IMPLEMENTATION OF KRONA INTO QIIME 2

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

KAAN BÜYÜKALTAY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF BIOINFORMATICS

MAY 2022

IMPLEMENTATION OF KRONA INTO QIIME 2

Submitted by Kaan Büyükaltay in partial fulfillment of the requirements for the degree of **Master of Science in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son
Supervisor, **Health Informatics Dept., METU**

Examining Committee Members:

Asst. Prof. Dr. Aybar Can Acar
Health Informatics Dept., METU

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics Dept., METU

Assoc. Prof. Dr. Tunca Doğan
Computer Engineering Dept., Hacettepe University

Date: *May 9, 2022*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Kaan Büyükaltay

Signature : __________________

ABSTRACT

IMPLEMENTATION OF KRONA INTO QIIME 2

Büyükaltay, Kaan
MSc., Department of Health Informatics
Supervisor: Assoc. Prof. Yeşim Aydın Son

May 2022, 48 pages

Microbiome studies mainly use two different approaches: metagenomic and amplicon analysis. The metagenomic analysis is based on a whole-genome shotgun approach. It is possible to acquire taxonomic resolution to species and strain level but requires high computational power and is expensive. In the amplicon analysis approach, marker genes like 16S, 18S, and ITS are used to identify the sequences. These genes are highly conserved but contain hypervariable regions, allowing researchers to acquire taxonomic resolution at the genus level. The downside of this approach is due to the biases introduced with PCR and its high error rate. QIIME 2 is a tool consisting of many plugins for amplicon analysis. QIIME 2 team has a decentralized development approach so that third parties can develop new plugins and share them with the community. Although there are different options for visualizations in QIIME 2, there is no tool for visualizing the contents of the features table for a single sample. On the other hand, Krona is a tool that visualizes hierarchical data as multi-layered pie charts. While this visualization is desired, it is not a straightforward task, as the QIIME 2 and Krona do not share the same data formats. This study aims to fill this gap by developing a plugin that will transfer QIIME2 outputs into Krona to ease the visualization of QIIME 2 analysis results with Krona without additional script or manipulation during amplicon analysis.

Keywords: metagenome, amplicon, plugin development

ÖZ

KRONA’NIN QIIME 2 İÇİN UYARLANMASI

Büyükaltay, Kaan

Yüksek Lisans, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Doç. Dr. Yeşim Aydın Son

Mayıs 2022, 48 sayfa

Mikrobiyom çalışmaları temel olarak iki farklı yaklaşım kullanır: metagenomik veya amplicon analizi. Metagenomik analiz tüm genom shotgun yöntemi üzerine temellendirilmiştir. Türlere ve suşlara ulaşabilecek taksonomik çözünürlük elde etmek mümkündür; fakat yüksek işlem gücü gerektirmektedir ve maliyetlidir. Amplikon analizinde 16S, 18S, ITS gibi marker genler kullanılarak sekanslar tanımlanmaktadır. Bu genler yüksek oranda korunumludur; fakat aşırı değişken bölgeler içermektedir, bu sayede araştırmacılar cins seviyesine kadar taksonomik çözünürlük elde edebilmektedir. Bu yaklaşımın dezavantajı PCR sapması sebebiyle yüksek hata oranı olmasıdır. QIIME 2 amplikon analizi için birçok eklentiden oluşan bir araçtır. QIIME 2 ekibi, üçüncü tarafların yeni eklentiler geliştirebilmesi ve bunları topluluk ile paylaşabilmesi için merkezi olmayan bir geliştirme yaklaşımına sahiptir. QIIME 2’de görselleştirmeler için farklı seçenekler bulunsa da tek bir örnek için özellikler tablosunun içeriğini görselleştirmeye yönelik bir araç bulunmamaktadır. Öte yandan Krona, hiyerarşik verileri çok katmanlı pasta grafikleri olarak görselleştirebilen bir araçtır. Bu görselleştirme talep ediliyor olsa da QIIME 2 ve Krona aynı veri formatlarını paylaşmadığı için bu basit bir görev değildir. Çalışmada amplikon analizi sırasında ek bir komut dosyası veya manipülasyon olmadan QIIME 2 ile elde edilen analiz sonuçlarının kolayca görselleştirilebilmesi için QIIME 2 çıktılarını Krona’ya aktaracak bir eklentinin geliştirilmesi hedeflenmiştir.

Anahtar Sözcükler: metagenom, amplikon, eklenti geliştirme

To My Family and Friends

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Assoc. Prof. Dr. Yeşim Aydın Son, for her guidance, encouragement, and moral support.

Besides my supervisor, I would like to thank all of QIIME 2 developers, especially Prof. Dr. Nicholas Bokulich and Matthew Dillon, for their support on the forum page, trying their best to answer all of the community's questions.

I would also like to thank all of my colleagues from BM Labosis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
QIIME	Quantitative Insights Into Microbial Ecology
QIIME 2	Quantitative Insights Into Microbial Ecology 2
DNA	Deoxyribonucleic Acid
Taq	Thermus aquaticus
E. coli	Escherichia coli
ddNTPs	Dideoxynucleotides
ATP	Adenosine Triphosphate
HTS	High-throughput Sequencing
RT-bases	Reversible terminator bases
SMS	Single-Molecule Sequencing
SMRT	Single-Molecule Real-Time Sequencing
ZMW	Zero-Mode Waveguide
ONT	Oxford Nanopore Technologies
SSU	Small-subunit
ITS	Internal Transcribed Spacer
OTU	Operational Taxonomic Unit
ASV	Amplicon Sequence Variant
qza	QIIME zipped artifact
qzv	QIIME zipped visualization
BIOM	Biological Observation Matrix
PD	Parkinson’s Disease
αSyn	Alpha-synuclein
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive

CHAPTER 1

INTRODUCTION

1.1 Motivation

The number of microbiome search results is increasing rapidly, as shown in Figure 1. They are considered a "new organ" in the human body (Baquero & Nombela, 2012). More research is needed to show the relationship between microbiomes and certain serious diseases (Gilbert et al., 2018).

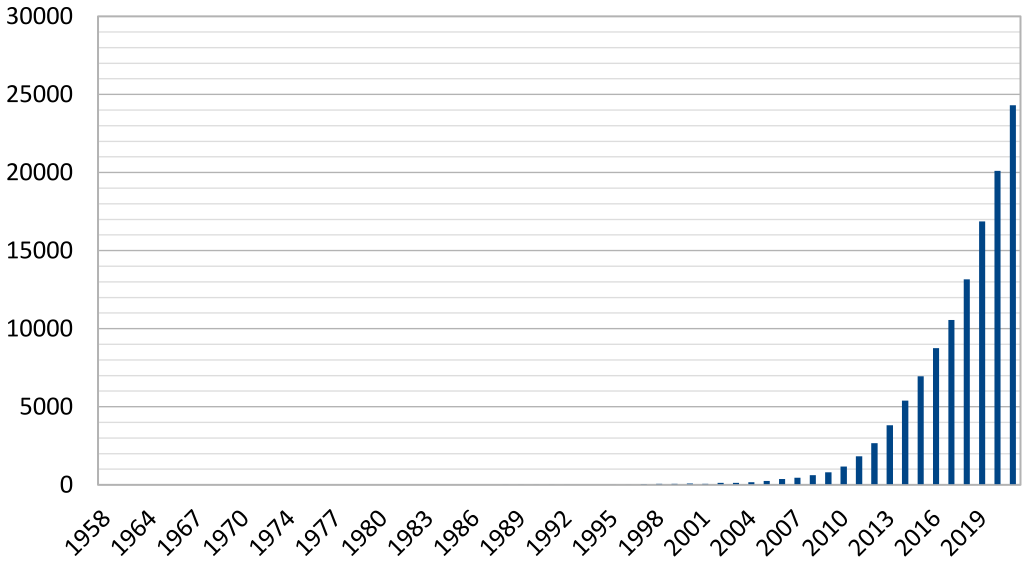


Figure 1: "microbiome" search results on PubMed by year

As the interest in microbiomes increases, the need for a better understanding of microbiomes emerges where different approaches to analysis and related tools come into the scene.

The high throughput DNA sequencing methodology (next-generation sequencing; NGS) has recently evolved like the microbiome. NGS allows the generation of high-dimensional biological data, which is vital for microbiome analysis. Microbiomes may contain millions of microbes. Advances in NGS made obtaining large amounts of genomic sequences easier and faster to be analyzed deeper (Jünemann et al., 2017). Also, the improvements led to better curated, publicly available databases, so many more systematic and efficient analysis methods can be developed.

There are two main genetic approaches to next-generation sequencing (NGS) for microbiome analysis, metagenomics and amplicon analysis (Liu et al., 2021).

Metagenomics is the technique or set of techniques whose main objective is to determine the microbial population in a determined environment studied in the context of its community. Shotgun metagenome can provide functional gene profiles directly and a taxonomic resolution to species or strain level. On the other hand, the shotgun metagenome is very expensive and time-consuming for analysis since it can generate vast amounts of data (Garrido-Cardenas & Manzano-Agugliaro, 2017).

Amplicon sequencing is the most widely used NGS method for microbiome analysis, and it can be applied to almost all sample types (Liu et al., 2021). Marker genes are used for amplification of the specified region of genomes. This method is quick and requires low biomass. However, PCR and primer biases affect the reliability, and taxonomic resolution is limited to the genus level (Liu et al., 2021).

QIIME 2 is a tool consisting of many plugins for amplicon analysis. QIIME 2 team has a decentralized development approach, so third parties can develop new plugins and share them with the community (Bolyen et al., 2019). Krona is a tool that visualizes hierarchical data as multi-layered pie charts (Ondov et al., 2011).

The motivation of this study was to develop a plugin for QIIME 2 that implements Krona so that it is possible to use them without the need for any other script or data structure manipulation.

CHAPTER 2

BACKGROUND

2.1 Microbiome

Microbes are simple organisms. They have no nucleus or bilayer membraned organelles and only present primary pathways. They are used as models for more complex mechanisms found in other organisms. Microbes help us answer questions about all organisms, especially with genetic research. The simplicity of their structure, especially having no nucleus, allows easy application of specific approaches to their genome (Srivastava, 2014).

Although there are many different definitions for the microbiome, and new ones are still proposed, the most cited definition describes microbiomes as a community of commensal, symbiotic, and pathogenic microorganisms within a body or other environment (Berg et al., 2020). Microbiomes are much more than individuals that are embodied within. We can think of microbes as "words," and microbiomes are "compositions." Their importance comes from the interactions between them.

2.2 Polymerase Chain Reaction

The double helix structure of DNA was discovered in the 1950s; however, it was impossible to amplify DNA sequences until the 1980s. The chemist Kary Mullis first devised the polymerase Chain Reaction (PCR) technique (Zhu et al., 2020).

PCR is based on the biological characteristic of DNA replication. Mullis aimed to use two primers that can bind to 5’ and 3’ ends, and with the help of the DNA polymerase enzyme, the sequence between the primers would be reproduced. Also, this could be done repeatedly to obtain the desired amount of amplified DNA from a minimal amount of original DNA (Zhu et al., 2020). The basic schema of PCR is shown in Figure 2.

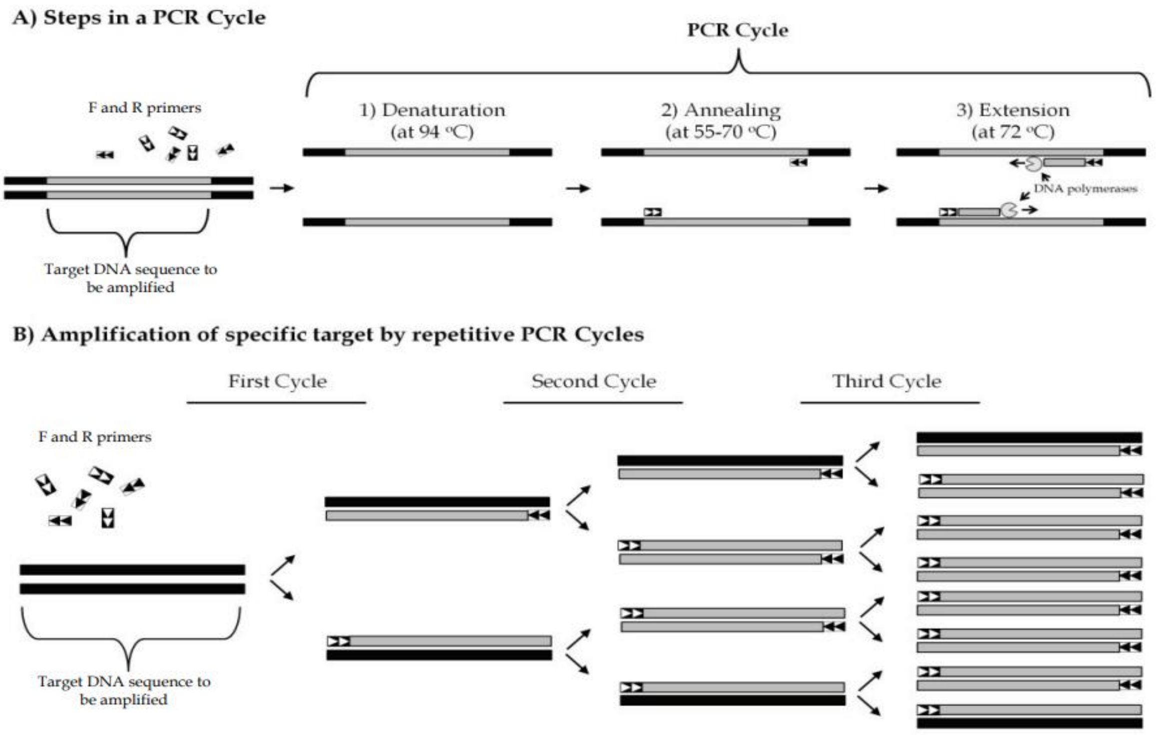

A) Steps in a PCR Cycle
PCR Cycle
F and R primers
1) Denaturation (at 94 °C)
2) Annealing (at 55-70 °C)
3) Extension (at 72 °C)
DNA polymerases
Target DNA sequence to be amplified
B) Amplification of specific target by repetitive PCR Cycles
First Cycle
Second Cycle
Third Cycle
F and R primers
Target DNA sequence to be amplified

Figure 2: Schematic representation of PCR. A: Steps in a PCR cycle. B: Exponential amplification of specific target by repetitive PCR cycles. (Yilmaz et al., 2012)

Mullis first used a DNA polymerase purified from Escherichia coli. However, both the enzyme and DNA were denatured at high temperatures, resulting in the addition of the enzyme at each cycle of PCR, which was costly. After some experimentation, Thermophilus aquaticus (Taq) emerged as a new candidate as an enzyme (Zhu et al., 2020).

The bacterium can resist extremely high temperatures, so Taq DNA polymerase replaced E. coli DNA polymerase. With Taq DNA polymerase, which is still active at the denaturing temperature of DNA, there was no need to re-add the enzyme in each PCR cycle (Zhu et al., 2020).

Mullis won the Nobel Prize in 1993 for the invention of PCR. The technique revolutionized molecular biology; more than 20 PCR-based strategies were invented later, including hot-start, reverse transcription, real-time, direct, and methylation-specific PCR.

2.3 Sequencing

With the invention of PCR, it is possible to amplify targeted regions in a nucleic acid which is the first step to obtaining genomic information. The next step, sequencing, is to obtain the information on the order of the nucleic acids. The technologies developed for sequencing can be reviewed in three generations.

Second and third-generation sequencing, broadly called next-generation sequencing (NGS. During NGS, millions of small DNA fragments are sequenced in parallel, which results in high throughput data. NGS offers cost and time-efficient solutions, so it has been adapted for many areas, such as whole-genome sequencing, transcriptomics, metagenomics, whole-exome sequencing, and epigenetics.

2.3.1 The first generation of sequencing

Watson and Crick solved the structure of DNA in 1953, but “sequencing” DNA took many more years.

Sanger developed the ‘chain-termination’ or dideoxy technique in 1977, which was the breakthrough for first-generation sequencing methods. The accuracy, robustness, and ease of use led Sanger sequencing to become the most common technology used to sequence DNA for years to come. Early methods used radioactively labeled ddNTPs; however, fluorescently labeled ddNTPs have outperformed radioactive methods (Shendure & Ji, 2008).

Sanger sequencing requires five components: DNA template, primers, DNA polymerase, deoxynucleotide triphosphates (dNTPs), and dideoxynucleotide triphosphates (ddNTPs) labeled with fluorescent dyes. The technique takes advantage of ddNTPs as ddNTPs lack a 3’-OH group. Therefore, it is impossible to form a phosphodiester bond between two nucleotides at 3’, resulting in the termination of the extension by performing four parallel reactions containing each individual ddNTP base and detecting fluorescent light as each reaction occurs (Shendure & Ji, 2008). It is only possible to sequence around 1 kb of genomic material with Sanger sequencing, which is not helpful for large-scale genomic projects.

2.3.2 The second-generation of sequencing

The challenge of first-generation sequencing was that the methods were not high-throughput, so there was a need for a method to generate large amounts of data efficiently.

The second-generation methods can be summed up as high throughput sequencing of short reads because the efficient length for every method capped around 500 bases, and the methods were tuned to be massively parallelized.

Pyrosequencing is the pioneer among the second-generation sequencing methods. The technique is based on pyrophosphate synthesis. The synthesis consisted of a two-enzyme process. The first step is done with ATP sulfurylase to convert pyrophosphate into ATP. In the second step, ATP is used as the substrate for luciferase, producing light proportional to the amount of pyrophosphate. The main advantages of pyrosequencing are that it can be performed with standard nucleotides and observed in real-time (Nyrén et al., 1993). Later improvements also added new features, but the primary method did not change much.

The first sequencing machines that allowed the mass parallelization of sequencing reactions were produced by 454. The machine worked with a plate containing approximately one million individually sequenced wells. The success of 454 made the High Throughput Sequencing (HTS) principle the foundation principle of the second-generation sequencers (Heather & Chain, 2016). The schema of 454 sequencing is shown in Figure 3.

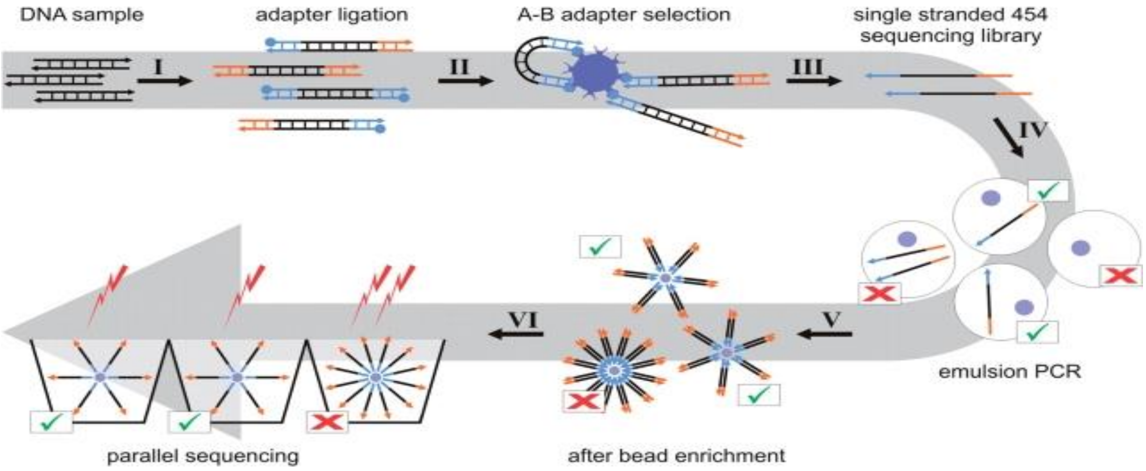


Figure 3: 454 sequencing (Meyer et al., 2008).

Although several second-generation sequencers have emerged, Illumina proved to be the most widely used worldwide as the company still dominates the market by owning 80% of the DNA sequencing market in 2021.

In Illumina sequencing, DNA molecules are attached to a flow cell and amplified with solid-state PCR to form DNA clusters. Reversible terminator bases (RT-bases) are added and washed away each cycle. A high-resolution camera keeps track of the clusters for fluorescent activity (Heather & Chain, 2016). The schema of Illumina sequencing is shown in Figure 4.

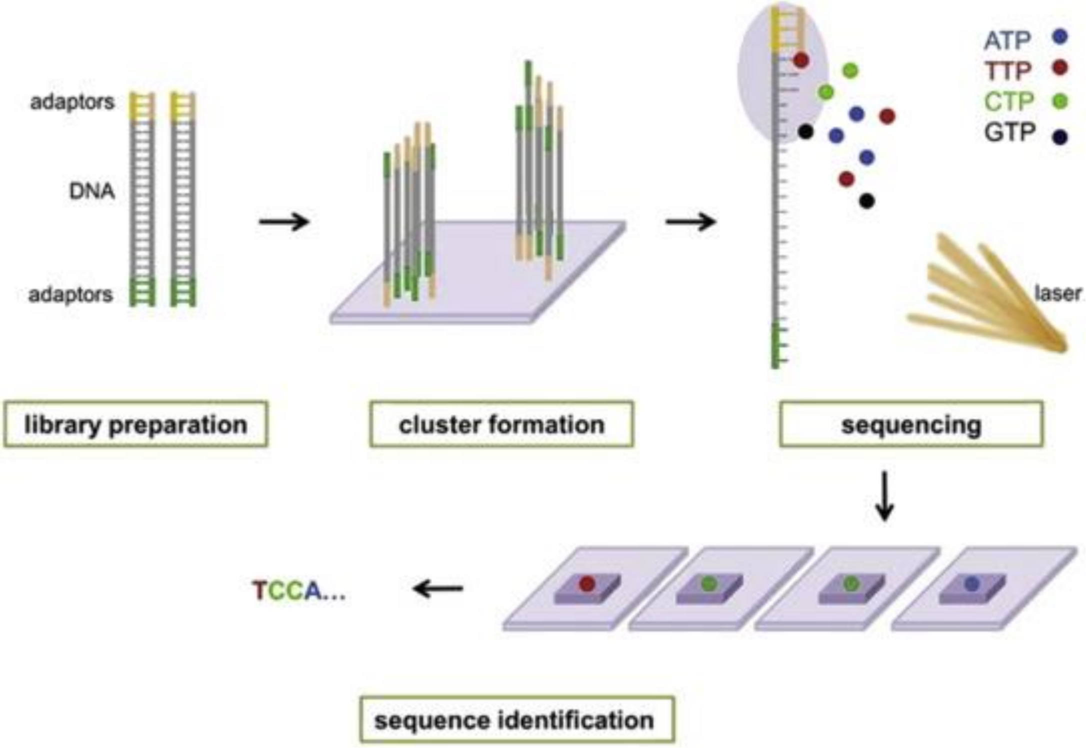


Figure 4: Illumina sequencing (Clark et al., 2018)

Three main Illumina sequencers, Miseq, Hiseq, and NovaSeq, can generate 30 million, 3 billion, and 13 billion reads per run (Heather & Chain, 2016). The latest machines can sequence the entire human genome in 48 hours at 1000$. The numbers alone summarize Illumina sequencers' power, yet error rates are pretty low, around once every thousand bases.

Despite the differences, Sanger, pyrosequencing, and Illumina methods are all classified as ‘sequence-by-synthesis’ (SBS) techniques, as they all require the direct action of DNA polymerase to produce the observable output.

2.3.3 The third-generation of sequencing

While second-generation sequencing is summed as short-read sequencing, third-generation sequencing represents long-read sequencing. The characteristics of third-generation methods are ‘single-molecule sequencing’ (SMS) and ‘real-time sequencing.’

The most widely used third-generation technology is the single-molecule real-time (SMRT) platform of PacBio. In PacBio SMRT runs, there are no flowcells. Arrays of microfabricated nanostructures called zero-mode waveguides (ZMWs), essentially tiny holes in a metallic film covering a chip, make the foundation of this technique. DNA polymerase is attached to the bottom of ZMWs, and as polymerization goes on, real-time bursts of the fluorescent signal are read. SMRT process can sequence single molecules exceeding 10 kb of reads in a short time (Heather & Chain, 2016). The schema of SMRT sequencing is shown in Figure 5.

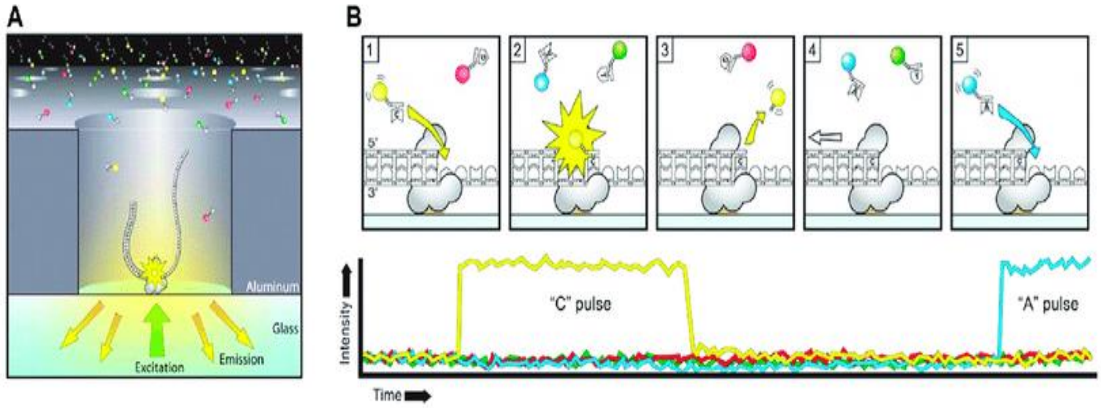


Figure 5: PacBio SMRT sequencing (Huo et al., 2021)

Another third-generation sequencing method, Nanopore, is developed by Oxford Nanopore Technologies (ONT). Nanopore sequencing is new compared to other techniques, first released to end-users in an early access trial in 2014. As the name suggests, nanopore sequencing relies on a flowcell containing numerous nanopores constructed with proteins. Single molecules are bound to the nanopores, and as the nucleotides pass by, the sequencer reads changes in electrical current. Nanopore sequencing suffers a high error rate per base, but as stated before, the technique is relatively new and has potential for improvements in the future. The schema of Nanopore sequencing is shown in Figure 6.

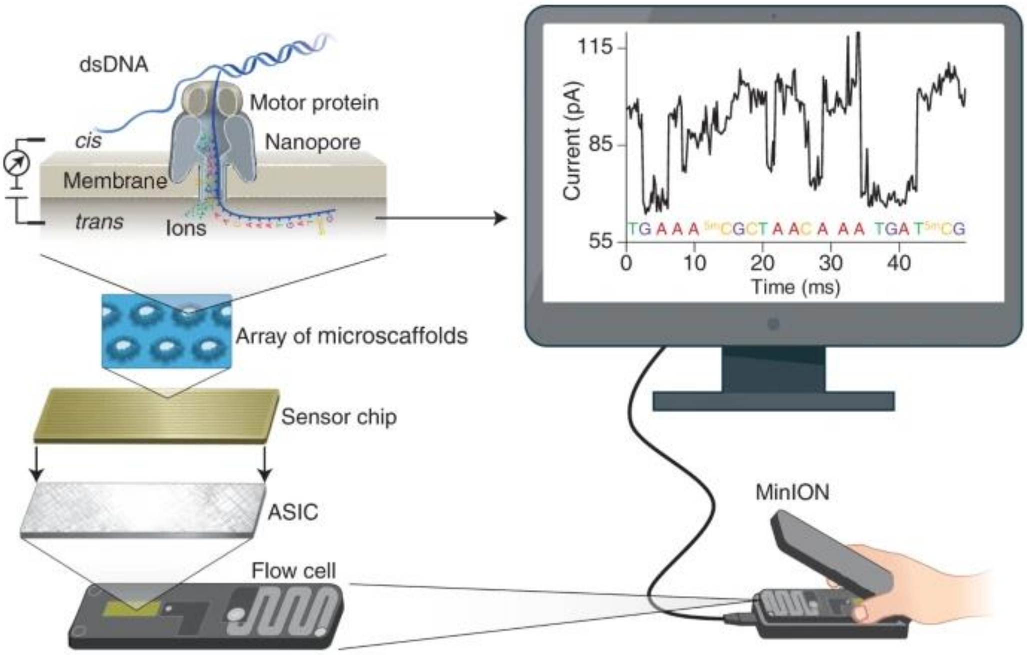


Figure 6: Nanopore sequencing (Wang et al., 2021)

There are many more sequencing techniques and platforms. The development of various NGS platforms is presented in Figure 7, and only the platforms that carried an impact today have been reviewed in this chapter.

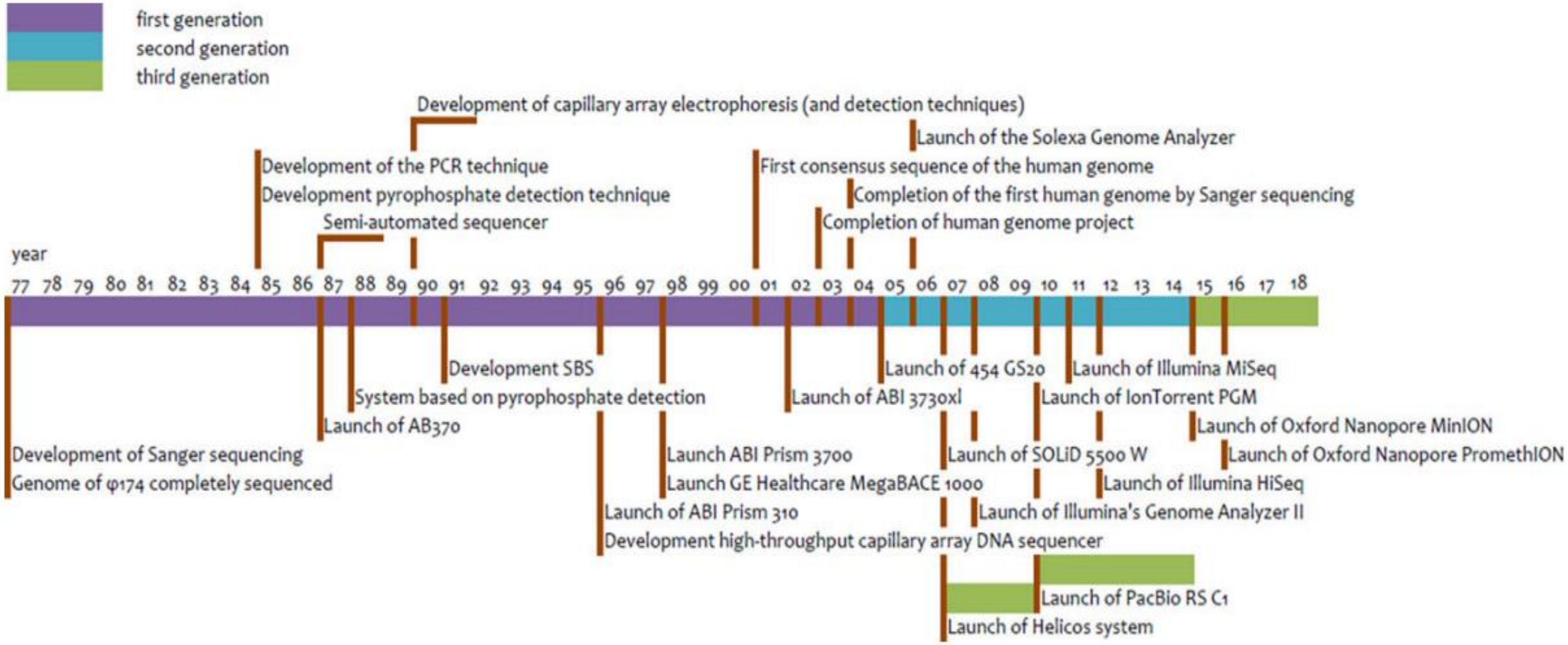

🟪 first generation
🟦 second generation
🟩 third generation
year
77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
Development of capillary array electrophoresis (and detection techniques)
Launch of the Solexa Genome Analyzer
Development of the PCR technique
Development pyrophosphate detection technique
First consensus sequence of the human genome
Completion of the first human genome by Sanger sequencing
Semi-automated sequencer
Completion of human genome project
Development SBS
System based on pyrophosphate detection
Launch of AB370
Development of Sanger sequencing
Genome of φ174 completely sequenced
Launch of ABI 3730xl
Launch of 454 GS20
Launch of Illumina MiSeq
Launch of IonTorrent PGM
Launch of Oxford Nanopore MinION
Launch ABI Prism 3700
Launch GE Healthcare MegaBACE 1000
Launch of SOLiD 5500 W
Launch of Oxford Nanopore PromethION
Launch of Illumina HiSeq
Launch of ABI Prism 310
Launch of Illumina's Genome Analyzer II
Development high-throughput capillary array DNA sequencer
Launch of PacBio RS C1
Launch of Helicos system

Figure 7: Timeline of NGS platforms. (Bruijns et al., 2018)

2.4 Amplicon Sequencing

A marker gene is an orthologous gene group that can be used to determine phylogeny in metagenomic terms. An orthologous gene group to be a marker gene needs two specifications. Firstly, a marker gene must be conserved in the research of interest. For example, the 16S rRNA gene is a highly conserved gene between bacteria and archaea. The following specification must have conserved and variable regions to design universal primers (Gohl et al., 2016).

Amplicon sequencing is the most cost-efficient and widely used method to characterize microbial communities. Taxonomic markers such as 16S SSU rRNA, 18S SSU rRNA, or ITS can be used to survey microbial diversity (Brumfield et al., 2020).

The length of these marker genes is mostly longer than 1 kb. For most NGS methods, amplicon sizes are restricted. For Illumina MiSeq, the maximum length of an amplicon is a 2×300 Paired-End sequence (Fadrosh et al., 2014). The schema of 16S amplicon sequencing is shown in Figure 8. In this figure V1-V3 region is amplified for analysis.

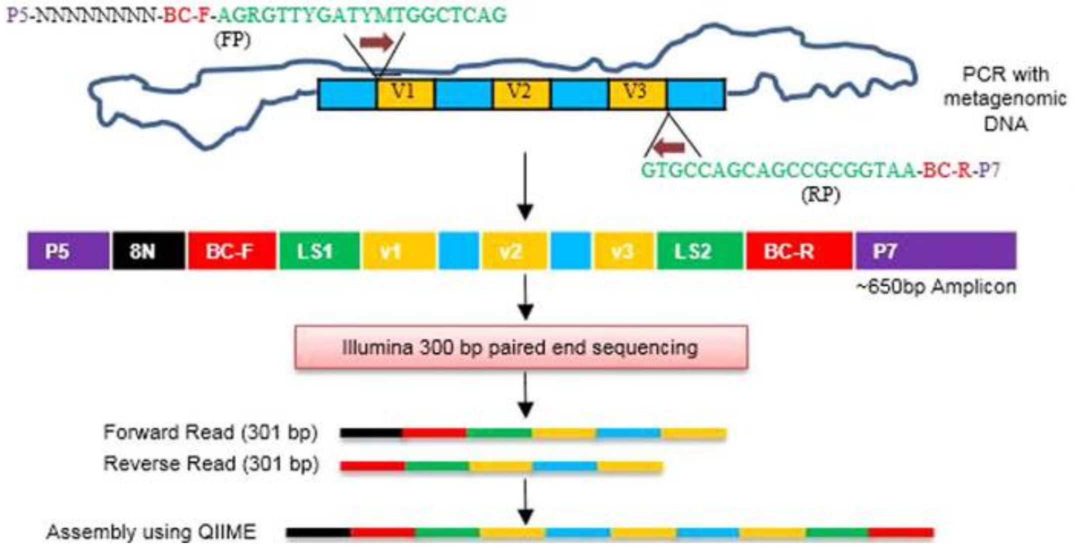


Figure 8: 16S amplicon sequencing. FP = forward primer, RP = reverse primer, V = variable region of 16S gene, P5 and P7 = Illumina sequencing primers, N = random nucleotide, BC-F = barcode of forward primer, BC-R = barcode of reverse primer, LS = linker sequence (Mandal et al., 2016)

However, these marker genes have conserved regions and variable regions. For example, 16S SSU rRNA has nine variable regions with an approximate length of 50 to 100 bases (Bodilis et al., 2012). The variability of the regions are shown in Figure 9.

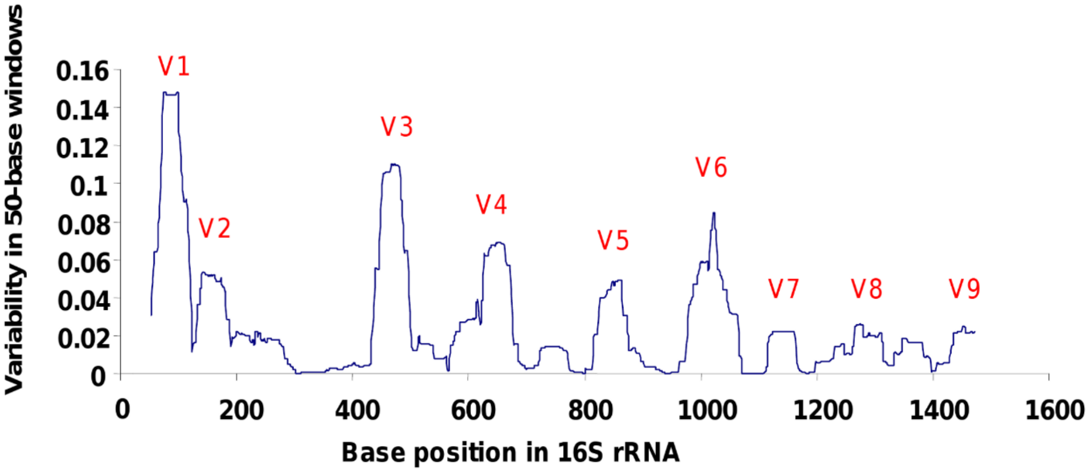


Figure 9: Variability of 16S regions. (Bodilis et al., 2012)

Targeting these variable regions allows us to characterize microbial diversity to a particular taxonomic level. Considering the PCR and primer biases polymorphisms of marker genes, the resolution is mainly limited to the genus level.

2.5 Bioinformatics Analysis of Amplicon Sequencing

Mothur and QIIME 2 are the most cited bioinformatics tools for amplicon sequencing analysis. Even though it is noted that these tools can perform with any dataset, it is shown that the results may differ immensely. There is a fundamental difference in how Mothur and QIIME 2 are developed (Prodan et al., 2020).

Mothur is a standalone tool that implements several algorithms and is available on all operating systems (Schloss et al., 2009). On the other hand, QIIME 2 is a python interface that orchestrates many different programs by altering inputs and outputs. Also, the interface has a plugin architecture that allows third parties to contribute functionality (Bolyen et al., 2019). QIIME 2 developers only develop the core plugins. So, QIIME 2 platform has different plugins for every analysis step.

As a result of the differences in the approaches, QIIME 2 features more programs than Mothur. There are 57 plugins available on QIIME 2, and only 22 are core plugins. A list of all plugins is available on QIIME 2 Library (https://library.qiime2.org/plugins/), and their descriptions are listed in the appendix.

2.5.1 QIIME 2

QIIME 2, the successor of QIIME (Caporaso et al., 2010), is an open-source data science platform, including interactive spatial and temporal analysis and visualization tools. QIIME 2 View (https://view.qiime2.org) is a service that allows researchers to view and share data without QIIME 2 clients. The basic QIIME 2 flowchart is shown in Figure 10. QIIME 2 can start the analysis from raw sequences and finish with publication-grade visualizations.

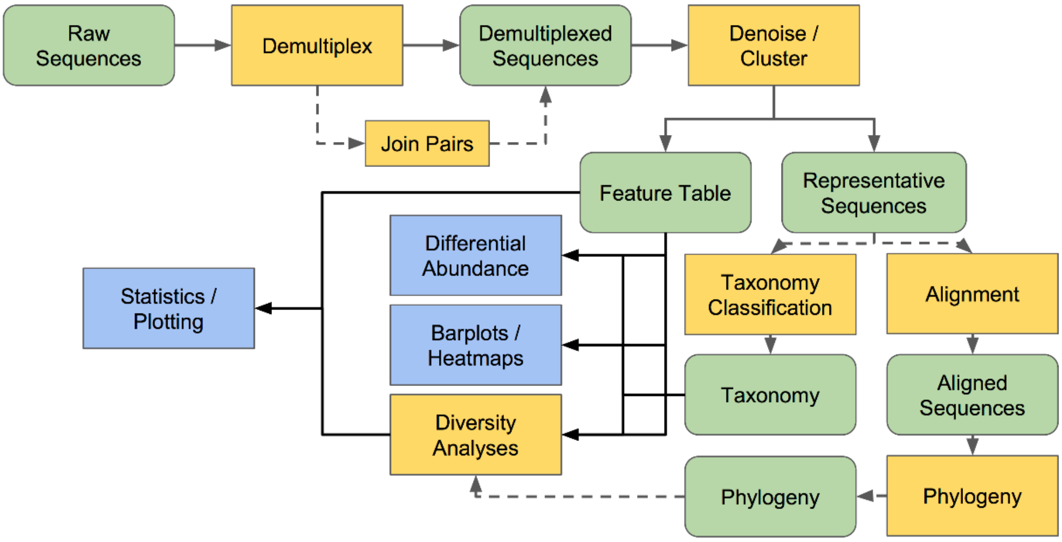


Figure 10: QIIME 2 Basic Flowchart (https://docs.qiime2.org/2022.2/tutorials/overview/)

For every method and visualization shown in the flow chart, multiple plugins are available in QIIME 2, making it easy for users to build their custom pipelines.

There are many steps and many different inclusions to amplicon sequencing analysis. The analysis is limited to taxonomic classification in this study. For taxonomic classification, it can be summarized in these steps: 1-Quality control, 2-Denoising or clustering, 3-Chimera detection, 4-Merging reads, 5-Taxonomic classification, 6-Visualization

2.5.1.1 Quality Control

q2-demux is a QIIME 2 plugin that can demultiplex sequence reads and visualize quality information. The quality information is based on PHRED scores (Cock et al., 2009). An example quality plot is shown in Figure 11.

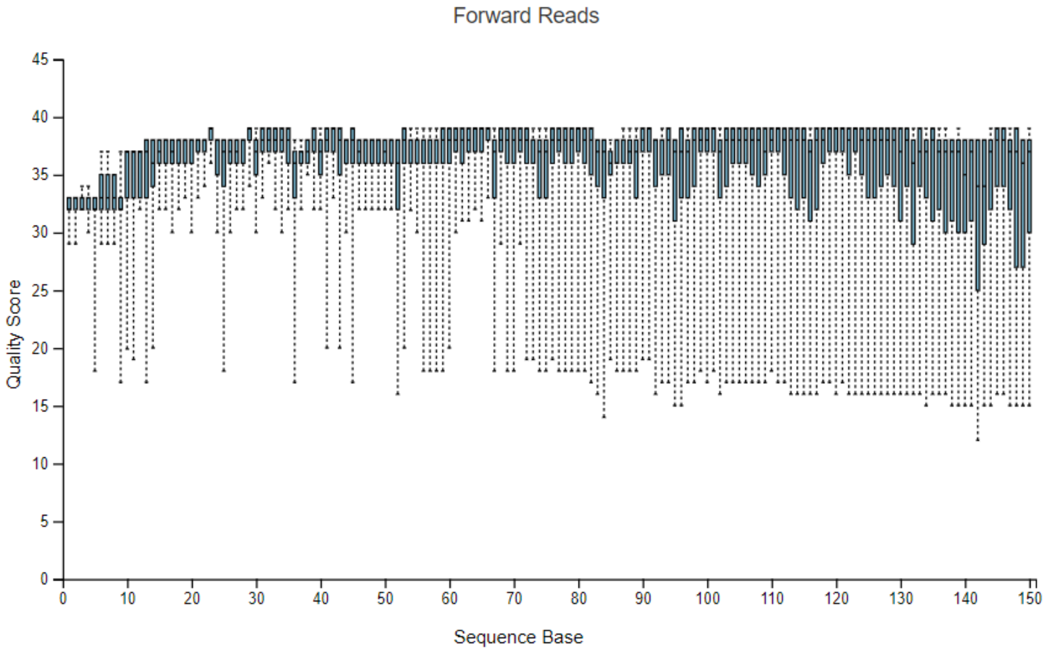


Figure 11: Example q2-demux visualization. Data to generate the plot is obtained from Parkinson's mouse study. (Sampson et al., 2016)

The box plots are generated for each position using a random sampling of 10000 out of the number of total sequences. Unlike other quality control tools, q2-demux plots data are obtained from all samples.

2.5.1.2 Denoising or Clustering

Denoising and clustering are two different methods to obtain a feature table in QIIME 2.

Clustering amplicon sequences based on a reference database or de novo is an older method to obtain Operational Taxonomic Units (OTUs). q2-vsearch (Rognes et al., 2016) and q2-dbOTU (Preheim et al., 2013) plugins are available for this method.

The denoising method is based on differences of a single nucleotide, outperforming OTUs, and generating Amplicon Sequence Variants (ASVs). q2-dada2 (Callahan et al., 2016) and q2-deblur (Amir et al., 2017) plugins are available for this method.

2.5.1.3 Chimera Detection

Chimeric reads are PCR products from multiple parent sequences. These reads are not biological sequences but artifacts. Chimeras inflate the apparent diversity, so they should be removed. The schema of chimera formation in PCR is shown in Figure 12.

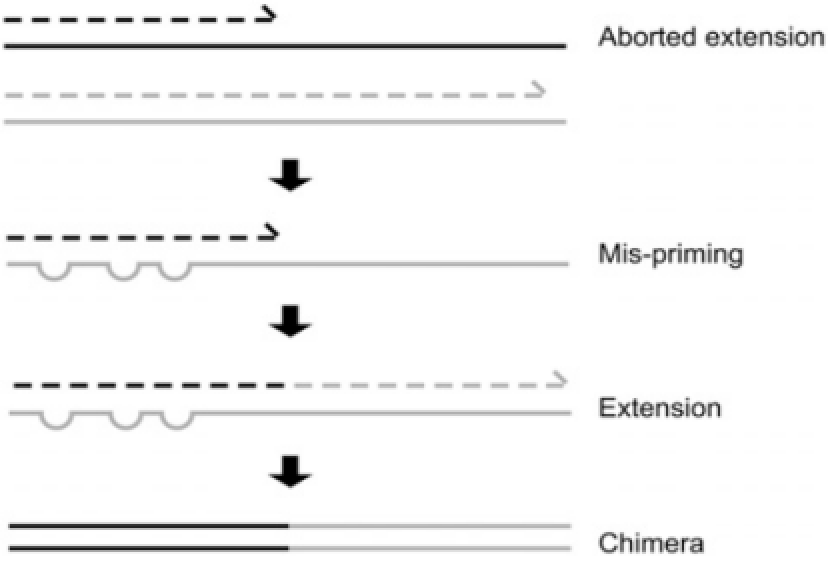


Figure 12: Chimera formation (Haas et al., 2011)

Both q2-vsearch and q2-dada2 have their options for chimera removal. q2-vsearch uses a reference-based method, and q2-dada2 uses a de novo approach.

2.5.1.4 Merging Reads

The reads must be merged to obtain full-length amplicon sequences if the sequencing method is paired-end. A global ends-free alignment is performed between forward and reverse reads. The minimum length of exact overlap is usually ten bases.

Both q2-vsearch and q2-dada2 have their options for merging reads. The general concept of read merging is shown in Figure 13.

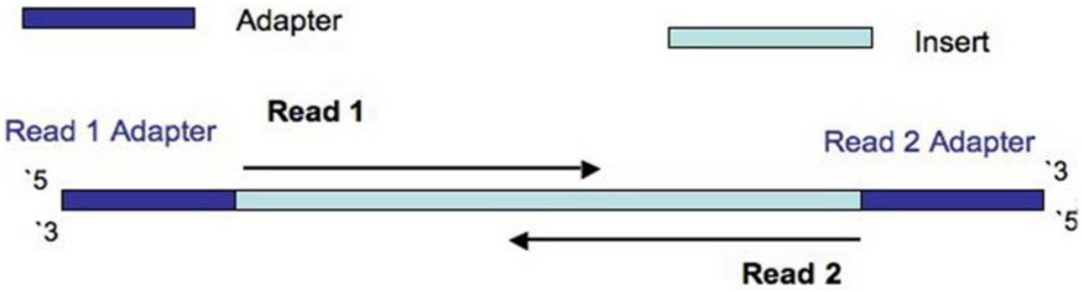


Figure 13: Merging paired end reads (Turner, 2014)

2.5.1.5 Taxonomic Classification

For taxonomic classification, a reference database is needed. There are several reference database projects, such as SILVA (Quast et al., 2013), Greengenes (DeSantis et al., 2006), and UNITE (Kõljalg et al., 2005).

There are two different approaches to the taxonomic classification.

The first one is alignment-based methods. These methods find a consensus assignment across top N hits. Alignment-based methods do not require pre-training.

The second one is a machine learning-based classification method. For this method, a classifier must be trained with a reference database, and then classification can be performed. Machine learning-based classification methods outperform alignment-based methods. However, machine learning-based classification methods can be memory intensive and slow.

Both methods are available in the q2-feature-classifier plugin (Bokulich et al., 2018).

2.5.1.6 Visualization

There are many visualization methods available for visualizing taxonomic classifications. Principally, three methods are preferred: Bar plots, heatmaps, and pie charts. An example heatmap is shown in Figure 14, and an example bar plot is shown in Figure 15.

Bar plots and heatmaps can be plotted using all samples and features, leading to a panoramic view. Pie charts are plotted for each sample which allows for individual checks.

The q2-taxa plugin can be used for bar plots, and the q2-feature-table plugin can be used for heatmaps. There is no plugin available for pie charts in QIIME 2.

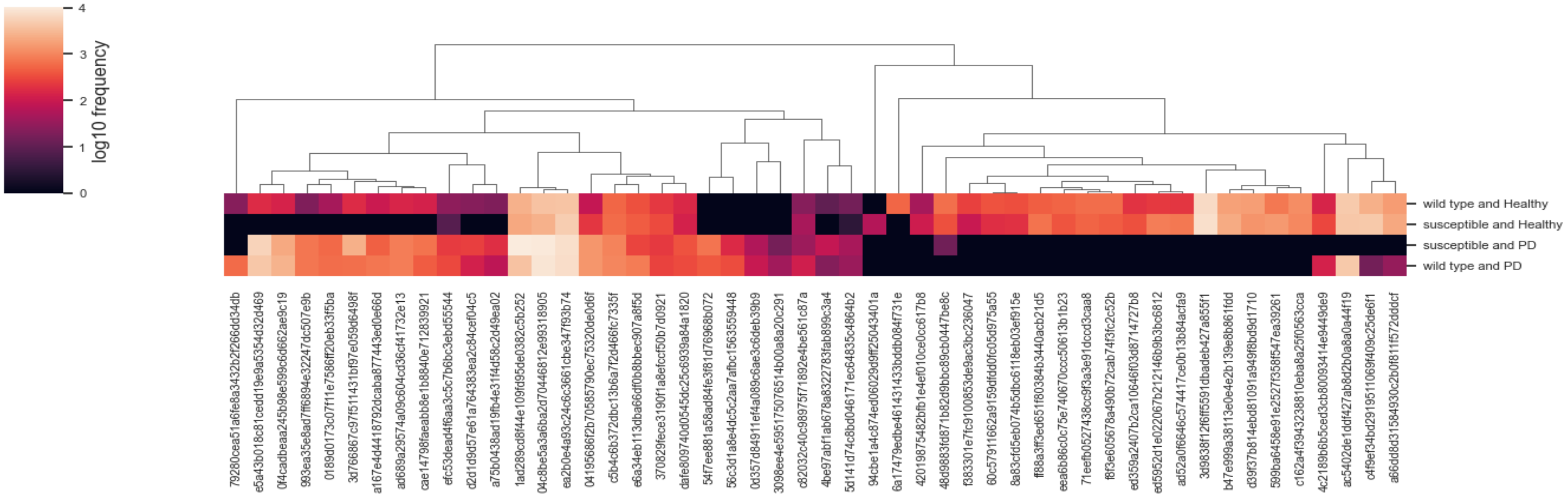


Figure 14: An example heatmap generated by QIIME 2. Data is obtained from Parkinson's Mouse study (Sampson et al., 2016)

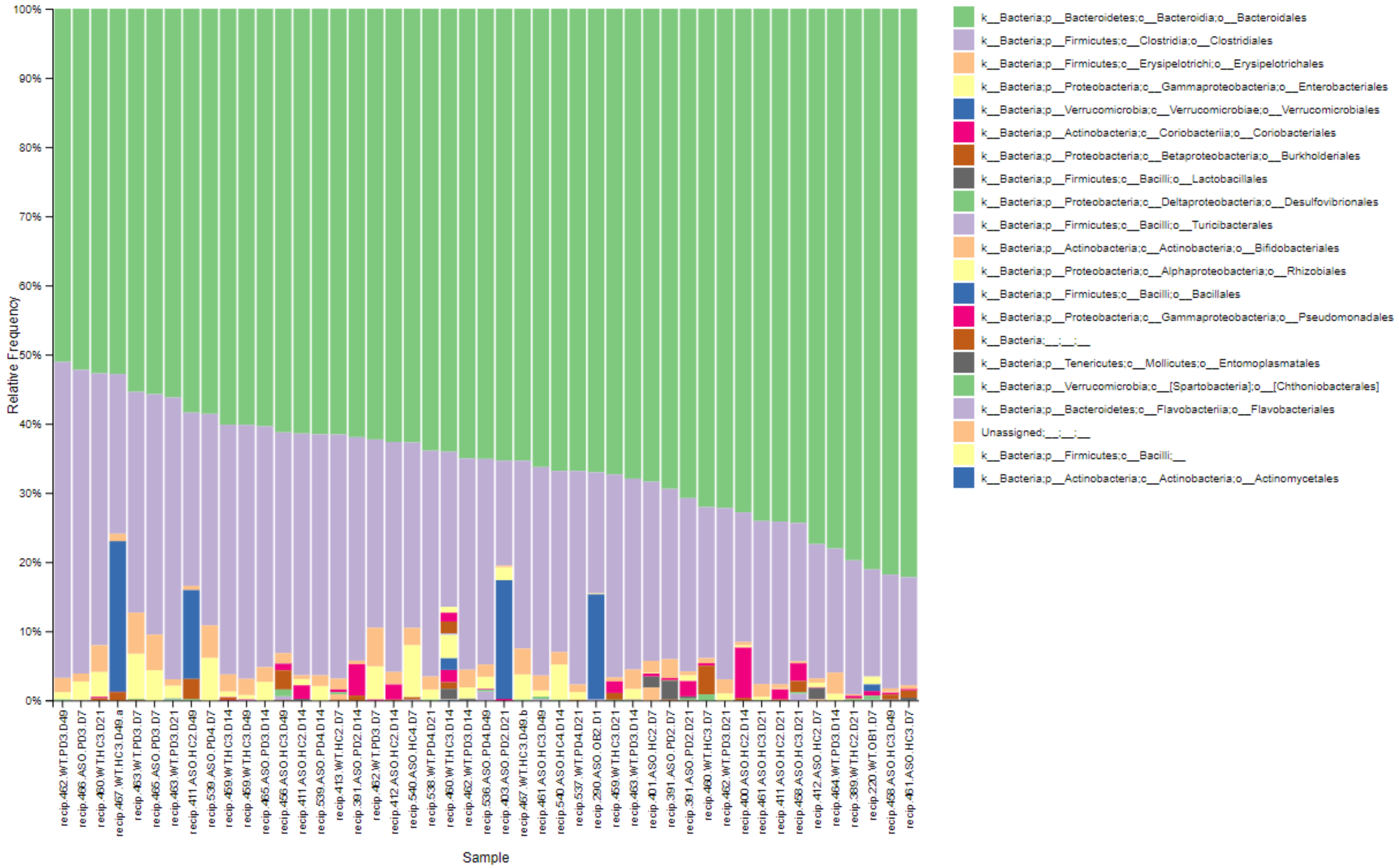


Figure 15: An example barplot generated by QIIME 2. Data is obtained from Parkinson's Mouse study (Sampson et al., 2016)

CHAPTER 3

METHOD

3.1 QIIME 2 Architecture

There are two types of data files in QIIME 2 architecture; artifacts and visualizations. Both data file types exist as compressed versions. There are semantic types, plugins, methods, and visualizers other than data files.

3.1.1 Artifacts

Artifacts contain data and metadata. The metadata describes the data stored in the artifact, such as its type, format, and provenance. Artifacts are used as inputs for methods. This file type has "qza" file extension.

3.1.2 Visualizations

Visualizations contain the metadata just as artifacts. Unlike artifacts, visualizations contain terminal outputs and visual representations of the data. This file type has a "qzv" file extension.

Visualization (qzv) files can be viewed with QIIME 2 View at https://view.qiime2.org. So, visualization files do not require a QIIME 2 Client, enabling easy sharing of visuals.

3.1.3 Semantic Types

Semantic types help users to use correct files for analyses. Each plugin in QIIME 2 accepts inputs and outputs as certain semantic types to prevent incorrect analyses. Usage of an example method is shown in Figure 16. Options have semantic types attributed to them in green-colored text.

Some of the common semantic types are as follows:

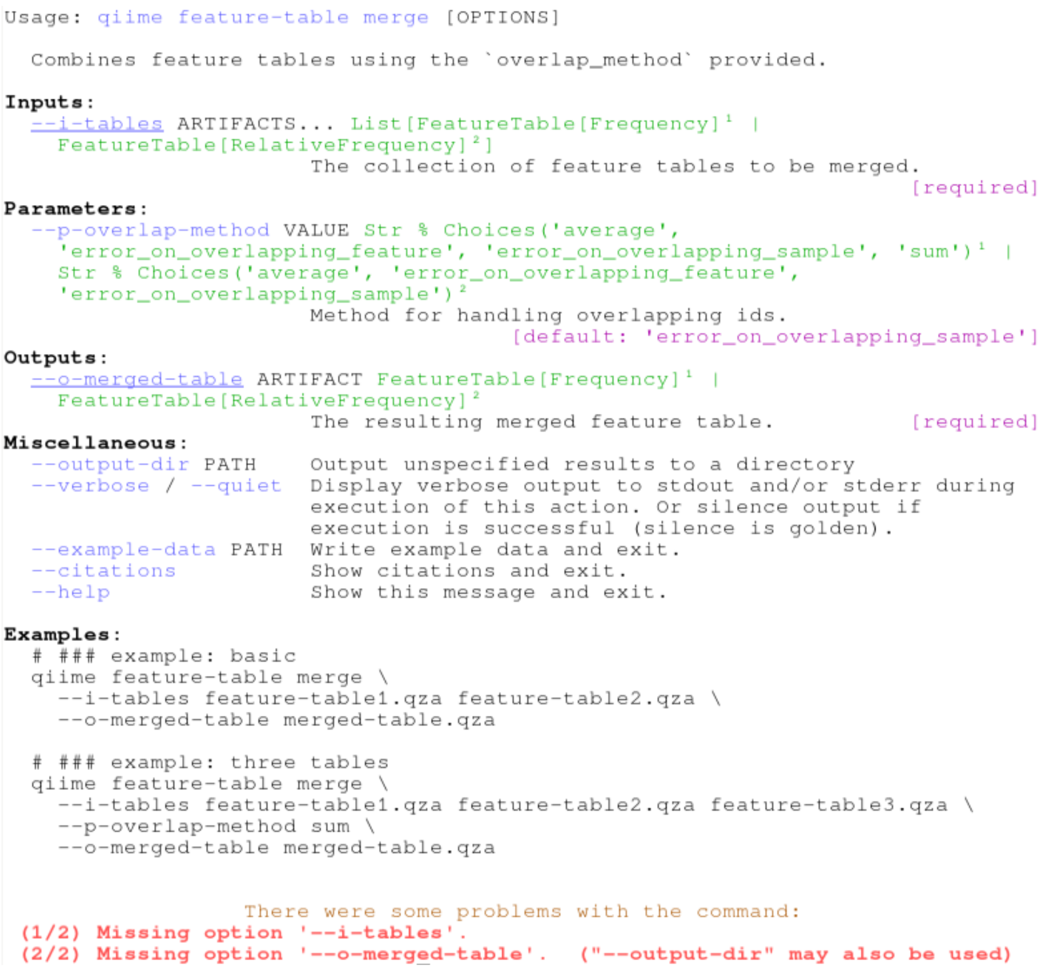
- FeatureTable[Frequency]: A feature table (e.g., samples by OTUs) where each value indicates the frequency of an OUT in the corresponding sample expressed as raw counts.

- Phylogeny[Rooted]: A rooted phylogenetic tree.

- Phylogeny[Unrooted]: An unrooted phylogenetic tree.

- DistanceMatrix: A distance matrix.

- PCoAResults: The results of a running principal coordinate analysis (PCoA).



```
Usage: qiime feature-table merge [OPTIONS]

  Combines feature tables using the `overlap_method` provided.

Inputs:
  --i-tables ARTIFACTS... List[FeatureTable[Frequency]¹ |
    FeatureTable[RelativeFrequency]²]
                       The collection of feature tables to be merged.
                                                                [required]
Parameters:
  --p-overlap-method VALUE Str % Choices('average',
    'error_on_overlapping_feature', 'error_on_overlapping_sample', 'sum')¹ |
    Str % Choices('average', 'error_on_overlapping_feature',
    'error_on_overlapping_sample')²
                       Method for handling overlapping ids.
                                      [default: 'error_on_overlapping_sample']
Outputs:
  --o-merged-table ARTIFACT FeatureTable[Frequency]¹ |
    FeatureTable[RelativeFrequency]²
                       The resulting merged feature table.      [required]
Miscellaneous:
  --output-dir PATH    Output unspecified results to a directory
  --verbose / --quiet  Display verbose output to stdout and/or stderr during
                       execution of this action. Or silence output if
                       execution is successful (silence is golden).
  --example-data PATH  Write example data and exit.
  --citations          Show citations and exit.
  --help               Show this message and exit.

Examples:
  # ### example: basic
  qiime feature-table merge \
    --i-tables feature-table1.qza feature-table2.qza \
    --o-merged-table merged-table.qza

  # ### example: three tables
  qiime feature-table merge \
    --i-tables feature-table1.qza feature-table2.qza feature-table3.qza \
    --p-overlap-method sum \
    --o-merged-table merged-table.qza


                  There were some problems with the command:
 (1/2) Missing option '--i-tables'.
 (2/2) Missing option '--o-merged-table'.  ("--output-dir" may also be used)
```

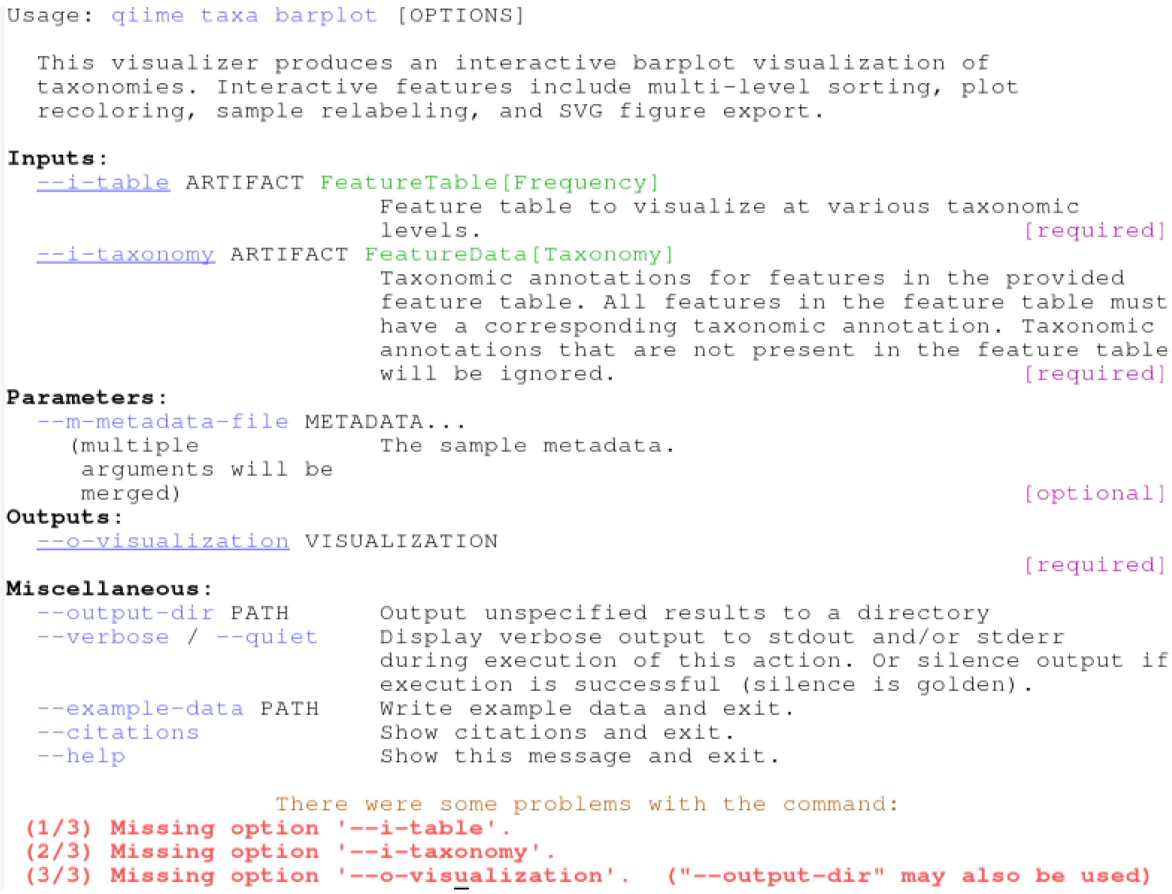
Figure 16: Example method. Input and output semantic types helps users to use the proper files.

3.1.4 Plugins

QIIME 2 platform consists of many plugins. QIIME 2 team has developed the core plugins, but third-party developers are encouraged to develop their plugins and publish them. This decentralized development approach has allowed QIIME 2 to be more accessible. As a result of this approach, QIIME 2 has a very supportive community. On the forum page (https://forum.qiime2.org/), developers and other users help each other with their research or plugin developments and improvements.

3.1.5 Methods and Visualizers

Methods are actions defined in QIIME 2 plugins, accepting inputs and outputs as artifacts. Visualizers are the end actions. They accept inputs as artifacts and outputting as single visualization. Usage of an example visualizer is shown in Figure 17.



```
Usage: qiime taxa barplot [OPTIONS]

  This visualizer produces an interactive barplot visualization of
  taxonomies. Interactive features include multi-level sorting, plot
  recoloring, sample relabeling, and SVG figure export.

Inputs:
  --i-table ARTIFACT FeatureTable[Frequency]
                       Feature table to visualize at various taxonomic
                       levels.                                    [required]
  --i-taxonomy ARTIFACT FeatureData[Taxonomy]
                       Taxonomic annotations for features in the provided
                       feature table. All features in the feature table must
                       have a corresponding taxonomic annotation. Taxonomic
                       annotations that are not present in the feature table
                       will be ignored.                           [required]
Parameters:
  --m-metadata-file METADATA...
    (multiple          The sample metadata.
     arguments will be
     merged)                                                      [optional]
Outputs:
  --o-visualization VISUALIZATION
                                                                  [required]
Miscellaneous:
  --output-dir PATH      Output unspecified results to a directory
  --verbose / --quiet    Display verbose output to stdout and/or stderr
                         during execution of this action. Or silence output if
                         execution is successful (silence is golden).
  --example-data PATH    Write example data and exit.
  --citations            Show citations and exit.
  --help                 Show this message and exit.

                  There were some problems with the command:
 (1/3) Missing option '--i-table'.
 (2/3) Missing option '--i-taxonomy'.
 (3/3) Missing option '--o-visualization'.  ("--output-dir" may also be used)
```

Figure 17: Example visualizer. As all visualizers do, ‘taxa barplot’ also ends with a visualization output.

3.2 Krona

Krona is a bioinformatics tool that visualizes hierarchical data as interactive multi-layered pie charts. Even though Krona is developed for metagenomics, it can be used for any hierarchical data.

Krona can use many different data sources, but only the "ktImportText" function is used in this study. This function takes tab-separated text files as input. An example input is shown in Table 1, and the result is shown in Figure 18.

Table 1: Example tab-separated file for Krona.

2	Fats	Saturated fat	
3	Fats	Unsaturated fat	Monounsaturated fat
3	Fats	Unsaturated fat	Polyunsaturated fat
13	Carbohydrates	Sugars	
4	Carbohydrates	Dietary sugar	
21	Carbohydrates		
5	Protein		
4			

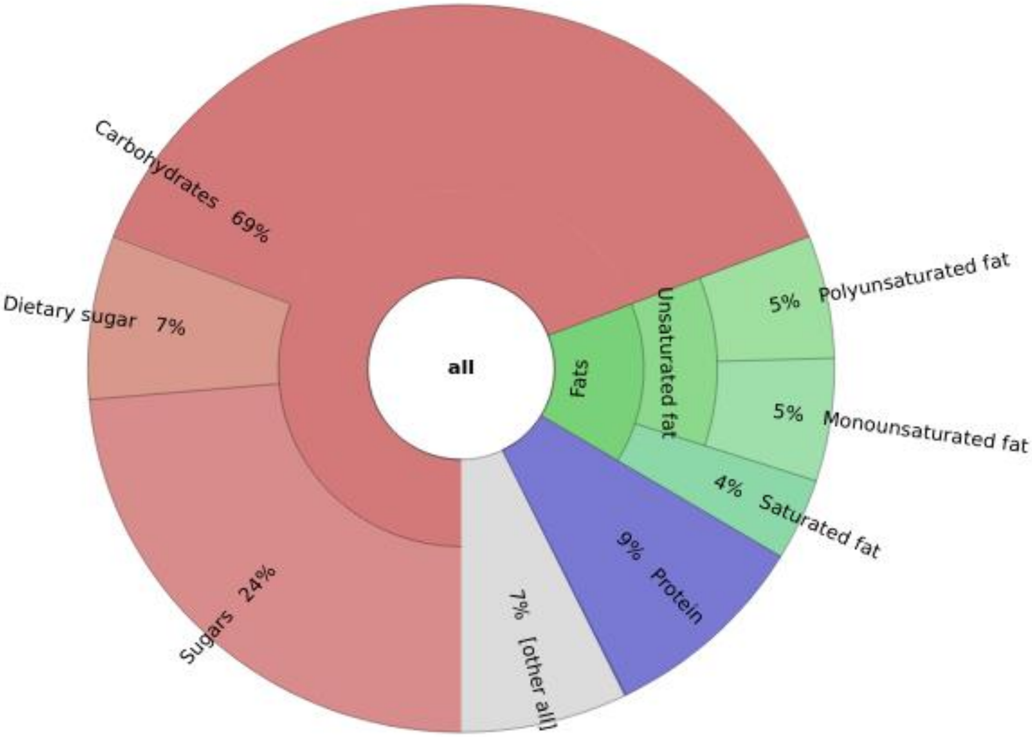


Figure 18: An example Krona plot

Today, many metagenomics tools and pipelines implement Krona for visualization purposes.

Tools that use Krona are:

- PhyloSift: http://phylosift.wordpress.com/
- metAMOS: https://github.com/treangen/metAMOS/wiki
- MGTAXA: http://mgtaxa.jcvi.org/
- Metavir: http://metavir-meb.univ-bpclermont.fr/
- PhymmBL: http://www.cbcb.umd.edu/software/phymm/
- MG-RAST: http://metagenomics.anl.gov/metagenomics.cgi?page=Home
- BacDive: https://bacdive.dsmz.de/isolation-sources
- Prophane: https://prophane.de/

3.3 QIIME 2 FeatureTable[Frequency] Semantic Type

FeatureTable[Frequency] contains BIOM files (McDonald et al., 2012). BIOM file format is converted to a tab-separated format so that the information it has can be used to generate Krona plots. An example tab-separated feature table is shown in Table 3.

Table 2: An example QIIME 2 feature table

#OTU ID	Sample 1	Sample 2	Sample 3
000c138572dbf0e0e1302483f21cffdf	1276	2395	1129
003d56badecf8011481e284fafb2a8b0	0	10	450
00628a58c966cbdb229afa401a691d2a	103	637	5642
007b31ffd1155f4384465acfebf23956	42	2	384
00867fa0fc82cd47f5baab8b286ab6d0	3320	3458	1209

CHAPTER 4

RESULTS

4.1 Development of q2-krona

QIIME 2 is originally written in Python, so Python has been selected as the language for q2-krona development.

Two tasks need to be completed before sending an input file to Krona. The first task is to assign taxonomy. A feature table consists of hashes and does not contain hierarchical data. An example taxonomy assigned feature table is shown in Table 4.

Table 3: Taxonomy assigned feature table

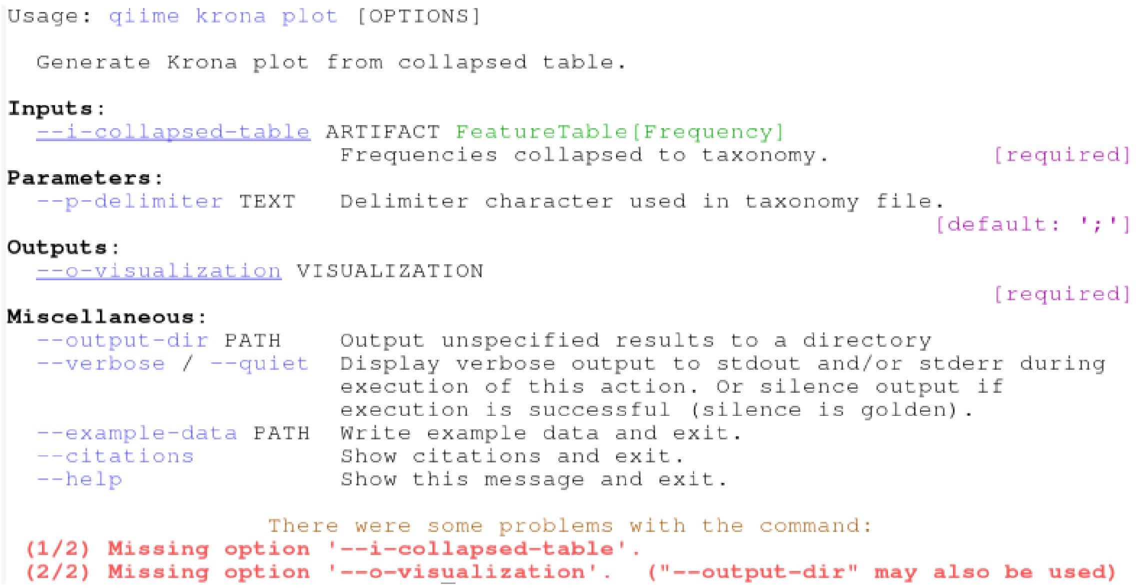
#OTU ID	Sample 1	Sample 2	Sample 3
k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales	1276	2395	1129
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales	103	647	6092
k__Bacteria;__;__;__	42	2	384
k__Bacteria;p__Firmicutes;c__Bacilli;__	3320	3458	1209

There are two essential steps to assigning taxonomy. The first step is to assign a taxonomy for each hash. The next step is to sum up ("collapse") the features assigned to the same taxonomy.

4.1.1 q2-krona “plot” method

The primary method created in the plugin q2-krona is named "plot." The "plot" method takes a FeatureTable[Frequency] semantic type file as an input regardless of the taxonomy. The method exports the BIOM file and handles the data to generate a tab-separated file for each sample in a temporary directory.

Using these files as input for the "ktImportText" method, Krona plots are generated. "plot" method also has a parameter for defining the delimiter of the taxonomy file. The delimiter is set to default as ";" Most reference taxonomies use ";" as a delimiter, as shown in Table 3, but there could be exceptions. Usage of the "plot" method is shown in Figure 19.



```
Usage: qiime krona plot [OPTIONS]

  Generate Krona plot from collapsed table.

Inputs:
  --i-collapsed-table ARTIFACT FeatureTable[Frequency]
                       Frequencies collapsed to taxonomy.           [required]
Parameters:
  --p-delimiter TEXT   Delimiter character used in taxonomy file.
                                                                [default: ';']
Outputs:
  --o-visualization VISUALIZATION
                                                                    [required]
Miscellaneous:
  --output-dir PATH    Output unspecified results to a directory
  --verbose / --quiet  Display verbose output to stdout and/or stderr during
                       execution of this action. Or silence output if
                       execution is successful (silence is golden).
  --example-data PATH  Write example data and exit.
  --citations          Show citations and exit.
  --help               Show this message and exit.

                There were some problems with the command:
 (1/2) Missing option '--i-collapsed-table'.
 (2/2) Missing option '--o-visualization'.  ("--output-dir" may also be used)
```

Figure 19: q2-krona plot

4.1.2 q2-krona "collapse_and_plot" pipeline

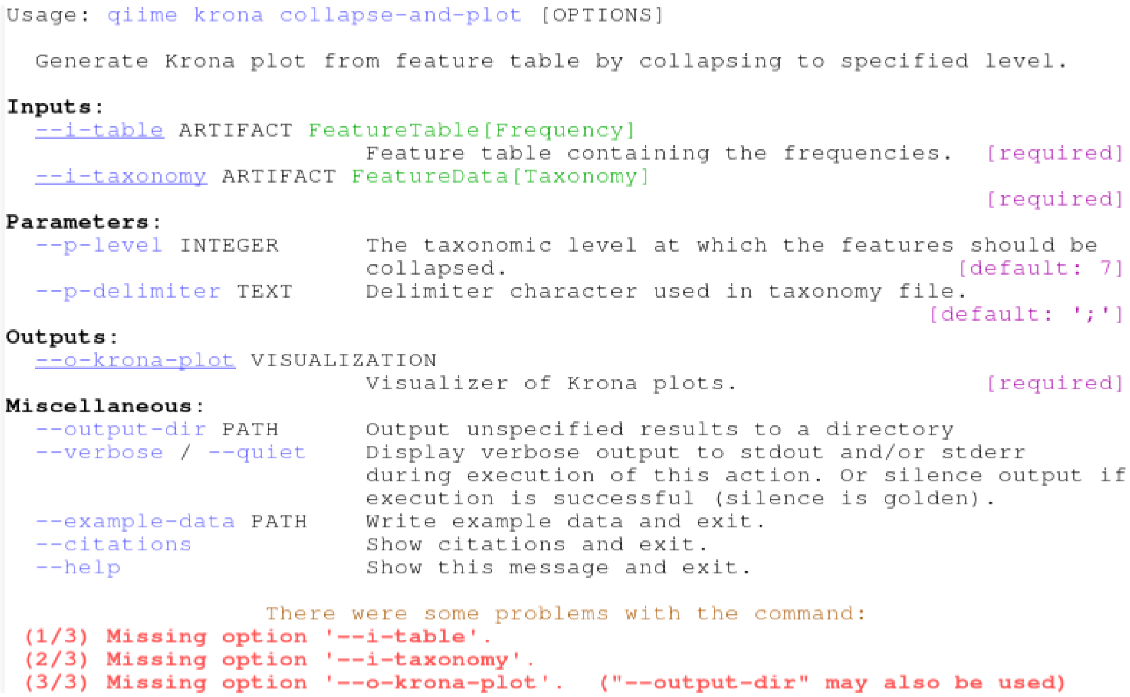
The semantic type of both feature tables is the FeatureTable[Frequency]. The action "plot" built for q2-krona only checks for the semantic type even though it is not sensible to plot the table containing hashes. q2-taxa plugin has the method "collapse," which assigns taxonomy to feature tables. A pipeline is created to use this method to generate Krona plots.

The pipeline created in the plugin q2-krona is named "collapse_and_plot." It pipes the "collapse" method to "plot."

The pipeline takes a FeatureTable[Frequency] and FeatureTable[Taxonomy] as input. The inputs are used in "collapse" to obtain hierarchical data.

The proposed pipeline uses the delimiter parameter of the "plot" method as the first parameter. The second parameter is the highest level to "collapse" feature tables.

Taxonomic levels (kingdom, phylum, class, order, family, genus, species) are available in most reference databases. The parameter is set to default as seven, pointing to species level. Usage of the "collapse_and_plot" pipeline is shown in Figure x.



```
Usage: qiime krona collapse-and-plot [OPTIONS]

  Generate Krona plot from feature table by collapsing to specified level.

Inputs:
  --i-table ARTIFACT FeatureTable[Frequency]
                         Feature table containing the frequencies.  [required]
  --i-taxonomy ARTIFACT FeatureData[Taxonomy]
                                                                    [required]
Parameters:
  --p-level INTEGER      The taxonomic level at which the features should be
                         collapsed.                               [default: 7]
  --p-delimiter TEXT     Delimiter character used in taxonomy file.
                                                                [default: ';']
Outputs:
  --o-krona-plot VISUALIZATION
                         Visualizer of Krona plots.                 [required]
Miscellaneous:
  --output-dir PATH      Output unspecified results to a directory
  --verbose / --quiet    Display verbose output to stdout and/or stderr
                         during execution of this action. Or silence output if
                         execution is successful (silence is golden).
  --example-data PATH    Write example data and exit.
  --citations            Show citations and exit.
  --help                 Show this message and exit.

                  There were some problems with the command:
 (1/3) Missing option '--i-table'.
 (2/3) Missing option '--i-taxonomy'.
 (3/3) Missing option '--o-krona-plot'.  ("--output-dir" may also be used)
```

Figure 20: q2-krona collapse_and_plot

4.2 Case Studies

Two case studies are chosen to present the performance of the plugin developed. Parkinson’s Mouse study (Sampson et al., 2016) is about differences in microbiomes between diseased and healthy hosts, and Sewage Treatment Systems (Hidalgo et al., 2021) study is about environmental issues. Even though these two studies cover different fields, they use QIIME 2 for amplicon sequencing analysis.

Also, each study focuses on specific taxa at certain analysis points where Krona plots can become very handy.

For demo purposes, raw sequences for each study have been retrieved from open access archives, and taxonomic analysis steps are redone as best as the relevant study describes. In addition, q2-krona is used for visualization.

4.2.1 Case 1: Parkinson's Mouse Study

The study is designed to reveal if the fecal microbiome affects Parkinson's Disease (PD) by 16S rRNA gene amplicon data analysis. Feces were collected from donors with PD and healthy controls to mice susceptible to developing PD due to a mutation (αSyn) or wild-type resistant to PD. The mice were checked for seven weeks to see if there were PD symptoms (Sampson et al., 2016).

The statement in the discussion part by the authors as follows:

> *"Several bacterial taxa are altered in mice receiving fecal transplants from PD patients compared to healthy controls. Additionally, number of bacterial genera are changed specifically in ASO [alpha-synuclein overexpressing] animals, but not WT [wild-type] mice, receiving microbes from the same donor. These include depletions in members of the family Lachnospiraceae and Ruminococceae in recipient mice, a notable finding as these same genera are significantly reduced in fecal samples directly from PD patients.* (Sampson et al., 2016)."

Any visualization does not support this statement about certain taxa.

Open-access data is obtained from EMBL ENA (http://www.ebi.ac.uk/ena) with the study accession ERP019564, and bioinformatic analyses are performed with QIIME 2 as described in the study. q2-krona plugin is used for visualization of taxonomic analysis.

The figures show the results of bar plots and searching *Lachnospiraceae* and *Ruminococceae* in Krona plots. In the quote, it is stated that there are depletions in the family Lachnospiraceae and Ruminococceae, but it is impossible to visualize members of the families with bar plots (Figure 21) whereas it can be seen very clearly in Krona plots (Figure 22-25).

For example, if we inspect sample recip.460.WT.HC3.D21, in barplots it is pretty noisy if it is desired to get information about lesser abundant taxa. Also, the color palette repeats itself, which could get us confused.

In the Krona plot in Figure 22, the sample is more intuitive as it has a continuous color palette, and the plot is interactive, allowing one to zoom in to view different levels.

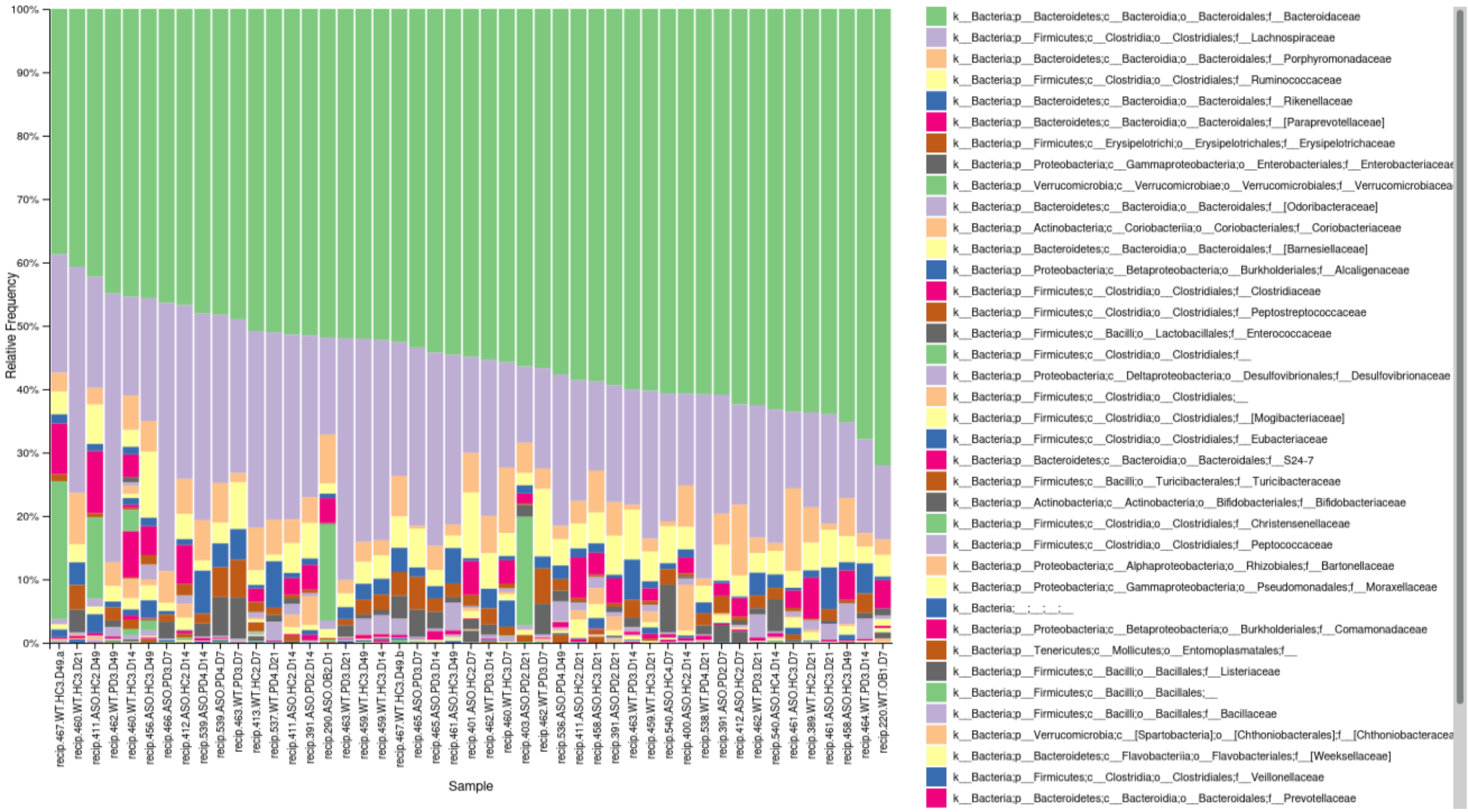
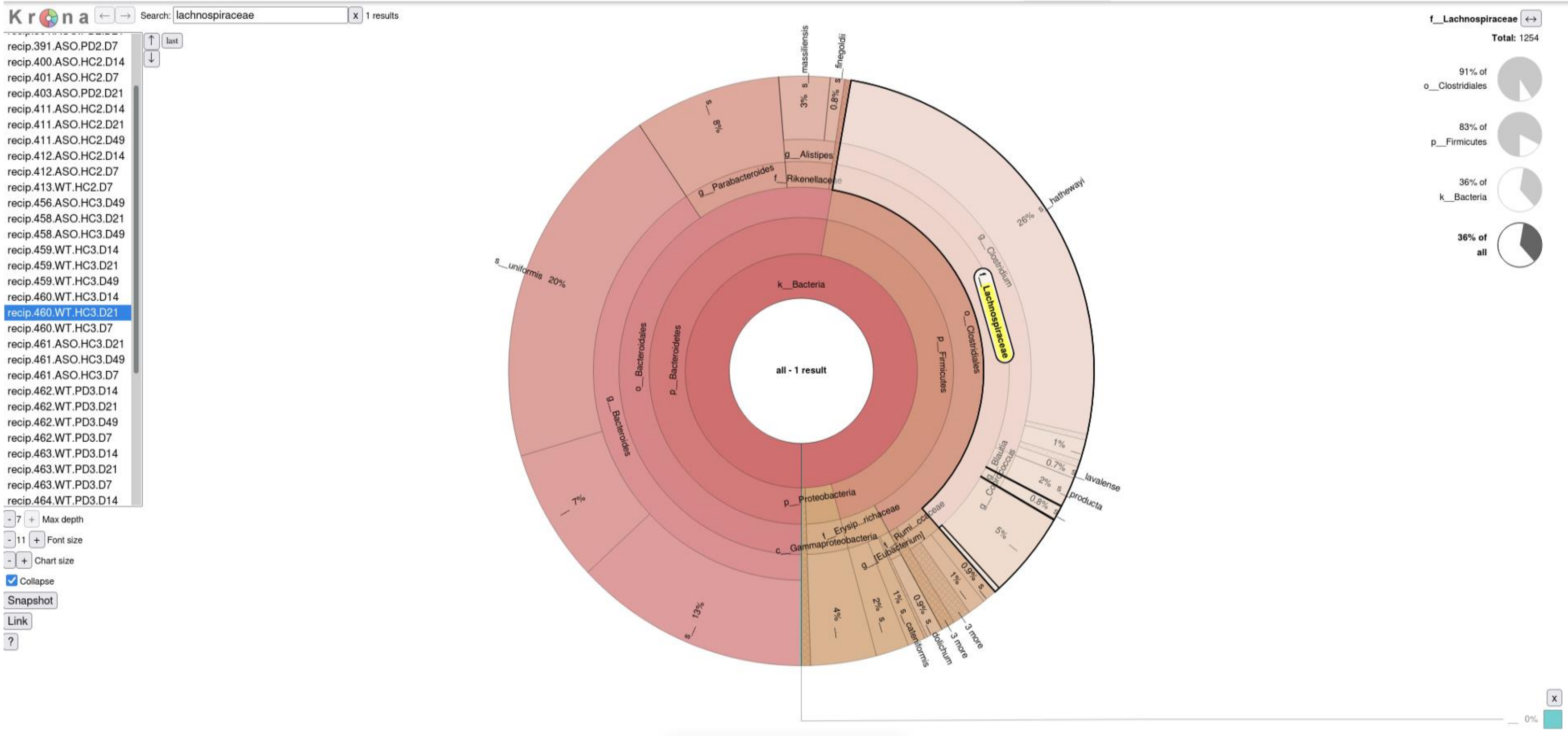


Figure 21: Barplots results for PD study (Sampson et al., 2016). Barplots are powerful at visualizing whole study. Krona plots are powerful at visualizing a single sample.



Search: lachnospiraceae x 1 results

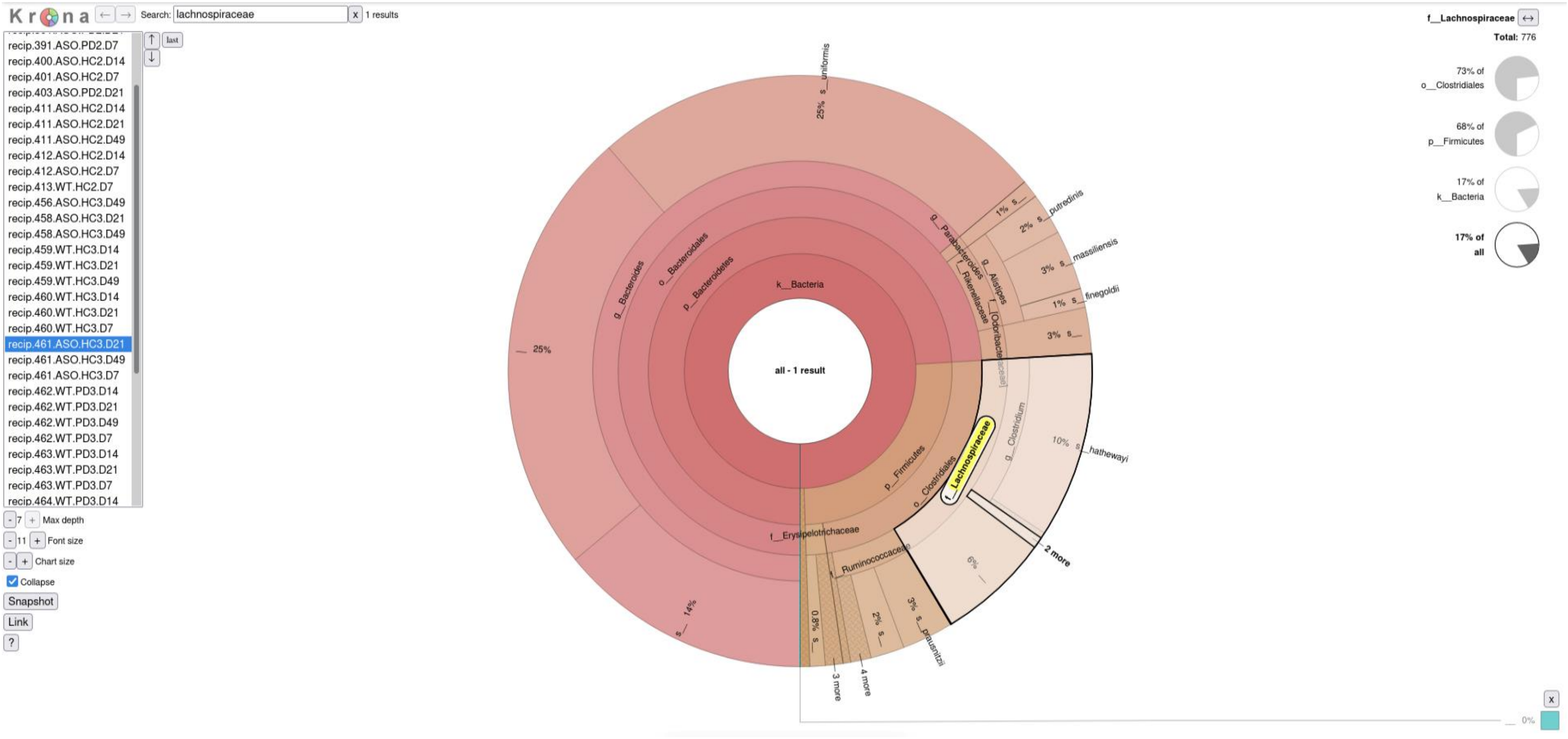
recip.391.ASO.PD2.D7
recip.400.ASO.HC2.D14
recip.401.ASO.HC2.D7
recip.403.ASO.PD2.D21
recip.411.ASO.HC2.D14
recip.411.ASO.HC2.D21
recip.411.ASO.HC2.D49
recip.412.ASO.HC2.D14
recip.412.ASO.HC2.D7
recip.413.WT.HC2.D7
recip.456.ASO.HC3.D49
recip.458.ASO.HC3.D21
recip.458.ASO.HC3.D49
recip.459.WT.HC3.D14
recip.459.WT.HC3.D21
recip.459.WT.HC3.D49
recip.460.WT.HC3.D14
recip.460.WT.HC3.D21
recip.460.WT.HC3.D7
recip.461.ASO.HC3.D21
recip.461.ASO.HC3.D49
recip.461.ASO.HC3.D7
recip.462.WT.PD3.D14
recip.462.WT.PD3.D21
recip.462.WT.PD3.D49
recip.462.WT.PD3.D7
recip.463.WT.PD3.D14
recip.463.WT.PD3.D21
recip.463.WT.PD3.D7
recip.464.WT.PD3.D14

- 7 + Max depth
- 11 + Font size
- + Chart size
☑ Collapse

f__Lachnospiraceae
Total: 1254

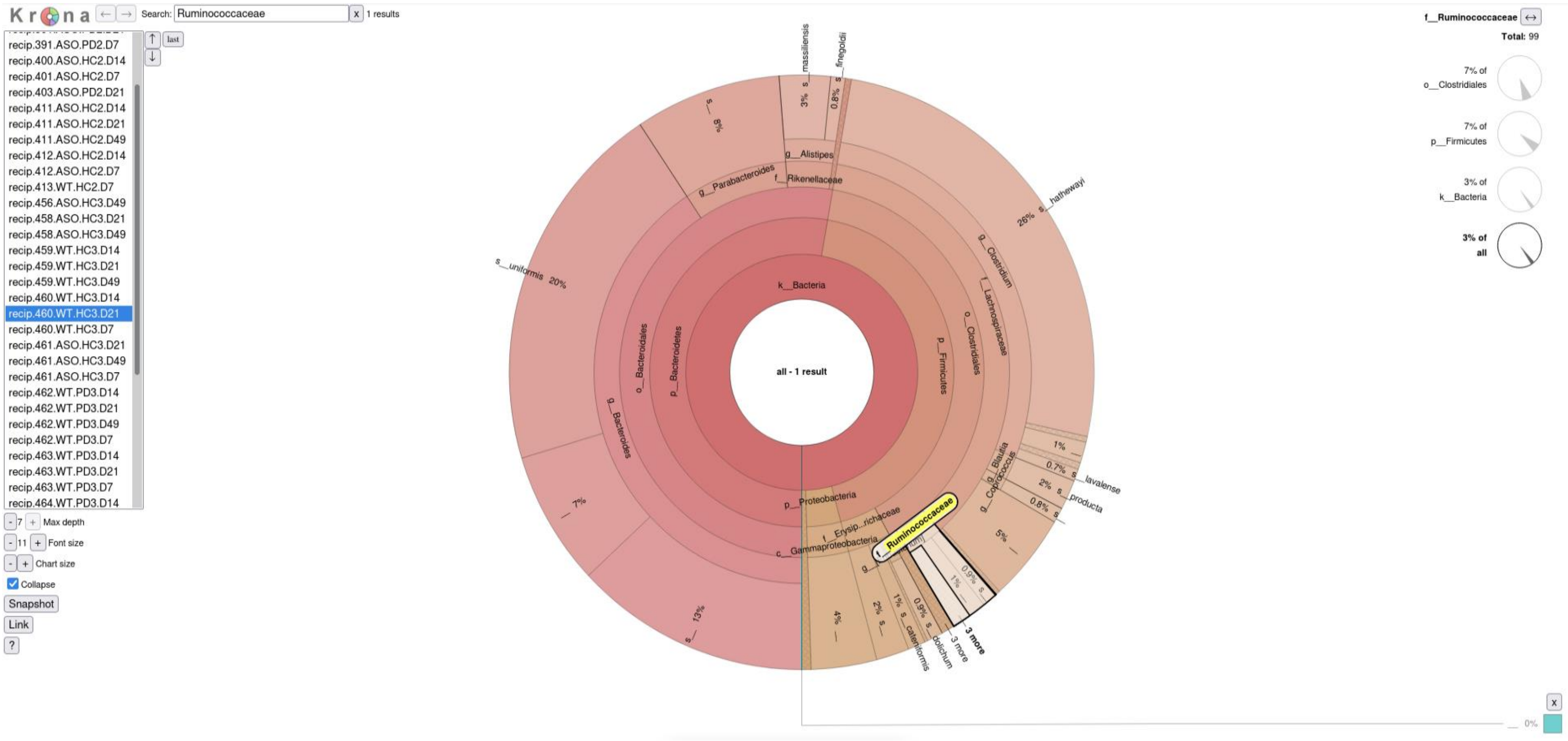
91% of o__Clostridiales
83% of p__Firmicutes
36% of k__Bacteria
36% of all

Figure 22: *Lachnospiraceae* search result of an example wild-type mouse



recip.391.ASO.PD2.D7
recip.400.ASO.HC2.D14
recip.401.ASO.HC2.D7
recip.403.ASO.PD2.D21
recip.411.ASO.HC2.D14
recip.411.ASO.HC2.D21
recip.411.ASO.HC2.D49
recip.412.ASO.HC2.D14
recip.412.ASO.HC2.D7
recip.413.WT.HC2.D7
recip.456.ASO.HC3.D49
recip.458.ASO.HC3.D21
recip.458.ASO.HC3.D49
recip.459.WT.HC3.D14
recip.459.WT.HC3.D21
recip.459.WT.HC3.D49
recip.460.WT.HC3.D14
recip.460.WT.HC3.D21
recip.460.WT.HC3.D7
recip.461.ASO.HC3.D21
recip.461.ASO.HC3.D49
recip.461.ASO.HC3.D7
recip.462.WT.PD3.D14
recip.462.WT.PD3.D21
recip.462.WT.PD3.D49
recip.462.WT.PD3.D7
recip.463.WT.PD3.D14
recip.463.WT.PD3.D21
recip.463.WT.PD3.D7
recip.464.WT.PD3.D14

Figure 23: *Lachnospiraceae* search result of an example susceptible mouse



Search: Ruminococcaceae x 1 results

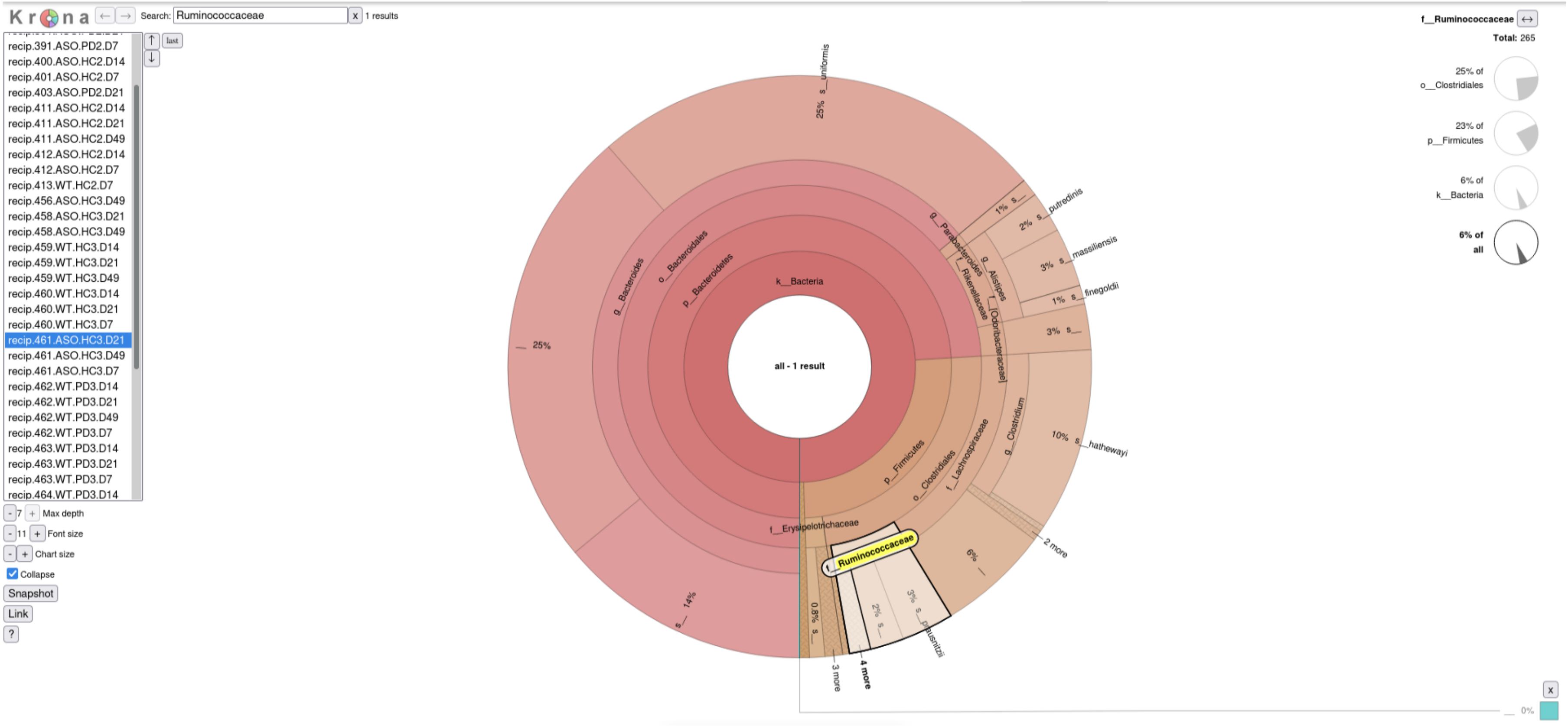
recip.391.ASO.PD2.D7
recip.400.ASO.HC2.D14
recip.401.ASO.HC2.D7
recip.403.ASO.PD2.D21
recip.411.ASO.HC2.D14
recip.411.ASO.HC2.D21
recip.411.ASO.HC2.D49
recip.412.ASO.HC2.D14
recip.412.ASO.HC2.D7
recip.413.WT.HC2.D7
recip.456.ASO.HC3.D49
recip.458.ASO.HC3.D21
recip.458.ASO.HC3.D49
recip.459.WT.HC3.D14
recip.459.WT.HC3.D21
recip.459.WT.HC3.D49
recip.460.WT.HC3.D14
recip.460.WT.HC3.D21
recip.460.WT.HC3.D7
recip.461.ASO.HC3.D21
recip.461.ASO.HC3.D49
recip.461.ASO.HC3.D7
recip.462.WT.PD3.D14
recip.462.WT.PD3.D21
recip.462.WT.PD3.D49
recip.462.WT.PD3.D7
recip.463.WT.PD3.D14
recip.463.WT.PD3.D21
recip.463.WT.PD3.D7
recip.464.WT.PD3.D14

7 Max depth
11 Font size
Chart size
☑ Collapse

f__Ruminococcaceae
Total: 99

7% of o__Clostridiales
7% of p__Firmicutes
3% of k__Bacteria
3% of all

Figure 24: *Ruminococceae* search result of an example wild-type mouse



Search: Ruminococcaceae x 1 results

recip.391.ASO.PD2.D7
recip.400.ASO.HC2.D14
recip.401.ASO.HC2.D7
recip.403.ASO.PD2.D21
recip.411.ASO.HC2.D14
recip.411.ASO.HC2.D21
recip.411.ASO.HC2.D49
recip.412.ASO.HC2.D14
recip.412.ASO.HC2.D7
recip.413.WT.HC2.D7
recip.456.ASO.HC3.D49
recip.458.ASO.HC3.D21
recip.458.ASO.HC3.D49
recip.459.WT.HC3.D14
recip.459.WT.HC3.D21
recip.459.WT.HC3.D49
recip.460.WT.HC3.D14
recip.460.WT.HC3.D21
recip.460.WT.HC3.D7
recip.461.ASO.HC3.D21
recip.461.ASO.HC3.D49
recip.461.ASO.HC3.D7
recip.462.WT.PD3.D14
recip.462.WT.PD3.D21
recip.462.WT.PD3.D49
recip.462.WT.PD3.D7
recip.463.WT.PD3.D14
recip.463.WT.PD3.D21
recip.463.WT.PD3.D7
recip.464.WT.PD3.D14

- 7 + Max depth
- 11 + Font size
- + Chart size
☑ Collapse

f__Ruminococcaceae
Total: 265

25% of o__Clostridiales
23% of p__Firmicutes
6% of k__Bacteria
6% of all

Figure 25: *Ruminococceae* search result of an example susceptible mouse

4.2.2 Case 2: Sewage Treatment Systems Study

The study is concluded to reveal the microbial composition of wastewater treatment systems to support future improvements. Wastewater treatment systems of a rural school and a factory are compared by analysis of 16S rRNA gene amplicon data (Hidalgo et al., 2021).

In the study, sludge samples were collected from the septic tank and the anaerobic filter of the wastewater treatment systems in duplicate and different seasons of the year. The study analyses and compares the predominant genera as a result of taxonomic analysis without visualization supporting the results.

> *"Meanwhile, Acinetobacter (7.7%), Smithella (4.4%), Arcobacter (2.9%), Methanobacterium (2.89%), and Syntrophobacter (2.4%) were the top five most abundant genera in the rural school. Acinetobacter has been related to fecal contamination (McLellan et al. 2010)."* (Hidalgo et al., 2021)

As for the PD study, open-access data is obtained with the study accession PRJEB36453, and bioinformatic analyses are performed with QIIME 2 described in the study. q2-krona plugin is used for visualization of taxonomic analysis.

Results of bar plots and searching *Acinetobacter* in Krona plots are shown in the figures for a rural school's anaerobic filter and septic tank in Autumn 2018.

In this study, the bar plots are much more crowded with information than PD study as shown in Figure 26. Repetition of colors makes it very unclear, and it is pretty hard to inspect a single sample however it is possible to get a picture of the whole study.

In Figure 27 and 28, Krona plots of sample AFP1-Autumn18 and STP1-Autumn18 are shown. With the power of the search function and interactivity, it is much more easier to navigate through the data.

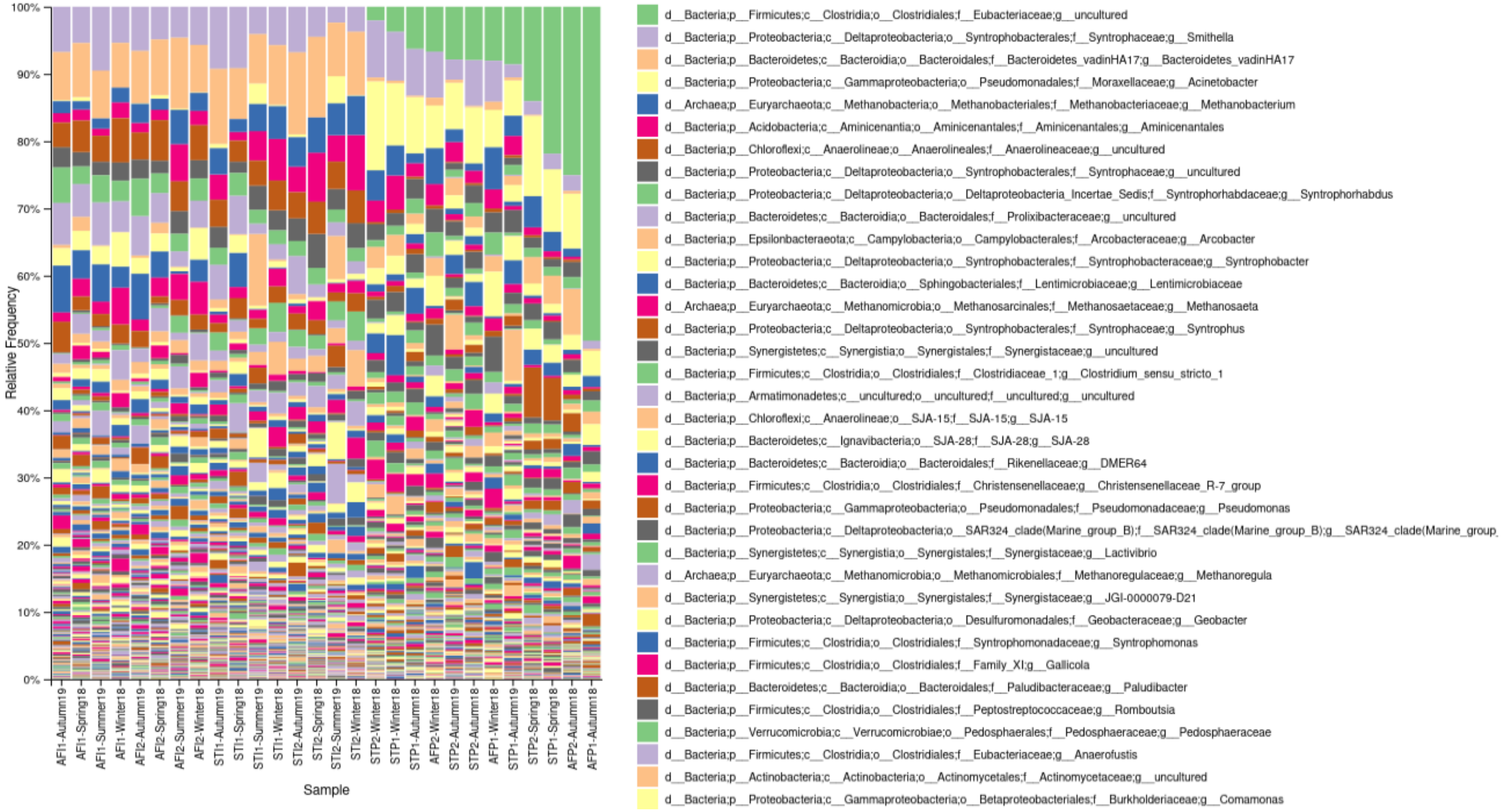


Figure 26: Barplot results of sewage study

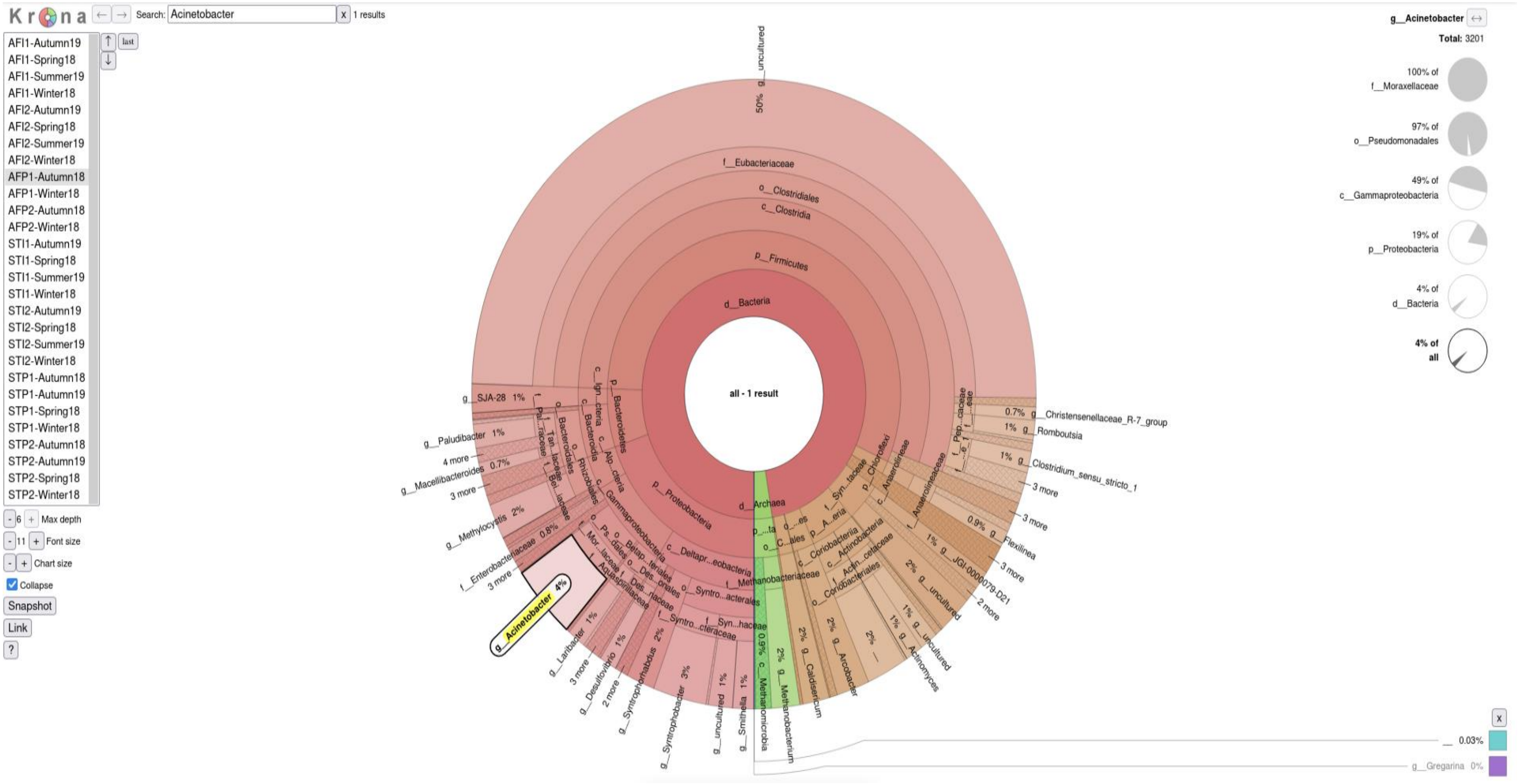


Figure 27: *Acinetobacter* search result of the anaerobic filter of rural school in Autumn 2018

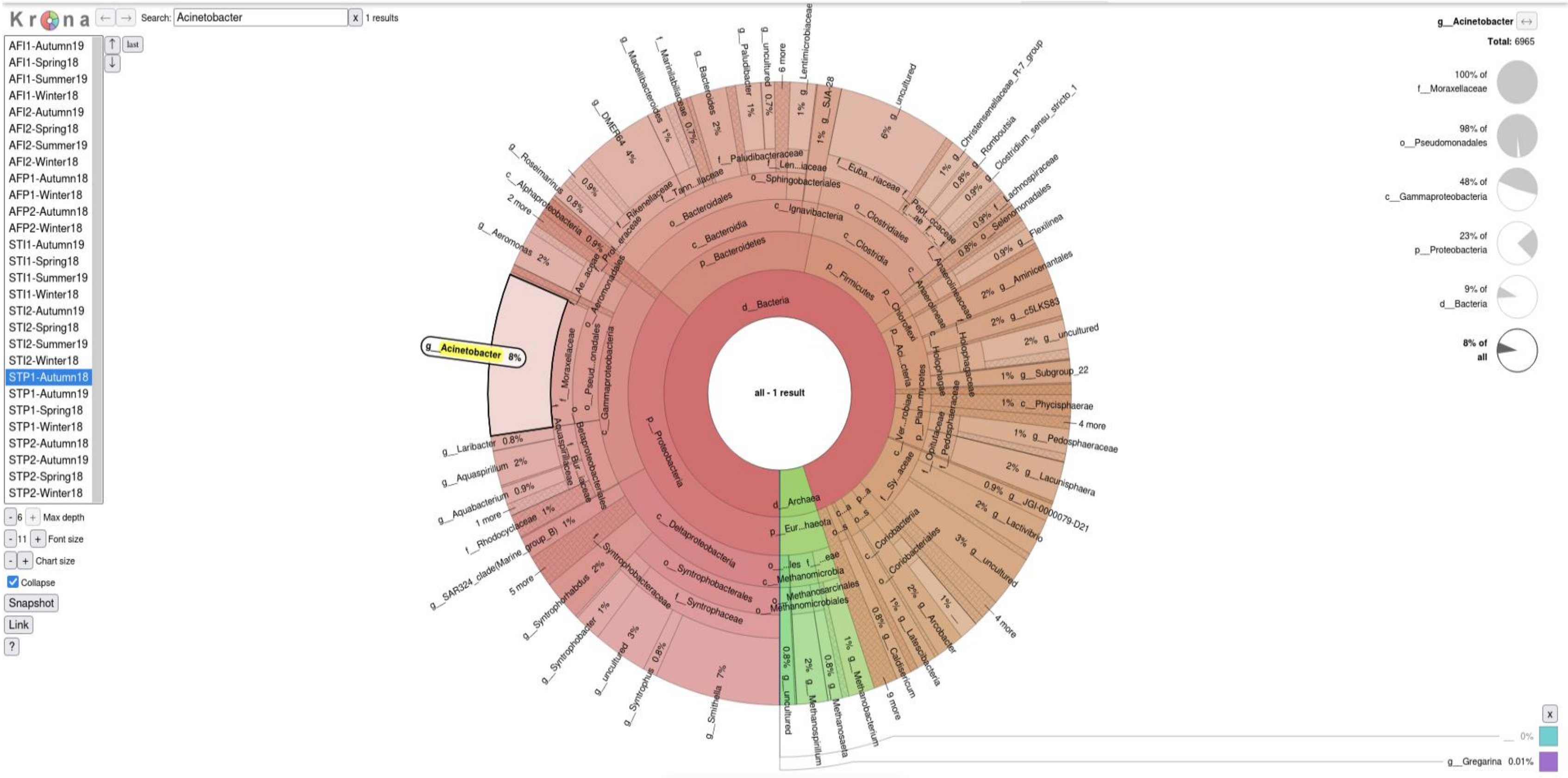


Figure 28: *Acinetobacter* search result of the septic tank of rural school in Autumn 2018

CHAPTER 5

DISCUSSION

Microbiome analysis is a new emerging topic that is getting attention from different areas such as medicine, ecology, environmental research, food, and wastewater treatment. The techniques available are comparably new and open for development.

Several tools are available for analysis and visualization, but Mothur and QIIME 2 are the most commonly used ones. Both can start with raw sequences and end with visualizations. Usually, there is no need for another tool; Mothur and QIIME 2 include their packages to manage the data for all analysis steps.

Increasingly, researchers from many different areas have started to utilize microbiome analysis in their studies, making user-friendliness one of the essential features. Dynamic figures have also become standard since microbiome studies have crowded data and messy figures. Interactive methods help users a lot in this manner. So, the main preferred features of the tools and visualizations can be summed to be user-friendly and draw dynamic figures.

The bar plots are the only visualization method available on QIIME 2 for taxonomic classifications. It is not as easy as Krona to check for each sample and taxa because the bar plots are designed to plot all the data gathered from the study and do not report at the individual sample level.

Considering the needs during visualization, a plugin for the visualization of microbial study results is developed, which is compatible with the QIIME 2 tool. The final product is available on https://github.com/kaanb93/q2-krona. The plugin consists of one method and one pipeline callable with QIIME 2 Client. The plugin is easy to install and open source. The plugin was released in November 2021. The developers of QIIME 2 have checked the plugin, and no issues have been reported yet.

We evaluated the proposed plugin's usefulness with two different case studies, where visualization of the presented results was missing. The published results were discussed at a family level in the first case study, which compared the microbiome of mice susceptible and resistant to PD. Also, in the second case study where samples were analyzed, the discussions are focused on genera. In both studies, the bar plots were generated according to the required level, and genera level plots were much more crowded than family level plots. These representations make it pretty hard to get a clean image of the discussed results.

For PD study, with q2-krona, the abundance of the family members discussed in taxonomic classification is made possible to visualize clearly. As for the sewage systems treatment study, it is easier to detect contamination levels in abundant taxa.

Reviewing the case studies shows that q2-krona presents a more observable output with the quick inclusion of the plugin. The main advantages of Krona are:

- It has a search option included. So, it is easier to navigate through crowded data.
- The plots are interactive, making it easier to visualize the smaller portions of the plots by zooming in.
- The results are shown for every sample individually.
- Krona uses a continuous color palette. The color palette and hierarchical plots make it much easier to distinguish each taxa.

Several features can be improved in the current version of our pipeline. Krona has a unique color scheme; users need to edit the output file to change it. This part can be changed to become much more user-friendly because editing the output file for such a task involves changing the values of multiple fields in the file.

There are other methods for this task in Krona, but there is no option for the “ktImportText” method. There is already an issue reported on the GitHub page of Krona (https://github.com/marbl/Krona/issues/147) about color blindness, which would be a potential enhancement to Krona itself. So, instead of developing a new implementation for Krona, q2-krona can be improved to include such an option.

Overall we have filled a gap by developing a plugin that will transfer QIIME2 outputs into Krona to ease the visualization of QIIME 2 analysis results with Krona without additional script or manipulation during amplicon analysis. Also showed the benefits of using the plugin in two case studies.

REFERENCES

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., & Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *MSystems, 2*(2). https://doi.org/10.1128/msystems.00191-16

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M. C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., … Schloter, M. (2020). Microbiome definition re-visited: old concepts and new challenges. In *Microbiome* (Vol. 8, Issue 1). https://doi.org/10.1186/s40168-020-00875-0

Bodilis, J., Nsigue-Meilo, S., Besaury, L., & Quillet, L. (2012). Variable copy number, intra-genomic heterogeneities and lateral transfers of the 16S rRNA gene in Pseudomonas. *PLoS ONE, 7*(4). https://doi.org/10.1371/journal.pone.0035647

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome, 6*(1). https://doi.org/10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Author Correction: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 (Nature Biotechnology, (2019), 37, 8, (852-857), 10.1038/s41587-019-0209-9). In *Nature Biotechnology* (Vol. 37, Issue 9). https://doi.org/10.1038/s41587-019-0252-6

Bruijns, B., Tiggelaar, R., & Gardeniers, H. (2018). Massively parallel sequencing techniques for forensics: A review. In *Electrophoresis* (Vol. 39, Issue 21). https://doi.org/10.1002/elps.201800082

Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L., & Leddy, M. B. (2020). Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLoS ONE, 15*(2). https://doi.org/10.1371/journal.pone.0228899

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7). https://doi.org/10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pẽa, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. In *Nature Methods* (Vol. 7, Issue 5). https://doi.org/10.1038/nmeth.f.303

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6). https://doi.org/10.1093/nar/gkp1137

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7). https://doi.org/10.1128/AEM.03006-05

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., & Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, *2*(1). https://doi.org/10.1186/2049-2618-2-6

Garrido-Cardenas, J. A., & Manzano-Agugliaro, F. (2017). The metagenomics worldwide research. In *Current Genetics* (Vol. 63, Issue 5). https://doi.org/10.1007/s00294-017-0693-8

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. v., & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine*, *24*(4). https://doi.org/10.1038/nm.4517

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. In *Genomics* (Vol. 107, Issue 1). https://doi.org/10.1016/j.ygeno.2015.11.003

Hidalgo, K. J., Saito, T., Silva, R. S., Delforno, T. P., Duarte, I. C. S., de Oliveira, V. M., & Okada, D. Y. (2021). Microbiome taxonomic and functional profiles of two domestic sewage treatment systems. *Biodegradation*, *32*(1). https://doi.org/10.1007/s10532-020-09921-y

Jünemann, S., Kleinbölting, N., Jaenicke, S., Henke, C., Hassa, J., Nelkner, J., Stolze, Y., Albaum, S. P., Schlüter, A., Goesmann, A., Sczyrba, A., & Stoye, J. (2017). Bioinformatics for NGS-based metagenomics and the application to biogas research.

In *Journal of Biotechnology* (Vol. 261). https://doi.org/10.1016/j.jbiotec.2017.08.012

Kõljalg, U., Larsson, K. H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A. F. S., Tedersoo, L., Vrålstad, T., & Ursing, B. M. (2005). UNITE: A database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist*, *166*(3). https://doi.org/10.1111/j.1469-8137.2005.01376.x

Liu, Y. X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., & Bai, Y. (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. In *Protein and Cell* (Vol. 12, Issue 5). https://doi.org/10.1007/s13238-020-00724-8

Mandal, R. K., Jiang, T., Al-Rubaye, A. A., Rhoads, D. D., Wideman, R. F., Zhao, J., Pevzner, I., & Kwon, Y. M. (2016). An investigation into blood microbiota and its potential association with Bacterial Chondronecrosis with Osteomyelitis (BCO) in Broilers. *Scientific Reports*, *6*. https://doi.org/10.1038/srep25882

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R., & Caporaso, J. G. (2012). The Biological Observation Matrix (BIOM) format or: How I learned to stop worrying and love the ome-ome. In *GigaScience* (Vol. 464, Issue 1). https://doi.org/10.1186/2047-217X-1-7

Nyrén, P., Pettersson, B., & Uhlén, M. (1993). Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical Biochemistry*, *208*(1). https://doi.org/10.1006/abio.1993.1024

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-385

Preheim, S. P., Perrott, A. R., Martin-Platero, A. M., Gupta, A., & Alm, E. J. (2013). Distribution-based clustering: Using ecology to refine the operational taxonomic unit. *Applied and Environmental Microbiology*, *79*(21). https://doi.org/10.1128/AEM.00342-13

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS ONE*, *15*(1). https://doi.org/10.1371/journal.pone.0227434

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1). https://doi.org/10.1093/nar/gks1219

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *2016*(10). https://doi.org/10.7717/peerj.2584

Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., Challis, C., Schretter, C. E., Rocha, S., Gradinaru, V., Chesselet, M. F., Keshavarzian, A., Shannon, K. M., Krajmalnik-Brown, R., Wittung-Stafshede, P., Knight, R., & Mazmanian, S. K. (2016). Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. *Cell*, *167*(6). https://doi.org/10.1016/j.cell.2016.11.018

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23). https://doi.org/10.1128/AEM.01541-09

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. In *Nature Biotechnology* (Vol. 26, Issue 10). https://doi.org/10.1038/nbt1486

Srivastava, S. (2014). Genetics of bacteria. In *Genetics of Bacteria*. https://doi.org/10.1007/978-81-322-1090-0

Turner, F. S. (2014). Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries. *Frontiers in Genetics*, *5*(JAN). https://doi.org/10.3389/fgene.2014.00005

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. In *Nature Biotechnology* (Vol. 39, Issue 11). https://doi.org/10.1038/s41587-021-01108-x

Yilmaz, A., Ilke, H., Alp, E., & Menevse, S. (2012). Real-Time PCR for Gene Expression Analysis. In *Polymerase Chain Reaction*. https://doi.org/10.5772/37356

Zhu, H., Zhang, H., Xu, Y., Laššáková, S., Korabečná, M., & Neužil, P. (2020). PCR past, present and future. In *BioTechniques* (Vol. 69, Issue 4). https://doi.org/10.2144/BTN-2020-0057

APPENDICES

APPENDIX A

LIST OF AVAILABLE QIIME 2 PLUGINS

Core plugins

q2-alignment: Plugin for generating and manipulating alignments.

q2-composition: Plugin for compositional data analysis.

q2-cutadapt: Plugin for removing adapter sequences, primers, and other unwanted sequence from sequence data.

q2-data2: Plugin for sequence quality control with DADA2.

q2-deblur: Plugin for sequence quality control with Deblur.

q2-demux: Plugin for demultiplexing & viewing sequence quality.

q2-diversity: Plugin for exploring community diversity.

q2-diversity-lib: Plugin for computing community diversity.

q2-emperor: Plugin for ordination plotting with Emperor.

q2-feature-classifier: QIIME 2 plugin for taxonomic classification of sequences.

q2-feature-table: Plugin for working with sample by feature tables.

q2-fragment-insertion: QIIME 2 plugin for insertion of fragments to reference taxonomies by using SEPP.

q2-gneiss: Plugin for building compositional models.

q2-longitudinal: Plugin for paired sample and time series analyses.

q2-metadata: Plugin for working with Metadata.

q2-phylogeny: Plugin for generating and manipulating phylogenies.

q2-quality-control: Plugin for quality control of feature and sequence data.

q2-quality-filter: Plugin for PHRED-based filtering and trimming.

q2-sample-classifier: Plugin for machine learning prediction of sample metadata.

q2-taxa: Plugin for working with feature taxonomy annotations.

q2-types: Plugin defining types for microbiome analysis.

q2-vsearch: Plugin for clustering and dereplicating with vsearch.

Others

DEICODE: Robust Aitchison PCA for sparse omics datasets, linking specific features to beta-diversity ordination through the use of compositional biplots.

EMPress: A fast and scalable phylogenetic tree viewer.

evident: Effect size and power calculations for microbiome diversity data.

gemelli: Gemelli is a toolbox for running tensor factorization on sparse compositional omics datasets.

mmvec: A software package for learning microbe-metabolite interactions.

q2-aldex2: Compositional differential abundance analysis. ALDEx2 provides a framework that encompasses essentially all high-throughput sequencing data types by modelling the data as a log-ratio transformed probability distribution rather than as counts.

q2-breakaway: `breakaway` is the premier package for statistical analysis of microbial diversity. `breakaway` implements the latest and greatest estimates of richness, as well as the most commonly used estimates. The `breakaway` philosophy is to estimate diversity, to put error bars on diversity estimates, and to perform hypothesis tests for diversity that use those error bars.

q2-clawback: Assembles taxonomic weights to increase classification accuracy with q2-feature-classifier. Can download data from Qiita or use your data.

q2-coordinates: A qiime2 plugin supporting methods for geographic mapping of qiime2 artifact data or metadata.

q2-coremicrobiome: Qiime2 plugin of COREMIC

q2-data-augment: Data augmentation is a very useful and widely used method in data science. Especially, it can increase the sample size of the training set for machine learning models. Rarefy for Augment uses a simple rarefaction method to achieve data augmentation. The data augmentation should only be implemented on the training data set.

q2-dbBact: dbBact term enrichment and wordcloud generation for amplicon experiment analysis.

q2-dbotu: QIIME 2 plugin for distribution-based clustering, which calls OTUs based on the similarity between their genetic sequences and their cound distribution across samples.

q2-fondue: Plugin for acquisition, re-use, and management of public nucleotide sequence (meta)data and while adhering to open data principles.

q2-gcn-norm: QIIME 2 plugin for normalizing sequences by 16S rRNA gene copy number (GCN) based on rrnDB database.

q2-health-index: QIIME 2 plugin for calculating the Health Index from microbiome data.

q2-ili: The plugin wrapping for *ili*.

q2-itsxpress: ITSxpress trims the conserved flanking regions of ITS sequences. ITSxpress is designed to support the calling of exact sequence variants rather than OTUs.

q2-krona: Plugin for creating Krona plots.

q2-metabolomics: q2-metabolomics is a tool to import metabolomics data into qiime2 to perform analysis.

q2-metaphlan2: MetaPhlAn is a computational tool for profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes, and Viruses) from metagenomic shotgun sequencing data with species-level resolution.

q2-micom: q2-micom provides an interface between Qiime 2 and MICOM. It allows creating and simulating metagenome-scale metabolic commuity models and predicting metabolic fluxes in a microbial consortium. It can incorporate abundance and environmental data.

q2-perc-norm: QIIME 2 plugin for percentile normalization to correct for batch effects in microbiome case-control studies.

q2-phylogenize: q2-phylogenize allows users to link microbial genes to environments using phylogenetic regression. Phylogenetic regression considers that species that share ancestry will often share phenotypic traits, including colonizing similar environments. This confounder can otherwise lead to high rates of spurious associations.

q2-picrust2: Plugin to run PICRUSt2 pipeline to get EC, KO, and MetaCyc pathway predictions based on 16S data. Either EPA-NG or SEPP can be used to place sequences into the required reference phylogeny.

q2-protein-pca: QIIME 2 plugin for principal component analysis of protein sequences, based on ranked multiple sequence alignments.

q2-repeat-rarefy: QIIME 2 plugin for generating the average rarefied table for library size normalization using repeated rarefaction. This simple "Average Rarefied Table" method can be an ideal alternative to the current one-shot rarefaction, as it can keep information and avoid variation of composition.

q2-sample-classifier: QIIME 2 plugin for machine learning prediction of sample data.

q2-SCNIC: SCNIC (Sparse Cooccurence Network Investigation for Compositional data) is a tools for building correlation networks from feature tables, finding modules in said networks and summarizing those modules.

q2-shogun: A QIIME 2 plugin wrapper for the SHOGUN shallow shotgun sequencing taxonomy profiler.

q2-sidle: The Short Multiple Reads Framework algorithm allows the scaffolding of multiple marker gene regions against a reference database;

q2-srs: QIIME 2 plugin for microbiome count data normalization by scaling with ranked subsampling (SRS).

q2-woltka: A versatile classifier of composition and function of metagenomes.

Qiime2 manifest & metadata generator: Generate metadata and manifest file for Qiime2 input.

Qurro: Interactively visualize feature rankings (differentials or feature loadings, sorted numerically) alongside the log-ratios of selected features' abundances.

RESCRIPt: Reference Sequence annotation and CuRatIon Pipeline RESCRIPt is a QIIME 2 plugin to support various operations for managing and curating reference sequence databases, DNA/RNA sequence data, and taxonomic data.