A 'Big Data' Architecture for Reproducibility in Microbiome research: Case Study on Diagnosis of Inflammatory Bowel Disease from Microbiome Data

Final Report

Billy Thornton

December 15, 2022

Abstract

Metagenomic sequencing has instigated a wave of new insights into the complexity and impact of micobial ecosystems in the gut through large, complex datasets. As these datasets have become more readily available understanding of the component microbes and the statistical power of potential analytics has grown and the potential insights grow further. This, however, raises a problem that all published results are version sensitive as they derived from reference datasets and analysis pipelines that are subject to regular change. This project provides a standardised and scalable computing architecture to enable formal links between analytical outcomes and their analysis pipelines to provide reproducibility, repeatability and replicability in microbiome analyses. This system is demonstrated through a case study reproducing a model for diagnosis of inflammatory bowel diseases with current data and showing the impact of changes between historical database versions on the diagnostic outcomes from the study.

I certify that all material in this dissertation which is not my own work has been identified.

Contents

1 Introduction

The human gut microbiota is the ecosystem of microbes that reside in the gut. This complex microbial ecology has broad impacts on health and contributes to multiple common diseases. The emergence of metagenomics, high-throughput sequencing of environmental samples, has facilitated many new insights into the biology of the microbiota in health and disease by providing large, complex metagenomics datasets. Microbiome research, has provided new insights into human biology and been labelled one of the new frontiers in life sciences although there are significant 'big data' challenges to support the reproducibility of research in this area.

The number of publicly available sequenced samples of the human microbiome has grown rapidly over recent years. In February 2021 NCBIs GenBank held 226,241,476 sequences an increase of 10,027,261 from the previous year, this number has doubled approximately every 18 months since the GenBanks inception in 1982 [**?**]. This large repository of pre-existing data may hold solutions to may rising health problems in the world today [1]. One such problem is the increasing number of cases of complex conditions such as Inflammatory Bowel Disease (IBD) [2–5].

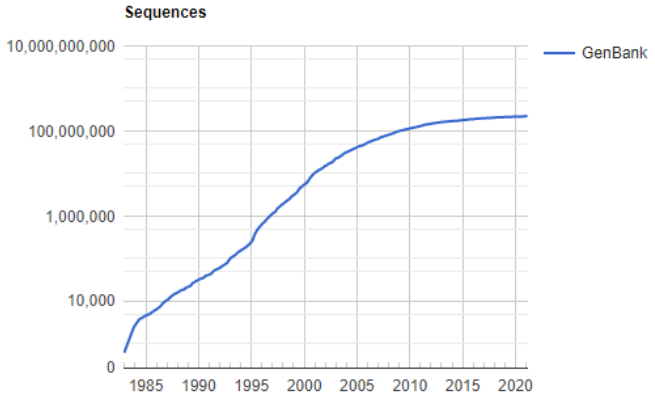


Figure 1: Growth of genetic sequences in the NCBI GenBank database demonstrating the doubling of the number of samples approximately every 18 months, Source - [**?**]

1.1 Diagnosis of IBD from Microbiome Data

IBD can be separated into two conditions Crohn's Disease (CD) and Ulcerative Colitis (UC). It has been estimated that the number of CD and UC cases in the UK rose by 81% and 46% respectively between 2000 and 2017 [6] and currently has over 6.8 million cases globally [7]. This is not only problematic for those affected but also the healthcare systems that support them with the average yearly cost of treatment per patient being €541 [8].

It often takes a long time to reach a diagnosis for these conditions, in one Swiss cohort there was an average diagnostic delay of 9 months from the initial symptom onset, with 25% having a delay of over 2 years [9]. It has been shown that this diagnostic delay is linked to poor clinical outcomes such as increased requirements for intestinal surgery [10].

Both diseases have a complex pathogenesis that isn't completely understood [11] which leads to difficulty differentiating the conditions [12, 13]. This is especially true when considering the role of the microbiota in perturbing the mucosa [14]. However, if we can detect patterns in the microbiome between UC and CD it may inform our understanding of the pathogenesis which would facilitate the development of better targeted medical interventions and provide new methods of diagnosis.

Gut microbiome data contains information about the composition of an individuals gut microbiota. This microbiome has been shown to be a contributing factor to the pathogenesis of IBD with diseased individuals presenting with reduced organism diversity and disproportional abundances of specific taxa [4, 15]. However components of the microbiome are highly dynamic and can change as a result of many factors [16–19]. This volatility presents challenges when attempting to identify factors that consistently contribute to each condition.

1.2 Microbiome Data and Biomarkers

A biomarker is a molecular signature that indicates the presence of a condition [20], an example would be the spike in insulin levels observed in those with diabetes. The biomarkers for IBD are significantly more nuanced and cannot often be reliant on signatures present in the blood. However the patterns present in microbiome datasets may provide a new avenue for biomarker discovery.

Microbiome datasets consist of raw sequence reads which are collected through the use of sequencing platforms. Raw sequences are represented as strings consisting of the nucleotide bases that are present in the sequence, Figure 2 provides an example of raw sequence data. These datasets are analysed using bioinformatics toolchains and pipelines implementing a wide variety of computational methodologies. The analysis process as a whole can be differentiated into 2 stages (show in figure 3).

- **Primary Analysis:** Processing the raw reads into a use able format by annotating them with taxonomic information. This is a multistage process involving demultiplexing, quality control, denoising and taxonomic classification, which results in lists of named species that were present in the biomes captured by the raw reads.
- **Secondary Analysis:** Taking the annotated species information and performing high level analysis. This is where the statistical methods are heavily used in order to derive understanding from the annotated species created during primary analysis. This stage is often focused on answering a question about the microbiome captured in the raw sequencing data. In the case of this project it is to evaluate its diagnostic potential.

```
1 @SRR578290.1 GNKV64B01BLSUE/4
2 TCTCATCCTGCTGCCTCCGTAGGCTGAGACTGCCAAGGCACACAGGGATAGGN
3 +
4 IIIIE???EEIIIEBB<<?IECCECIIIIEEEGICCCCIIIIIIFFHIIIGG!
```

Figure 2: An example of a raw read. Each read consists of 4 lines, Line 2 represents the sequence data

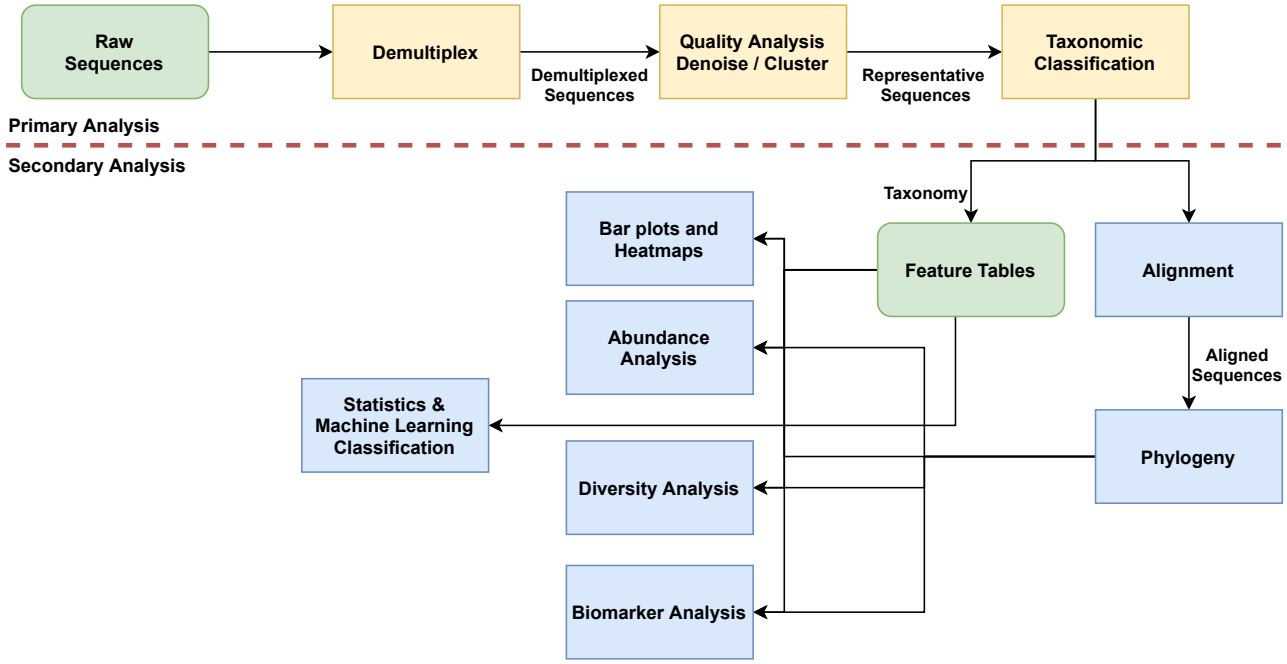


Figure 3: Overview of the key stages when processing sequence data with QIIME 2. Green denote data objects, yellow are primary analysis stages and blue are secondary analysis stages

1.3 Data Ecosystem

Taxonomic classification is a key stage of primary analysis, whose goal is to assign a taxonomic annotation (name and classification) to the raw sequences. This can be done in many ways but it is

common to use reference databases. These databases contain a number of pre annotated taxa. Using statistical methods the pipeline determines which taxa in the database is most similar to the raw sequence data, thus assigning a taxonomic annotation.

These datasets and indeed the analysis tools themselves are part of a data eco-system which are subject to regular change. These changes include discovery of new taxa or alterations to specific tool implementations. The changes will effect how samples are classified and impact already published results [21] this is especially true for classifications at the genus level and below [22]. This impact will alter the inferred representation of gut microbiomes and our understanding of them which could invalidate all published historical studies. However it is unclear how this may impact our understanding of the underlying biology and the derived diagnostic classifications.

1.4 Dynamic Data - Repeatability, Reproducibility and Replicability

The goal of **repeatability** is to ensure that the same results can be achieved on multiple trials or iterations if the same workflow and data is used. By ensuring high levels of repeatability results can be used as the basis for further investigation as they are not effected by runtime variations. This is especially beneficial when handling computationally expensive microbiome data and methods as it allows results to be used in later stages without reprocessing. As an example if the primary analysis stage of the computing architecture is repeatable then it should not need to be reprocessed when changes are made to the secondary stage. This allows the architecture to vastly reduce the amount of computation required.

Low levels of **reproducibility** are a large problem within scientific research [23] and microbiome research is no exception. A study is deemed reproducible if the same results can be obtained using the data and methods described in a study [24]. Reproducibility can effect microbiome research at multiple stages, both in the lab and in the in-silico downstream analysis.

Within the lab issues of reproducibility focus on the handling of samples, ensuring no contamination and processing using the same techniques. It has been shown that changes in these conditions can lead to differences in measurements that are unrelated to the biology, this so-called "batch effect" is a problem in microbiome studies that can prevent the generalization of findings [25–27]. Downstream computational analysis contains additional hurdles to reproducibility with knowledge of the software stack, parameters, datasets and associated metadata required to accurately reproduce results. However this information is rarely included [27] with studies and so the results become nearly impossible to reproduce.

A study is deemed **replicable** if the same outcomes can be obtained by generating new results with new data whilst analysing with the same methods. In order for high levels of replicability to be achieved studies must be both repeatable and reproducible whilst providing enough information about the methods used to allow for changes to be made by new research groups where appropriate, such as changing the microbiome samples being analysed.

These hurdles to reproducibility along with low repeatability means that the results of these studies become static, unable to be used in any real way beyond the studies themselves. They cannot be used as the basis for further investigation due to inability to verify result accuracy. Additionally without high replicability it becomes impossible to compare the findings of different studies when using new data, this is a result of the inability to reproduce studies due to the lack required information. A system by which repeatability, reproducibility and replicability can be assured would be highly beneficial, providing a gateway to deeper research questions without having to struggle to reprocess or reproduce previous findings.

1.5 Updated Project Premise

As the project progressed an additional area of interest became apparent, this being the effect of reference databases on results. To reflect this a greater emphasis was placed on understanding how changes to the data ecosystem may effect the interpretation of the biological complexities and how these changes may impact historical studies. This updated focus promoted the development of a

computational architecture which focuses on repeatability, reproducibility and replicability in order to provide a platform for more rapid and robust microbiome analyses.

This architecture is then demonstrated through use of a case study reproducing a model for diagnosis of IBD using data from a pre-existing study and explore what impacts changes to the underlying data ecosystem have on the diagnostic outcomes. Thus showing clearly why these styles of pipelines are important for maintaining result validity as the reference datasets and computational tools advance, allowing the use of previous studies to form the basis of further research.

2 Summary of Literature Review and Updated Project Specification

There are many computational challenges for handling microbiome datasets beyond the repeatability and reproducibility problems stated above, these datasets are computationally expensive to process, stemming from their properties which can be likened to those of any other big data set [28]. Microbiome datasets satisfy the 3 Vs model of big data outlined by Doug Laney [29] these being:

- Huge volume, increasing rapidly due to more frequent sequencing and also the high dimensionality of sequenced datasets due to the complex biology they represent .
- High velocity, a result of the rapid changes that take place in the microbiome and the frequent sample sequencing that is required to capture them [16–19] .
- High variety both in the type of data and in the data storage methods that are used [30].

2.1 Previous Biomarker Studies

There have been many prior studies that have successfully determined IBD biomarkers from gut microbiome data.

Papa et al successfully analysed gut microbiome data to discover key taxa associated with each disease (CD and UC) and using random forest classified healthy samples against IBD with an AUC (Area under the ROC Curve) of 0.73 [31].

Pascal et al [32] demonstrated in their 2017 study how there are differences between the microbiomes of those with CD and UC. They showed using diversity analysis and factor identification how CD has a greater dysbiosis and lower microbial diversity than UC.

Zhou et al [33] ran analysis on 3 large datasets (Chinese IBD patients and PRISM and RISK datasets) and implemented RF with a prediction accuracy of 87.5% and 79.1% for classifying CD and UC respectively.

The problem with these studies is that many did not provide both the sample and metadata or omitted large proportions of their methodologies meaning it was impossible to reproduce them. This project uses the data from a study by Morgan et al [5], which identified several taxa that were more abundant in healthy individuals and also documented lowered diversity in those with IBD. This study was chosen as it had readily available data and metadata and had a small enough cohort size that it could be analysed in a reasonable time. This data was also later reprocessed by Duvallet et al [34] who implemented a random forest classification model for diagnosis of IBD. They did however note poor reproducibility of Morgan et al findings.

2.2 Requirements and Hypotheses

As stated earlier the requirements and hypothesis of this project have changed since the original project specification, table 2 shows these updated requirements. These changes reflect less emphasis on raw method comparison and instead capture the new goals of analysing the changes in biological representation when using different reference databases.

2.3 Hypotheses

- By using secondary microbiome analysis techniques it is possible to determine biomarkers for IBD from gut microbiome datasets allowing for the prediction of host disease phenotype using machine learning classification techniques.
- **The use of different reference databases effect the results of individual stages of microbiome analyses and also effects the accuracy of machine learning classification**.

2.4 Project Requirements

Index	Description of Requirement
Case Study - Non-Functional Requirements	
CS-1a	Can a biomarker distinguish between UC and CD?
CS-1b	What are the boundaries between UC and CD?
CS-1c	How do models compare when using different inputs (e.g. different reference databases)?
CS-2	Are current studies repeatable, reproducible and replicable?
CS-3a	Does changing the reference database greatly effect the inferred representation of the data or impact classification accuracy?
CS-3b	Is there are case for designing experiments and frameworks with reference databases changes in mind?
System Architecture Functional Requirements	
SA-1	The pipeline should be able to run a number of microbiome analysis steps i.e. diversity measures in a repeatable, reproducible and replicable way
SA-2	Each stage of the pipeline should be runnable independently whilst ensuring repeatability. Parameters of these stage should also be changeable to allow for evaluation of the impact of different parameters (different reference databases)
SA-3	In order to facilitate the previous requirement the results from a given stage will need to be stored in a uniform way so that they can be called upon as needed.
SA-4	The output of the pipeline should then be mapped into useable plots to represent the results.
SA-5	The framework should be able to apply Random forest to the results of the Qiime2 pipeline to attempt to classify IBD. If this can be done with an area under receiver operating characteristic curve of 70% this will be deemed as a valid biomarker (see non-functional requirement 1a).
SA-6	The framework should also facilitate the testing and comparison of the biomarker by using the same methods on different different data.
Evaluation Requirements	
EV-1	A combination of methodologies are used successfully to determine IBD biomarkers from microbiome data sets.
EV-2	All functional and non-functional requirements are met.
EV-3	Biomarkers can be used within prediction based algorithms to predict the disease phenotypes of a sample.
EV-4	An understanding of the biology can be derived, including how this may be effected by the use of different methods and databases

Table 2: A table showing all the functional and non-functional requirements for the project along with the evaluation criteria bold IDs represent a new or altered requirement

3 Design - System Architecture for Reproducibility of Metagenomics Analysis

The proposed analysis architecture is designed to wrap pre-existing tools and frameworks through the use of a containerised service oriented approach. This maintains the native abilities of the tools

used (QIIME2 [35] and the bioBakery [36]) whilst also providing additional functionality focusing on maintaining strong links between analytical results, their respective methods and data sources. This architecture is designed to be highly scalable thus allowing additional tools or processes to be incorporated with minimal alterations needed while ensuring high levels of reproducibility, repeatability and replicability.

3.1 Services Overview

The "big data" nature of metagenomic datasets and the high computational load associated with their analyses makes it highly beneficial to reduce data re-processing where possible. This can be made easier by separating the individual stages of analysis into discrete computational services whilst following a single responsibility principle. The metagenomic workflow lends itself to a service oriented approach as it consists of interconnected analysis steps (Figure 3), these steps can be converted into individual computational services without much issue. Following a single responsibility principle allows each service to be handled independently as each encapsulates a single part of the analysis workflow.

Figure 4 shows the hierarchical structure that results from this separation, each service is now able to be executed and modified independently. The key benefit to this is that later stages, such as those in secondary analysis, can now be modified without needing to reprocess each previous stage. As an example the reference database used during taxonomic classification could be changed without reprocessing the prior stages.

This is extremely effective at reducing computational costs associated with this style of analysis because the earlier stages are often the most computationally expensive. It also provides an added layer of flexibility to rerun the more exploratory secondary analysis stages where the ability to experiment with different parameters may have the greatest impact. It should be noted this is only possible when each stage is repeatable, as it allows confidence in the consistency of results as long as there are no changes. In order to achieve this repeatability each service is designed to be containerised.

In this project 14 containerised services are built that combined constitute a microbiome analysis pipeline. The pipeline takes raw data from metagenomic reads though to a modelling architecture which produces classifications from a random forest model.

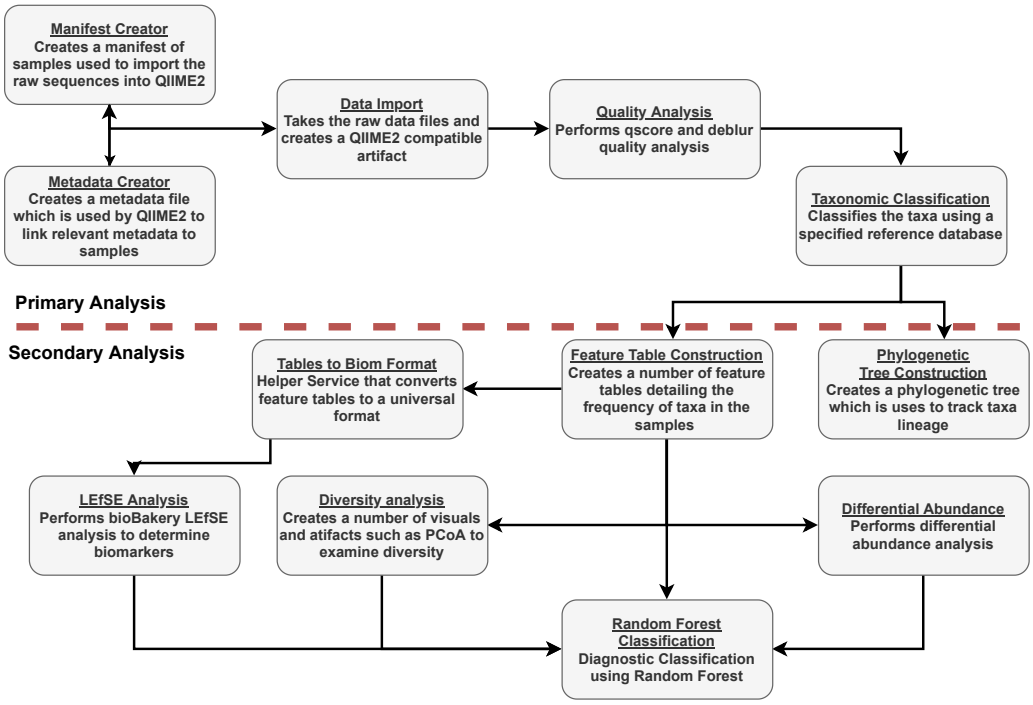


Figure 4: Hierarchical structure of the individual services that constitute the analysis pipeline

3.2 Containerised Workflow

The software stack for metagenomic analysis can include many dependencies, which may have many different versions, which may update over time. This massive variability makes it difficult to reproduce a papers workflow and results in several hurdles to high levels of reproducibility and repeatability. Without the ability to recreate the workflow reproducing results becomes nearly impossible, as such it is beneficial to run the services in an isolated environment.

This project uses Docker containers which provide a lightweight virtual machine that can be configured to include all the prerequisites for a given service including the runtime, tools and libraries with the additional ability to specify versions of each. Containers are therefore completely system agnostic (as long as docker is installed) not effected by the host OS or any tools that may be installed there. Each time a container is run it starts the runtime environment from a clean state this makes the containers highly repeatable as they cannot be effected by any previous executions. These containers can then be distributed and rebuilt on any system making them highly reproducible as they install the exact same environment originally used.

Each of the created services in this project shared a container pattern, first collecting the data for processing, processing that data and then storing the results and running information. This was done by designing a docker image for each container which included the necessary requirements for the analysis tools and also implementing common functionality for data handling within the architecture itself such as database access. These docker images can be found in the codebase for the project ().

3.3 Data Linkage

In order to provide high levels of reproducibility, repeatability and replicability formal linkage between results, methods, metadata and data is required, providing a clearer understanding of how each result is generated. To achieve this a central database (shared between all services) is used. This database stores information about each execution of a service, linking it to its respective dataset, information about its generated result and also logging the used parameters. This allows the result of any given experiment to be tracked back to its initial dataset by querying the database.

The database is designed using a NoSql schema, as shown in figure 5. Using a NoSql database affords an additional layer of flexibility which is needed in order to capture the wide variety of data used. The benefits of this can be seen in figure 5 by looking at the metadata object which contains the metadata for a given sample. The metadata collected for a sample may vary widely based on the purpose of the original study, with any number of possible attributes needing to be captured, by using NoSql these data objects can be entered into the database without needing to create additional tables.

Each time a service is run an entry is made to the database detailing its passed parameters, the service that provided its input and the file locations for its outputs (separated into data files and visual files, charts and images) along with an experiment ID. The manifest creator service is the one exception to this format as it also stores the location of all the sample files used. As this service is the root of the hierarchical tree all subsequent services can be tracked back and the sample information can be retrieved.

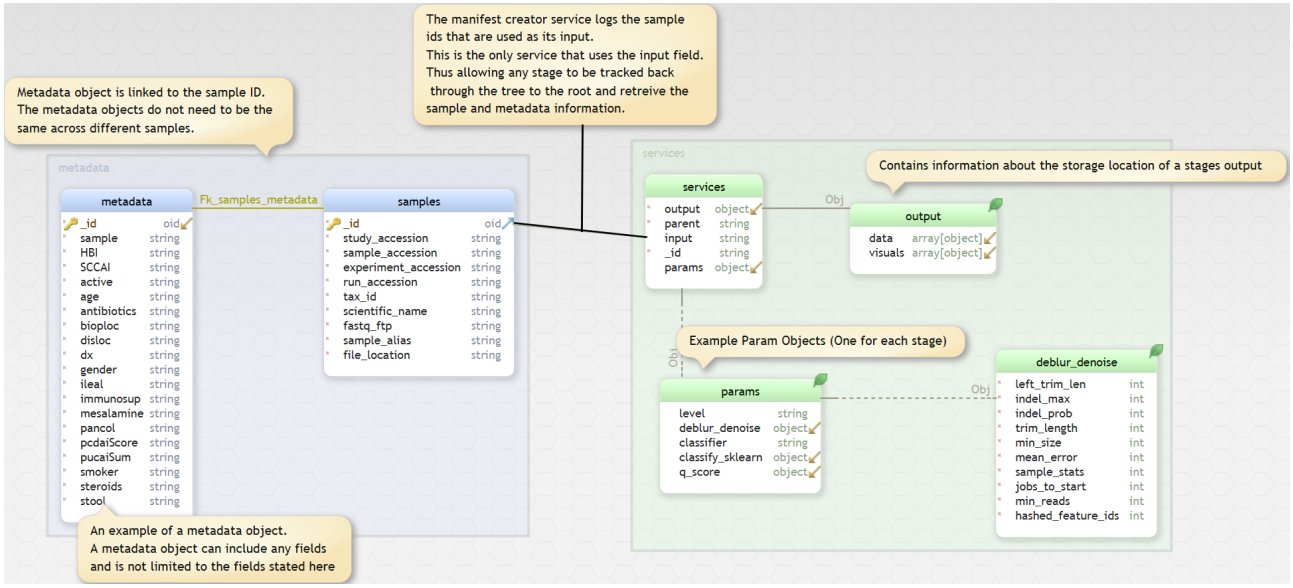


Figure 5: Database schema demonstrating the links between different collections and examples of objects within each collection

This database was implemented using MongoDB and a python file was written to provide a customised system to interact with the database. This file was included in all the services docker file and included a number of methods to collect specific and generalised data. This included a breadth first searching algorithm which is able to search the database tree for specific previous executions of a service using its ID, thus allowing the results of previous experiments to be retrieved.

3.4 Mounted Volume

The results produced by metagenomic tools cannot easily be stored in a database as they can consist of visuals and large quantities of data. As such these files need to be stored on the hosts system disk. This space on disk is connected to each docker containerised service by mounting it as a volume, becoming permanent storage that the container can use to save files. The file path of each result in this volume is stored in the central database. This not only provides links from result files to the database and the information stored there in but also allows each service to access the result files from any previous service without prior knowledge of its location. This allows flexible inter-service communication as file locations do not need to be hard coded. This is made possible by the tree searching methods implemented in the database helper class as described earlier.

4 Design - Case Study Diagnosis UC and CD from Metagenomics Data

In order to perform the analysis as part of the diagnostics case study a vast array of tools were utilised, all tools used were parts of the Qiime 2 platform with the exception of Linear discriminant analysis Effect Size (LEfSe) which is a part of the bioBakery. Qiime 2 is an open source platform designed for microbiome bioinformatics, it consists of a number of plugins which all follow the Qiime 2 design pattern. Qiime2 can be executed both from the command line but also provides an api for python 3 (the artifact api). Using this api a number of Qiime2 tools were included as service within the analysis architecture outlined earlier. The version of Qiime 2 used throughout the project was 2020.8, and was implemented using the docker "qiime2 core" images provided by the Qiime development team which includes all the core Qiime2 dependencies and plugins.

4.1 The Datasets

All analysis was performed using publicly available data which was initially analysed by Morgan et al in 2012 [5]. It consists of a cohort of 119 CD patients, 74 UC patients and 27 healthy controls. This study was chosen for its data availability both in the form of samples and metadata. All the samples were downloaded from the European Nucleotide Archive and are publicly available along with their respective metadata APPENDIX SOMETHING. This dataset has also been processed by Duvallet et al [34], who in addition to analysing the microbiome also used random forest machine learning in order to perform disease classification.

In addition to sample datasets two reference databases were chosen, SILVA 111 which was released in June 2012 and SILVA 138 released in December 2019. The large time period between the two datasets releases helps demonstrate how incremental updates over time may cause disparity between the results of new and old studies.

4.2 Primary Analysis Methods

1. All sample data was imported using a manifest file following qiime2s manifest format, the samples collected from the ENA were pre demultiplexed so this stage of processing was not required. The data was imported as "SingleEndFastqManifestPhred33V2", a format specifier used by qiime2 to identify the location of the quality score in the raw reads. *This was run from the manifest creator container (LINK).*

2. As part of the primary analysis all data sets were **quality controlled**. Qiimes implementation of dada2 was used for denoising, specifically the denoise_pyro function using the parameters outlined by the original study [5]. All samples were trimmed (truncated) to a minimum length of 200 nucleotides and maximum length of 600. Any reads with a quality score of less than 25 were discarded. *Implemented by the quality analysis container (LINK)*

3. All taxonomic features were **assigned a taxonomy** using qiime2s feature-classifier plugin using the classify-sklearn method. This makes use of a Naive Bayes classifier that is trained using a taxonomic reference database, two classifiers were trained using the fit-classifier-naive-bayes method of the same package using the two reference databases (Silva 111 and Silva 138). All of these methods utilise scikit-learn version 0.23.1 . Training of the classifiers took place in a stand-alone service which was run only one, this is due to the extremely high system requirements needed to train these classifiers (a peak of 41 gigabytes of ram was used during the training process with an average run time of 7 hours, 42 minutes per classifier), more detailed usage statistics are provided in table **??** later. *Implemented by the taxonomic classification container (LINK)*

4.3 Secondary Analysis Methods

1. Feature tables were constructed from the classified taxa information using qiime2s featuretable.collapse methods both a relative abundance table and standard frequency tables were utilised. *Feature tables container (LINK)*

2. **Diversity analyses** were performed using the diversity.pipelines.core_metrics_phylogenetic function. A sampling depth of 1153 was used in conjunction with a rooted phylogenetic tree created using the fasttree method *PhloTree Container (LINK).* This produced a number of diversity metrics of which the weighted_unifrac distance matrix and principle coordinate analysis (PCoA) results were used for beta diversity and PCoA analysis respectively. Beta-group-significance plots were created using the method of the same name using the permanova group significance metric and pairwise tests. PCoA plots were generated using the empress.visualisers.comunity_plot method, the filter_extra_samples, filter_missing_features and ignore_missing_samples arguments were all set to true to filter out any missing information and the number_of_features shown in the biplot was set to 10. The diversity.pipelines.alpha function was used to generate alpha diversity

plots using default parameters with the shannon metric. All other input data for these methods were created via the previous services.

3. **LEfSe** was implemented using the biobakery/lefse docker image (version 0.0.1) and was used to determine features most likely to explain the differences between the samples (CD, UC, healthy). Feature tables were converted from qiime2 format into comma separated file formats using pandas and biom packages provide by the qiime2 docker image. These files were then additionally formatted using LEfSes' built in function (format_input.py) with a normalisation parameter set to 1 million to convert the absolute taxa counts into relative values. Diagnosis classification was used as the class row for the data and no subclass were specified. When running the LEfSe analysis (via run_lefse.py) a linear discriminant analysis threshold of 2 was used.

4. In order to distinguish between CD, UC and Healthy samples results from several stages of the secondary analysis process were used to train a **random forest classifier** using sklearn.ensemble RandomForestClassifier with cross-validation. Training and testing data were created using python scikit-learn StratifiedKFold function with shuffling of the data. The classifier was trained using 100 estimators and a balanced class weight. The weighting is used to try to counteract the effect that the imbalanced classes has on the model performance. This imbalance is a remnant of the poor design for the original study.

5 Results and System Deployment

5.1 Case Study - Primary Analysis Results

Within the dataset there were 220 samples with each sample containing an average (median) of 4824.5 reads. A total of 1,077,809 individual reads were present amongst all samples. As part of primary analysis these samples were quality controlled, after this process the number of reads per sample was reduced to a median of 1137. The distribution of these read frequencies can be seen in figure 6. 872 unique features were identified from the remaining reads, where a feature represents an unique sequence variant (USV), which is a group of very closely related individuals. Essentially an USV represents a taxa that is present amongst the samples.

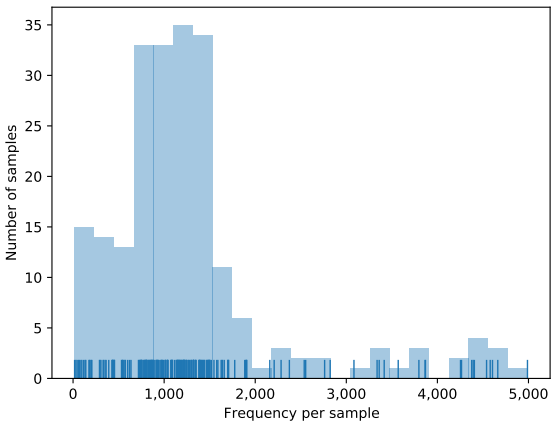


Figure 6: Histogram showing the frequency of reads per sample

USVs were then assigned taxonomies, using both the reference databases. Of the 872 USVs classified by the Silva111 database (using the methods outlined above) 823 changed their classification when using Silva138. It should be noted that changes in classification may be a result of changes in name and not a change in the taxonomy assigned by the classifier, i.e. the new classification may be the same microbe but the exact name of the microbe may have changed. However 185 changes did add a new taxonomic level to the classification many of which were new genus information that was not classified by the older database. Even excluding these changes this demonstrates the important

of using recent databases and updating old data, as the disconnect that is caused by the changing of taxonomic names and hierarchy can cause confusion between results. The impact of these changes is explored in later sections.

5.1.1 Computational Costs

The computational costs of this primary analysis are large, even when not including the time required to download and pre-process the data. Table **??** shows a summary of the resources used by each service.

TABLE COMING SOON - ANECDOTALLY I KNOW WHAT EACH USES, ITS CLEAR WHICH ARE MORE RESOURCE INTENSIVE, JUST NEED TO GET THAT DATA FROM DOCKER PROPERLY

However due to the implementation of the pipeline these services do not need to be reprocessed again resulting in significantly reduced processing cost. The benefits of this are clear, especially when considering that the study data used here is a relatively small sample size, reducing costs allows for larger datasets to be processed resulting in more generalized and robust findings.

5.2 Case Study - Secondary Analysis Results

Using the taxonomies assigned during the primary analysis stage a number of frequency tables were created. These tables detail the abundance of each taxa in each sample (POSSIBLE EXAMPLE FIGURE) thus allowing comparisons between samples and the taxa that are contained within them. Alpha diversity is one such comparison.

5.2.1 Alpha diversity

Alpha diversity is an estimate of how diverse the microbiome of a particular sample is, using richness (the number of different species) and evenness (the relative abundance of the species) in the sample. Shannon entropy is a metric to calculate the alpha diversity score and was used in conjunction with Kruskal-Wallis tests in order to compare the alpha diversity of the different conditions, figure 7 shows this distribution.

As the figure shows the alpha diversity of all groups are fairly similar, CD patients show a slight reduction in diversity relative to the healthy controls, as has been documented in the past [5]. However the significance of these differences are rather low (based on Kruskal-Wallis tests), this indicates that the differences between alpha diversity values are unlikely to be a result of individual conditions and would not be able to be used by themselves as a biomarker indicator.

Additionally the indication of slight significance between CD and healthy individuals (P value of 0.003063) may in fact be a result of many other factors not directly caused by the condition. An example of this might be the changes in diet that CD patients often adapt or the therapeutics they are prescribed, changing the microbiome and causing the significance seen in the data. These effects will be seen in the CD samples but are not the cause or result of the condition directly and thus could not be used as a primary method of diagnosis.

5.2.2 Beta Diversity

Beta diversity is a similar metric however it looks instead at how the compositions of each conditions microbiome differ from one another by using distance matrices where by the greater the distance between two groups the larger the dissimilarity between them. Figure 8 shows the distances between each diagnosis and the healthy group, here the distances between each group are again marginal, the major composition of the microbiomes between each group are similar. This is likely due to the general composition of the GIT microbiome which is highly diverse, indicating that a large proportion of the diversity captured is irrespective of the hosts disease status. Diagnosis using a microbiome cannot be based wholly on the overall characteristics of the microbiome and instead requires significantly more detailed comparisons.

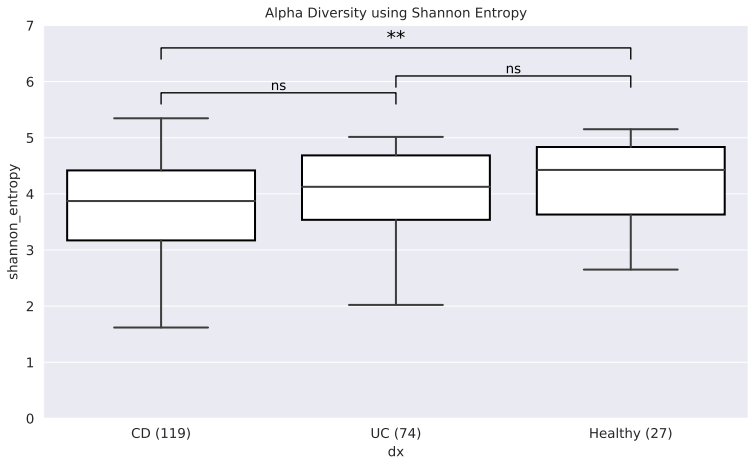


Figure 7: Alpha Diversity using Shannons entropy metric. Significance between groups denoted using the following system: $ns : p > 0.05, * : 0.01 < p < 0.05, ** : 10 - 4 < p < 0.01$ The p value of the Kruskal-Walis tests

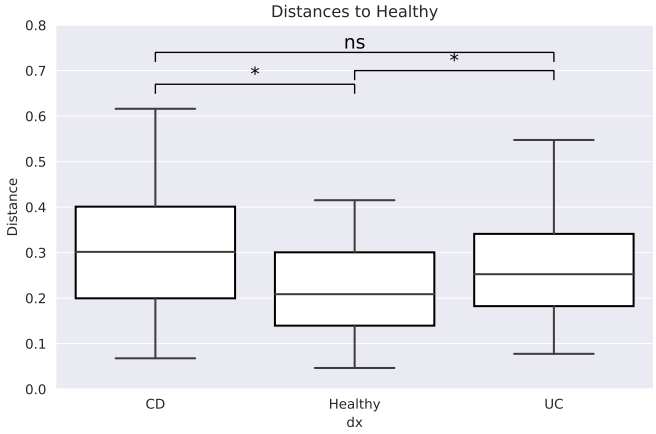


Figure 8: Beta diversity distances using the PERMANOVA metric with 999 permutations. Significance between groups denoted using the following system: $ns : q > 0.05, * : 0.01 < q < 0.05, ** : 10-4 < q < 0.01$ where q represents the p value of the pairwise PERMANOVA tests corrected for the permutations

5.2.3 Principle Coordinates Analysis (PCoA)

PCoA plots are methods of visualising the similarities and dissimilarities between communities, it does this by reducing the highly dimensional feature tables into 3 dimensional space using unsupervised clustering methods based on eigenanalysis. In addition to PCoA, biplots are used to demonstrate the key features that determine the position of the points on the PCoA plot, a visual representation of the weighting given to each feature. The PCoA plot in figure 9 shows the distribution of samples while using the **Silva 111** reference database along with its respective biplot.

There is no clear separation between groups (conditions) in the PCoA plot, even despite the first latent variable (axis 1) capturing 54.48% of the variance within the data. This implies that a large proportion of the variation in the microbiomes are not related to diagnosis at all, which reinforces the findings of the beta diversity tests. Despite this there are some trends between the clusters. Healthy individuals are less represented by the second latent variable (axis 2) tending to cluster at the bottom of the plot. The PCoA biplot arrows indicate the key taxa responsible for this clustering; *Faecalibacterium, Prevotella* and an unspecified genus of the *Lachnospiraceae family.*

Morgans original study on this data showed a reduced abundance of *Faecalibacterium* in those with CD which may indicate why its presence is causing lower clustering of healthy individuals and

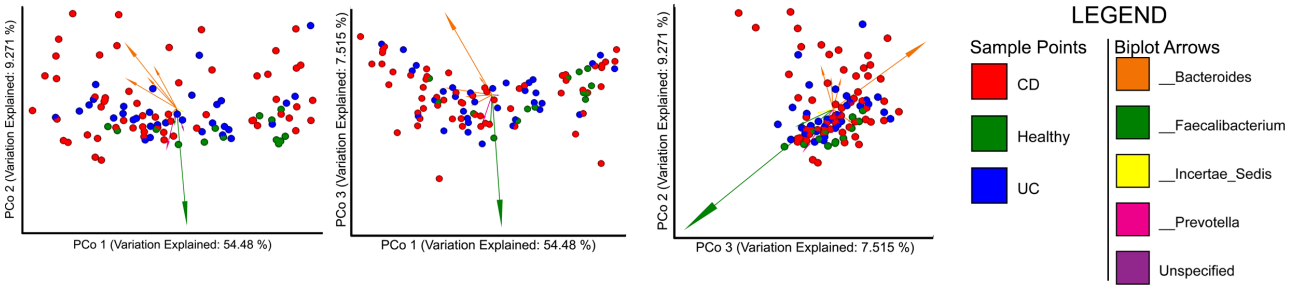


Figure 9: PCoA plots demonstrating the distribution of samples using a weighted unifrac distance matrix

thus higher clustering of those with IBD. However it is also one of the most abundant bacteria in the gut microbiome which may be the cause of its statistical power. Without obvious clustering based on diagnosis its unclear how well the taxa identified by PCoA will perform as biomarkers. This is especially true when comparing the results after changing the taxonomic database.

The results of PCoA when using the **Silva 138** reference database are shown in Figure 10 which demonstrates an overall similar clustering of samples with each latent variable even capturing similar percentages of the overall variation. However the biplot has changed drastically. *Ruminococcus.gnavus* has is now identified as most impactful taxa for positioning on the second latent variable (axis 2) instead of *Faecalibacterium*. The presence of *Ruminococcus.gnavus* has been linked to IBD specifically CD in multiple studies [37, 38] but was not identified specifically in the results by Morgan et al.

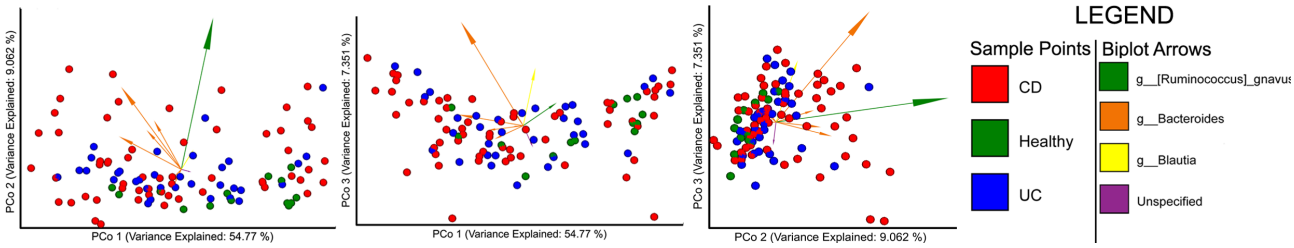
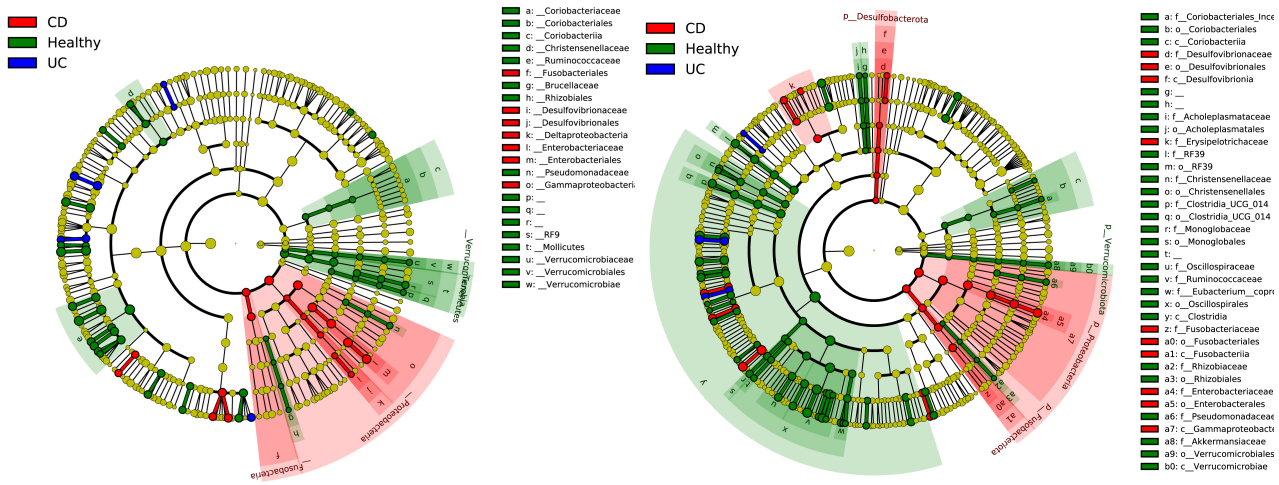


Figure 10: PCoA plots demonstrating the distribution of samples using a weighted unifrac distance matrix

This change in interpretation is a direct result on the new taxonomic information and resolution that is provided by the more recent database. The feature tables created using Silva 111 did not classify any ASV as *Ruminococcus.gnavus* as it is not present within the database as a possible taxonomy. This shows the power of updated databases, simply reprocessing the data has changed our understanding of how the microbiomes are represented by providing a new genus which may be related to IBD. This does not completely undermine the older database, the overall distributions remain the same and are captured accurately but these broad characteristics do not inform our understanding of the pathogenesis of the conditions.

5.2.4 LEFsE

LEFsE helps demonstrate the effects of different reference databases even more clearly. It is a tool specifically designed to identify biomarkers from microbiome data. When performed on the data classified using Silva 111 LEFsE identified 82 discriminative features however when using Silva 138 it identified 129. By looking at the cladogram (a tree representation of taxonomic lineage) produced by LEFsE (figure 11) we can interpret why this may be.



(a) Silva 111 reference database

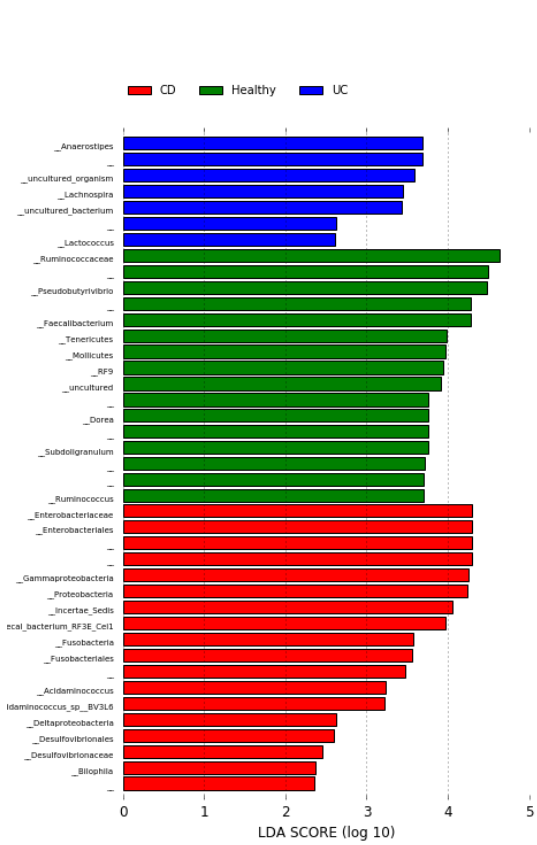
(b) Silva 138 reference database

Figure 11: Two cladograms generated for the dataset by lefse showing the discriminate taxa for each host phenotype and its taxonomic tree structure. Two different reference datasets were used.

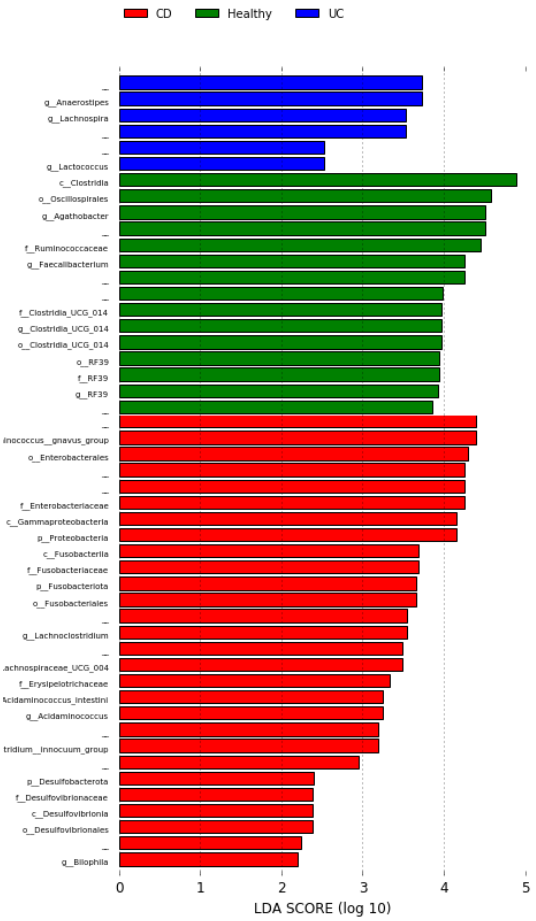
The structures of the two cladograms are subtly different. The branches of the Silva 138 have moved as a result of the additional branching at the 4th taxonomic level (taxonomic class). These additional branches are the result of the increase taxonomic resolution and precision provided by the more up to date reference database. This change in structure directly effects the taxanomic lineage of individual genus a species which allows LEfSe is able to identify more discriminant features in the Silva 138 dataset than was possible when using Silva 111.

In addition to cladograms LEfSe provides the Linear Discriminant Analysis (LDA) score for each of the features it identifies which is a measure of how discriminant the feature is. A subset of these scores can be seen in figure 12. LEfSe identifies significantly more features for the healthy samples than those with IBD and the LDA drop off of these features is also significantly more gradual.

The large discrimantory power and number of the healthy features indicates that classifying healthy vs. IBD would be possible with a high level of success. However it captures the difficulty of differentiating between CD and UC quite clearly. With a only a small number of discriminant features identified it would be extremely challenging to differentiate between CD and UC. This was also seen within the PCoA plots with CD and UC overlapping each other quite heavily with no clear clustering.



(a) Silva 111 reference database



(b) Silva 138 reference database

Figure 12: Two LEfSe LDA bar plots identifying the key discriminant features of each condition. FULL DATA IN APPENDIX

The increased precision of Silva 138 has allowed LEfSe to determine significantly more features for CD compared to Silva 111. Interestingly, many of the biomarkers identified in figure 12a are also present in figure 12b and hold close to the same LDA score. This is important as it means that biomarkers found using older datasets are not deemed invalid by more recent results and in fact the new results are just adding to the understanding that is already being presented. It should be noted that this may only be true for the specific data presented here.

In contrast, changing the reference database actually removes a UC biomarker. This is likely because the UC taxa identified are extremely specific (only at the genus level) with no broader taxonomic groups so when the structure of the data changes these biomarkers lose some of their significance. This may indicate that the taxa that was removed was in fact noise present in the data and not an acutal valid biomarker.

Of the CD biomarkers identified by LEfSe (using Silva 138) the second highest is *Ruminococcus.gnavus* which is also a key factor determined by PCoA. However when using Silva 111 the highest identified taxa is *Enterobacteriaceae*, this bacteria is not even present within the PCoA biplot. The reason for its lack of presence is unclear but shows that using newer databases with more accurate representations of the underlying biology not only improves the results of specific tools like LEfSe but also potentially the broader statistical tools like PCoA.

5.2.5 Machine Learning - Random Forest Classification

Random forest is a supervised ensemble learning method which can be used for data classification, it does this by creating many decision trees and then uses the result of majority voting from all the trees as the final classification. It has been used by several studies to distinguish between IBD and healthy

controls [33, 34, 39] and performs better than many other methods [40].

As random forest is a generic machine learning method any combination of inputs could be used to train and test the model. For this project 3 different inputs were chosen each of which were processed for both Silva 111 and Silva 138, those inputs were:

1. The classified ASV relative abundance table, the input for many of the secondary analysis steps
2. The classified ASV relative abundance table, with the alpha diversity (shannon entropy) scores
3. A relative abundance table only containing the ASV's that LEfSe determined to be significant

The area under the Receiver operating characteristic curve (AUROC) was used as the primary performance metric as it very commonly used in similar studies and allows for a direct comparison to Duvallet et al [34] who also performed Random forest on this dataset achieving an AUROC of 0.81 they did however deem values > 0.7 as successful. Additionally due to the large class imbalance within the dataset (only 27 healthy samples in contrast to the 119 CD samples) a precision recall curve is also used as it has been shown to work well with imbalanced classification models. Table 4 presents the summary of these scores.

	Area Under ROC (Macro Averaged)		Area Under precision-recall (Macro Averaged)	
Data input	**Silva 111**	**Silva 138**	**Silva 111**	**Silva 138**
Just the ASV relative abundance table (1)	**0.71**	0.69	**0.55**	0.52
The ASV relative abundance table with alpha diversity (2)	0.68	**0.69**	0.52	**0.54**
The ASV relative abundance table filtered to the taxa identified by LEfSe (3)	0.69	0.69	**0.55**	0.52

Table 4: A table showing all area under curve scores for the two scoring metrics (ROC and precision-recall) The highest score of each category are highlighted in bold

The scores for both metrics are extremely close (within 0.03) regardless of input data format or Silva version. The average of all the AUROC scores is 0.69 within margin or error to the 0.7 achieved by Duvallet, meaning the pipeline can successfully replicate the study. However when using the precision recall curve the accuracy is show to be significantly worse, barely better than a random classifier (0.5 area under curve score). This clearly shows how choice of performance metric can radically change the interpretation of the accuracy of the classifier.

The performance of the LEfSe filtered table when using Silva 138 is the same as unfiltered data despite having 93 fewer features. This indicates that the features identified by LEfSe do indeed provide higher levels of discriminant power. Although this reduction in features did not drastically improve training time for this dataset reducing the dimensionality of larger datasets may prove beneficial, especially if the relative performance of the models remains the same.

The increased resolution provided by the newer reference database did not improve the overall performance of the model. This is likely due to most significant features being accurately captured in both the old and new data. It is the more novel taxa that are most likely to change as result of the increased taxonomic resolution and these features carry less significance for classification as a whole. This does not make these features unimportant as they be linked to specific types of cases of disease, for example disease that is especially active, but does mean that they will only impact the classification of specific individuals. It may also be that these less common taxa are closely linked to individual conditions but in order to capture the significance much larger datasets are required.

6 Project Evaluation and Discussion

The following section discusses the primary contributions of this project whilst evaluating the project process and outcomes as a whole. This is done by comparing the final project outcomes against the original requirements outlined in table 2 of section 2.4.

6.1 Reproducibility of Existing Systems

This project has assessed the lack of reproducibility of current systems for microbiome research and how they do not currently enable reproducible outcomes for the majority of diagnostic studies (requirement CS-2). The original studies covered in the literature review all failed to be easily reproducible. Despite all studies providing public access to the raw sequence files only one provided the meta data as well. These weak connections between the samples and the necessary metadata to analyse them are one of the primary hurdles to reproducibility.

Additionally many studies failed to provide the necessary information to reproduce analyses methods. Current systems rely on authors providing detailed information about methods or providing code files to allow for reproducibility. This is far from a perfect solution as small gaps in information or differences in analysis environments can inhibit reproducibility. Overall this projects case study demonstrated that many of the findings of the original study, by Morgan et al, were not reproducible this was also noted in the report by Duvallet et al who also failed to reproduce Morgans findings. Furthermore this project failed to reproduce a random forest model with the same AUROC accuracy as the one created by Duvallet, likely a result of the outdated tools they used.

These issues with reproducibility have to be addressed with more transparent result provenance, correctly tracking the data and metadata used, and improved analyses workflows which allow for reproducibility.

6.2 System Architecture for Reproducibility of Metagenomics Analysis

A primary contribution and requirement (SA-1) of this project was the proposal and development of a computational system which promotes reproducibility, repeatability and replicability of microbiome analyses. The developed system successfully achieved this requirement by utilising a combination of containerised services and standardised data handling.

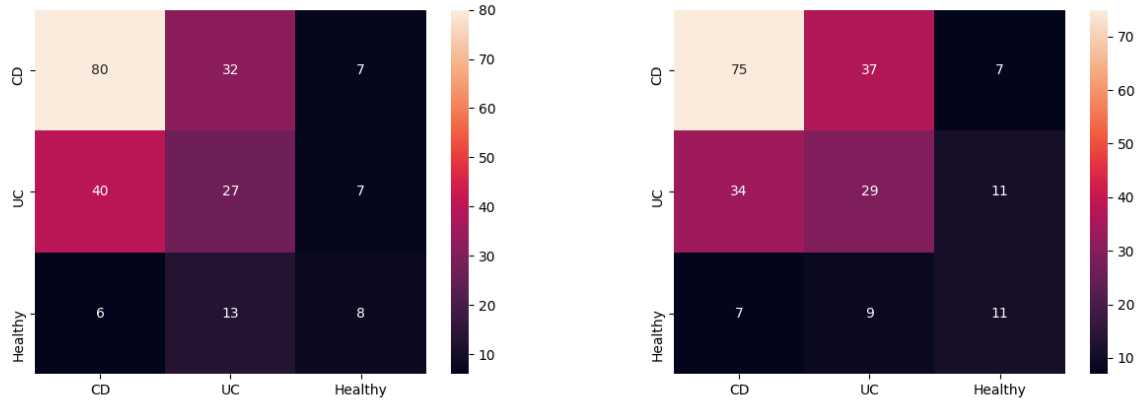
All stages of the analyses process were implemented using a containerised architecture which allowed for service independence. This independence allowed each service to be executed individually which was a primary requirement for the system (requirement SA-2). By meeting this requirement the system is able to reprocess individual stages of analysis without reprocessing the entire pipeline which massively reduces the computational costs when repeating and reproducing results. Each services parameters are defined externally from the code of the container itself, which provided flexibility when reprocessing stages with different parameters and data which was an essential part of the investigation that took place in the project (requirements SA-2, SA-6). Additionally by allowing the data to be changed it massively reduces the development overhead for replicating studies as new data can be processed using the exact same methods without having to rewrite large amount of software.

This was possible by utilising a combination of a centralised database, which was used to store the runtime parameters for each service execution, and local system file storage (a mounted volume) which stored the generated results (SA-3). This was necessary as each service generates not only data results but also graphs and plots as required by SA-4. This uniform methods of data handling in addition to a generic system for utilising the data storage systems allows new services to be written without specific knowledge of the data architecture providing easier scalability of the system.

Overall this service provides massive benefits for reducing both the development and computational time associated with microbiome research. This is not only true when completing initial investigations, with easier data handling and system scalability but also when distributing and reproducing pre existing results.

6.3 Diagnosis of UC and CD from Metagenomics Data

Using the developed analysis system it was possible to create a random forest model that was able to classify UC, CD and healthy samples with an average AUROC of 0.69, within margin of error from the initial specification requirement (SA-6). However this does not allow for a conclusive answer as to whether these biomarkers can be used to distinguish between UC and CD (requirement CS-1a) or what the exact boundaries between the conditions are (requirement CS-1b). By looking at confusion matrices for the models, in figure 13, it becomes clear why this is the case.



(a) Silva 111 reference database

(b) Silva 138 reference database

Figure 13: The confusion matrices from a random forest model trained using ASV relative abundance tables, using StratifiedKFold for cross validation

These confusion matrices show that high levels of confusion between CD and UC are still occurring (regardless of reference database version). Although the models perform well based on overall AUROC score the remaining confusion demonstrates that the boundaries between conditions still remains unclear. This is the result of the similar nature of the conditions and thus the large similarities and overlaps of the microbiomes. This overlap was demonstrated nicely by the PCoA plots of section **??** and also the small difference between the various diversity measures. Additionally although LEfSe was able to determine discriminant taxa for both conditions they were few in number and had low discriminatory power compared to those of the healthy control group.

Overall the results show the potential of using biomarkers to distinguish between UC and CD (requirements CS-1a) but due to the complexity of the conditions and their respective microbiomes the results generated cannot be used to perfectly separate conditions. Further research into other methods of identifying biomarkers along with larger datasets may prove combined with new methods of classification (for example deep learning) may prove beneficial in using metagenetic for diagnosis.

6.4 Changes to Reference Databases and their Impact on Classification

One of the primary goals of the project was to determine whether changing the reference database greatly impacted the classification accuracy of the random forest model (CS-3a). Based on the performance of the models using AUROC and precision-recall curves its clear that changes to the reference databases do not greatly impact the overall model performance. Additionally the confusion seen between UC and CD in the models were present in both Silva 111 and Silva 138.

However, this does not mean that the changes did not impact the inferred representation of the microbiome, in fact 60 samples changed classification when changing the reference database. This is a direct result of differences to how the composition of the microbiome was inferred between the two databases. As stated in section **??**, of the 872 ASVs identified within the samples 823 changed classification and of those that changed 185 gained additional levels of taxonomic resolution. These changes effect how individual samples are classified within the model due to changes in their inferred composition, the stacked barplot in figure 14 provides an visual of how the composition can change.

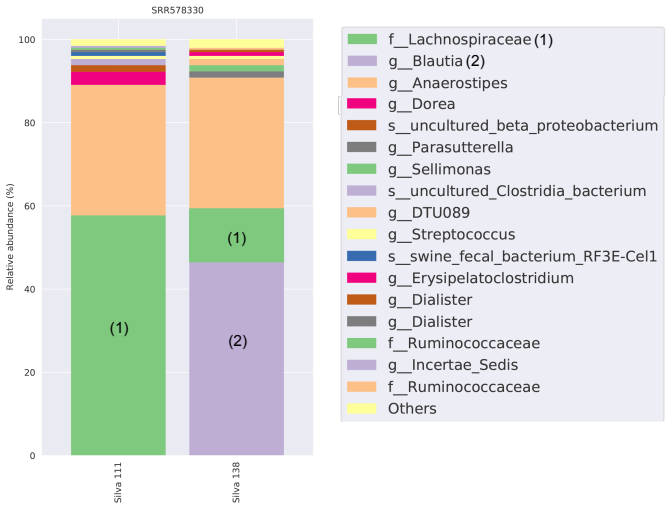


Figure 14: Stacked bar plot detailing the relative abundance of bacteria in sample SRR578330. This sample is a UC patient misclassified as CD when using Silva 111 but correctly classified when using Silva 138. Legend is ordered relative to the bottom of the stacked bar

The differences between the composition of the sample are quite dramatic. When using Silva 111 58% of the relative abundance of the sample was considered to be *Lachnospiracace*(1) however this percentage changed as a result of additional information present in Silva 138. Due to Silva 138s increased taxonomic resolution a large proportion of the *Lachnospiracace* was classified to include additional genus information labelled *Blautia*(2), which is a sub genus of the *Lachnospiracace* family. In addition to the increased resolution for features with large relative abundance the changes to the composition are also present in the less abundant species, with many of the species changing classification entirely.

These changes to the composition directly impact how the sample was classified by the random forest model. This sample was initial misclassified as CD when using Silva 111 but received the correct classification when using Silva 138. This clearly shows that classifications are susceptible to reference database changes even if the global performance of the model is not effected. This lack of impact on the global performance despite lower level changes could indicate that the microbiome may be an unreliable data source for precise classification but could also be a result of how models interpret the statistical significance of data features.

Regardless of which may be the case it is clear that systems that attempt to create this style of classification model would benefit from being able to update or change reference databases as appropriate (Requirement CS-3b). This ability would allow the robustness of the models when subject to lower level changes to be assessed and also allow the models to remain relevant as the field progress as a whole.

6.5 Overall Project Process Evaluation

Overall this project has been successful in implementing a computational analysis system and answering many questions about the microbiome of patients with IBD. Two hypotheses were presented at the beginning of the project and broken down into key functional and non-functional goals which were all met (requirement EV-2) in addition to a final set of evaluation requirements. The project satisfied requirements EV-1 and EV-3 by identifying potential biomarkers for UC and CD and successfully predicting disease phenotypes using a random forest model. However, the results demonstrated that this style of classification is still prone to classification confusion as a result of the difficult to define boundaries between conditions.

EV-4, which encapsulated the second hypothesis, was to determine whether changes to reference databases would effect the results of the microbiome analyses. This requirement has also been met

with changes to the inferred microbiome composition shown to highly prevalent across many samples which in turn altered several results from the analyses stages. However, these changes did not impact the accuracy of machine learning classification due to a combination of factors, which were covered extensively in the previous section.

On reflection this project was impacted by many unforeseen issues throughout the development process mainly as a result of its complicated specification, incorporating both a software development and research based focus. One such issue was the acquirement of data for processing which took significantly longer than originally expected due to the poor metadata availability. Although this issue did inform massive positive changes to the overall goals of the project, focusing more on reproducibility, it did prevent the exploration of several additional areas initially planned, namely the comparison of individual machine learning methods.

A lot of time was also consumed trailing different bioinformatics tools that did not provide relevant information for answering the hypothesis of the project. Although the methods were selected in the design stages it became apparent that there were many different implementations, due to an unfamiliarity with bioinformatics and microbiome studies, time was unnecessarily spent implementing methods that did not provide the exact results needed. In future projects it would be beneficial to spend more time research individual implementations of these analyses methods to reduce this impact on time.

Despite these issues the overall process of the project worked well, first focusing on the creation of the analysis system architecture which then allowed rapid creation of results. This meant that even though the early stages of the project faced several delays all results were still produced and analysed. This was only possible thanks to the reduced computation time allowed by the system and also the parameter flexibility which enabled the entire pipeline to be customised and then left to execute. This meant that once the pipeline had been used for Silva 111 changing a single line in the docker configuration allowed the pipeline to be run again with Silva 138, this would not have been possible without the time committed to ensuring the pipeline was fit for purpose.

7 Conclusion

Microbiome research has provided new insights into human biology and been labelled on of the frontiers of life sciences although it has significant 'big data' challenges. These challenges have resulted in poor data management and as a result led to low levels of result reproducibility. This project address some of the reproducibility challenges associated with the analyses of these 'big data' sets by providing and implementing a scalable standardised computing architecture. This architecture provides methods of data and method management whilst also aimed at reducing computational and development costs. Application of this architecture could enable microbiome research to be a more robust, established frontier of life sciences.

References

[1] C. M. Cullen, K. K. Aneja, S. Beyhan, C. E. Cho, S. Woloszynek, M. Convertino, S. J. McCoy, Y. Zhang, M. Z. Anderson, D. Alvarez-Ponce, E. Smirnova, L. Karstens, P. C. Dorrestein, H. Li, A. S. Gupta, K. Cheung, J. G. Powers, Z. Zhao, and G. L. Rosen, "Emerging priorities for microbiome research," *Frontiers in Microbiology*, vol. 11, feb 2020.

[2] L. Jostins, , S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J.-P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Büning, A. Cohain, S. Cichon, M. D'Amato, D. D. Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. D. Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhao, M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, E. E. Schadt, M. J. Daly, A. Franke, M. Parkes, S. Vermeire, J. C. Barrett, and J. H. Cho, "Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease," *Nature*, vol. 491, no. 7422, pp. 119–124, oct 2012.

[3] C. Manichanh, N. Borruel, F. Casellas, and F. Guarner, "The gut microbiota in IBD," *Nature Reviews Gastroenterology & Hepatology*, vol. 9, no. 10, pp. 599–608, aug 2012.

[4] D. N. Frank, C. E. Robertson, C. M. Hamm, Z. Kpadeh, T. Zhang, H. Chen, W. Zhu, R. B. Sartor, E. C. Boedeker, N. Harpaz, N. R. Pace, and E. Li, "Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases," *Inflammatory Bowel Diseases*, vol. 17, no. 1, pp. 179–184, jan 2011.

[5] X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, A. Bousvaros, J. Korzenik, B. E. Sands, R. J. Xavier, and C. Huttenhower, "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment," *Genome Biology*, vol. 13, no. 9, p. R79, 2012.

[6] D. King, R. C. Reulen, T. Thomas, J. S. Chandan, R. Thayakaran, A. Subramanian, K. Gokhale, N. Bhala, K. Nirantharakumar, N. J. Adderley, and N. Trudgill, "Changing patterns in the epidemiology and outcomes of inflammatory bowel disease in the united kingdom: 2000-2018," *Alimentary Pharmacology & Therapeutics*, vol. 51, no. 10, pp. 922–934, apr 2020.

[7] S. Alatab, S. G. Sepanlou, K. Ikuta, H. Vahedi, C. Bisignano, S. Safiri, A. Sadeghi, M. R. Nixon, A. Abdoli, H. Abolhassani, V. Alipour, M. A. H. Almadi, A. Almasi-Hashiani, A. Anushiravani, J. Arabloo, S. Atique, A. Awasthi, A. Badawi, A. A. A. Baig, N. Bhala, A. Bijani, A. Biondi, A. M. Borzì, K. E. Burke, F. Carvalho, A. Daryani, M. Dubey, A. Eftekhari, E. Fernandes, J. C. Fernandes, F. Fischer, A. Haj-Mirzaian, A. Haj-Mirzaian, A. Hasanzadeh, M. Hashemian, S. I. Hay, C. L. Hoang, M. Househ, O. S. Ilesanmi, N. J. Balalami, S. L. James, A. P. Kengne, M. M. Malekzadeh, S. Merat, T. J. Meretoja, T. Mestrovic, E. M. Mirrakhimov, H. Mirzaei, K. A. Mohammad, A. H. Mokdad, L. Monasta, I. Negoi, T. H. Nguyen, C. T. Nguyen, A. Pourshams, H. Poustchi, M. Rabiee, N. Rabiee, K. Ramezanzadeh, D. L. Rawaf, S. Rawaf, N. Rezaei, S. R. Robinson, L. Ronfani, S. Saxena, M. Sepehrimanesh, M. A. Shaikh, Z. Sharafi, M. Sharif, S. Siabani, A. R. Sima, J. A. Singh, A. Soheili, R. Sotoudehmanesh, H. A. R. Suleria, B. E. Tesfay, B. Tran, D. Tsoi, M. Vacante, A. B. Wondmieneh, A. Zarghi, Z.-J. Zhang, M. Dirac, R. Malekzadeh, and M. Naghavi, "The global, regional, and national burden of inflammatory

bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017," *The Lancet Gastroenterology & Hepatology*, vol. 5, no. 1, pp. 17–30, jan 2020.

[8] M. E. van der Valk, M.-J. J. Mangen, M. Leenders, G. Dijkstra, A. A. van Bodegraven, H. H. Fidder, D. J. de Jong, M. Pierik, C. J. van der Woude, M. J. L. Romberg-Camps, C. H. Clemens, J. M. Jansen, N. Mahmmod, P. C. van de Meeberg, A. E. van der Meulen-de Jong, C. Y. Ponsioen, C. J. Bolwerk, J. R. Vermeijden, P. D. Siersema, M. G. van Oijen, and B. O. and, "Healthcare costs of inflammatory bowel disease have shifted from hospitalisation and surgery towards anti-TNFα therapy: results from the COIN study," *Gut*, vol. 63, no. 1, pp. 72–79, nov 2012.

[9] S. R. Vavricka, S. M. Spigaglia, G. Rogler, V. Pittet, P. Michetti, C. Felley, C. Mottet, C. P. Braegger, D. Rogler, A. Straumann *et al.*, "Systematic evaluation of risk factors for diagnostic delay in inflammatory bowel disease," *Inflammatory bowel diseases*, vol. 18, no. 3, pp. 496–505, 2012.

[10] A. M. Schoepfer, M.-A. Dehlavi, N. Fournier, E. Safroneeva, A. Straumann, V. Pittet, L. Peyrin-Biroulet, P. Michetti, G. Rogler, and S. R. Vavricka, "Diagnostic delay in crohn's disease is associated with a complicated disease course and increased operation rate," *American Journal of Gastroenterology*, vol. 108, no. 11, pp. 1744–1753, nov 2013.

[11] K. Geboes, "Pathology of inflammatory bowel diseases (IBD): variability with time and treatment," *Colorectal Disease*, vol. 3, no. 1, pp. 2–12, jul 2008.

[12] A. B. Price, "Overlap in the spectrum of non-specific inflammatory bowel disease–'colitis indeterminate'." *Journal of Clinical Pathology*, vol. 31, no. 6, pp. 567–577, jun 1978.

[13] R. K. Yantiss and R. D. Odze, "Diagnostic difficulties in inflammatory bowel disease pathology," *Histopathology*, vol. 48, no. 2, pp. 116–132, jan 2006.

[14] R. Boyapati, J. Satsangi, and G.-T. Ho, "Pathogenesis of crohn's disease," *F1000Prime Reports*, vol. 7, apr 2015.

[15] I. Sekirov, S. L. Russell, L. C. M. Antunes, and B. B. Finlay, "Gut microbiota in health and disease," *Physiological Reviews*, vol. 90, no. 3, pp. 859–904, jul 2010.

[16] L. A. David, C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, S. B. Biddinger, R. J. Dutton, and P. J. Turnbaugh, "Diet rapidly and reproducibly alters the human gut microbiome," *Nature*, vol. 505, no. 7484, pp. 559–563, dec 2013.

[17] Z. Xu and R. Knight, "Dietary effects on human gut microbiome diversity," *British Journal of Nutrition*, vol. 113, no. S1, pp. S1–S5, dec 2014.

[18] L. Dethlefsen, S. Huse, M. L. Sogin, and D. A. Relman, "The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rRNA sequencing," *PLoS Biology*, vol. 6, no. 11, p. e280, nov 2008.

[19] M. Blaser, "Stop the killing of beneficial bacteria," *Nature*, vol. 476, no. 7361, pp. 393–394, aug 2011.

[20] K. Strimbu and J. A. Tavel, "What are biomarkers?" *Current Opinion in HIV and AIDS*, vol. 5, no. 6, pp. 463–466, nov 2010.

[21] M. Balvočiūtė and D. H. Huson, "SILVA, RDP, greengenes, NCBI and OTT — how do these taxonomies compare?" *BMC Genomics*, vol. 18, no. S2, mar 2017.

[22] M. A. Sierra, Q. Li, S. Pushalkar, B. Paul, T. A. Sandoval, A. R. Kamer, P. Corby, Y. Guo, R. R. Ruff, A. V. Alekseyenko, X. Li, and D. Saxena, "The influences of bioinformatics tools and reference databases in analyzing the human oral microbial community," *Genes*, vol. 11, no. 8, p. 878, aug 2020.

[23] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604, pp. 452–454, may 2016.

[24] S. L. McArthur, "Repeatability, reproducibility, and replicability: Tackling the 3r challenge in biointerface science and engineering," *Biointerphases*, vol. 14, no. 2, p. 020201, mar 2019.

[25] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, sep 2010.

[26] C. Poussin, N. Sierro, S. Boué, J. Battey, E. Scotti, V. Belcastro, M. C. Peitsch, N. V. Ivanov, and J. Hoeng, "Interrogating the microbiome: experimental and computational considerations in support of study reproducibility," *Drug Discovery Today*, vol. 23, no. 9, pp. 1644–1657, sep 2018.

[27] P. D. Schloss, "Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research," *mBio*, vol. 9, no. 3, jun 2018.

[28] M. Cheng and K. N. Le Cao, "Microbiome big-data mining and applications using single-cell technologies and metagenomics approaches toward precision medicine," *Frontiers in genetics*, vol. 10, 2019.

[29] D. Laney, "3D data management: Controlling data volume, velocity, and variety," META Group, Tech. Rep., February 2001. [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[30] National Center for Biotechnology Information. (2020, Accessed On: 07, Nov. 2020) File format guide. [Online]. Available: https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/bam-files

[31] E. Papa, M. Docktor, C. Smillie, S. Weber, S. P. Preheim, D. Gevers, G. Giannoukos, D. Ciulla, D. Tabbaa, J. Ingram, D. B. Schauer, D. V. Ward, J. R. Korzenik, R. J. Xavier, A. Bousvaros, and E. J. Alm, "Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease," *PLoS ONE*, vol. 7, no. 6, p. e39242, jun 2012.

[32] V. Pascal, M. Pozuelo, N. Borruel, F. Casellas, D. Campos, A. Santiago, X. Martinez, E. Varela, G. Sarrabayrouse, K. Machiels, S. Vermeire, H. Sokol, F. Guarner, and C. Manichanh, "A microbial signature for crohn's disease," *Gut*, vol. 66, no. 5, pp. 813–822, feb 2017.

[33] Y. Zhou, Z. Z. Xu, Y. He, Y. Yang, L. Liu, Q. Lin, Y. Nie, M. Li, F. Zhi, S. Liu, A. Amir, A. González, A. Tripathi, M. Chen, G. D. Wu, R. Knight, H. Zhou, and Y. Chen, "Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction," *mSystems*, vol. 3, no. 1, jan 2018.

[34] C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nature Communications*, vol. 8, no. 1, dec 2017.

[35] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, Y. Bai, J. E. Bisanz, K. Bittinger, A. Brejnrod, C. J. Brislawn, C. T. Brown, B. J. Callahan, A. M. Caraballo-Rodríguez, J. Chase, E. K. Cope, R. D. Silva, C. Diener, P. C. Dorrestein, G. M. Douglas, D. M. Durall, C. Duvallet, C. F. Edwardson, M. Ernst, M. Estaki, J. Fouquier, J. M. Gauglitz, S. M. Gibbons, D. L. Gibson, A. Gonzalez, K. Gorlick, J. Guo, B. Hillmann, S. Holmes, H. Holste, C. Huttenhower, G. A. Huttley, S. Janssen, A. K. Jarmusch, L. Jiang, B. D. Kaehler, K. B. Kang, C. R. Keefe, P. Keim, S. T.

Kelley, D. Knights, I. Koester, T. Kosciolek, J. Kreps, M. G. I. Langille, J. Lee, R. Ley, Y.-X. Liu, E. Loftfield, C. Lozupone, M. Maher, C. Marotz, B. D. Martin, D. McDonald, L. J. McIver, A. V. Melnik, J. L. Metcalf, S. C. Morgan, J. T. Morton, A. T. Naimey, J. A. Navas-Molina, L. F. Nothias, S. B. Orchanian, T. Pearson, S. L. Peoples, D. Petras, M. L. Preuss, E. Pruesse, L. B. Rasmussen, A. Rivers, M. S. Robeson, P. Rosenthal, N. Segata, M. Shaffer, A. Shiffer, R. Sinha, S. J. Song, J. R. Spear, A. D. Swafford, L. R. Thompson, P. J. Torres, P. Trinh, A. Tripathi, P. J. Turnbaugh, S. Ul-Hasan, J. J. J. van der Hooft, F. Vargas, Y. Vázquez-Baeza, E. Vogtmann, M. von Hippel, W. Walters, Y. Wan, M. Wang, J. Warren, K. C. Weber, C. H. D. Williamson, A. D. Willis, Z. Z. Xu, J. R. Zaneveld, Y. Zhang, Q. Zhu, R. Knight, and J. G. Caporaso, "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2," *Nature Biotechnology*, vol. 37, no. 8, pp. 852–857, jul 2019.

[36] L. J. McIver, G. Abu-Ali, E. A. Franzosa, R. Schwager, X. C. Morgan, L. Waldron, N. Segata, and C. Huttenhower, "bioBakery: a meta'omic analysis environment," *Bioinformatics*, vol. 34, no. 7, pp. 1235–1237, nov 2017.

[37] A. B. Hall, M. Yassour, J. Sauk, A. Garner, X. Jiang, T. Arthur, G. K. Lagoudas, T. Vatanen, N. Fornelos, R. Wilson, M. Bertha, M. Cohen, J. Garber, H. Khalili, D. Gevers, A. N. Ananthakrishnan, S. Kugathasan, E. S. Lander, P. Blainey, H. Vlamakis, R. J. Xavier, and C. Huttenhower, "A novel ruminococcus gnavus clade enriched in inflammatory bowel disease patients," *Genome Medicine*, vol. 9, no. 1, nov 2017.

[38] M. T. Henke, D. J. Kenny, C. D. Cassilly, H. Vlamakis, R. J. Xavier, and J. Clardy, "Ruminococcus gnavus, a member of the human gut microbiome associated with crohn's disease, produces an inflammatory polysaccharide," *Proceedings of the National Academy of Sciences*, vol. 116, no. 26, pp. 12 672–12 677, jun 2019.

[39] L. J. Marcos-Zambrano, K. Karaduzovic-Hadziabdic, T. L. Turukalo, P. Przymus, V. Trajkovik, O. Aasmets, M. Berland, A. Gruca, J. Hasic, K. Hron, T. Klammsteiner, M. Kolev, L. Lahti, M. B. Lopes, V. Moreno, I. Naskinova, E. Org, I. Paciência, G. Papoutsoglou, R. Shigdel, B. Stres, B. Vilne, M. Yousef, E. Zdravevski, I. Tsamardinos, E. C. de Santa Pau, M. J. Claesson, I. Moreno-Indias, and J. Truu, "Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment," *Frontiers in Microbiology*, vol. 12, feb 2021.

[40] B. D. Topçuoğlu, N. A. Lesniak, M. T. Ruffin, J. Wiens, and P. D. Schloss, "A framework for effective application of machine learning to microbiome-based classification problems," *mBio*, vol. 11, no. 3, jun 2020.

A DATA SOURCES

B Additional Graphics

B.1 3 Dimensional PCoA

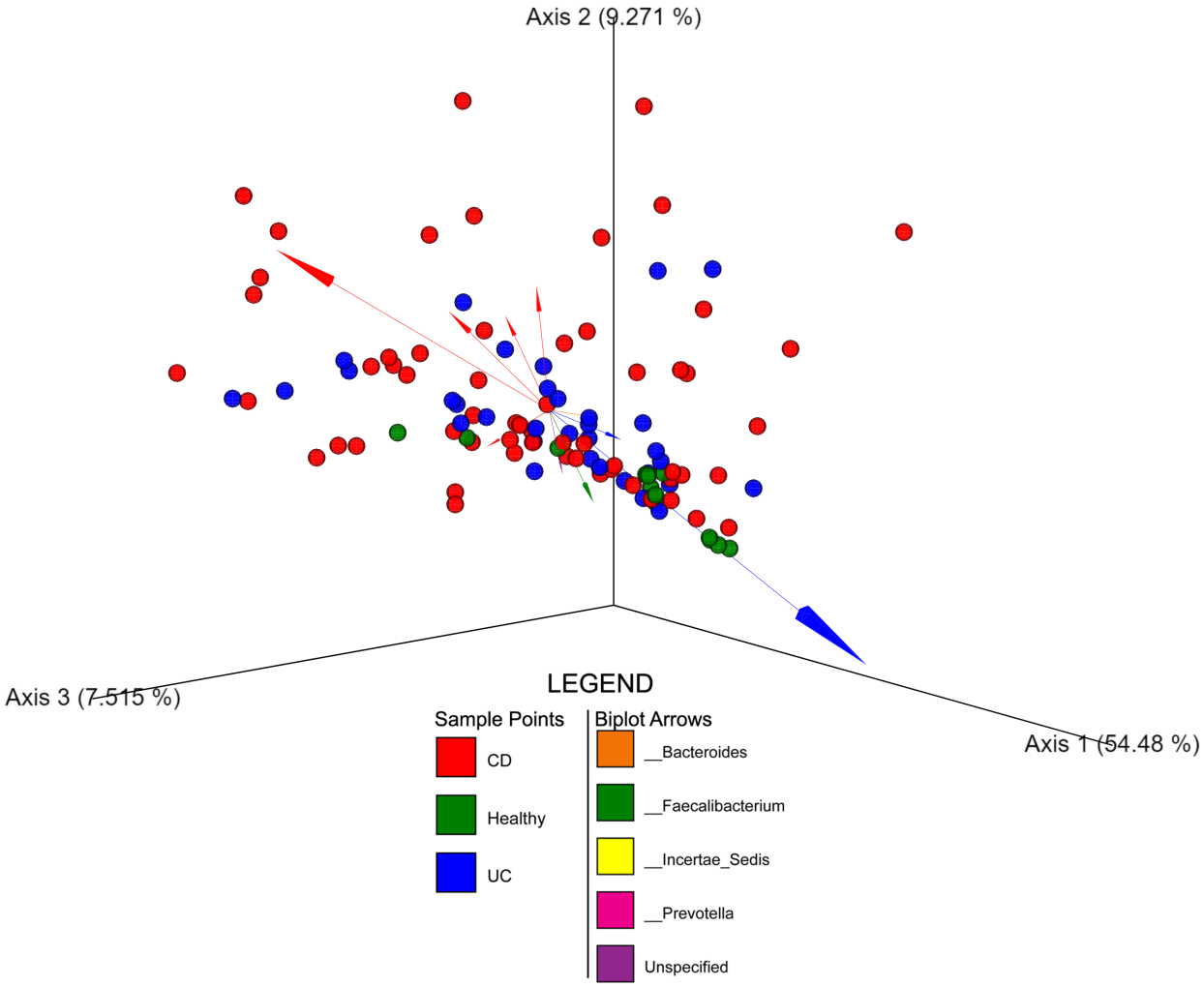


Figure 15: 3 Dimensional version of the Silva 111 PCoA Plot

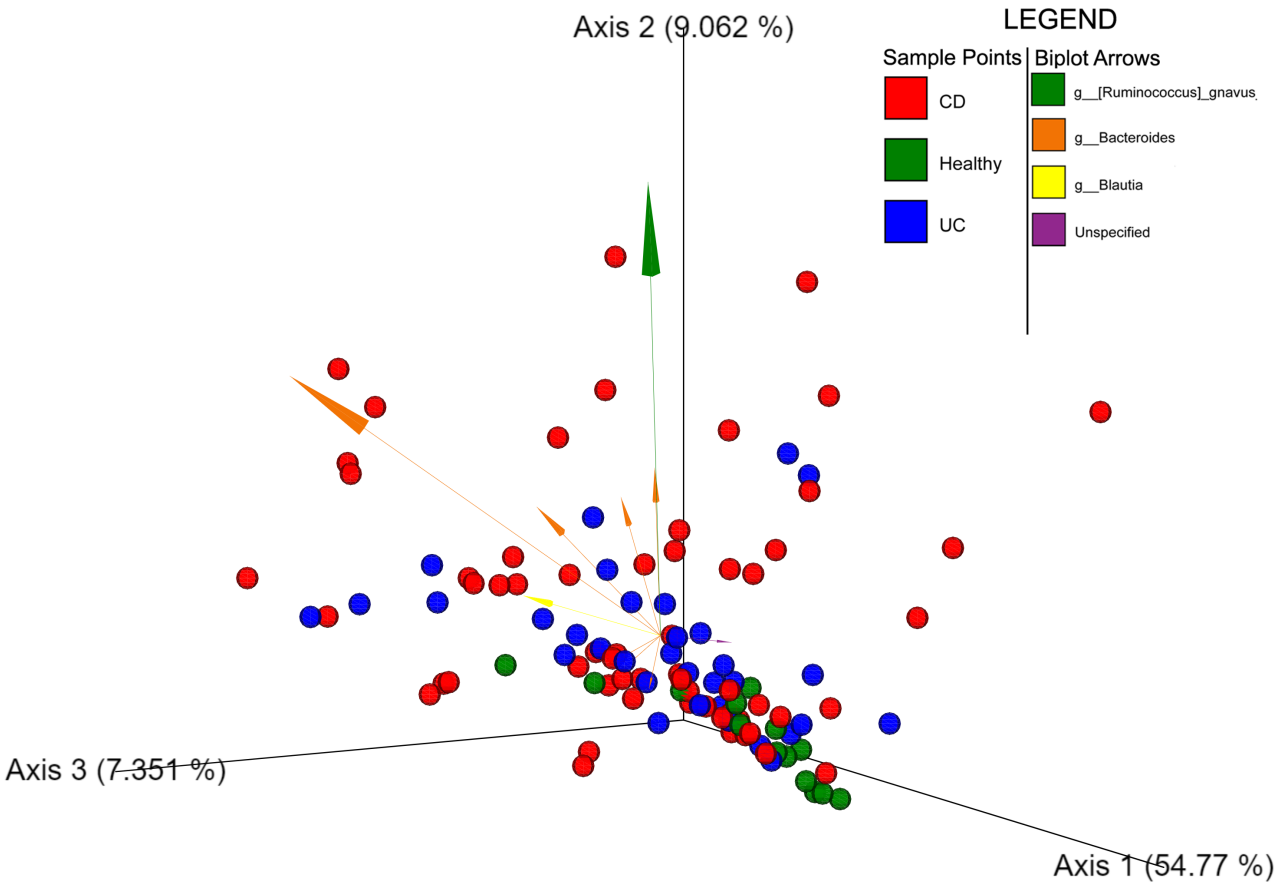


Figure 16: 3 Dimensional version of the Silva 138 PCoA Plot

B.2 Full Size LEfSe Graphics

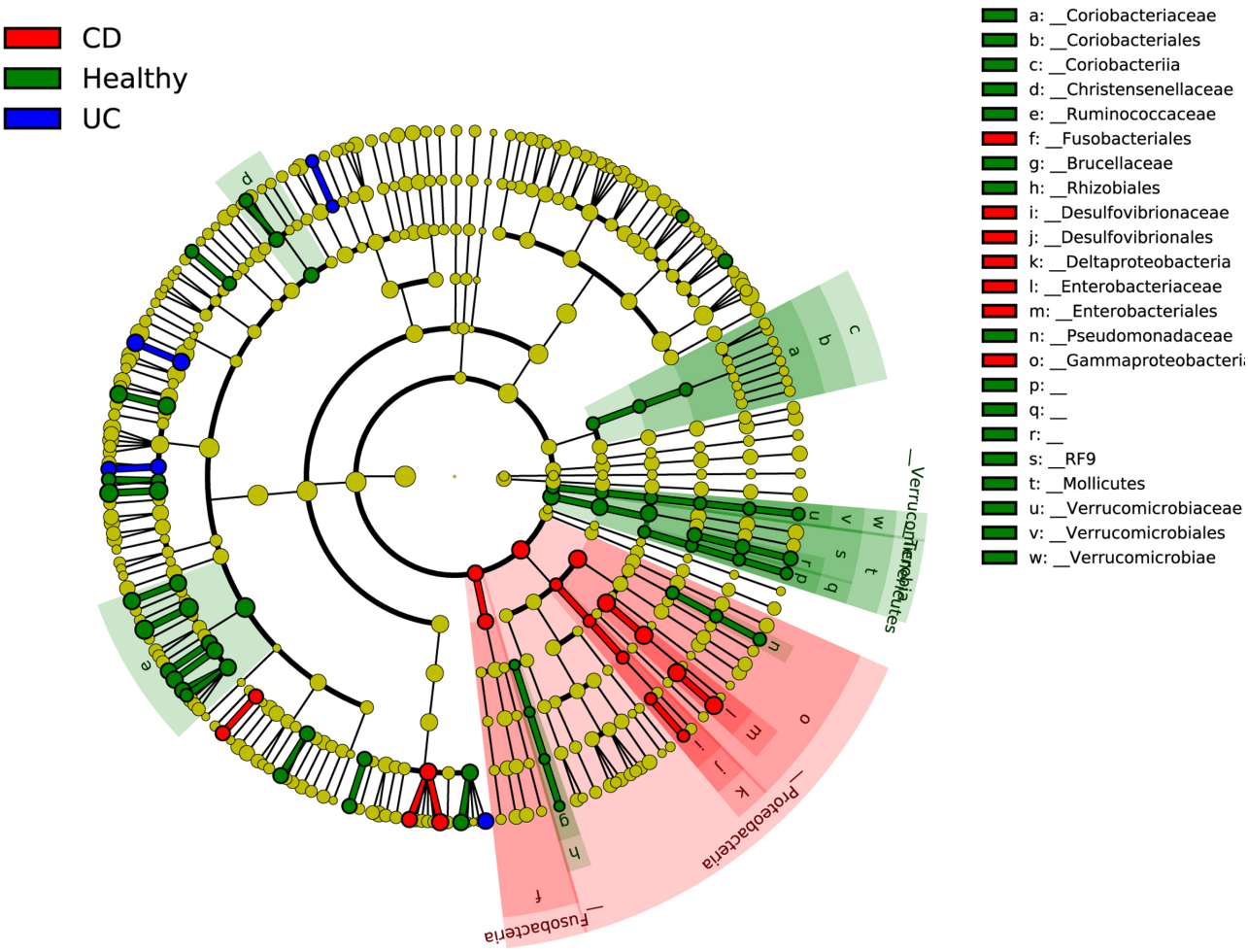


Figure 17: Full size Silva 111 LEfSe Cladogram

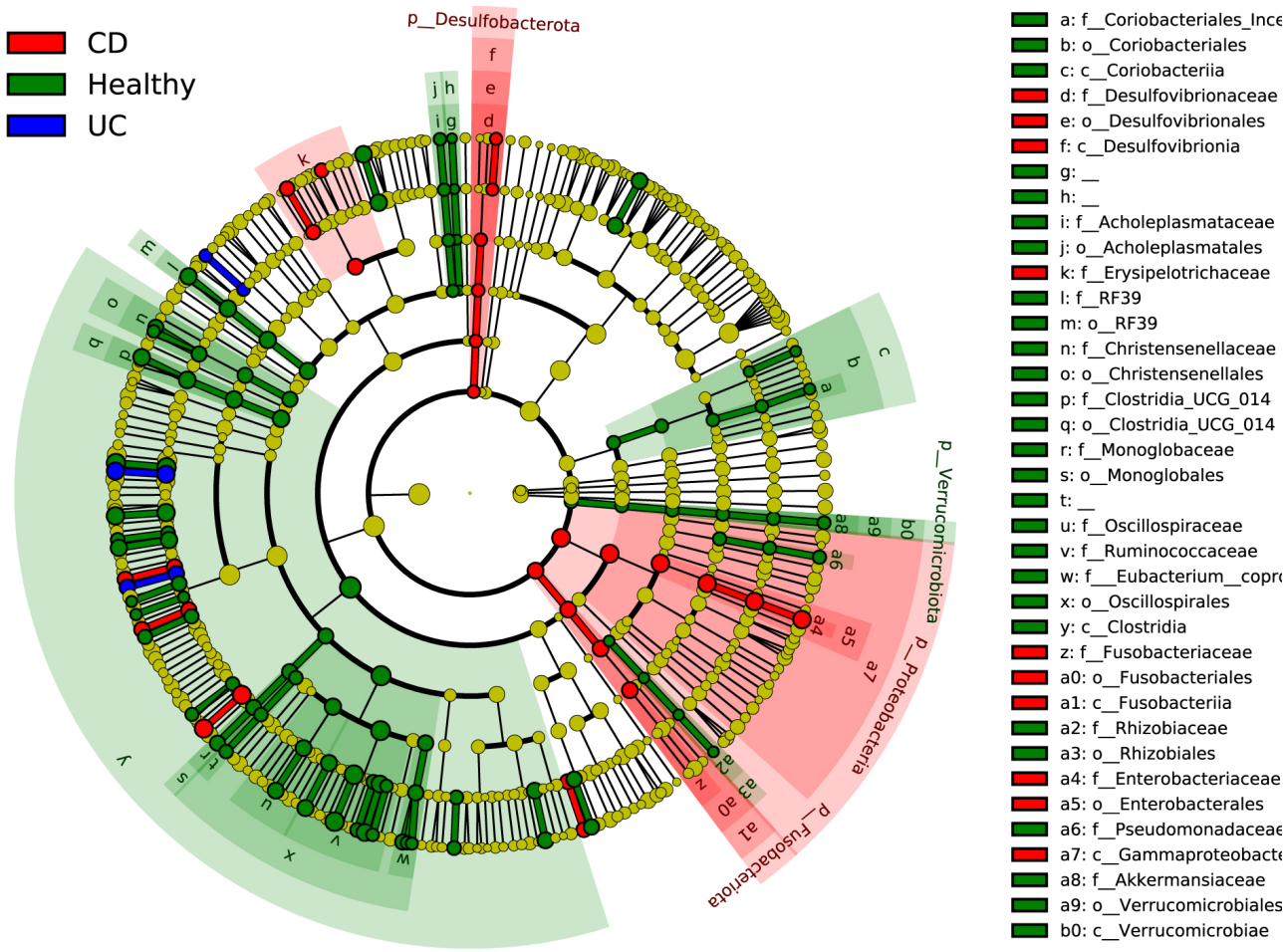


Figure 18: Full size Silva 138 LEfSe Cladogram