* (June 2012). DOI: 10.1038/nature11225. URL: https://doi.org/10.1038/nature11225.

[85] Scikit-bio documentation. *Newick format*. Dec. 2016. URL: http://scikit-bio.org/docs/0.4.2/generated/skbio.io.format.newick.html.

[86] Nicola Doniselli et al. "The identification of an integral membrane, cytochrome c urate oxidase completes the catalytic repertoire of a therapeutic enzyme". In: *Scientific Reports* 5.1 (Sept. 2015). DOI: 10.1038/srep13798. URL: https://doi.org/10.1038/srep13798.

[87] GABRIEL R. DRAPEAU and ROBERT A. MACLEOD. "A Role for Inorganic Ions in the Maintenance of Intracellular Solute Concentrations in a Marine Pseudomonad". In: *Nature* 206.4983 (May 1965), pp. 531–531. DOI: 10.1038/206531a0. URL: https://doi.org/10.1038%5C%2F206531a0.

[88] Mirjana Rajilič-Stojanovič E. G. Zoetendal. "High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota." In: *Gut* 57.11 (Nov. 2008), pp. 1605–1615.

[89] Robert C. Edgar. *Can SINTAX predict species?* Jan. 2017. URL: http://drive5.com/usearch/manual/sintax_species.html.

[90] Robert C. Edgar. *Chimera Formation*. Feb. 2017. URL: http://drive5.com/usearch/manual/chimera_formation.html.

[91] Robert C. Edgar. *Cross-talk*. Feb. 2017. URL: http://drive5.com/usearch/manual/crosstalk.html.

[92] Robert C. Edgar. *manual Rarefaction*. 2016. URL: http://drive5.com/usearch/manual/rare.html.

[93] Robert C. Edgar. *Operational Taxonomic Units (OTUs)*. Nov. 2106. URL: http://www.drive5.com/usearch/manual/otu_definition.html.

[94] Robert C. Edgar. *Singletons*. July 2016. URL: http://www.drive5.com/usearch/manual/singletons.html.

[95] Robert C. Edgar. *UCLUST algorithm*. Jan. 2017. URL: http://drive5.com/usearch/manual/uclust_algo.html.

[96] Robert C. Edgar. "UNCROSS: Filtering of high-frequency cross-talk in 16S amplicon reads". In: *bioRxiv* (Nov. 2016).

[97] Robert C. Edgar. "UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing". In: *bioRxiv* (Oct. 2016), p. 2016.

[98] Robert C. Edgar. *Uparse OTU analysis pipeline*. Online; accessed 28-November-2016. URL: http://drive5.com/usearch/manual/uparse_pipeline.html.

[99] Robert C. Edgar. *UPARSE-OTU algorithm*. Online; accessed 28-November-2016. URL: http://www.drive5.com/usearch/manual/uparseotu_algo.html.

[100] A. Everard et al. "Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity". In: *Proceedings of the National Academy of Sciences* 110.22 (May 2013), pp. 9066–9071. DOI: 10.1073/pnas.1219451110. URL: https://doi.org/10.1073%5C%2Fpnas.1219451110.

[101] Brent Ewing and Phil Green. "Base-Calling of Automated Sequencer Traces UsingPhred.II. Error Probabilities". In: *Genome Research* 8.3 (Mar. 1998), pp. 186–194. DOI: 10.1101/gr.8.3.186. URL: https://doi.org/10.1101/gr.8.3.186.

[102] Daniel P. Faith. "Conservation evaluation and phylogenetic diversity". In: *Biological Conservation* 61.1 (1992), pp. 1–10.

[103] Daniel P. Faith. "Conservation evaluation and phylogenetic diversity". In: *Biological Conservation* 61.1 (1992), pp. 1–10. DOI: 10.1016/0006-3207(92)91201-3. URL: https://doi.org/10.1016%5C%2F0006-3207%2892%2991201-3.

[104] Anthony P. Fejes. *What is a bioinformatician*. Jan. 2014. URL: http://blog.fejes.ca/?p=2418.

[105] Felsenstein and Seattle Kuhner lab Department of Genome Sciences University of Washington. *The Newick tree format*. June 2016. URL: http://evolution.genetics.washington.edu/phylip/newicktree.html.

[106] Davide Festi et al. "Gut microbiota and metabolic syndrome". In: *World Journal of Gastroenterology* 20.43 (2014), p. 16079. DOI: 10.3748/wjg.v20.i43.16079. URL: https://doi.org/10.3748%5C%2Fwjg.v20.i43.16079.

[107] National Science Foundation. *Collaborative Research: ABI Development: Extensible, reproducible and documentation-driven microbiome data science.* [Online; accessed 27-October-2016]. 2016. URL: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1565100.

[108] KEGG: Kyoto Encyclopedia of Genes and Genomes. *Caffeine metabolism - Reference pathway.* 2016. URL: http://www.genome.jp/dbget-bin/www_bget?pathway+map00232.

[109] KEGG: Kyoto Encyclopedia of Genes and Genomes. *Carotenoid biosynthesis - Reference pathway.* 2016. URL: http://www.genome.jp/dbget-bin/www_bget?pathway+map00906.

[110] KEGG: Kyoto Encyclopedia of Genes and Genomes. *Circadian rythm - Reference pathway.* 2016. URL: http://www.genome.jp/dbget-bin/www_bget?ko04712.

[111] KEGG: Kyoto Encyclopedia of Genes and Genomes. *Fatty acid elongation - Reference pathway.* 2016. URL: http://www.genome.jp/dbget-bin/www_bget?pathway+map00062.

[112] KEGG: Kyoto Encyclopedia of Genes and Genomes. *Flavonoid biosynthesis - Reference pathway.* 2016. URL: http://www.genome.jp/dbget-bin/www_bget?pathway+map00941.

[113] KEGG: Kyoto Encyclopedia of Genes and Genomes. *Fluorobenzoate degradation - Reference pathway.* 2016. URL: http://www.genome.jp/dbget-bin/www_bget?pathway+map00364.

[114] KEGG: Kyoto Encyclopedia of Genes and Genomes. *Penicillin and cephalosporin biosynthesis: Orthology.* 2016. URL: http://www.genome.jp/dbget-bin/get_linkdb?-t+orthology+path:map00311.

[115] KEGG: Kyoto Encyclopedia of Genes and Genomes. *Steroid biosynthesis - Reference pathway.* 2016. URL: http://www.genome.jp/dbget-bin/www_bget?pathway+map00100.

[116] KEGG: Kyoto Encyclopedia of Genes and Genomes. *Systemic lupus erythematosus pathway: map05322.* June 2016. URL: http://www.genome.jp/dbget-bin/www_bget?map05322.

[117] Kyoto Encyclopedia of Genes and Genomes. *Sulfur metabolism.* Feb. 2017. URL: http://www.genome.jp/dbget-bin/www_bget?pathway+ko00920.

[118] C Gordon et al. "Abnormal sulphur oxidation in systemic lupus erythematosus". In: *The Lancet* 339.8784 (Jan. 1992), pp. 25–26. DOI: 10.1016/0140-6736(92)90144-r. URL: https://doi.org/10.1016%5C%2F0140-6736%5C%2892%5C%2990144-r.

[119] J. C. GOWER. "Some distance properties of latent root and vector methods used in multivariate analysis". In: *Biometrika* 53.3-4 (1966), pp. 325–338. DOI: 10.1093/biomet/53.3-4.325. URL: https://doi.org/10.1093/biomet/53.3-4.325.

[120] HDF Group. *HDF5 Data Model*. Nov. 2016. URL: https://support.hdfgroup.org/HDF5/.

[121] Dr. Robert Heckendor. *Newick Format*. July 2016. URL: http://marvin.cs.uidaho.edu/Teaching/CS515/newickFormat.html.

[122] Gough SC Heward J. "Genetic susceptibility to the development of autoimmune disease." In: *Clinical Science* 93.6 (Dec. 1997), pp. 479–491.

[123] Lora V. Hooper, Tore Midtvedt, and Jeffrey I. Gordon. "How host-microbial interactions shape the nutrient environment of the mammalian intestine". In: *Annual Review of Nutrition* 22.1 (July 2002), pp. 283–307. DOI: 10.1146/annurev.nutr.22.011602.092259. URL: https://doi.org/10.1146/annurev.nutr.22.011602.092259.

[124] Elaine Y. Hsiao et al. "Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders". In: *Cell* 155.7 (Dec. 2013), pp. 1451–1463. DOI: 10.1016/j.cell.2013.11.024. URL: https://doi.org/10.1016%5C%2Fj.cell.2013.11.024.

[125] Lin B.F. Hsieh C.C. "Dietary factors regulate cytokines in murine models of systemic lupus erythematosus". In: *Autoimmunity Reviews* 1 (Nov. 2011), pp. 22–27.

[126] J. B. Hughes et al. "Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity". In: *Applied and Environmental Microbiology* 67.10 (Oct. 2001), pp. 4399–4406. DOI: 10.1128/aem.67.10.4399-4406.2001. URL: https://doi.org/10.1128/aem.67.10.4399-4406.2001.

[127] Susan M. Huse et al. "A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters". In: *PLoS ONE* 7.6 (June 2012). Ed. by Niyaz Ahmed, e34242. DOI: 10.1371/journal.pone.0034242. URL: https://doi.org/10.1371%5C%2Fjournal.pone.0034242.

[128] University of Idaho. *Principles of Vegetation Measurement & Assessment and Ecological Monitoring & Analysis*. 2016. URL: http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%5C%209(Composition%5C&Diversity)/9_2_Biodiversity.htm.

[129] Illumina. *bcl2fastq Conversion User Guide*. June 2016. URL: https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/bcl2fastq/bcl2fastq_letterbooklet_15038058brpmi.pdf.

[130] Illumina. *MiSeq Reagent Kit v2 Preparation Guide*. Oct. 2012. URL: https://support.illumina.com/content/dam/illumina-support/documents/myillumina/cebf8b82-b1d9-4384-a64b-002db4193cbe/miseqreagentkit_v2_reagentprepguide_15034097_b.pdf.

[131] Illumina. *MiSeq Sequencing Chemistry*. Dec. 2016. URL: https://support.illumina.com/content/dam/illumina-support/courses/MiSeq_Sequencing_Chemistry/story.html?iframe.

[132] Illumina. *MiSeq Sequencing Fundamentals*. Dec. 2016. URL: https://support.illumina.com/content/dam/illumina-support/courses/MiSeq_Sequencing_Fundamentals/story.html?iframe.

[133] Illumina. *Using a PhiX Control for HiSeq® Sequencing Runs*. Feb. 2017. URL: https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/hiseq-phix-control-v3-technical-note.pdf.

[134] Illumina Inc. *Illumina Sequencing by Synthesis*. Oct. 2016. URL: https://www.youtube.com/watch?v=fCd6B5HRaZ8.

[135] Open Osmosis Initiative. *Systemic lupus erythematosus (SLE) - causes, symptoms, diagnosis & pathology*. May 2016. URL: https://www.youtube.com/watch?v=0junqD4BLH4.

[136] NIH: National Human Genome Research Institute. *DNA Sequencing Costs: Data*. Dec. 2016. URL: https://www.genome.gov/sequencingcostsdata/.

[137] T. Ishizuka et al. "Human Carboxymethylenebutenolidase as a Bioactivating Hydrolase of Olmesartan Medoxomil in Liver and Intestine". In: *Journal of Biological Chemistry* 285.16 (Feb. 2010), pp. 11892–11902. DOI: 10.1074/jbc.m109.072629. URL: https://doi.org/10.1074/jbc.m109.072629.

[138] the Jackson Laboratory. *NZB Mice*. Aug. 2014. URL: https://www.jax.org/strain/000684.

[139] J. M. Janda and S. L. Abbott. "16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls". In: *Journal of Clinical Microbiology* 45.9 (July 2007), pp. 2761–2764. DOI: 10.1128/jcm.01228-07. URL: https://doi.org/10.1128%5C%2Fjcm.01228-07.

[140] Qingyun Yan. Yuhe Yu Jiajia Ni. "How much metagenomic sequencing is enough to achieve a given goal?" In: *Scientific Reports* 3 (Mar. 2013).

[141] M. E. V. Johansson et al. "The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria". In: *Proceedings of the National Academy of Sciences* 105.39 (Sept. 2008), pp. 15064–15069. DOI: 10.1073/pnas.0803124105. URL: https://doi.org/10.1073%5C%2Fpnas.0803124105.

[142] David Siegmund John D. Storey Jonathan E. Taylor. "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach". In: *Journal of Royal Statistical Society* 66.1 (Jan. 2004), pp. 187–205. URL: https://www.jstor.org/stable/3647634?seq=1#page_scan_tab_contents.

[143] Anouk Kaulmann and Torsten Bohn. "Carotenoids, inflammation, and oxidative stress—implications of cellular signaling pathways and relation to chronic disease prevention". In: *Nutrition Research* 34.11 (Nov. 2014), pp. 907–929. DOI: 10.1016/j.nutres.2014.07.010. URL: https://doi.org/10.1016/j.nutres.2014.07.010.

[144] Martin Kircher, Susanna Sawyer, and Matthias Meyer. "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform". In: *Nucleic Acids Research* 40.1 (Oct. 2011), e3–e3. DOI: 10.1093/nar/gkr771. URL: https://doi.org/10.1093/nar/gkr771.

[145] Dan Knights. *Discovering Patterns in the Microbiome Course*. Mar. 2016. URL: http://metagenome.cs.umn.edu/microbiomecodebrowser/doc/index.html.

[146] Adam P. Kohm, Kevin G. Fuller, and Stephen D. Miller. "Mimicking the way to autoimmunity: an evolving theory of sequence and structural homology". In: *Trends in Microbiology* 11.3 (Mar. 2003), pp. 101–105. DOI: 10.1016/s0966-842x(03)00006-4. URL: https://doi.org/10.1016%5C%2Fs0966-842x%2803%2900006-4.

[147] K. T. Konstantinidis and J. M. Tiedje. "Genomic insights that advance the species definition for prokaryotes". In: *Proceedings of the National Academy of Sciences* 102.7 (Feb. 2005), pp. 2567–2572. DOI: 10.1073/pnas.0409727102. URL: https://doi.org/10.1073/pnas.0409727102.

[148] MO BIO LABORATORIES. *PowerFecal® DNA Isolation Kit*. Mar. 2017. URL: https://mobio.com/media/wysiwyg/pdfs/protocols/12830.pdf.

[149] N. Lane. "The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1666 (Mar. 2015), pp. 20140344–20140344. DOI: 10.1098/rstb.2014.0344. URL: https://doi.org/10.1098%5C%2Frstb.2014.0344.

[150] Morgan G. I. Langille. *Analysis of Metagenomic Data: Introduction to PICRUSt*. July 2016. URL: https://www.youtube.com/watch?v=0eK2ycMgVmY.

[151] BINF Bioinformatics lessons and tutorials. *Pyrosequencing*. Feb. 2017. URL: https://binf.snipcademy.com/lessons/ngs-techniques/454-roche-pyrosequencing.

[152] BINF Bioinformatics lessons and tutorials. *Semiconductor Sequencing*. Feb. 2017. URL: https://binf.snipcademy.com/lessons/ngs-techniques/sequencing-by-ligation.

[153] BINF Bioinformatics lessons and tutorials. *Sequencing by Ligation*. Feb. 2017. URL: https://binf.snipcademy.com/lessons/ngs-techniques/sequencing-by-ligation.

[154] Z. Liu et al. "Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers". In: *Nucleic Acids Research* 36.18 (Sept. 2008), e120–e120. DOI: 10.1093/nar/gkn491. URL: https://doi.org/10.1093/nar/gkn491.

[155] Julien Loubinoux et al. "Sulfate-reducing bacteria in human feces and their association with inflammatory bowel diseases". In: *FEMS Microbiology Ecology* 40.2 (May 2002), pp. 107–112. DOI: 10.1111/j.1574-6941.2002.tb00942.x. URL: https://doi.org/10.1111/j.1574-6941.2002.tb00942.x.

[156] C. Lozupone and R. Knight. "UniFrac: a New Phylogenetic Method for Comparing Microbial Communities". In: *Applied and Environmental Microbiology* 71.12 (Dec. 2005), pp. 8228–8235. DOI: 10.1128/aem.71.12.8228-8235.2005. URL: https://doi.org/10.1128/aem.71.12.8228-8235.2005.

[157] M LYTE et al. "Induction of anxiety-like behavior in mice during the initial stages of infection with the agent of murine colonic hyperplasia Citrobacter rodentium". In: *Physiology & Behavior* 89.3 (Oct. 2006), pp. 350–357. DOI: 10.1016/j.physbeh.2006.06.019. URL: https://doi.org/10.1016%5C%2Fj.physbeh.2006.06.019.

[158] Citartan Marimuthu et al. "Single-stranded DNA (ssDNA) production in DNA aptamer generation". In: *The Analyst* 137.6 (2012), p. 1307. DOI: 10.1039/c2an15905h. URL: https://doi.org/10.1039%2Fc2an15905h.

[159] Frédéric Marin et al. "Molluscan Shell Proteins: Primary Structure, Origin, and Evolution". In: *Current Topics in Developmental Biology*. Elsevier, 2007, pp. 209–276. DOI: 10.1016/s0070-2153(07)80006-8. URL: https://doi.org/10.1016%5C%2Fs0070-2153%2807%2980006-8.

[160] Micah T McClain et al. "Early events in lupus humoral autoimmunity suggest initiation through molecular mimicry". In: *Nature Medicine* 11.1 (Dec. 2004), pp. 85–89. DOI: 10.1038/nm1167. URL: https://doi.org/10.1038%5C%2Fnm1167.

[161] Paul J. McMurdie and Susan Holmes. "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible". In: *PLoS Computational Biology* 10.4 (Apr. 2014). Ed. by Alice Carolyn McHardy, e1003531. DOI: 10.1371/journal.pcbi.1003531. URL: https://doi.org/10.1371/journal.pcbi.1003531.

[162] M.D. Melvyn R. Werbach. *Nutritional Influences on Illness*. 2002. URL: http://www.tldp.com/lupus/nutrition.html.

[163] National Research Council (US) Committee on Metagenomics. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. National Academies Press, 2007.

[164] Arin L. Zirnheld Michele M. Kosiewicz and Pascale Alard. "Gut microbiota, immunity, and disease: a complex relationship." In: *Frontiers in Microbiology* 2.180 (Sept. 2011), p. 2011.

[165] Sun Microsystems. *Open Grid Scheduler*. [Online; accessed 07-November-2016]. URL: http://gridscheduler.sourceforge.net/.

[166] Bradley S. Moore et al. "Plant-like Biosynthetic Pathways in Bacteria: From Benzoic Acid to Chalcone1". In: *Journal of Natural Products* 65.12 (Dec. 2002), pp. 1956–1962. DOI: 10.1021/np020230m. URL: https://doi.org/10.1021/np020230m.

[167] K. M. Neufeld et al. "Reduced anxiety-like behavior and central neurochemical change in germ-free mice". In: *Neurogastroenterology & Motility* 23.3 (Nov. 2010), 255–e119. DOI: 10.1111/j.1365-2982.2010.01620.x. URL: https://doi.org/10.1111%5C%2Fj.1365-2982.2010.01620.x.

[168] J. Gregory Caporaso et al. Nicholas A Bokulich. *mockrobiota: a public resource for microbiome bioinformatics benchmarking*. May 2016. DOI: 10.7287/peerj.preprints.2065v1/supp-1. URL: https://doi.org/10.7287/peerj.preprints.2065v1/supp-1.

[169] D. H. Parks et al. "STAMP: statistical analysis of taxonomic and functional profiles". In: *Bioinformatics* 30.21 (July 2014), pp. 3123–3124. DOI: 10.1093/bioinformatics/btu494. URL: https://doi.org/10.1093%5C%2Fbioinformatics%2Fbtu494.

[170] J. Petri M. Allbritton. "Antibiotic allergy in systemic lupus erythematosus: A case-control study." In: *Journal of Rheumatology* 19.2 (1992), pp. 265–269.

[171] Louis Legendre Pierre Legendre. *Numerical ecology*. second. Elsevier Science B.V., 2003.

[172] Earth Microbiome Project. *16S Illumina Amplicon Protocl*. Nov. 2016. URL: http://www.earthmicrobiome.org/protocols-and-standards/16s/.

[173] QIIME. *OTU picking strategies in QIIME*. Dec. 2016. URL: http://qiime.org/tutorials/otu_picking.html.

[174] QIIME. *Quantitative Insights Into Microbial Ecology*. [Online; accessed 27-October-2016]. 2016. URL: http://qiime.org/index.html.

[175] Santasabuj Das Rahul Shubhra Mandal Sudipto Saha. "Metagenomic Surveys of Gut Microbiota". In: *Genomics Proteomics Bioinformatics* 13.3 (June 2015), pp. 148–158.

[176] "Raising standards in microbiome research". In: *Nature Microbiology* 1.7 (June 2016), p. 16112. DOI: 10.1038/nmicrobiol.2016.112. URL: https://doi.org/10.1038/nmicrobiol.2016.112.

[177] "Raising standards in microbiome research". In: *Nature Microbiology* 1.7 (June 2016), p. 16112. DOI: 10.1038/nmicrobiol.2016.112. URL: https://doi.org/10.1038%5C%2Fnmicrobiol.2016.112.

[178] Howard Amital et al. Ram Reifen. "Linseed Oil Suppresses the Anti-beta-2-glycoprotein-I in Experimental Antiphospholipid Syndrome". In: *Journal of Autoimmunity* 15.3 (2000), pp. 381–385.

[179] MG RAST. *Illumina four-color sequencing by synthesis*. May 2014. URL: https://www.youtube.com/watch?v=tuD-ST5B3QA.

[180] Jacques Ravel and K Wommack. "All hail reproducibility in microbiome research". In: *Microbiome* 2.1 (2014), p. 8. DOI: 10.1186/2049-2618-2-8. URL: https://doi.org/10.1186%5C%2F2049-2618-2-8.

[181] Justus Reunanen et al. "Akkermansia muciniphila Adheres to Enterocytes and Strengthens the Integrity of the Epithelial Cell Layer". In: *Applied and Environmental Microbiology* 81.11 (Mar. 2015). Ed. by H. Goodrich-Blair, pp. 3655–3662. DOI: 10.1128/aem.04050-14. URL: https://doi.org/10.1128%5C%2Faem.04050-14.

[182] F. E. Rey et al. “Metabolic niche of a prominent sulfate-reducing human gut bacterium”. In: *Proceedings of the National Academy of Sciences* 110.33 (July 2013), pp. 13582–13587. DOI: 10.1073/pnas.1312524110. URL: https://doi.org/10.1073/pnas.1312524110.

[183] Fiona Powrie Richard Blumberg. “Microbiota, Disease, and Back to Health: A Metastable Journey”. In: *Science Translational Medicine* 6.4 (June 2012).

[184] June L. Round and Sarkis K. Mazmanian. “The gut microbiota shapes intestinal immune responses during health and disease”. In: *Nature Reviews Immunology* 9.5 (May 2009), pp. 313–323. DOI: 10.1038/nri2515. URL: https://doi.org/10.10382Fnri2515.

[185] G. D. Ruxton. “The unequal variance t-test is an underused alternative to Student’s t-test and the Mann-Whitney U test”. In: *Behavioral Ecology* 17.4 (Apr. 2006), pp. 688–690. DOI: 10.1093/beheco/ark016. URL: https://doi.org/10.1093%5C%2Fbeheco%2Fark016.

[186] Dieter Schlee et al. “Numerical Taxonomy. The Principles and Practice of Numerical Classification”. In: *Systematic Zoology* 24.2 (June 1975), p. 263. DOI: 10.2307/2412767. URL: https://doi.org/10.2307/2412767.

[187] Thermo Fisher Scientific. *Next Generation Sequencing Library Preparation*. Oct. 2015. URL: https://www.youtube.com/watch?v=_yC0Bzw3WbQ.

[188] Nicola Segata. “Computational meta’omics for microbial community studies”. In: *Molecular Systems Biology* 14.9 (May 2013).

[189] Ron Sender, Shai Fuchs, and Ron Milo. “Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans”. In: *Cell* 164.3 (Jan. 2016), pp. 337–340. DOI: 10.1016/j.cell.2016.01.013. URL: https://doi.org/10.1016%5C%2Fj.cell.2016.01.013.

[190] Ron Sender, Shai Fuchs, and Ron Milo. “Revised Estimates for the Number of Human and Bacteria Cells in the Body”. In: *PLOS Biology* 14.8 (Aug. 2016), e1002533. DOI: 10.1371/journal.pbio.1002533. URL: https://doi.org/10.13712Fjournal.pbio.1002533.

[191] Fergus Shanahan. “The gut microbiota-a clinical perspective on lessons learned”. In: *Nature Reviews Gastroenterology & Hepatology* 9.10 (Oct. 2012), pp. 609–614.

[192] Thomas J. Sharpton. “An introduction to the analysis of shotgun metagenomic data”. In: *Frontiers in Plant Science* 5 (June 2014). DOI: 10.3389/fpls.2014.00209. URL: https://doi.org/10.3389%5C%2Ffpls.2014.00209.

[193] Na-Ri Shin et al. “An increase in the Akkermansia spp. population induced by metformin treatment improves glucose homeostasis in diet-induced obese mice”. In: *Gut* 63.5 (June 2013), pp. 727–735. DOI: 10.1136/gutjnl-2012-303839. URL: https://doi.org/10.1136%5C%2Fgutjnl-2012-303839.

[194] Stewart Shuman and Michael S. Glickman. "Bacterial DNA repair by non-homologous end joining". In: *Nature Reviews Microbiology* 5.11 (Nov. 2007), pp. 852–861. DOI: 10.1038/nrmicro1768. URL: https://doi.org/10.1038/nrmicro1768.

[195] O.E. Pagovich S.M. Vieira and M.A. Kriegel. "Diet, microbiota and autoimmune diseases". In: *LUPUS* 23 (2014), pp. 518–526.

[196] Andres Baselga (British Ecological Society). *What is Beta Diversity?* May 2015. URL: https://methodsblog.wordpress.com/2015/05/27/beta_diversity/.

[197] Vassilis L. Souliotis et al. "Defective DNA repair and chromatin organization in patients with quiescent systemic lupus erythematosus". In: *Arthritis Research & Therapy* 18.1 (Aug. 2016). DOI: 10.1186/s13075-016-1081-3. URL: https://doi.org/10.1186%5C%2Fs13075-016-1081-3.

[198] J. D. Storey and R. Tibshirani. "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16 (July 2003), pp. 9440–9445. DOI: 10.1073/pnas.1530509100. URL: https://doi.org/10.1073%5C%2Fpnas.1530509100.

[199] D.L. Sun et al. "Intragenomic Heterogeneity of 16S rRNA Genes Causes Overestimation of Prokaryotic Diversity". In: *Applied and Environmental Microbiology* 79.19 (July 2013), pp. 5962–5969. DOI: 10.1128/aem.01282-13. URL: https://doi.org/10.1128%5C%2Faem.01282-13.

[200] Louise E. Tailford et al. "Mucin glycan foraging in the human gut microbiome". In: *Frontiers in Genetics* 6 (Mar. 2015). DOI: 10.3389/fgene.2015.00081. URL: https://doi.org/10.3389%5C%2Ffgene.2015.00081.

[201] T Tansey. "Structure and regulation of mammalian squalene synthase". In: *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1529.1-3 (Dec. 2000), pp. 49–62. DOI: 10.1016/s1388-1981(00)00137-2. URL: https://doi.org/10.1016%5C%2Fs1388-1981%2800%2900137-2.

[202] Agilent Technologies. *TapeStation Video: 2200 TapeStation System*. Oct. 2016. URL: TapeStation%20Video:%202200%20TapeStation%20System.

[203] Xin-Ping Tian. "Gastrointestinal involvement in systemic lupus erythematosus: Insight into pathogenesis, diagnosis and treatment". In: *World Journal of Gastroenterology* 16.24 (2010), p. 2971. DOI: 10.3748/wjg.v16.i24.2971. URL: https://doi.org/10.3748%5C%2Fwjg.v16.i24.2971.

[204] Sara Tucci and William Martin. "A novel prokaryotictrans-2-enoyl-CoA reductase from the spirocheteTreponema denticola". In: *FEBS Letters* 581.8 (Mar. 2007), pp. 1561–1566. DOI: 10.1016/j.febslet.2007.03.013. URL: https://doi.org/10.1016/j.febslet.2007.03.013.

[205] Haiting Wang et al. "Neutrophil Extracellular Trap Mitochondrial DNA and Its Autoantibody in Systemic Lupus Erythematosus and a Proof-of-Concept Trial of Metformin". In: *Arthritis & Rheumatology* 67.12 (Nov. 2015), pp. 3190–3200. DOI: 10.1002/art.39296. URL: https://doi.org/10.1002%5C%2Fart.39296.

[206] Q. Wang et al. "Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy". In: *Applied and Environmental Microbiology* 73.16 (June 2007), pp. 5261–5267. DOI: 10.1128/aem.00062-07. URL: https://doi.org/10.1128%5C%2Faem.00062-07.

[207] David I. Warton and Francis K. C. Hui. "The arcsine is asinine: the analysis of proportions in ecology". In: *Ecology* 92.1 (Jan. 2011), pp. 3–10. DOI: 10.1890/10-0340.1. URL: https://doi.org/10.1890/10-0340.1.

[208] George M. Weinstock. "Genomic approaches to studying the human microbiota". In: *Nature* 489.7415 (Sept. 2012), pp. 250–256. DOI: 10.1038/nature11553. URL: https://doi.org/10.1038%5C%2Fnature11553.

[209] Wikipedia. *Arcsine distribution*. 2017. URL: https://en.wikipedia.org/wiki/Arcsine_distribution.

[210] Wikipedia. *Cephalosporin-C deacetylase*. Aug. 2016.

[211] Wikipedia. *FASTQ*. Feb. 2017. URL: https://en.wikipedia.org/wiki/FASTQ_format.

[212] Wikipedia. *Metagenomics*. [Online; accessed 24-October-2016]. 2016. URL: https://en.wikipedia.org/wiki/Metagenomics.

[213] Wikipedia. *Phred quality score*. Feb. 2017. URL: https://en.wikipedia.org/wiki/Phred_quality_score.

[214] Erik Scott Wright and Kalin Horen Vetsigian. "Quality filtering of Illumina index reads mitigates sample cross-talk". In: *BMC Genomics* 17.1 (Nov. 2016). DOI: 10.1186/s12864-016-3217-x. URL: https://doi.org/10.1186%5C%2Fs12864-016-3217-x.

[215] Bo Yang, Yong Wang, and Pei-Yuan Qian. "Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis". In: *BMC Bioinformatics* 17.1 (Mar. 2016). DOI: 10.1186/s12859-016-0992-y. URL: https://doi.org/10.1186%5C%2Fs12859-016-0992-y.

[216] Baoli Zhu, Xin Wang, and Lanjuan Li. "Human gut microbiome: the second genome of human body". In: *Protein & Cell* 1.8 (Aug. 2010), pp. 718–725. DOI: 10.1007/s13238-010-0093-z. URL: https://doi.org/10.10072Fs13238-010-0093-z.