Gut Microbiota of HIV patients:

Investigation by IlluminaMiSeq Shallow Sequencing

MD A B M SHARIFUZZAMAN

Master degree in Applied Biotechnology

HEDMARK UNIVERSITY COLLEGE

2014

ACKNOWLEDGMENT

I hereby give thanks to all of the persons involved throughout this thesis and the masters program.

The key person of this research work was Professor Knut Rudi, the person whose brilliant guidance provides me every direction and supervision of this thesis. I also thank full to my Co supervisor Professor Robart Wilson.

Special thanks to Felix ChinweijeNwosu (PhD student of Professor Knut Rudi) for your enormous help both during lab work and the analysis. Your suggestion helps me a lot to understand the metagenomic pipeline MG-RAST.

I am indebted to Ekaterina Avershina (PhD student of Professor Knut Rudi) to give your valuable time during lab work.

TABLE OF CONTENTS

Abstract

The colonization of a healthy gastrointestinal (GI) tract is a primary target for HIV infection. HIV infection causes inflammation, which results in disruption of the mucosal surface of the gut and the breakdown of the gut wall integrity which causes the free flow of pathogenic microbes to the lymph nodes. HIV infection causes alteration of gut microbiota in affected subjects, compared with healthy subjects. However the pattern of gut microbiota depends on many variables. Studies supported that probiotic supplement could bring back the normal gut microbiota to some extent as well as possibly suppressing pro inflammatory responses. This study is aimed at investigating the relationship between the gut microbiota and HIV infection progression in the Swedish patients. There were fifteensubjects under this study and divided into five groups (three subjects in each) such as Baseline, Follow up, Elite control, Immune-deficient and Control. One subject of Elite control was rejected as it was failed QC of MG-RAST.The analysis of gut microbiota is carried out through library preparation and sequencing using theNextera XT DNA kit and MiSeqsystem, respectivelyat the in Norwegian Univeristyof Life Science. The taxonomical and functional abundance was analyzed by the help of the MG-RAST pipeline. The abundance between groups were significantly different in different taxonomical levels from phylum to genus. Firmicuteswas the only bacterial phyla whose abundance was found significantly different between the subject groups. *Akkermansiamuniciphila*was the only bacterial species which represent the bacterial phylum Verrucomicrobia, and weremost abundant in the immune-deficient group. Clostridia, bacteria that play a critical role in the body's immune defense mechanism, were less abundant in the control group than the infected groups. Bacterial families showed much more variety of abundance among the groups. Bacterial genes functioning for membrane transport, defense, virulence and disease were less abundant in immune-deficient group.

Keywords: HIV, gut microbiota, probiotics, MG-RAST, taxonomical abundance, functional abundance

ABBBREVIATIONS

µl- Microliter

ºC- Degree Celsius

AIDS- Acquired Immunodeficiency Syndrome

ANOVA- Analysis of Variance

ART- Anti Retroviral Treatment

ATP- Adenosine Tri Phosphate

BLAST- Basic Local Alignment Search Tool

bp- Base Pair

COG- Clusters of Orthologous Group

dATP-deoxyadenosine triphosphates

dCTP-deoxycytidine triphosphates

dGTP-deoxyguanosine triphosphates

DNA- Deoxyribonucleic acid

dNTO-deoxynucleotide triphosphates

dTTP-deoxythymidine triphosphates

eggNOG-evolutionary genealogy of genes: Non-supervised Orthologous Groups

FGA-Functional Gene Array

HIV- Human Immunodeficiency Virus

IL- Interleukin

IMG- Integrated Microbial Genome

KEGG: Kyoto Encyclopedia of Genes and Genomes

LPS- Lipopolysaccharide

MGA- Metagenome micro array

NAC-Normalized abundance count

ng-Nano gram

NGS- Next Generation Sequencing

NOD- Nucleotide-binding oligomerization domain

NOG- Non-supervised Orthologous Groups

NTA- Nextera XT TagmentAmplicon

PCA- Principle Component Analysis

PCR- Polymerase Chain Reaction

PPi- Inorganic pyrophosphate

RDP- Ribosomal Database project

RNA- Ribonucleic acid

SBS- Sequencing by Synthesis

SIV - Simian immunodeficiency virus

ssDNA-signle stranded DNA

sttdev – Standard deviation

Th- T helper

TLR- Toll like receptor

tRNA- transfer RNA

1. Introduction

1.1. Human microbiota

Human body harbors at least 100 trillion microbial cells(Whitman, Coleman, & Wiebe, 1998) which is far more than human cells (Goodman & Gordon, 2010) and quadrillion virus within and on our body(Haynes and Rohwer, 2011). They constitute the microbiota of human body and the genes encoded by them are referred as microbiome(Clemente, Ursell, Parfrey, & Knight, 2012). The role of microbiome in health and disease has been revealed by the application of advanced genomic processing technology and computational mapping of biological data from human microbial environment(Sweeney & Morton, 2013).Apparently there area complex interactions among these microbial community and with the host. So it is normal that human health and physiology experienced a great impact from this interaction (Clemente et al., 2012). Certainly as a consequence microbiota plays key role on human health and disease (O'Hara & Shanahan, 2006). Genetics of the host is one of the key factor for the establishment and shaping of the gut microbiota because host genomic loci influence the bacterial composition (Benson et al., 2010); (Spor, Koren, & Ley, 2011). Bacteria are the dominating microbes in the gut microbiota. While Bacteroidetes and Firmicutes are the most abundant division of bacteria in the gut (Turnbaugh et al., 2006). There are three types of relationship arises between human and bacteria residing into host from the dynamic evolutionary process e.g., symbiosis, commensalism and pathogenicity (Backhed, Ley, Sonnenburg, Peterson, & Gordon, 2005). It is estimated that human gut microbiota comprises approximately 500-1000 species (Ley, Peterson, & Gordon, 2006), (Dethlefsen, Huse, Sogin, & Relman, 2008) among them a small number dominate the community (Arumugam et al., 2011) which encode 100 folds more unique gene than the human genome(Ley et al., 2006). In recent years bacterial component of microbiota studied under large scale projects such as the Human Microbiome Projects (Group et al., 2009); (Turnbaugh et al., 2007) and MetaHIT(Qin et al., 2010). Accumulating evidence suggested that the gut microbiome has profound roles in modulating the host metabolism(Holmes, Li, Athanasiou, Ashrafian, & Nicholson, 2011), (Claus et al., 2011), (Diaz Heijtz et al., 2011). The intestinal microflora possess diverse complex of functional enzymes that are detrimental and beneficial (Z. J. Shu, Cao, & Halmurat, 2011). Gut bacteria plays a role in immune system development(Z. Shu, Ma, Tuerhong, Yang, & Upur, 2013). Human gastrointestinal tract can accommodate about half or more of our total immune cells in the body (Mowat & Viney, 1997). Interaction between gut microbiota and immune system give signals to promote the maturation of immune cells and

normal development of immune functions(Chow, Lee, Shen, Khosravi, & Mazmanian, 2010). There is a proposition that enteric viral replication and systematic pathogenesis has been promoted by intestinal microbiota (Kuss et al., 2011).Innate and adaptive immunity both play an important role in shaping of gut microbiota composition (Carvalho et al., 2012). Natural gut bacteria secrets one type molecule that interacts with the toll-like receptor 4 (TLR-4). Furthermore virus can make infection after this interaction (Pennisi, 2011).In the early stage of HIV infection, there is a very fast and significant damage to gut associated lymphoid tissues (GALTs) with profound depletion of Th17 cells. Intestinal bacterial is controlled by Th17 cells, a subset of CD4+ T cells (Brenchley et al., 2004). Depletion of Th17 cells leads the microbial translocation in the gastrointestinal tract (Gautreaux, Deitch, & Berg, 1994b) and (Gautreaux, Deitch, & Berg, 1994a)

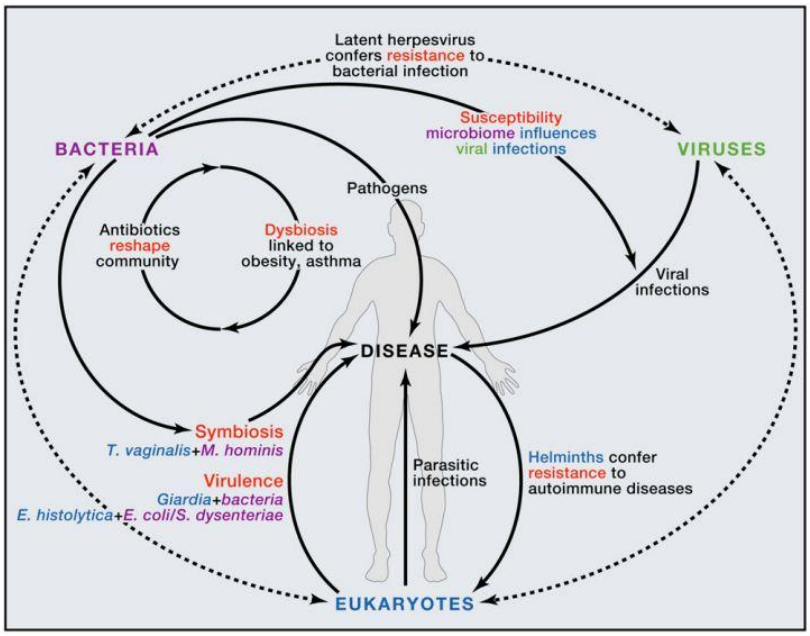


Figure 1.1: Effect of Interactions of Bacteria, Viruses, and Eukaryotes in Health and Disease."One microbe, one disease"- diseases were studied under this classical concept. Nevertheless now it is well established that a complex interaction between bacteria, virus and eukaryotes as well as the drug and some drugs shapes the disease phenotypes(Clemente et al., 2012).

There is a great variation in taxa present in gut along with the interindividual variation. Although the microbiota is stable within the individual throughout the life, external perturbation e.g., antibiotic can alter the composition by long term decrease in bacterial diversity (Clemente et al., 2012). There is a an evidence that reduction of antibiotic resistance pathogens following the reduced number of antibiotic treatment (Goossens, Ferech, Vander Stichele, Elseviers, & Group, 2005) which postulated that the repeated use of antibiotic can

increase the reservoir antibiotic resistance gene in our own microbiome (Sommer, Dantas, & Church, 2009) (figure 1.1). Evidence suggests that bacteria present in the amniotic fluid of anotherwith lower numbers and diversity, though babies are considered as sterile in the uterus (Jimenez et al., 2008).Immediately upon birth babies experience diverse environments and colonized by microbes either from mother's vagina or from skin depending on the delivery mode of the baby(Adlerberth & Wold, 2009); (Dominguez-Bello et al., 2010). Babies born normally have the vaginal microbiota present in their mother(figure 1.2). On the other hand caesarean babies have communities resemble to the skin(Dominguez-Bello et al., 2010). Caesarean babies have lower bacterial cells counts in their fecal samples but their antibody secreting cells are higher. Thus it is assumed that gut microbiota development influence the immunological development during the first year of life(Huurre et al., 2008).Initially the bacterial and viral load in the infant gut is low, and then there is a sharp increase during early development(Adlerberth & Wold, 2009); (Breitbart et al., 2008); (Koenig et al., 2011) and (Vaishampayan et al., 2010). It takes approximately 1.5 years to reach an adult like the bacterial composition in the gut, which starts from the end first year of life (Palmer, Bik, DiGiulio, Relman, & Brown, 2007)that is stable throughout the life(figure1.2)

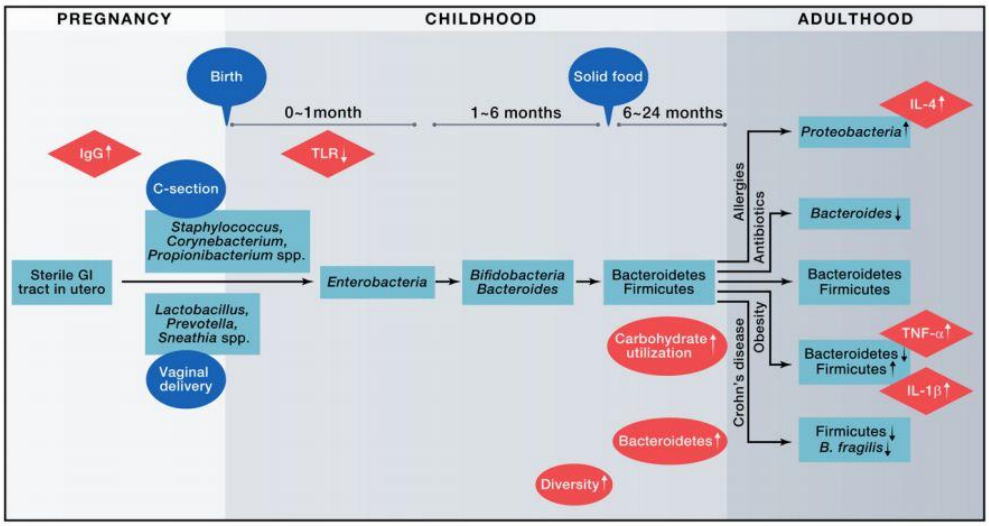


Figure 1.2: Development of microbiota during the life time(Clemente et al., 2012).The newborn baby is colonized while its gastrointestinal tract is sterile when it was in fetus status. The colonization varies either vaginal or skin like pattern of mother depending upon its delivery mode. The reduced activity of Toll-Like Receptor (TLR) allows the formation of stable bacterial microbiota in the gut during first 4 weeks of birth. The community goes towards an adult like pattern as the baby grows up. During this period the immune system can recognize the commensal and pathogenic bacteria. Although different disease conditions can change the pattern of microbiota of the gut despite it is quite stable during adulthood where Bacteroidetes and Firmicutes are most abundant divisions (Clemente et al., 2012).

1.2.Host immune system and microbiota:

The mammalian immune system specially the mucosal immune system in human has a complex correlation with his microbiota.IgA plays a fundamental role in mucosal immunity (Clemente et al., 2012). The colonization of specific commensal bacteria can induce IgA and protects mucosal surfaces and contribute to host-microbiota mutuality (Bouskra et al., 2008); (Macpherson, Geuking, & McCoy, 2011) and (Peterson, McNulty, Guruge, & Gordon, 2007). An experiment on germ free mice with reduced gut secretary IgA, showed deficiency in gut-associated lymphoid tissue during developmental stage, smaller Peyer's patches and mesenteric lymph nodes (Round & Mazmanian, 2009). Bacterial cell wall (lipopolisaccharide and peptidoglycane) and flagellin contains specific microbe-associated molecular patters (MAMPs) those are recognized by innate immune system. Host uses several proteins e.g., Toll-like receptors (TLRs) to recognize these types of antigens. The gut and mucosal immune system do not response properly if these proteins are absent or mutated (O'Hara & Shanahan, 2006). The commensal bacteria interact with TLRs to suppress the inflammatory response and improve the immunological tolerance (O'Hara & Shanahan, 2006) and (Round et al., 2011).There is different multi-protein complexes (inflammasomes) expressed in myeloid cells (Martinon, Burns, & Tschopp, 2002). As they are responsible for the activation of inflammatory process (Mariathasan et al., 2004) these multi-protein oligomers can sense exogenous and endogenous damage associated molecular patterns. NOD-like receptors (NLRs) recognize microbial molecules and can form inflammasomes. Reduced level of IL-18 has been observed as a consequence of deficient NLPRP6 which results alteration in microbiota and intestinal hyperplasia (Elinav et al., 2011). Commensal microbiota can also modulate the host adaptive immune system. Differentiation of T-cell population can also be affected by the microbiota within the host (Lathrop et al., 2011).

1.3.Microbial translocation:

In pathogenic human immunodeficiency virus (HIV) infections, microbial translocations are proposed to be an important influential event in chronic immune activation. Disease progression is closely associated with the disease progression. A series of immune pathological events takes place in the gastrointestinal tract prior to HIV associated microbial translocation. The early and severe mucosal $CD4^{+}$ depletion, hyperactivity of mucosal immune system, disruption in intestinal epithelium integrity along with enterocyte apoptosis and tight junction disruption and finally downfall of microbiome where the opportunistic

bacteria dominate the community (Marchetti, Tincati, & Silvestri, 2013). In normal condition the gastrointestinal tract is functionally and anatomically integrates. Hence the translocating microorganisms and microbial products are phagosytosed within the lamina propia and mesenteric lymph nodes. In contrary microbial translocation is a nonphysiological passage of gastrointestinal microflora and microbial products through the breach of intestinal epithelial barrier and lamina propia and in the end local mesenteric lymph nodes and finally to the systemic circulation (Wolochow, Hildebrand, & Lamanna, 1966), (Fuller & Jayne-Williams, 1970a) and (Fuller & Jayne-Williams, 1970b). Hence microbial translocation modifies the gut microbiome by antibiotics, pre and probiotics as well as transfer of microbial bioproducts from the gut (figure 1.3). In addition to it can reduce mucosal immune activation such as IL-7, IL-17 and IL-22 (Marchetti et al., 2013).

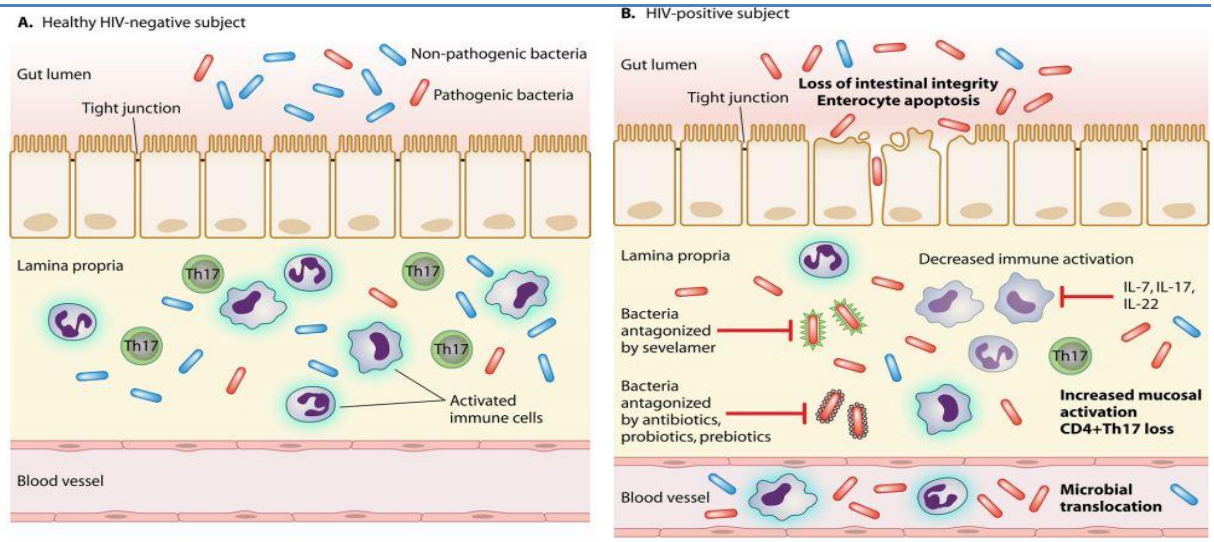


Figure 1.3: HIV associated damage to the gastrointestinal tract. (A) The gastrointestinal mucosal barrier of a healthy, HIV-negative subject is anatomically and functionally integrated is such a way that microbial translocation induces mucosal immune activation and limits the microbial translocation to the peripheral blood. (B) On the other hand HIV-positive patients have intestinal barrier which loses its tight junctions, enterocyte apoptosis, local immune activation and depletion of $CD4^+$ Th17 cells. Therefore pathogenic bacteria and microbial products are translocated from the gut lumen to systemic circulation through lamina propia(Marchetti et al., 2013).

1.3.1. Immune activation due to microbial translocation:

Microbial translocation results T-cell activation. Subsequently, it persuades the HIV disease progression. Chronically HIV-infected patients exhibit higher level of circulating LPS than the HIV negative individuals suggesting increased level of microbial translocation in the patients. The LPS level in plasma has positive correlation with innate and adaptive immune

activation (Brenchley et al., 2006). Another *in vitro* experiment showed that exposure of microbial TLR ligands affect T-lymphocyte activation thus increase the immune activation to the extreme level in the HIV- infected patients. Thus microbial translocation is a critical factor for immune activation (Funderburg et al., 2008).

1.3.2. Characterization of stool microbiota in association with translocating flora in HIV infected patients:

Gut microbiota has a profound effect on the maintenance of immunity and health. To maintain both local and systemic immunity in the normal state it is important that the composition of gut microbiota should to be specific(Paiardini, Frank, Pandrea, Apetrei, & Silvestri, 2008). Experiments on germ free mice showed that it gains different components of mucosal immunity after the repopulation of microbial flora in the gut(Tlaskalova-Hogenova et al., 2004) and (Umesaki, Okada, Matsumoto, Imaoka, & Setoyama, 1995) although it is difficult to determine the magnitude of the effects or causes of gut microbiota in different disease conditions. But in the animal models the gut microbiota has been altered dramatically both quantitative and compositional level(McKenna et al., 2008). In the early stages of HIV infection patients exhibit damaged fecal flora than the normal condition where it is dominated by the pathogens such as*Pseudomonas aeruginosa* and *Candida albicans* then the beneficial bacteria such as bifidobacteria and lactobacilli(Gori et al., 2008). HIV positive patients have higher contribution of proinflammatory/ inflammatory inducing bacterial order, e.g.,*Enterobacteriales*and*Bacteroidales* in their gut compared to healthy controls.Deudenal T-cell activation is negatively correlated with the stool bacterial load. The duodenal CD4+ T-cell loss and peripheral CD8+ T-cell activation has been identified by the level of fecal DNA of bacterial order *Enterobacteriales*and*Bacteroidales(*Ellis et al., 2011). Another study demonstrated that after the ART treatment on HIV patient's gut has enriched withimmune modulator and antiinflammatory bacterial family *Enterobacteriaceae*(Merlini et al., 2011).

1.3.3. Modification of gastrointestinal microbiota as a treatment of microbial translocation:

Probiotics are living microorganism which upon administration provides beneficial health benefits to the host (Reid, 2010) by inhibiting proinflammatory cytokines, decreasing gut permeability and stimulating the mucosal immunity (Marchetti et al., 2013). Therefore beside antibiotic treatment the use of probiotics is considered as an alternative way to alter the gut microbiota. Studies on HIV-infected patients in Tanzania demonstrated that probiotic yogurt

consumption was associated with $CD4^+$ T-cell augmentation (Irvine et al., 2010). Another studies demonstrated that ART along with probiotic supplementation provide a benefits to the SIV infected macaque by increment of antigen presenting cell frequency and function, enhanced T-cell immunity in intestinal mucosa, Th17 cells becoming more polyfuctional and reduces T-cell activation (Klatt et al., 2012).

1.4.Response of gastrointestinal immune system to the HIV infection:

The mucosal surface of the gastrointestinal tract serves as a protective barriers as a part of body's immune system. In addition to this mucosal immune system contains the greater part of the body's T-cells. HIV specific CD8+ T-cells plays a role in controlling the viral replication which persists in chronic infection (Koup et al., 1994). Experiments suggested that gastrointestinal tract is a major site of HIV replication following infection to the mucosal surface which results destruction of $CD4^+$ T-cell (Brenchley & Douek, 2008) due to weaker immune activity in gut. It is assumed that the CD4+ T-cells in the gastrointestinal tract are infected, 10 times more by the virus than the peripheral blood(Mehandru et al., 2007).

1.5.**Effect of probiotics on gastrointestinal tract of HIV patients:**

In the early stage of HIV disease the level of gut CD4+ Th17 declines in response to the alteration of the gut in normalmicroflora. A study of model animal having inflammatory bowel disease showed that combination of probiotic bacteria can up regulate the Tregactivation, which suppress the proinflammatory immune response in the animals. Preliminary studies showed that probiotic supplements enhance the growth and protection of CD4+ T-cells(Cunningham-Rundles et al., 2011). Therefore, it is assumed that probiotic bacteria can provide benefit in HIV treatment specially children who were infected before the development of their gut flora(Dicks, Fraser, ten Doeschate, & van Reenen, 2009).

1.6.Metagenomics:

In Greek, *meta* means "transcendent". In the field of microbial ecology metagenomics is one of the most remarkable events to understand the microbial genetic recourses in a given environment. It is a combination of genomic technologies and Bioinformatics tools. Hence metagenomics enables the researchers to analyze the functional gene composition in addition to this phylogenetic analysis of a microbial community. There are a huge number of microorganisms which are not culturable. So to understand the above mentioned composition of a microbial community, it is necessary to get genetic information from all microorganisms

present over there. The great potentiality of metagenomics is that it overcomes this limitation, therefore researchers can go through all the genetic information from culturable or unculturable microorganism under examination. A typical sequence-based metagenomics study involves sampling, processing for DNA sequencing, assembly, binning, annotation and analysis.

1.6.1. Sampling:

Sampling is the first and crucial step of a metagenomic study in a sense that all the representative cells should contribute to the extracted DNA and subsequent library preparation for sequencing. Based on the host or environmental condition from where these are collected. Each sample requires separate protocols for DNA extraction (Venter et al., 2004), (Burke, Kjelleberg, & Thomas, 2009) and (Delmont, Robe, Clark, Simonet, & Vogel, 2011). The microbial community closely associated with the host such as the gut microbiota of human, fractionation or selective lysis is a suitable method of DNA extraction from the microbes rather than the host (Burke et al., 2009; Thomas et al., 2010). For the metagenomic study of viruses from sea water, physical fractionation is suitable because a certain part of the total community will undergo for study. In this circumstances selective filtration, centrifugation or flow cytometry is suitable to enrich the target sample fraction (Angly et al., 2006; Palenik, Ren, Tai, & Paulsen, 2009; Venter et al., 2004). The representative extraction of microbial DNA from the soil direct lysis might cause the loss of DNA prior to sequencing thus make a bias on the result. Therefore, to maximize the DNA extraction, physical separation or isolation of cells should be employed. Enzyme inhibitor can interfere the subsequent processing there for this technique might helps help to avoid coexistence (Delmont et al., 2011). In some metagenomic studies where the total yield of DNA is too little so that it is not a suitable starting material for library preparation because modern sequencing technologies require high nano grams or micrograms of DNA. To overcome this situation amplification of the DNA sample is a good alternative but amplification of DNA sample might cause a bias on result due to reagent contamination, chimera formation or bias during amplification (Thomas, Gilbert, & Meyer, 2012). Multiple displacement amplification (MDA) is a tool which uses random hexamers and phage phi29 polymerase (Lasken, 2009).Therefore, it is essential to determine this strategy is suitable prior to sampling of a metagenomic study.

1.6.2. Sequencing technologies for metagenomic analysis:

The purpose of DNA sequencing is to determine the sequential order of the nucleotide bases on a target DNA molecule. Therefore scientist can use it for molecular breeding, cloning, finding pathogenic genes, and finding comparative and evolutionary studies. In that case the technologies should be fast, accurate, cheap and user friendly.Fredrick Snager developed the first DNA sequencing technologies(Sanger, Nicklen, & Coulson, 1977). The sequence chemistry was based on the chain termination method where Walter Gilbert developed another sequencing technology based on chemical modification of DNA and subsequent cleavage at specific bases (Liu et al., 2012). But Sanger's method was well accepted and adopted as the primary technology because it was efficient. In addition to this it used less radioactive molecules than Gilbert's method. But both of them were laborious in that time because they require expert handling. In 1987 Applied Biosystems introduced the first automatic sequencing technology machine (AB370). This technology adopted capillary electrophoresis to make sequencing faster and more accurate. Human genome project where Sanger sequencing technology was used as a main tool is a milestone for DNA sequencing technologies. This project stimulated the necessity of the development of the existing sequencing technology in different aspects such as increasing the speed parallel to accuracy, reducing the cost and increasing the automation. In addition to this it has accelerated the development of next generation sequencing (NGS) technologies (Collins, Morgan, & Patrinos, 2003). Consequently 454 lunched 454 in 2005 and Solexa released Genome Analyzer the next year and Agencourt provide the SOLiD platform. There were the primary NGS technologies with good performance in throughput, speed, accuracy and cost compare to the Sanger sequencing platform. Later Applied Biosystems purchased Agencourt on 2006, Roche purchased 454 on 2007 and Solexa purchased by Illumina.

1.6.2.1.Roche 454 System:

It is the first commercially successful NGS platform based on 'sequencing by synthesis' chemistry. It is called 'pyrosequencing technology' based on the detection of pyrophosphate released during nucleotide incorporation (Mardis, 2008). The template DNA with 454 specific adaptors are single stranded and immobile by amplification beads (figure 1.4). Then the dNTP's (dATP, dTTP, dCTP and dGTP) are incorporated to the template DNA by a PCR reaction with the help of ATP sulfurylase, luciferase, luciferin, DNA polymerase and adenosine 5′ phosphosulphate (APS) and released pyrophosphate (PPi) which equals to the

amount of incorporating nucleotide.After incorporating the correct complementary dNTP's to the ssDNA template bypolymersit releases pyrophosphate (PPi) stoichiometrically. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5´ phosphosulfate which later converts the luciferin to oxyluciferin with the help of luciferase enzyme. This conversion generates visible lights that are proportional to the amount of ATP detected by a specific camera and analyzed. In the last step of this PCR reaction apyrase degrades and removs the unincorporated nucleotides and ATP and the reaction continues with another nucleotide(Ronaghi, Uhlen, & Nyren, 1998) (figure 1.5). 454 GS FLX Titanium system was lunched in 2008 with increased read length and accuracy. In 2009 Roche upgraded their system which simplified the library preparation and data processing steps(Roche, 2011).

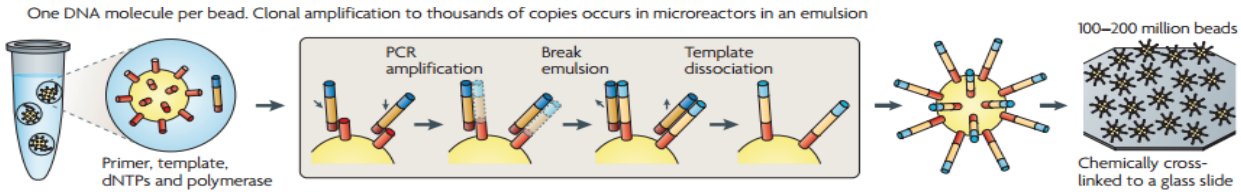


Figure 1.4: Template immobilization in Roche 454 platform. A bead-bound library of ssDNA is emulsified with amplification reagents thus several thousands of copies of the same template sequencehave been created by PCR amplification. These beads are chemically crosslinked to a glass slide or deposited into PicoTiterPlate wells(Metzker, 2010).

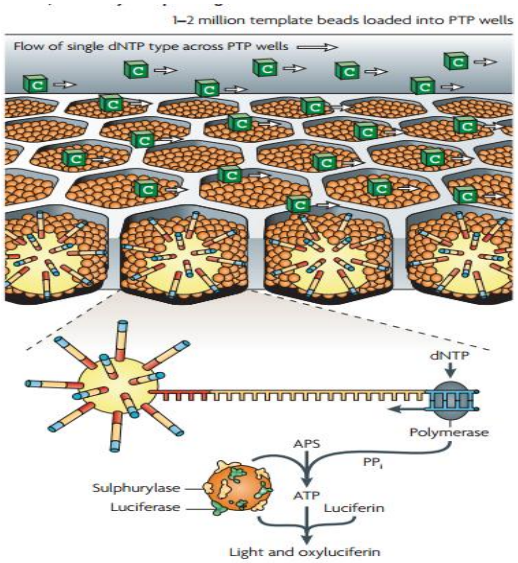


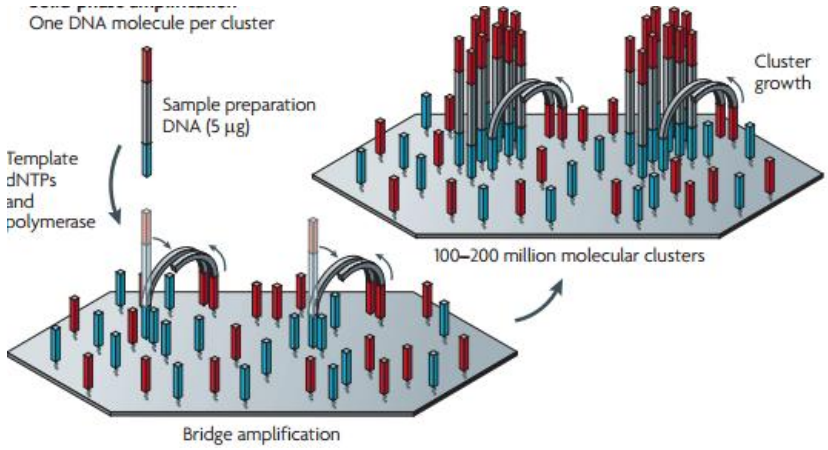
Figure 1.5:Pyrosequencing using Roche/454's Titanium platform. For understanding a single type of 2′-deoxyribonucleoside triphosphate (dNTP) — cytosine — is shown here. When the cytosine makes a bond with the complementary base on the template it release a pyrophosphate (PPi) which converted to ATP by ATP sulphurylase. This ATP converts luciferin to oxyluciferin. This conversion generates visible lights which are detected by the charge coupled device (CCD) (Metzker, 2010).

1.6.2.2.AB SOLiD system:

This system adopts ‘sequencing by ligation’ where the libraries can be sequenced by 8-base probe ligation on a SOLiDflowcell. The probe contains ligation site on its first base, cleavage site on its fifth base and four different fluorescent dyes attached to the last base. These probes are annealed and ligated; the preferential ligation by DNA ligase for matching sequences leads to a signal informing of the nucleotide at that position. Prior to ligation the template DNA is PCR amplified and attached to a glass slide. A fluorescent signal will be captured during the annealing and cleavage of the last three bases of the probes. After five rounds of sequencing using the loader primer set the template is deduced (Huse, Huber, Morrison, Sogin, & Welch, 2007). SOLiD system preferentially applies on whole genome resequencing, targeted resequencing, transcriptome research and epigenome studies(Liu et al., 2012).

1.6.2.3.Illumina GA/HiSeq system:

This system adopts the technology of, sequencing by synthesis (SBS) where a library of DNA molecule with fixed adaptors are denatured and attached to the flowcell. A cluster of clonal DNA fragments form after a bridge amplification (figure 1.6). Before sequencing the cluster of libraries are single stranded by a linearization enzyme. Then all four kinds of dNTP’s which contain differential cleavable dyes and removable blocking groups would complement the template one base at a time and emit a signal which is captured by a charge-coupled device (CCD) (figure 1.7). It can also be applied for whole-genome and region sequencing, transcriptome analysis, small RNA discovery, methylation profiling, and genome -wide protein-nucleic acid interaction analysis (M. Meyer & Kircher, 2010).HiSeq 2000 system in the early 2010 which adopts the same system of sequencing but with increased per run and introduced multiplexing where it could handle thousands of samples simultaneously. Among these three NGS systems IlluminaHiSeq 2000 have the biggest output with lowest reagent cost, SOLiD system has the highest accuracy and Roche system has the longest read length(Liu et al., 2012).



Figue 1.6: Primer immobilization and template clustering in Illumina GA platform. A primer immobilized to a solid support where a single molecule template is primed and bridge amplified to its adjacent primers to form clusters (Metzker, 2010).

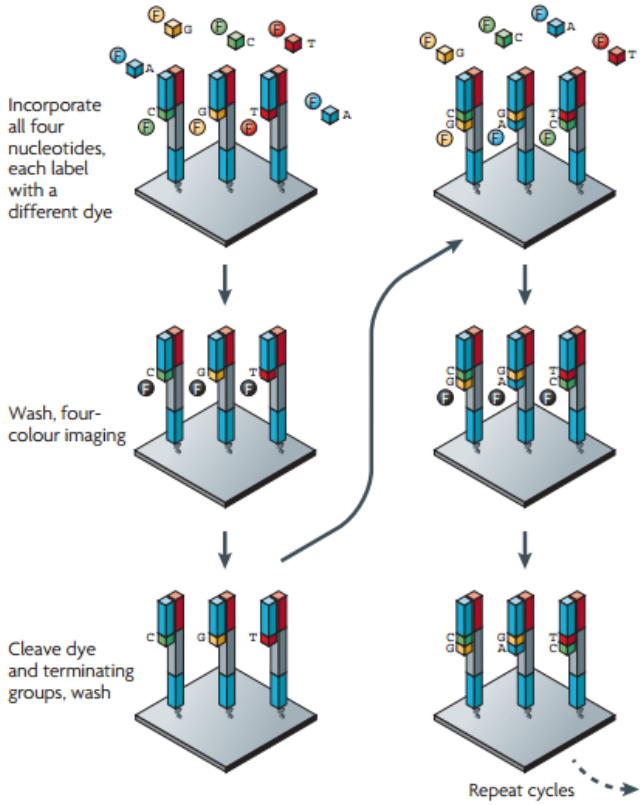


Figure 1.7: The four-colour cyclic reversible termination (CRT) method used by Illumina were. Following washing and imaging each successful incorporation of radiolebeled bases has been captured. A cleavage step removes the fluorescent dyes and regenerates the 3′-OH group using the reducing agent tries (2-carboxyethyl) phosphine (TCEP)(Metzker, 2010).

1.6.2.4.Compact personal genome machine (PGM) sequencer:

Ion Personal Genome Machine (PGM) and MiSeq were launched by Ion Torrent and Illumina respectively. The main advantages of these two systems is they are small in size, feature first turnover rates, but limited data throughput comfortable for clinical application and small labs(Liu et al., 2012).Ion PGM was first released in 2010 which is a first commercial sequencing machine that does not use fluorescence and camera scanning actually using semiconductor technology. By detecting the changes of pH when a nucleotide has been incorporated to the DNA molecule by the polymerase enzyme, PGM recognized whether the nucleotide is added or not. If the correct nucleotide has incorporated, change of voltage will be found. Change of voltage is equal to the number of nucleotides incorporated (Flusberg et al., 2010). There was an outbreak of exceptionally virulent Shiga-toxin (Stx) producing *Escherchia coli* O104:H4 in Germay at the middle of the year 2011(Mellmann et al., 2011) and (Rohde et al., 2011). Scientist used Ion Torrent PGM and HiSeq 2000 for the whole genome sequencing to identify the type of *E. coli*. Therefore PGM shows its potentiality in faster sequencing but limited throughput when there is an outbreak of new disease (Liu et al., 2012).

1.6.2.5.MiSeq Sequencer:

For rapid and cost-effective genetic alasysMiSeq sequencing system is a good choice which is based on synthesis by sequencing technology (SBS) employed by the same company Illumina(Illumina, 2014a). This single instrument integrates the functions of cluster generation, SBS and data analysis where it can complete the sequence within a day (8 hours)(Liu et al., 2012). The reversible terminator based method detects the incorporated single bases into the cluster of DNA strands. The bases are fluorescently labeled that are imaged upon incorporation and then cleaved to allow the incorporation of the next bases(Illumina, 2014a). In addition to this Nextera andTruSeq are used in this modern and innovative system. It has broder range of applicability which makes it promising for the next era of sequencing such as amplicon sequencing, clone checking, ChIP-Seq and small genome sequencing(Liu et al., 2012).

Table 1.1:comparison of different next-generation sequencing platforms.

	Roche454's GS FLX Titanium	Illumina/ Solexa's GA	Life/APG's SOLiD	HiSeq 2000*
NGS chemistry*' **	Pyrosequencing**	Real time**	Cleavable probe SBL**	Sequencing By Synthesis (SBS)*
Read length (bases)	330**	75 or 100**	50**	Up to 150 bases***
Accuracy	99.9%*	N/A	99.94% (raw data)*	98% (101PE)*

*(Liu et al., 2012)

**(Metzker, 2010)

***(Quail et al., 2012)

Instrument price may differ among the region. FR=Fragment run, Mate-pair run=MPR

Table 1.2: Technical specifications of benchtop instruments MiSeq 2000 and Ion Torrent PGM (Quail et al., 2012)

	IlluminaMiSeq	Ion Torrent PGM
Sequence/run	1.5-2 Gb	20-50 Mb on 314 chip 100-200 Mb on 316 chip 1 Gb on 318 chip
Run time	27 hours***	2 hours
Observed raw error rate	0.80%	1.71%
Read length	Up to 150 bases	~ 200 bases

*All cost calculations are based on list price quotations obtained from the manufacturer and assumed expected sequence yield stated.

***Includes two hours of clusters generation.

1.6.3. Assembly:

To obtain a longer genomic contigs from the genome of uncultured organisms than assembly of a short read fragments is a preferable strategy, although the development of a true

assembler is still an early stage. Researchers utilize two different strategies for the assembling of metagenomic samples: reference-based (co-assembly) and *de novo* assembly. If the reference genome of the organism which is under metagenomic study is available than reference based assembly is a suitable strategy for assembling of the contigs. The main advantages of this type of assembly is it can be done in the small machine for instance a laptop. They are also less time consuming. The algorithms used by these packages are fast and suitable for small machines. The available software packages are Newbler, AMOS, or MIRA (Thomas et al., 2012)**.** As metagenomics is newly evolved sequencing techniques to explore the microbial ecology therefore over 90% of microbes in metagenomic data are unknown (Wooley, Godzik, & Friedberg, 2010). *De novo* assembly doesn't utilize any reference genomes for this purpose. In addition to this it uses a large machine setup and all are based on the de Bruijn graphs such as EULER, Abyss, Velvet, SOAP. These all are for single genome therefore they do not work well for metagenomic datasets, except for some very small datasets containing specific species (Pop, 2009). Longer sequences make them easier to compare with the know genomic dataset which is better for gene annotation(Wommack, Bhavsar, & Ravel, 2008). The taxonomic assignment or phylogenetic analysis needs a specific cutoff length of the sequence, i.e., MG-RAST requires only 75 bp or longer for gene prediction for taxonomic binning and functional classification. Another objective of the assembler is to reduce data while reads are assembled to cluster rather than contigs. The MG - RAST pipeline uses clustering for data reduction (Thomas et al., 2012). None of the strategies are bias-free, but with increased accuracy.

1.6.4. Binning:

The process of sorting the DNA sequences into groups based on the conserved nucleotide composition (compositional binning) or the representative genes that represents a known gene in a reference database (similarity based binning) of an individual genome or genomes from a closely related organism. Phylopythia, S-GSOM, PCAHIER and TACAO are the compositional based binning algorithm, but these are not suitable for short reads as smaller the length the bias frequency is increased. IMG/M, MG-RAST, MEGAN, CARMA, Sort-ITEMS and MetaPhyler are the example of similarity based binning. Programs like MetaCluster and PhymmBL employed both compositional and similarity based binning alogorithms. Self-organising maps (SOMs) or hierarchical clustering or user defined binning are different methods of binning the DNA sequences (Thomas et al., 2012).

1.6.5. **Annotation**:

If the minimal length of contigs produced after the assembly of the reads are 30000 bp or longer and the objective of the study is to reconstruct the genome, then the existing pipeline for genome annotation is preferable such as RAST (Aziz et al., 2008) or IMG (Markowitz et al., 2009). On the other hand, if the reads are unassembled or in short contigs in length, then the annotation can be performed on the entire community. In this case the tools for genome annotation is less useful for metagenomic analysis. In these circumstances special tools for the annotation of the metagenomic short reads are developed.. In metagenomics study this type of sequences is annotated in two steps e.g., feature prediction and functional annotation (Thomas et al., 2012). A DNA sequence whose protein or function is unknown, but the open reading frames of the DNA sequences reflects the characteristics of a gene than this is called a putative gene. A number of algorithms have been developed to identify this CDS from the whole genome (Lukashin & Borodovsky, 1998). FragGeneScan(Rho, Tang, & Ye, 2010), MetaGeneMark(McHardy, Martin, Tsirigos, Hugenholtz, & Rigoutsos, 2007), MetaGeneAnnotator (MGA)/Metagene(Noguchi, Taniguchi, & Itoh, 2008) and Orphelia(Hoff, Lingner, Meinicke, & Tech, 2009)are used as a tool to predict the metagenomic CDS's. But these tools miss a significant subset of genes for example FragGeneScan which is better than most other methods gives 70% true positive rates but BLAST-based searches can potentially annotate these missing genes (Thomas et al., 2012). The pipeline for featuring the non protein coding genes require significant computational resources because beside the protein coding gene, there are a lot of non protein coding genes reside into the genome such as tRNAs(Lowe & Eddy, 1997), signal peptides (Bendtsen, Nielsen, von Heijne, & Brunak, 2004) or CRISPERs (Bland et al., 2007). The MG - RAST pipeline uses FragGeneScan (FGS) to find out the genes in short reads and similarity searches for ribosomal RNAs it employes SILVA (Pruesse et al., 2007), Greengenes(DeSantis et al., 2006), and RDP (Cole et al., 2009) databases. SILVA is an online tool for the alignment of small and large subunit of ribosomal RNA sequences from bacteria, archaea and eukarya. FGA and MGA employed by both CMERA's RAMCAPP pipeline (Sun et al., 2011) and IMG/M. Beside these IMG/M uses other tools also. It is estimated that 20 to 50% metagenomic sequences can be annotated so there is a major computational challenge on

functional annotation (Gilbert et al., 2010). ORFans are the sequence which cannot be mapped with the know sequences may be due to error in algorithms or the gene product is involved in an unknown biochemical function or may be their product (protein) may have structural similarity with a know protein though there is no homology between their DNA sequences (Godzik, 2011). KEGG (Kanehisa, Goto, Kawashima, Okuno, & Hattori, 2004), eggNOG(Muller et al., 2010), COG/KOG (Tatusov et al., 2003), PFAM (Finn et al., 2010) and TIGRFAM (Selengut et al., 2007)aremetagenomic database for functionalannotation. To maximize the functional annotation of metagenomic dataset it is necessary to merge them on a common platform because a single reference database can not cover all the biological functions e.g., MG-RAST and IMG/M. These platforms should share their data with other platforms. MG-RAST is a fully automated and data analysis pipeline, which is optimized between accuracy and computational efficiency for short reads. Provides feature prediction and functional annotation in the form of abundance profiles with quality controlling facilities (Thomas et al., 2012). MG-RAST has more than 12000 users and 131,567 data sets of which 19,000 are publicly available and 58.81 Terabases analyzed as of August 2014 reflects the centralization and standardization of resources and data sets. IMG/M performs hidden Markov model (HMM) and BLASTX searches which give higher sensitivity and comparison is based on all Vs all genes. The user of CAMERA (Sun et al., 2011) requires good knowledge on annotation of data and the analysis pipeline to interpret the result with better confidence level. MEGAN uses BLAST searches to visualize the annotated results as functional and taxonomic dendogram which makes the analysis easy (Huson, Auch, Qi, & Schuster, 2007).

1.7.Nextera® XT DNA sample preparation kit:

The Nextera XT DNA sample preparation kit enables the researchers to make a sequencing ready library by a single tagmentation enzymatic reaction following an optimized limited cycle PCR reaction. This protocol is suitable to make sequencing ready libraries for bacteria, archaea and viruses. Those having small genome, PCR amplicons and plasmids. It requires a very small amount of DNA (1 ng) which makes it suitable to sequence the sample which have very limited availability. This sample preparation kit along with MiSeq sequencing platform enables the researchers to get the sample data within a single day (8 hours). Following tagmentation of the sample where the samples are independently fragmented and tagged with adaptors(figure 1.8), the sample normalization steps is very simple compare to other protocols. The bead-based sample normalization steps avoid the quantification of thelibrary

before sample pooling because this sample preparation kit enables the researchers to prepare the libraries at equivalent concentrations. This sample preparation kit uses an indexed primer set which enables Barcoding up to 384 samples uniquely in a single experiment. Therefore up to 384 samples can be pooled together and are ready for sequencing. These indexed primers assign the reads from to the proper sample (Illumina, 2014b).

After the simultaneous fragmentation of the sample DNA by an enzymatic reaction following a limited cycle PCR reaction, both ends of the PCR sample are indexed with duel indexing strategy with two 8 base sequences (appendix 1). Dual indexing is enabled by adding a unique Index 1 (i7) adaptors (N701-N712) adjacent to the P7 sequence and Index 2 (i5) adaptors (S501-S508) adjacent to the P5 sequence to each sample for the 96 samplesNextera XT Index Kit. Arrange index 1 (i7) primer tubes (orange caps) in horizontal and index 2 (i5) primers (white caps) in vertical order so that N701 is in column 1 and N712 is in column 12 and that S501 is in row A and S508 is in row H(Illumina, 2012) (appendix 2).

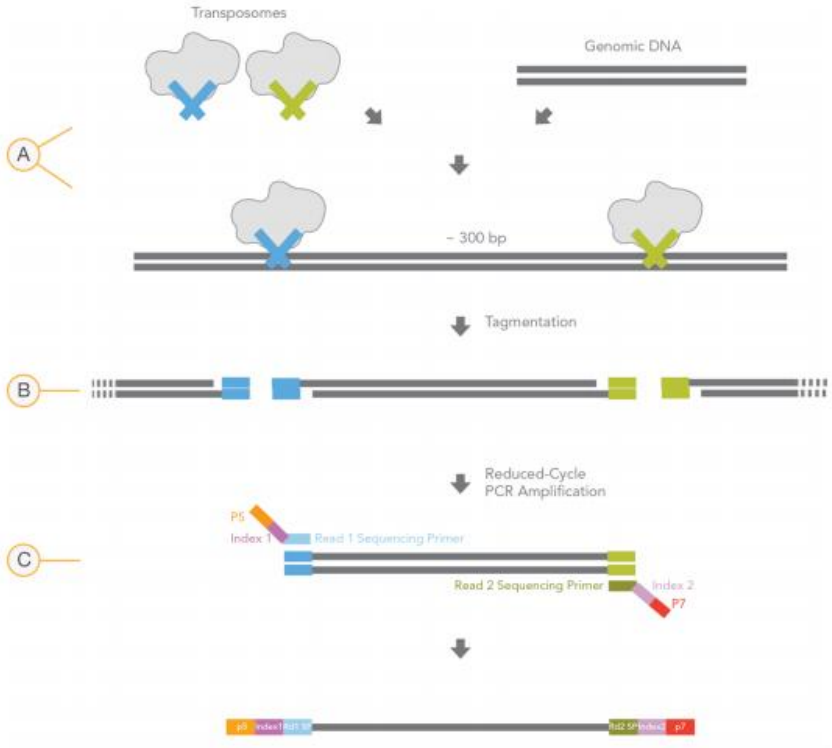


Figure 1.8:Library preparation by Nextera XT sample preparation kit. a) Binding of template DNA with Nextera XT transposom with adaptor. b) random fragmentation of template DNA, adaptors added on both sides. c) Limited cycle PCR to add sequencing primer sequences and indices.(Illumina, 2012).

1.8.The metagenomic RAST (MG-RAST) server:

RAST stands for Rapid Annotation using Subsystems Technology where MG-RAST means RAST for metagenomes. MG-RAST server is a Minimum Information about a Genome Sequence (MIGS) compliant proposed by The Genomic Standards Consortium provides web-based interface with easy access. The interface is designed in way to easy handling of browsing and analyzing the data. The platform of open source metagenomic RAST high-throughput pipeline provides many sequence analysis services in addition to functional annotation and phylogenetic comparisons, which is based on a SEED (McNeil et al., 2007)framework for comparative genomics by comparing both protein and nucleotide databases. In addition to SEED framework other open source and publicly available components such as NCBI BLAST(Altschul et al., 1997), other SEED subsystems, SEED nr, FIGfam,SQLite and Sun Grid Engine is used by this pipeline. It provides a framework so that the user can share others data sets for analysis, therefore it provides a new model for metagenomic data annotation and analysis removing the requirement the high performance computational set up as before. This pipeline normalizes (removing the duplicate sequences) and process the raw data (in FASTA format) automatically, but the user can upload the raw unassembled data or assembled contigs. In the second step the screening of potential protein encoding genes (PEGs) present in the normalized data is performed by using BLASTX (Altschul et al., 1997)search where the e-value cutoff value set to 0.01 just because of increasing the number of potential genes. It also uses rDNA, RNA and boutique databases. INSDC databases sourced the SEED comprehensive non-redundant databases, other sources are used for searching(Overbeek et al., 2005). In the next step the computation of the matched data with the other databases are analyzed. The phylogenetic information stored in SEED nr database and similarities of the data to the ribosomal RNA database are used to compute the phylogenetic reconstruction of the metagenomeand functional annotation of the sequence were done by SEED FIGfams(F. Meyer, Overbeek, & Rodriguez, 2009)and subsyatems. The number of pagesis identified by subsystem comparison tool connected to subsystem. Each sample is scored by dividing the number of sequences that are similar to a protein in each subsystem and the number of sequences from the sample that are similar to any protein in a subsystem. The taxonomic profiles of different samples are highlighted in the taxonomic heat

map which is determined by the phylogenetic or phylogenomicapproces. In both cases samples may be grouped in a nonquantitative fashion. In case of functional annotation the samples are grouped by the subsystem scores, but in taxonomic assignments the predominant phylogenetic groups are highlighted (F. Meyer et al., 2008).

1.8.1. P-value tool:

It allows to carry out a statistical test to conclude if there is a "significant" difference in the abundance in a given category across the specified group of samples. It is necessary to separate the samples into two or groups to perform a statistical test that determine the presence of the level of significance. In MG-RAST the group selection can only perform from the PCA analysis tool. The metagenomes were divided into five groups and stored. Each time after log in in this step should repeat if the samples undergoing significance tests. Four types of tests are available in MG-RAST version 3 where the tests of significance are selected based on two criteria but they are selected automatically.

Table 1.4:P-value test selections. This tool automatically chooses the best out of four statistical tests. Testes are selected based on the data type and number of groups.

	Normalized data Use Parametric Testes	Raw data Use non-parametric tests
Tests for 2 groups	Use Non-parametric Tests	Mann-Whitney test
Tests for 3 or more groups	One way ANOVA	Kruskal-Wallis test

1.8.2. Normalization and standardization:

Normalization reshapes a primary distribution with log transformation. As the biological variables exhibit log-normal distribution, therefore the data exhibit a normal distribution after normalization. In addition to this normalized data more satisfy the assumptions of downstream tests such as ANOVA or t-test. It is necessary to remove the inter sample variability among the groups to make the data more comparable. Therefore standardization is required, which is a kind of transformation applied to a group of distribution so that all exhibit same mean and standard deviation.The analysis page calculates the ordination

visualizations with either raw or normalized counts, at the user's option. The normalization procedure is as follows.

$$\text{Normalized value}_i = \log_2 (\text{raw counts}_i + 1)$$

The standardized values are then calculated from the normalized values by subtracting the mean of each samples normalized values and dividing by the standard deviation of each sample normalized values.

$$\text{Standardized}_i = (\text{normalized}_i - \text{mean (normalized}_i)) / \text{stddev (normalized}_i)$$

As all the taxonomic and functional analysis of this study is based on normalized data and there are 5 groups of metagenomes under comparison so the significance test is one way ANOVA test.

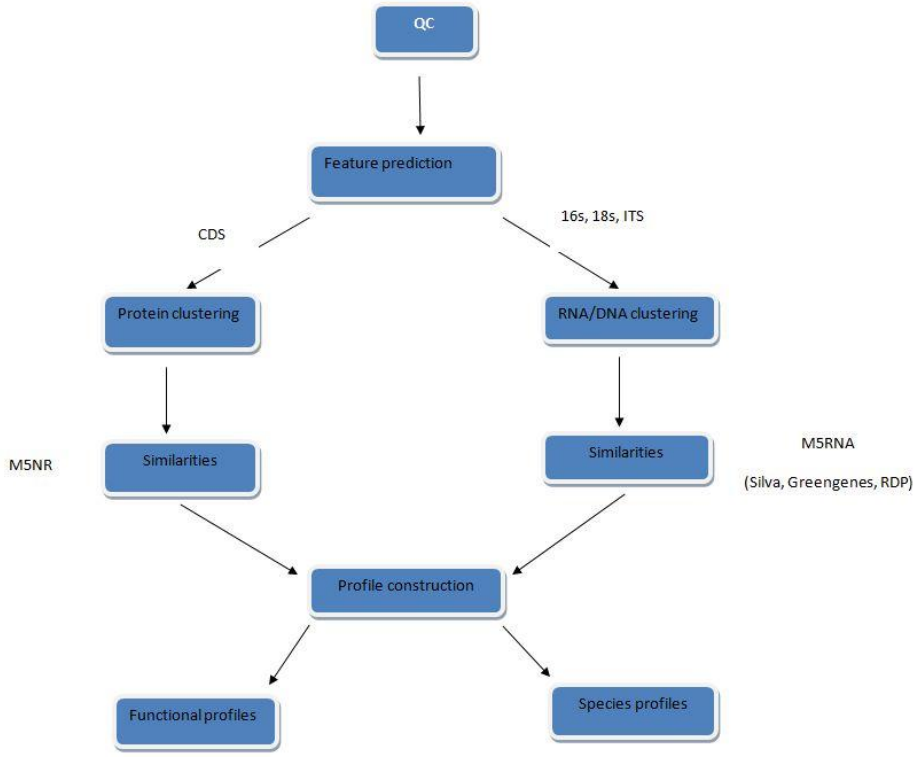


Figure 1.9: Overview of processing pipeline MG-RAST v3. In older versions where BLAST had used for computing similarities, here BLAT is employed which faster than the BLAST. In this pipeline after uploading the metadata it has normalized (it is also possible to analyze the raw data) and resulting abundance profiles are fed into different databases. There the organisom and functional abundance level is constructed (Andreas Wilke, 2013).

1.9.Significance and aim of the study:

Human Immune Deficiency (HIV) causes AIDS in human along with some other mammalians. Its primary site of infection is the gut. The human gut harbor microorganisms which are beneficial, some are commercial and some are pathogenic. The pattern of this microbiota altered due various disease conditions such as diarrhea, depression, inflammatory bowel disease, etc. In addition to these gut microbiota can alter due to various non disease related conditions such as geography, food habit. Much evidence abounds of the contribution of microbial community in the human gut to many disease conditions. HIV causes inflammation in the gut wall, therefore the pattern of the gut microbiota has changed. Understanding of alteration of gut microbiota due to HIV is still evolving and many experiments are running in different part of the world considering different parameters. However, according to previous studies, AIDS patients experienced improved condition into their gut after giving a probiotic supplement.

The aim of this project is to:

a. Identify the gut microbiota of Swedish HIV patients.
b. Understand whether any differences exist in the bacterial community among HIV infected and uninfected subjects.

2. Material and Methods:

2.1.Study Design:

The study runs on the Swedish patients who were divided into five groups, e.g., baseline, follow up, immune deficient, elite control and control. Baseline and follow up groups consist of the same patients. The difference between them was that base line is the condition prior to the probiotic treatment. The immune deficient group consistedwithHIV infected patients who developed intoAIDS. The patients in the elite control group had HIV but did not progress to AIDS. Lastly the control group consist of the people who did not have HIV. The stool sample from three people from each group were collected and stored under -20°C. DNA was extracted from the stool of these subjects and examined for metagenomic study (figure 2.1).

2.2.Library preparation by Nextera XT DNA sample preparation kit:

In the first step input DNA was tagmented (tagged and fragmented) by the Nexter XT transposom where the transposom fragments the input DNA and add adaptor sequences to the ends and subsequent PCR amplification. Only 1 ng DNA is needed for tagmentation. The DNA concentration of each sample was determined by Qubit® Fluorometerbefore tagmentation so that each sample has same contribution to the pool. After the adaptor sequences were added to the end of the template DNA were PCR amplified to generate multiplexed sequencing libraries. As the DNA concentrations were too low to make a pool so they undergo amplification by Illumina colony primers. The DNA amplicons were run on an agarose gel of 1% concentration at 80 voltages for forty minutes to get sure that they were successfully amplified.

2.2.1. Determination of the concentration of input DNA:

Prior to the tagmentation of input DNA it is necessary to determine their concentration. Thus, each sample can make the same contribution to the pool ready for sequencing. At first working solution was prepared by mixing Qubit reagent and Qubit buffer by 1:199 ratio. To draw a standard curve two standard solutions were made by this working solution at the ratio of 10:190 where first one had (10 µl in volume) DNA negative and the second one had 10 µl DNA solution (1ng/µl). After drawing a standard curve another solution was prepared for each sample where the ration was 1:199 (sample:working solution). The fluorescence of each solution was detected and by comparing with the previous fluorescence v's concentration curve their DNA concentration was also measured.

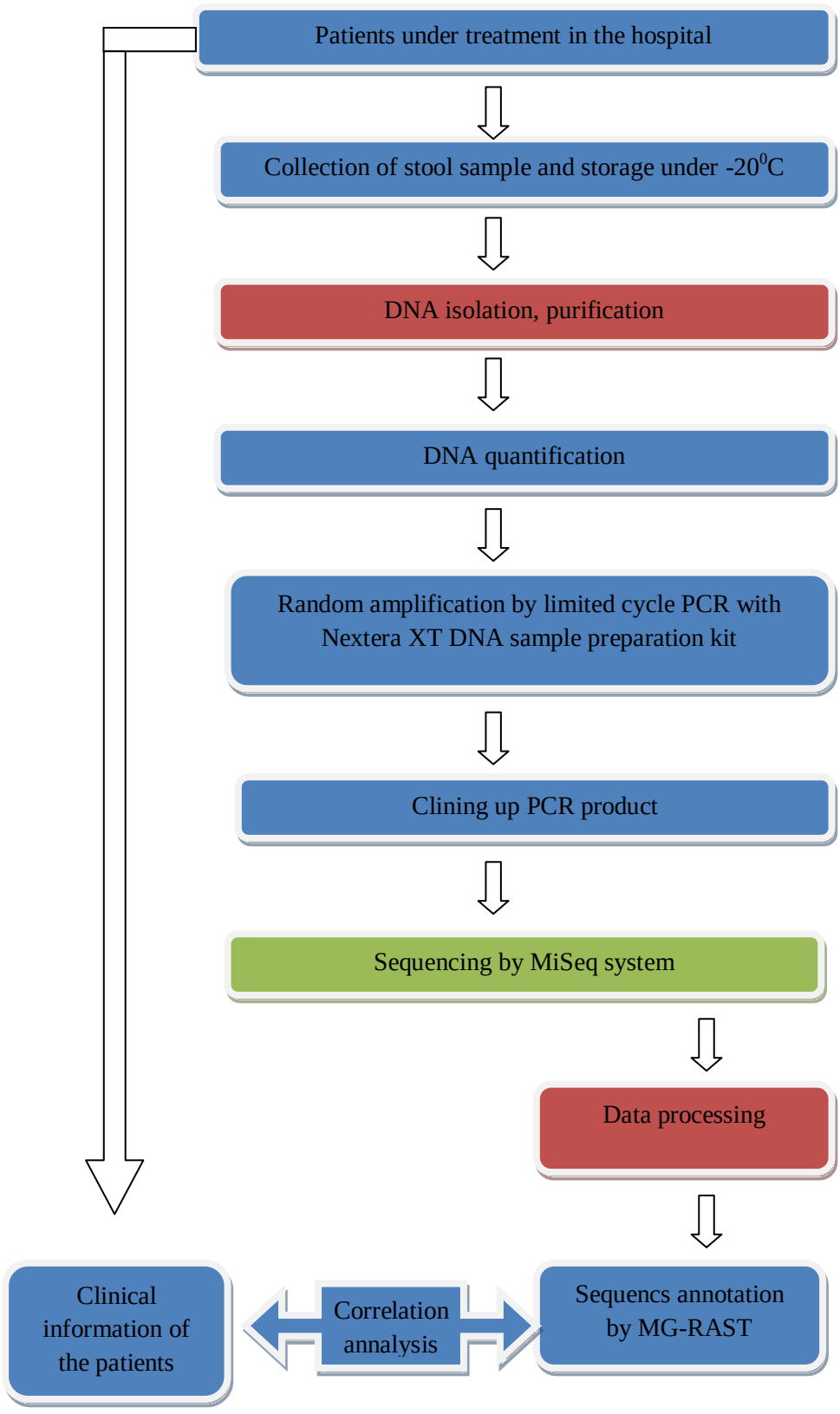


Figure 2.1: The outline of the scheme. Procedure in the red boxes done by Felix Nwosu, PhD candidate of Professor Knut Rudi. The procedure in the green box was done in NMBU.

2.2.2. Tagmentation of input DNA:

To make an NTA plate,5 µl of input DNA at 0.2 ng/ µl (1 ng total) was added to 10 µl of tagment DNA buffer and finally added to 5 µl of amplicontagment mix (ATM). The mixing was taken place in a 96-well TCY plate. A multichannel pipette was used to mix them well. Gently pipette 5 times up and down. Following centrifugation in a plate centrifuge machine for 1 minute, the sealed plate with a plastic seal underwent a short run on a thermocycler (55ºC for 5 minutes). The mix was ready for neutralization when it reaches to 10ºC. Then 5 µl of neutralize tegment buffer was added to each well and centrifugation for 1 minute the palette was kept at room temperature for 5 minutes. After the incubation 15 µl of the Nextera PCR master mix was added to each well of NTA plate. The combination of primers was determined earlier and 5 µl of the primer solution was added according to the list. The tips was changed before going to next the row or column. Finally the orange and white caps were changed to avoid cross contamination in the future. The plate was centrifuged for 1 minute after gentle pipetting for 3-4 times and sealing with a plastic and perform a PCR reaction where initial denaturation was 72ºC for 3 minutes then the denaturation temperature was increased to 95ºC for 30 seconds. Then a cycle of (12x) of reaction was performed at the temperature 95ºC, 55ºC and 72ºC where the duration was 10 seconds, 30 seconds and 30 seconds respectively. The final elongation step was carried out at the same temperature of 72ºC for 5 minutes in order to ensure full extension of any remaining single-stranded DNA. The reaction was then finished and kept at the temperature between 2ºC-8ºC (up to 2 days). The NTA PCR product was run at 1% agarose gel for 40 minutes at 80 voltages to check its quality and the DNA concentration of each PCR product was measured by Qubitfluorometer.

2.2.3. PCR clean up:

AMPure XP bead was used topurify the library DNA in addition to this remove very short library fragment from the population. For this purpose at first 50 µl of NTA PCR product was transferred to another plate. Before mixing the PCR product with the 50 µl 1X AMPure XP beads and incubation for 5 minutes at room temperature, the AMPure XP beads were vortexed for 30 seconds. Then the plat was placed on a magnetic stand until the supernatant had cleared. The supernantent was removed and discarded carefully while the plate was still on the magnetic stand. The beads were washed twice by freshly prepared 80% ethanol by following steps while the plate was on a magnetic plate. 200 µl freshly prepared 80% ethanol was added to each well, but the beads were not resuspended and incubated for 30 seconds.

After incubation, supernatent was removed and discarded. The beads undergo the second wash following removing the excess ethanol. The plate was kept on the magnetic plate and air dried for 15 minutes. After air drying, the plate was removed from the magnetic plate and 52.5 µl Resuspension Buffer (RSB) was added to each well and gently pipette for 10 times. After incubation for 2 minutes at room temperature, it was then placed on a magnetic plate again until the supernantent was cleared (at least 2 minutes). Then the clean PCR product was obtained after transferring the supernatant from the plate while it was still on the magnetic plate. Following the measurement by Qubitfluorometer, the clean PCR products were analyzed by running on 1% agarose gel for 40 minutes at 80 volts. As the concentration of the clean PCR product was too low for pooling it undergo colony PCR by illumine colony primers. A PCR master mixwas firstly prepared on ice in DNA-free area of the laboratory to avoid any form of DNA contamination (table 2.1). The gently vortexed 6.75 µl of master-mix, 2 µl DNA template and 16.25 µl of dH_2O was then gently poured into different wells of the plate. PCR reaction was performed where initial de-naturation was 95ºC for 15 minutes, Then a cycle of (12x) three step PCR reaction was performed where de-naturation temperature continued at the same temperature for 30 minutes, 55ºC and 72ºC where the duration was 30 seconds in both steps. The final elongation step was carried out at the same temperature of 72ºC for 5 minutes in order to ensure full extension of any remaining single-stranded DNA. The reaction was then finished and kept in the thermo-cycler at the temperature 10ºC. The concentration of the PCR products was measured by Qubitfluorometer. A pool was prepared according to the concentration of the PCR products where each PCR product has the same contribution to the pool. Each of them contributed 20 ng DNA. After making a pool the mix undergoes PCR clean up process again described earlier.

2.2.Sequencing, assembly and binning:

The prepared DNA libraries were sequenced in MiSeq system installed in Norwegian University of Life Sciences, Norway, and assembly and binning of the sequenced data were done by Felix Nowsu. The data were normalized prior to analysis. Then the taxonomic and functional abundance was annotated by MG-RAST. All files uploaded to MG-RAST was named by using alphanumeric and.-_ characters without spaces. As there were no files larger than 50 MB or less than 1 MB so it is unnecessary to make a zip file. There are three different files such as sequence files (FASTA, FASTAQ or SFF formats), metadata files (filled out spreadsheet) and barcode files (plain text ASCHII).

2.3.Pre-processing of raw mixed DNA sequence:

The raw sequence data generated from the samples have variable start and end points. Other undesirable characteristics of raw spectral sequences are: non-synchronization of peaks or peaks' shifts and scaling variation between samples. In order to correct these anomalies, some pre-processing procedures were carried out. First, all sample sequences were aligned to each other, followed by trimming of the sequences to a region comprising a common start and end fragments. Thereafter, normalization of sequence data to remove scale variation and correlation optimized warping (COW) to rectify non-synchronization of the peaks were carried out.

2.4.Functional abundance:

The functional abundance count is an integer (0 or positive) based count of a particular function that have been detected. Sequence similarity searches were computed against a protein database derived from theM5NR, which provides nonredundant integration of many databases: GenBank, SEED, IMG, UniProt, KEGG, and eggNOGs. MG-RAST v3 now supports many complementary views into the data with one similarity search, including different functional hierarchies: SEED Subsystems, IMG terms, COG, eggNOGs, and ontologies such as GO (Gene Ontology Consortium, 2013). The annotated source was SubSystem with the maximum e-value cutoff 1e-5. The minimum % identity cut of was fixed in 60 which means minimum 60% identity between the selected metagenome and existing sBLAT sequences are under consideration. The minimum alignment length cutoffis set to 15 which means minimum 15 aa for protein databases that predefined in MG-RAST.

2.5.Species and strain identification and short read:

For the analysis of metagenomic data into lower taxonomic level, it is important to resolve it to go down to the species level. Random reads gives a level of resolution by which someone can distinguish between closely related species and strains due to the fact that random sequences has targeted species or strain specific genes that are not usually use in phylogenetic markers. Therefore phylogenetic marker cannot provide this level of resolution. The comparison between very short reads (less than ~50 bp) and reads of length ~100 bp shows higher read length reflects a reasonable level of confidence (BLASTX bit-score of 30 and higher) whereas short read length results rigorous under prediction. In parallel to this, it is

observed that reads of length 35 bp and 100 bp are long enough to identify species (Huson et al., 2007)

2.6.Lowest common ancestor analysis:

MG-RAST provides taxonomic annotations based on Lowest Common Ancestor (LCA) method introduced by MEGAN (Huson et al., 2007). The advantage of this method is it avoids multiple annotations for a single feature. This program uses an algorithm where all hits are collected that assigns to the lowest common ancestor (LCA) of the set of taxa. Species specific sequences are annotated to taxa near the leaves of the NCBI tree. Conserved sequences are assigned to high-order taxa closer to the root (Huson et al., 2007)(Fig. 2.2).

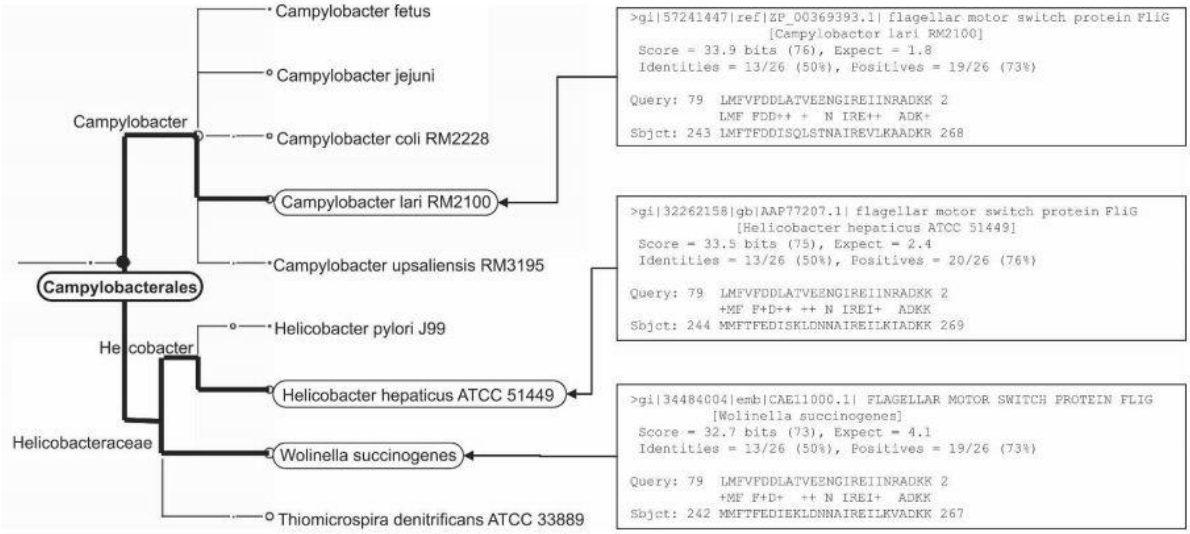


Figure 2.2:Species identification ofmetagenomicdata.From right to the left, three BLASTX matches obtained for a specific read "r" from the mammoth data set, to sequences representing *Campylobacter lari*, *Helicobacter hepaticus*, and *Wolinellasuccinogenes* respectively. The LCA-assignment algorithm assigns "r" to the taxonCampylobacterales, as it is the lowest-common taxonomical ancestor of the three matched species (Huson et al., 2007).

3. Results

3.1. Library preparation:

After DNA library preparation by Nextera® XT DNA sample preparation kit, the libraries underwent analysis by agarose gel electrophoresis to perform a visual examination where a particular type of bands will ensure the presence of DNA libraries. The band showed a smear throughout different lanes which means the DNA libraries are generated with varying lengths and the negative control that did not give the visible band on the gel, indicates absence of DNA contamination in the extraction and purification procedures (figure 3.1).

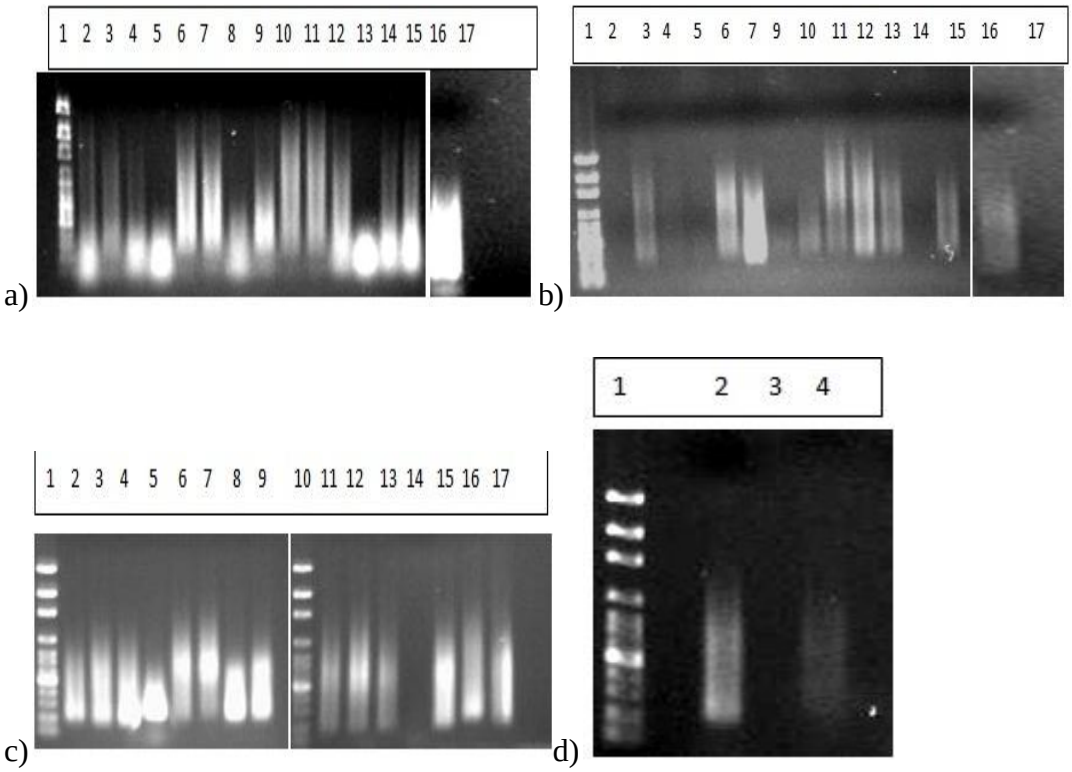


Figure 3.1: Gel run of different products during library preparation by Nextera XT DNA sample preparation kit from Illumina. All the gel analysis was performed on 1% agarose in 1X TAE buffer running for 40 minutes at 80 volts. a) gel run of NutralizedTagmentAmplicons (NTA) on 1% agarose with 1X TAE buffer. Lane 1: 100bp ladder, lane 2-16: NTA of each sample and 17: negative control. b) Gel run of Clean amplified NTA (NTA) on 1% agarose with 1X TAE buffer. Lane 1: 100bp ladder, lane 2-16: NTA of each sample and 17: negative control. c) Gel run of re-amplified Clean amplified NTA (CAN) on 1% agarose with 1X TAE buffer. Lane 1 & 10: 100bp ladder, lane 2-9 and 11-17: reamplified NTA of each sample and 18: negative control. d) Gel run of reamplified Clean amplified NTA (CAN). Lane 1: 100bp ladder, lane 2: pooled reamplified CAN product, lane 3: negative control and lane 4: discarded supernatent after step 8 during reamplified CAN processing.

3.2. Statistics of raw and processed data:

There were fifteen libraries prepared for metagenomic analysis in MG-RAST. One libraries from elite control produced large amount of short reads. Therefore, failed QC process in MG-RAST.The raw data comprised of 482,699,994 bp in total among the 15 metagenomes which comprises 4051675 sequences and after processing 398,544,691 bp were counted, which comprises 3574938 sequences (details in appendix 4). Which means 84155303 bp (17.43% of total) failed the quality control process due to some reasons, e.g., duplication (figure 3.2a,b). On average, there are 207690 protein features had been predicted from the processed data of each metagenome. Among these predicted protein features 188636 are assigned an annotation using at least one of the protein database shared by MG-RAST (figure 3.3b) (details in appendix 5).

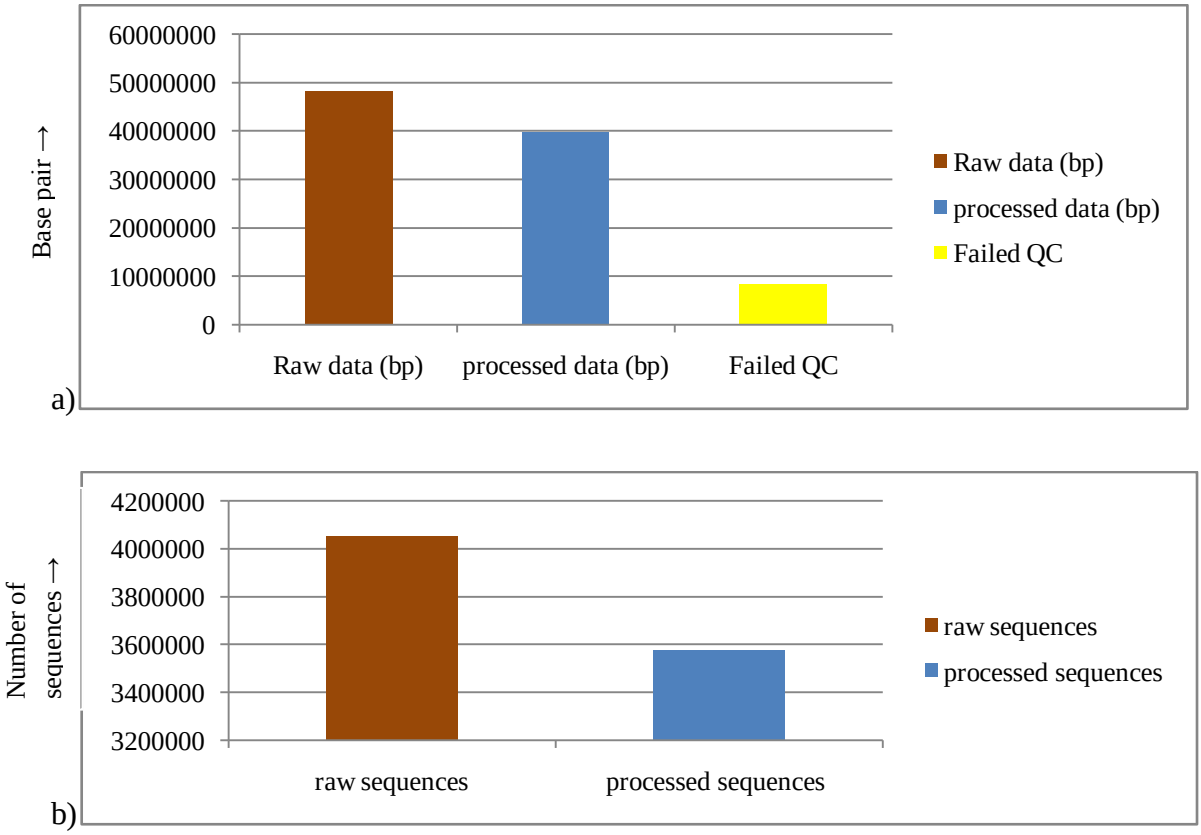
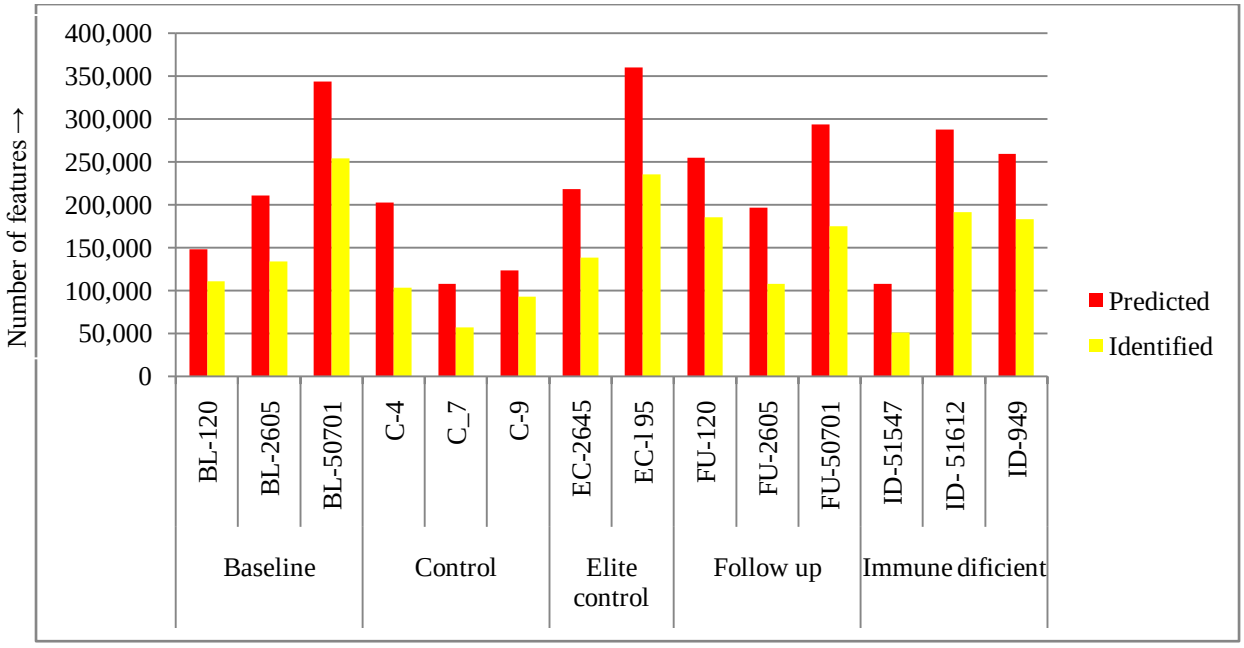
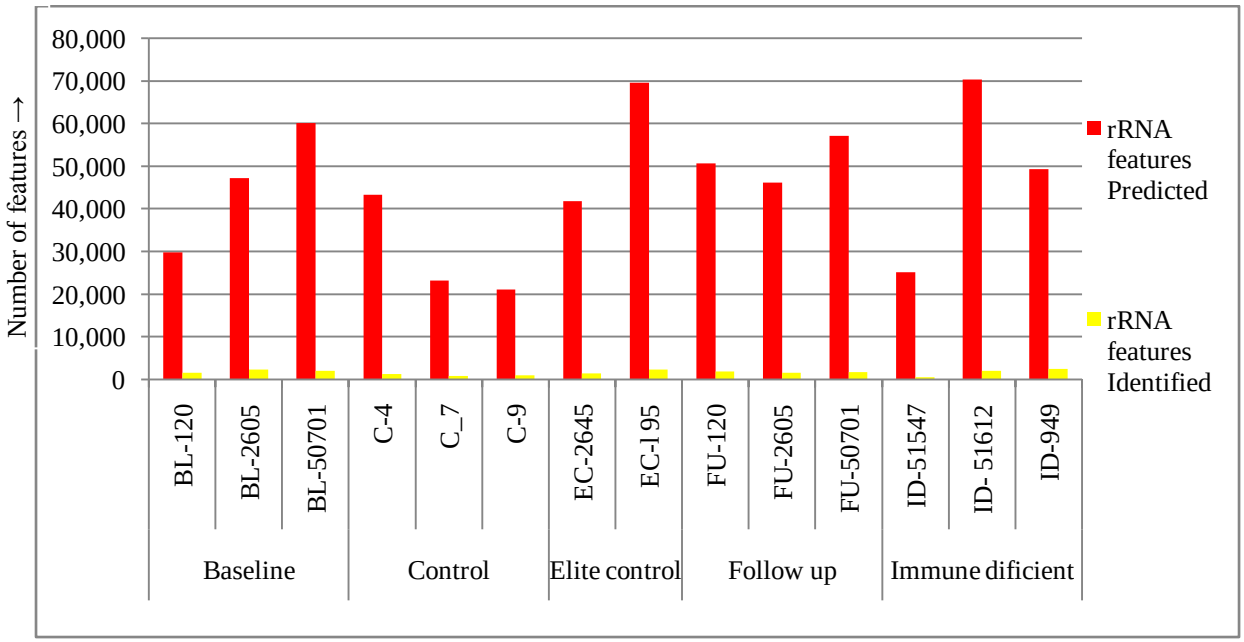


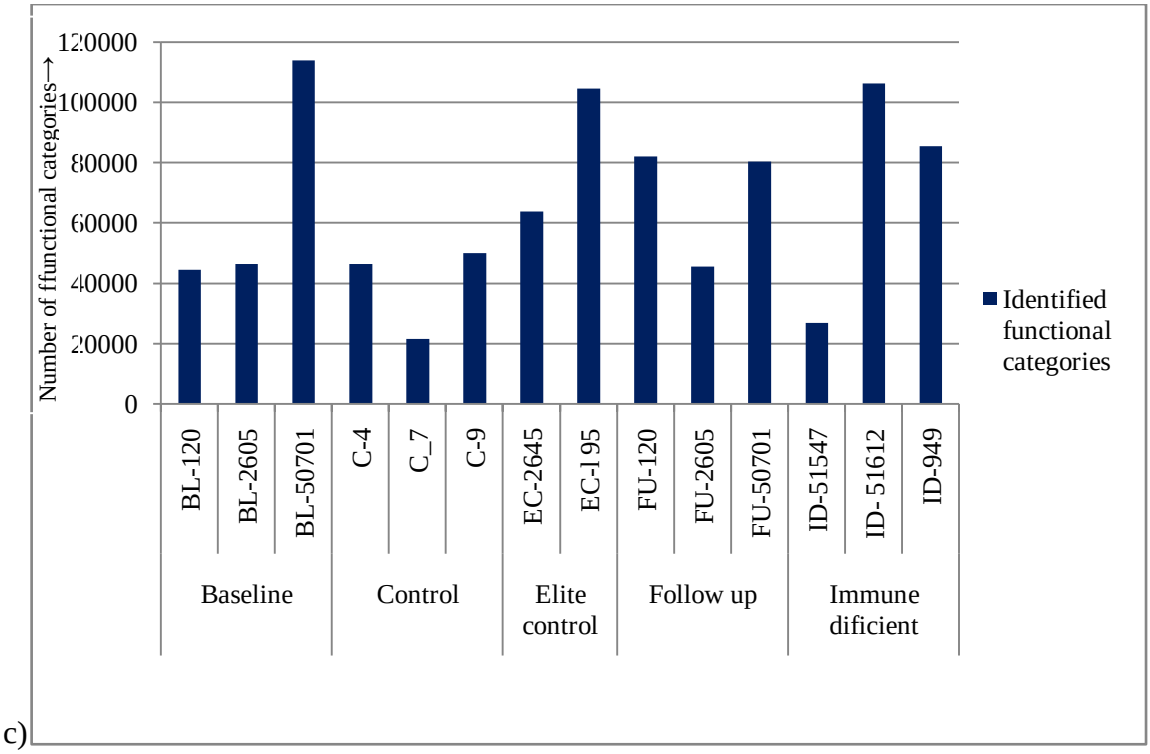
Figure 3.2: The bar chart represents a) the number of raw data (bp) and the processed data (bp) after data normalization and the amount of data that was failing in the normalization process. b)Represents the number of raw sequences by the raw data and the number of processed sequences after data normalization.



a) Predicted and identified protein features.



b) Predicted and identified rRNA features.



c)

Figure 3.3**:** The data represent predicted (red bar) and identified (yellow bar) a)protein features and b)rRNA features. c) identified functional categories. Predicted protein features were annotated with similarity to a protein of known function using M5NR database. In addition, ribosomal RNA genes were mapped to the rRNA databases.

3.3. Taxonomic abundance by lowest common ancestor method:

There were five different groups of patient stool sample from the Swedish cohort. Among the sixty Swedish stool specimen samples we had selected fifteen of them and each of them corresponding to one of these five groups. We analyzed all the data groups and compared using a maximum e-value cutoff 1e-5, minimum identity cutoff 95% and a minimum alignment length cutoff 35 measured in aa for protein and bp for RNA database. The data had been normalized to values between 0 and 1. Five groups of samples, e.g., Immune deficient, follow-up, elite control, control and base line were analyzed and compared with each other. The bar charts (figure 3.4 to 3.15) used to visualize the approximate membership percentage within each taxonomic level included in each metagenomic sample. In addition to this we performed significance tests to identify taxonomic levels that are significantly different between these groups. The p-values were calculated using ANOVA-one-way. Data analysis showed that any microbial domain is not significantly different between the groups, but Archaea is closer to the significant level where its p value is slightly higher than $P< 0.05$. It was observed that bacterial domain is not significant (figure 3.4) among the groups, but in its

lower taxonomic level, such as phylum, class, order, family, genus, species significance found in every level.

3.3.1. Distribution of bacteria:

The bacteria was the most abundant domain in all the groups, two subjects each from immune deficient and follow up group had higher abundance of bacterial abundance compare to other groups. The normalized abundance counts (NAC) of immune-deficient subjects were 0.973, 0.937 and 0.796 (ID-945, ID-51612 and ID-51547 respectively; the NAC of the follow up subjects (FU-50701, FU-2605 and FU-120) were 0.966, 861 and 0.958. Their corresponding state in the baseline group (BL-50701, BL-2605 and BL-120) showed literally similar NAC e.g., 1.00, 0.87 and 0.912. The NACof C-9 (control)was 0.774 which was higher than C-7 (control). Overall, these two subjects from the control group showed lowerNACcompared to all other groups (figure 3.4).

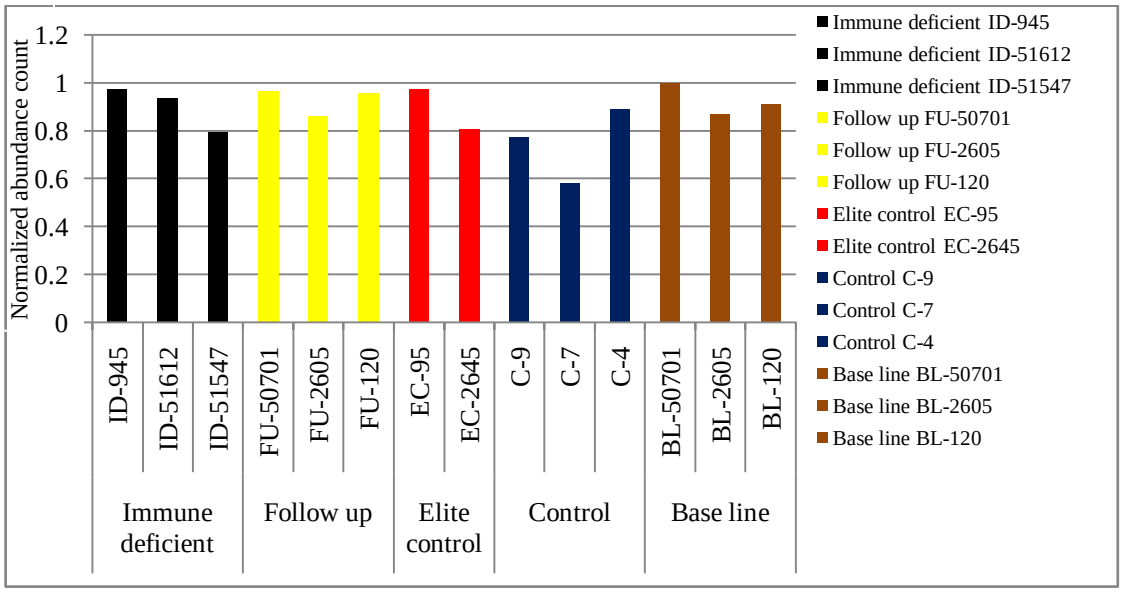


Figure 3.4: The bar chart represents the normalized abundance count of bacteria in fourteen samples. X-axis represents different subjects under five groups mentioned at the bottom. Y-axis represents the normalized abundance count score (0 to 1) of each subjects. Different subjects under same group are in same colour.

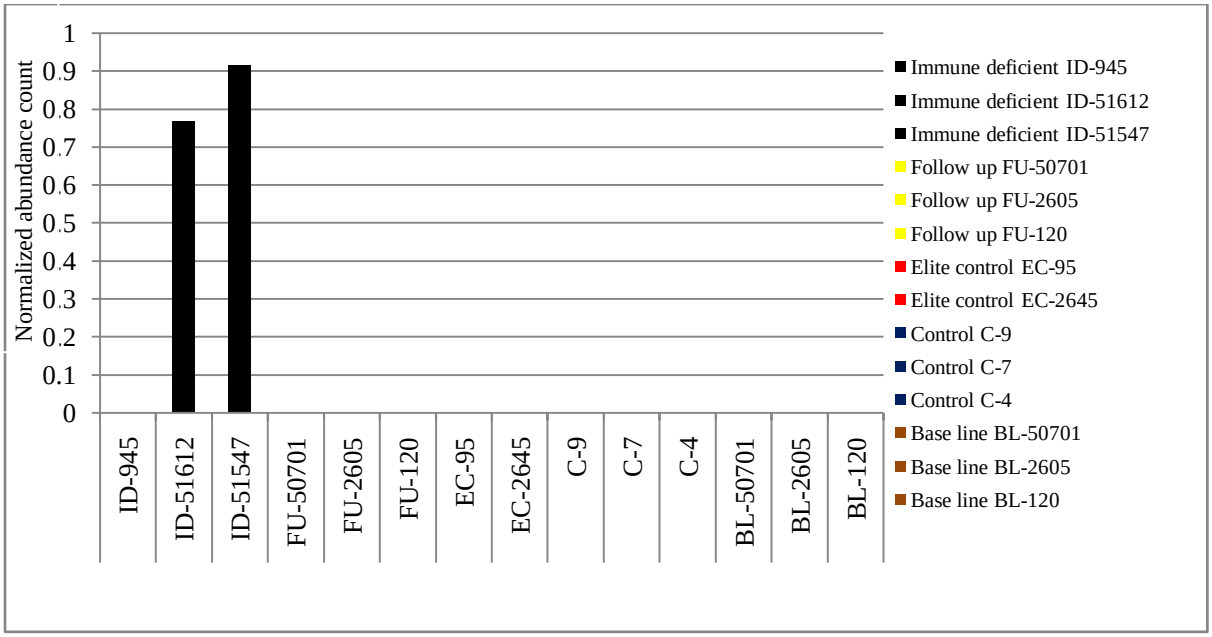


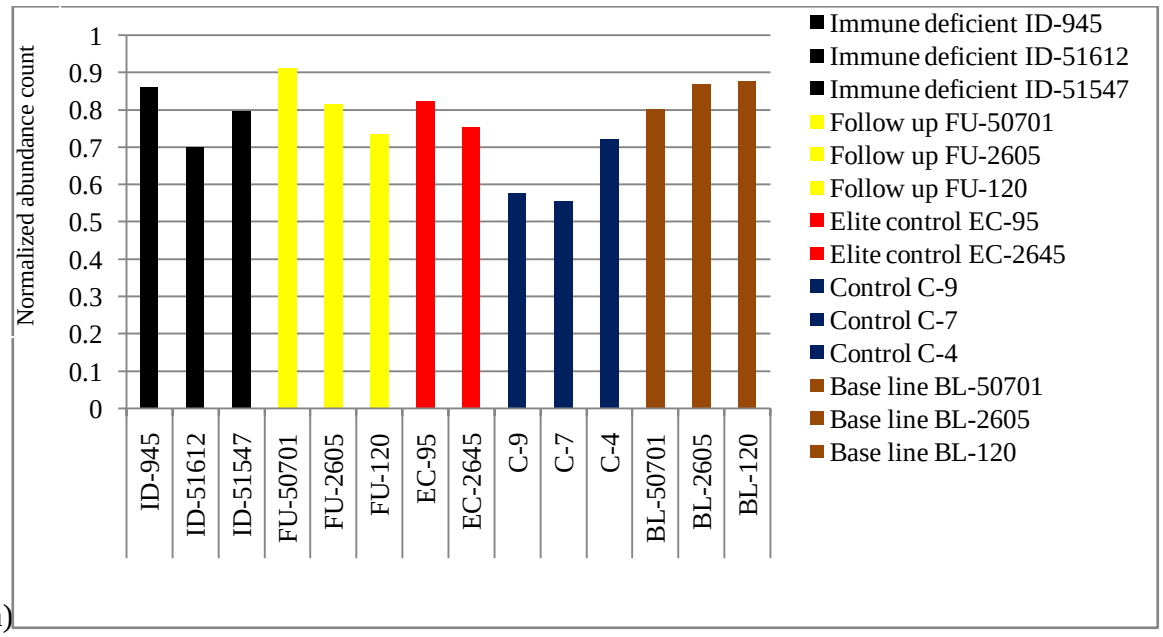
Figure 3.5: The bar chart presentation the normalized abundance count of Archeae ($P<0.05$) in fourteen sampleswhich was abundant only in the Immune deficient group. All the subjects have a code number with their group code at the X-axis. Y-axis represents the normalized abundance count score (0 to 1) of each subjects. Different subjects under same group are in same colour.

3.3.2. Phyla distribution:

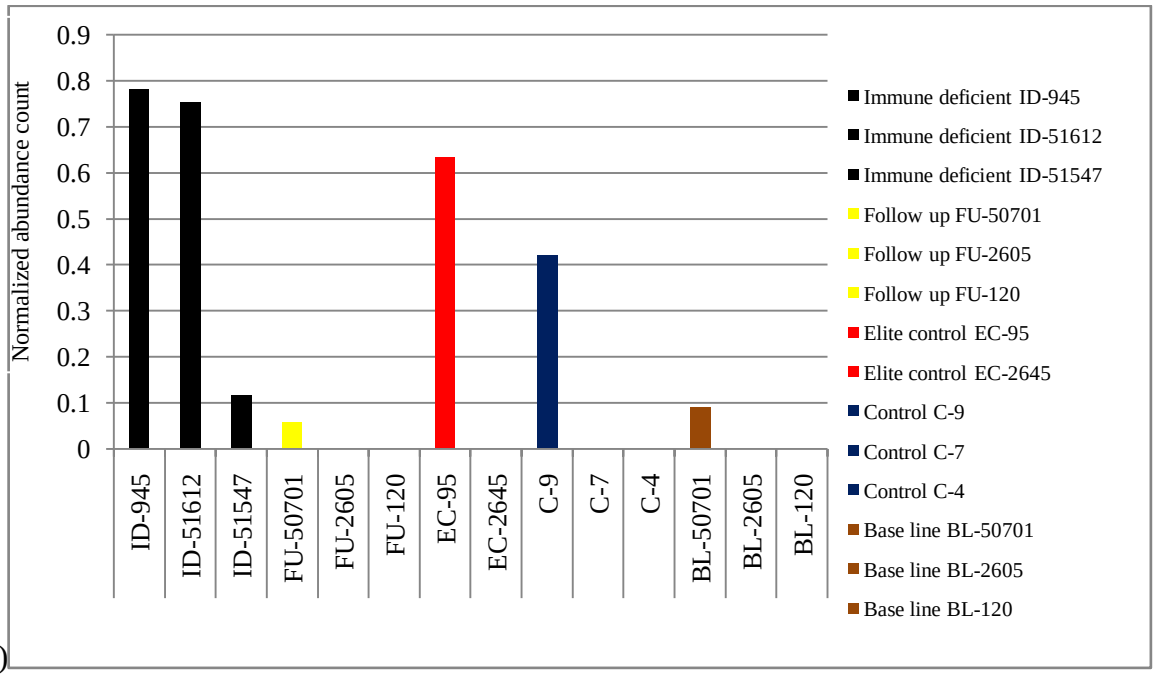
Though bacterial distribution as a domain didn't show any significance between thegroups, butit showsthe significanceat lower taxonomic levels such as bacterial phylum Firmicutes.It was distributed among all the groups. The NACof the C-9, C-7 and C-4allof the control group were 0.578, 0.558 and 0.722 respectively. A greaterNAC wasobserved in the subjects in all other groups. The value of NAC of EC-95 and EC-2645were 0.824 and 0.755. Whereas NAC of ID-945, ID-51512 and ID-51547 0.86, 0.70 and 0.787 respectively. Two subjects from the baseline group (BL-120 and BL-2605) showed a higher NAC compared to the follow up (FU-120 and FU-2605) group where same subjects participated. The NAC were 0.877 and 0.869 in the baseline, 0.733 and 0.816 in the follow up respectively(figure 3.6a). Bacterial phylum Verrucomicrobia showed a different pattern of abundance counts. The NAC of ID-945 and ID-51612 (immune deficient group) were 0.781 and 0.754. Compared to this, the NACof EC-95 and C-9 were 0.634 and 0.421. The NAC of a subject which perform both in follow up (FU-50701) and baseline (BL-50701) was 0.058 and 0.091 respectively. The NAC of the immune deficient subjects was comparatively much higherthen the other groups(figure 3.6b). Actinobacteria was less abundant in the control group. The NACof C-9 and C-7 (control group) were 0.273 and 0.315 (figure 3.6c). The NAC of Proteobacteria in

FU-50701, FU-2605, FU-120, EC-95, EC-2645, C-9, C-7, C-4, BL-50701, BL-2605 and BL-120 were 0.738, 0.513, 0.45, 0.475, 0.454, 0.588, 0.382, 468, 0.516, 0.454 and 0.487 respectively(figure 3.6d). FU-2605, FU-120, EC-95, EC-2645, BL-2605 and BL-120 show comparatively similar NAC. The NAC of Bacteroidetes in two subjects of the follow up group (FU-2605 and FU-120) were 0.643 and 0.96, whereas it was 0.519 and 0.839 respectively in BL-2605 and BL-120. Therefore, the subjects in the follow up group had higher NAC.

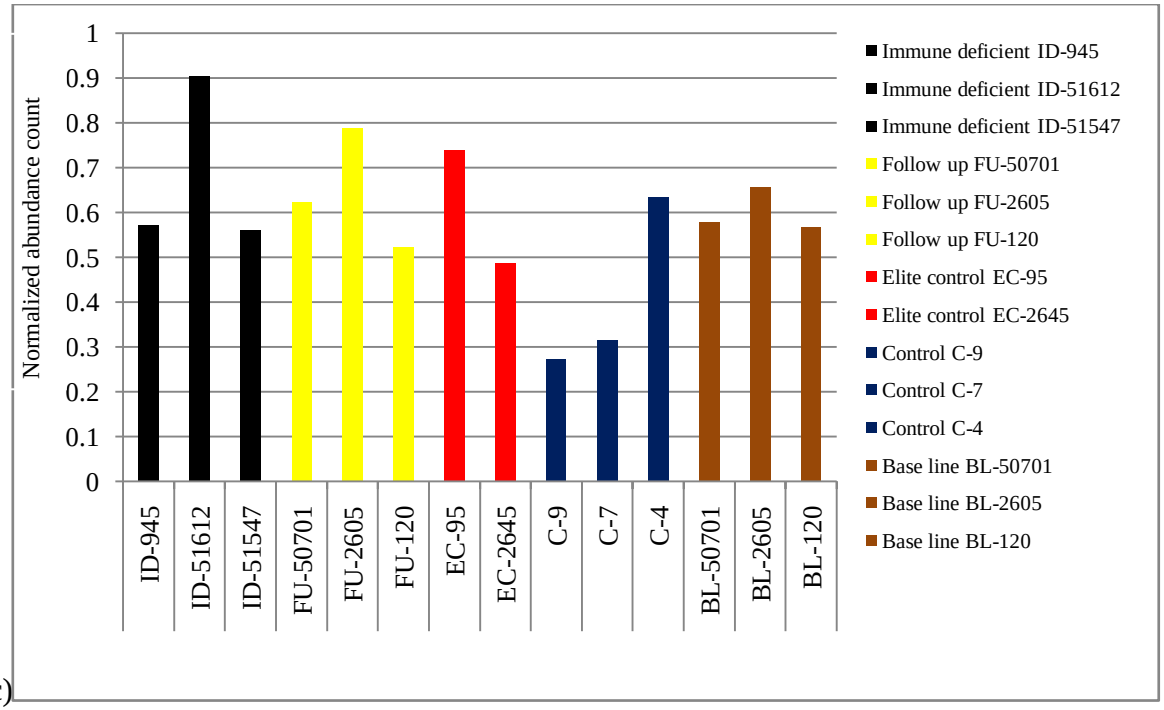
a)



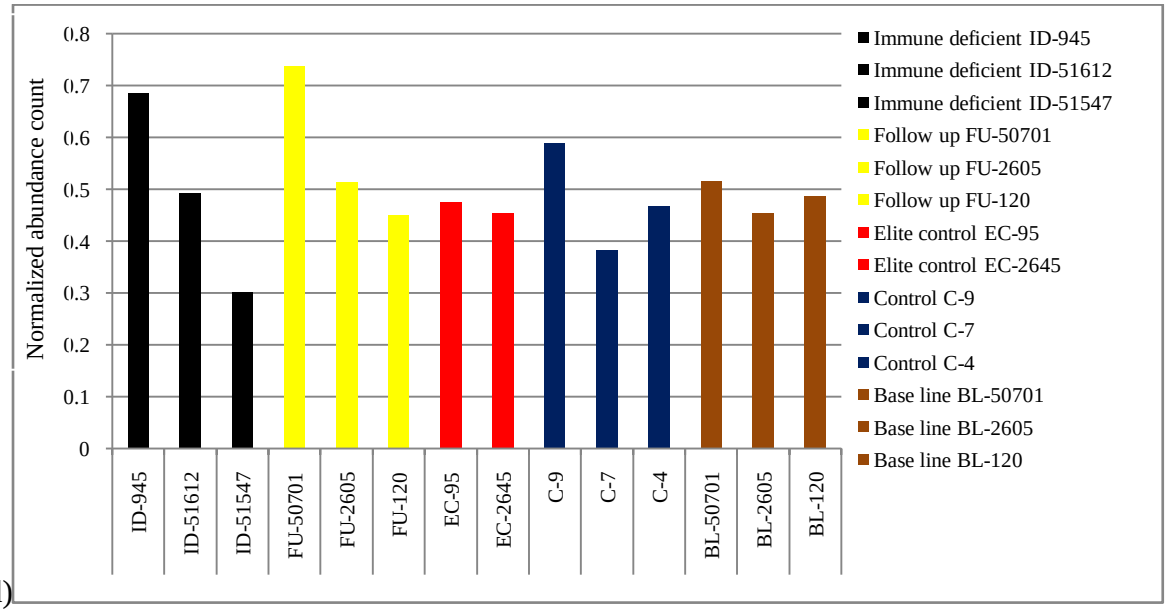
b)



c)


Normalized abundance count
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
ID-945
ID-51612
ID-51547
FU-50701
FU-2605
FU-120
EC-95
EC-2645
C-9
C-7
C-4
BL-50701
BL-2605
BL-120
Immune deficient ID-945
Immune deficient ID-51612
Immune deficient ID-51547
Follow up FU-50701
Follow up FU-2605
Follow up FU-120
Elite control EC-95
Elite control EC-2645
Control C-9
Control C-7
Control C-4
Base line BL-50701
Base line BL-2605
Base line BL-120

d)


Normalized abundance count
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
ID-945
ID-51612
ID-51547
FU-50701
FU-2605
FU-120
EC-95
EC-2645
C-9
C-7
C-4
BL-50701
BL-2605
BL-120
Immune deficient ID-945
Immune deficient ID-51612
Immune deficient ID-51547
Follow up FU-50701
Follow up FU-2605
Follow up FU-120
Elite control EC-95
Elite control EC-2645
Control C-9
Control C-7
Control C-4
Base line BL-50701
Base line BL-2605
Base line BL-120

e)

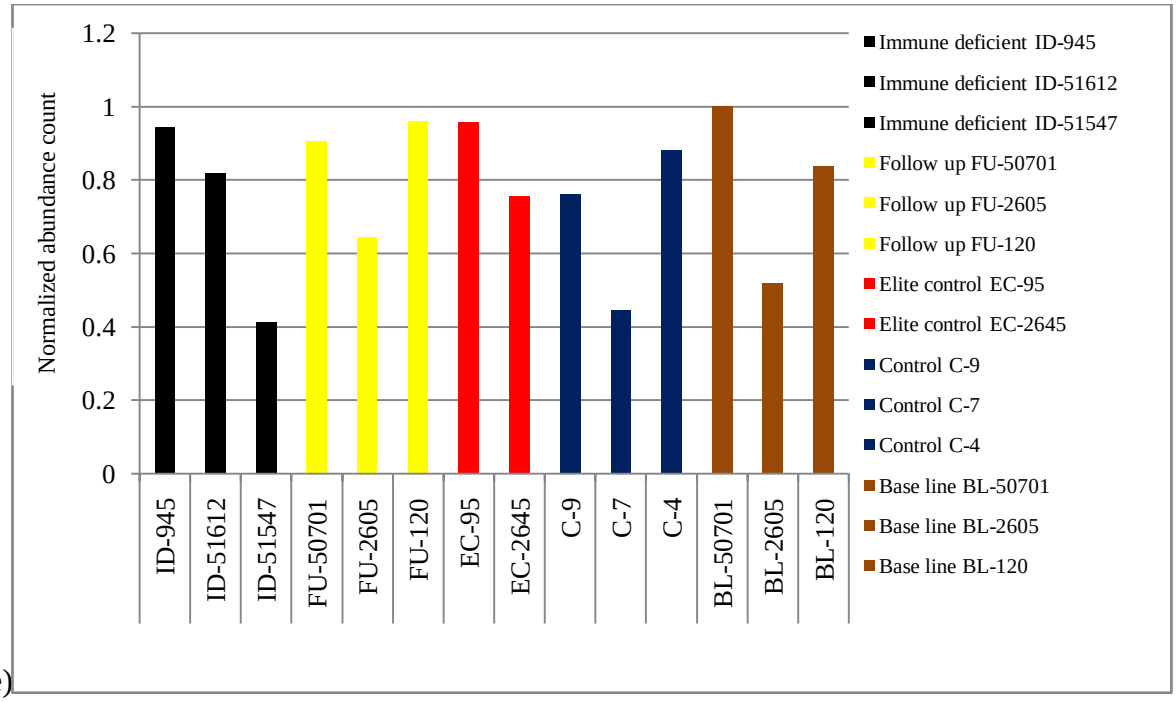


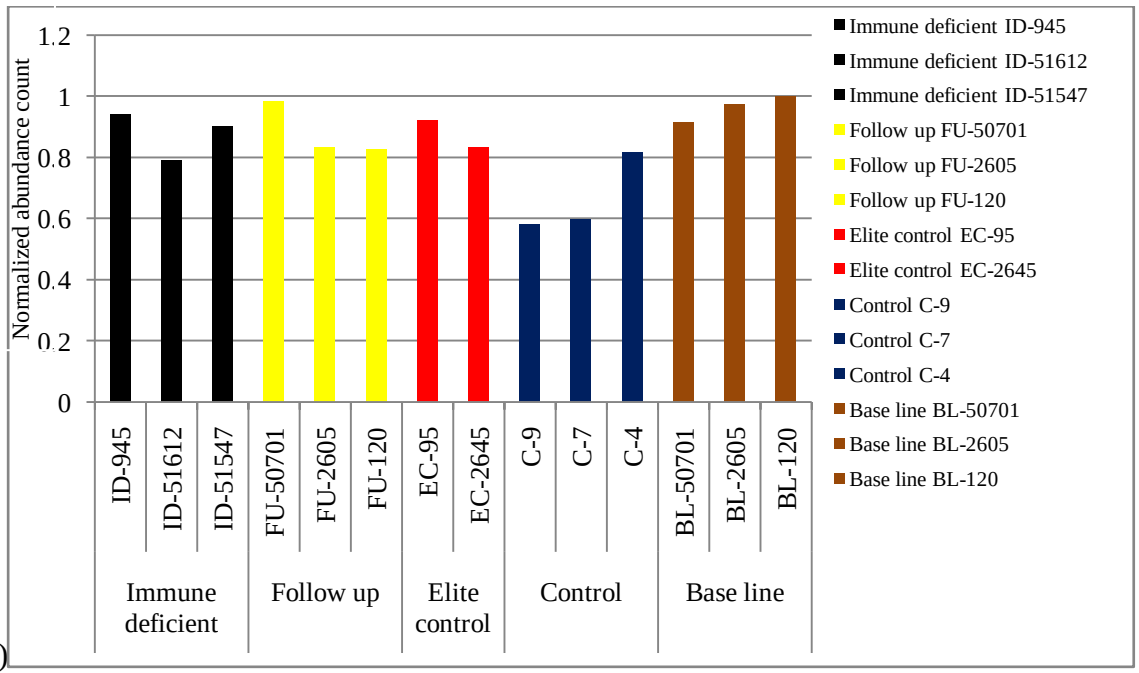
Figure 3.7: Normalized abundance count of the bacterial phyla. a) Firmicutes, b) Verrucomicrobia,c) Actinobacteria, d) Proteobacteria and e) Bacteroidetes. All the subjects have a code number with their group code on the X-axis.and Y-axis: Normalized abundance count (0 to 1) of each subject. Different subjects under same group are in same colour.

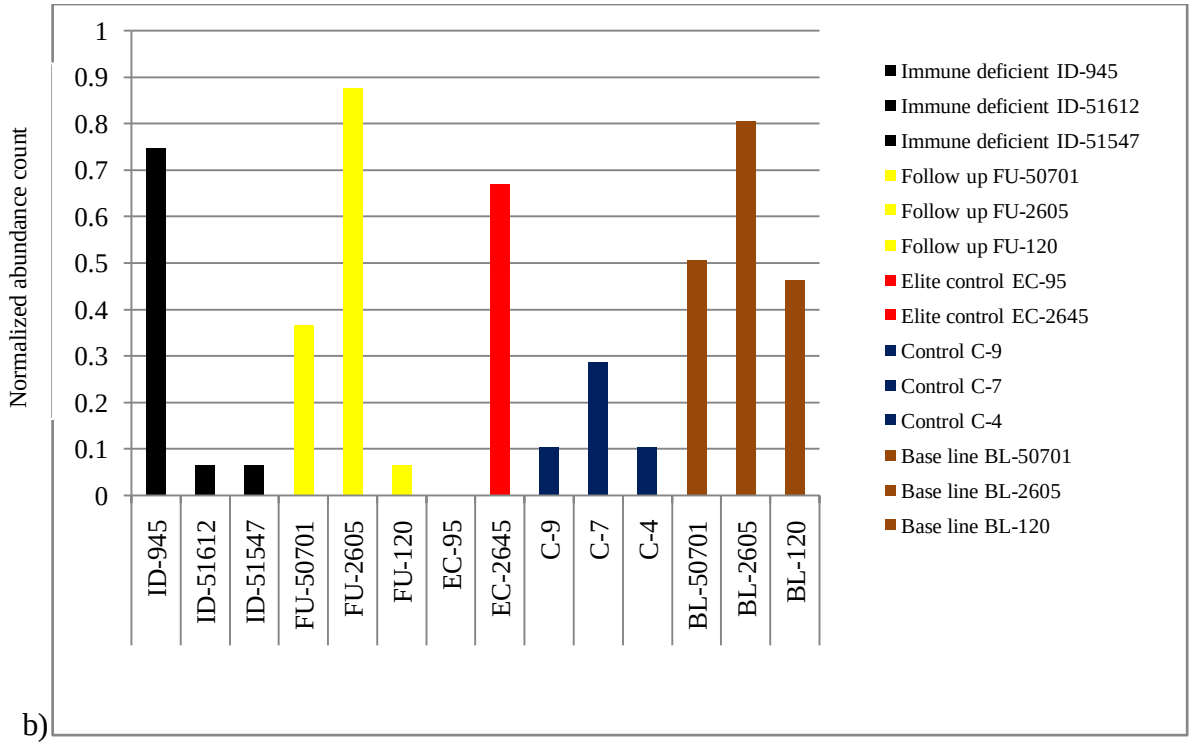
3.3.3. Class distribution

There were four bacterial classes under the Firmicutes phylum distributed in different subjects. NAC of Clostridia in different subject groups were significantly different ($P<0.05$). It was less abundant in the control group. The NAC of C-9 and C-7 were 0.582 and 0.599 which was lower compared to the values of other subjects. The NAC of BL-120 and BL-2605 were 1.00 and 0.974. In contrast to this the NAC of FU-120 and FU-2605 were lowered to 0.827 and 0.836. All the subjects whose were HIV infected had higher NAC then the C-9 and C-7(figure: 3.8a). The NAC of Erysipelotrichiwasfluctuated among the subjects, even within a group. The NAC of BL-120 and BL-50701 were 0.463 and 0.506 respectively, whereas it lowers to 0.066 and 0.366 in their corresponding follow up subjects (figure 3.8b). Furthermore, it was also abundant in the baseline group at lower taxonomic level (figure: 3.15). The NAC of the bacterial class Betaproteobacteriawas significantly different among the groups. The NAC of BL-120, BL-2605 and BL-50701 were 0.221, 0.612 and 0.699, and the NAC of FU-120, FU-2605 and FU-50701 were 0.125, 0.378 and 0.461 respectively. The NAC of the subjects in the baseline grouphad higher NAC compared to its corresponding subjects in the follow up group. Both subjects in the elite control group (EC-95 and EC-2645)

had a same NAC value which was 0.613. The NAC of ID-945, ID-51612 and ID-51547 were 0.183, 0.157 and 0.079(figure 3.9a) which was lower compared to other groups. The NAC of ofGammaproteobacteria was lower subjects of baselines, elite control and immune deficient groups. The NAC of FU-2605 and FU-50701 were 0.521 and 1.00 respectively. In contrast to this, the NAC of BL-2605 and BL-50701 were 0.221 and 0.472. The NAC of both ID-945 and ID-50701 were 0.125. The control group showed higher NCA to the follow up groups. The NCA of two control subjects (C-9 and C-4) were 0.739 and 0.472 (figure 3.9b).

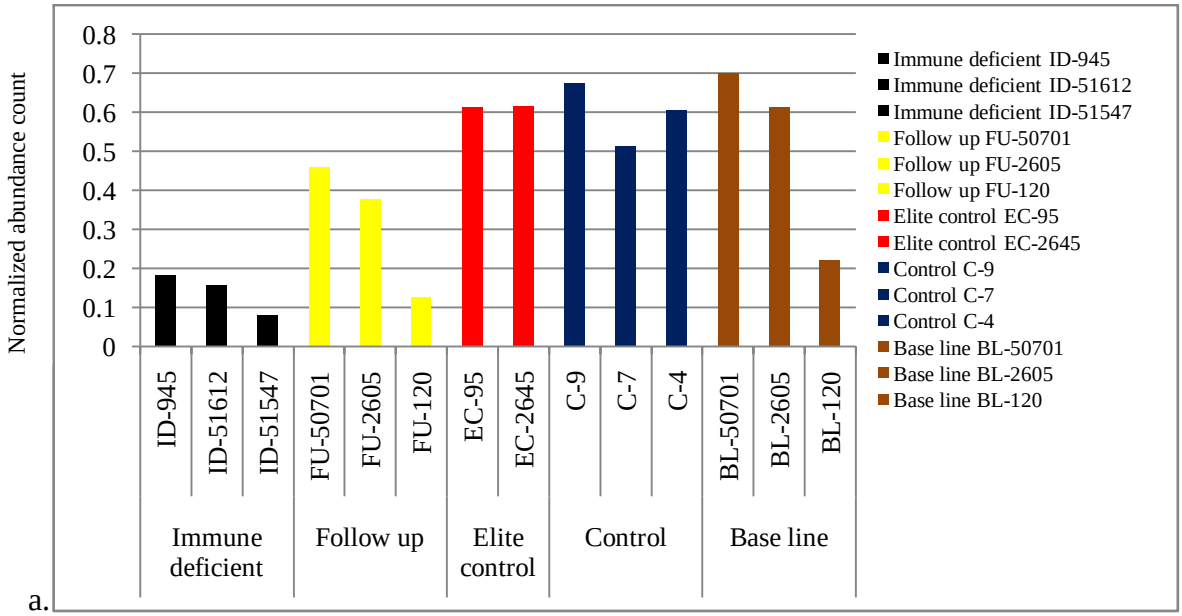
a)





b)

Figure 3.8: Normalized abundance count of the bacterial class under the Firmicutes phylum. a) Clostridia and b) Erysipelotrichi.All the subjects have a code number (e.g., 50701) with their group code (e.g., FU for Follow up) on the X-axis.and Y-axis: Normalized abundance count (0 to 1) of each subject.



a.

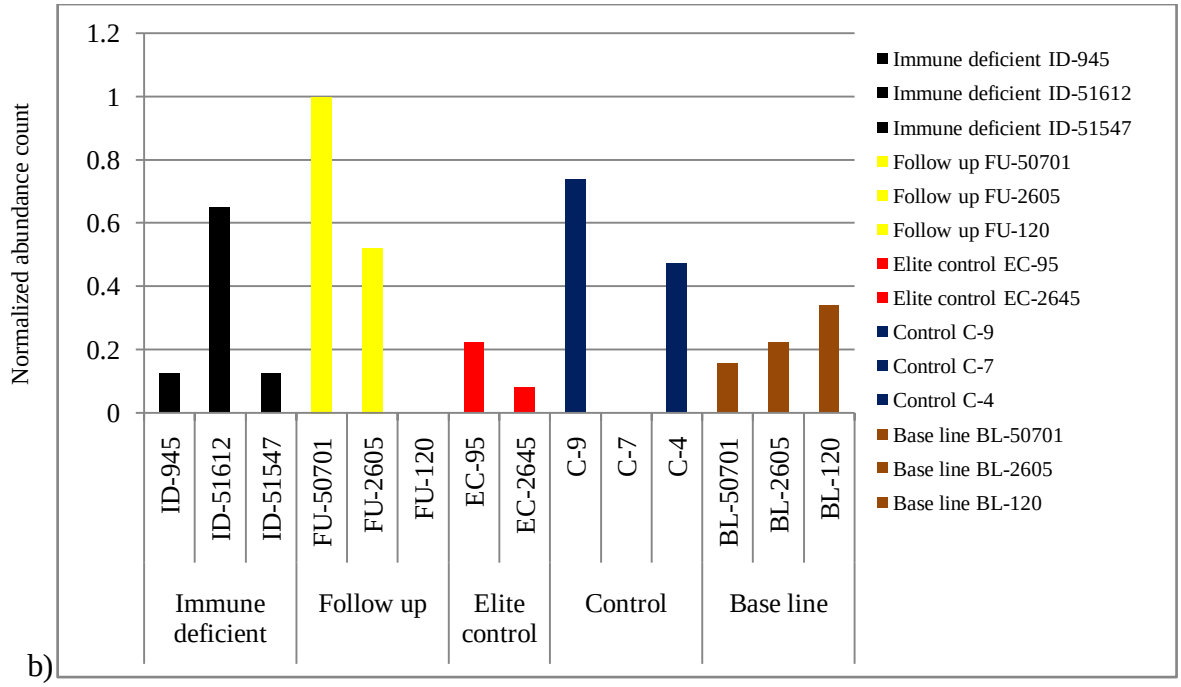


Figure 3.9: Normalized abundance count of the bacterial classes under the phylum Proteobacteria. a) Betaproteobacteria and d) Gammaproteobacteria. All the subjects have a code number (e.g., 945) with their group code (e.g., ID for Immune-deficient) on the X-axis and Y-axis: Normalized abundance count (0 to 1) of each subject. Different subjects under same group are in same colour.

3.3.4. Order distribution:

The normalized abundance count (NCA) of Clostridialeswas significantly different ($P<0.05$) among the groups. It was observed that the NAC of BL-120 and BL-2605were 1.00 and 0.974 where as its corresponding members in follow up group FU-120 and FU-2606 had 0.827 and 0.836 respectively which were lower thanthe baseline group. The NAC of C-9 and C-7 were 0.582 and 0.60 respectively, which is much lower compared to all other subjects (figure 3.10). The NAC of Erysipelotrichales in BL-50701 was 0.544 whereas in FU-50701 it was lowered to 0.371. FU-120 did not depict any NAC. In contrast to this NAC of BL-120 was 0.491. The NAC of C-9 and C-4 were 0.048. Two subjects (ID-51612 and ID-51147) from the immune deficient group did not show any abundance counts (figure 3.11).

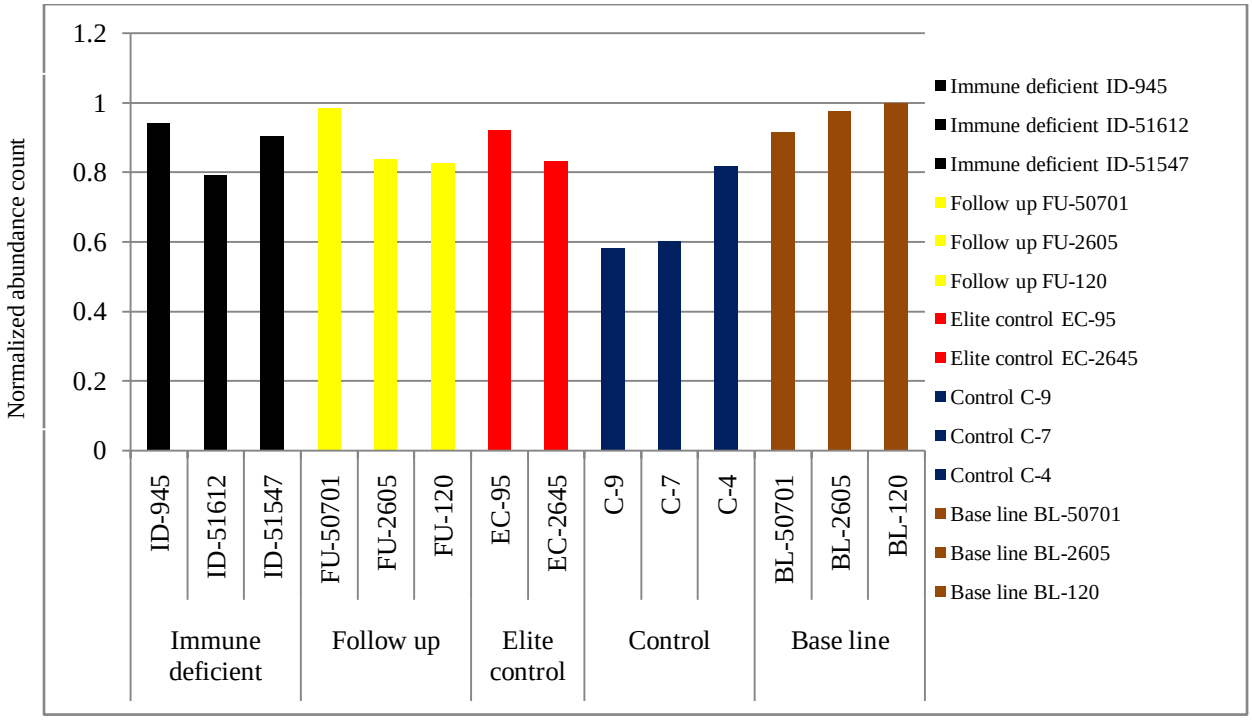


Figure 3.10: Distribution of Clostridiales under the phylum Firmicutes.All the subjects have a code number (e.g., 95) with their group code (e.g., EC for elite control) on the X-axis and Y-axis: Normalized abundance count (0 to 1) of each subject. Different subjects under same group are in same colour.

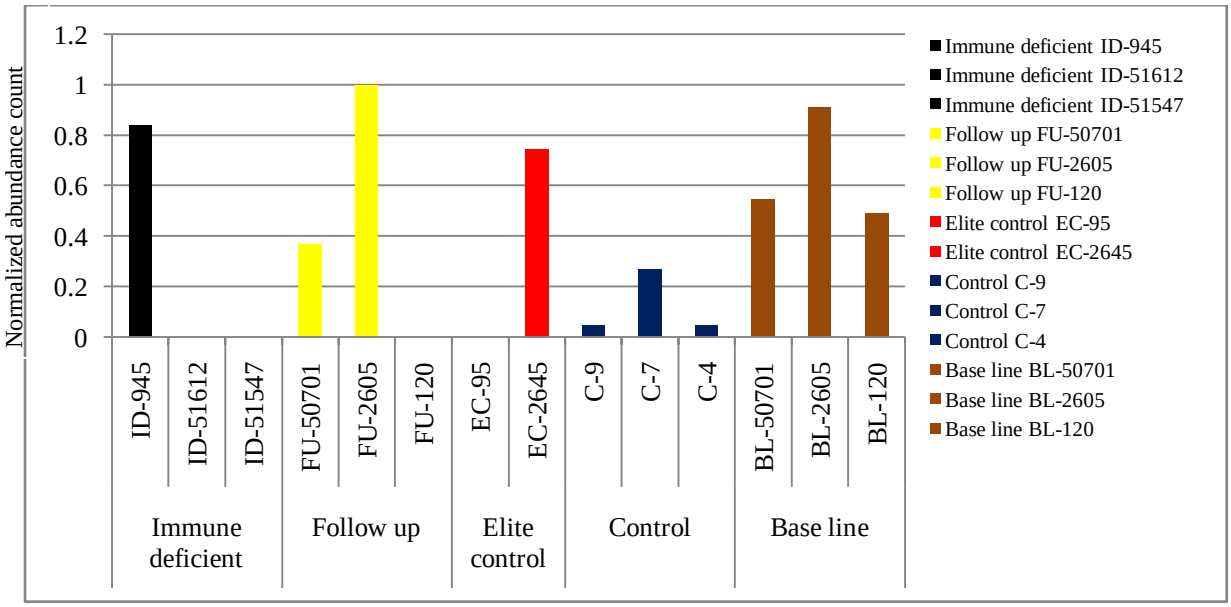
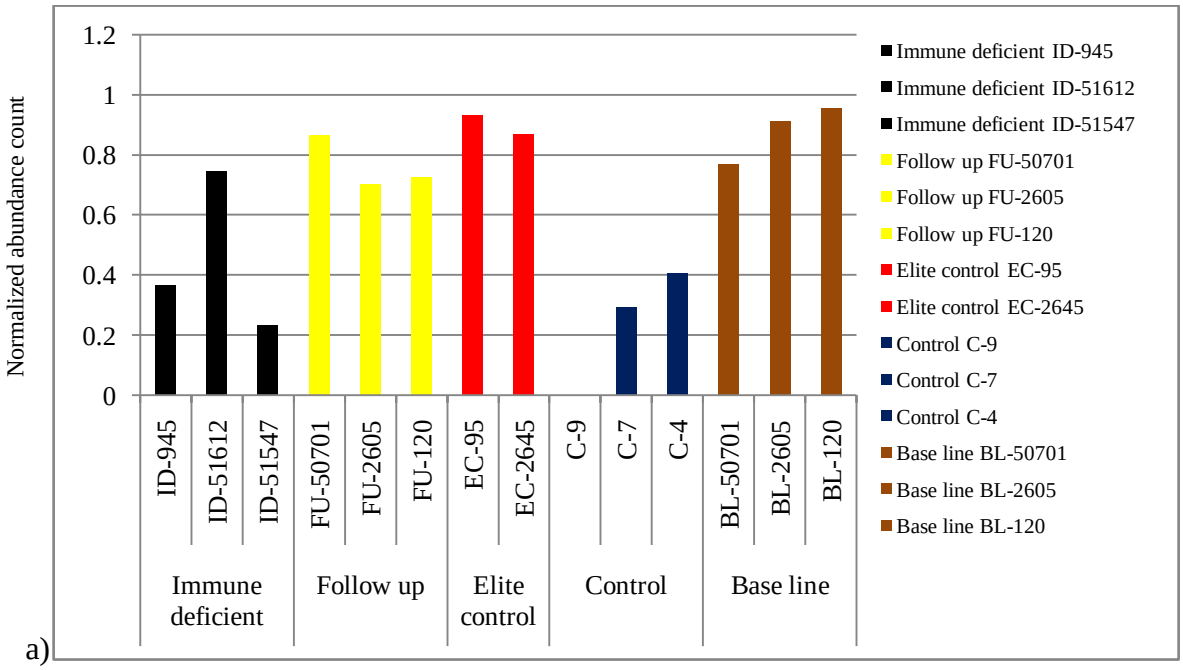


Figure 3.11: Normalized abundance count of Erysipelotrichales in different subjects.All the subjects have a code number (e.g., 9) with their group code (e.g., C for Control) on the X-axis and Y-axis: Normalized abundance count (0 to 1) of each subject. Different subjects under same group are in same colour.

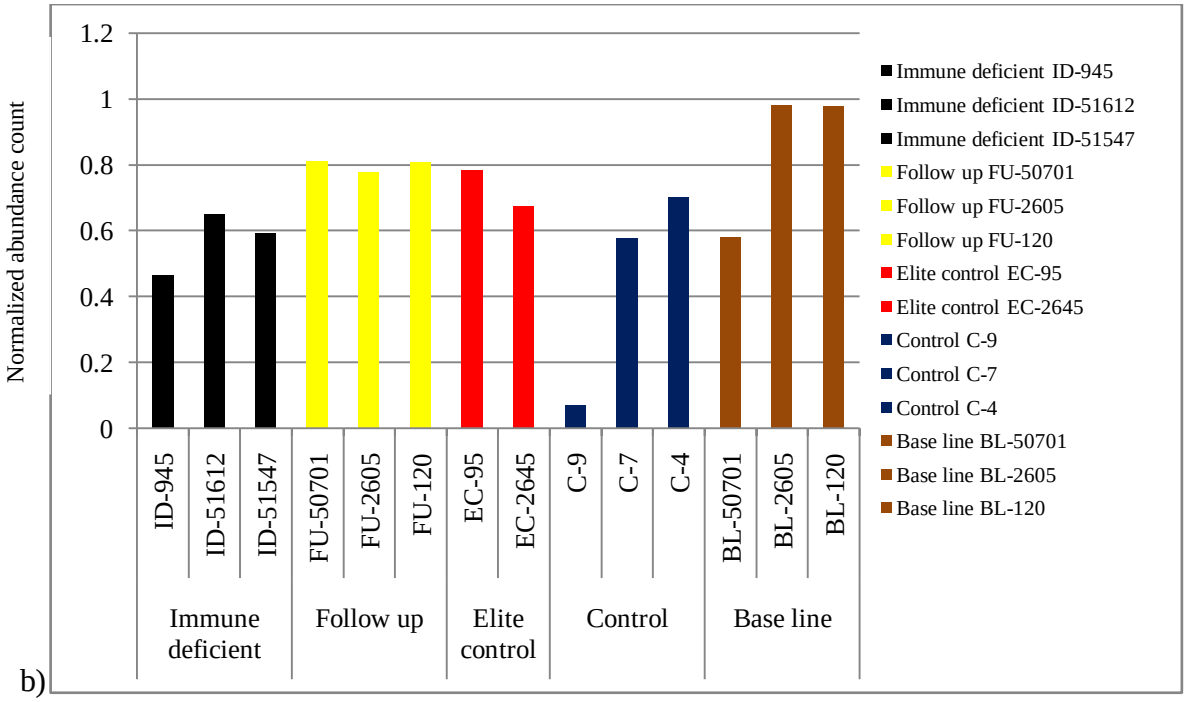
3.3.5. Family distribution:

Four bacterial Family *Eubacteriaceae*, *Lachnospiraceae*, *Ruminococcaceae* and *Clostridiaceae* under the order Clostridiales order are abundant. Among them only *Eubacteriaceae*was significantly different betweenthe groups. The NAC of ID-945 and ID-51547 were 0.366 and 0.235. In contrast to this NAC of C-7 and C-4 were 0.296 and 0.408 which shows literally similar NAC with the subjects of the immune deficient group mentioned earlier. C-9 (control group)didn't show any indication of NAC. The NAC of FU-120 and FU-2605 were 0.727 and 0.706. The NAC of BL-120 and BL-2605 were 0.956 and 0.912 which were higher compared to their follow up group. EC-95 and EC-2645 both from the elite control group had NAC of 0.934 and 0.869 which were close to the subjects in the baseline group (figure 3.12a). Normalized abundance count of*Lachnospiraceae*within the follow up group was similar. The NAC of FU-2605 andFU-120 were 0.779 and 0.809 whereas NAC of BL-2605 and BL-120 were 0.983 and 0.978 (figure 3.12b). The NAC of *Ruminococcaceae*in C-9 and C-7 were 0.447 and 0.501. Which was comparatively higherin the other groups (figure 3.12c).The NAC of *Clostridiaceae*in ID-945, FU-50701 and BL-50701 were 0.922, 1.00 and 0.958 respectively, which is comparatively higher within each group. In contrast to this FU-2605 and FU-120 had a lower NAC value which was 0.33 and 0.235 respectively. In addition to this NAC of BL-2605 and BL-120 were 0.277 and 0.454. NAC of EC-95 and EC-2645 were 0.213 and 0.183 (figure 3.12d).

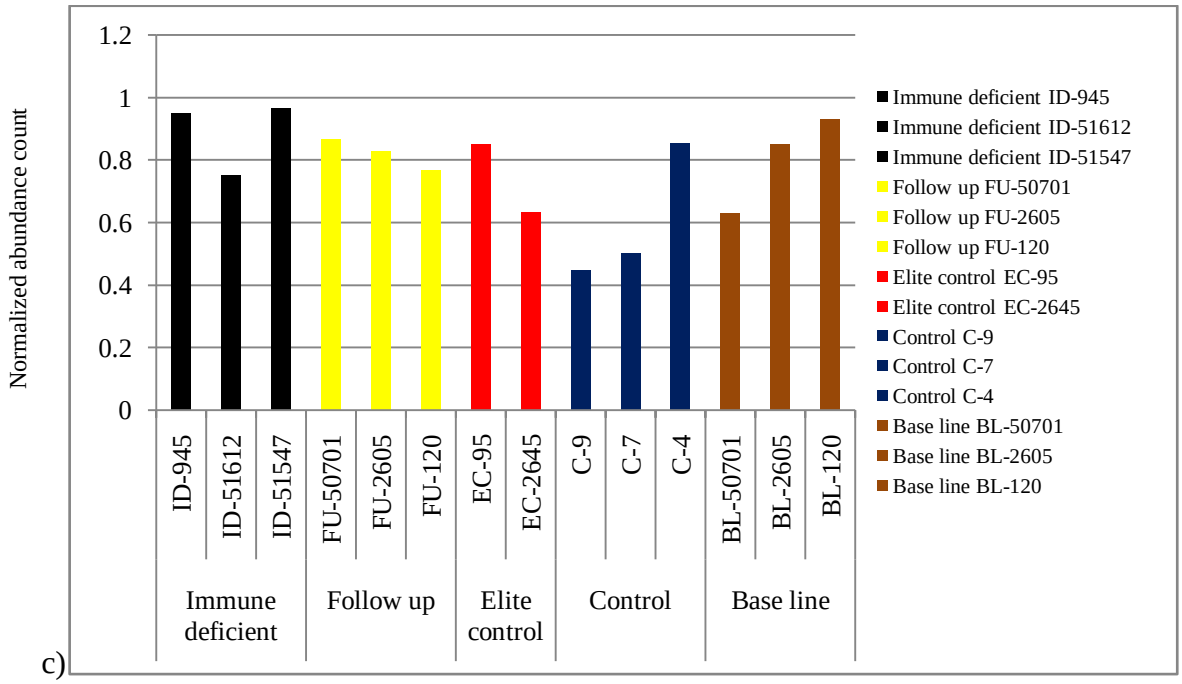
a)

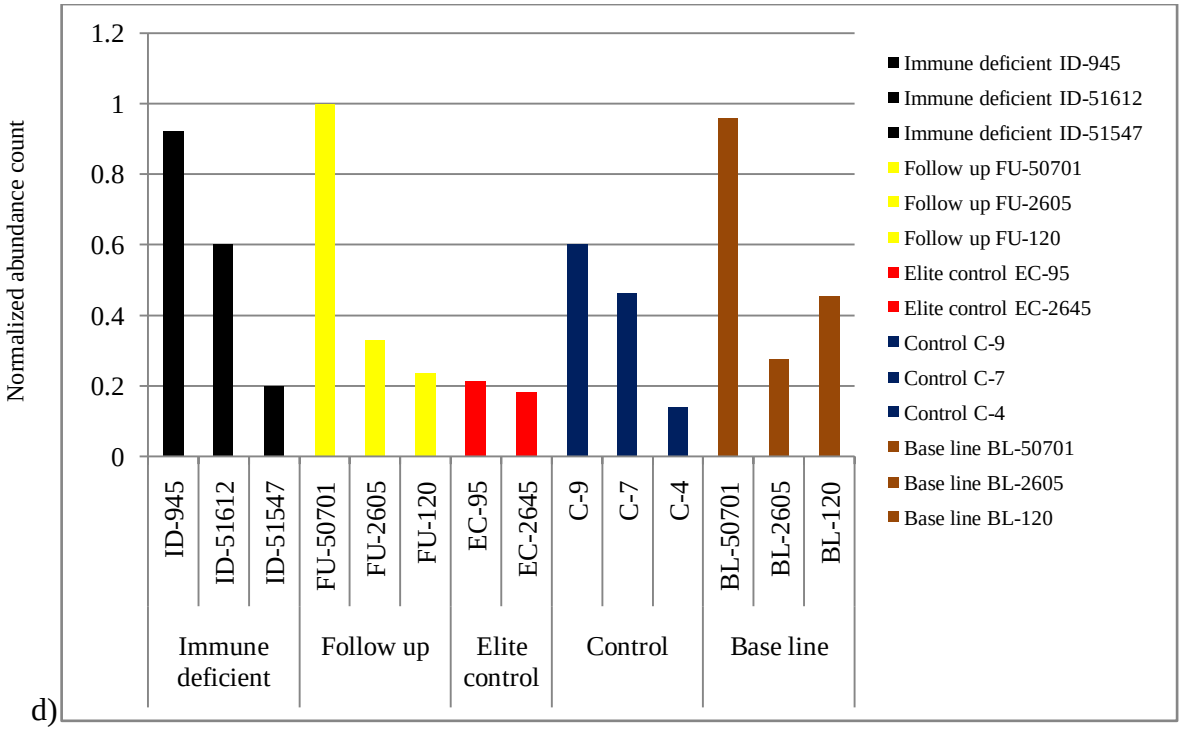


b)


Normalized abundance count
1.2
1
0.8
0.6
0.4
0.2
0
ID-945
ID-51612
ID-51547
FU-50701
FU-2605
FU-120
EC-95
EC-2645
C-9
C-7
C-4
BL-50701
BL-2605
BL-120
Immune deficient
Follow up
Elite control
Control
Base line
Immune deficient ID-945
Immune deficient ID-51612
Immune deficient ID-51547
Follow up FU-50701
Follow up FU-2605
Follow up FU-120
Elite control EC-95
Elite control EC-2645
Control C-9
Control C-7
Control C-4
Base line BL-50701
Base line BL-2605
Base line BL-120

c)


Normalized abundance count
1.2
1
0.8
0.6
0.4
0.2
0
ID-945
ID-51612
ID-51547
FU-50701
FU-2605
FU-120
EC-95
EC-2645
C-9
C-7
C-4
BL-50701
BL-2605
BL-120
Immune deficient
Follow up
Elite control
Control
Base line
Immune deficient ID-945
Immune deficient ID-51612
Immune deficient ID-51547
Follow up FU-50701
Follow up FU-2605
Follow up FU-120
Elite control EC-95
Elite control EC-2645
Control C-9
Control C-7
Control C-4
Base line BL-50701
Base line BL-2605
Base line BL-120



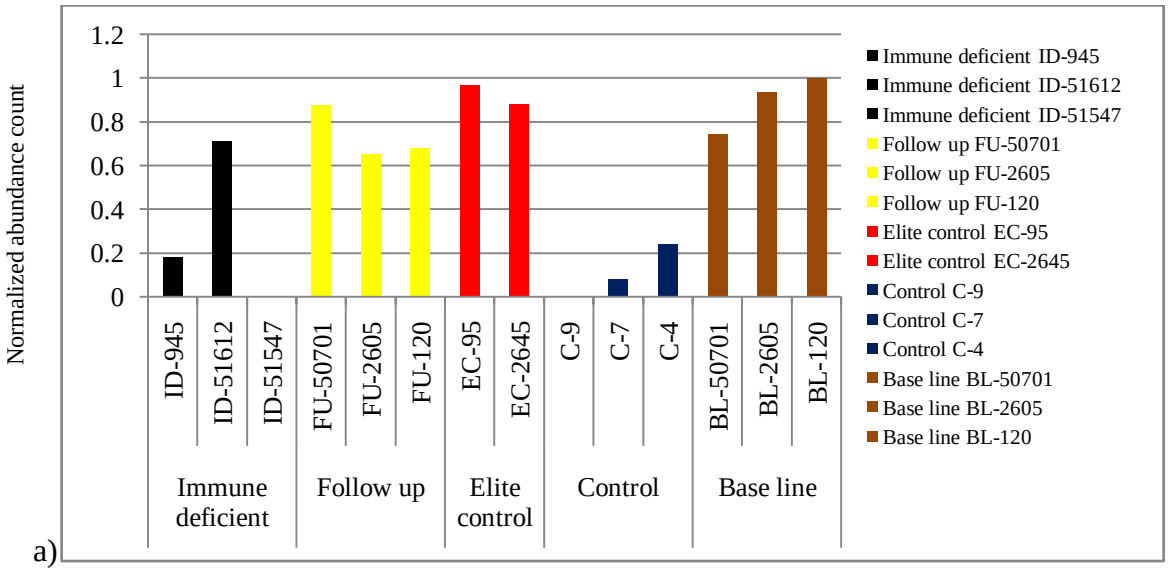
d)

Figure 3.12: Normalized abundance count of different families amongthe subjects. a) *Eubacteriaceae*, b) *Lachnospiraceae*, c) *Ruminococcaceae*and d) *Clostridiaceae*.Different subjects under same group are in same colour. X-axis represents fourteen code number (e.g., 9) with their group code (e.g., C for Control) and Y-axis: Normalized abundance count (0 to 1) of each subject.

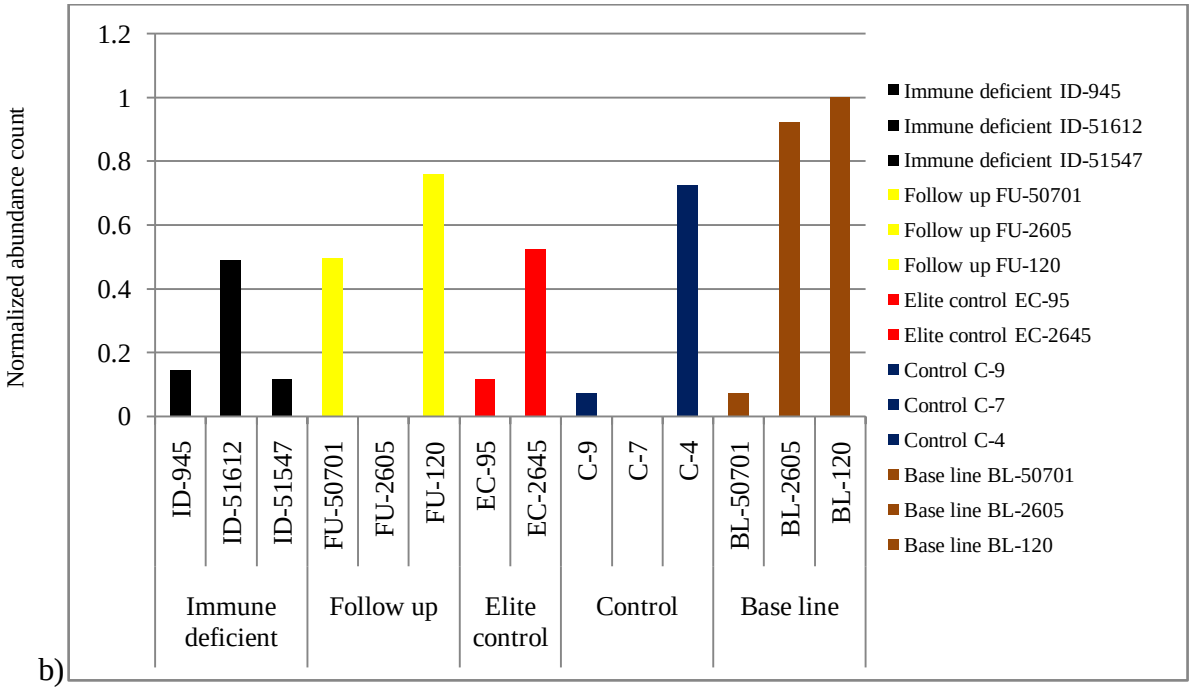
3.3.6. Genus distribution:

The NAC of *Eubacterium*into different groups were significantly different. The NAC of the immune deficient group was lower compared to the other HIV infected groups. ID-945 and ID-51547 had NAC of 0.182 and 0 respectively. NAC of C-9, C-7 and C-4 were 0, 0.083 and 0.24 which is much lower than the follow up, elite control and baseline group. The NAC of FU-50701, FU-2605 and FU-120 were 0.876, 0.653 and 0.682. The NAC of BL-50701, BL-2605 and BL-120 were 0.741, 0.94 and 1.00. BL-120 and BL-2605 had higher NAC compare to FU-120 and FU-2605. The NAC of EC-95 and EC-2645 from the elite control group were 0.97 and 0.879 which was closer to the BL-120 and BL-2605 (figure 3.13a). Bacterial genus R*oseburia* which was highly abundant in two the baseline group members are less abundant in the follow up group, even absent in one follow up sample. The NAC of BL-120 and BL-2605 were 1.00 and 0.922 which is much higher than their corresponding subjects in the follow up group such as FU-120 and FU-2605 had NAC of 0.757 and 0. In addition to this, C-9 and C-7 had much more lower NAC value such as 0.073 and 0 compare to the other subjects whoes depict an abundance (fig 3.13b).The NAC of *Ruminococcus* was significantly

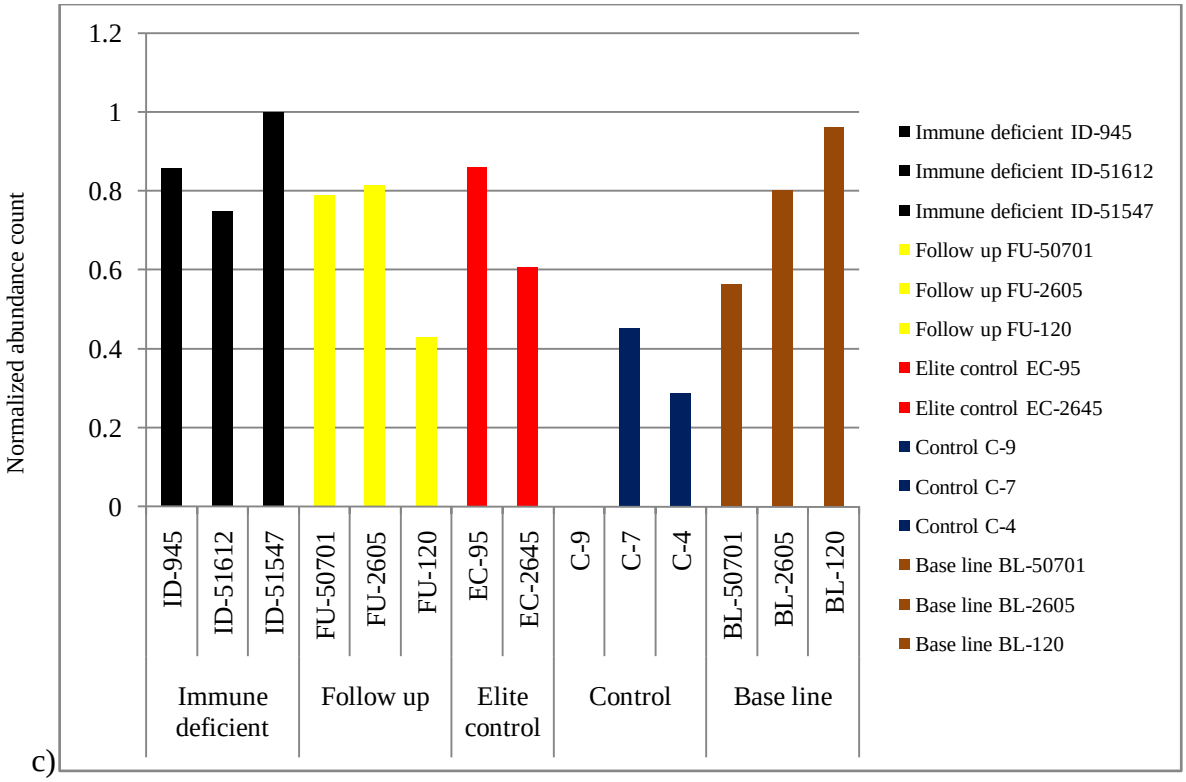
different among the groups. Whereas the NAC of C-9, C-7 and C-4 were 0, 0.453 and 0.287. Subjects from the control group showed the relatively lower NAC value than other HIV infected subjects (figure 3.13c).



a)



b)



c)

Figure 3.13: NAC of different genus in different groups.a) *Eubacterium,* b) *Rusoburia* and c)Ruminococcus.*Different subjects under same group are in same colour.X-axis represents fourteen code number (e.g., 9) with their group code (e.g., C for Control) and Y-axis: Normalized abundance count (0 to 1) of each subject. *= P<0.05.

3.3.7. Species distribution:

The NAC of two species from the genus *Eubacterium* weresignificantly different among the groups. *Eubacterium rectal*was abundant in elite control samples. The NAC of EC-95 and EC-2645 were 0.908 and 0.658. The NAC of FU-50701, FU-2605 and FU-120 were 0.9, 0.715 and 0.76. Compare with these subjects NAC of BL-50701, BL-2605 and BL-120 0.807, 0.956 and 1.00 respectively, depicts that NAC was lowered in the patient 2605 and 120. C-9, C-7 and C-4 had a lower NAC score such as 0, 0.29 and 0.074 respectively. The NAC score was 0 to all of the immune deficient patients(figure 3.14a). The NAC score of *Eubacteriumsiraeum*in ID-945, ID-51612 and ID-51547 were 0.118, 0.783 and 0 which depict this species had lower abundance. FU-50701, FU-120, C-9, C-7, BL-50701 and BL-2605 did not show NAC score. Two subjects of elite control group such as EC-95 and EC-2645 showed highest NAC score, which is 1.00 (figure 3.14b).

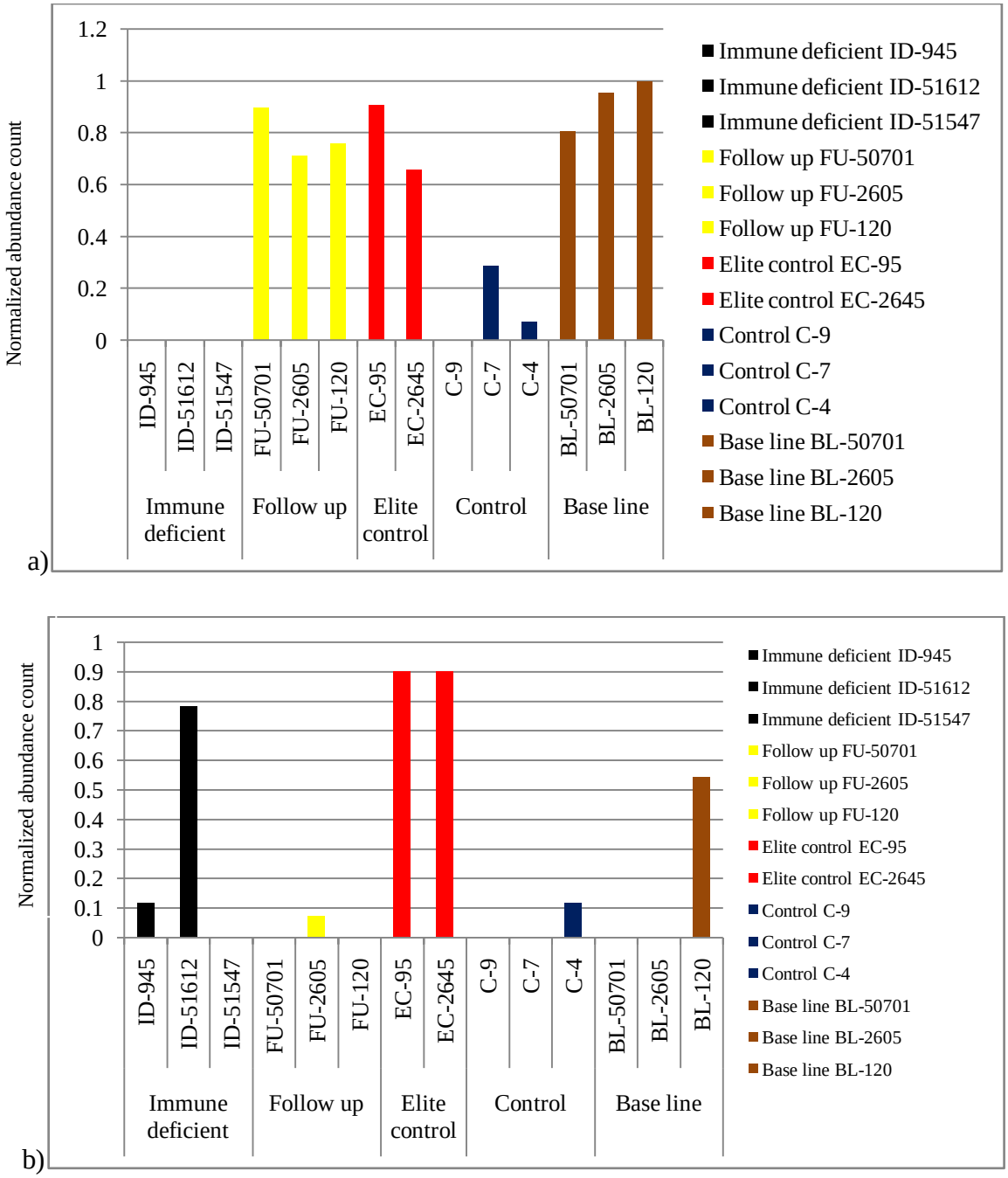


Figure 3.14: NAC scores of the species of the genus *Eubacterium*. a)*Eubacterium rectal and* b) *Eubacteriumsiraeum.* Different subjects under same group are in same colour. X-axis represents fourteen code number (e.g., 9) with their group code (e.g., C for Control) and Y-axis: Normalized abundance count (0 to 1) of each subject.

3.3.8. Distribution of Verrucomicrobia phylum at lower taxonomic level:

The bacterial phylum Verrucomicrobia showed an intriguing pattern of distribution where it is more abundant in the immune-deficient group which continues to its lower taxonomic level. There was only one species found in these samples which

is*Akkermansiamuciniphila*The NAC score of ID-945, ID-51612 and ID-51547 were 1.00, 0.96 and 0.08. The NAC score of EC-95, C-9 and BL-50701 were 0.797, 0.502 and 0.047. This species was completely absent in follow up group, BL-50701 showed very lower abundance (figure 3.15).

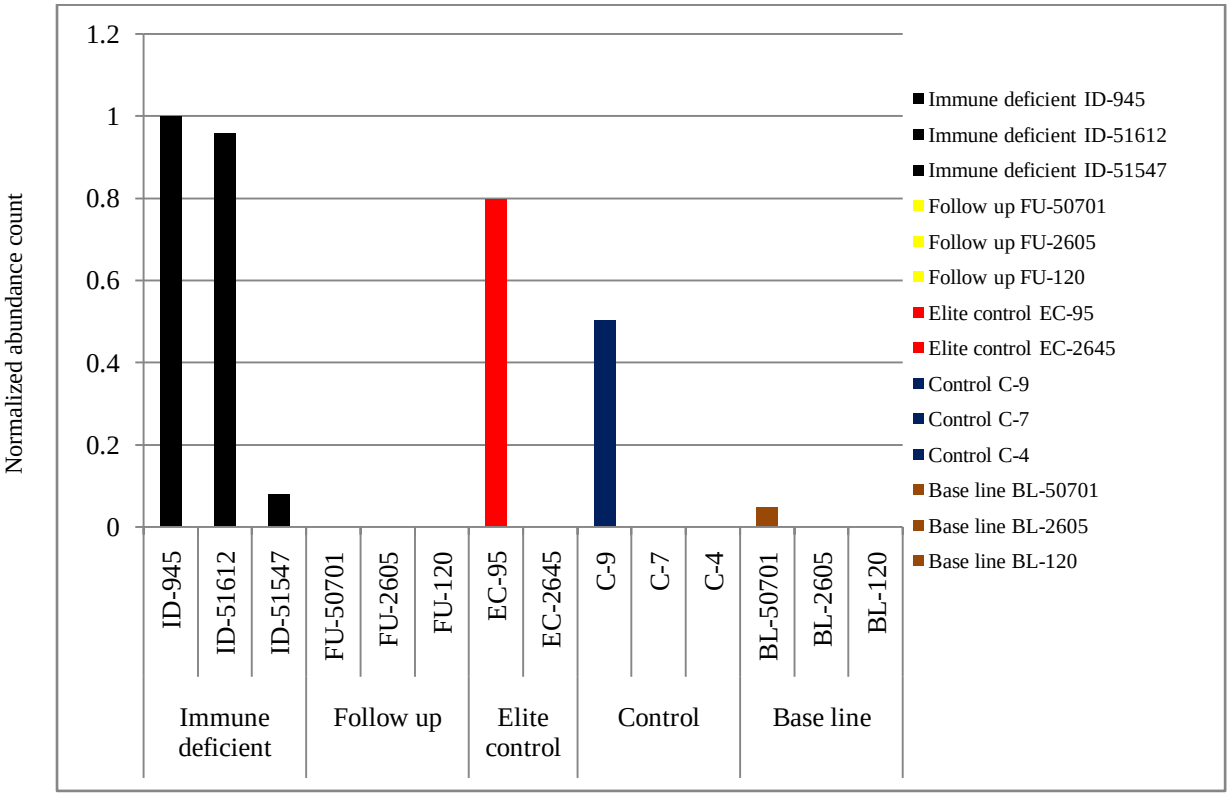


Figure 3.15: The normalized abundance counts of*Akkermansiamuciniphila*in fourteen subjects. Different subjects under same group are in same colour. X-axis represents fourteen code number (e.g., 9) with their group code (e.g., C for Control) and Y-axis: Normalized abundance count (0 to 1) of each subject.

3.4. Functional abundance:

It is necessary to understand the functional predictions along with the microbial community composition to understand the community information.Sequencing of total microbial community DNA by shotgun metagenomics provides the window for functional screening. Genes involved in different metabolic functions are predicted by matching the known functional gene sequences in the databases. Shotgun metagenomic sequencing approaches measures the levels of mRNA which in turn measures the level of functional genes expressed under specific conditions such as diet, disease etc(Lozupone, Stombaugh, Gordon, Jansson, & Knight, 2012).

3.4.1. Functional category hits distribution:

After grouping annotated sequence sets into higher level functional groups, these sequence sets were compared with a number of protein database projects e.g., SEED, SubSystem, COGs, NOGs, KEGG orthologs can curate those functional hierarchies. Protein coding genes were analyzed using FragGeneScan provided by MG-RAST pipeline. The highest number of predicted function was 126,400 in ID-51612, followed by EC-95 which had 119,865 predicted functions. And next to this subject BL-50701 showed 118,400 predicted functions. Lower number of functions predicted from the subjects of the control group (figure 3.16a).

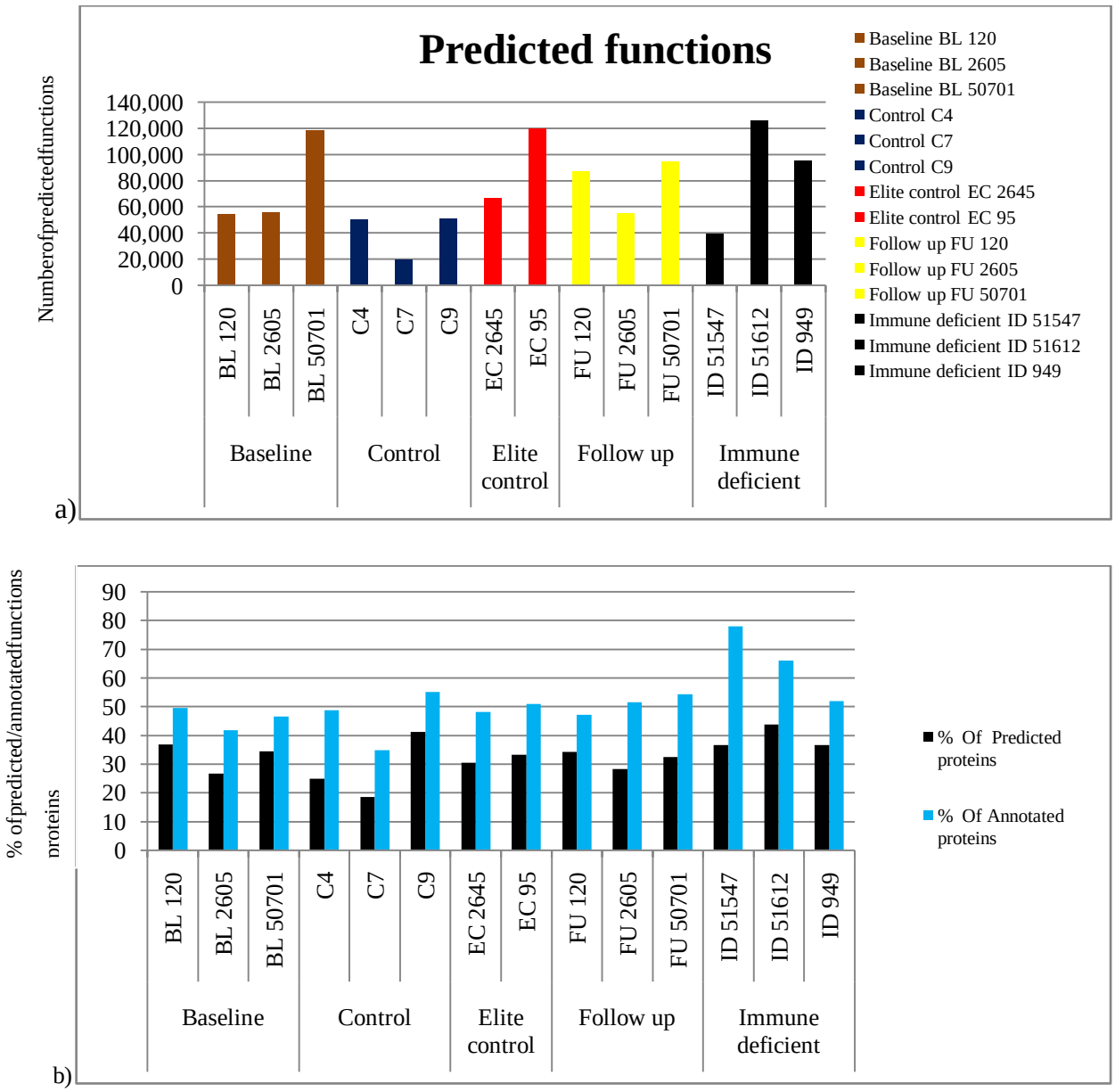
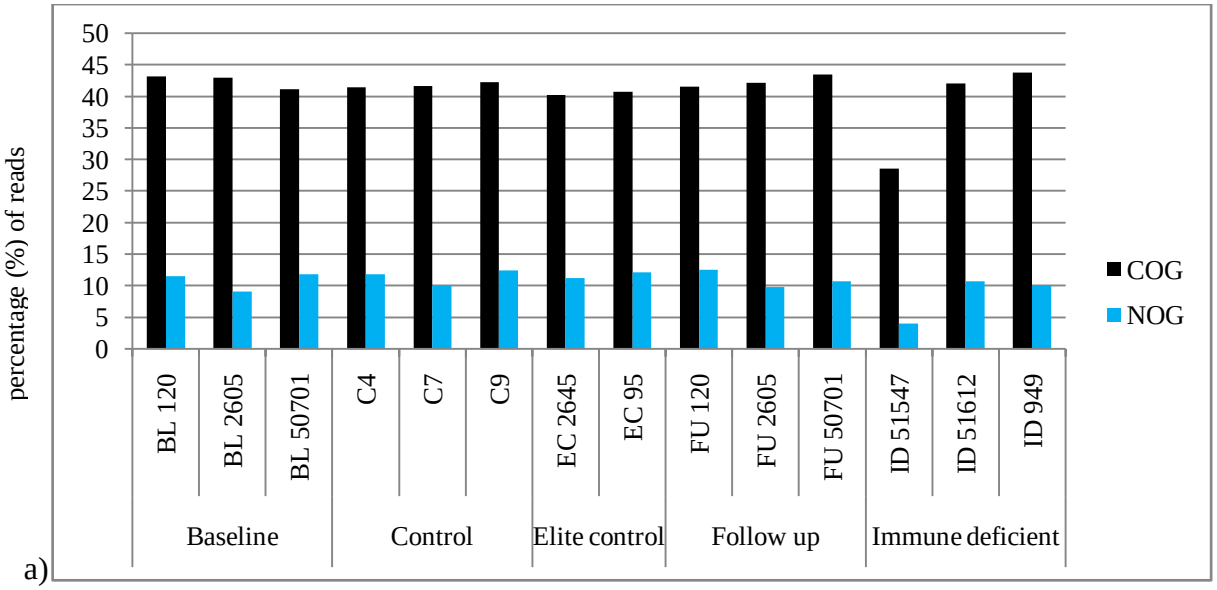
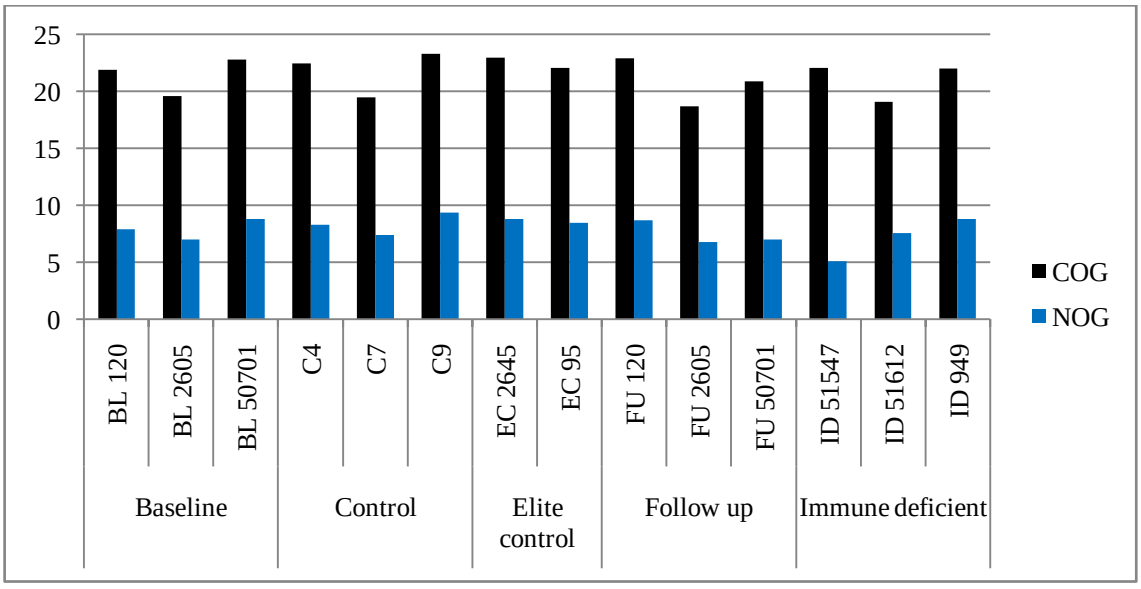


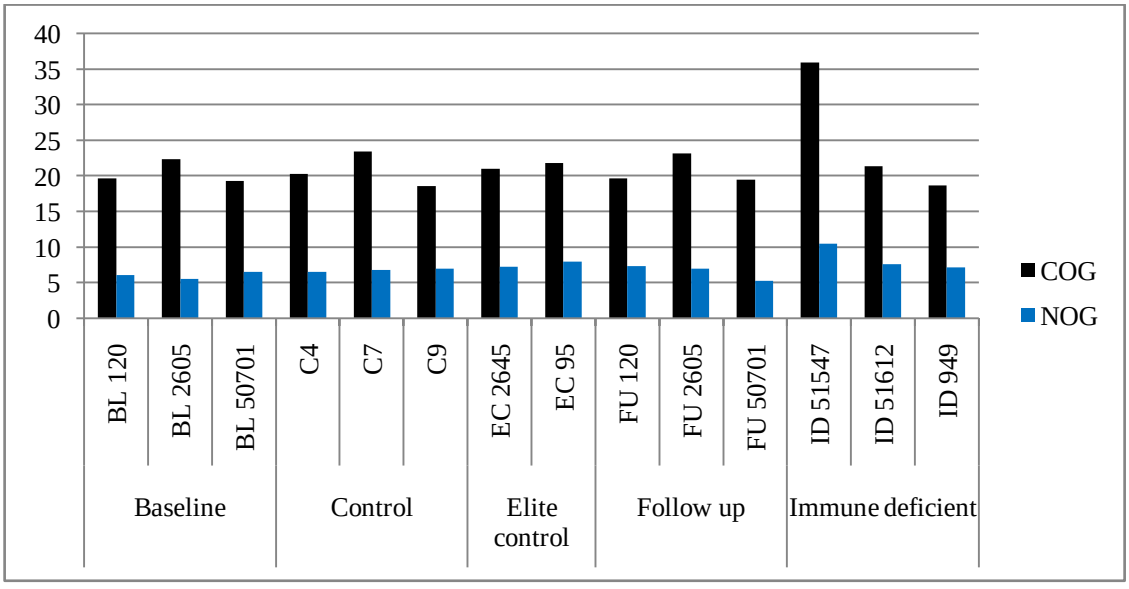
Figure 1.16: Distribution of a)predicted functions, b) predicted and annotated proteins at the highest level supported SubSystem. X-axis: percentage (%)


percentage (%) of reads
50
45
40
35
30
25
20
15
10
5
0
BL 120
BL 2605
BL 50701
C4
C7
C9
EC 2645
EC 95
FU 120
FU 2605
FU 50701
ID 51547
ID 51612
ID 949
Baseline
Control
Elite control
Follow up
Immune deficient
COG
NOG

a)


percentage (%) of reads
25
20
15
10
5
0
BL 120
BL 2605
BL 50701
C4
C7
C9
EC 2645
EC 95
FU 120
FU 2605
FU 50701
ID 51547
ID 51612
ID 949
Baseline
Control
Elite control
Follow up
Immune deficient
COG
NOG

b)


percentage (%) of reads
40
35
30
25
20
15
10
5
0
BL 120
BL 2605
BL 50701
C4
C7
C9
EC 2645
EC 95
FU 120
FU 2605
FU 50701
ID 51547
ID 51612
ID 949
Baseline
Control
Elite control
Follow up
Immune deficient
COG
NOG

c)

Figure 3.17**:** The distribution of functional categories a) metabolism, b) cellular processing and signaling and c) information storage and processing for at the highest level supported by NOG and COG functional hierarchies. The numberson the Y axisindicates the percentage (%) of reads with predicted protein functions annotated to the category for the given source.

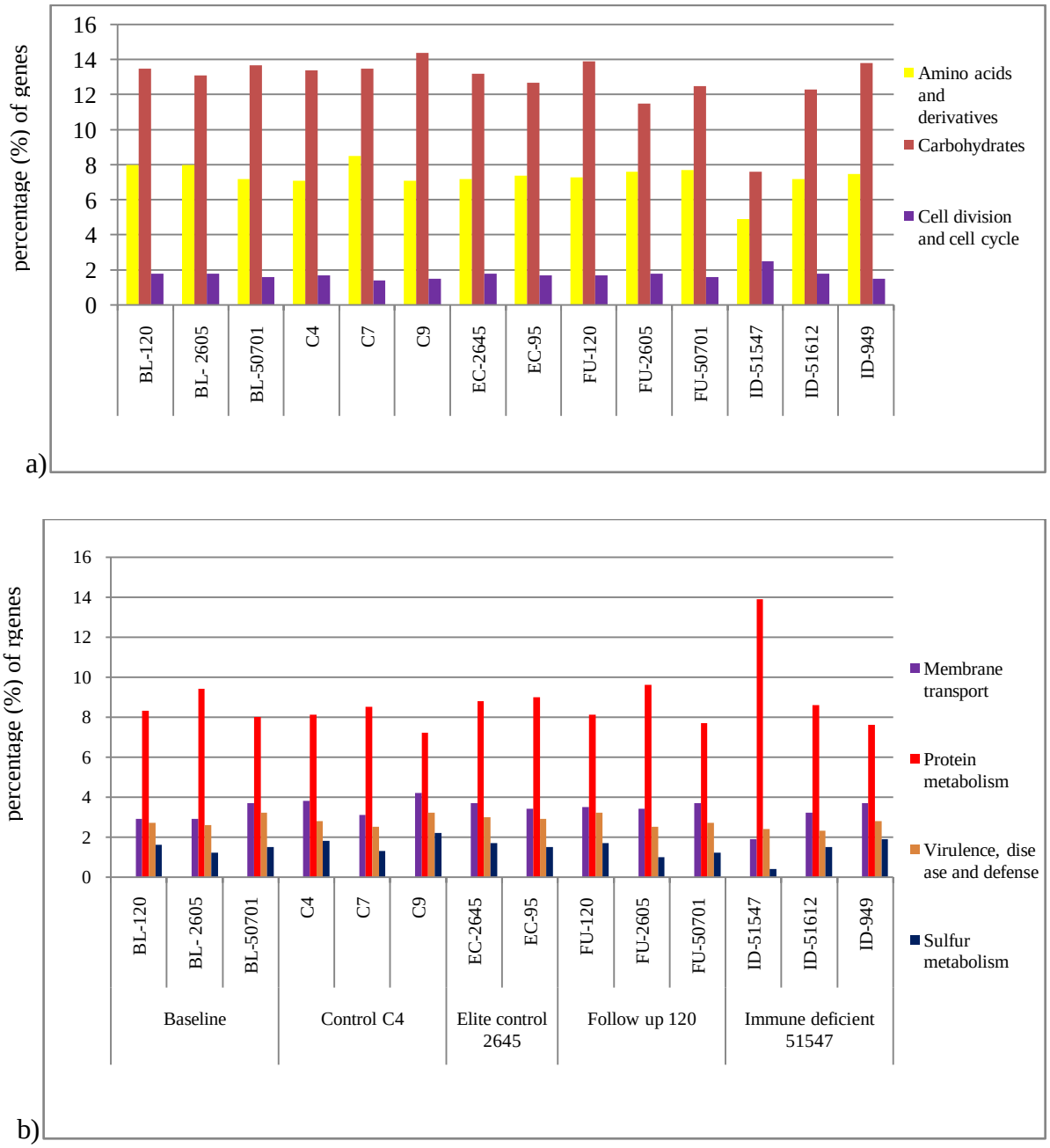


Figure 3.18: Distribution of Subsystems functional categories. A) Genes function for amino acid derivatives and derivatives, carbohydrates and, cell divisions and cell cycle. b) Genesfunction for membrane transport, protein metabolism, virulence, disease and defense and sulfur metabolism in all 14metagenomes. The y-axis represents the percentage of gene functioning in each category.

4. Discussion:

4.1. Taxonomic and functional abundance:

There are many factors that can alter the composition of gut microbiota over time such as geography, food culture, disease condition, treatment etc.Looking at the phylum level, it shows that the bacterial phyla Actinobacteria and Bacteroidetes interplay between them. Actinobacteria and Bacteroidetes abundance analysis showed that the latter has a higher abundance than the former among the groups except the immune-deficient group.Bacterial Phylum Verrucomicrobia was dominated only in immune-deficient group. *Akkermansiamuniciphila*was the only species of this phylum that was abundant in the same manner throughout the taxonomic levels. One study suggested that broad spectrum antibiotic treatment can induce high level colonization of Verrucomicrobia(Dubourg et al., 2013)but it is unclear that the immune-deficient patients received any antibiotic treatment earlier.

Gammaproteobacteriahas a higher abundance in the fecal sample of HIV infected patients than the healthy control (Mutlu et al., 2014)whereas the follow up group in our study showed higher abundance compared with its corresponding baseline group and higher than the control group. In contrast to these,the elite control group had lower abundance compared withall the group.

Fecal microbiota of HIV infected patients in the US where the untreated HIV individuals with chronic infection hada significantly higher relative abundance of Prevotellaceae (Prevotella), Erysipelotrichaceae (Catenibacterium and Bulleidia), Veillonellaceae (Dialister and Mitsuokella) and Clostridium cluster XIII and genus Bifidiobacterium, and*Bacteroidaceae*(Bacteroides), *Rikenellaceae* (Alistipes) and *Porphromonadaceae*(Parabacteroidetes) have higher abundance levels in the gut of HIV negative controls(Lozupone et al., 2013). In the present study, we found thatthe bacterial class Erysipelotrichi is abundant in the baseline group. In contrast to this, the follow up group comprised of the same patients showed lower abundance compared with baseline group. In addition to this the bacterial phyla Bacteroidetes was relatively abundant in the elite control group compared to others. Other sample groups showed some variation within the group, but the follow up group had increased abundance levels than the baseline group. Transgenic expression of DEFA5 (α-defense) causes an increase of Bacteroidetes which is significant and loss of Clostridia, Bacilli and Erysipelotrichias well as IL-17 producing T cells in the intestinal LP (Salzman et al., 2010). The substantial part of the total gut microbiota is

comprised of gram positive and rod shaped commensal bacterial class Clostridia which exerts a strong influence on the host immune system(Lopetuso, Scaldaferri, Petito, & Gasbarrini, 2013). According to our study, it was demonstrated that the abundance level of Clostridia was lower in the control group than the other HIV infected group. The immune stressed condition of these groups might be induced their abundance as Clostridia plays a critical role in the body's immune defense mechanism. However Clostridia also consists of some pathogenic bacteria e.g., *Clostridium difficile* which in turn causes difficulty in HIV affected patients such as diarrhea (Sivapalasingam & Blaser, 2005).Studies on gut microbiota suggested that normal gut microbiota is enriched with *Ruminococcaceae*and *Lanchnospiraceae* families (Lozupone et al., 2012). In our study both of them have a lower abundance in the control groupthan the other groups. A study on vaginal microbiota after a treatment with probiotics demonstrated a decrease in abundance of bacterial family *Lachnospiraceae*(Ling et al., 2013). A similar pattern was observed in the gut of the subjects of this study between the Baseline and Follow up groups. However, there were alsointer individual variation. However environment between the gut and vagina are different and having different environments they might be expected to show a different pattern, butthey showed the same pattern. The species from the family *Ruminococcaceae* increased in relative abundance with HIV infection(Lozupone et al., 2013). The inter-group variation among the group in our study showed that it was less abundant in the control group, while the immune deficient showed higher abundance than the others. The comparison between follow up and baseline showed that one of them experienced decreased abundance (BL-120) than the earlier state where as another sample (BL-2605) was in steady state and another one (BL-50701) showed increased abundance. So the probiotic supplement showsnoticeable impact on *Ruminococcaceae*.There was a declining pattern of *Roseburia*abundance observed between baseline and follow up groups. Although it was demonstrated that *Roseburia* showed an increase in abundance in association to the biomarkers related to microbial translocation, it does not mean it is a probable cause, might be it is a reflection of HIV disease progression (Lozupone et al., 2014).In our study the abundance level was decreased in follow up group compared withbaseline group. Therefore, it is assumed that the probiotics may have an effect on microbial translocation. There was a declining pattern of Eubacterium observed in highly active antiretroviral therapy (HAART)-naive HIV-1-infected adults when they were supplemented with prebiotic oligosaccharides (Gori et al., 2011). Our study reflects the same effects, although the supplement was different as well as the patients did not receive antiviral therapy.

It is very hard to come a general conclusion about the overall effect of probiotics on the abundance of genes that function in the metabolism of amino acids and their derivatives because there are a number of metabolic reactions involved. The data did not show remarkable differences among these groups. Two metagenomic samples from the follow up group (FU-120 and FU-2605) showed decreased abundances of genes that function in the metabolism of amino acids and their derivatives compared with its corresponding state in baseline group. Studies on identical twins and germ free mice demonstrated that probiotics (FMP-fermented milk products) changes the expression of microbial-encoded enzymes, especiallyrelated to carbohydrate metabolism (McNulty et al., 2011) but in our study these geneswere decreased at follow up group whilethe control group showed higher abundance than the immune deficient group. Abundance patterns of genesthat function in protein metabolism were similar both in baseline and follow up group, but that does not mean that there was no effect of probiotics. Probiotics might restore the abundance of genes that function in protein metabolism, which was less abundant in early stage (baseline) because the control group had the similar pattern of abundance where as the elite control group had relatively higher abundances than the control. In addition to this, the abundance of gene that function in carbohydrate metabolism was higher than those functions in protein metabolism, leading to the assumption that the food diet of the patients in this study was carbohydrate-rich. Genes that function in membrane transport had higher abundance in follow up group than their corresponding members in the baseline group. The control and follow up groups exhibited the same pattern while the immune deficient group had lower abundance level, which suggests that the probiotic supplement can improve the genesfunction in membrane transport. Genesthat function in virulence, disease and defense had lower abundance in the immune deficient group than the other four groups which seems to be normal as the immune deficient group's progress to AIDS.

4.2 Limitation of this study:

It’s very hard to come an affirmative conclusion from this study as the data setwas small. The study of human gut microbiota is a new and important topic to understand its correlation with different disease conditions. Therefore, availability of data is limited. In addition, metagenomic studies on the correlation between gut microbiota, HIV/AIDS and probiotics are very few. However, there are many factors that can alter the pattern of gut microbiota other than the aforementioned factors. One metagenomicdata from the elite control group was discarded due to bad quality. To get a widespread view in this area, more study is required.

There is very little review literature is available to make conclusions from an identified pattern both in taxonomic abundance and functional abundance.

4.3 Suggestion of future work:

To give a strong statistical power to this study, it is recommended to repeat this experiment with a large sample size. Since the probiotics are reported to play significant role in immune response regulation, it is suggested to measure different immunological markers which are fluctuate due to HIV infection such as $CD4^{+}$T-cell, different pro inflammatory and anti inflammatorymolecules. In addition to this, detection of the markers related to microbial translocation through the gut wall can provide a better view of the effect of probiotic on translocation. The metagenomic study of gut microbiota greatly varies on many factors such as sampling locations, type of samples as well as geographical positions, food habit, lifestyle, age, sequencing techniques, library preparation protocols, PCR primers, how the data are analyzed. So combination the study can provide a wider view of the probiotic induced alteration of gut microbiota.

5. Conclusion:

Firmicuteswas the only bacterial phylum which issignificantly different among the groups, and less abundant in the control group. Verrucomicrobia was dominated only in the immune deficient group where as it had very low abundance in the follow up and the baseline group. Bacterial phylum Actinobacteria was less abundant than the bacteroidets. The gut microbiota of the HIV infected patients along with the control group. The gut microbiota alteration due to probiotic treatment influence the family, which might reflect the effect of the probiotics on the gut microbiota. Abundance of the genus *Eubacterium* was relatively decreased, which was also observed in the ART treated following prebiotic supplemented patients. A probiotic supplement can improve the gene functions related to membrane transport and the gene functioning of defense mechanism was less abundant in the immunodeficient group. These findings suggested a probable relationship among gut microbiota, HIV infection and probiotics. In addition, an extended work on immunological markers and larger sample sizesto elucidate a correlation with them.

6. References

Adlerberth, I., & Wold, A. E. (2009). Establishment of the gut microbiota in Western infants. *Acta Paediatr, 98*(2), 229-238. doi: 10.1111/j.1651-2227.2008.01060.x

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Andreas Wilke, Elizabeth M. Glass, Jared Bischof, Daniel Braithwhaite, mark DSouza, Wolfgang Gerlach, Travis Harrison, Kevin Keegan, Hunter Matthews, Tobias Paczian, Wei Tang, William L. Trimble, Jared Wilkening, Narayan Desai and Folker Meyer. (2013). MG-RAST Technical report and manual for version 3.3.6 – rev. 1.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., . . . Rohwer, F. (2006). The marine viromes of four oceanic regions. *PLoS Biol, 4*(11), e368. doi: 10.1371/journal.pbio.0040368

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., . . . Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature, 473*(7346), 174-180. doi: 10.1038/nature09944

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., . . . Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics, 9*, 75. doi: 10.1186/1471-2164-9-75

Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., & Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science, 307*(5717), 1915-1920. doi: 10.1126/science.1104816

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol, 340*(4), 783-795. doi: 10.1016/j.jmb.2004.05.028

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., . . . Pomp, D. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci U S A, 107*(44), 18933-18938. doi: 10.1073/pnas.1007028107

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., & Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics, 8*, 209. doi: 10.1186/1471-2105-8-209

Bouskra, D., Brezillon, C., Berard, M., Werts, C., Varona, R., Boneca, I. G., & Eberl, G. (2008). Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. *Nature, 456*(7221), 507-510. doi: 10.1038/nature07450

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R. A., Felts, B., . . . Rohwer, F. (2008). Viral diversity and dynamics in an infant gut. *Res Microbiol, 159*(5), 367-373. doi: 10.1016/j.resmic.2008.04.006

Brenchley, J. M., & Douek, D. C. (2008). HIV infection and the gastrointestinal immune system. *Mucosal Immunol, 1*(1), 23-30. doi: 10.1038/mi.2007.1

Brenchley, J. M., Price, D. A., Schacker, T. W., Asher, T. E., Silvestri, G., Rao, S., . . . Douek, D. C. (2006). Microbial translocation is a cause of systemic immune activation in chronic HIV infection. *Nat Med, 12*(12), 1365-1371. doi: 10.1038/nm1511

Brenchley, J. M., Schacker, T. W., Ruff, L. E., Price, D. A., Taylor, J. H., Beilman, G. J., . . . Douek, D. C. (2004). CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *J Exp Med, 200*(6), 749-759. doi: 10.1084/jem.20040874

Burke, C., Kjelleberg, S., & Thomas, T. (2009). Selective extraction of bacterial DNA from the surfaces of macroalgae. *Appl Environ Microbiol, 75*(1), 252-256. doi: 10.1128/AEM.01630-08

Carvalho, F. A., Koren, O., Goodrich, J. K., Johansson, M. E., Nalbantoglu, I., Aitken, J. D., . . . Gewirtz, A. T. (2012). Transient inability to manage proteobacteria promotes chronic gut

inflammation in TLR5-deficient mice. *Cell Host Microbe, 12*(2), 139-152. doi: 10.1016/j.chom.2012.07.004

Chow, J., Lee, S. M., Shen, Y., Khosravi, A., & Mazmanian, S. K. (2010). Host-bacterial symbiosis in health and disease. *Adv Immunol, 107*, 243-274. doi: 10.1016/B978-0-12-381300-8.00008-3

Claus, S. P., Ellero, S. L., Berger, B., Krause, L., Bruttin, A., Molina, J., . . . Nicholson, J. K. (2011). Colonization-induced host-gut microbial metabolic interaction. *MBio, 2*(2), e00271-00210. doi: 10.1128/mBio.00271-10

Clemente, J. C., Ursell, L. K., Parfrey, L. W., & Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell, 148*(6), 1258-1270. doi: 10.1016/j.cell.2012.01.035

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., . . . Tiedje, J. M. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res, 37*(Database issue), D141-145. doi: 10.1093/nar/gkn879

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science, 300*(5617), 286-290. doi: 10.1126/science.1084564

Cunningham-Rundles, S., Ahrne, S., Johann-Liang, R., Abuav, R., Dunn-Navarra, A. M., Grassey, C., . . . Cervia, J. S. (2011). Effect of probiotic bacteria on microbial host defense, growth, and immune function in human immunodeficiency virus type-1 infection. *Nutrients, 3*(12), 1042-1070. doi: 10.3390/nu3121042

Delmont, T. O., Robe, P., Clark, I., Simonet, P., & Vogel, T. M. (2011). Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods, 86*(3), 397-400. doi: 10.1016/j.mimet.2011.06.013

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., . . . Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol, 72*(7), 5069-5072. doi: 10.1128/AEM.03006-05

Dethlefsen, L., Huse, S., Sogin, M. L., & Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol, 6*(11), e280. doi: 10.1371/journal.pbio.0060280

Diaz Heijtz, R., Wang, S., Anuar, F., Qian, Y., Bjorkholm, B., Samuelsson, A., . . . Pettersson, S. (2011). Normal gut microbiota modulates brain development and behavior. *Proc Natl Acad Sci U S A, 108*(7), 3047-3052. doi: 10.1073/pnas.1010529108

Dicks, L. M., Fraser, T., ten Doeschate, K., & van Reenen, C. A. (2009). Lactic acid bacteria population in children diagnosed with human immunodeficiency virus. *J Paediatr Child Health, 45*(10), 567-572. doi: 10.1111/j.1440-1754.2009.01566.x

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A, 107*(26), 11971-11975. doi: 10.1073/pnas.1002601107

Dubourg, G., Lagier, J. C., Armougom, F., Robert, C., Audoly, G., Papazian, L., & Raoult, D. (2013). High-level colonisation of the human gut by Verrucomicrobia following broad-spectrum antibiotic treatment. *Int J Antimicrob Agents, 41*(2), 149-155. doi: 10.1016/j.ijantimicag.2012.10.012

Elinav, E., Strowig, T., Kau, A. L., Henao-Mejia, J., Thaiss, C. A., Booth, C. J., . . . Flavell, R. A. (2011). NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell, 145*(5), 745-757. doi: 10.1016/j.cell.2011.04.022

Ellis, C. L., Ma, Z. M., Mann, S. K., Li, C. S., Wu, J., Knight, T. H., . . . Asmuth, D. M. (2011). Molecular characterization of stool microbiota in HIV-infected subjects by panbacterial and order-level 16S ribosomal DNA (rDNA) quantification and correlations with immune activation. *J Acquir Immune Defic Syndr, 57*(5), 363-370. doi: 10.1097/QAI.0b013e31821a603c

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., . . . Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Res, 38*(Database issue), D211-222. doi: 10.1093/nar/gkp985

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., . . . Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods, 7*(6), 461-465. doi: 10.1038/nmeth.1459

Fuller, R., & Jayne-Williams, D. J. (1970a). Resistance of the fowl (Gallus domesticus) to invasion by its intestinal flora. I. The effect of hormonal bursectomy on the invasiveness of the intestinal microflora of the fowl. *Res Vet Sci, 11*(4), 363-367.

Fuller, R., & Jayne-Williams, D. J. (1970b). Resistance of the fowl (Gallus domesticus) to invasion by its intestinal flora. II. Clearance of translocated intestinal bacteria. *Res Vet Sci, 11*(4), 368-374.

Funderburg, N., Luciano, A. A., Jiang, W., Rodriguez, B., Sieg, S. F., & Lederman, M. M. (2008). Toll-like receptor ligands induce human T cell activation and death, a model for HIV pathogenesis. *PLoS One, 3*(4), e1915. doi: 10.1371/journal.pone.0001915

Gautreaux, M. D., Deitch, E. A., & Berg, R. D. (1994a). Bacterial translocation from the gastrointestinal tract to various segments of the mesenteric lymph node complex. *Infect Immun, 62*(5), 2132-2134.

Gautreaux, M. D., Deitch, E. A., & Berg, R. D. (1994b). T lymphocytes in host defense against bacterial translocation from the gastrointestinal tract. *Infect Immun, 62*(7), 2874-2884.

Gilbert, J. A., Field, D., Swift, P., Thomas, S., Cummings, D., Temperton, B., . . . Muhling, M. (2010). The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One, 5*(11), e15545. doi: 10.1371/journal.pone.0015545

Godzik, A. (2011). Metagenomics and the protein universe. *Curr Opin Struct Biol, 21*(3), 398-403. doi: 10.1016/j.sbi.2011.03.010

Goodman, A. L., & Gordon, J. I. (2010). Our unindicted coconspirators: human metabolism from a microbial perspective. *Cell Metab, 12*(2), 111-116. doi: 10.1016/j.cmet.2010.07.001

Goossens, H., Ferech, M., Vander Stichele, R., Elseviers, M., & Group, Esac Project. (2005). Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet, 365*(9459), 579-587. doi: 10.1016/S0140-6736(05)17907-0

Gori, A., Rizzardini, G., Van't Land, B., Amor, K. B., van Schaik, J., Torti, C., . . . Clerici, M. (2011). Specific prebiotics modulate gut microbiota and immune activation in HAART-naive HIV-infected adults: results of the "COPA" pilot randomized trial. *Mucosal Immunol, 4*(5), 554-563. doi: 10.1038/mi.2011.15

Gori, A., Tincati, C., Rizzardini, G., Torti, C., Quirino, T., Haarman, M., . . . Clerici, M. (2008). Early impairment of gut function and gut flora supporting a role for alteration of gastrointestinal mucosa in human immunodeficiency virus pathogenesis. *J Clin Microbiol, 46*(2), 757-758. doi: 10.1128/JCM.01729-07

Group, Nih Hmp Working, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., . . . Guyer, M. (2009). The NIH Human Microbiome Project. *Genome Res, 19*(12), 2317-2323. doi: 10.1101/gr.096651.109

Hoff, K. J., Lingner, T., Meinicke, P., & Tech, M. (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res, 37*(Web Server issue), W101-105. doi: 10.1093/nar/gkp327

Holmes, E., Li, J. V., Athanasiou, T., Ashrafian, H., & Nicholson, J. K. (2011). Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. *Trends Microbiol, 19*(7), 349-359. doi: 10.1016/j.tim.2011.05.006

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., & Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol, 8*(7), R143. doi: 10.1186/gb-2007-8-7-r143

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res, 17*(3), 377-386. doi: 10.1101/gr.5969107

Huurre, A., Kalliomaki, M., Rautava, S., Rinne, M., Salminen, S., & Isolauri, E. (2008). Mode of delivery - effects on gut microbiota and humoral immunity. *Neonatology, 93*(4), 236-240. doi: 10.1159/111102

Illumina. (2012). Nextera® XT DNA Sample Preparation Guide.

Illumina. (2014a). MiSeq Systems. Focused power. Speed and simplicity for targeted and small genome sequencing.

Illumina. (2014b). Nextera® XT DNA sample preparation kit. The fastest and easiest sample prep workflow for small genomes, PCR amplicons and plasmids. .

Irvine, S. L., Hummelen, R., Hekmat, S., Looman, C. W., Habbema, J. D., & Reid, G. (2010). Probiotic yogurt consumption is associated with an increase of CD4 count among people living with HIV/AIDS. *J Clin Gastroenterol, 44*(9), e201-205. doi: 10.1097/MCG.0b013e3181d8fba8

Jimenez, E., Marin, M. L., Martin, R., Odriozola, J. M., Olivares, M., Xaus, J., . . . Rodriguez, J. M. (2008). Is meconium from healthy newborns actually sterile? *Res Microbiol, 159*(3), 187-193. doi: 10.1016/j.resmic.2007.12.007

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res, 32*(Database issue), D277-280. doi: 10.1093/nar/gkh063

Klatt, N. R., Canary, L. A., Vanderford, T. H., Vinton, C. L., Engram, J. C., Dunham, R. M., . . . Brenchley, J. M. (2012). Dynamics of simian immunodeficiency virus SIVmac239 infection in pigtail macaques. *J Virol, 86*(2), 1203-1213. doi: 10.1128/JVI.06033-11

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., . . . Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A, 108 Suppl 1*, 4578-4585. doi: 10.1073/pnas.1000081107

Koup, R. A., Safrit, J. T., Cao, Y., Andrews, C. A., McLeod, G., Borkowsky, W., . . . Ho, D. D. (1994). Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J Virol, 68*(7), 4650-4655.

Kuss, S. K., Best, G. T., Etheredge, C. A., Pruijssers, A. J., Frierson, J. M., Hooper, L. V., . . . Pfeiffer, J. K. (2011). Intestinal microbiota promote enteric virus replication and systemic pathogenesis. *Science, 334*(6053), 249-252. doi: 10.1126/science.1211057

Lasken, R. S. (2009). Genomic DNA amplification by the multiple displacement amplification (MDA) method. *Biochem Soc Trans, 37*(Pt 2), 450-453. doi: 10.1042/BST0370450

Lathrop, S. K., Bloom, S. M., Rao, S. M., Nutsch, K., Lio, C. W., Santacruz, N., . . . Hsieh, C. S. (2011). Peripheral education of the immune system by colonic commensal microbiota. *Nature, 478*(7368), 250-254. doi: 10.1038/nature10434

Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell, 124*(4), 837-848. doi: 10.1016/j.cell.2006.02.017

Ling, Z., Liu, X., Chen, W., Luo, Y., Yuan, L., Xia, Y., . . . Xiang, C. (2013). The restoration of the vaginal microbiota after treatment for bacterial vaginosis with metronidazole or probiotics. *Microb Ecol, 65*(3), 773-780. doi: 10.1007/s00248-012-0154-3

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., . . . Law, M. (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol, 2012*, 251364. doi: 10.1155/2012/251364

Lopetuso, L. R., Scaldaferri, F., Petito, V., & Gasbarrini, A. (2013). Commensal Clostridia: leading players in the maintenance of gut homeostasis. *Gut Pathog, 5*(1), 23. doi: 10.1186/1757-4749-5-23

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res, 25*(5), 955-964.

Lozupone, C. A., Li, M., Campbell, T. B., Flores, S. C., Linderman, D., Gebert, M. J., . . . Palmer, B. E. (2013). Alterations in the gut microbiota associated with HIV-1 infection. *Cell Host Microbe, 14*(3), 329-339. doi: 10.1016/j.chom.2013.08.006

Lozupone, C. A., Rhodes, M. E., Neff, C. P., Fontenot, A. P., Campbell, T. B., & Palmer, B. E. (2014). HIV-induced alteration in gut microbiota: Driving factors, consequences, and effects of antiretroviral therapy. *Gut Microbes, 5*(4).

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature, 489*(7415), 220-230. doi: 10.1038/nature11550

Lukashin, A. V., & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res, 26*(4), 1107-1115.

Macpherson, A. J., Geuking, M. B., & McCoy, K. D. (2011). Immunoglobulin A: a bridge between innate and adaptive immunity. *Curr Opin Gastroenterol, 27*(6), 529-533. doi: 10.1097/MOG.0b013e32834bb805

Marchetti, G., Tincati, C., & Silvestri, G. (2013). Microbial translocation in the pathogenesis of HIV infection and AIDS. *Clin Microbiol Rev, 26*(1), 2-18. doi: 10.1128/CMR.00050-12

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet, 9*, 387-402. doi: 10.1146/annurev.genom.9.081307.164359

Mariathasan, S., Newton, K., Monack, D. M., Vucic, D., French, D. M., Lee, W. P., . . . Dixit, V. M. (2004). Differential activation of the inflammasome by caspase-1 adaptors ASC and Ipaf. *Nature, 430*(6996), 213-218. doi: 10.1038/nature02664

Markowitz, V. M., Mavromatis, K., Ivanova, N. N., Chen, I. M., Chu, K., & Kyrpides, N. C. (2009). IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics, 25*(17), 2271-2278. doi: 10.1093/bioinformatics/btp393

Martinon, F., Burns, K., & Tschopp, J. (2002). The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-beta. *Mol Cell, 10*(2), 417-426.

McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods, 4*(1), 63-72. doi: 10.1038/nmeth976

McKenna, P., Hoffmann, C., Minkah, N., Aye, P. P., Lackner, A., Liu, Z., . . . Bushman, F. D. (2008). The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog, 4*(2), e20. doi: 10.1371/journal.ppat.0040020

McNeil, L. K., Reich, C., Aziz, R. K., Bartels, D., Cohoon, M., Disz, T., . . . Stevens, R. (2007). The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Res, 35*(Database issue), D347-353. doi: 10.1093/nar/gkl947

McNulty, N. P., Yatsunenko, T., Hsiao, A., Faith, J. J., Muegge, B. D., Goodman, A. L., . . . Gordon, J. I. (2011). The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Sci Transl Med, 3*(106), 106ra106. doi: 10.1126/scitranslmed.3002701

Mehandru, S., Poles, M. A., Tenner-Racz, K., Manuelli, V., Jean-Pierre, P., Lopez, P., . . . Markowitz, M. (2007). Mechanisms of gastrointestinal CD4+ T-cell depletion during acute and early human immunodeficiency virus type 1 infection. *J Virol, 81*(2), 599-612. doi: 10.1128/JVI.01739-06

Mellmann, A., Harmsen, D., Cummings, C. A., Zentz, E. B., Leopold, S. R., Rico, A., . . . Karch, H. (2011). Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One, 6*(7), e22751. doi: 10.1371/journal.pone.0022751

Merlini, E., Bai, F., Bellistri, G. M., Tincati, C., d'Arminio Monforte, A., & Marchetti, G. (2011). Evidence for polymicrobic flora translocating in peripheral blood of HIV-infected patients with poor immune response to antiretroviral therapy. *PLoS One, 6*(4), e18580. doi: 10.1371/journal.pone.0018580

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet, 11*(1), 31-46. doi: 10.1038/nrg2626

Meyer, F., Overbeek, R., & Rodriguez, A. (2009). FIGfams: yet another set of protein families. *Nucleic Acids Res, 37*(20), 6643-6654. doi: 10.1093/nar/gkp698

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., . . . Edwards, R. A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics, 9*, 386. doi: 10.1186/1471-2105-9-386

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc, 2010*(6), pdb prot5448. doi: 10.1101/pdb.prot5448

Mowat, A. M., & Viney, J. L. (1997). The anatomical basis of intestinal immunity. *Immunol Rev, 156*, 145-166.

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., . . . Bork, P. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res, 38*(Database issue), D190-195. doi: 10.1093/nar/gkp951

Mutlu, E. A., Keshavarzian, A., Losurdo, J., Swanson, G., Siewe, B., Forsyth, C., . . . Landay, A. (2014). A compositional look at the human gastrointestinal microbiome and immune activation parameters in HIV infected subjects. *PLoS Pathog, 10*(2), e1003829. doi: 10.1371/journal.ppat.1003829

Noguchi, H., Taniguchi, T., & Itoh, T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res, 15*(6), 387-396. doi: 10.1093/dnares/dsn027

O'Hara, A. M., & Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Rep, 7*(7), 688-693. doi: 10.1038/sj.embor.7400731

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., . . . Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res, 33*(17), 5691-5702. doi: 10.1093/nar/gki866

Paiardini, M., Frank, I., Pandrea, I., Apetrei, C., & Silvestri, G. (2008). Mucosal immune dysfunction in AIDS pathogenesis. *AIDS Rev, 10*(1), 36-46.

Palenik, B., Ren, Q., Tai, V., & Paulsen, I. T. (2009). Coastal Synechococcus metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. *Environ Microbiol, 11*(2), 349-359. doi: 10.1111/j.1462-2920.2008.01772.x

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., & Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biol, 5*(7), e177. doi: 10.1371/journal.pbio.0050177

Pennisi, E. (2011). Microbiology. Gut bacteria lend a molecular hand to viruses. *Science, 334*(6053), 168. doi: 10.1126/science.334.6053.168

Peterson, D. A., McNulty, N. P., Guruge, J. L., & Gordon, J. I. (2007). IgA response to symbiotic bacteria as a mediator of gut homeostasis. *Cell Host Microbe, 2*(5), 328-339. doi: 10.1016/j.chom.2007.09.013

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief Bioinform, 10*(4), 354-366. doi: 10.1093/bib/bbp026

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glockner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res, 35*(21), 7188-7196. doi: 10.1093/nar/gkm864

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., . . . Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature, 464*(7285), 59-65. doi: 10.1038/nature08821

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics, 13*, 341. doi: 10.1186/1471-2164-13-341

Reid, G. (2010). The potential role for probiotic yogurt for people living with HIV/AIDS. *Gut Microbes, 1*(6), 411-414. doi: 10.4161/gmic.1.6.14079

Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res, 38*(20), e191. doi: 10.1093/nar/gkq747

Roche. (2011). A New GS jnuior sequencer.

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N. J., Hentschke, M., . . . Consortium, E. coli O104:H4 Genome Analysis Crowd-Sourcing. (2011). Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N Engl J Med, 365*(8), 718-724. doi: 10.1056/NEJMoa1107643

Ronaghi, M., Uhlen, M., & Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. *Science, 281*(5375), 363, 365.

Round, J. L., Lee, S. M., Li, J., Tran, G., Jabri, B., Chatila, T. A., & Mazmanian, S. K. (2011). The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science, 332*(6032), 974-977. doi: 10.1126/science.1206095

Round, J. L., & Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nat Rev Immunol, 9*(5), 313-323. doi: 10.1038/nri2515

Salzman, N. H., Hung, K., Haribhai, D., Chu, H., Karlsson-Sjoberg, J., Amir, E., . . . Bos, N. A. (2010). Enteric defensins are essential regulators of intestinal microbial ecology. *Nat Immunol, 11*(1), 76-83. doi: 10.1038/ni.1825

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A, 74*(12), 5463-5467.

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., . . . White, O. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res, 35*(Database issue), D260-264. doi: 10.1093/nar/gkl1043

Shu, Z. J., Cao, Y., & Halmurat, U. (2011). Gut flora may offer new therapeutic targets for the traditional Chinese medicine enteric dialysis. *Expert Opin Ther Targets, 15*(10), 1147-1152. doi: 10.1517/14728222.2011.614234

Shu, Z., Ma, J., Tuerhong, D., Yang, C., & Upur, H. (2013). How intestinal bacteria can promote HIV replication. *AIDS Rev, 15*(1), 32-37.

Sivapalasingam, S., & Blaser, M. J. (2005). Bacterial diarrhea in HIV-infected patients: why Clostridium difficile, and why now? *Clin Infect Dis, 41*(11), 1628-1630. doi: 10.1086/498037

Sommer, M. O., Dantas, G., & Church, G. M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science, 325*(5944), 1128-1131. doi: 10.1126/science.1176950

Spor, A., Koren, O., & Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat Rev Microbiol, 9*(4), 279-290. doi: 10.1038/nrmicro2540

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., . . . Wooley, J. (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res, 39*(Database issue), D546-551. doi: 10.1093/nar/gkq1102

Sweeney, T. E., & Morton, J. M. (2013). The human gut microbiome: a review of the effect of obesity and surgically induced weight loss. *JAMA Surg, 148*(6), 563-569. doi: 10.1001/jamasurg.2013.5

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., . . . Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics, 4*, 41. doi: 10.1186/1471-2105-4-41

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp, 2*(1), 3. doi: 10.1186/2042-5783-2-3

Thomas, T., Rusch, D., DeMaere, M. Z., Yung, P. Y., Lewis, M., Halpern, A., . . . Kjelleberg, S. (2010). Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J, 4*(12), 1557-1567. doi: 10.1038/ismej.2010.74

Tlaskalova-Hogenova, H., Stepankova, R., Hudcovic, T., Tuckova, L., Cukrowska, B., Lodinova-Zadnikova, R., . . . Kokesova, A. (2004). Commensal bacteria (normal microflora), mucosal immunity and chronic inflammatory and autoimmune diseases. *Immunol Lett, 93*(2-3), 97-108. doi: 10.1016/j.imlet.2004.02.005

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature, 449*(7164), 804-810. doi: 10.1038/nature06244

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature, 444*(7122), 1027-1031. doi: 10.1038/nature05414

Umesaki, Y., Okada, Y., Matsumoto, S., Imaoka, A., & Setoyama, H. (1995). Segmented filamentous bacteria are indigenous intestinal bacteria that activate intraepithelial lymphocytes and induce MHC class II molecules and fucosyl asialo GM1 glycolipids on the small intestinal epithelial cells in the ex-germ-free mouse. *Microbiol Immunol, 39*(8), 555-562.

Vaishampayan, P. A., Kuehl, J. V., Froula, J. L., Morgan, J. L., Ochman, H., & Francino, M. P. (2010). Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. *Genome Biol Evol, 2*, 53-66. doi: 10.1093/gbe/evp057

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., . . . Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science, 304*(5667), 66-74. doi: 10.1126/science.1093857

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A, 95*(12), 6578-6583.

Wolochow, H., Hildebrand, G. J., & Lamanna, C. (1966). Translocation of microorganisms across the intestinal wall of the rat: effect of microbial size and concentration. *J Infect Dis, 116*(4), 523-528.

Wommack, K. E., Bhavsar, J., & Ravel, J. (2008). Metagenomics: read length matters. *Appl Environ Microbiol, 74*(5), 1453-1463. doi: 10.1128/AEM.02181-07

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput Biol, 6*(2), e1000667. doi: 10.1371/journal.pcbi.1000667

Appendix

Appendix 1: Different index sequences to generate sample sheets to demultiplex the samples

Index 1 (i7)	Sequence	Index 2 (i5)	Sequence
N701	TAAGGCGA	S501	TAGATCGC
N702	CGTACTAG	S502	CTCTCTAT
N703	AGGCAGAA	S503	TATCCTCT
N704	TCCTGAGC	S504	AGAGTAGA
N705	GGACTCCT	S505	GTAAGGAG
N706	TAGGCATG	S506	ACTGCATA
N707	CTCTCTAC	S507	AAGGAGTA
N708	CAGAGAGG	S508	CTAAGCCT
N709	GCTACGCT		
N710	CGAGGCTG		
N711	AAGAGGCA		
N712	GTAGAGGA		

Appendix 2: Arrangement of index primers. Index 1 (i7) in horizontal order and Index 2 (i5) in vertical order. It enables 96 unique combinations of duel indexing.

Index 1→		1	2	3	4	5	6	7	8	9	10	11	12
Index 2↓		N701	N702	N703	N704	N705	N706	N707	N708	N709	N710	N711	N712
A	S501						*						
B	S502					*	*						
C	S503					*	*						
D	S504					*	*						
E	S505					*	*						
F	S506					*	*						
G	S507					*	*						
H	S508					*	*						

*this combination of indiceswere used in this study.

Appendix 3: Sample grouping with their identification number and concentration (ng/µl).

Identification number	Group	Concentration (ng/µl)	Identification number	Group	Concentration (ng/µl)
120	Baseline	0.431	949	Immune deficient	0.343
120	Follow up	0.204	95	Elite control	0.866
2605	Baseline	0.14	2645	Elite control	0.207
2605	Follow up	0.11	232	Elite control	0.708
50701	Baseline	Too low	C4	Control	Too low
50701	Follow up	0.18	C7	Control	1.73
51612	Immune deficient	Too low	C9	Control	3.05
51547	Immune deficient	0.15			

Appendix 4: The number of base pairs, number of sequences, GC content (%) and mean length of each sequence (bp) before and after the quality control of 15 metagenomic data sets

Metagenome	MG-RAST ID	bp counts		Sequence counts		Mean GC content (%)		Mean length (bp)	
		Uploaded	Post QC	Uploaded	Post QC	Uploaded	Post QC	uploaded	Post QC
Baseline 120	4547096.3	22,626,182	21,010,126	191,397	176,567	46±9	46±8	118±36	118±34
Baseline 2605	4547098.3	32,991,643	29.929,230	292,270	260,703	46±10	46±9	112±38	114±5
Baseline 50701	4547100.3	54,037,976	49,467,034	434,472	398,929	48±9	48±9	124±35	124±33
Control C4	4547102.3	34,239,223	29,237,441	281,480	237,132	44±11	45±11	121±36	123±33
Control C7	4547104.3	18,593,816	15,597,400	150985	125366	44±9	47±8	123±35	124±33
Control C9	4547106.3	19,362,561	17,880,172	150,470	141,486	45±8	45±8	128±32	126±32
Elite control 2645	4547108	34,095,965	31,187,963	283,273	259,273	47±10	47±9	120±36	120±34
Elite control 95	4547110.3	56,576,419	51,586,270	470,575	428,325	48±10	49±9	120±36	120±34
Follow up 120	4547112	39,928,949	36,439,656	337,656	306,603	47±9	47±9	118±37	118±34
Follow up 2605	4547114.3	30,890,501	27,353,161	281,097	244,888	47±11	47±11	109±37	111±33
Follow up 50701	4547116.3	50,660,354	42,324,860	408,018	339,428	44±9	45±8	124±34	124±32
Immunodificient 51547	4547118.3	16,815,981	15,472,651	136,070	126,450	38±10	38±10	123±34	122±32
Immunodificient 51612	4547120.3	49,254,242	39,977,831	442,515	353,221	48±11	49±10	111±37	113±33
Immuno deficient 949	4547122.3	22,626182	21,010,126	191,397	176,567	46±9	46±8	118±36	118±34
Total		482,699,994	398,544,691	4051675	3574938				

Appendix 5: The number ofprocessed protein features and rRNA features, identified protein feature and rRNA features. Predicted protein features could be annotated with similarity to a protein of known function using M5NR database. In addition, ribosomal RNA genes are mapped to the rRNA databases.

		Protein features		rRNA features		
Metagenome	MG-RAST ID	Predicted	Indentified	Predicted	Identified	Identified functional categories
Baseline 120	4547096.3	148,086	110,363	29,707	1,431	44421
Baseline 2605	4547098.3	210,572	133527	47,251	2,245	46200
Baseline 50701	4547100.3	343,555	254,320	60,222	1,894	113860
Control C4	4547102.3	202,507	103,417	43,383	1,214	46203
Control C7	4547104.3	107,339	56,838	23,187	694	21550
Control C9	4547106.3	123,680	92,509	21,009	853	49816
Elite control 2645	4547108.3	218,395	138,044	41,844	1,378	63783
Elite control 95	4547110.3	360,440	235,699	69,659	2,196	104438
Follow up 120	4547112.3	255,141	185,187	50,654	1,866	81940
Follow up 2605	4547114.3	196,850	107,433	46,148	1,442	45510
Follow up 50701	4547116.3	293,486	174,922	57,195	1,713	80292
Immuno deficient 51547	4547118.3	107,725	50,725	25,108	521	26657
Immuno deficient 51612	4547120.3	288,153	191,300	70,361	1,967	106257
Immuno deficient 949	4547122.3	259,431	183,361	49,330	2,410	85408