Metagenomic taxonomic classification, binning, and genetic composition of the vaginal microbiome through metaWRAP and HUMAnN 3.0

Sneh Koul

Abstract—The vaginal microbiome is an important element to women's health. If imbalances do arise this leads to a variety of vaginal diseases, which is why it is important to study the vaginal microbiome. Alongside the rising concern of women's health, we have the rise of metagenomic sequences. With the cost of Next Generation Sequencing reducing significantly, there are many metagenomic datasets being sequenced. By combining the concept of metagenomics and vaginal microbiome, we were able to study the vaginal microbiome through analyzing metagenomic data obtained from the mid vagina and posterior fornix from the Human Microbiome Project. Metagenomic pipeline anlaysis was conducted through metaWRAP and HUMAnN to examine the taxonomic classification, binning, and genetic composition. Taxonomic classification revealed an overwhelming amount of Lactobacillus bacteria present in all samples. Binning analysis revealed three bins along with similar taxonomy. A number of gene families were also identified along with numerous pathways associated with energy metabolism. These results are consist with findings in literature but more research needs to be done for advancements to be made in vaginal health.

Index Terms—Metagenomics, Taxonomy, Binning, metaWRAP, HUMAnN 3.0, Vaginal microbiome

1 Introduction

Metagenomics is the field of examining genomes within a particular community. The community could be defined in a variety of settings like the environment or the human body. This field initially began with the examination of environmental DNA in order to better understand the microbial world around through uncovering the genetic diversity [1]. Besides exploring the genetic compositions of these communities, researchers are also able to grasp a better understanding of the phylogeny of the organisms and with both these elements many advancements have been made within the fields of microbial ecology and evolution [1]. Additionally, with the significant reduction of the cost of next generation sequencing (NGS), many advancements are yet come as a results of the growth of metagenome datasets.

As a result of the growth of sequence based datasets it is important to understand the basic pipeline of sequence based metagenomic analysis. The initial step is obtaining samples and processing. The processing is specifically based of the sample, but the overall goal is the same which is DNA extraction. The next step in the pipeline is sequencing which could be completed through Sanger sequencing or NGS, but in the past 10 years we have seen a shift from Sanger to NGS. Specifically of the NGS technologies, 454/Roche and Illumina have been extensively used for metagenomic studies [1]. Following sequencing is assembly which is the assembly of short DNA sequences into genes or organisms [2]. The two main strategies for assembly of metagenomic samples are reference-based or *de novo* assembly [1]. After the assembly process, a researcher may consider binning. Binning is the process of grouping DNA sequences based of off how closely they might portray the genome of a related organism. [1]. If a researcher decides not to go the route of binning, annotation is the next major step. Annotation is usually completed in two steps: 1) Identifying genes 2) Identifying gene functions. For both the binning and annotation there are a variety of computational tools to help like MG-RAST, MEGAN, and MetaGeneAnnotator. Overall this pipeline is able to provide researchers with novel information with regards to taxonomy and genetic classification.

Now as previously stated this metagenomic pipeline application was initially used in the environment setting and for this reason we wanted to branch out and conduct a metagenomic analysis of the human vaginal microbiome specifically examining the taxonomy, binning classification, and genetic composition.

The human vaginal microbiome are microorgansims which are established in the vagina. Research has shown that the inhabitants of the vagina are mostly (around 70%) *Lactobacillus* [3]. *Lactobacilli* are known, specifically in the vagina, for inhibiting the binding of other bacteria along with lactic acid production [4]. Lactic acid production leads to several benefits like "enhancing gene transcription and DNA repair" and promotes homeostasis [4]. The abundance of *Lactobacilli* are influenced by a variety of internal and external factors. For example, emotional stress can potentially reduce the abundance of *Lactobacilli* which can lead to increased levels of inflammation [4]. Now why would *Lactobacilli* be the most abundant bacteria in the vaginal microbiome? There are quite a few theories including the unique "reproductive physiology, risk of STDs, and risk of microbial complications linked to pregnancy and birth" [3]. This shows us how important it is to study the vaginal micobiome as it plays a huge role in women's health.

If there are bacterial imbalances in the microbiome this can lead to a lot of complications like diseases. With the need to better understand the vaginal microbiome we employed a metagenomic pipeline analysis (metaWRAP and HUMAnN) on vaginal samples obtained from the Human Microbiome Project (HMP) to examine the taxonomic classification, genetic composition, and binning analysis. Along with gaining further knowledge about vaginal microbiome we also evaluated the similarities and differences among the pipeline analysis tools.

2 Methodology

2.1 Dataset

The dataset consisted of three samples obtained from the HMP. Two of the samples were from the mid vagina (SRS014466 and SRS015072) while the other sample was from the posterior fornix (SRS014343).

2.2 Databases

The databases required for these pipelines include: CheckM, MiniKraken DB, NCBI_nt BLAST database, NCBI taxonomy, ChocoPHlAn, and uniref90.

2.3 Metagenomic pipeline analysis

Two metagenomic pipelines were used to analyze the vaginal microbiome: metaWRAP and HUMAnN. The metaWRAP pipeline is specifically known for the binning analysis, while HUMAnN is known for taxonomy classification and genetic composition/functional annotation.

2.3.1 metaWRAP

metaWRAP tool consists of numerous steps which generally follows the sequence-based metagenomic pipeline discussed earlier. Figure 1 shows the basic pipeline workflow [5].

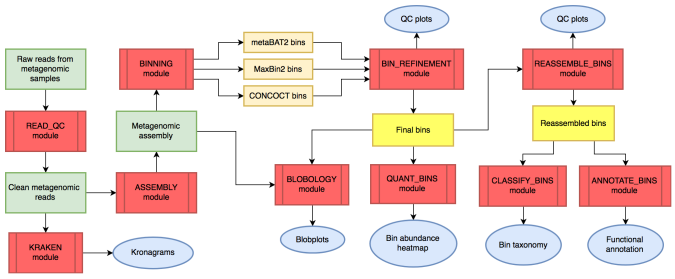

Fig. 1: metaWRAP pipeline workflow [5]

Briefly, after the samples were downloaded from the HMP the reads went through a trimming process, which removes and trims low quality reads, using the *read_qc* function. Following the read trimming, all samples were co assembled using the metaSPAdes feature which is apart of the *assembly* function. Now in order to determine the taxonomic composition the KRAKEN tool was utilized. It is important to note that due to computational constraints the KRAKEN standard database was not utilized instead a smaller version of the database, the MiniKraken database, was used. After determining the taxonomic classification of the samples, the

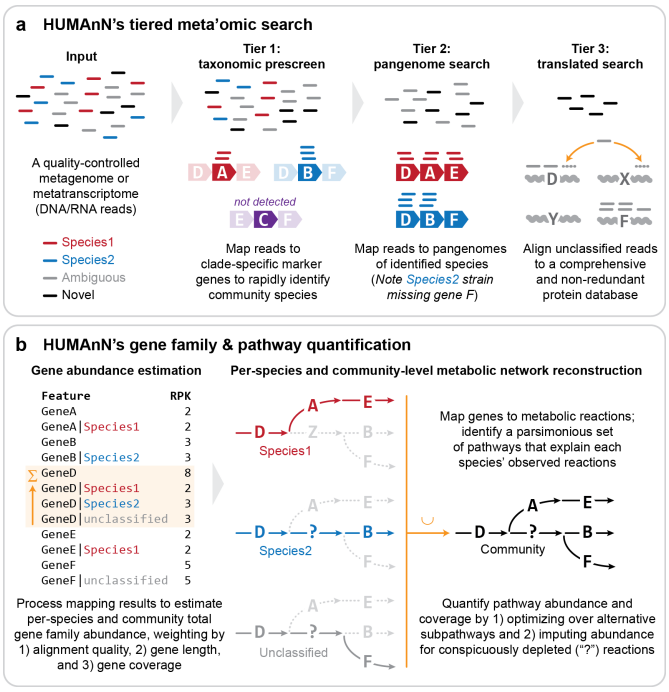

Fig. 2: HUMAnN pipeline workflow [6]

binning module was applied on the co-assembled reads. The initial phase of the binning process is binning with three different algorithms: CONCOCT, MaxBin, and metaBAT. This is then followed by the bin_refinement module to generate a single bin. It is important to emphasis that the parameters for minimum completion and maximum contamination were set to 50% and 10% due to relative small size of these samples. Now in order to visualize the bins across the samples, the *quant_bins* is run to provide the abundances of the bins across samples through a heatmap. The visualization of the bins is followed by reassembly of the bins in order to further improve the bins. With these reassembled bins one can now examine the taxonomy as well as the genetic composition of the bins through the functions *classify_bins* and *annotate_bins*, respectively.

2.3.2 HUMAnN 3.0

The HUMAnN pipeline is more focused on taxonomy along with gene classification. Figure 2 shows a representation of the pipeline [6]. Briefly, input reads go through taxonomy screen by using the tool MetaPhlAn3. This is followed by mapping the reads to identified species through using Bowtie2. Finally, unmapped reads are aligned to the uniref90 database with DIAMOND.

3 Results

3.1 Taxonomy Classification

The first vaginal sample revealed, through the Kraken module of metaWRAP, about 58% classification with the Bacteria kingdom and about 42% remained unclassified. The largest specie classification was *Lactobacillus crispatus* with about a 42% match as shown in Figure 3. The second vaginal sample revealed about a 45% classification with the Bacteria kingdom and about 55% remained unclassified. The

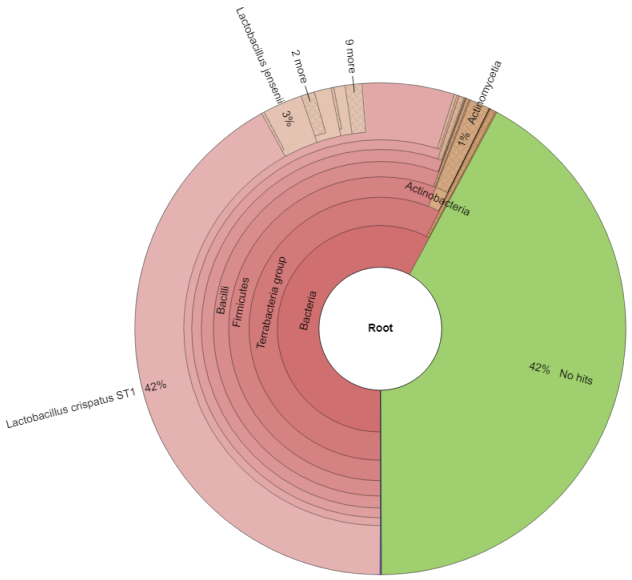

Fig. 3: Taxonomy Classification of sample SRS014466

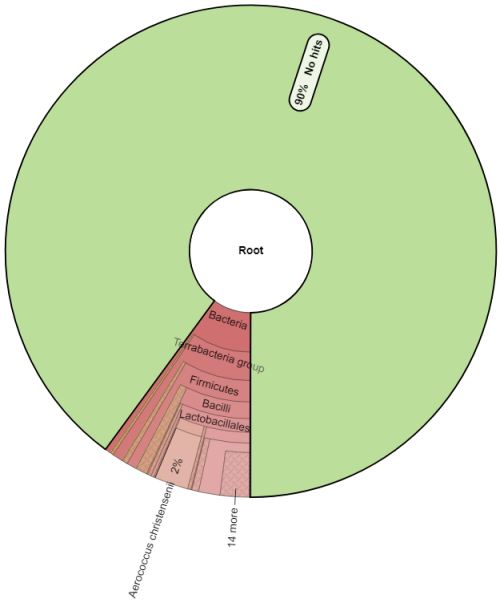

Fig. 5: Taxonomy Classification of sample SRS014343

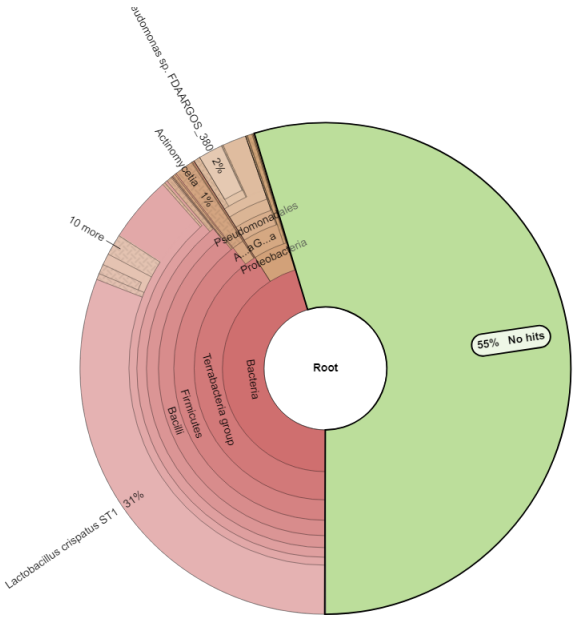

Fig. 4: Taxonomy Classification of sample SRS015072

largest specie classification was also *Lactobacillus crispatus* with about a 31% match as shown in Figure 4. The final sample from the posterior fornix revealed only around a 10% classification with the Bacteria kingdom and about 90% remained unclassified. The largest group classified was Lactobacillales with about a 7% match. A full report of the taxonomy of each sample is included in the supplemental.

Now taxonomy classification determined through HUMAnN revealed 7 species for the first vaginal sample (SRS014466): *Lactobacillus crispatus*, *Lactobacillus iners*, *Lactobacillus jensenii*, *Lactobacilluscoleohominis*, *Actinomyces oris*, *Prevotella bivia*, and *Mycobacteroides chelonae*. The second

Species	metaWRAP (%mapped reads)	HUMAnN (%mapped reads)
Lactobacillus crispatus	41.96	78.12
Lactobacillus iners	N/A	18.17
Lactobacillus jensenii	2.81	3.19
Lactobacilluscoleohominis	N/A	0.34
Actinomyces oris	0.10	0.12
Prevotella bivia	N/A	0.04
Mycobacteroides chelonae	N/A	0.02

Fig. 6: Differences in % mapped reads of species identified in SRS014466: metaWRAP vs HUMAnN

vaginal sample (SRS015072) revealed 7 species: *Lactobacillus crispatus*, *Lactobacillus iners*, *Pseudomonas fluorescens*, *Gardnerella vaginalis*, *Mycobacteroides chelonae*, *Lactobacillus jensenii*, and *Ureaplasma parvum*. The posterior fornix sample revealed 6 species: *Lactobacillus iners*, *Lactobacillus jensenii*,*Lactobacillus vaginalils*, *Lactobacillus gasseri*, *Lactobacilluscoleohominis*, *Staphylococcus epidermidis*, and *Lactobacillus crispatus*. Overall we can see that metaWRAP, specifically KRAKEN, was able to identify more species compared to the HUMAnN tool. It is also important to note that species identified by HUMAnN were not identified by metaWRAP. In Figure 6, for example we can specifically see this along with the differences in terms of the percent of mapped reads between the two tools when specifically looking at the first vaginal sample (SRS014466).

3.2 Binning Analysis

The initial binning revealed 31, 3, 5 bins with CONCOCT, MaxBin, and metaBAT. Following bin refinement, we found 5,3,5, and 3 bins with CONCOCT, MaxBin, metaBAT, and metaWRAP, respectively. Now in order to examine bin sets for completion and contamination Figure 7 reveals the ranking among the different binning algorithms. This reveals that the bin refining process does somewhat help produce a okay bin set (metaWRAP) when examining the completion and contamination.

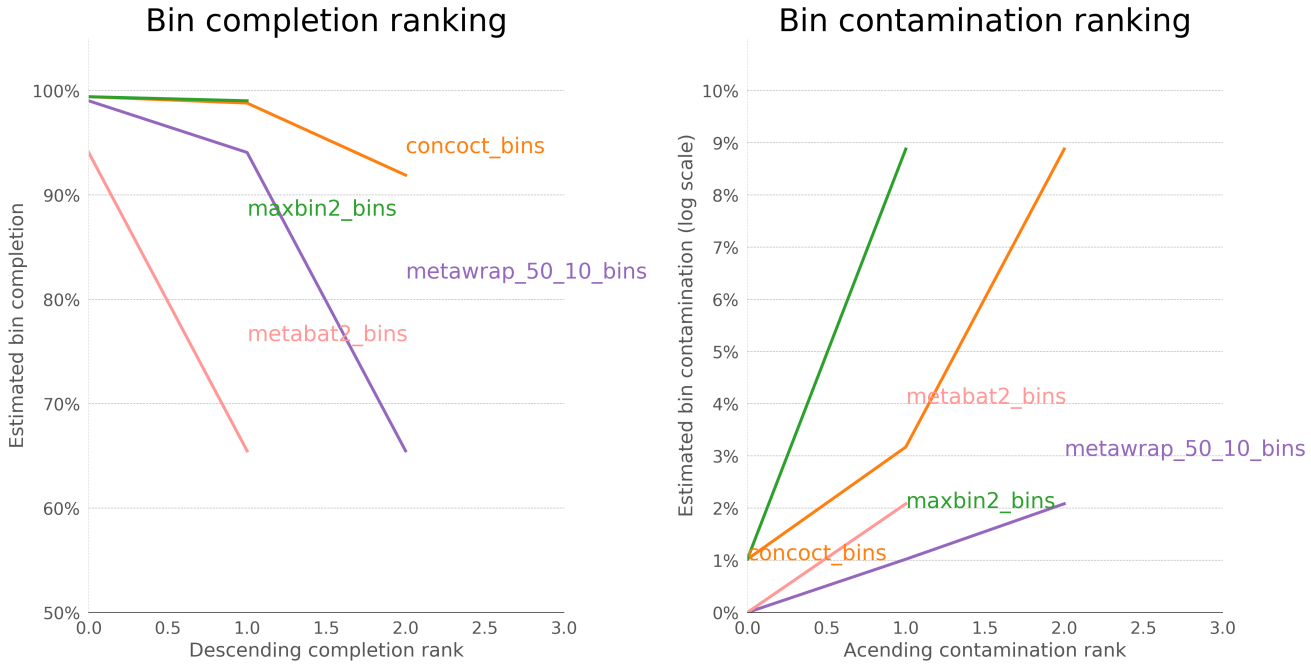

Fig. 7: Binning contamination and completion of samples from the vaginal microbiome

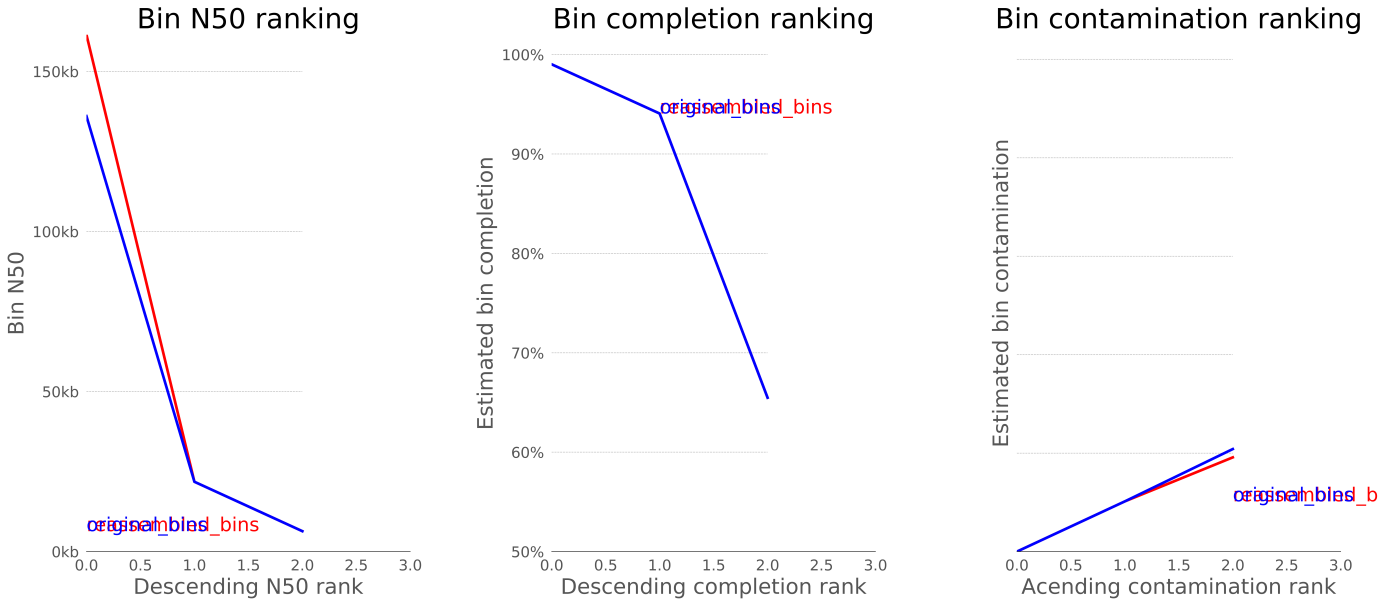

Fig. 8: Binning contamination and completion of samples from the vaginal microbiome followed by reassembly

Now with one solid bin set reassembly was applied to further improve the bin set. Figure 8 reveals that reassembly didn't make much of further improves as we can overlaps between the reassembled bins and original bins. Among these reassembled bins the taxonomy has shown to include *Lactobacillus crispatus* and *Limosilactobacillus vaginalis*. The number of translated genes identified are 1672, 728, and 2476 for bin 1, bin 2, and bin 3 respectively.

3.3 Gene Families and Pathway Abundance

The first vaginal sample revealed 13,727 gene families. The second vaginal sample revealed 16,249 gene families. The posterior fornix sample revealed 18,763 gene families. As one can see there is quite a lot of gene families among the samples but it important to note that some of these can be identified as unclassified.

Now looking at the pathways found to be associated amongst these samples there seems to be quite a few. Among all samples we see pathways like pyruvate fermentation, fatty acid biosynthesis initiation, tRNA charging, coenzyme A biosynthesis, UMP biosynthesis being highlighted. Full reports of the gene families and pathways are included in the supplemental.

4 Discussion

4.1 Taxonomy Classification

The vaginal microbiome is a unique environment in terms of what microorganisms inhabit it. Through our taxonomic classification, of a small dataset representing the vaginal microbiome, we saw the dominance of numerous *Lactobacillus bacteria*. As discussed in the introduction, *Lactobacilli* are known to be dominating in the vaginal microbiome and this work along with other metagenomic studies confirm this finding [7], [8], [9]. It is interesting though to see some bacteria which are not commonly found in a healthy vaginal microbiome like *Staphylococcus epidermidis* and *Gardnerella vaginalis*. Both of these bacteria are usually found in women who may face a vaginal disease like Bacterial Vaginosis (BV) [10]. Now when comparing metaWRAP and HUMAnN in terms of taxonomic classification we see that the metaWRAP was able to identify a lot more species. The reason is possibly due to the databases used. Some of the databases may not have all species included but it is interesting that metaWRAP was able to find more considering the fact that a significantly smaller (4G) database was used. Now we also see the taxonomy results of the bins to also be similar to results from KRAKEN and HUMAnN which shouldn't be surprising.

4.2 Binning Analysis

Following taxonomic classification was the binning analysis. We saw that metaWRAP uses three different algorithms initially followed by a combined approach. Now with a simple literature search there really isn't any paper published comparing these algorithms specifically for the vaginal microbiome. It is known that CONCOCT, MaxBin, and metaBAT are well known and reliable binning algorithms, but it is important to note what makes a good binning algorithm is based of off completion and contamination. Completion means "the level of coverage of a population genome" while contamination is "the amount of sequence that does not belong to this population from another genome" [5]. From Figure 7, we can see that both MaxBin and CONCOCT perform better compared to metaWRAP in terms of bin completion, but metaWRAP performs better in terms of bin contamination. This is interesting because from the implementation of metaWRAP relative to these other algorithms, metaWRAP has been shown to perform better but also have the highest completion and lowest contamination [5]. One potential reason metaWRAP may not have performed at the top could be due to the relatively small dataset used in this study.

4.3 Pathway analysis

We saw several gene families and pathways represented among all the samples. Looking at these pathways we see that most of them are associated with energy metabolism and research has shown a strong link between energy metabolites and the vaginal microbiome. It has been shown that quite a lot of external factors can influence the metabolism in relation to the vaginal microbiome. For example, our analysis revealed fatty acid biosynthesis to be a prominent pathway and research has shown that fatty acids, specifically short chain, can lead to dysbiosis of the vaginal microbiome [11]. Additionally, stress and ethnicity have also been shown to influence metabolic profiles which is indirectly able to impact the vaginal microbiome [11], [12].Now these pathways described are only a small glimpse of truly what the vaginal microbiome has to show. Through more metagenomic studies we can gather more knowledge in terms of species and genetic composition which will can hopefully help in advancements of vaginal related diseases.

5 Future directions

The dataset used in this study was relatively small and in the future utilizing a larger datasets would provide more a holistic view of the vaginal microbiome. Additionally, we would use a machine which would be able to handle more rigorous computations along with databases.

6 Supplemental

In the supplemental folder, one will find FastQC report of all samples (, QUAST assembly report, bind abundance heatmap, KRAKEN report for each sample (folder), KRONA taxonomy representation, functional annotation of each bin (folder), and gene families/pathway abundance/coverage for each sample(folder). Scripts used have also been included.

7 Additional Information

metaWRAP pipeline can be found here https://github.com/bxlab/metaWRAP and the HUMAnN pipeline can be found here https://github.com/biobakery/biobakery/wiki/humann3 .

REFERENCES

[1] T. Thomas, J. Gilbert, and F. Meyer, "Metagenomics-a guide from sampling to data analysis," *Microbial informatics and experimentation*, vol. 2, no. 1, pp. 1–12, 2012.

[2] J. S. Ghurye, V. Cepeda-Espinoza, and M. Pop, "Focus: Microbiome: Metagenomic assembly: Overview, challenges and applications," *The Yale journal of biology and medicine*, vol. 89, no. 3, p. 353, 2016.

[3] E. A. Miller, D. E. Beasley, R. R. Dunn, and E. A. Archie, "Lactobacilli dominance and vaginal ph: why is the human vaginal microbiome unique?," *Frontiers in microbiology*, vol. 7, p. 1936, 2016.

[4] S. S. Witkin and I. M. Linhares, "Why do lactobacilli dominate the human vaginal microbiota?," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 124, no. 4, pp. 606–611, 2017.

[5] G. V. Uritskiy, J. DiRuggiero, and J. Taylor, "Metawrapa flexible pipeline for genome-resolved metagenomic data analysis," *Microbiome*, vol. 6, no. 1, pp. 1–13, 2018.

[6] F. Beghini, L. J. McIver, A. Blanco-Míguez, L. Dubois, F. Asnicar, S. Maharjan, A. Mailyan, P. Manghi, M. Scholz, A. M. Thomas, *et al.*, "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3," *Elife*, vol. 10, p. e65088, 2021.

[7] F. Liu, Y. Zhou, L. Zhu, Z. Wang, L. Ma, Y. He, and P. Fu, "Comparative metagenomic analysis of the vaginal microbiome in healthy women," *Synthetic and systems biotechnology*, vol. 6, no. 2, pp. 77–84, 2021.

[8] H. Berman, M. McLaren, and B. Callahan, "Understanding and interpreting community sequencing measurements of the vaginal microbiome," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 127, no. 2, pp. 139–146, 2020.

[9] O. Mehta, T. S. Ghosh, A. Kothidar, M. R. Gowtham, R. Mitra, P. Kshetrapal, N. Wadhwa, R. Thiruvengadam, S. Bhatnagar, B. Das, *et al.*, "Vaginal microbiome of pregnant indian women: Insights into the genome of dominant lactobacillus species.," *Microbial ecology*, vol. 80, no. 2, 2020.

[10] M. Wilks, R. Thin, and S. Tabaqchali, "Quantitative bacteriology of the vaginal flora in genital disease," *Journal of medical microbiology*, vol. 18, no. 2, pp. 217–231, 1984.

[11] S. Baldewijns, M. Sillen, I. Palmans, P. Vandecruys, P. Van Dijck, and L. Demuyser, "The role of fatty acid metabolites in vaginal health and disease: application to candidiasis," *Frontiers in microbiology*, p. 1656, 2021.

[12] E. Jašarević, C. L. Howerton, C. D. Howard, and T. L. Bale, "Alterations in the vaginal microbiome by maternal stress are associated with metabolic reprogramming of the offspring gut and brain," *Endocrinology*, vol. 156, no. 9, pp. 3265–3276, 2015.