MSc Bioinformatics, Polyomics and Systems Biology

Title: "Metagenomic Contigs Visualisation Tool."

Student: Orges Koci, 2155480

Supervisor: Umer Zeeshan Ijaz

A report submitted in partial fulfilment of the requirements for the MSc Bioinformatics, Polyomics and Systems Biology Degree at The University of Glasgow, August 2015

Contents

1. Summary

The purpose of this project is to develop a novel visualization tool for whole genome shotgun sequencing metagenomics that gives the ability to integrate the generated outputs from the recently published metagenomic contigs binning software called CONCOCT and an annotation tool called PROKKA, to explore different aspects of the metagenomics datasets with sample information through a unified Java-based viewer which to author's knowledge was not done before. The CONCOCT pipeline takes a collated assembly of all the samples from a metagenomic assembly software such as IBDA, uses this assembly (contigs), coverage information (reads mapping to each contig), along with the KMER contents to represent each contig as a feature vector, and then performs ordination in reduced space (principle component analysis) followed by clustering the contigs in this reduced space using an unsupervised method called Gaussian mixture model. The developed visualization software in addition to loading the tables from CONCOCT also has built-in parser to visualize the annotation of these contigs which provides the ability of a focused and an interactive exploration of the features of contigs in the context of sample coverages.

The design of such a tool is challenging due to the enormity of metagenomic data in terms of loading numerous features of metagenomic contigs along with their DNA content, and hence in the design and development of a Graphical User Interface, all the necessary optimizations have been implemented. The interface itself was designed to be user-friendly, easily manageable with a high level of interaction and adjustment between the system and the user to spot patterns of interest and to provide sufficient insights for research purposes.

2. Abbreviations

Some abbreviations are extensively used in the text of the report and define several techniques and features. The mostly used of them are listed and annotated right below:

- **CD** – Crohn's Disease
- **CDS -** Coding DNA Sequence
- **COGs –** Clusters of Orthologous Genes
- **GBK –** GenBank File
- **GUI** – Graphical User Interface
- **PCA -** Principal Components Analysis

3. Analysis

3.1 Problem Statement

Metagenomics has emerged as a very useful technique that can be used to analyze microbial communities without the need to culture them in the laboratory. This has all been made possible by the recent advances in Next Generation Sequencing (NGS) technologies that now allow millions of DNA fragments to be sequenced in parallel with a very high throughput. Additionally, majority of the microbes in some environments cannot be cultured and so metagenomics offers a path to the study of their community structures, phylogenetic composition, species diversity, metabolic capacity, and functional diversity.

A widely used approach for the profiling of microbial communities, due to its cost-effectiveness has been the high-throughput short-read sequencing of the PCR amplified *amplicons* such as 16S/18S rRNA genes [8]. However, this approach only tells who is there in the community and not their functional potential, and is affected by its perceived lack of precision in characterizing a microbial community in the presence of amplification biases for various hypervariable regions of the 16S/18S rRNA gene [7].

The drawbacks of PCR amplification based approach have led to the usage of another popular approach for studying the species composition and diversity of natural bacterial communities, the approach of Whole Genome Shotgun (WGS) sequencing of environmental DNA, where the total genomic DNA content is sheared/fragmented (ultrasonification, nebulization etc) and is then sequenced [4]. The WGS analysis then requires an additional step to assemble overlapping DNA fragments. WGS analysis provides robust estimates of microbial community composition and diversity while it does not involve the amplification bias associated with the choice of hypervariable regions for 16S/18S rRNA genes [11]. More importantly, the WGS approach has benefits over the 16S/18S rRNA analysis, where the former is just a barcoding study and provides only diversity information of a microbial community and, the later approach gives additional information on the genomic content as well as metabolic potential of the species thus offering a more thorough investigation into microbial community profiling.

To cover all the metagenomes in a given sample, deep sequencing and high throughput NGS technologies such as HiSeq is often required which generates an enormous amount of data that requires more space and memory than 16S/18S rRNA metagenomics, highly parallelized bioinformatics algorithms, and efficient data processing techniques to obtain meaningful scientific results. There is also a need to merge taxonomic information with the functional annotation of the genomic contigs and therefore, there is a lot of stress on the visualization aspect of WGS sequencing data. This is what we intend to address through this project by having a single interface unifying the diversity information in terms of sample coverages with the DNA content of the contigs.

Nevertheless, despite the importance for the development of a software that meets the requirement described, not many attempts have been made to have a single interface to see different analyses in entirety. To date, the existing visualization software that

incorporate data from WGS sequencing analysis are small in number while in general terms they are limited from the aspect of the information they provide. Particularly, in conventional genome viewers like MGAViewer [12] it is possible for one to analyze at most two genomes together for comparative genomics. Moreover, another tool that implements an advanced analysis and visualization platform for 'omics data called Anvi'o [2] lacks an interface, offers very little interactivity, doesn't allow merging of annotations of genomic regions although one can visualize phylogenetic trees and coverages. These drawbacks leave a room for improvement in terms of the functionalities to be offered by a metagenomic contigs viewing software and create a higher need for a software with graphical user interface that can hook onto multiple data outputs from WGS sequencing analyses tools.

3.2 Aims of the project

In the current project, the main objective was to develop a novel viewer for Whole-Genome Shotgun sequencing based analysis and incorporate all sources of information covering different aspects of this analysis. We have chosen the output from CONCOCT pipeline that enables one to bin the contigs for a species together on a reduced space feature based representation of these contigs. This reduced representation can then be visualised as PCA plots in the viewer with cluster labels. Additionally, beyond looking at contigs and their clustering, a variety of features were pursued to be implemented, partially or totally not offered in existing metagenomic visualisation tools, in order to develop a program with a wide range of abilities for a research in metagenomics.

CONCOCT software [5] is a pipeline developed in the research group of Dr. Umer Zeeshan Ijaz, and is useful for binning metagenomic contigs by coverage and composition and provides the majority of information visualized in our viewer. Even though CONCOCT does a good job at binning contigs, it produces data tables and does not provide a structured interface to present the large amount of results from the generated data. For this reason, the tool developed in this project principally aims to extend CONCOCT by implementing the interface that will visualise the data generated from the CONCOCT pipeline, as well as integrate annotation data for the contigs

through annotation tools such as PROKKA [9] and PRODIGAL [3], and also present appropriately these results in a coherent and structured way for multiple samples.

4. Product

4.1 Software Design

The software developed in this project is intended to be used for research purposes, in combination possibly with other tools and targeting audience with little or no background in bioinformatics. As a result, it is very important for the software to be designed in such a way that it allows flexibility to the user, has simple dependencies and provides a wide range of features in a single and powerful interface. For these reasons, the viewer was developed as a Java-based tool which can satisfy the above conditions. Java has cross-platform portability since a compiled Java program is able to run on all platforms for which there exists a Java Virtual Machine (JVM). This holds for all major operating systems, including Windows, Mac OS and Linux. In addition, Java is very flexible for designing GUIs by providing a rich set of packages for graphics manipulation and it is relatively easy to place the graphical components on the interface.

We already had an idea about the final interface in mind from very little visualization that was done as 2D plots in R for data tables from CONCOCT. This included visualisation of clusters in the first two PCA dimensions, as well as generation of annotated diagrams for contigs with enzymes information and other characteristics from an independent python script written by Dr Ijaz. Thus, the general design and graphical interface of the current model is mainly based on these examples and attempts to follow their structure in terms of data representation. Nevertheless, further improvements (e.g., changing the view on sample basis) and additional features (i.e. coverage information plots) were implemented during the project in order to create a very handy and quick tool to look up relevant information. In addition, a set of adjustment tools were implemented such as zoom-in/out buttons, buttons to change the size of the elements in the panel and the background colour and functionality to save the displayed outputs

(plots/sequences) to an external file for further use. All these features guide the user to find patterns in the samples more effectively.

Loading large-scale data

The first design consideration in any visualisation tool is the loading and management of possibly large amounts of data. For this reason, the software is developed in such a way that it attempts to minimize loading time by having multiple passes, some extracting only summary of the data, some executed as a result of user interactivity where the details are required (e.g., when there is a large sample space). Thus, advanced data structures and techniques were considered to improve the general performance of the software such as dynamically set multidimensional `ArrayLists` and an optimized polymorphic object-oriented design to extract multiple information from the same data files (e.g., GenBank file).

Another optimization is a threshold that was applied to limit the dimensionality of the PCA plots (the lower dimensions offer very little variability and can be removed). Also, in order to avoid memory deadlocks, the loading of coverages per sample was done interactively without the need to keep all the coverages of the samples in the memory. The similar strategy was adopted in other views of the software to optimize for speed (without buffering) as well as to reduce memory footprint.

However, it is worth mentioning here that there is a limit to all these optimizations and beyond a certain point (for large-scale metagenomic WGS datasets comprising more than hundreds of samples and thousands of contigs), the current software will start buffering that may lead to a reduced performance.

Graphical User Interface

We have considered a multiple document interface (MDI) as opposed to the single document interface (SDI) where you have one window per functionality to avoid cluttering up of windows on the desktop and also to have a single window with multiple panes to visualize all the information. This makes the GUI more coherent, simpler to user and easier to operate.

Upon the initial execution of the software, the basic window is presented to the user which includes a basic panel on the top side for the main toolbars, a bottom panel which contains a slider in one side with a `ComboBox` and two main panels which essentially split the GUI in two main windows. Moreover, two additional panels hidden from view were also loaded on the left side and were used for coverage information and get displayed only when the user enables them. On the upper side of the GUI, a `FileBar` is included with a menu option to import all the required files to the software which not only contain the data to be displayed initially but also update and adjust several elements of the GUI later after they are successfully imported. Thus, the basic parts of the software are initially blank, some tools e.g. `JSliders` are not set, they contain default sample values and the interaction and visualisation starts when the user has provided the necessary files (figure 1).

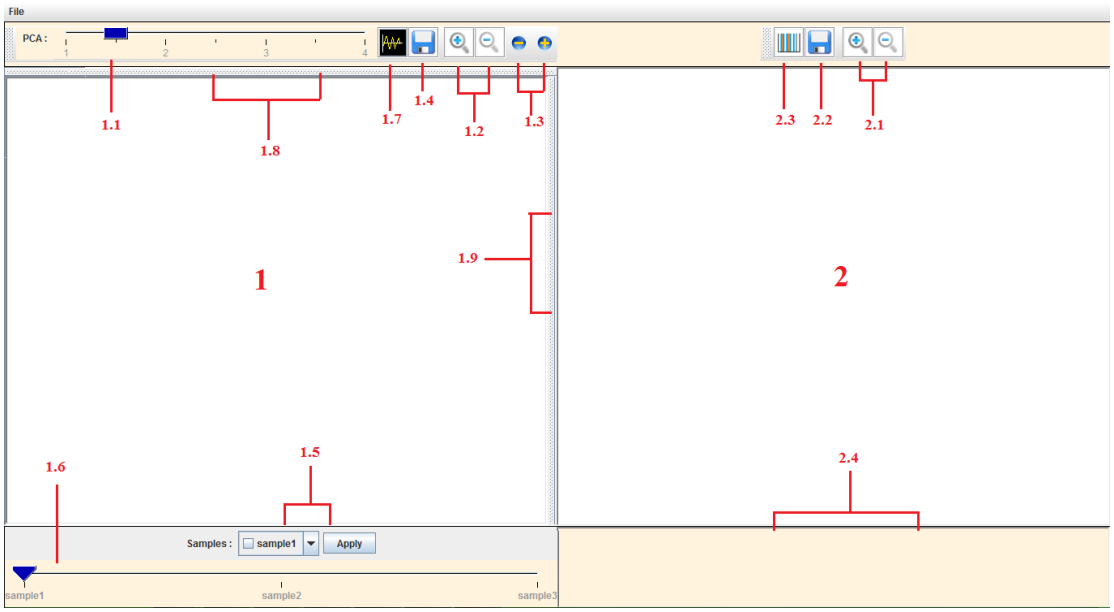


Figure 4.1.1: The basic GUI of the model upon its initial execution. Annotation numbers indicate the separation of functionalities between the two main parts.

As mentioned before and can be seen in figure 1, the model is split in two main parts (numbers 1 and 2) each one of them is connected to a specific toolbar which is located right above the main visualisation panels and the one below them. The functionality of each annotated object will be explained accordingly right below:

- **1** : Located in the left part contains the space where the data of the PCA analysis and the visualisation of contigs get displayed to the user after loading the PCA tables from CONCOCT. In the toolbar on the upper side there is a slider and several buttons which can update and change the characteristics of the PCA plot.

- **1.1** : This slider shows the number of dimensions of the PCA analysis, and the total number in view can change depending on the file that is loaded (something that is variable from CONCOCT software). The user can scroll between the dimensions (two at a time) selecting and inspecting the variation of contigs cluster while doing so.

- **1.2** : Zoom in/out buttons that can change the size of the plot accordingly to remove cluttering of points.

- **1.3** : Buttons to change the size of the individual points for contigs with the same aim of removing cluttering as in 1.2.

- **1.4** : Button used to save the current clustering view as a PNG image for further use.

- **1.5** : `ComboBox` containing the labels of all the samples used for the prior analysis through CONCOCT with a `CheckBox` beside each of one them. The user can choose any of the samples he wants to inspect and then by pressing the "Apply" button can update the slider window (1.6) with only the chosen samples.

- **1.6** : The slider is updated with samples chosen after selecting samples in 1.5. In the initial execution of the software this slider is disabled and contains some default example labels.

- **1.7** : Button which is used to enable/disable the hidden panels that display the coverage values of the contigs as barplots superimposed onto both axes for a chosen sample from the slider in 1.6.

- **1.8 - 1.9** : The location of hidden panels to display coverages.

- **2** : The right main part of the GUI where all the annotations and the features of a chosen contig (by left-clicking on a particular contig on the left panel) are displayed. Similar to the left panel, there is a toolbar on the top side and a panel, blank initially, drawn in the bottom.

- **2.1** : Zoom in/out buttons with the same functionality as described in 1.2.

- **2.2** : Button to save the displayed annotation track as a PDF file.

- **2.3** : Button to save as a FASTA file the selected CDS regions (user can select multiple CDS regions on the length of a contig by left-clicking on them) on the annotation track.

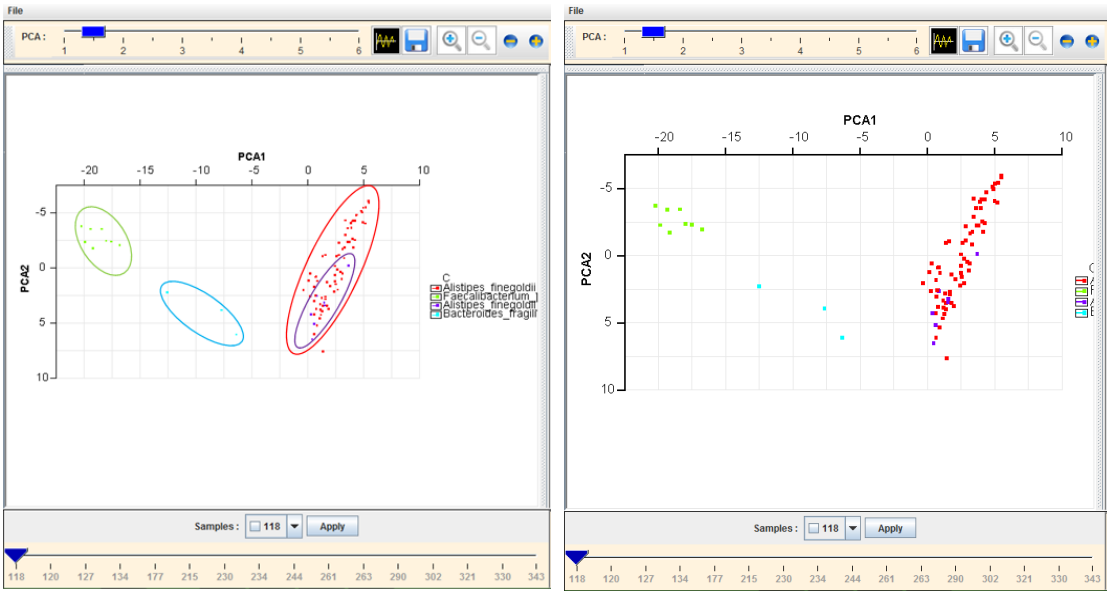
- **2.4** : Panel where the complete nucleotide sequence of the chosen contig is shown.

4.2 Features

In this section we run the visualization tool on a dataset (default dataset shipped with CONCOCT) and show how the interface looks like. This dataset comprises contigs belonging to four species for a total of 16 samples.

Principal Component Analysis

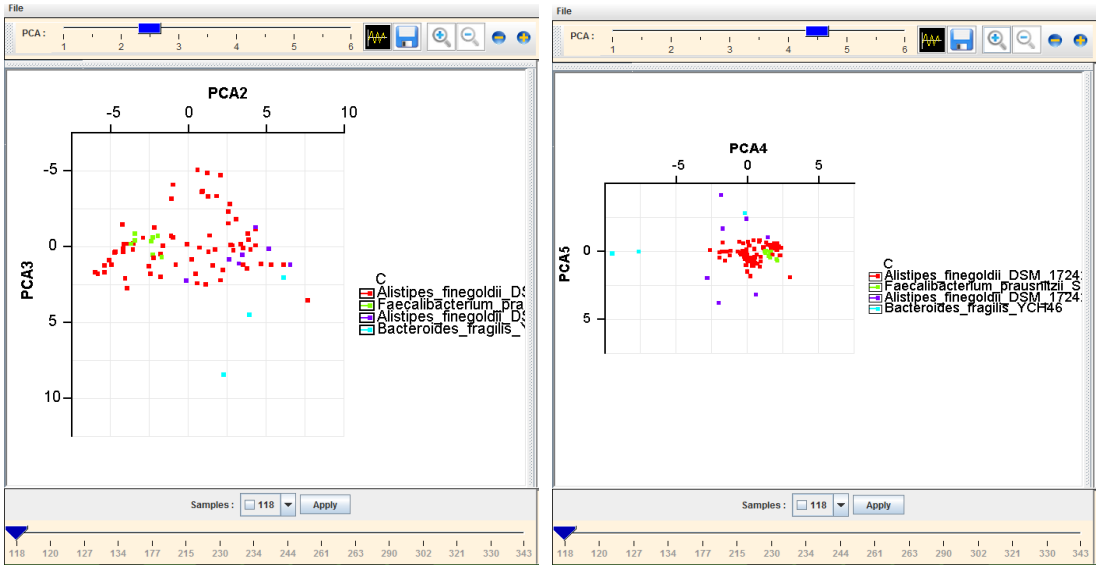
One of the major features that are implemented in the tool is the visualisation for PCA analysis. Below are presented some examples of this feature:



(a) *(b)*

Figure 4.2.1: Two different screenshots of the same execution displaying the main output for the PCA projection over the first two principal components. The main right panels are omitted since they are blank at this point in time. On the left side (a), the analysis detected four clusters which are annotated with ellipses that were manually added for description purposes. The clusters are annotated on the right side of the panel below the "C" letter with distinct colours accordingly and with taxonomical information at species level (e.g. Alistipes_finegoldii, Bacteroides_fragilis etc.). On the right side (b) there is a better focus in the PCA plot as a result of zooming in and increasing the size of the contigs (1.2 and 1.3 in the previous section)

After the demonstration of one of the basic visualisation parts of the software another useful feature is the ability of the software to browse through different PCA dimensions. There were a total of six PCA dimensions obtained from CONCOCT with plots of two sets of dimensions given below:



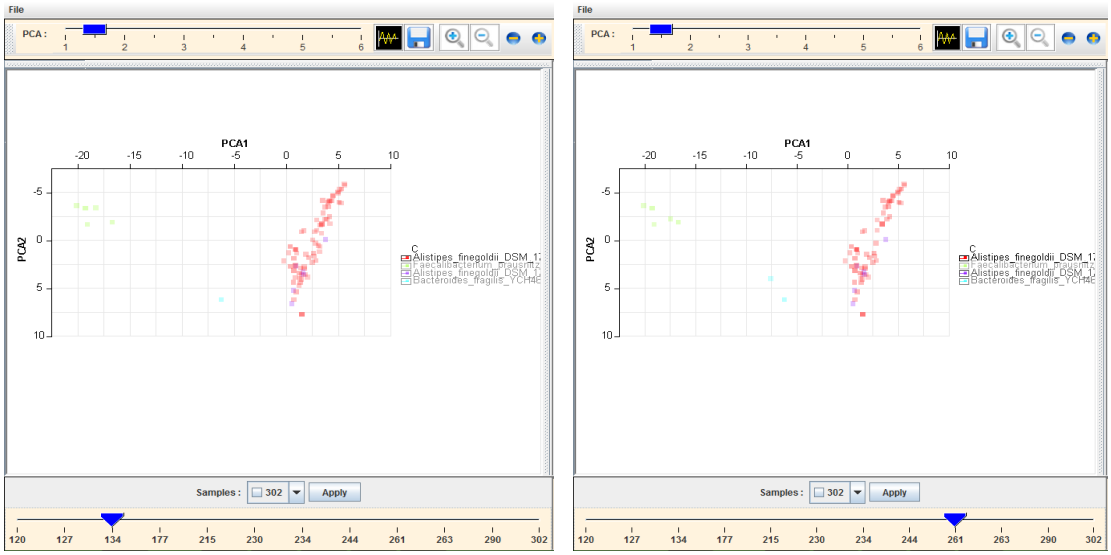
(a) *(b)*

Figure 4.2.2: Two example screenshots of PCA plots over a set of different dimensions. On the left side (a) the choice for the components to be plotted is set respectively to the slider to be 2-3 and the corresponding plot is generated. Similarly, on the right side (b) the selection for the components is set to 4-5 and the corresponding plot is displayed. The lower order dimensions have smaller eigenvalues and show less variability as can be seen in (b).

Coverage Information

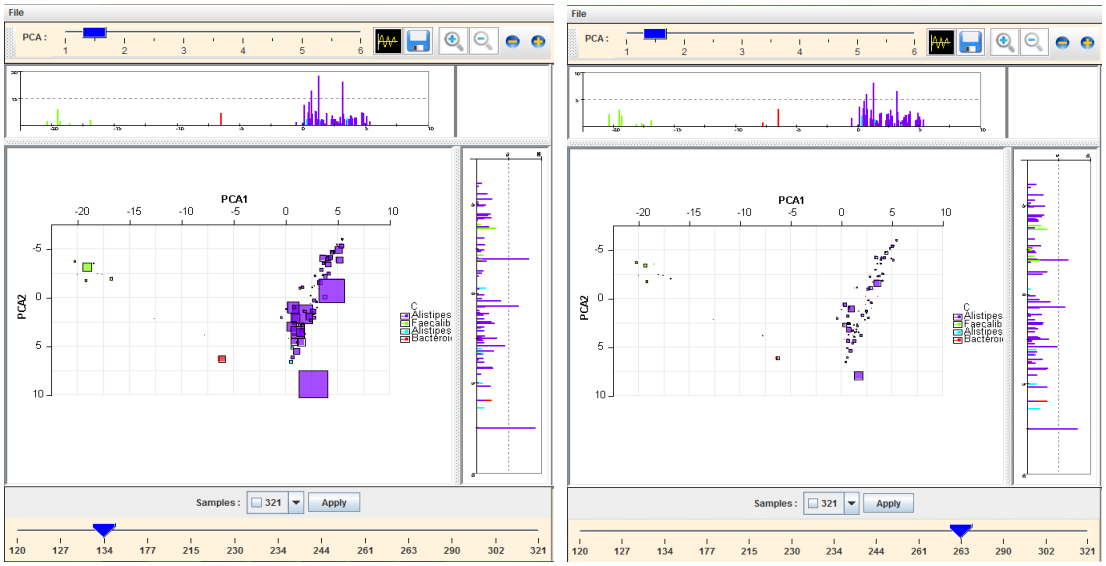
To display the coverage information, a variety of visualisation approaches were implemented to allow maximum discrimination between contigs in terms of coverages for a given sample and between different conditions: adjustment of the opacity of the contigs being one and adjustment also of the size of the contigs being the other one, although the latter is not recommended for a larger dataset. Another approach and basic feature of the software is the projection of the coverage information as barplots on the x and y axis of the PCA plot.

Additionally, in order to distinguish the species that are most abundant, the opacity value of the labels of the cluster with the taxonomical information in the legend is also changed to highlight the important species. This approach can be quite useful for a dataset with numerous clusters where it would be difficult to distinguish between the clusters on the PCA plot with naked eye. Some examples are given below:



(a) (b)

Figure 4.2.3: Example outputs displaying coverage information for two different samples. The level of opacity represents the value of coverage. On the left side (a) it can be seen that for sample "134", the cluster represented by red colour is the most abundant i.e. "Alistipes_finegoldii" while the labels of the other clusters are dimmed. Similarly, on the right side (b) "Alistipes_finegoldii" is still the most abundant one but other species are also highlighted slightly.



(a) (b)

Figure 4.2.4: Example screenshots for displaying coverage information based on contig sizes and also their projections on both axis. In the left figure (a), sample "134" seems to have higher presence of "Alistipes_finegoldii" which is coloured in purple and smaller sizes of boxes for other clusters. In the right figure (b) even though "Alistipes_finegoldii" is still higher in abundance although relatively less as compared to other species.

It is worth mentioning that there are numerous combinations one can consider to get the best view of the coverages and some of these are bundled up as themes in a right-click pop-up menu for easy changing.

Annotation Tracks

Since the GUI of our model is split between two main windows and the second panel gives detailed information for the contigs displayed on the left panel, the software supports "Listeners" which capture the position and the motion of the mouse cursor. Thus, when the user moves the cursor the position is obtained and its coordinates are checked if they "hover" above a certain object on the PCA plot i.e. above a contig. In case this holds, then a label is shown for that contig with its name.

Next, if the user wishes to inspect the specific contig then he can click on that and this action activates the right panel where the data for the annotation tracks get displayed. The same approach with the mouse motion capture for defining the selection of the visualised objects is also followed in the right panel. Below, there are some example screenshots highlighting these features:

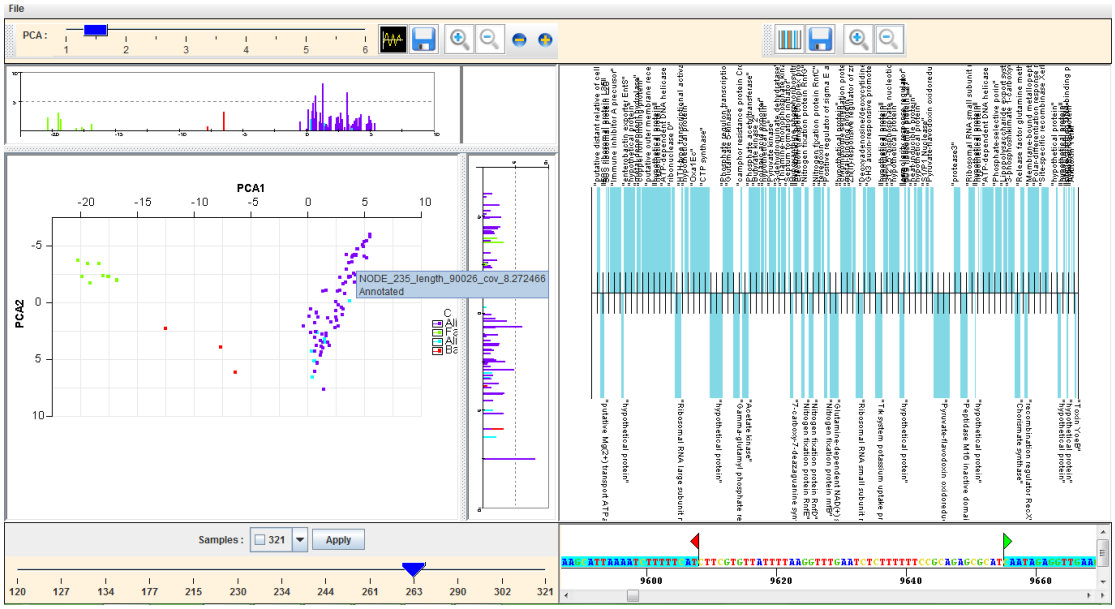
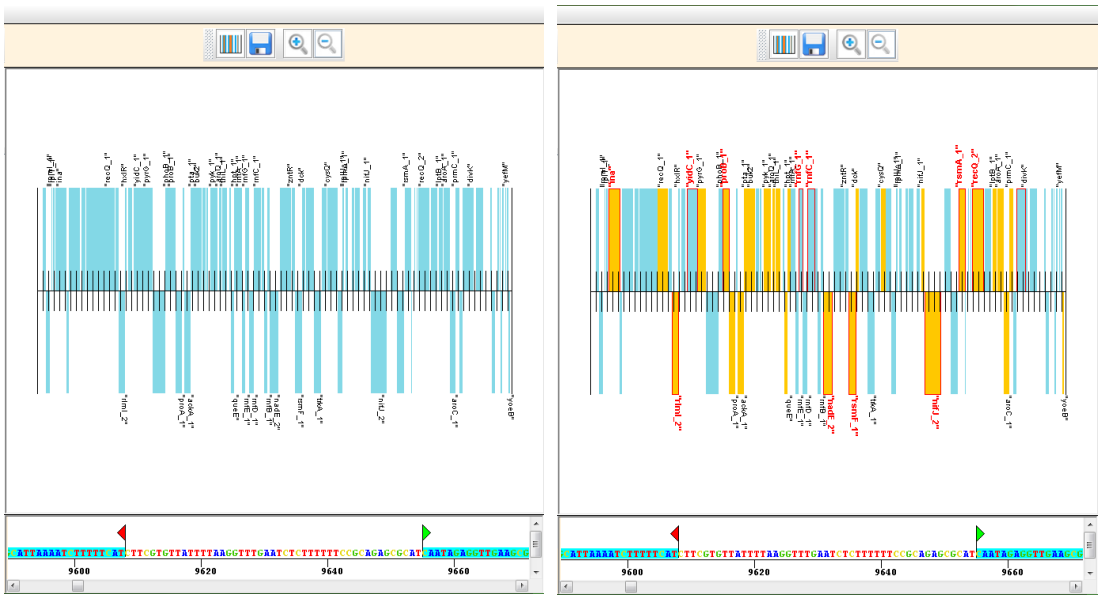


Figure 4.2.5: Screenshot showing the selection of a specific contig from user and the visualisation of the annotation tracks for that contig. The label in left panel displays the contig's id. In the right panel the cyan boxes indicate the CDS regions along the length of the contig followed by their product as a label above each one of them. Similarly, below the right panel, the nucleotide sequence of the contig is displayed visualising CDS regions in cyan colour and specifying their start and end by a green and red flag respectively.

The default labels that get displayed for the CDS regions mentions their "product" picked from GenBank file, although it is possible to switch to labels showing gene information instead. Moreover, the user can inspect the enzyme and/or COGs information within the CDS blocks by just right-clicking and choosing from a pop-up menu. In addition, if there is a need to explore some specific CDS regions further, then a set of them can be selected manually by the user and can get extracted as an output file with their sequence details (say a user wants to analyse how a particular homologous gene differs between different species through 3rd party tools). Example screenshots are shown below:



(a) *(b)*

Figure 4.2.6: In figure (a), the labels for the CDS regions are set to gene labels. In figure (b), the yellow boxes indicate CDS regions that contain enzyme information (not all protein encoding regions are enzymes, and those that are have an EC number associated with them). The selected regions to be exported to a file have a thick red outline with red labels thus distinguishing from those that are not exported.

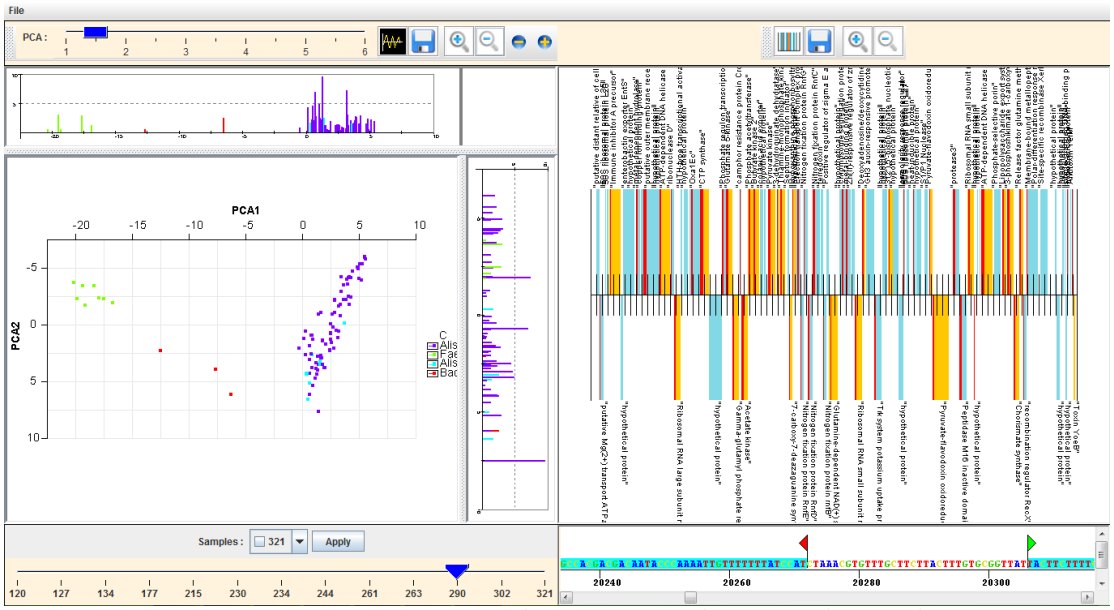


Figure 4.2.7: Screenshot displaying most of the described features of the software. On the left panel, the feature for COGs information is also shown where COGs are visualised by red boxes within the CDS regions.

4.3 Documentation

Execution and Dependencies

In order to run the viewer, the requirements are quite simple. At first, it is necessary to ensure that the computer where the software is going to be executed meets the java version required. The versions that support the requirements of the software are these of "1.8.0" e.g. "1.8.0_51" etc. The user can either click the associated JAR file or from the command-line interface by typing:

- `java –jar CViewer.jar`

Another important consideration is the issue of memory available to the software. In order to give more memory to the software one can manually set the size of Java heap. A way to achieve that is to set the Java Environment Settings (for windows) :

- `-Xms512m` - to assign a minimum of 512MB memory for the Java.
- `-Xmx1024m` - to assign a maximum of 1GB memory for the Java.

The above can also be given as an argument to the command line instruction.

The settings for Java described above were optimized on a Windows machine with 3 GB of RAM memory, and on Intel Pentium P6200 2.13GHz processor and a 64-bit operating system, and may vary for other machines.

List of files required for the tool

After the initial execution of the software, the GUI will be displayed as described in previous sections. At this point, in order to proceed with the visualisation there is need to import a set of files that will provide the necessary visualizations. These are listed below:

- PCA file : The N x D CSV file that contains the D PCA components for each contig and is obtained from CONCOCT software.

- Clustering file : The N x 1 CSV file that gives the labelling of which cluster each contig belongs to and is also obtained from CONCOCT software.

- Coverage file : The N x S TSV file that contains the information for the coverage values for the S samples of the dataset, also obtained from CONCOCT software.

- GenBank file : The file that contains all the annotation tracks for the N contigs of the dataset. This is obtained from PROKKA software.

- COGs file : The files that contains the information for the COGs found in the dataset. Obtained by running the CDS region extracted from PROKKA against NCBI's CDD database with output in BLAST format.

- TAXA file : The N x 6 CSV file that contains the taxonomical information of each of the contigs at Phylum, Class, Order, Family, Genus, and Species level. It

is obtained through TAXAassign software [5], a tool for annotating nucleotide sequences at different taxonomic levels using NCBI's Taxonomy.

5. Evaluation on Real Dataset

We tested the pipeline to deal with large-scale WGS dataset and incorporated changes to the software design and the interface both in terms of underlying algorithms but also placement of graphical elements without the loss of information.

5.1 Prior work, Dataset Preparation

The dataset used for our testing the pipeline is related to Crohn's disease patients and healthy individuals and was provided by Dr. Ijaz from his ongoing research [1]. Crohn's disease, is a type of inflammatory bowel disease (IBD) that may affect any part of the gastrointestinal tract from mouth to anus [6]. It is caused by a combination of environmental, immune and bacterial factors in genetically susceptible individuals. It results in a chronic inflammatory disorder, in which the body's immune system attacks the gastrointestinal tract possibly directed at microbial antigens.

The WGS dataset comprised of two types of samples: from individuals who have Crohn's disease but who have also gone through a two-months long nutritional therapy called Exclusive Enteral Nutition (EEN); and from healthy individuals just to compare how the gut microbiota changes as a result of treatment and how it differs from healthy baseline. The WGS sequencing dataset comprised twenty individuals with 12 samples for CD patients at the end of treatment and 12 for healthy individuals. The shotgun metagenomics samples were prepared with the Nextera XT Prep Kit (Illumina, FC-131-1096, UK), and Illumina dual-barcoding Nextera XT Index kit (Illumina, FC-131-1002, UK). Sequencing libraries were pooled in equimolar concentration and quantified with the KAPA SYBR® FAST qPCR Kit (Kapa biosystems, KK4824, UK) then, loaded onto both lanes of a rapid run flow cell at 10 pM concentration. Clusters were generated on-board a HiSeq 2500 (Illumina) instrument and sequencing performed using TruSeq

Rapid SBS Kit reagents (Illumina, FC-402-4001, FC-402-4002). Sequencing was performed following a paired-end 150 cycle recipe. All these details were provided by Dr Ijaz.

5.2 Results

In order to assess the efficiency of the software in terms of execution time, Crohn's disease dataset comprising all the files mentioned before (N=158.432) were used. In the first pass to load summary information (all files), the tool took on average 11,745 s for 3 runs (each run differs by loading several other software in addition to the visualizaton tool). If we also include the time it takes to initialize and populate the graphical elements on the interface, the average time then increases to 30,729 s for 3 runs. Thus half-a-minute of loading time for 156K contigs is reasonable for most WGS datasets. Part of the evaluation of the software was also to verify that the tool is indeed able to describe in a proper way the large amounts of data contained in the dataset used, to distinguish between clusters (260 in this case), show DNA content information for the entirety of the specific contig and generally to check the functionality of the implemented features (figure 5.2.1). We also wanted to see that with the visualization implemented, we are able to pick up subtle differences between CD and healthy individuals in terms of contigs that are up/down regulated in one of these conditions.

The PCA plots were examined (a total of 12 dimension) to see if we can find any distinct clusters in the datasets and it can be seen from figure 5.2.2 that indeed that is the case as two clusters are sticking out. The distinction goes away as we look at lower PCA dimensions.

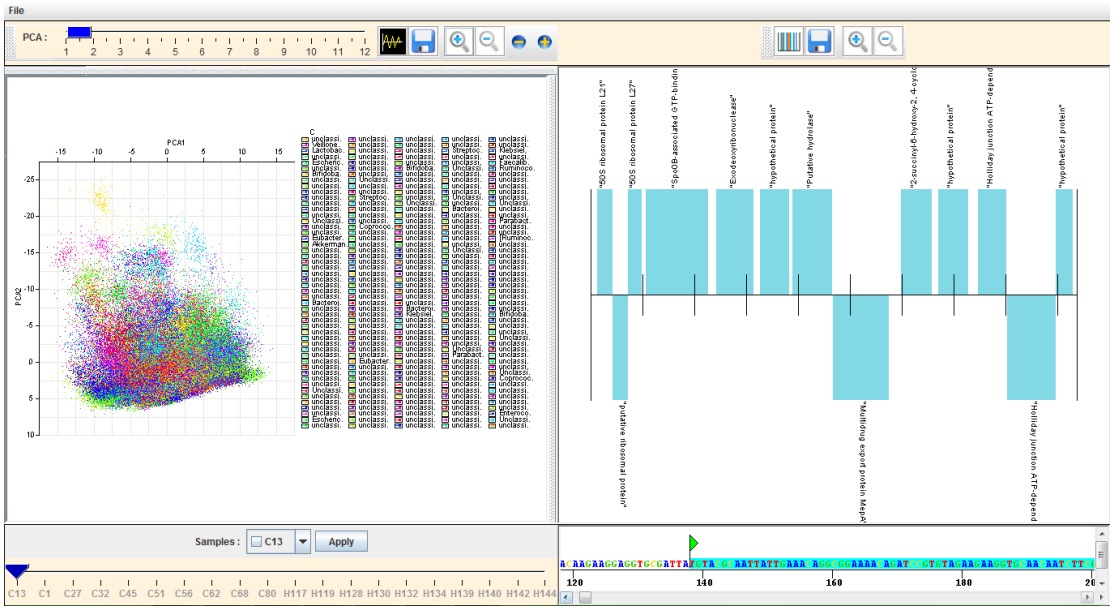
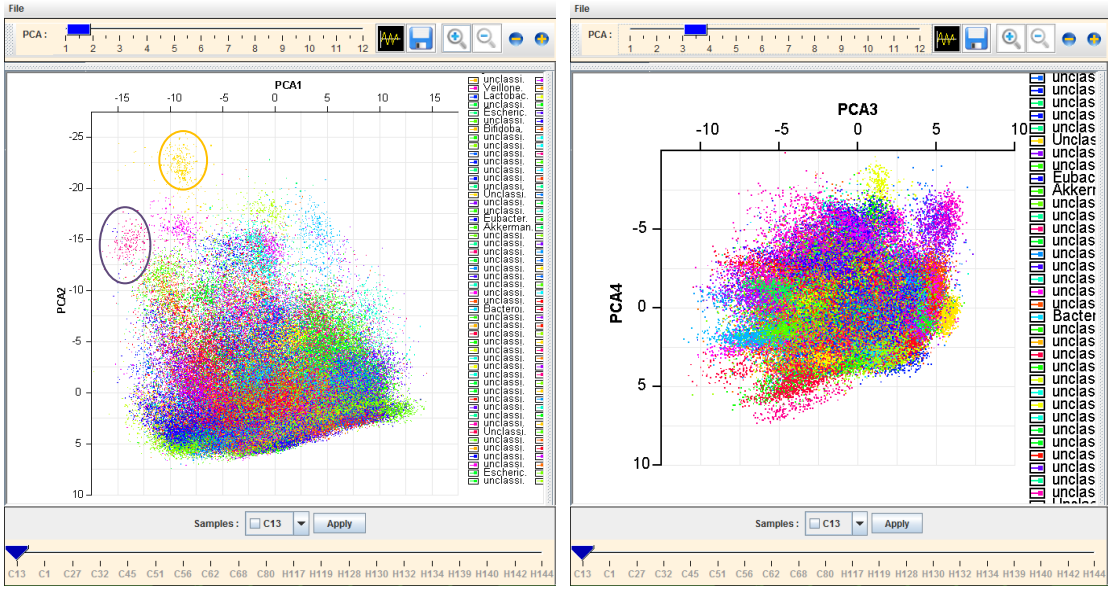


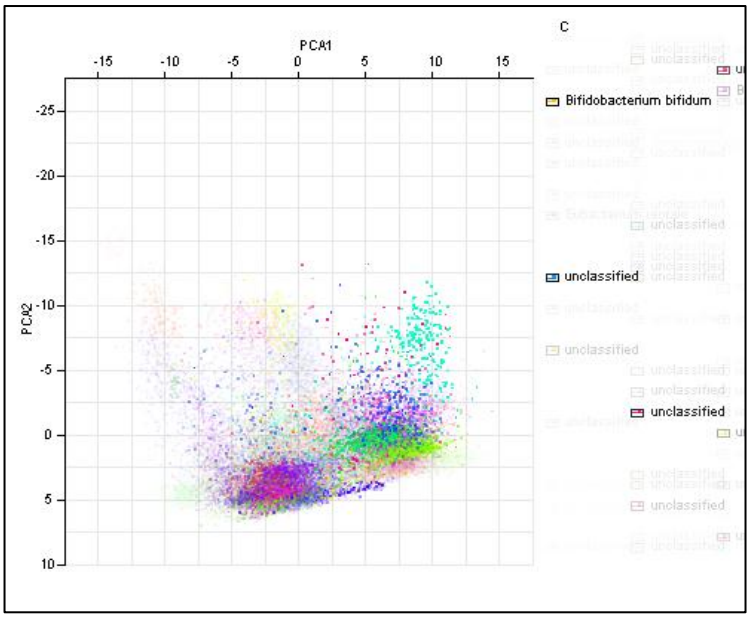
Figure 5.2.1: Screenshot generated from the viewer upon the initial execution. We can notice in the left panel that the contigs are displayed properly in a PCA plot while they are coloured according to the cluster they belong to and with taxonomy information. On the right side are displayed the CDS regions and the DNA sequence for a chosen contig.



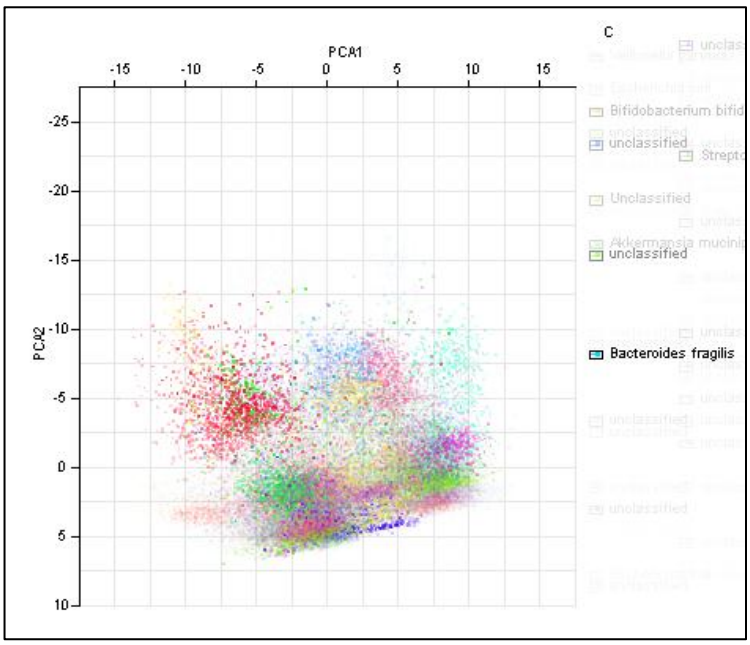
(a) *(b)*

Figure 5.2.2: Example screenshots for investigation of patterns among the data through PCA. We can notice that two clusters are distinct in the first two dimension (a) and not otherwise (b).

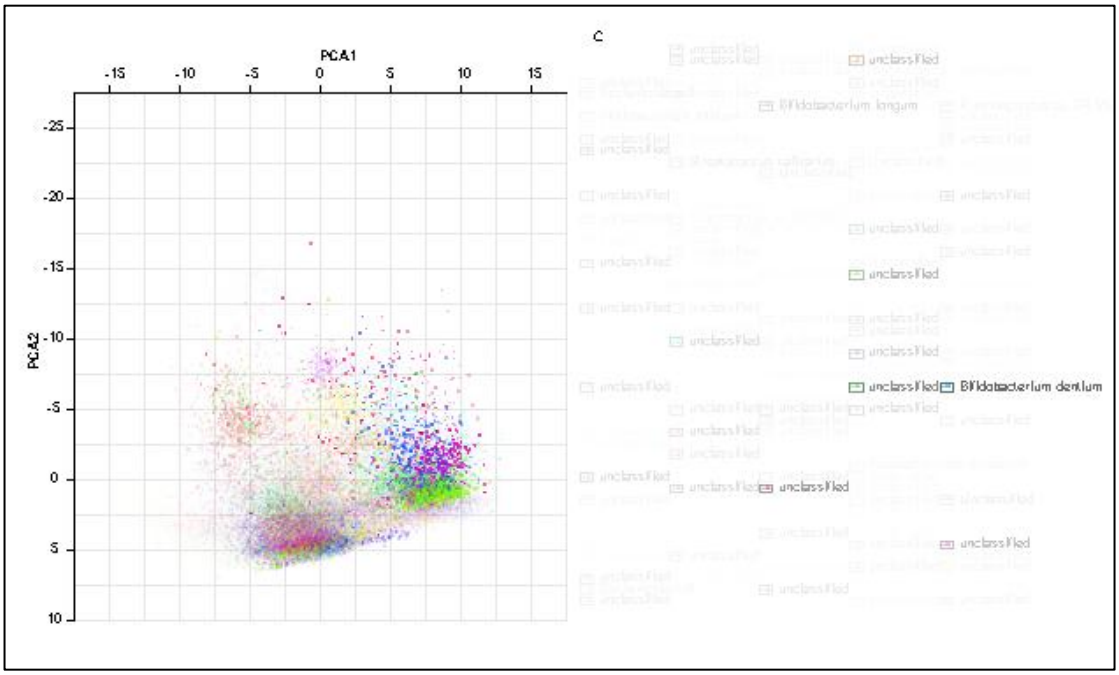
Next, the analysis continued by focusing on specific samples which described either a CD patient or a healthy individual and by investigating the coverage levels of the contigs for these samples. The analysis for these samples revealed for the case of CD patients some of the microbes such as *"Bifidobacterium bifidum"* and *"Bacteroides fragilis"* to be up regulated. Moreover, for the case of healthy individuals microbe *"Bifidobacterium dentum"* was found to be up regulated. These results can also be supported by relevant clinical trials that verify the influence of *Mucosal Bifidobacteria* in Crohn's disease [10]. Some representative parts of this analysis are displayed and annotated below:



(a)



(b)



(c)

Figure 5.2.3: Screenshots presenting the microbes found to be up regulated in Crohn's disease dataset. The first two figures describe the results for two CD patients where "Bifidobacterium bifidum" (a) and "Bacteroides fragilis" (b) are up regulated as it is indicated from the level of the opacity of the corresponding clusters. Similar, figure (c) describes a healthy individual where "Bifidobacterium dentum" is highlighted indicating that it is up regulated.

6. Discussion

We have developed a Java-based visualization tool for whole genome shotgun sequencing metagenomics analysis which offers a single interface to explore differences between samples generated under different conditions (e.g., samples for CD individuals VS healthy individuals). The developed tool is optimized in terms of underlying data structures to quickly load and integrate wealth of information for metagenomic contigs from the analyses tools such as PROKKA, CONCOCT, and TAXAassign. With the per-sample view, the user has the ability to pick out contigs that are up/down regulated between different conditions based on coverage information. Furthermore, the sequencing content of the contigs along with annotated functional profile

(proteins/enzymes/COGs) can also be explored further. The interactivity provided by the software will enable the user to selectively focus on samples and contigs that are interesting from the point-of-view of the study. If he finds some coding sequence regions important, then he also has the ability to extract and analyze the sequences further with third party tools.

7. Further Work

It is important to mention that the software developed during this project and described in this report was designed and implemented in a short time with specific goals in mind. There is still a room for improvement with new features to be integrated and functionalities to be adopted. These additional implementations could make the software more powerful and capable to produce further results for research purposes.

One of these features of further work that could be integrated in the viewer could be the ability to generate a phylogenetic tree for those contigs that can be fully resolved taxonomically. For now, the tool provides some basic taxonomical information without involving any details for the relation and closeness among the contained clusters in the dataset. This information could be obtained through the generation of an appropriate phylogenetic tree by utilising cluster-of-orthologus genes (COGs).

In addition, the software could be enriched with a variety of extra information which may include GC information and a set of statistical features obtained from protein sequences. In the future, we would like to update the tool to calculate the GC content for each contig and combine it with coverage information to identify GC bias. Moreover, we will also endeavour to incorporate statistics for CDS regions such as isoelectric point, aliphatic index, net charge, boman index, hydrophobicity index etc that are often used to identify antimicrobial resistant genes.

8. References

1. Christopher, Q., Nick, L., Umer, Z. I., Murat A. E., Delphine, S., Julie, R., et al. Extensive modulation of the fecal metagenome in children with Crohn's disease during exclusive enteral nutrition. *Submitted for publication*

2. Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Sogin, M. L., Delmont, T. O. (2015). Anvi'o: An advanced analysis and visualization platform for 'omics data. PeerJ PrePrints. **3**: e1566

3. Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. **11**: 119

4. James, L. W. and Eugene, W. M. (1997). Human Whole-Genome Shotgun Sequencing. Genome Res. 7: 401-409

5. Johannes, A., Brynjar, S. B., Bruijn, .I, Melanie, S., Joshua, Q., Umer, Z. I., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*. **11**: 1144-1146

6. Kansal, S., Wagner, J., Kirkwood, C. D., Catto-Smith, A. G. (2013). Enteral Nutrition in Crohn's Disease: An Underused Therapy. Gastroenterology Research and Practice. **2013**: 11 pages

7. Ong, S. H., Kukkillaya, V. U., Wilm, A., Lay, C., Ho, E. X. P., Low, L., et al. (2013). Species Identification and Profiling of Complex Microbial Communities Using Shotgun Illumina Sequencing of 16S rRNA Amplicon Sequences. PLoS ONE. **8**: e60811

8. Poretsky, R., Rodriguez, L. M., Luo, C., Tsementzi, D., Konstantinidis, K. T. (2014). Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. PLoS ONE. **9**: e93827

9. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics. **30**: 2068-2069

10. Steed, H., Macfarlane, G. T., Blackett, K. L., Bahrami, B., Reynolds, N., Walsh, S. V., et al. (2010). Clinical Trial: The Microbiological and Immunological Effects of Synbiotic Consumption – A Randomized Double-blind Placebo-controlled Study in Active Crohn's Disease. Aliment Pharmacol Ther. **32**: 872-883.

11. Shah, N., Tang, H., Doak, T.G., Ye, Y. (2011). Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. Pac Symp Biocomput. **2011**: 165–176.

12. Zhengwei, Z., Beifang, N., Jing, C., Sitao, W., Shulei, S., Weizhong Li. (2012). MGAviewer: A desktop visualisazation tool for analysis of metagenomics alignment data. Bioinformatics. **29**: 122-123