Original Article

Metagenomic Analysis Of Human Gut Microbiota

1) Muhammad Moazzam Bin Rasheed
2) Marvah Mehmood Rana
3) Mian Talha Sarfraz
4) Sana Shamim
5) Irshad Begum
6) Safia Khan
7) Dr. Noor Jahan
8) Muhammad Hassan Taj*
9) Muhammad Adil Aziz

ranamoazzamuaf@gmail.com[1]
marvahmahmood@gmail.com[2]
talhasarfraz29@gmail.com[3]
sana.shamim@duhs.edu.pk[4]
niazi.irshad70@gmail.com[5]
saphiyaahmed@yahoo.com[6]
noor.jahan@duhs.edu.pk[7]
hassan.84np@gmail.com[8]
azizadil556@gmail.com

1,8,9) University of Agriculture Faisalabad Pakistan
2,3) School of Interdisciplinary Engineering & Science (SINES), NUST, Islamabad
5,6) Department of Chemistry, University of Karachi 75270, Karachi
4,7) Department of Pharmaceutical Chemistry, Dow College of Pharmacy, faculty of Pharmaceutical Sciences, Dow university of health Sciences
*Correspondence
hassan.84np@gmail.com
DOI: 10.47750/pnr.2023.14.04.78

Abstract

The human gut harbors a complex and dynamic community of microorganisms collectively known as the gut microbiota, which plays a crucial role in maintaining host health. Advancements in the field of metagenomic analysis of the human gut microbiota, shedding light on its potential implications for human health. the study of the human gut microbiota through metagenomics has unveiled an astounding diversity of microbial species, encompassing bacteria, archaea, viruses, and eukaryotic microorganisms. The gut microbiota's composition varies significantly between individuals, influenced by factors such as diet, age, genetics, geographical location, and overall health status. Advances in sequencing technologies and bioinformatics tools have facilitated comprehensive profiling of the gut microbiome, enabling researchers to explore the intricate relationships between specific microbial taxa and their functions. The gut microbiota's influence extends beyond the gastrointestinal tract, impacting various physiological processes, including metabolism, neurodevelopment, and even mental health. Dysbiosis, an imbalance in the gut microbiota composition, has been associated with various health conditions such as inflammatory bowel diseases, obesity, diabetes, and allergies. In conclusion, metagenomic analysis of the human gut microbiota has revolutionized our understanding of the microbial world residing within us. This research has paved the way for personalized approaches to healthcare, such as targeted dietary interventions, fecal microbiota transplantation, and the development of novel therapeutics aimed at modulating the gut microbiota to promote health and combat disease. Continued research in this field promises to unlock new frontiers in microbiology, advancing our knowledge of the intricate relationship between humans and their gut microbial companions.

Keywords ; Gut Microbiota, Metagenomics, Phylogenetic tree.

Introduction

Microorganisms are present in this universe from the creation of the world, they are 3.5 billion years old.[1] Scientists believed that the human body made up of different types of microorganism or on different locations in our body. These different locations in our body might be some respiratory tract, lungs, gastrointestinal tract, skin and blood of the human body. These microorganisms affect the efficiency of different functions, a process in our body and also somehow they are necessary for our life. Now you are able to check out all the process and problems which occur in our body.[2] Microbes are the essential and key elements of the worlds. According to the researchers, these microbes are a basic part of performing the function in this universe. The microbes are present in our body at different places, we cannot see them with a naked eye. Microbes that are the basic/essential part of our environment will carry out most of the processes which occur in nature.

Metagenomics is the genomics analysis of microorganism by in an environmental sample. Another definition of it, metagenomics tells us about the presence and non-presence of microbes and their genomic potential. The development of the metagenomics is due to the certain evidence of the uncultured microorganisms. The evidence about the uncultured microorganism is derived from analysis of 16S rRNA gene which is amplified from the environment. Metagenomics provided a way to study the physiology and ecology of environmental microorganisms. Through metagenomics, you are able to find the novel genes and their products.[4] Metagenomics, still a new field of to find uncultured microorganism and it is also produced well knowledge about the microbial communities. All metagenomics process will occur in the same way. Firstly, DNA is extracted. Secondly, the mixed sample could be analyzed or cloned. Then creating the libraries of these cloned genomes group and studied these libraries in different ways. The world of metagenomics is so far being very large in itself like the microbiology. Recent developments in genomic technologies have enabled the creation of innovative approaches to environmental health monitoring and risk assessment. In contrast to more traditional investigations, which can be specific to species and genes, high throughput sequencing of entire microbial communities provides snapshots of network and functional composition on a global scale.[5]

Since there have been millions of years, human and gut microbiota, which refers to the entire population of microorganisms that populate the gastrointestinal system, have developed in each other. As a consequence of this, humans are reliant on certain bacteria for the provision of essential nutrients for their diet, the formation of their immune systems, and the provision of protection against opportunistic diseases. There are broad links between the properties of the microbiome and the host genotype, despite the fact that the structure and composition of a person's gut microbiota exhibit a great amount of change. The human microbiome is a potentially adaptable phenotype that has significant consequences for human health, as indicated by patterns of disease and exposure to maternal microbiota[6-8]. In spite of the lack of microbial diversity, developed countries nearly universally gift a marked reduction in the occurrence of human gut parasites. This is a gift from evolution[9]. Studies examining the role that parasites play in

forming the intestinal microbiota are scarce, despite the fact that it is estimated that 3.5 billion people throughout the world are infected with at least one parasite (either a protozoan or a helminth) [10]. Yet, at some point in the course of evolution, intestine microbes and gut-living parasites have co-inhabited the human gastrointestinal tract. Furthermore, it is likely that the community dynamics are determined by using cutting-edge and beyond interactions (both throughout a person's lifespan and in the course of evolutionary records) among microbiota, protozoa, helminths, and the host immune reaction[11,12].

Methodology

Sample Retrieval

Sample retrieval is the first step in metagenomics undertaking. There are different unique data containing sites, from wherein we collect our data. One of them is the SRA database. SRA database is available in NCBI (National Center for Biotechnology Information). NCBI contains information about different databases relevant to biotechnology, biomedicine and bioinformatics. The accession number of our SRA data set was #ERR261975.

Sample Processing

In the process of metagenomic analysis, the preprocessing of collection reads prior to assembly, gene prediction, and annotation is an important step that is frequently skipped over. The bottom calling of the raw data that is produced by the sequencing machines is one of the steps included in preprocessing. Other steps include the elimination of cloning vector collection through vector screening, the removal of low-best bases through best trimming (which is determined by using base calling), and the elimination of verifiable collection contaminants through contaminant screening. The process of removing cloning vector sequences from base-referred to as series reads is referred to as vector screening. In metagenomic information units, the complete and accurate deletion of cloning vector series is especially significant since those record sets frequently contain large sections of extremely low coverage, and each read uniquely reflects a portion of a genome. since of this, it is especially vital to get rid of cloning vector series in the right way. The coming together of these data without first vector trimming can result in the formation of chimera contigs. In these contigs, the vector collection, which is typical of most reads, functions to bring together sequences that are not connected to one another.

The SRA data set is extracted in Galaxy Europe (https://usegalaxy.eu) from NCBI. Galaxy is a popular genomic workbench that is accessible on the web. It gives users the ability to do computational analysis of genomic data [15]. Any researcher with Internet connection can use the analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services that are made accessible via the public Galaxy service. Downloading the Galaxy application and modifying it to address specific requirements allows users to set up their own local Galaxy servers. Galaxy has cultivated a huge community of contributors consisting of both users and programmers. This program retrieves information from the NCBI's Short Read Archive (SRA) and saves it in the FASTQ file format. It derives its functionality from the FASTQ-dump software that is included in the SRA Toolkit [17]. The program will automatically get data for you when you type a single accession number into the box labeled Accession and click the Execute button. It is important to keep the following points under consideration,

- If data is paired-ended (or mate-paired) the tool will generate a single Interleaved dataset, in which forward and reverse mates are alternating.
- If data is single ended, a standard single FASTQ dataset will be produced.

Raw sequence data that is being produced by high throughput sequencing pipelines can be checked for quality using FastQC, which is designed to make the process of doing so as straightforward as possible. This program provides a modular set of analyses, which may be utilized by us to get a rapid picture of whether or not our data has any flaws of which we should be aware before undertaking any additional study. Picard-tools libraries, which are used for SAM/BAM processing, are incorporated into FastQC as well. As the input read file that needs to be checked, FastQC will check a Galaxy FASTQ, fastq.gz, sam, or bam file. Additionally, it will take an optional file that has a list of contaminants' information. This file will be tab-delimited and will have two columns: the name of the contaminant and its sequence. The program additionally accepts a user-defined limits.txt file as an additional configuration option. This file allows the user to establish the warning thresholds for the various modules and also specifies which modules should be included in the output. The FASTQ splitter is applied to the joined paired-end readings as the sample is being paired. It does so by dividing a single FASTA dataset that represents paired-end run into two separate datasets, one for each end of the run. This tool is only useful for datasets in which both ends have the same amount of information. In order to distinguish between the split left-hand and right-hand readings, sequence identifiers will be prefixed with either a /1 or /2.

Assembly

The term "assembly" refers to the process of reconstructing genomes from smaller pieces of DNA, known as "reads," which are produced as a byproduct of a sequencing experiment [20]. The assembly of short read fragments will be performed in order to obtain larger genomic contigs if the research intends to recover the genome of uncultured organisms or obtain full-length CDS for further characterization rather than providing a functional description of the community. This will be the case if the research is being conducted. There are two primary methods that are typically employed when putting together sequences [21,22]. The overlap-based strategy, also known as the traditional overlap layout consensus (OLC) method, the more sophisticated string graph, and the de Bruijn graph approach, which is used more frequently. Both methods make use of a data structure known as a "graph," which represents all of the connections (edges) between the many fundamental sequence elements, such as reads, that are extracted from the sequence dataset. Reads are referred to as nodes. The assembly of lengthy sequencing reads is where overlap-based algorithms shine, however when applied to high throughput short read sequencing data, it was shown that these approaches are computationally too expensive to be practical. On the other hand, techniques based on De Bruijn graphs made it possible to assemble short read data in an effective manner .[23,24]. On galactic Europe, I assembled a unicycle, a spade, and a shovel using the respective assemblers [25].

The Unicycler pipeline is a hybrid assembly method that is used for bacterial genomes. To build accurate and comprehensive assemblies, it employs both the short reads produced by Illumina and the long reads produced by Pac Bio or Nano pore. Unicycler takes inputs in the form of short reads generated by Illumina and saves them in FASTQ

format. Galaxy requires that they be in FASTQ format with Sanger encoding of the quality ratings as an additional requirement. Long readings obtained from either Oxford Nano pore or Pac Bio can be in either the FASTQ or FASTA format. The genomes of individual bacteria are pieced together by Shovill using Illumina's paired-end reads. For Illumina whole genome sequencing data from bacteria and many other tiny organisms, the SPAdes genome assembler has emerged as the de facto standard de novo genome assembler. SPAdes was a significant advancement when compared to earlier assemblers such as Velvet; nonetheless, some of its components can be sluggish, and it has typically been unable to deal with overlapping paired-end readings in an effective manner[26].

SOAPdenovo and SOAPdenovo2 is a component of the Short Oligonucleotide Analysis Kit and served as the primary focus of its development when it was first conceived for single genome assemblies. [27]In order to accommodate a variety of k-mer lengths, two distinct variations of the assembler have been designed. The first version is limited to k-mer lengths that are relatively small, but it has the benefit of consuming less memory. The second version, on the other hand, takes k-mer lengths that are as long as 128 characters. Their method is de Brujin single k-mer and read pair format is interleaved or separate. Multiple libraries and extensive instructions for this tool are also available and are well documented. A wide range of input sequence file format is accepted by it.

IDBA-UD is a collection of distinct de Bruijn graph-based assemblers, each of which is dedicated to performing a certain function[28]. By splitting the assembly graph, this program makes an effort to preserve and rebuild minute differences between sub-strains that are closely related to one another. This assembler has a feature called the multi k-mer assembly technique. This approach iterates through a variety of different k-mer values in an effort to enhance the de Bruijn graph and the assembly that is produced as a result. IDBA-UD is unusual among assemblers that are based on the de Bruijn graph because it supports even values for the k-mer length k, which are typically avoided because of the possibility that palindromes would occur. However, there is no program manual can be found, and the product itself is not well documented. It would have been more useful if it accepted FASTQ, which is the most prevalent sequence file type, however it only accepted FASTA files instead of any other format. Despite that, this instrument had a great deal of success in the assembly industry, which led to its widespread adoption.

Ray Meter The de Bruijn graph-based assembler is known as Meta, and it did not make use of predefined coverage cut-offs. Instead, it examines the k-mer coverage the distribution in the dataset to describe the minimum coverage value (for which the majority of k-mers can still be expected to be correct), and the average coverage value (displayed by the majority of k-meters) individually for each continuous read path within the de Bruijn graph. This is done so that the minimum coverage value and the average coverage value can be compared to one another. The assembler is compatible with mate-paired, paired-end, and single-end libraries that are provided in a wide variety of file and compression formats. In addition, a number of helpful built-in options for downstream analysis are provided, such as the ability to identify reference species and determine their relative abundance. A value that must be supplied in order for compilation and installation to take place is what places a limit on the longest possible k-mer[30].

The only assembly tool in this comparison that is not a de Bruijn graph-based assembler is **Omega**[31]. The other tools are all de Bruijn graph-based. In its place, it makes use of an approach known as overlap-based string graph, which is

often utilized for the construction of long sequencing read data. For efficient and computationally competitive handling of high throughput sequencing datasets, its overlaps are recognized by employing indexed tables of reading prefixes and suffixes of prescribed length. This allows for faster processing. By adopting the string graph approach, it is possible to cope with short read lengths from Illumina and the associated challenges with repeat resolution in a manner that is analogous to that of the de Bruijn graph approach. Only the FASTA and FASTQ file formats are acceptable for input reads[32].

Megahit employs a newly developed data structure, the "succinct de Bruijn graph", which substantially reduces memory requirements. The default cut-off value is 2, so k-mers that occur at least twice are selected while singleton k-meters are discarded. Megahit accepts single and paired-end readings in compressed and uncompressed, FASTA or FASTQ format, and also supports piping input data from standard input. Its usage is simple and well-documented, and issues can be addressed via email forum or the GitHub page for the project. The k-mer figure can be adjusted between 15 and 127 characters in length[34]. **SPAdes** was first intended for single-cell sequencing data, and it was developed to address two key challenges that arise with single-cell sequencing data. These concerns include the unequal read coverage of amplified DNA and the requirement to detect and resolve chimera sequences. SPAdes was developed to address these issues. But while being based on de Bruijn graphs, it has a significant memory consumption. By default, Bayes Hammer is used to correct Illumina reads before they are assembled, and an iterative multi-k-mer method that is very similar to IDBA-UD is utilized. Unlike most assemblers, which implement paired-end information for simplification steps following standard de Bruijn graph construction, SPAdes directly incorporate this data in the graph by using k-bimers, which are sets of k-mers taken from reading pairs and separated by an estimated distance value. In contrast, most assemblers implement paired-end information for simplification steps after standard de Bruijn graph construction. Afterwards, SPAdes will iteratively correct and alter the distance estimation of each k-bimer. This will allow for the non-uniform insert length distribution that is present in the majority of shotgun sequencing libraries to be taken into consideration. The k-mer range for the iterative creation of a de Bruijn graph is computed automatically based on the reading duration and the sequence data type. However, the k-mer range can also be defined directly, utilizing k-mers with a length of up to 128 base pairs. It is able to process a wide variety of data types and formats in either their compressed or uncompressed forms[35].

Pipelines A great number of the aforementioned tools have been combined into freely accessible pipelines, which integrate assembly with initial read processing or later analysis processes. These pipelines are available to the public. MetAMOS[39] is a modular framework for metagenome assembly, analysis, and validation. It is one of these pipelines that is among the most helpful and adaptable of these pipelines. MOCAT is a different pipeline that can handle quality trimming, decontamination, assembly, assembly revision, and gene prediction. However, in comparison to MetAMOS[40], MOCAT has significantly less flexibility. It was released before SOAPdenovo2 was published, and it contains SOAPdenovo versions 1.05 and 1.06 as integral assembly components, making use of optimized parameters and enhanced error correction, in addition to scaffolding phases.

The "divide and conquer" strategy was utilized in the creation of the **SLICEMBLER**[41] pipeline, which was designed for ultra-deep sequencing datasets. The read dataset is then partitioned into subgroups of equal size, each of which is assembled independently using an assembler of the user's choosing. After that, the components are coupled with one another and rebuilt in an iterative manner. Both the well-known **IMG/M**[42] and **MG-RAST**[43] metagenome sequence databases are equipped with integrated pipelines for the analysis of metagenomes. There is also a great deal more pipelines available, and continuous construction is being done on new ones. However, a fundamental understanding of the Unix bash scripting language is all that is required to set up personal pipelines, which not only ensures complete control over the entire workflow but also makes it possible to cater to individual requirements and answer certain biological concerns.

The process of connecting a certain sequence with an organism is referred to as "binning." For the purpose of binning, we made use of the PATRIC metagenomic binning service. The Pathosystem Resource Integration Centre is an all-bacterial Bioinformatics Resource Centre (BRC) (http://www.patricbrc.org). This information may be found on their website. PATRIC was initially responsible for storing and combining data on eight distinct bacterial and viral pathogen families. This allowed PATRIC to provide researchers with thousands of reliably annotated bacterial genomes, integrate the accompanying omics data, and provide a site with analysis tools to promote research into infectious diseases. PATRIC uses automated scripts for the gathering and assimilation of data in order to obtain genomes from GenBank and RefSeq on a monthly basis. Either reads or contigs are obtained by the Metagenomic Binning Service, which then makes an effort to "bin" the data into a collection of genomes. The bacterial and archaeal genomes contained in environmental samples can be reconstructed with the help of this service. The PATRIC workspace is where the input to the binning service needs to be located. Either paired reads of assembly from us, or assembled contigs from another source, are acceptable alternatives[44].

Phylogenetic analysis

Phylogenetic analysis is performed to find the evolutionary relationship among species. First of all, Multiple Sequence alignment is performed using MEGAX(Molecular Evolutionary genetics analysis) software [49] and then the phylogenetic tree is also constructed in it. Then iTOL is used for tree annotation. Another method for this purpose is to do “Taxonomic Classification” by the assembled files or from accession number of SRA, which also includes the phylogenetic tree of the species and Taxonomy.

Results

Description of the sample

We analyzed our sample from the southwest Cameroon rural area. The common age of the player is 50-70 years. This pattern therefore no longer handiest having the subsistence mode however also have the genetic records. We choose this area and pattern from there's because the preceding paintings are completed to many years ago, based totally on the dietary questionnaires and isotopes evaluation, showed they have wonderful dites. There are different other

samples are present but we get one of them to study its taxonomy and other features of it. We get the sample raw reads from the SRA database in the NCBI, where these data set are present and many other volunteers use this in their studies. The data which we get from SRA database its accession number is ERR2619756.

Quality Check and removal of low-quality reads

We upload our records at the galaxy server and follow a few tools on it. We take a look at the excellent of our reads record by the use of "FASTQC" (Figure 1). This tool gets the file in the form of paired reads. In the output, it shows the graph approximately the fine of your study document. The quality graph of our data shown below in the figure. Then we use "SICKLE" use to remove the low-quality reads from our data set. After that, we use "SPLITTER" to separate our forward and reverse data file set. It gives us a way to easily manage our obtained sequence file from this tool. This tool also made two separate files of our data set so we can easily manage it well.

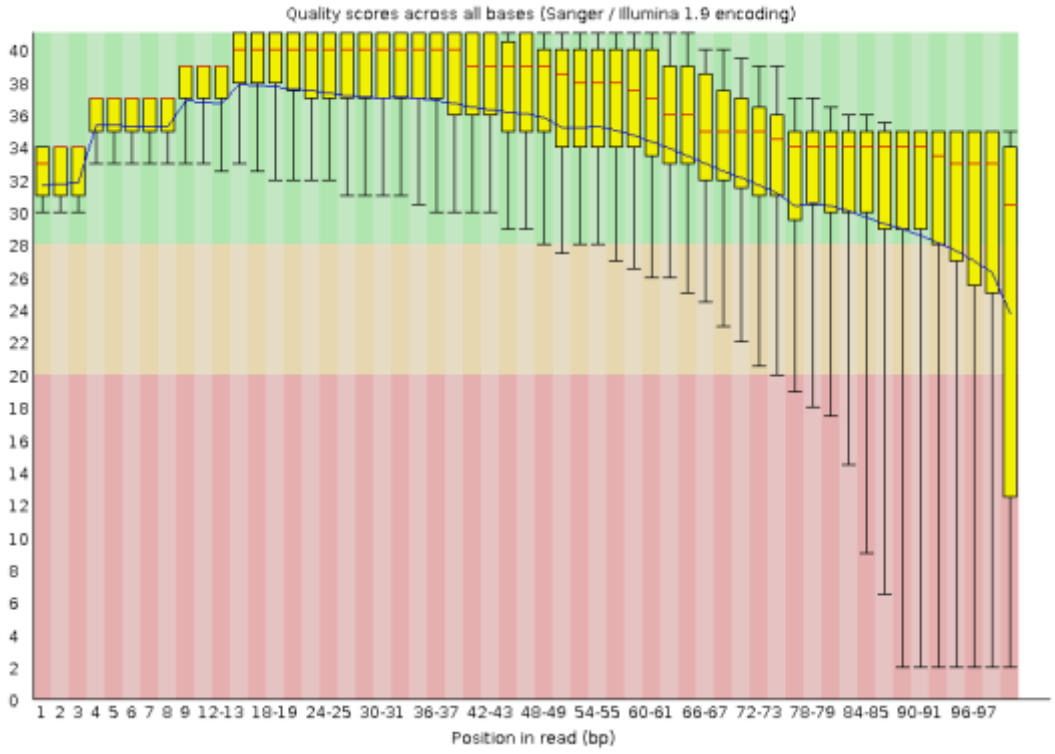

Figure 1: FASTQC quality checking graph of the sample

Assembly

It is the next step closer to our result. In this, there are specific tools on galaxy Europe to carry out assembly of your terrific reads. We deliver separate reads report to any of the tools and it gives us one output report with the aid of combing these two reads report. We used "SPADES" assembler and locate it good amongst all the different assembly tools. You also performed assembly in "PATRIC". PATRIC is better than "SPADEs" for short length reads. The resultant file of the assembly gives us the contigs.

Binning

After assembled our files with Galaxy Europe download and upload on the PATRIC for the binning purpose. There are some parameters which we follow while performing the binning of our data.

- Completeness >=80%
- Fine consistency >=87%
- Contamination =<10%
- A single PheS protein

In the result of binning we get a different number of bins for a single file, there may be bad bins (Table 2) or good bins (Table 1).

Table 1: Good bin extracted from the data

Score	Genome ID	Genome Name	Reference Genome	Coarse consistency (%)	Fine consistency (%)	Completeness (%)	Contamination (%)	Contig count	DNA size (bp)	Contigs N50 (bp)	Mean Coverage
1928	2053619.53	Succinivibrio sp. strain clonal population	2053619.3	91.4	90.7	100	1.2	183	2096965	18695	12.34

Bad bins or genomes are in data which didn't meet the requirements of the server. We found 11 bad in the sample mean the 11 genomes which exist in our data but the binning parameters didn't fulfil.

Table 2: *Bad bin extracted from the sample*

Score	Genome ID	Genome Name	Reference Genome	Coarse consistency (%)	Fine consistency (%)	Completeness (%)	Contamination (%)	Contig count	DNA size (bp)	Contigs N50 (bp)	Mean Coverage
1291	1262929.313	Prevotella sp. CAG:520 clonal population	1262929.3	96.7	93.4	92.3	13	301	3456861	29748	24.95

1220	2049022.261	Catenibacterium sp. strain clonal population	2049022.3	91.8	90.9	43.5	4.1	200	868761	5973	6.66
1220	1262930.126	Prevotella sp. CAG:5226 clonal population	1262930.3	97.9	93.6	100	16	361	3365777	22196	37.13
489	2079531.161	Prevotella sp. Marseille-P4119 clonal population	2079531.3	94.0	87.1	89.5	27.1	847	4417521	12866	17.74
469	2022527.433	Prevotella sp. 885 clonal population	2022527.3	97.9	90.0	97.8	29.8	591	4708192	20200	15.21
156	360807.857	Roseburia inulinivorans clonal population	360807.4	95.6	85.7	92.2	34	639	4215783	14203	6.8
145	626940.990	Phascolarctobacterium succinatutens clonal population	626940.20	92.1	86.9	75.8	31.2	666	1702317	4247	7.09
76	1262925.262	Prevotella sp. CAG:386 clonal population	1262925.3	96.0	85.4	90.1	35.1	699	3791376	12616	18.97
-78	2292365.102	Prevotella sp. TF12-30 clonal population	2292365.3	88.2	87.3	34	27.4	795	2957822	8619	65.52
-645	2292054.53	Prevotella sp. AM23-5 clonal population	2292054.3	91.0	90.2	15.1	35.6	493	1548121	7231	59.34

-21 08	165179.2239	Prevotella copri clonal population	1651739.39 1651739.408	97.0	68.1	97.4	76.5	5176	13275232	5636	54.66

Succinivibrio

Succinivibrio is rod-shaped motile organisms with polar flagella. They have a curved spiral shape. *Succinivibrio* is anaerobic bacteria that ferment glucose and they obtain nitrogen through ammonia. *Succinivibrio* has been shown to cause disease, but they are rarely pathogenic in humans.

Table 3: Feature view of Succinivibrio from PATRIC

Genome ID	Accession	Feature ID	Feature Type	Start	End	AA Length	Product
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.1947.2471.fwd	CDS	1947	2471	174	Pyruvate:ferredoxin oxidoreductase, gamma subunit (EC 1.2.7.1)
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.2481.2768.fwd	CDS	2481	2768	95	Pyruvate:ferredoxin oxidoreductase, delta subunit (EC 1.2.7.1)
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.2768.3928.fwd	CDS	2768	3928	386	Pyruvate:ferredoxin oxidoreductase, alpha subunit (EC 1.2.7.1)
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.3931.4818.fwd	CDS	3931	4818	295	Pyruvate:ferredoxin oxidoreductase, beta subunit (EC 1.2.7.1)
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.4827.6158.fwd	CDS	4827	6158	443	Coenzyme A ligase
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.6265.7818.rev	CDS	6265	7818	517	Na+/H+ antiporter

				65	18		
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.648.1583.fwd	CDS	648	1583	311	Methionine biosynthesis and transport regulator MtaR, LysR family
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.7925.8905.rev	CDS	7925	8905	326	hypothetical protein
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.90.527.rev	CDS	90	527	145	Iron-sulfur cluster regulator IscR
2053619	QAMS01000001	PATRIC.2053619.3.QAMS01000001.CDS.9343.9561.fwd	CDS	9343	9561	73	hypothetical protein

Prevotella

Prevotella is a genus of Gram-negative bacteria. *Prevotella* spp. are members of the oral, vaginal, and gut microbiota and are often recovered from anaerobic infections of the respiratory tract. These infections include aspiration pneumonia, lung abscess, pulmonary empyema, and chronic otitis media and sinusitis.

Table 4: Feature view of Prevotella

Genome	Feature Type	Start	End	Length	Strand	FIGfam ID	AA Length	Product
Prevotella sp. CAG:520	CDS	1620	2066	447	-		148	hypothetical protein
Prevotella sp. CAG:520	CDS	450	1598	1149	-		382	hypothetical protein
Prevotella sp. CAG:520	CDS	10010	10291	282	-	FIG001874 1	93	ATP-dependent Clp protease adaptor protein ClpS

Prevotella sp. CAG:520	CDS	10393	11430	1038	-	FIG009378666	345	hypothetical protein
Prevotella sp. CAG:520	CDS	11804	12862	1059	-	FIG006382834	352	hypothetical protein
Prevotella sp. CAG:520	CDS	1254	2714	1461	-	FIG000017500	486	Carbon starvation protein A
Prevotella sp. CAG:520	CDS	13201	13929	729	-	FIG012285117	242	hypothetical protein
Prevotella sp. CAG:520	CDS	2880	3200	321	+	FIG000096377	106	PaaD-like protein (DUF59) involved in Fe-S cluster assembly
Prevotella sp. CAG:520	CDS	3311	4075	765	+	FIG000009433	254	UDP-2,3-diacylglucosamine diphosphatase (EC 3.6.1.54)
Prevotella sp. CAG:520	CDS	4233	5318	1086	-	FIG000000555	361	Tryptophanyl-tRNA synthetase (EC 6.1.1.2)
Prevotella sp. CAG:520	CDS	461	1177	717	+	FIG000374331	238	UPF0758 family protein
Prevotella sp. CAG:520	CDS	5451	6902	1452	-	FIG001656666	483	Arylsulfatase (EC 3.1.6.1)
Prevotella sp. CAG:520	CDS	6889	7086	198	-		65	hypothetical protein
Prevotella sp. CAG:520	CDS	7134	7784	651	-	FIG000007488	216	Leucyl/phenylalanyl-tRNA--protein transferase (EC 2.3.2.6)
Prevotella sp. CAG:520	CDS	7796	10006	2211	-	FIG001400 0	736	ATP-dependent Clp protease ATP-binding subunit ClpA

Prevotella sp. CAG:520	CDS	1	87	87	-	29	hypothetical protein

Catenibacterium

Catenibacterium is a Gram-positive, non-spore forming and anaerobic genus from the family *Erysipelotrichidae,* with one known species. It is a Gram-positive, non-spore forming and anaerobic genus from the family *Erysipelotrichidae,* with one known species.

Table 5: Feature view of Catenibacterium

Genome	Feature Type	Start	End	Strand	AA Length	Product
Catenibacterium sp. strain UBA9492	CDS	2356	2577	+	73	hypothetical protein
Catenibacterium sp. strain UBA9492	CDS	1342	1563	+	73	Protein translocase membrane subunit SecG
Catenibacterium sp. strain UBA9492	CDS	1615	3744	+	709	3'-to-5' exoribonuclease RNase R
Catenibacterium sp. strain UBA9492	CDS	3750	4196	+	148	tmRNA-binding protein SmpB
Catenibacterium sp. strain UBA9492	CDS	1	1209	+	402	Hydrolase (HAD superfamily) in cluster with DUF1447

Catenibacterium sp. strain UBA9492	CDS	1257	2123	+	288	Permease of the drug/metabolite transporter (DMT) superfamily
Catenibacterium sp. strain UBA9492	CDS	2142	3485	-	447	hypothetical protein
Catenibacterium sp. strain UBA9492	CDS	3625	4848	+	407	hypothetical protein

Roseburia inulinivorans

Roseburia inulinivorans is a bacterium first isolated from human faeces. It is anaerobic, gram-positive or gram-variable, slightly curved rod-shaped and motile. The cell range in size from 0.5-1.5 to 5.0 micrometers. A2-194 is the type of strains.

Table 6: Feature view of Roseburia inulinivorans

Genome	Accession	Feature Type	Start	End	Strand	AA Length	Product
Roseburia inulinivorans	CYXX01000001	CDS	100555	101769	-	404	Isocitrate dehydrogenase [NADP] (EC 1.1.1.42)
Roseburia inulinivorans	CYXX01000001	CDS	101815	102477	-	220	Transcriptional regulator, GntR family
Roseburia inulinivorans	CYXX01000001	CDS	102480	103943	-	487	Re face-specific citrate synthase (EC 2.3.3.3)
Roseburia inulinivorans	CYXX01000001	CDS	107898	109988	-	696	Elongation factor G-like protein TM_1651
Roseburia inulinivorans	CYXX01000001	CDS	110376	111470	+	364	Prephenate dehydrogenase (EC 1.3.1.12)
Roseburia inulinivorans	CYXX01000001	CDS	111500	112456	-	318	Fructokinase (EC 2.7.1.4)

Roseburia inulinivorans	CYXX01 000001	CDS	112 518	114 035	-	505	beta-fructofuranosidase (EC 3.2.1.26)
Roseburia inulinivorans	CYXX01 000001	CDS	114 052	115 722	-	556	Sucrose ABC transporter, substrate-binding protein
Roseburia inulinivorans	CYXX01 000001	CDS	115 39	125 22	+	327	LSU rRNA pseudouridine(955/2504/2580) synthase (EC 5.4.99.24)
Roseburia inulinivorans	CYXX01 000001	CDS	115 741	116 637	-	298	Sucrose ABC transporter, permease protein 2
Roseburia inulinivorans	CYXX01 000001	CDS	116 650	117 588	-	312	Sucrose ABC transporter, permease protein 1
Roseburia inulinivorans	CYXX01 000001	CDS	117 845	118 870	+	341	Sucrose operon repressor ScrR, LacI family

Phylogenetic tree Construction

Phylogenetic analysis is for checking the relatedness among the different biological species. The tree graphically represents the evolutionary relationship among the different species and organisms. For the taxonomy classification, we construct a phylogenetic tree of our SRA raw-reads data set.

Here in the conclusion of our report, we have investigated the taxonomical and functional diversity (Figure 2), as well as to unveil the interactions between the eukaryotic and prokaryotic components of the human gut microbiota. We take previously records set from SRA database. Then upload it in galaxy Europe to perform exclusive tools in step with our requirements. First test the satisfactory of our records set by the usage of FASTQC. We set the parameters of this device and then add our file there. In the output, it offers us two different documents. Then we use tool for trimming of our statistics to separate paired give up files. We apply this tool to the first file of our data set. For this, we use splitter. Then appearing the assembly of our resultant file with any of the given equipment in the galaxy Europe.

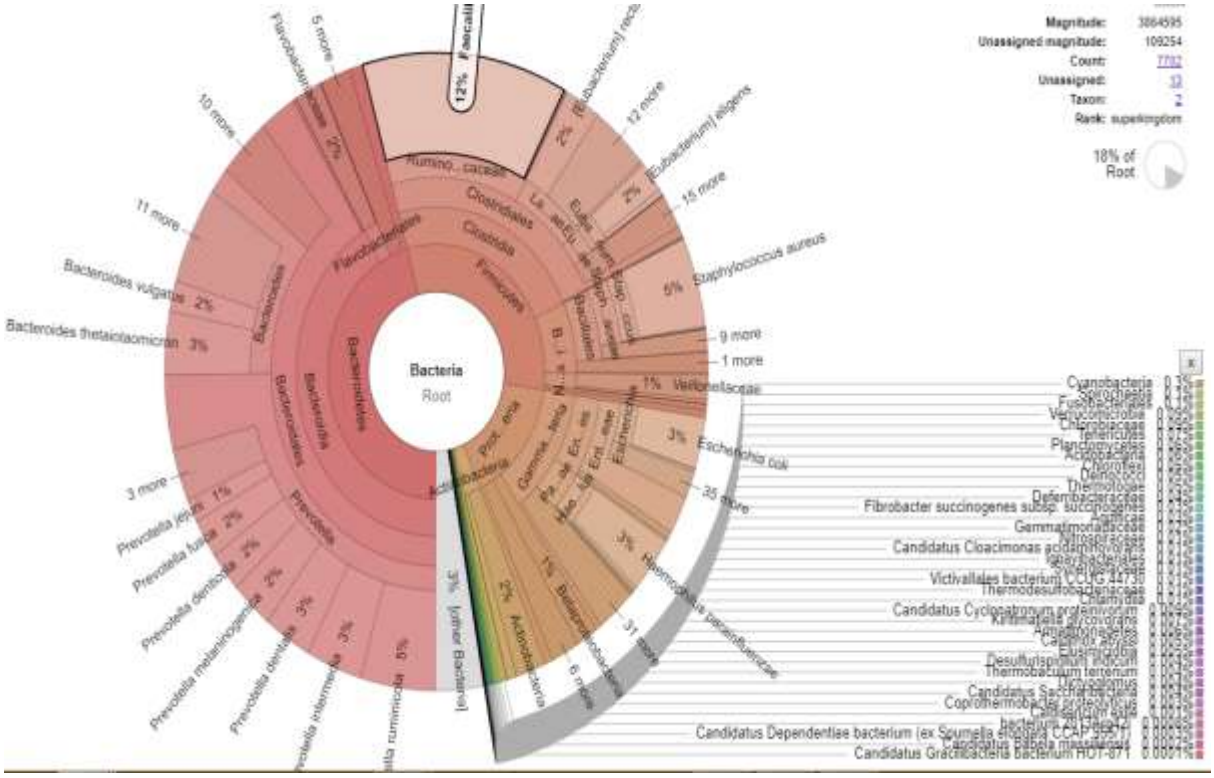

Figure 2: Phylogenetic tree of the sample

For assembly we also set the parameters and give the reads record to the assembler and in end result it deliver us the contigs record. Simply down load this contigs and record and upload it on PATRCI for metagenomic binning. When we get effects from this report, we interpret our end result by way of the documents and the defaults parameters of this equipment. In the PATRIC, we also make the phylogenetic tree of our data set. This show the graphical representation of our genome and its sub divisions. And at last we interpret our data using the default parameters of all the tools and then write the result of our report

References

1. Schopf JW, Packer BM. Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. Science 1987;237(4810):70-73.

2. Isenberg H, Painter B. Indigenous and pathogenic microorganisms of humans. Indigenous and pathogenic microorganisms of humans 1980:25-39.

3. Xu J. Invited review: microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Molecular ecology 2006;15(7):1713-1731.

4. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 2004;68(4):669-685. (In eng). DOI: 10.1128/MMBR.68.4.669-685.2004.

5. Council NR. The new science of metagenomics: revealing the secrets of our microbial planet: National Academies Press, 2007.

6. David LA, Maurice CF, Carmody RN, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 2013;505:559. DOI: 10.1038/nature12820

7. Goodrich Julia K, Waters Jillian L, Poole Angela C, et al. Human Genetics Shape the Gut Microbiome. Cell 2014;159(4):789-799. DOI: https://doi.org/10.1016/j.cell.2014.09.053.

8. Blekhman R, Goodrich JK, Huang K, et al. Host genetic variation impacts microbiome composition across human body sites. Genome Biology 2015;16(1):191. DOI: 10.1186/s13059-015-0759-1.

9. Elliott DE, Summers RW, Weinstock JV. Helminths as governors of immune-mediated inflammation. International Journal for Parasitology 2007;37(5):457-464. DOI: https://doi.org/10.1016/j.ijpara.2006.12.009.

10. Kay GL, Millard A, Sergeant MJ, et al. Differences in the Faecal Microbiome in Schistosoma haematobium Infected Children vs. Uninfected Children. PLOS Neglected Tropical Diseases 2015;9(6):e0003861. DOI: 10.1371/journal.pntd.0003861.

11. Costello EK, Stagaman K, Dethlefsen L, Bohannan BJM, Relman DA. The Application of Ecological Theory Toward an Understanding of the Human Microbiome. Science 2012;336(6086):1255. DOI: 10.1126/science.1224203.

12. Fumagalli M, Pozzoli U, Cagliani R, et al. Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. The Journal of Experimental Medicine 2009;206(6):1395. DOI: 10.1084/jem.20082779.

13. Buffie CG, Pamer EG. Microbiota-mediated colonization resistance against intestinal pathogens. Nature Reviews Immunology 2013;13:790. (Review Article). DOI: 10.1038/nri3535.

14. Hayes KS, Bancroft AJ, Goldrick M, Portsmouth C, Roberts IS, Grencis RK. Exploitation of the Intestinal Microflora by the Parasitic Nematode <em>Trichuris muris</em&gt. Science 2010;328(5984):1391. DOI: 10.1126/science.1187703.

15. Goecks J, Nekrutenko A, Taylor J, The Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 2010;11(8):R86. DOI: 10.1186/gb-2010-11-8-r86.

16. Blankenberg D, Taylor J, Schenck I, et al. A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. 2007;17(6):960-964.

17. Leinonen R, Sugawara H, Shumway M, Collaboration obotINSD. The Sequence Read Archive. Nucleic Acids Research 2010;39(suppl_1):D19-D21. DOI: 10.1093/nar/gkq1019 %J Nucleic Acids Research.

18. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data2014.

19. Blankenberg D, Gordon A, Von Kuster G, et al. Manipulation of FASTQ data with Galaxy. Bioinformatics 2010;26(14):1783-1785. DOI: 10.1093/bioinformatics/btq281 %J Bioinformatics.

20. Reich JG, Drabsch H, Däumler A. On the statistical assessment of similarities in DNA sequences. Nucleic acids research 1984;12(13):5529-5543. (In eng). DOI: 10.1093/nar/12.13.5529.

21. Li Z, Chen Y, Mu D, et al. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. 2012;11(1):25-37.

22. Nagarajan N, Pop MJNRG. Sequence assembly demystified. 2013;14(3):157.

23. Chaisson MJ, Pevzner PAJGr. Short read fragment assembly of bacterial genomes. 2008;18(2):324-330.

24. Pevzner PA, Tang H, Waterman MS. A new approach to fragment assembly in DNA sequencing. Proceedings of the fifth annual international conference on Computational biology: ACM; 2001:256-267.

25. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Current protocols in bioinformatics 2010;Chapter 11:Unit-11.5. (In eng). DOI: 10.1002/0471250953.bi1105s31.

26. Bankevich A, Nurk S, Antipov D, et al. ***"SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing"***. Journal of computational biology : a journal of computational molecular cell biology 2012;19(5):455-477. DOI: 10.1089/cmb.2012.0021.

27. Li R, Zhu H, Ruan J, et al. De novo assembly of human genomes with massively parallel short read sequencing. 2010;20(2):265-272.

28. Peng Y, Leung HC, Yiu S-M, Chin FYJB. Meta-IDBA: a de Novo assembler for metagenomic data. 2011;27(13):i94-i101.

29. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil JJGb. Ray Meta: scalable de novo metagenome assembly and profiling. 2012;13(12):R122.

30. Boisvert S, Laviolette F, Corbeil JJJocb. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. 2010;17(11):1519-1533.

31. Haider B, Ahn T-H, Bushnell B, Chai J, Copeland A, Pan CJB. Omega: an overlap-graph de novo assembler for metagenomics. 2014;30(19):2717-2722.

32. Myers EWJB. The fragment assembly string graph. 2005;21(suppl_2):ii79-ii85.

33. Simpson JT, Durbin RJGr. Efficient de novo assembly of large genomes using compressed data structures. 2012;22(3):549-556.

34. Li D, Liu C-M, Luo R, Sadakane K, Lam T-WJB. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. 2015;31(10):1674-1676.

35. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. 2012;19(5):455-477.

36. Rinke C, Schwientek P, Sczyrba A, et al. Insights into the phylogeny and coding potential of microbial dark matter. 2013;499(7459):431.

37. Kaster A-K, Mayer-Blackwell K, Pasarelli B, Spormann AMJTIj. Single cell genomic study of Dehalococcoidetes species from deep-sea sediments of the Peruvian Margin. 2014;8(9):1831.

38. Nikolenko SI, Korobeynikov AI, Alekseyev MA. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC genomics: BioMed Central; 2013:S7.

39. Treangen TJ, Koren S, Sommer DD, et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. 2013;14(1):R2.

40. Kultima JR, Sunagawa S, Li J, et al. MOCAT: a metagenomics assembly and gene prediction toolkit. 2012;7(10):e47656.

41. Mirebrahim SH. Efficient Methods for Analysis of Ultra-Deep Sequencing Data. UC Riverside; 2015.

42. Markowitz VM, Ivanova NN, Szeto E, et al. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Research 2007;36(suppl_1):D534-D538. DOI: 10.1093/nar/gkm869 %J Nucleic Acids Research.

43. Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer FJCSHP. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. 2010;2010(1):pdb. prot5368.

44. Wattam AR, Abraham D, Dalay O, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic acids research 2014;42(Database issue):D581-D591. (In eng). DOI: 10.1093/nar/gkt1099.

45. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EWJNar. GenBank. 2011;40(D1):D48-D53.

46. Pruitt KD, Tatusova T, Brown GR, Maglott DRJNar. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. 2011;40(D1):D130-D135.

47. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microbial informatics and experimentation 2012;2(1):3-3. (In eng). DOI: 10.1186/2042-5783-2-3.

48. Brettin T, Davis JJ, Disz T, et al. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. 2015;5:8365.

49. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution 2018;35(6):1547-1549. DOI: 10.1093/molbev/msy096 %J Molecular Biology and Evolution.