Middle East Technical University
Informatics Institute

METATRANSCRIPTOMICS ANALYSIS USING MICROBIOME RNA-SEQ DATA

Advisor Name: Aybar Can ACAR
(METU)

Student Name: Aysu KOSEOGLU
(Program Name – MI)

June 2024

TECHNICAL REPORT
METU/II-TR-2024-

Orta Doğu Teknik Üniversitesi
Enformatik Enstitüsü

MİKROBİYOM RNA-SEQ VERİSİ KULLANARAK METATRANSCRIPTOMİK ANALİZ

Danışman Adı: Aybar Can ACAR
(ODTÜ)

Öğrenci Adı: Aysu KÖSEOĞLU
(Program Adı- MI)

Haziran 2024

TEKNİK RAPOR
ODTÜ/II-TR-2024-

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY (Internal Use)	2. REPORT DATE 14.06.2024

3. TITLE AND SUBTITLE

METATRANSCRIPTOMICS ANALYSIS USING MICROBIOME RNA-SEQ DATA

4. AUTHOR (S) Aysu KOSEOGLU	5. REPORT NUMBER (Internal Use) **METU/II-TR-2024-**

6. SPONSORING/ MONITORING AGENCY NAME(S) AND SIGNATURE(S)

Informatics Online Master's without Thesis Program, Department of Medical Informatics, Informatics Institute, METU

Advisor: Aybar Can ACAR Signature:

7. SUPPLEMENTARY NOTES

8. ABSTRACT (MAXIMUM 200 WORDS)

This project aimed to understand Coprothermobacter proteolyticus and explore meta transcriptomics analysis using microbiome RNA sequencing (RNA-seq). Additionally, the project emphasized the importance of training data for meta transcriptomics analysis, facilitating comprehensive study of microbial communities and gene expression patterns. Overall, this project expands our knowledge of Coprothermobacter proteolyticus, meta transcriptomics analysis, and their practical applications in various fields.

9. SUBJECT TERMS Meta transcriptomics analysis, RNA-seq	10. NUMBER OF PAGES **20**

TABLE OF CONTENTS

REPORT DOCUMENTATION PAGE iii
METU/II-TR-2024- iii
LIST OF FIGURES v
INTRODUCTION 1
CHAPTER 1: DATA 3
CHAPTER 2: PRE - PROCESSING 4
CHAPTER 3: EXTRACTION OF THE COMMUNITY PROFILE 11
CONCLUSION 16

LIST OF FIGURES

Figure 1: Metabolic function and pathway reconstruction.....1

Figure 2: MultiQC: FastQC mean quality scores 4

Figure 3: MultiQC: FastQC per sequence quality scores 5

Figure 4: MultiQC: FastQC per base sequence content.....5

Figure 5: MultiQC: FastQC per base GC content.....6

Figure 6: MultiQC: FastQC per base N content 6

Figure 7: MultiQC: Sequence duplication levels 7

Figure 8: MultiQC: Overrepresented sequences.....7

Figure 9: MultiQC: Adapter content 8

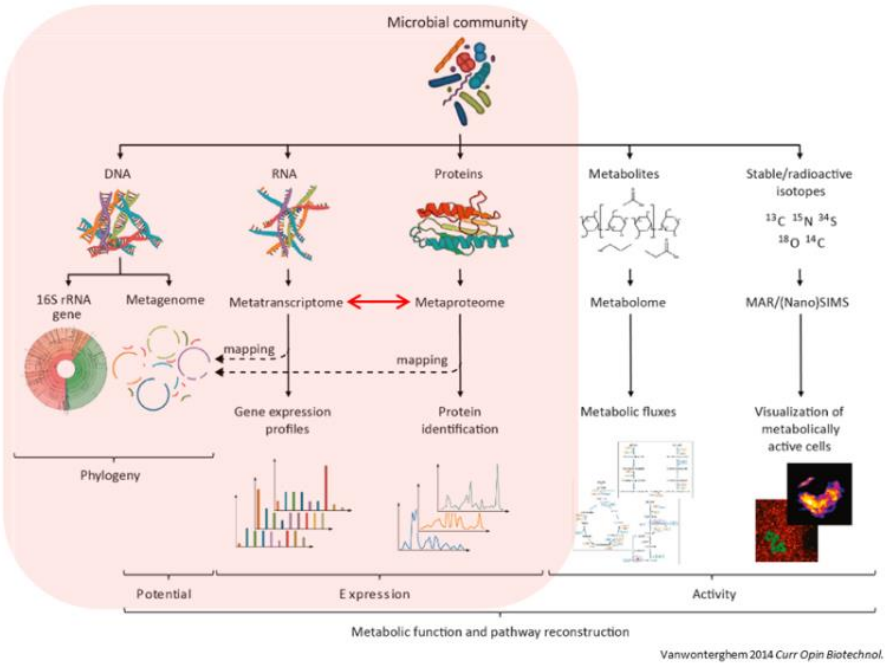
Figure 10: Cutadapt report output.....9

Figure 11: Krona pie chart 12

Figure 12: Result of combining MetaPhlAn2 and HUMAnN2 15

INTRODUCTION

Microbiomes play a crucial role in the well-being of hosts, the development of diseases, and the overall environment. The advancement of sequencing technologies has greatly facilitated the study of microbiota and microbial communities. These technologies allow us to explore the dynamics of the microbiome by examining various aspects such as DNA content (metagenomics), RNA expression (metatranscriptomics), protein expression (metaproteomics), and small molecules (metabolomics).



(Figure 1)

The continuous evolution of sequencing platforms and the availability of numerous bioinformatic tools have resulted in significant progress in metagenomics and metatranscriptomics, enabling the investigation of complex microbial communities. These techniques provide valuable insights into the taxonomic profiles and genomic components of microbial communities. While metagenomics offers information about the taxonomies present in a microbiome, it does not provide detailed insights into important functions. This is where metatranscriptomics and metaproteomics play a significant role.

This project focuses specifically on metatranscriptomics analysis. Metatranscriptomics allows us to understand how the microbiome

responds to the environment by studying the functional analysis of genes expressed by the microbial community. It also helps in estimating the taxonomic composition of the microbial population. Additionally, metatranscriptomics provides confirmation of predicted open-reading frames (ORFs) and potential identification of novel sites of transcription and/or translation within microbial genomes. Furthermore, it aids in the generation of more comprehensive protein sequence databases for metaproteomics analysis.

For the purpose of this tutorial, utilizing metatranscriptomics data from a time-series analysis of a microbial community within a bioreactor (Kunath et al., 2018). The study generated metatranscriptomics data from 3 replicates across 7 time points. Prior to amplification, rRNA was reduced using rRNA depletion techniques. The sequencing library was prepared using TruSeq stranded RNA sample preparation, which involved the creation of a cDNA library.

In this project, the focus will be on biological replicate A of the first time point. In a subsequent tutorial, the project demonstrates how to compare the results across different time points and replicates. The input files utilized here have been trimmed from the original files to save time and resources.

To analyze the data, the project employs the ASaiM workflow, which will be explained in a step-by-step manner. ASaiM (Batut et al., 2018) is an open-source Galaxy-based workflow designed for microbiome analyses. This workflow offers a streamlined and reproducible environment within the Galaxy platform for users to explore metagenomic and metatranscriptomic data. The ASaiM workflow has been updated by the GalaxyP team at the University of Minnesota to enable metatranscriptomics analysis of large microbial datasets (Mehta et al.,

CHAPTER 1: DATA

Microbiomes play a crucial role in maintaining host health, influencing disease outcomes, and impacting the environment. Analysing the functional aspects of microbial communities, beyond just their taxonomic composition, has become increasingly important. Meta transcriptomics, which involves studying RNA expression in microbial communities, allows researchers to gain insights into the functional groups expressed by the microbiome and their correlation with the studied conditions.

This tutorial, offered by the Galaxy Training Network, aims to introduce researchers to the fundamental concepts of meta transcriptomics data analysis. The tutorial utilizes paired-end datasets of raw shotgun sequences in FastQ format as input, and guides users through the following steps:

- Pre-processing the data
- Extracting and analysing the community structure to obtain taxonomic information
- Extracting and analysing the community functions to gain functional insights
- Integrating taxonomic and functional information to understand the contributions of specific taxa to expressed functions.

The dataset used in the tutorial is derived from a time-series analysis of a microbial community within a bioreactor (Kunath et al., ISME, 2018). For the purpose of saving time and resources, only the data from one time point (1st) and a single biological replicate (A) are analyzed in this tutorial. The provided dataset has been trimmed from the original file.

CHAPTER 2: PRE - PROCESSING

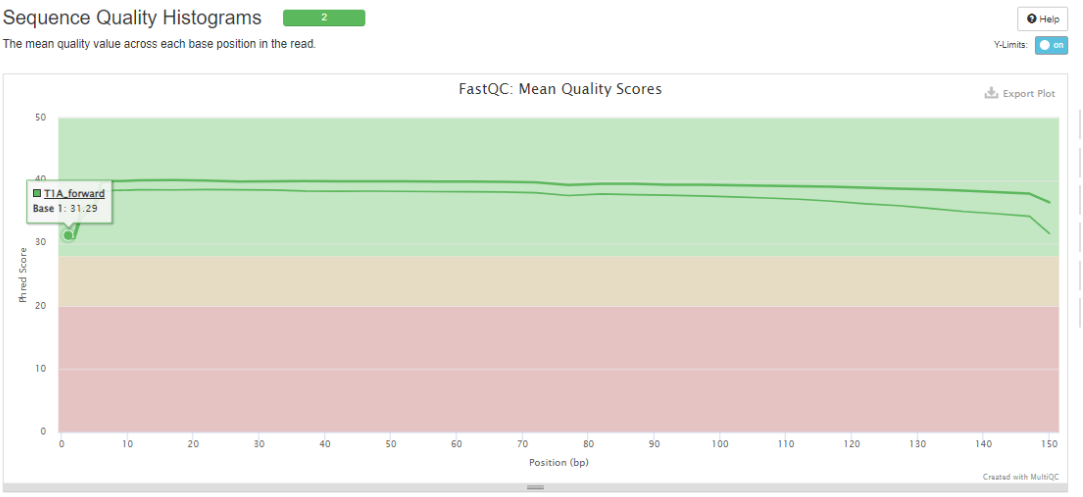
Quality Control:

FastQC

FastQC is a quality control tool used to assess the quality of raw sequencing data. It provides a comprehensive analysis of various quality metrics, including per-base sequence quality, per-base sequence content, sequence length distribution, and overrepresented sequences. The primary goal of using FastQC is to identify any potential issues or biases in the raw data that may affect downstream analysis. By examining the FastQC report, researchers can make informed decisions on whether any pre-processing steps, such as trimming or filtering, are necessary to improve the data quality.

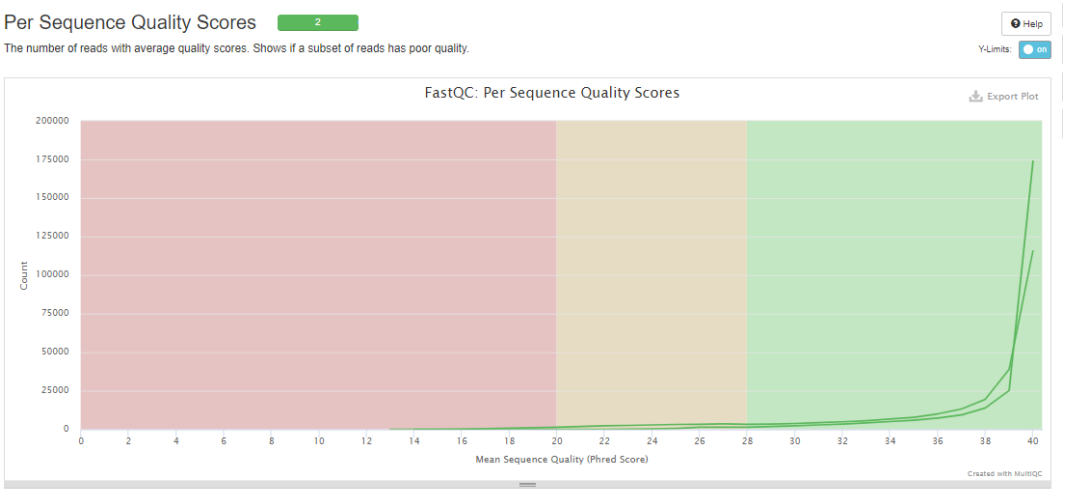
MultiQC

MultiQC is a useful tool for generating a comprehensive quality control report by aggregating results from multiple tools, including FastQC. It provides a summarized overview of the quality metrics obtained from various steps of the analysis pipeline. By utilizing MultiQC, researchers can easily compare and interpret the quality control results of multiple samples in a unified and standardized format. This allows for efficient identification of common issues across the dataset and aids in the decision-making process for further pre-processing or analysis steps.



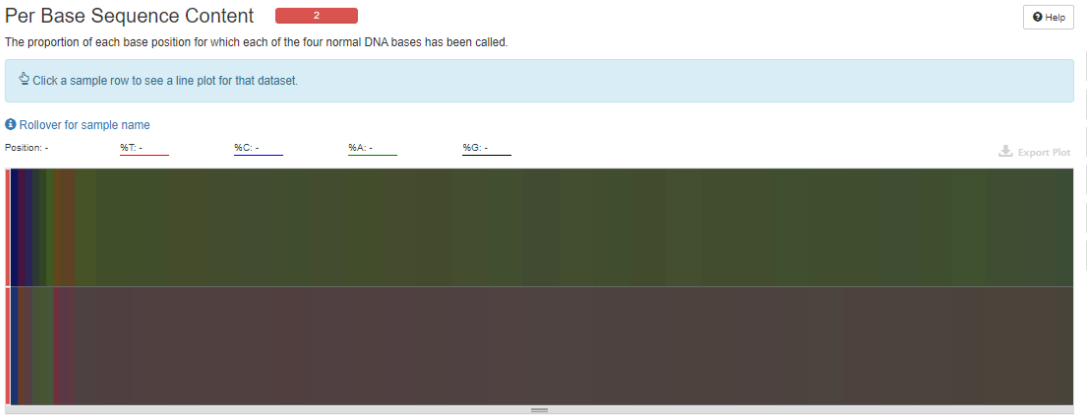
(Figure 2)

This graph represents the mean quality value is successful across each base position in the read of T1A_reverse T1A_ forward data.



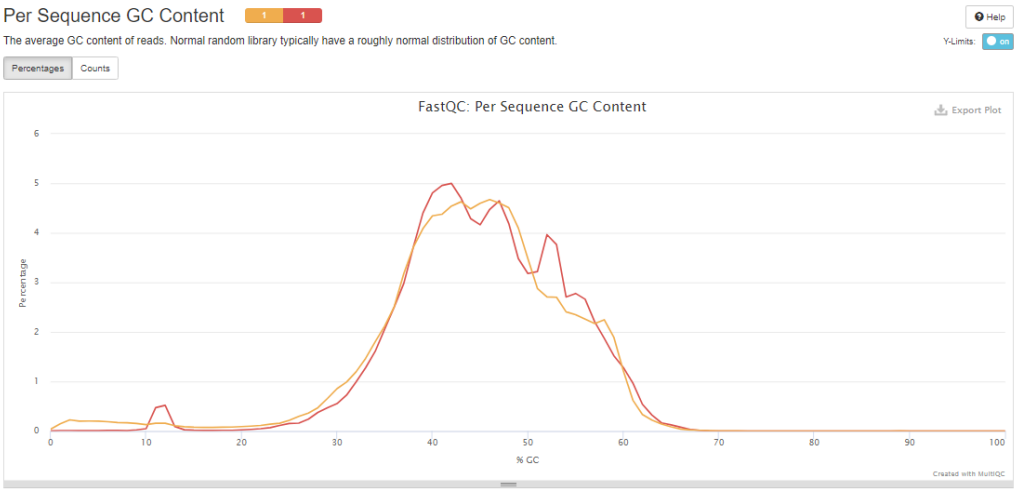
(Figure 3)

The average quality scores for each data are higher enough.



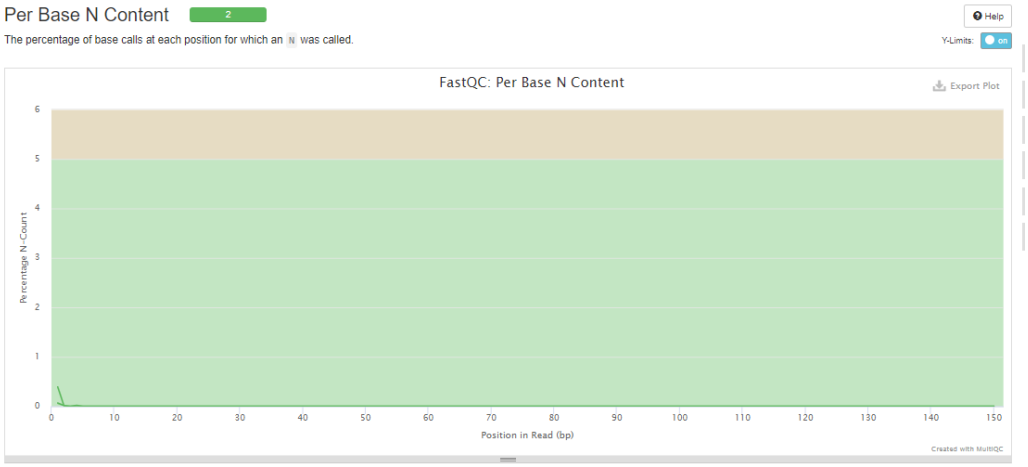
(Figure 4)

This results indicates that the observed nucleotide composition at certain positions within the reads does not meet the expected patterns.



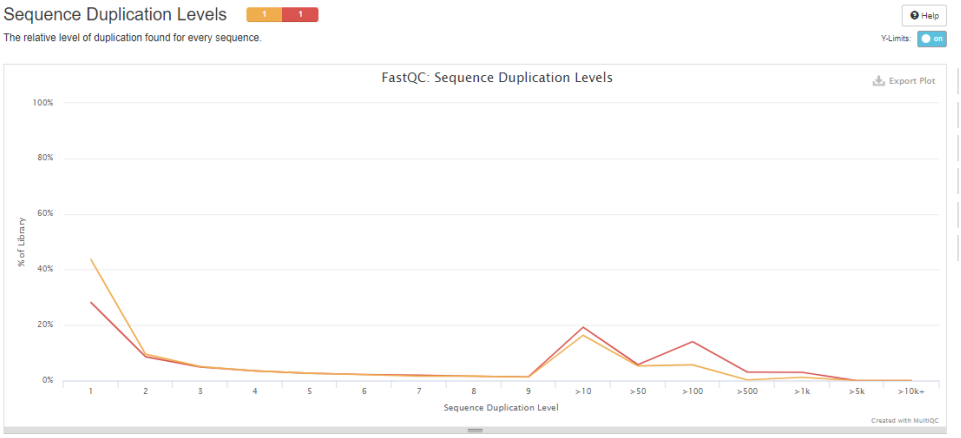
(Figure 5)

The result indicates that there might be potential irregularities in the sequencing process.



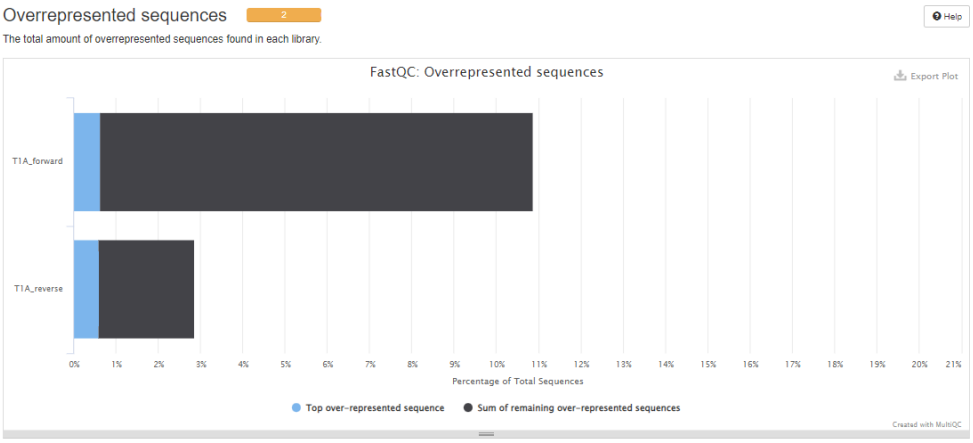
(Figure 6)

This indicates that the sequencing reads contain a low number of ambiguous bases are typically encountered when there are gaps or uncertainties in the sequence information.



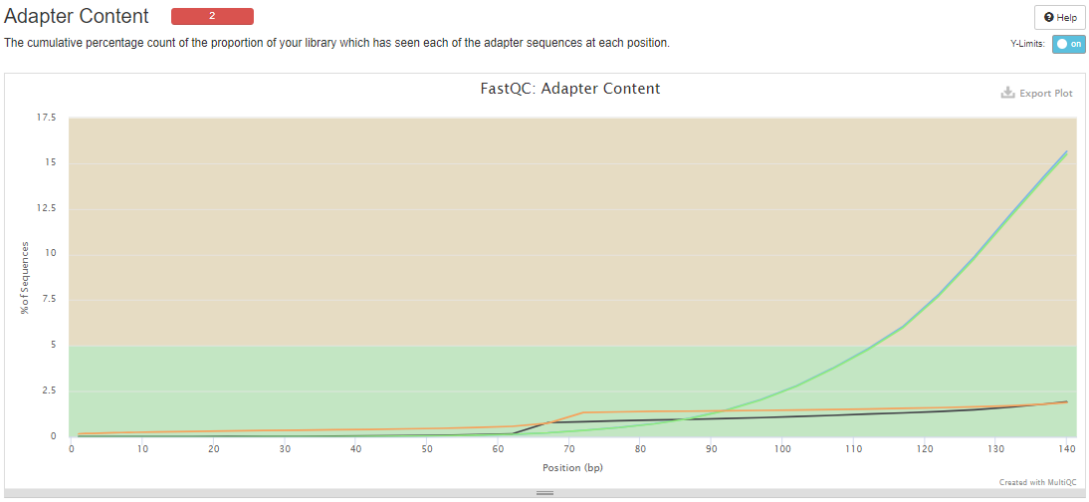
(Figure 7)

This indicates that the sequencing reads in the dataset have consistent.



(Figure 8)

This indicates the presence of sequences that are significantly overrepresented in the data compared to what would be expected by chance.



(Figure 9)

This graph represents the adapter content of the data. It can be understood that the data need trimming to increase their quality.

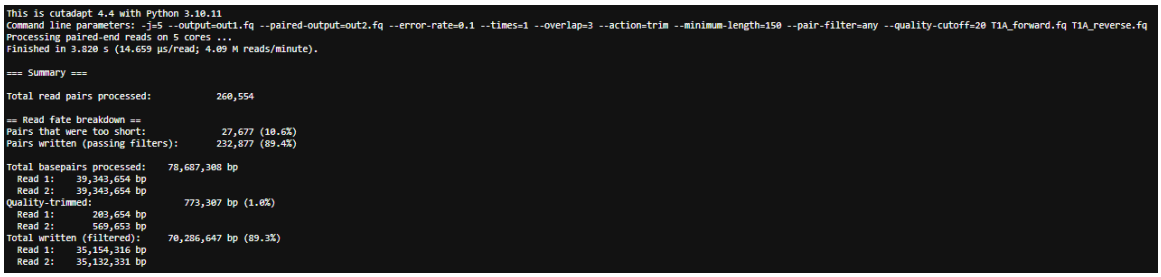
Trimming

Trimming is a pre-processing step that involves the removal of low-quality bases or sequences from the raw data. This process helps to improve the overall data quality and reliability of downstream analyses. Trimming can be performed using tools such as Trimmomatic or Cutadapt, which identify and remove adapter sequences, low-quality bases, and/or reads below a certain quality threshold. The purpose of trimming is to enhance the accuracy of subsequent analyses, reduce the presence of sequencing errors, and increase the efficiency of alignment or assembly processes.

By incorporating FastQC, MultiQC, and trimming in the pre-processing step, researchers can ensure that the raw sequencing data is of high quality, allowing for more accurate and reliable downstream analysis. FastQC provides a detailed assessment of quality metrics, MultiQC streamlines the visualization and comparison of quality control results, and trimming removes low-quality sequences, resulting in improved data quality for subsequent analysis steps.

In Cutadapt part of this project, in filter options part, the minimum length parameter is assigned 150, in read modification options, the quality

cutoff value assigned as 20. One of three output of Cutadapt which is Cutadapt on data 2 and data 1 report represents trimming and filtering information.



(Figure 10)

Ribosomal RNA Fragments Filtering

In meta transcriptomics data analysis, it is common to have a significant portion of reads originating from rRNA fragments. These rRNA fragments do not provide relevant information about the functional gene expression of the microbial community and can interfere with downstream analyses. Therefore, a pre-processing step involves filtering out these rRNA fragments from the dataset.

By employing tools such as “SortMeRNA” or Bowtie, the raw meta transcriptomics data can be aligned against a reference rRNA database. The aligned reads that correspond to rRNA fragments are then removed, resulting in a dataset that is depleted of rRNA sequences. This filtering step helps to enhance the accuracy and specificity of downstream analyses by focusing on the functional gene expression of interest within the microbial community.

SortMeRNA has five outputs which are aligned forward reads, aligned reverse reads, unaligned forward reads, unaligned reverse reads, and log file. The number of sequences in the aligned forward read file is 1947 and the number of unaligned sequences in the unaligned file is 2858. The number of identified reads as rRNA is 1947, and 2858 reads not identified as non RNA.

Interlace Forward and Reverse Reads

Meta transcriptomics data is often generated using paired-end sequencing, where forward and reverse reads correspond to different ends of the same transcript. To utilize these paired end reads effectively, a pre-processing step involves interlacing the forward and reverse reads into a single file.

By using tool FastQ interlacer the forward and reverse reads can be combined into a single file, maintaining the pairing information. This step allows for the subsequent analysis to consider both ends of the transcripts and maximize the information extracted from the paired-end sequencing.

In summary, the pre-processing steps in this project include filtering out rRNA fragments to focus on functional gene expression and interlacing forward and reverse reads to fully utilize paired-end sequencing data. These steps contribute to improving the quality and specificity of the meta transcriptomics analysis by removing non-functional sequences and maximizing the utilization of paired-end information.

CHAPTER 3: EXTRACTION OF THE COMMUNITY PROFILE

The aim of the "Extraction of the Community Profile" step is to obtain information about the taxonomic composition and abundance of the microbial community within a meta transcriptomics dataset. This step involves analysing the sequencing data to identify and classify the different microbial taxa present in the sample.

By extracting the community profile, researchers can gain insights into the taxonomic structure and diversity of the microbial population. This information is valuable for understanding the composition of the microbiome, identifying dominant or rare taxa, and investigating the potential functional capabilities of the community.

Community Structure Visualization

Metaphlan

Metaphlan is a widely used tool for taxonomic profiling of metagenomic and meta transcriptomic data. It utilizes a unique marker-based approach to assign taxonomic labels to the sequencing reads. By comparing the reads against a comprehensive reference database, Metaphlan provides information about the taxonomic composition and abundance of different microbial taxa present in the dataset. This taxonomic profiling serves as a basis for understanding the community structure of the microbial population in the meta transcriptomics dataset.

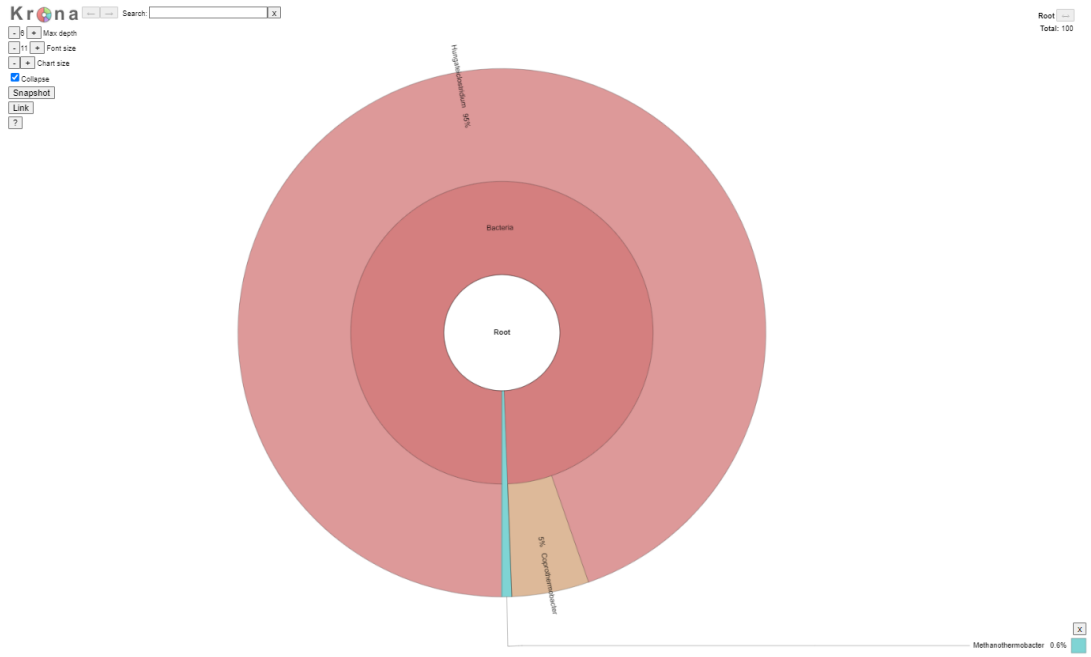
Metaphlan has 5 outputs in Galaxy which are BIOM file, SAM file, Bowtie2 output, predicted taxon relative abundances, and predicted taxon relative abundances at each taxonomic levels. Predicted taxon level abundances output includes the lineage with different taxonomic levels, the previous lineage with NCBI taxon id, the relative abundance found for our sample for the lineage, and any additional species. The high level taxa kingdom (_k) is set on the beginning of the file.

Krona Pie Chart

Krona is a visualization tool that allows for the creation of interactive and informative pie charts to represent the taxonomic composition of the community structure. The Krona pie chart provides a hierarchical

representation, where each level represents a taxonomic rank (e.g., phylum, class, genus). The size of each pie slice corresponds to the abundance or relative proportion of the taxonomic group it represents. By interactively exploring the Krona pie chart, you can gain insights into the distribution and dominance of different taxonomic groups within the microbial community. This visualization method simplifies the interpretation and communication of the community structure findings.

By utilizing Metaphlan for taxonomic profiling and generating Krona pie charts, you were able to visualize and understand the community structure of the microbial population within your meta transcriptomics dataset. This approach provided insights into the taxonomic composition and relative abundance of different microbial taxa. The interactive nature of the Krona pie chart facilitated the exploration and interpretation of the community structure, enabling you to identify key taxonomic groups of interest for further analysis or investigation.



(Figure 11)

This pie chart exhibits the percentages of the bacterias Hungateiclostridium thermocellum (%95) and the Coprothermobacter proteolyticus (%5).

Extract the Functional Information

HUMAnN (HMP Unified Metabolic Analysis Network)

HUMAnN is a widely used tool for functional profiling of metagenomic and meta transcriptomic data. It leverages a comprehensive reference database to assign functional annotations to the sequencing reads. By aligning the reads against this database, HUMAnN provides information about the functional pathways and genes expressed by the microbial community. It allows for the quantification of functional features, such as pathway abundances and gene expression levels.

Renormalize a HUMAnN-generated table

Renormalization is a process that adjusts the abundance values in a HUMAnN-generated table to account for differences in sample read depths or library sizes. This step ensures that the functional data remains comparable across samples by normalizing the abundances to a common scale. Renormalization helps to account for potential biases and variations in sequencing depth between samples, allowing for more accurate and meaningful comparisons of functional profiles.

Unpack Pathway Abundances to Show Genes Included

Pathway abundances obtained from HUMAnN represent the overall functional potential of the microbial community. Unpacking pathway abundances involves further exploration and identification of the specific genes that contribute to each pathway. By dissecting the pathway abundances, you can gain insights into the specific genes or gene families that are associated with particular functional pathways. This helps to elucidate the molecular mechanisms and processes involved in the community's functional activities.

Regroup HUMAnN Table Features and Select Lines That Match an Expression

To streamline the analysis and focus on specific functional features of interest, regroupe the HUMAnN table features based on specific criteria or metadata. This regrouping allows for the aggregation or grouping of functional information based on shared characteristics, such as specific

genes, pathways, or functional categories. Additionally, selected lines or data points from the HUMAnN table that matched a particular expression or threshold. This step enables the identification of functional features that meet specific criteria or exhibit differential expression patterns.

Through these processes, the "Extract the Functional Information" step aims to uncover the functional potential and characteristics of the microbial community within your meta transcriptomics dataset. It involves functional profiling using HUMAnN, renormalizing data for comparability, exploring gene contributions to pathways, regrouping features for analysis, and selecting specific lines or data points based on expression criteria. These steps help to reveal the functional landscape of the microbial community and provide insights into the molecular activities and metabolic potential of the organisms present.

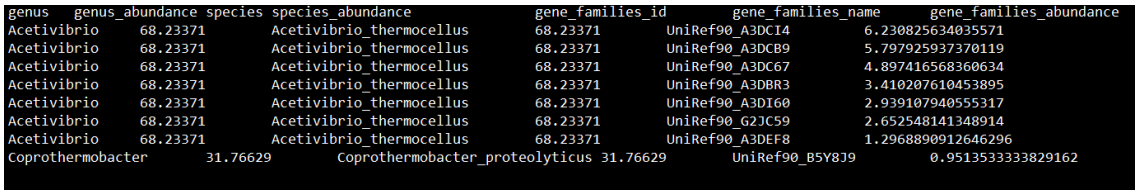
Combine MetaPhlAn2 and HUMAnN2

MetaPhlAn2 is a tool used for taxonomic profiling that provides valuable information about the microbial composition of the community. It utilizes a unique set of marker genes to accurately identify and quantify the abundance of different microbial taxa present in the sample. By running MetaPhlAn2 on our dataset, we obtained taxonomic profiles, revealing the relative abundance of specific microbial groups and their potential roles within the community.

On the other hand, HUMAnN2 focuses on functional profiling by analyzing the gene families, pathways, and functional annotations within the meta transcriptomics data. It offers insights into the functional potential and activities of the microbial community, shedding light on the underlying molecular processes and metabolic functions.

To obtain a more comprehensive view, we integrated the outputs of MetaPhlAn2 and HUMAnN2. By combining the taxonomic and functional information, we gained a deeper understanding of the interplay between microbial composition and functional capabilities. This integration allowed us to explore the functional contributions of specific taxonomic groups and understand how they shape the overall metabolic landscape of the community.

The merged outputs provided a holistic perspective, linking taxonomic composition with functional profiles. This comprehensive analysis was crucial in unravelling the intricate relationships between the microbial community's structure, taxonomic diversity, and its potential functional activities. Ultimately, this step enhanced our understanding of the meta transcriptomics dataset, revealing valuable insights into the complex dynamics and functional potential of the microbial community under study (Franzosa et al., 2018).



genus	genus_abundance	species	species_abundance	gene_families_id	gene_families_name	gene_families_abundance
Acetivibrio	68.23371	Acetivibrio_thermocellus	68.23371	UniRef90_A3DCI4		6.230825634035571
Acetivibrio	68.23371	Acetivibrio_thermocellus	68.23371	UniRef90_A3DCB9		5.797925937370119
Acetivibrio	68.23371	Acetivibrio_thermocellus	68.23371	UniRef90_A3DC67		4.897416568360634
Acetivibrio	68.23371	Acetivibrio_thermocellus	68.23371	UniRef90_A3DBR3		3.410207610453895
Acetivibrio	68.23371	Acetivibrio_thermocellus	68.23371	UniRef90_A3DI60		2.939107940555317
Acetivibrio	68.23371	Acetivibrio_thermocellus	68.23371	UniRef90_G2JC59		2.652548141348914
Acetivibrio	68.23371	Acetivibrio_thermocellus	68.23371	UniRef90_A3DEF8		1.2968890912646296
Coprothermobacter	31.76629	Coprothermobacter_proteolyticus	31.76629	UniRef90_B5Y8J9		0.9513533333829162

(Figure 12)

CONCLUSION

In conclusion, the research conducted in this project has significantly contributed to the understanding of Coprothermobacter proteolyticus and the analysis of meta transcriptomics data using microbiome RNA sequencing (RNA-seq). The study by Kunath et al. (2018) shed light on the lifestyle and genetic evolution of Coprothermobacter proteolyticus, elucidating the intricate process of converting proteins to polysaccharides. This research has important implications in various fields, including biotechnology, environmental science, and microbial ecology.

The findings presented in the study demonstrate the adaptability of Coprothermobacter proteolyticus, highlighting its ability to utilize diverse protein sources for growth and energy production. The investigation of its genetic evolution provides valuable insights into the mechanisms behind its proteolytic activities and offers potential avenues for further exploration and exploitation of its enzymatic capabilities.

Additionally, the project report reviewed the work of Batut et al. (2019), which focused on the development and availability of training data for meta transcriptomics analysis using microbiome RNA-seq data. This contribution is crucial in advancing the field of meta transcriptomics and facilitating accurate and comprehensive analysis of complex microbial communities. The availability of such training data sets opens up opportunities for researchers to improve their understanding of the functional dynamics and gene expression patterns within microbial ecosystems.

Overall, this project has underscored the significance of studying microbial communities and their functional potential. By investigating the lifestyle and genetic evolution of Coprothermobacter proteolyticus, as well as exploring the utilization of meta transcriptomics data for microbiome analysis, the project has expanded our knowledge in these areas. These findings pave the way for future research and practical applications, such as developing novel biotechnological approaches, optimizing industrial processes, and enhancing our understanding of environmental microbiology.

It is important to note that while this project has made significant contributions, there are still many avenues for further research and exploration. Future studies could delve deeper into the specific mechanisms employed by Coprothermobacter proteolyticus in protein degradation and polysaccharide synthesis. Furthermore, the development of more comprehensive training data sets for meta transcriptomics analysis will continue to refine our understanding of microbial gene expression and functional dynamics within complex ecosystems.

In conclusion, the research presented in this project report has provided valuable insights into Coprothermobacter proteolyticus and the analysis of meta transcriptomics data using microbiome RNA sequencing. These findings contribute to our knowledge of microbial biology and have practical implications in various fields. The project opens up new possibilities for future research and applications, driving advancements in biotechnology, environmental science, and microbial ecology.

REFERENCES

Batut, B., Mehta, S., Kumar, P., Jagtap, P., & Hiltemann, S. (2019, August 8). Training data for "metatranscriptomics analysis using microbiome rnaseq data." Zenodo. https://doi.org/10.5281/zenodo.4776250

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., ... & Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods, 15(11), 962-968.

Kunath, B. J., Delogu, F., Naas, A. E., Arntzen, M. Ø., Eijsink, V. G. H., et al. (2018). From proteins to polysaccharides: lifestyle and genetic evolution of Coprothermobacter proteolyticus. The ISME Journal, 1. https://doi.org/10.1038/s41396-018-0290-y

Jagtap, P., Mehta, S., Sajulga, R., Batut, B., Leith, E., Kumar, P., Hiltemann, S., & Zierep, P. (2023, May 17). Galaxy training: Metatranscriptomics analysis using microbiome RNA-Seq Data. Galaxy Training Network. https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/metatranscriptomics/tutorial.html#overview