Metagenomic Analysis of Army Ant Guts

ECES 480/490

DHANTHA GUNARATHNA, FEIYANG XUE, ARIANA ENTEZARI

Abstract

Metagenomic analysis is a powerful tool for analysis of the taxonomic classification and function annotation of environmental samples. In this study, we performed a metagenomic analysis of army ant gut samples from ants collected at two sites in Costa Rica. Filtering and trimming of the Illumina sequencing data was first done. Taxonomic classification was then performed using unassembled and assembled sequences. Preliminary functional annotation was done using HUMAnN2. Abundant bacteria in the samples include Lactobacillales and Propionibacteria from the phyla Firmicutes and Actinobacteria, respectively. Overall, taxonomic classification varied widely by sample and by analysis pipeline.

Background

Metagenomic analysis gives valuable insight to the genetic content of communities of organisms. In recent years, it has provided a pathway for progress in fields such as microbial ecology, evolution, and diversity [1]. Not only can classification of communities of microbes be identified, but their functional gene composition can also be analyzed. This gives a more comprehensive report than a typical 16S rRNA analysis [1].

Ants are well-studied due to the symbiotic relationship they have with bacteria, including in their guts, which help with important processes such as digestion and nutrition [2]. These gut bacteria vary widely among ant species and may be related to environmental factors. Ant species such as *Cephalotes varians* have been found to have low species-level diversity in their guts [2]. In addition, the genus *Cephalotes* has been found to have little variation between colonies [2]. Army ants are known for their nomadic behavior and the substantial ecological impact they have as they move from location to location. They are also known for their characteristic group predation, in which they move in swarms to predate and collect food for their colonies [3].

This paper focuses on the metagenomic analysis of 5 samples of army ants from sites in Costa Rica. The lab of Dr. Jacob Russell collected samples of *Eciton burchelli parvispinum* from 2 colonies in Santa Rosa and 2 colonies in Monteverde and *Labidus praedator* from a colony in Monteverde (Table 1).

Table 1. Samples collected from Costa Rica.

Sample ID	Species	Colony	Site (in Costa Rica)
1	*Eciton burchelli parvispinum*	MVEbp1	Monteverde
2	*Eciton burchelli parvispinum*	MVEbp2	Monteverde
3	*Eciton burchelli parvispinum*	SREbp1	Santa Rosa
4	*Eciton burchelli parvispinum*	SREbp2	Santa Rosa
5	*Labidus praedator*	MVLprae1	Monteverde

They then isolated the guts of these ants and performed DNA extraction prior to sequencing. Samples were sent to UC Berkeley for Illumina Sequencing. Sequencing data was received in the form of paired-end sequencing files in fastq format.

Metagenomic analysis was done on these samples using various pipelines, focusing on the differences in results between unassembled and assembled sequencing data. Results showed that taxonomic classification varied widely between unassembled and assembled sequences, and MetaPhlAn2 gave very different results than the other packages.

Methods

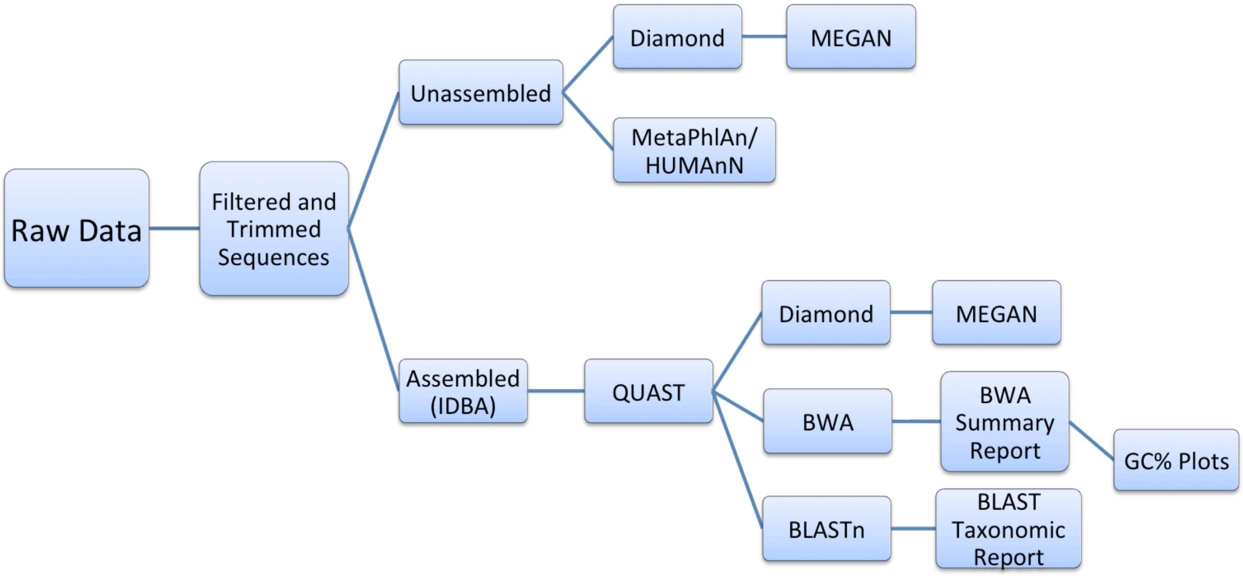

Figure 1. Flow chart of metagenomics analysis pipeline used in the study.

FastQC

FastQC was used to generate quality reports of the raw sequencing data. After inputting the 10 fastq files, outputs for each file gave a quality report of the sequences. The reports contained per base sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, any overrepresented sequences and adapter contents, and Kmer content. FastQC reports before and after filter and trimming can be found on our Github.

KneadData Pipeline

KneadData, a quality control tool for metagenomic sequencing data, was performed on all 5 samples using paired-end inputs. The KneadData pipeline used BowTie2-Build, Trimmomatic, and BowTie2.

BowTie2-Build (Custom Database)

A custom database of host genomes needed to be built to remove possible contamination. Ant genomes from the family *Formicidae* were compiled from an NCBI search to form a reference database while a human reference database was already present in BowTie2 [4]. To check that when the ant genomes were sequenced, they did not have bacterial contamination, a megablast BLASTn search was done using a conserved 16S rRNA gene from an Entomoplasmataceae bacterium against the compiled ant genomes. Not hits were found, so it can be inferred that there is unlikely to be significant bacterial contamination in the sequences genomes.

.

Trimommatic

The KneadData pipeline first uses Trimmomatic to trim adapter sequences and remove leading and trailing low quality reads [5]. It also uses a sliding window approach to trimming by scanning the sequences with a window of 4 nucleotides and trimming the sequences when the average quality score is below 20. It also drops reads less than 70 bases long.

BowTie2

KneadData then uses BowTie2 to remove human and host contamination [4]. It does so by removing sequences that align to the reference databases built from BowTie2-Build, as this indicates contamination.

Unassembled Sequences Pipeline

MetaPhlAn2

MetaPhlAn2 was used to profile sequences from bacteria, archaea, eukaryotes, and viruses, using around 1 million clade-specific marker genes [6]. The sequences are profiled into relative abundances of reads per sample, analyzed by taxonomic classification.

HUMAnN2

HUMAnN2 is unified metabolic analysis network used for functional annotation [7] . It can annotate metabolic pathways in metagenomic data to help understand the functionality of a microbial community.

Assembled Sequences Pipeline

IDBA

IDBA was used to assemble the paired-end sequences with iterative de Brujn graph assembly. The scaffold fasta file output was used as the assembled data for further methods. Due to a hardcoded limit on read length [8], the source code of IDBA was modified in order to use pair end reads longer than 128.

QUAST

QUAST was used for quality assessment of the assembly by IDBA [9]. The IDBA scaffold file was input. n50 numbers were taken to assess the length that the collection of all contigs of that length of longer accounts for at least 50% of the assembly.

Burrows Wheeler Alignment (BWA)

BWA is an aligner used to map sequences against a reference genome [10]. In this instance, it was used to map the paired-end, filtered, trimmed sequences to the scaffold file from IDBA.

Summary Report

A summary report of BWA was generated to find the length of each scaffold, its read-depth coverage, and its GC content [11].

BLASTn

Scaffolds from the scaffold file output from IDBA were blasted against NCBI non-redundant nucleotide (nt) database using the megablast option to find best species hit [12].

Taxonomy Report

Since BLASTn only gives a taxonomic ID, the taxonomic classification needs to be given. This was done using a script to convert the taxonomic ID for each scaffold to a taxonomic classification [11].

GC Plots

GC plots were created with a modified script from Kumar *et al.* [13]. The input for these plots were GC content, read coverage, taxonomic classification, and scaffold length.

Diamond and MEGAN on Unassembled and Assembled Sequences Pipeline

This pipeline is for processing either assembled or unassembled reads that have been aligned to perform taxonomy analysis. Furthermore, functional analysis can also be achieved in this pipeline. The map below provides an overview to the pipeline (Figure 2).

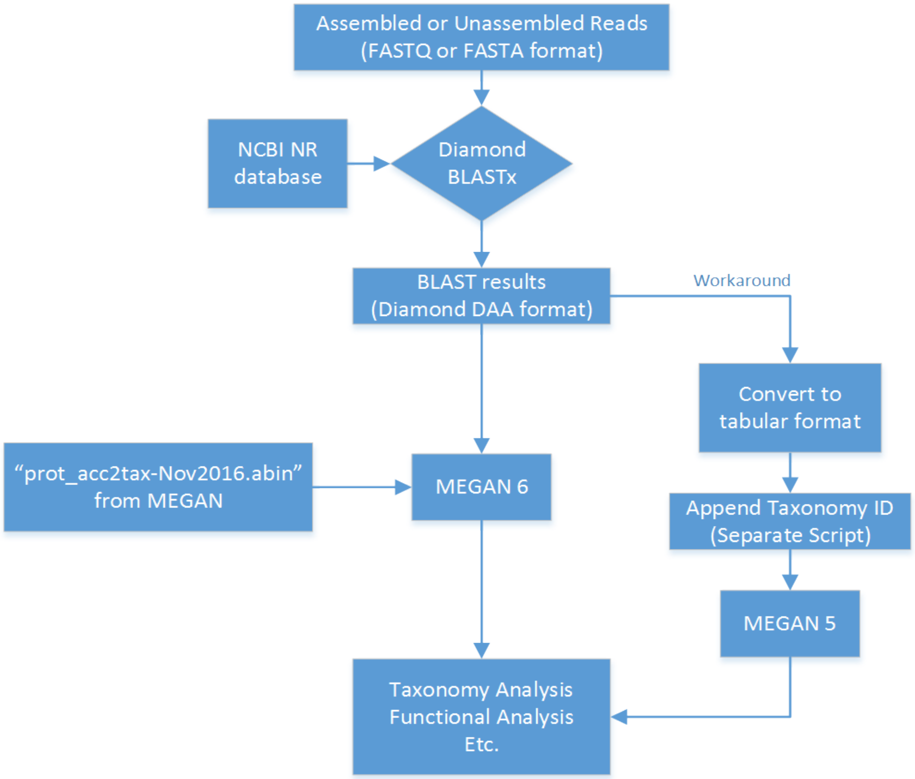

Figure 2. Flowchart for Diamond-MEGAN pipeline

This process was designed and carried out under the constraints that MEGAN has to be run on a local computer, and Diamond can be run on a high performance computing machine. This situation exists when a high performance computing machine is available without graphical interface including X11, while the command-line-enabled MEGAN 6 Ultimate Edition is unavailable due to cost of license.

Diamond

Diamond is an open-source software for sequence reads alignment against a protein reference database, and it functions as a drop-in replacement to BLASTX [14]. The tool runs offline, and it can reach about 20,000 times faster than BLASTX with short reads.

In this project, Diamond was used on both unassembled and assembled reads. Inputs to diamond are either fastq file from KneadData (as unassembled) or fasta file from IDBA (as assembled). In addition, diamond uses a NR protein database from NCBI. Diamond is a resource demanding software, especially in terms of memory and temporary storage. It is able to take advantage of a parallel environment and large memory.

MEGAN6

MEGAN is a toolbox for microbiome data analysis in an graphical and interactive fashion. Recent releases of this software include MEGAN 6 Ultimate Edition, MEGAN 6 Community Edition and MEGAN 5 [15]. It is worth noting that MEGAN 6 Community Edition is a free software with GPL license, and MEGAN 5 offers free academic license. The MEGAN 6 branch in this pipeline was carried out with the community edition.

Import of data can be slow with MEGAN 6, so it includes a feature of importing DAA files from Diamond, during which taxonomy ID is looked up with an accession number. This lookup is necessary as Diamond output is labelled with an NCBI accession number, while MEGAN relies on Taxonomy ID for its analysis. A mapping file is provided on MEGAN website for this purpose. Since this lookup process can be slow on certain computers, the MEGAN 5 branch in the map serves as an workaround, as it uses another server to perform lookup and append taxonomy ID. This workaround should be avoided unless the local computer is unable to import data within a reasonable timeframe.

Results/Discussion

Various pipelines were used to classify taxonomy in the ant gut samples. A flow diagram of the process shows that all pipelines used filtered, trimmed sequences output from KneadData (Figure 1).

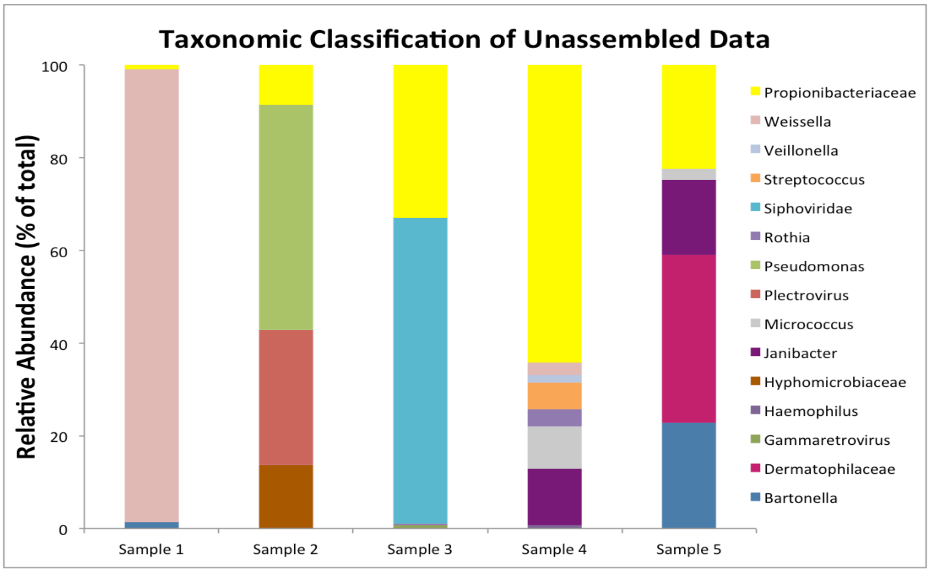

Figure 3. MetaPhlAn2 analysis of unassembled sequences that were filtered and trimmed, organized by order. Note the variation in relative abundance, defined by number of reads out of the total for the sample, among samples. There is a high abundance of viral sequences in samples 2, 3, and 5. In Samples 1-5, the highest relative abundance of reads were from the orders Weisella, Pseudomonas, Siphoviridae, Propionibacteriaceae, and Plectrovirus, respectively.

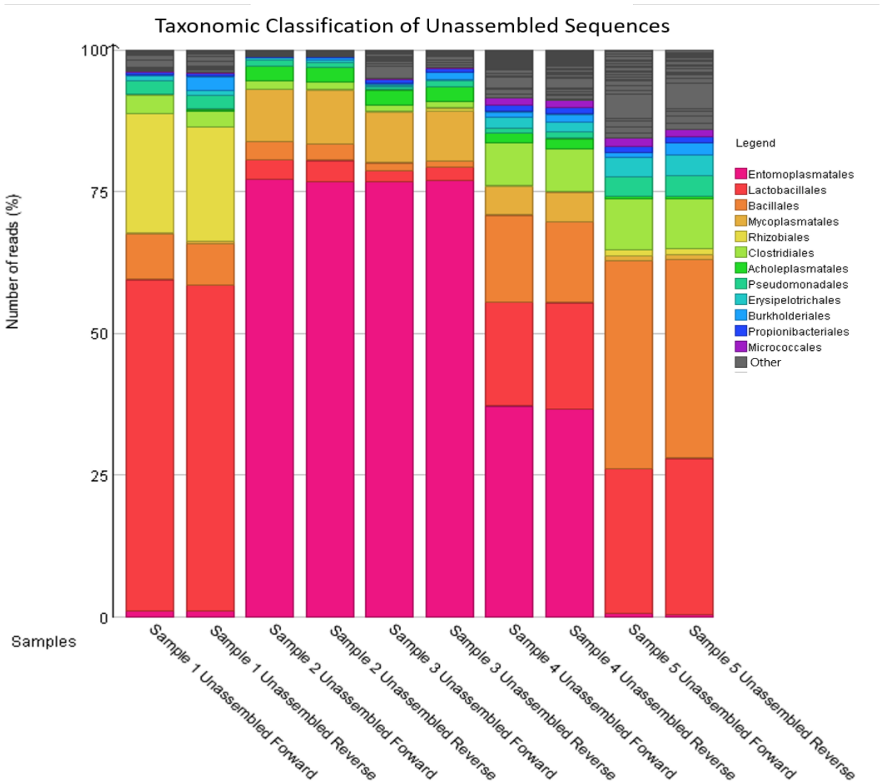

Figure 4. Taxonomic classification of unassembled sequences using Diamond and MEGAN6. Diamond was used to align sequences to the NCBI protein database and MEGAN6 was used to taxonomically classify them. The highest relative abundance of reads from Sample 1 was from the order Lactobacillales. In Samples 2, 3, and 4, the highest relative abundance of reads were from the order Entomoplasmatales. In Sample 5, the highest relative abundance of reads was from the order Mycoplasmatales.

Differences between MetaPhlAn2 and MEGAN6 can be noted in the high abundance of viral data in MetaPhlAn2, which is lacking in MEGAN6. In addition, MetaPhlAn2 does not classify the ant contamination that is abundantly seen in MEGAN6. For purposes of focusing on microbial data, this contamination was removed from MEGAN6 plots (Figures 3, 6).

Classification was also done on data assembled with IDBA. These assemblies were assessed for quality by running QUAST. N50 numbers were assessed to be of high quality (Table 2), so the assemblies were used to align the sequences using BWA.

Table 2. N50 numbers from the summary of the IDBA assembly.

Sample	N50
1	36261
2	23255
3	16459
4	24970
5	18541

The assembled data was run on BLASTn and the best hit was found for each contig. It was then run through a taxonomic report to find the taxonomic classification.

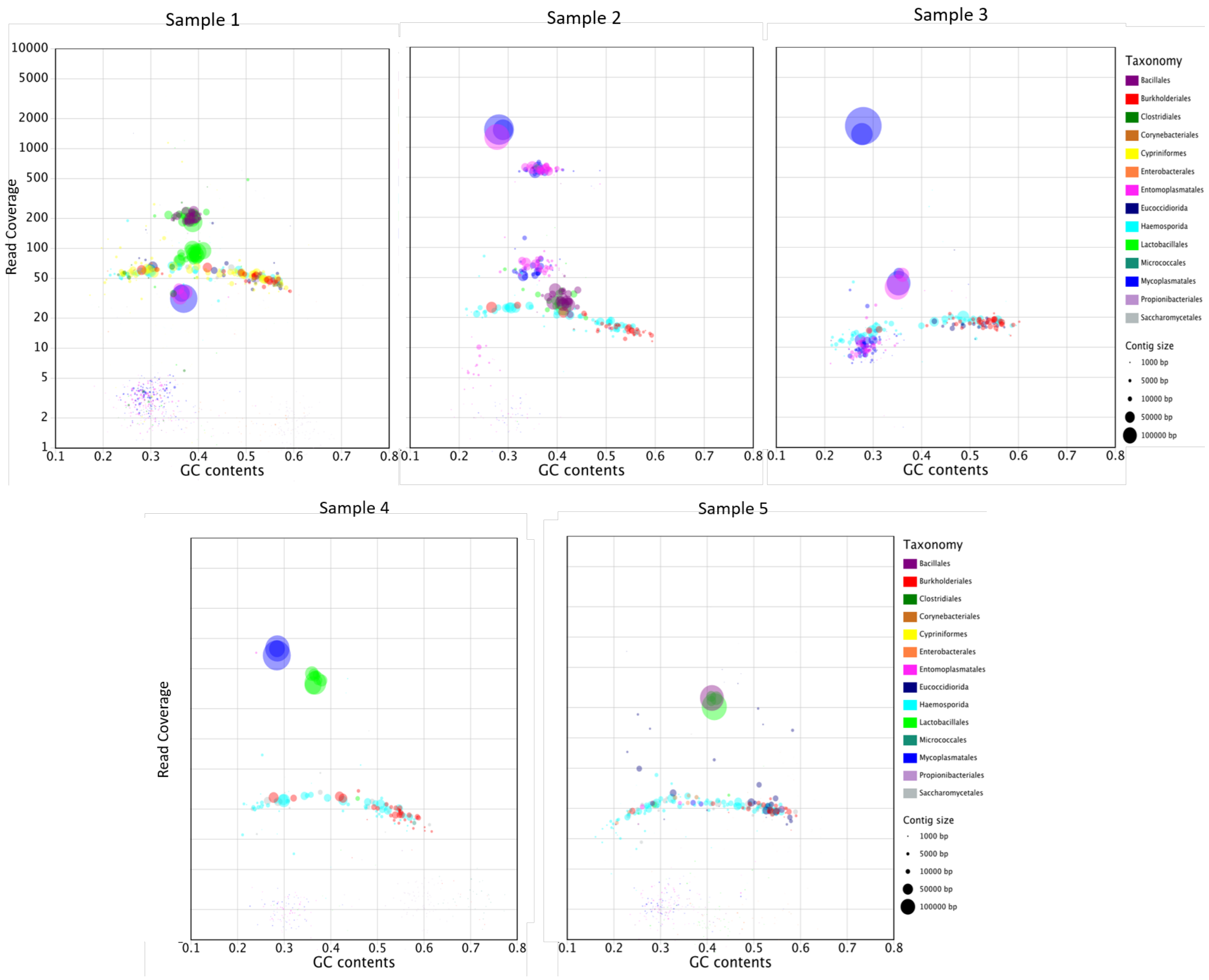

Figure 5. GC Plots of the 5 samples. Taxonomic classification is color-labelled, and circle size of each contig is proportional to its size. This classification was combined with the length of each contig, read coverage, and GC composition generated from a BWA summary report to generate

the plots. A similar spread of Haemosporida is seen across all 5 samples, suggesting this bacteria is conserved among ant species Similarly, a cluster of Mycoplasmatales is seen in 3 of the 5 samples, and it has a high read coverage, emphasizing its abundance in the samples. There are clusters of different colored contigs, such as Entomosplasmatales and Bacillales clustered together, which may be a result of inaccuracy in the BLASTn search. It chooses the top hit, but there may have been a similarly accurate next hit. A similar clustering is unsurprisingly seen between Bacillales and Lactobacillales, likely because of their close relatedness. It is important to note that to make these GC plots, eukaryotic contigs were removed. These contigs were likely from ant DNA contamination, and took up a majority of the plot. By removing these contigs, it was easier to visualize the trends of gut data using the GC plots.

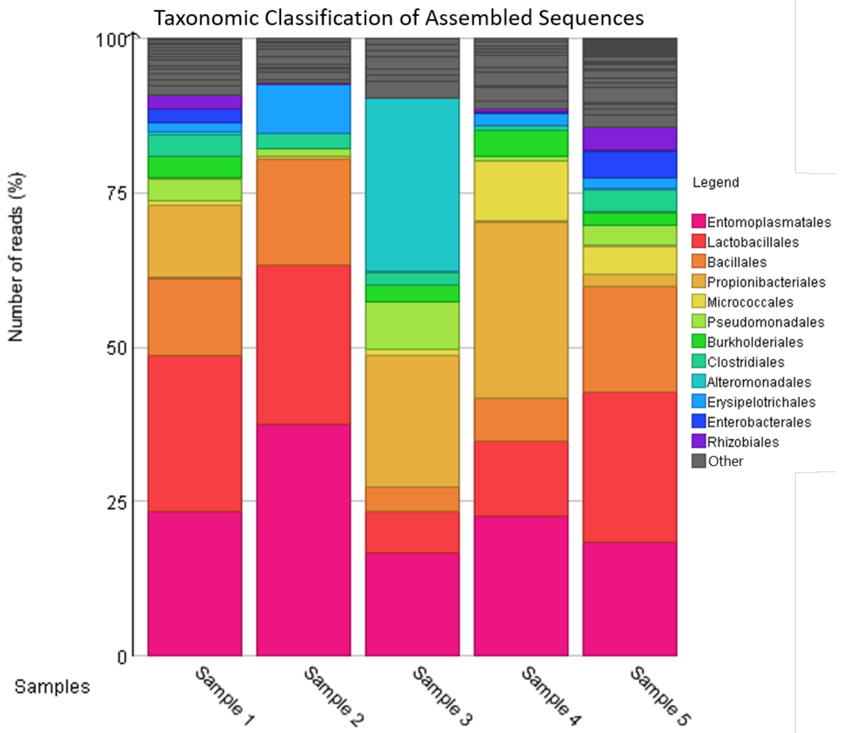

Figure 6. Taxonomic classification of assembled sequences using Diamond and MEGAN6. In this assembled sequences pipeline, IDBA results were run on Diamond followed by MEGAN6 to generate taxonomic classification. There is a high relative abundance of Entomoplasmatales and Lactobacillales in all of the samples.

Two of the most common bacteria in Sample 1 generated by the GC plots were Lactobacillales and Propionibacteria (Figure 5). Similarly, Lactobacillales were abundant in Sample 1 according to MEGAN6 (Figure 6). A similar high abundance of Entomoplasmatales is seen in both the GC plots and MEGAN6 for all 5 samples (Figures 5, 6). Similarities between the two pipelines include the presence of the bacteria Burkholderiales, Propionibacteriales, Enterobacteriales, and Enterobacterales. There are, however, some differences between the pipelines. In sample 3, there is a high abundance of Alteromonadales according to MEGAN6

(Figure 6), but this is not seen in the GC plots. The underlying differences in these results are due to differences between using a BLASTn search and the Diamond search, which does BLASTX.

Unassembled and assembled data can be compared from MEGAN6. In sample 1, a majority of the gut data were Lactobacillales in unassembled data (Figure 3), but in assembled data, there were equally abundant Entomoplasmatales and Lactobacillales (Figure 6). In the unassembled data, samples 2 and 3 look very similar (Figure 3). In the assembled data, there are distinct differences, especially in the large presence of Alteromonadales in sample 3 (Figure 6). Overall, it is important to note that unassembled and assembled data gave vastly different results.

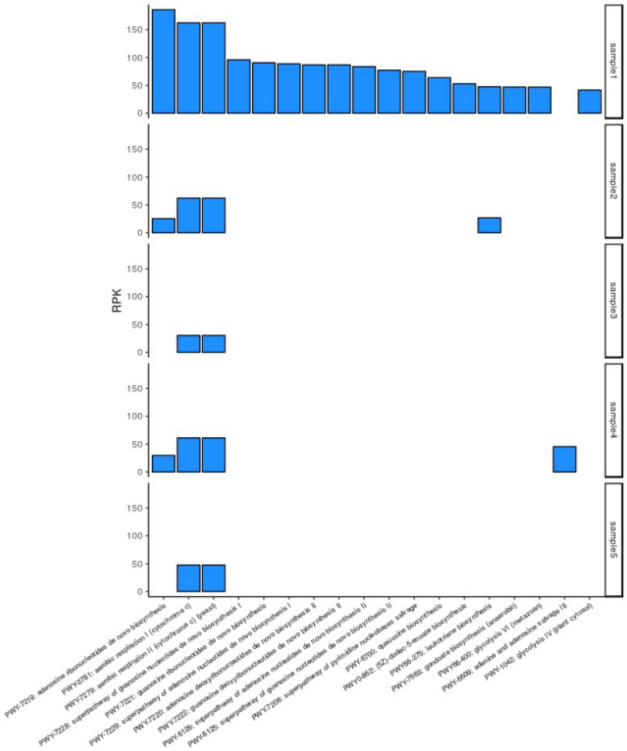

Figure 7. HUMAnN2 analysis of all 5 samples. HUMAnN2 was used for functional annotation. Results showed limited pathways, possibly due to the limited MetaPhlAn2 classifications. Commonly, all samples had PWY-3781: aerobic respiration I (cytochrome c) and PWY-7279: aerobic respiration II (cytochrome c).

Conclusion

Through the use of multiple packages, a comprehensive metagenomics analysis was completed. There was comparison between taxonomic analysis in unassembled and assembled data, and between packages within both. Future work should be done to complete functional annotation, possibly using MEGAN6. There is also an open-ended question of the reasons behind differences bacterial composition between samples, even in the same site but a different colony.

References

[1] Thomas, T., Gilbert, J., & Meyer, F. (2012). “Metagenomics - a guide from sampling to data analysis.” *Microbial Informatics and Experimentation*, *2*, 3.

[2] Hu, Y., Łukasik, P., Moreau, C. S. and Russell, J. A. (2014). “Correlates of gut community composition across an ant species (*Cephalotes varians*) elucidate causes and consequences of symbiotic variability.” *Mol Ecol, 23*: 1284–1300.

[3] Brady, S.G., Fisher, B.L., Schultz, T.R., Ward, P.S. (2014). “The rise of army ants and their relatives: diversification of specialized predatory doryline ants”. *BMC Evolutionary Biology, 14*(93).

[4] Langmead B, Salzberg SL. (2012). “Fast gapped-read alignment with Bowtie 2.” *Nat Methods, 9*: 357-359.

[5] Bolger, A. M., Lohse, M., & Usadel, B. (2014). “Trimmomatic: A flexible trimmer for Illumina Sequence Data.” *Bioinformatics*, btu170.

[6] Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., Segata, N. (2015). “MetaPhlAn2 for enhanced metagenomic taxonomic profiling.” *Nature Methods 12*, 902–903.

[7] Franzosa E.A., McIver L.J., Rahnavard G., Thompson L.R., Schirmer M., Weingart G., Schwarzberg Lipson K., Knight R., Caporaso J.G., Segata N., Huttenhower C. “Functionally profiling metagenomes and metatranscriptomes at species-level resolution.” (Submitted).

[8] Y. Peng, "Google Groups: MiSeq 250bp reads: modifying short_sequence.h,". [Online]. Available: https://groups.google.com/d/msg/hku-idba/gijtWg-I6dM/6Gjubh4-wZIJ. [Accessed 20 03 2017].

[9] Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. (2013). “QUAST: quality assessment tool for genome assemblies”. *Bioinformatics,* 29(8): 1072-1075.

[10] Li H. and Durbin R. (2009). “Fast and accurate short read alignment with Burrows-Wheeler Transform”. *Bioinformatics, 25*:1754-60.

[11] Kumar, S., & Blaxter, M. L. (2011). “Simultaneous genome sequencing of symbionts and their hosts.” *Symbiosis (Philadelphia, Pa.)*, *55*(3), 119–126.

[12] Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. (2008). "BLAST+: architecture and applications." *BMC Bioinformatics 10*:421.

[13] Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., & Blaxter, M. (2013). Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in Genetics*, *4*, 237.

[14] Benjamin Buchfink, Chao Xie & Daniel H. Huson. (2015). “Fast and Sensitive Protein Alignment using DIAMOND.” *Nature Methods, 12*, 59-60.

[15] Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C. (2016). "User Manual for MEGAN V5.11.3."

[16] Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C. (2007). “MEGAN analysis of metagenomic data.” *Genome Research 17*(3): 377-386.

[17] Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L. (2010). “IDBA- A Practical Iterative de Bruijn Graph De Novo Assembler.” In: Berger B. (eds) Research in Computational Molecular Biology. RECOMB 2010. Lecture Notes in Computer Science, vol 6044. Springer, Berlin, Heidelberg

[18] Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L (2012) “IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth”. *Bioinformatics, 28*: 1420-1428.