**Experimental Report: Comparison between different strategy in human tissue AMR identification with Metatranscriptomics**

# Introduction

Human microbiota contains various composition across body habitats. It is a dynamic and diverse ecosystem that closely related to human health. Perturbations of human microbiota has been indicated associated with numerous disease and disorder. Localized colonization of various pathogens such as Fusobacterium are also suggested to have casual effects on diseases like colorectal cancer.

Large-scale microbial community profiling techniques have broadened our insights on human microbiota. Two of the most commonly used tools are meta-taxonomy and metagenomics. Meta-taxonomy sequences marker genes, in order to identify microbial species. With the method, bacteria and archaea could be recognized based on the 16S subunit of small ribosomal RNA, while nuclear ribosomal internal transcribed spacer (ITS), 18S rRNA or 26S rRNA regions can be used for fungi detection. Regions on the subunits can be highly conserved common to majority of species, therefore allows sequencing with universal primers to taxonomic classified microbes present in a community. However, meta-taxonomy has limited accuracy in taxonomic resolution. To overcome some of the limitations, metagenomics could be used which allows the study microbial community at strain level. The method extracts whole genomes from bacterial cells in the community and could accurately locating a genome in phylogenetic tree thus identifying novel species or predicting functional potential of the members.

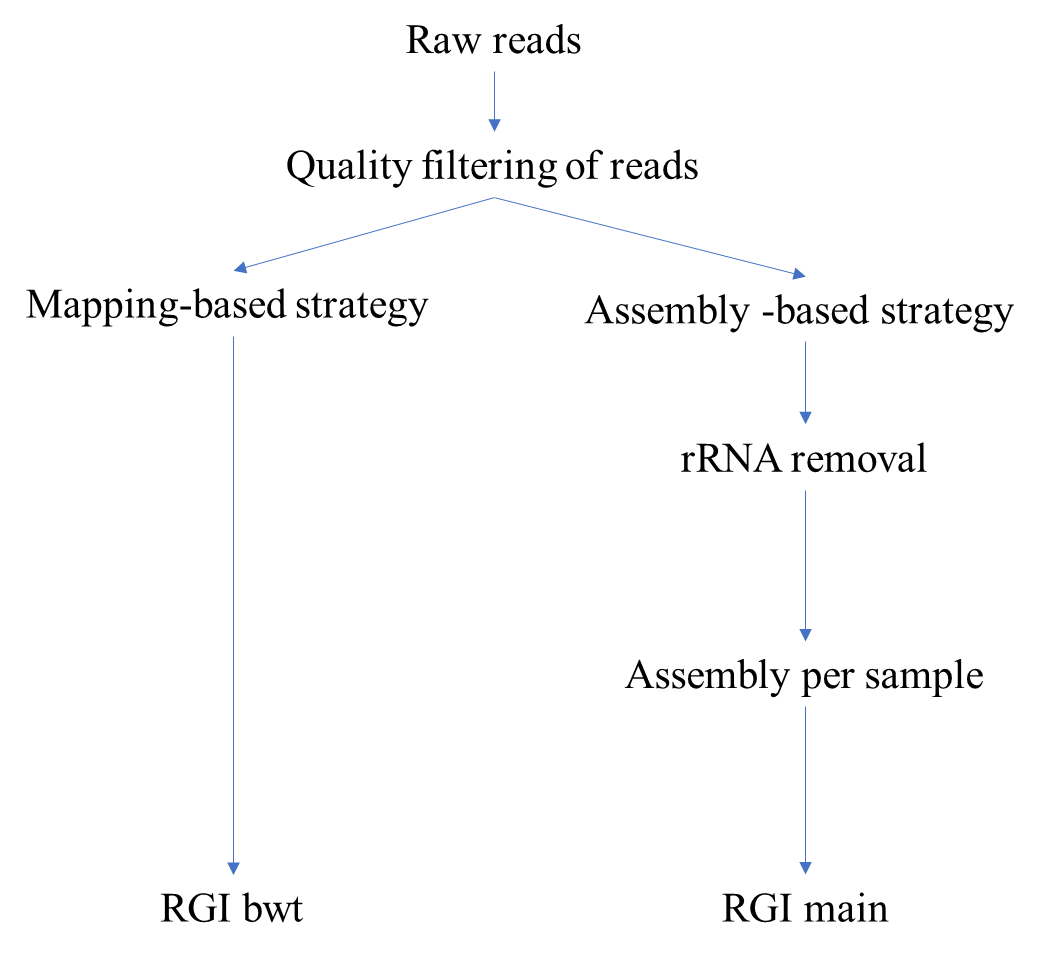
Although the mentioned technologies provide information on the presence of microorganisms and genes in the community, they do not give a closer look at active microbial members. In contrast, metatranscriptomics, which refers to the collective transcriptomes of a given habitats, enables to explore cell viability and transcriptional activity under different conditions, clinical parameters and overtime, therefore better represent host-microbe interactions. Research also indicates that the functional activity may vary widely across host (Franzosa et al., 2014). However, technical and sampling-related challenges remains to be addressed. Due to experimental and analytical limitations, messenger RNA (mRNA) usually has short life-time, contaminating DNA may easily introduced during sampling and samples could contains high abundance of rRNA (usually accounts for over 95% with transfer RNA). Also, the absence of poly(A)-tails in prokaryotes makes selection of microbial transcripts especially hard (Pereira-Marques et al., 2024).

The analysis methods can also be controversial during analysis of metatranscriptomics data. Generally, two types of method: mapping-based analyses and assembly-based analyses could be used in metatranscriptomics. Mapping-based analyses is more commonly in microbial studies, it maps high quality reads to reference genomes to detect potential genes and gene expression. Compared to assembly-based methods, the type of method is far less computationally intensive but well-suited for investigation of low-abundance transcripts. Also, mapping relies heavily on the availability and accuracy of reference sequences and is challenging in communities with novel organisms, close related strains, ambiguous orthologous or paralogous gene families. Assembly-based analyses assemble the reads into contigs, and map the contigs to reference genomes. These analyses can be challenging and requires higher quality sequencing data. The variable coverage of transcripts and the presence of closely related organisms may lead to chimeric contigs. Additionally, most assemblers are not specially designed for metatranscriptome and these methods are not customized to the task and quantitatively evaluated. However, it has the advantages of not requiring references, remove redundancy among read data, and allow the identification of novel expressed protein-coding genes. These methods are also more informative in subsequent analysis.

Antimicrobial resistant genes (AMR genes) are ubiquitous in natural niches. The collection of all AMR genes in microorganisms are regard as resistome, which has attracted more global attention due to their risk to public health. The rapid emergence of strains with new AMR genes or multi-drug resistance mechanisms indicate there is possibly no drugs can be used in treatment. Therefore, studying of resistome in a host-associated environment could be clinical important when transferred to pathogens. In localized tumor, existence of genes may guide further usage of antibiotics and improve survival. RGI is a software that able to predict antibiotic resistomes from both contigs or short DNA reads. With the tool, predicted AMR genes from both mapping-based or assembly-based methods could be compared and evaluated, therefore, provide more accurate information of the existence of the AMR genes, and reduce false-positive result.

# Methods and Materials

In this study, RNA sequencing data from the tissues of a total of 162 patients with CRC were collected from research (Joanito et al., 2022). Reads unmapped to the human genome (microbial reads) were further processed in FastQC (version 0.11.9) (Lo & Chain, 2014), Trimmomatic (version 0.33) (Bolger et al., 2014), SortMeRNA (Kopylova et al., 2012), to do quality check, quality filtering and rRNA sequences removal, respectively. The reads were trimmed based on a sliding window trimming approach. In the method, reads will be cut when the average base Phred quality within a 4-base sliding window is less than 15. The reads were discarded when the length was less than 100 and more than half of the read length was trimmed. Further, SortMeRNA was used to remove rRNA reads. After trimming, a quality check was applied with FastQC. Megahit was used to assemble the remained reads into longer contigs. In the research, the AMR genes were explored with the software RGI (6.0.3) (Alcock et al., 2023). RGI is a software able to recognize antibiotic resistance genes from sequence data: both short reads and contigs. In mapping-based method, high-quality reads after trimming were used as input of RGI, which was able to align the short DNA sequences using KWA against a comprehensive antibiotics resistance database (CARD) database. For assembly-based method, assembled contigs were used to predict open reading frames (ORFs) using prodigal and protein sequences were used to predict AMR genes with rgi\_main function. Only results with a cutoff of strict or perfect and a ‘Best\_Identities’ over 50 were remained. Therefore, results could be compared and analyzed.



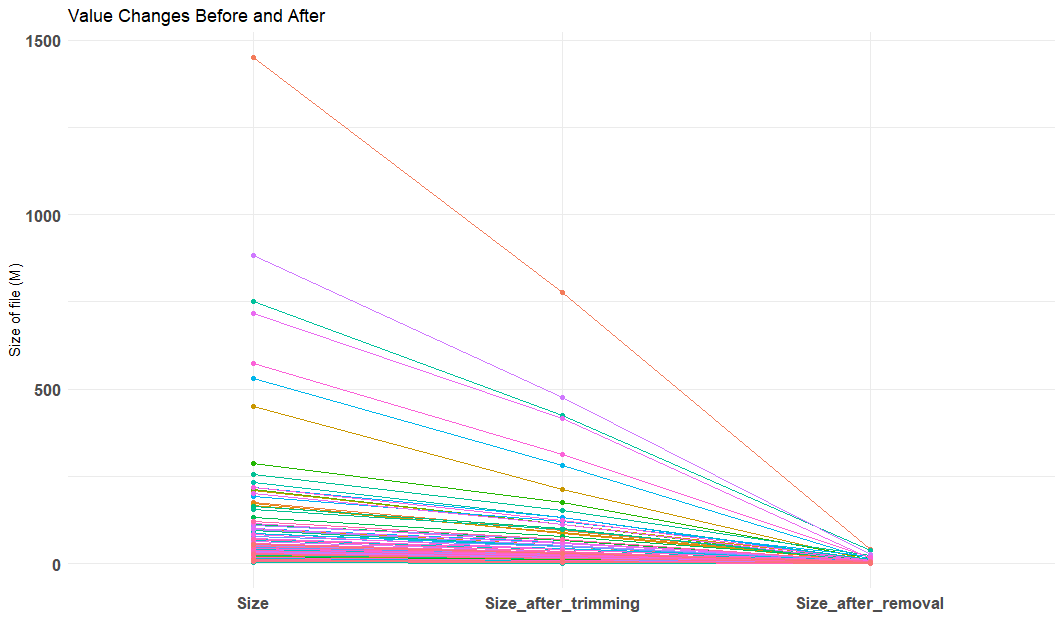
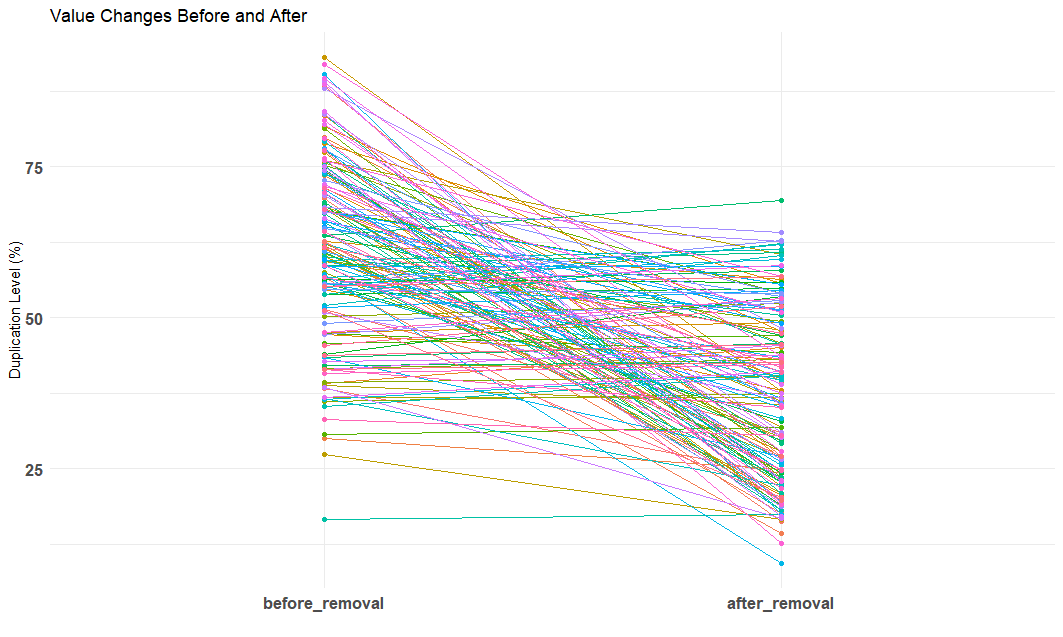
*Figure 1. Scheme of workflow*

# Result and discussion

## Quality of reads before and after the removal of rRNA

Trimming of low-quality sequences and rRNA removal were two common strategies in metatranscriptomics, however, very few research was done about the read quality changes during the process. Therefore, in the section, we will further discuss the condition of sequences after these two steps.

After the removal of low-quality sequences and rRNA, the After quality evaluation with MultiQC after trimming of low-quality reads, all samples were suggested pass the quality thresholds for ‘per base sequence quality’, ‘Per Tile sequence Quality’ and ‘Per Sequence Quality Scores’. However, about 30% of samples were rejected by means of ‘per base sequence content’, 21% of samples did not pass ‘sequence duplication levels’, 77% of samples got rejection in ‘sequence duplication level’. This result is very similar to research of (Gerard & María, 2022), in which all of the parameters also failed passing the threshold. After removal of rRNA, the ‘Sequence Duplication Levels’ of sample seems to be improved a lot, where only 39 samples were rejected. However, some samples were observed with even higher duplication level. This could happen due to that the deletion of rRNA sequence lower the complexity and diversity of overall sample, therefore highly duplicated mRNA become more apparent. Also, due to the lack of sequences and inadequate coverage, the rest of sequence may concentrate in several highly expressed transcripts, thus increase duplication level.

*Figure 2. (A)Change of file size after trimming and rRNA removal (B) Change of duplication level before and after rRNA removal*

In contrast, the ‘Per Base Sequence Content’ and ‘Per Sequence GC content’ were observed with worse performance, about 57 % and 81% of samples were rejected. Notably, the distribution of GC content per sequence (40-48%) were more unified after removal of rRNA (Hu et al., 2022). Before rRNA removal, the GC content was multimodal distribution with two peaks around 40%-48% and 50%-60% respectively. This phenomenon however reasonable. As in bacteria, the GC content of rRNA mostly distributed around 0.5-0.7, while for whole genome, GC content could be varied from 0.3 to 0.7 according to species (Hu et al., 2022). As most of sequences with GC content over 50% were removed, it also indicated the rRNA removal process was successful.

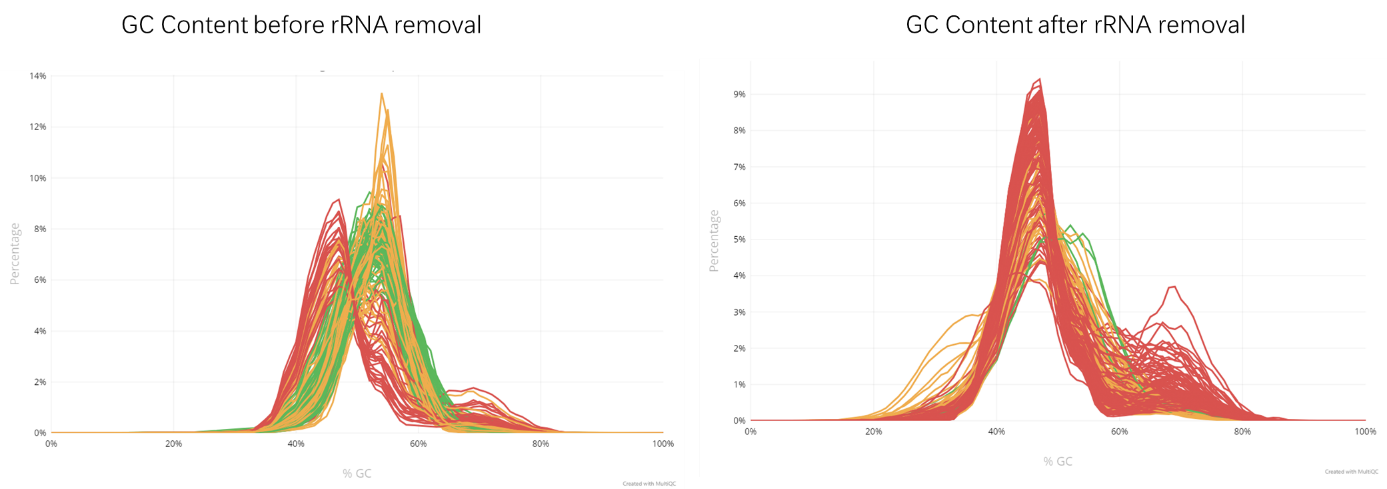


Figure3. (A) GC content distribution of samples before rRNA removal and (B) after rRNA removal.

To investigate changes of function of duplicated level, we performed BLAST against Refseq RNA database on top 20 overrepresented sequences. The results were illustrated in Table 1. It can be seen that most of the sequence after trimming were rRNA. This result can be reasonable considering the high fraction of rRNA in microbial genome (80-90%). After removal, duplicated sequence over 50bp were almost gone. These overrepresented sequences were mostly included in bacterial genomes, which indicated that an rRNA removal process could be necessary.

*Table 1. Top 20 Overrepresented sequence*

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | Phocaeicola faecicola strain AGMB03916 16S ribosomal RNA | Fretibacterium fastidiosum draft genome |
| 2 | Bacteroides fragilis strain ATCC 25285 23S ribosomal RNA gene | Klebsiella variicola strain FDAARGOS\_628 chromosome |
| 3 | Bacteroides fragilis strain ATCC 25285 23S ribosomal RNA gene | Citrobacter freundii strain FAHZZU8106 chromosome |
| 4 | Hallella mizrahii strain LKV-178-WT-2A 16S ribosomal RNA | Klebsiella pneumoniae strain KP12 chromosome, complete genome |
| 5 | Bacteroides helcogenes strain P 36-108 23S ribosomal RNA gene | Unknown |
| 6 | Phocaeicola faecicola strain AGMB03916 16S ribosomal RNA | Unknown |
| 7 | Bacteroides fragilis strain ATCC 25285 23S ribosomal RNA gene | Unknown |
| 8 | Salmonella enterica subsp. enterica serovar Typhimurium strain LT2 23S ribosomal RNA | Unknown |
| 9 | Bacteroides fragilis NCTC 9343 16S ribosomal RNA | Unknown |
| 10 | Bacteroides fragilis strain ATCC 25285 23S ribosomal RNA gene | Unknown |
| 11 | Phocaeicola faecicola strain AGMB03916 16S ribosomal RNA | Unknown |
| 12 | Salinivibrio socompensis S35 16S ribosomal RNA | Unknown |
| 13 | Salmonella enterica subsp. enterica serovar Typhimurium strain LT2 23S ribosomal RNA | Unknown |
| 14 | Bacteroides fragilis strain ATCC 25285 23S ribosomal RNA gene | Homo sapiens IGH c274\_heavy\_\_IGHV1-69\_IGHD1-20\_IGHJ5 mRNA |
| 15 | Yersinia similis strain Y228 23S ribosomal RNA | Homo sapiens IGH c369\_heavy\_\_IGHV4-39\_IGHD1-7\_IGHJ4 mRNA |
| 16 | Citrobacter cronae strain Tue2\_1 16S ribosomal RNA | Homo sapiens isolate 317.4C.L076.06\_W2 immunoglobulin heavy chain variable region mRNA |
| 17 | Bacteroides fragilis strain ATCC 25285 23S ribosomal RNA gene | Unknown |
| 18 | Bacteroides fragilis strain ATCC 25285 23S ribosomal RNA gene | Homo sapiens partial immunoglobulin heavy chain VDJ rearrangement |
| 19 | Bacteroides fragilis strain ATCC 25285 23S ribosomal RNA gene | Homo sapiens clone 2b2-015 immunoglobulin heavy chain variable region mRNA |
| 20 | Salmonella enterica subsp. enterica serovar Typhimurium strain LT2 23S ribosomal RNA | Shewanella chilikensis strain DC57 chromosome |

# Existence of certain human gene fraction in samples may indicated insufficient decontamination procedure. This also indicate after deletion of rRNA, the human genes became more apparent. Therefore, mRNA enrichment kits were suggested here in order to at least increase mRNA content.

## Assembly or Mapping

As mentioned in the method part, we used two types of strategies in AMR genes prediction with RGI: mapping-based method (mapping of short reads directly), assembly-based methods (contigs assembled per sample). Both strategies have their advantages and disadvantages.

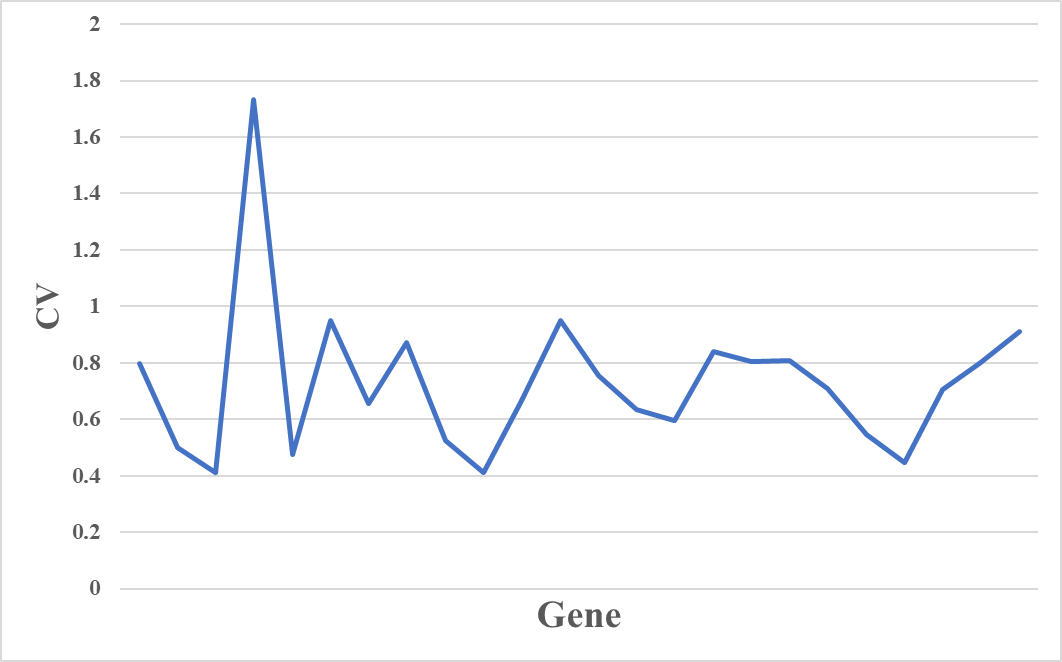
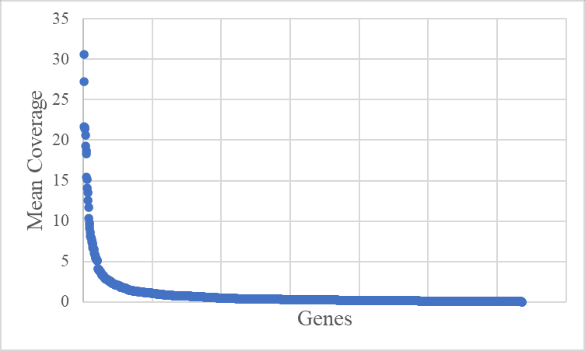
* Mapping-based method

Advantages using mapping-based method:

1. Read mapping approaches were deem superior to assembly methods for AMR gene detection in the past decades. It allows the identification of low abundance organisms in a community.
2. It could avoid problem of chimeric genes happened during assembly.

Risk in this study:

Sequencing depth could significantly impact on AMR profiles. In a research using metagenomics, the number of AMR gene families observed stabilized at a sequence depth of ~80 million reads per sample, the richness of AMR allelic did not appear to have plateaued even at a sequencing depth of 200 million reads per sample (Gweon et al., 2019) . This number, is much higher than our samples, in which the highest reads number were around 3 million. This brings huge limitation in our detection procedure, as their types could be underestimated. This inadequate read number also leads to the result that most of recognized AMR genes have an average depth per nucleotide lower than 5 (Figure 4A). To further check the quality of genes, the coverage uniformity within AMR genes with identity over 90% was evaluated using coefficient of variation (CV, standard deviation nucleotide depth/mean nucleotide depth). The distribution of CV was indicated in Figure4B. As it indicated, most genes have very high CV value, even over 1.5. Notably, this number usually supposed to be lower than 0.5. This could be technical bias brought by rRNA removal process, or problems happened during sequencing process (coverage is too low to bring useful expression information of microbial genes).



*Figure4. (A) Dustribution of mean coverage depth of nucleotides in mapped genes (B) Distribution of CV values in mapped genes with over 90% identity.*

* Assembly-based method

Advantages:

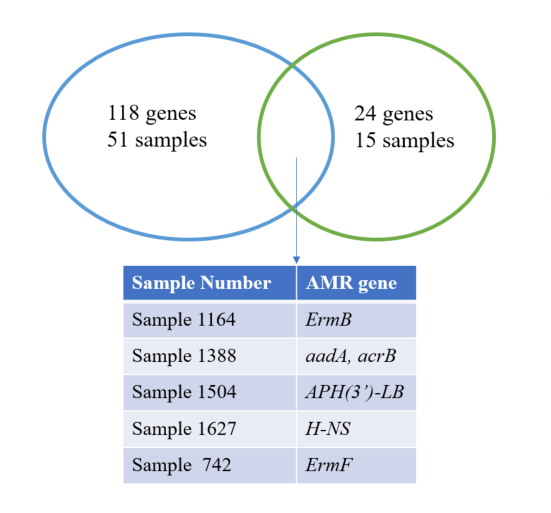
1. Able to include the genetic context around resistance genes, therefore provide further information on the interplay with host.
2. Relative complete genes to provide better functional information.

Risks:

Chimeric contigs could be a common problem during assembly caused by the variable coverage and presence of closely related organisms, therefore reduce the accuracy in following analysis. However, there are also papers indicating assembled contigs could also capture the majority of high-abundant transcripts even there are close phylogenetic distances between microbes in high diversity samples (Cooper et al., 2014; Toseland et al., 2014). Therefore, assembly-based methods could potentially provide complementary results to mapping-based methods, and further confirm the research question. In this study, the low coverage found in samples could be the largest obstacle. As mentioned, the risk of low coverage in genes may also impact on the performance of assembler. However, it will still provide useful information of highly expressed genes.

## Comparison of result of RGI between assembly and mapping-based methods

With further filtration with a deep-learning method: DeepARG. A total of 118 AMR genes from 51 samples with gene identity over 50% from all samples were recognized in mapping-based method. For assembly-based methods, assembly-per-sample method only found 24 AMR genes from 15 sample. Common genes found between mapping-based method and assembly-per-sample method, which wer*e ErmB, aadA, acrB, APH(3’)-lb, H-NS* and *ErmF* from five samples. The result may indicate that there is very high possibility to have these six genes in the samples.



*Figure5. Common genes found between mapping-based method (blue) and assembly-per-sample method (green)*

# Conclusion

From the report, several challenges were identified in utilizing metatranscriptome data from cancer tissue. First, the coverage of reads is crucial for the functional annotation of the data, regardless of whether mapping or assembly-based strategies are applied. Insufficient reads in samples can lead to an underestimation of AMR gene species and poor coverage uniformity within genes. This issue may also result from false mapping, where homologous reads are incorrectly aligned to reference genes.

Second, the decontamination of human genes and removal of rRNA is critical in metatranscriptome. After rRNA removal, the distribution of GC content of samples became more uniformed and overrepresented sequences were more informative. However, we found that overrepresented sequences can still aligned with human genes, as identified through BLAST analysis.

Third, there was little agreement between the genes identified by different methods. This finding suggests a high likelihood of these genes existing in the samples, yet it also indicates that a low number of reads may negatively impact the assembly outcomes.

The report supports the presence of certain genes in the tissue samples, particularly those identified by both mapping and assembly-based methods, which suggests a higher confidence in their existence. However, due to limited read coverage and overall read depth, the reliability of both approaches is compromised, making it difficult to draw definitive conclusions from this available dataset.

**Recommendations for Future Research:**

**mRNA Enrichment**: Implementing an mRNA enrichment kit during the sampling process could help increase the yield of relevant transcripts, thus improving the accuracy of both mapping and assembly-based approaches.

**Higher Coverage Sequencing**: Achieving higher sequencing coverage is crucial. It would ensure more comprehensive representation of the transcriptome, leading to better functional annotation and more reliable identification of AMR genes.

**Complementary Experimental Methods**: Techniques such as Fluorescence In Situ Hybridization (FISH) or Polymerase Chain Reaction (PCR) could be used to validate the presence of AMR genes in tissue samples. These methods offer a more direct and specific means of confirming the findings from metatranscriptomic analyses.

\*\*Therefore, the AMR genes analysis in the final report can be very unreliable using this dataset on Singapore cohort. This is also one of the reasons I write down this experimental report, as there is little chance to have reliable results from the data and it is limited to really know what genes and whether they are truly present unless experiments are performed. However, due to AMR genes were identified based on different method: mapping-based, assembly-based RGI method and HUMAnN3, there is high chance to have them in tissues.

# References

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., Edalatmand, A., Petkau, A., Syed, S. A., Tsang, K. K., Baker, S. J. C., Dave, M., McCarthy, M. C., Mukiri, K. M., Nasir, J. A., Golbon, B., Imtiaz, H., Jiang, X., Kaur, K., … McArthur, A. G. (2023). CARD 2023: Expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research*, *51*(D1), D690–D699. https://doi.org/10.1093/nar/gkac920

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Cooper, E. D., Bentlage, B., Gibbons, T. R., Bachvaroff, T. R., & Delwiche, C. F. (2014). Metatranscriptome profiling of a harmful algal bloom. *Harmful Algae*, *37*, 75–83. https://doi.org/10.1016/j.hal.2014.04.016

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., Giannoukos, G., Boylan, M. R., Ciulla, D., Gevers, D., Izard, J., Garrett, W. S., Chan, A. T., & Huttenhower, C. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences*, *111*(22), E2329–E2338. https://doi.org/10.1073/pnas.1319284111

Gerard, A., & María, C. (2022). *Functional exploration of natural product metabolism and metatranscriptomic analyses in uncultivated soil bacteria from the Talabre-Lejía transect (Atacama desert)*. https://repositorio.uchile.cl/handle/2250/191742

Gweon, H. S., Shaw, L. P., Swann, J., De Maio, N., AbuOun, M., Niehus, R., Hubbard, A. T. M., Bowes, M. J., Bailey, M. J., Peto, T. E. A., Hoosdally, S. J., Walker, A. S., Sebra, R. P., Crook, D. W., Anjum, M. F., Read, D. S., Stoesser, N., Abuoun, M., Anjum, M., … on behalf of the REHAB consortium. (2019). The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environmental Microbiome*, *14*(1), 7. https://doi.org/10.1186/s40793-019-0347-1

Hu, E.-Z., Lan, X.-R., Liu, Z.-L., Gao, J., & Niu, D.-K. (2022). A positive correlation between GC content and growth temperature in prokaryotes. *BMC Genomics*, *23*(1), 110. https://doi.org/10.1186/s12864-022-08353-7

Joanito, I., Wirapati, P., Zhao, N., Nawaz, Z., Yeo, G., Lee, F., Eng, C. L. P., Macalinao, D. C., Kahraman, M., Srinivasan, H., Lakshmanan, V., Verbandt, S., Tsantoulis, P., Gunn, N., Venkatesh, P. N., Poh, Z. W., Nahar, R., Oh, H. L. J., Loo, J. M., … Tan, I. B. (2022). Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nature Genetics*, *54*(7), 963–975. https://doi.org/10.1038/s41588-022-01100-4

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Lo, C.-C., & Chain, P. S. G. (2014). Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics*, *15*(1), 366. https://doi.org/10.1186/s12859-014-0366-2

Pereira-Marques, J., Ferreira, R. M., & Figueiredo, C. (2024). A metatranscriptomics strategy for efficient characterization of the microbiome in human tissues with low microbial biomass. *Gut Microbes*, *16*(1), 2323235. https://doi.org/10.1080/19490976.2024.2323235

Toseland, A., Moxon, S., Mock, T., & Moulton, V. (2014). Metatranscriptomes from diverse microbial communities: Assessment of data reduction techniques for rigorous annotation. *BMC Genomics*, *15*(1), 901. https://doi.org/10.1186/1471-2164-15-901

