Metagenomic Analysis of Army Ant Guts

Dhantha Gunarathna

Feiyang Xue

Ariana Entezari

Metagenomic Analysis

- Use bioinformatics tools to understand genetic content of communities of organisms
- Broad spectrum of analysis gives understanding of links between genetic function and evolution for uncultured organisms
- Can give a novel understanding of microbial function

Army Ant Guts

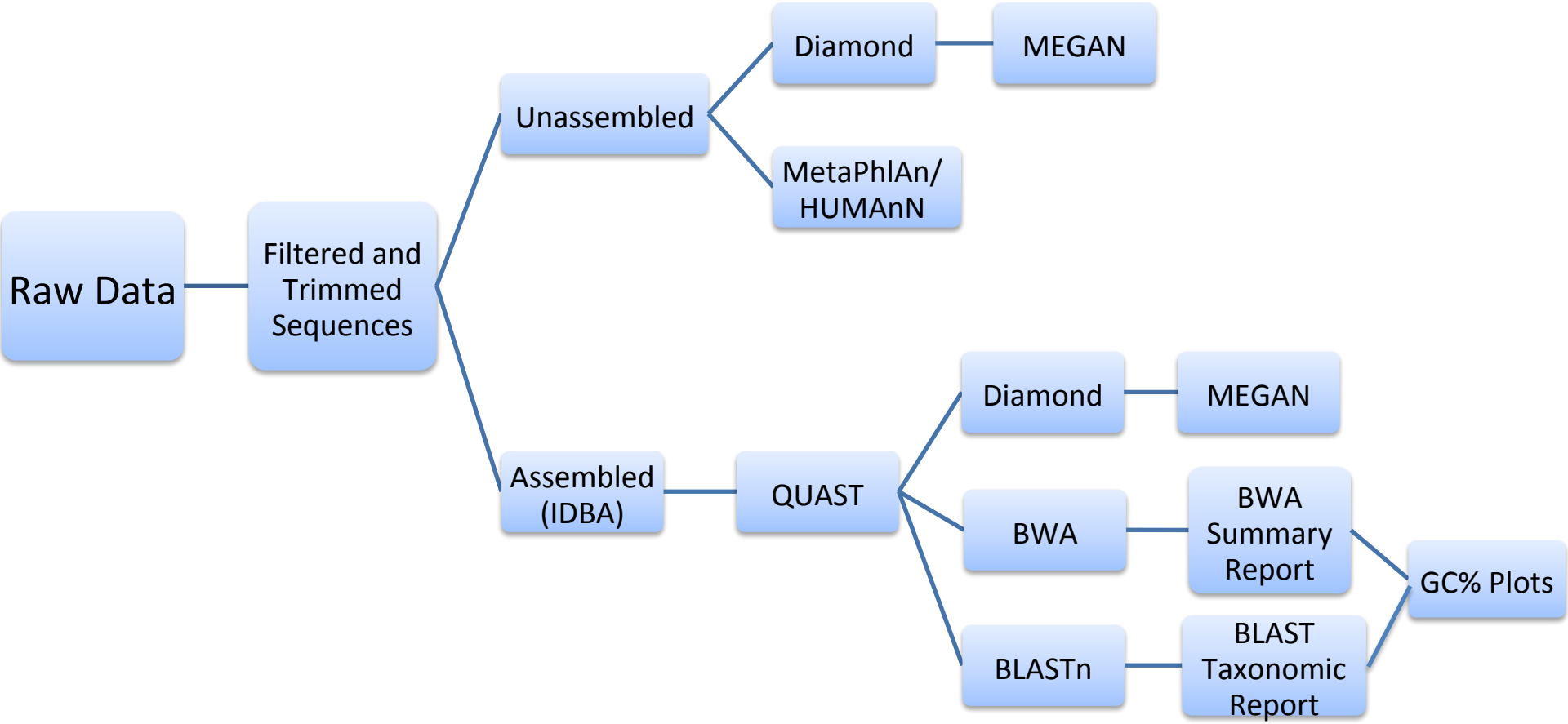
- Ants have a symbiotic relationship with bacteria
  - many of these bacteria are in ant guts
  - ecological and physiological factors are what is dictating the diversity of bacteria
    - varies between species
- Army ants - characterized by nomadism (move colonies), group predation (work together to feed), and are carnivorous

Raw Sequencing Data

- Army ant samples collected in Costa Rica by Russell lab
- Ileums of ants were dissected and DNA was extracted
- Whole genome shotgun samples sent for Illumina sequencing to UC Berkeley
- 10 paired-end sequencing files were received (in fastq format)
  - forward and reverse sequences for the 5 samples

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

Metagenomic Analysis Pipeline



KneadData

- Used to perform quality control on metagenomic sequencing data
- BowTie2-Build
  - *in silico* separation of bacterial reads from contaminant reads
  - Built a custom database of host genomes
    - Human genome and ant genomes from family *Formicidae*
  - Ensured that custom database did not include bacterial data

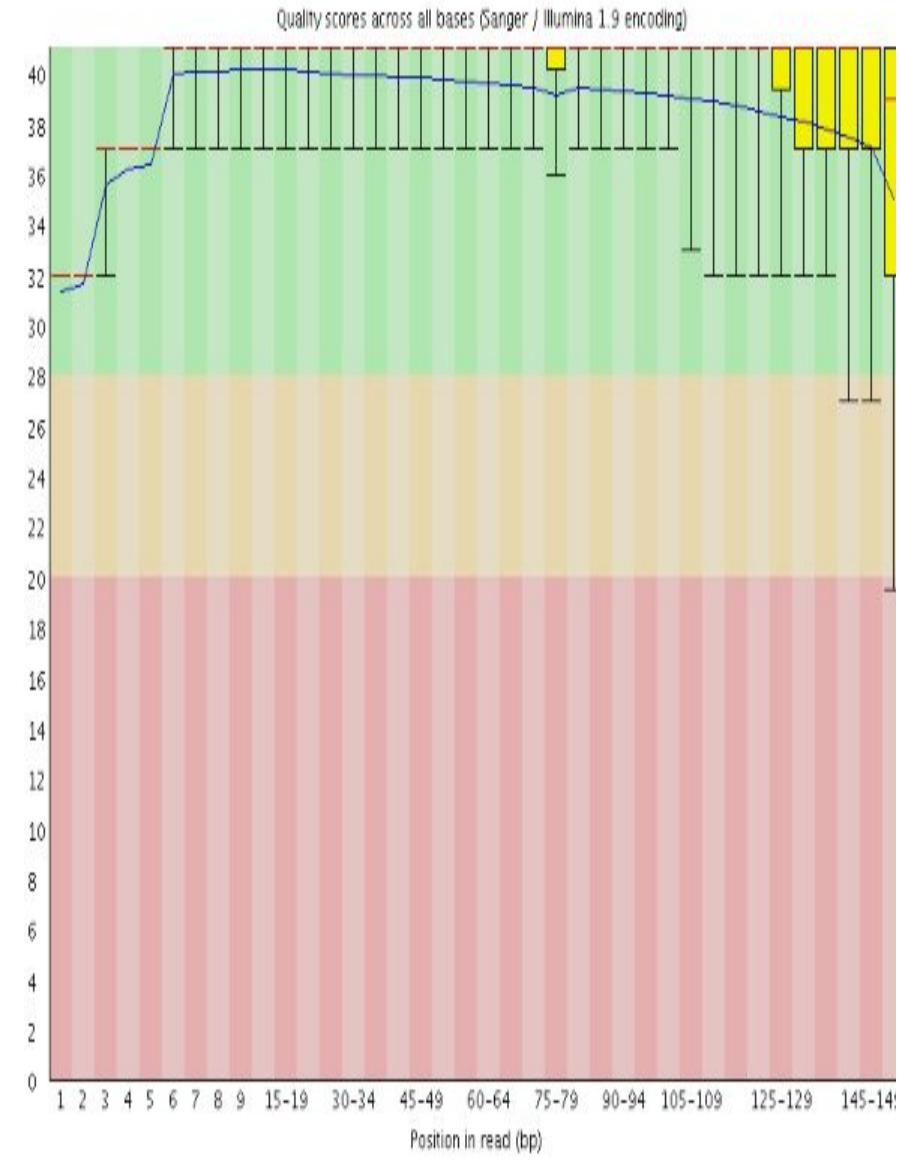
```
bowtie2-build metaphlan2/markers.fasta metaphlan2/db_v21/mpa_v21_m200
```

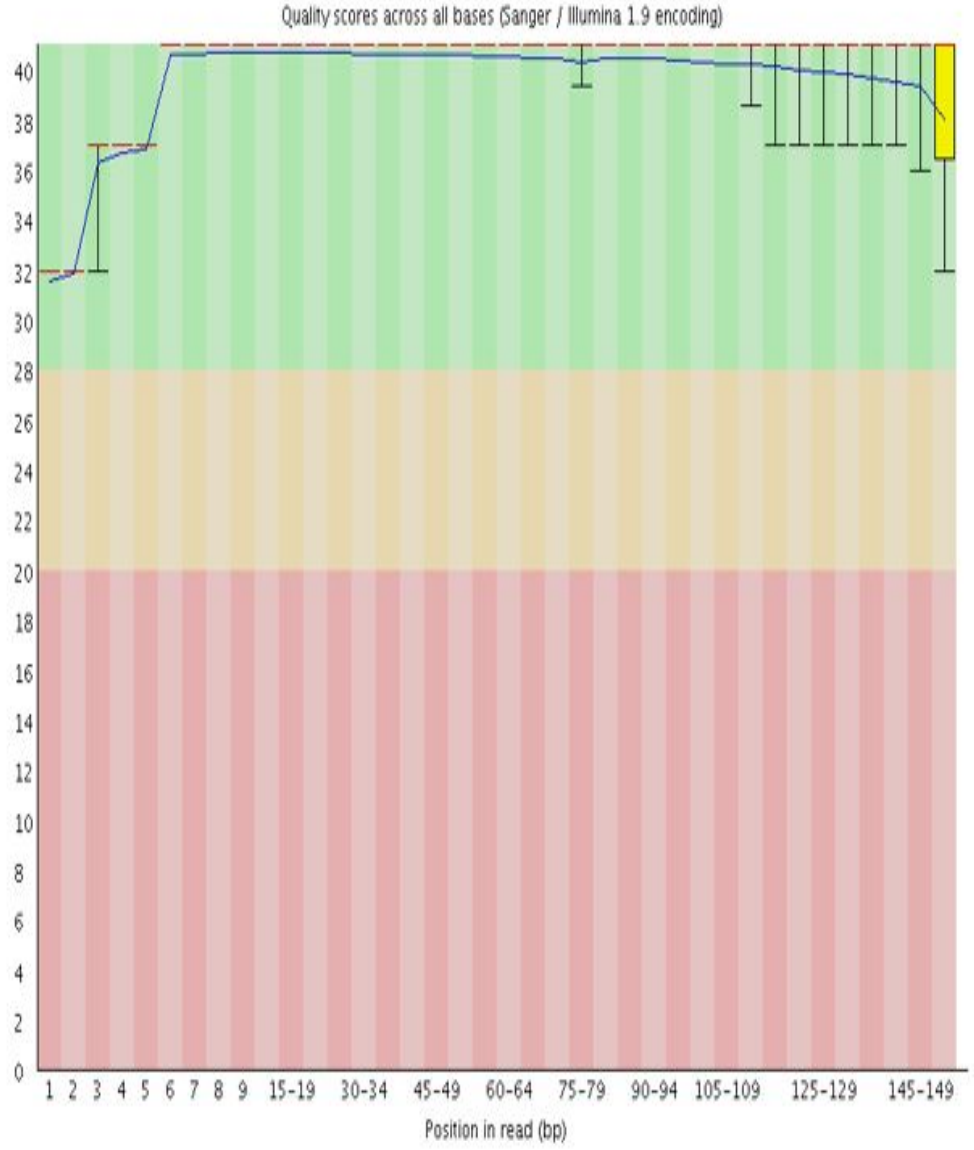
KneadData

- Trimmomatic
  - trim adapter sequences from Illumina
  - remove leading and trailing low quality reads
  - Uses a sliding window approach (4 nucleotides wide) to remove windows that have a quality score below 20
  - Remove reads less than 70 bases long

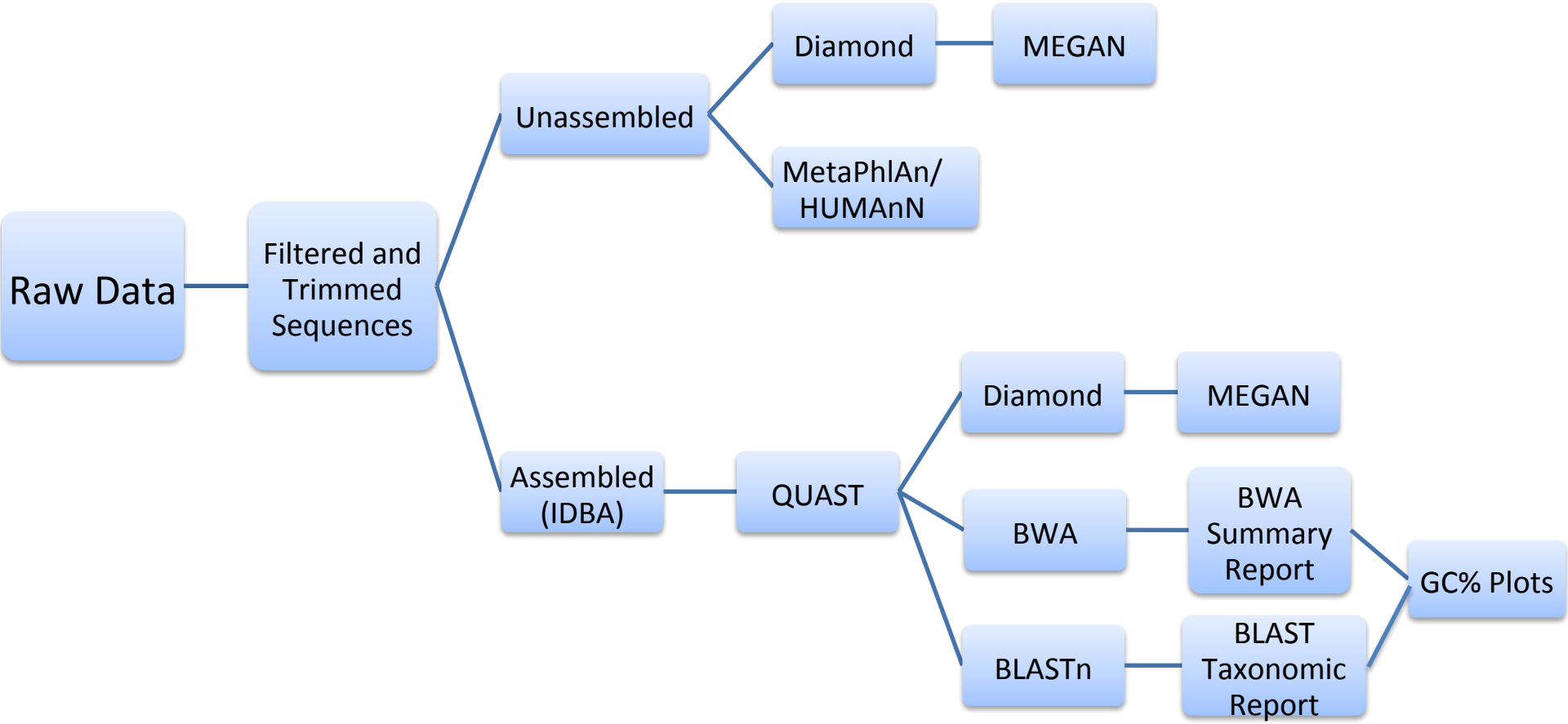
```
kneaddata --input seq1.fastq --input seq2.fastq -db $DATABASE --output kneaddata_output
```

FastQC


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 125-129 145-14
Position in read (bp)


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 125-129 145-149
Position in read (bp)

Metagenomic Analysis Pipeline



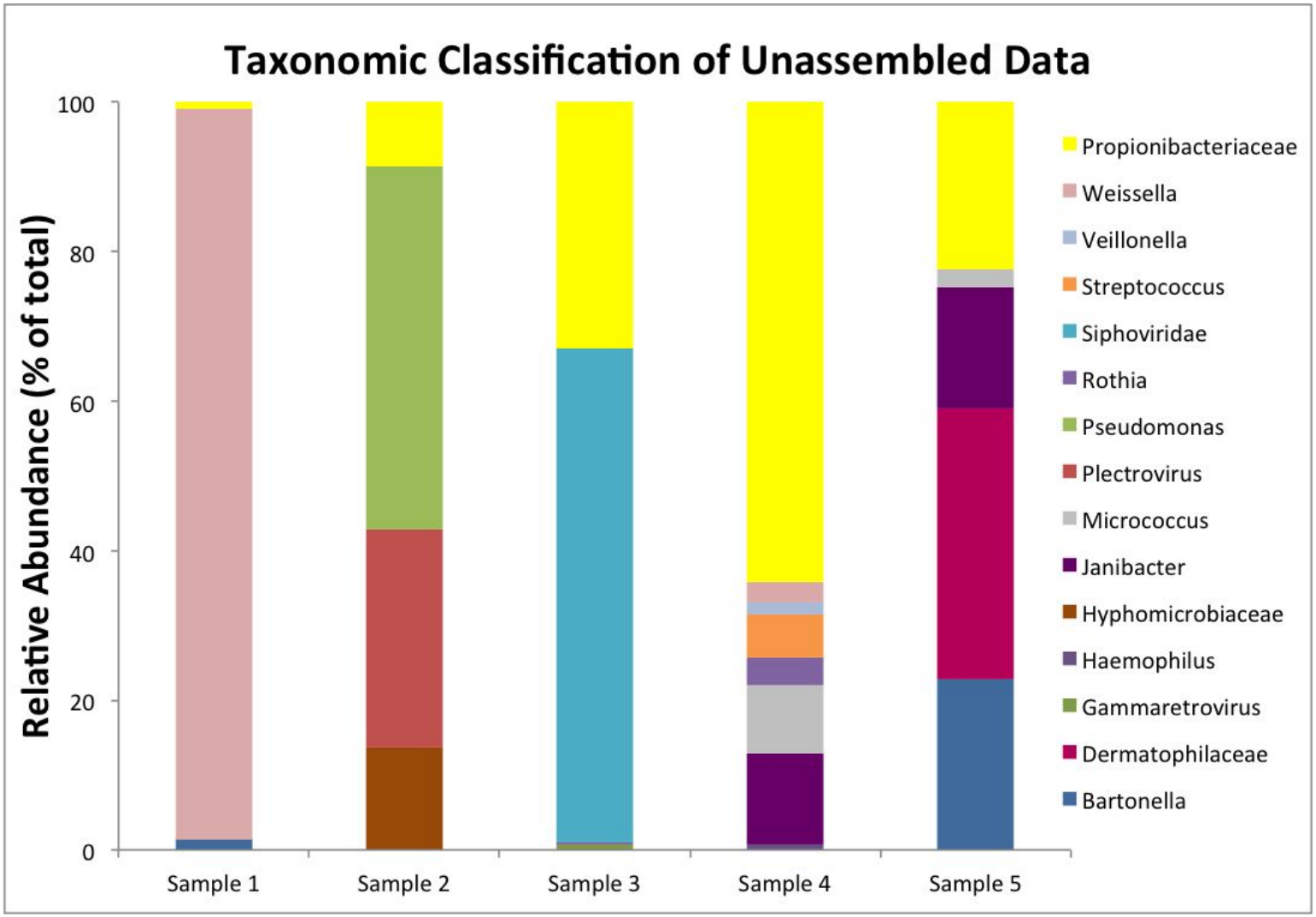
Unassembled Sequences Analysis

MetaPhlAn2

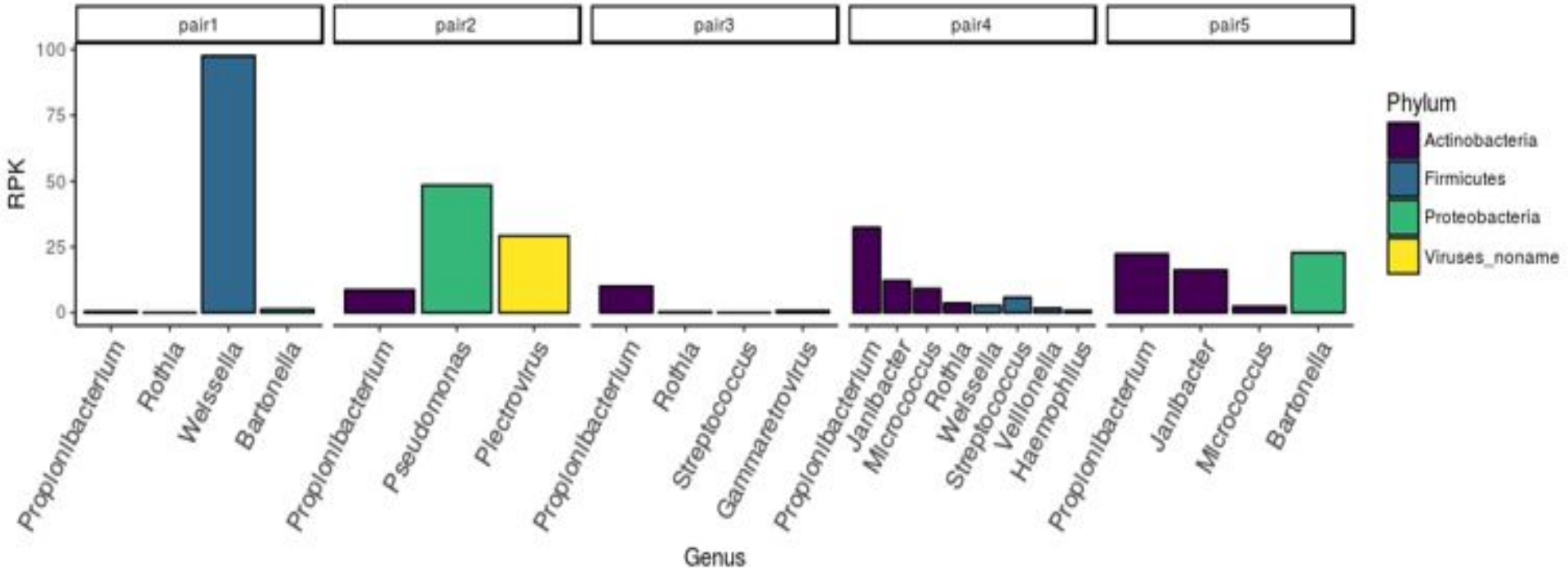
- Align paired-end sequences and profile their relative abundances
- BowTie2: align sequence reads to reference sequences
- Profiles sequences from bacteria, archaea, eukaryotes, and viruses, using around 1 million clade-specific marker genes

```
metaphlan2.py metagenome_1.fastq,metagenome_2.fastq --bowtie2out metagenome.bowtie2.bz2
--nproc 5 --input_type fastq > profiled_metagenome.txt
```

MetaPhlAn2 Taxonomic Classification



MetaPhlAn2 Output



HUMAnN2

- HUMAnN2 is unified metabolic analysis network
- It’s a pipeline for efficiently and accurately profiling the presence/absence of microbial pathways
- This process, referred to as functional profiling, aims to describe the metabolic potential of a microbial community and its members.

HUMAnN2 Outputs

1. **Gene families file**

This file details the abundance of each gene family in the community

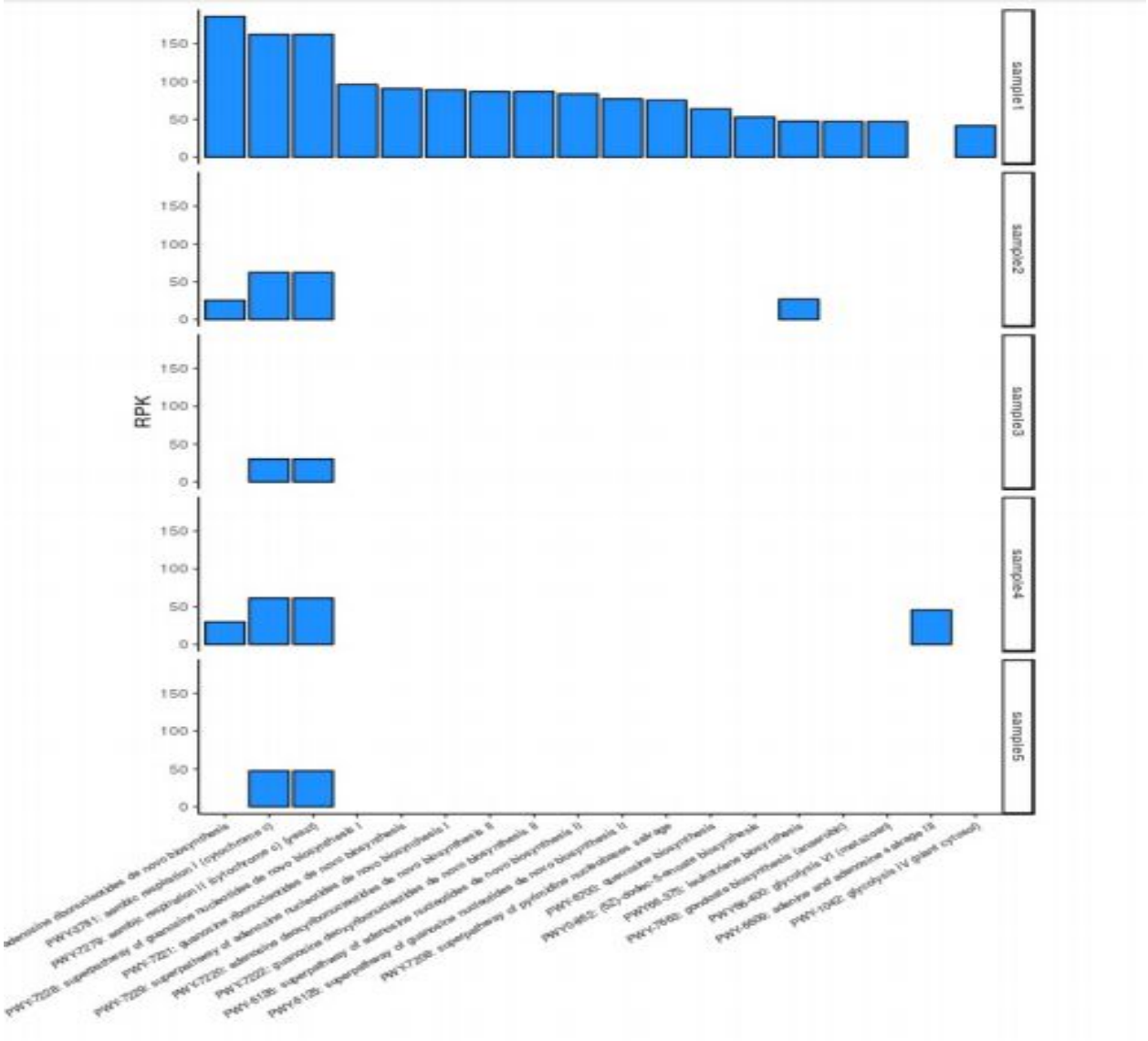
2. **Pathway abundance file**

This file details the abundance of each pathway in the community as a function of the abundances of the pathways component reactions, with each reaction's abundance computed as the sum over abundances of genes catalyzing the reaction.

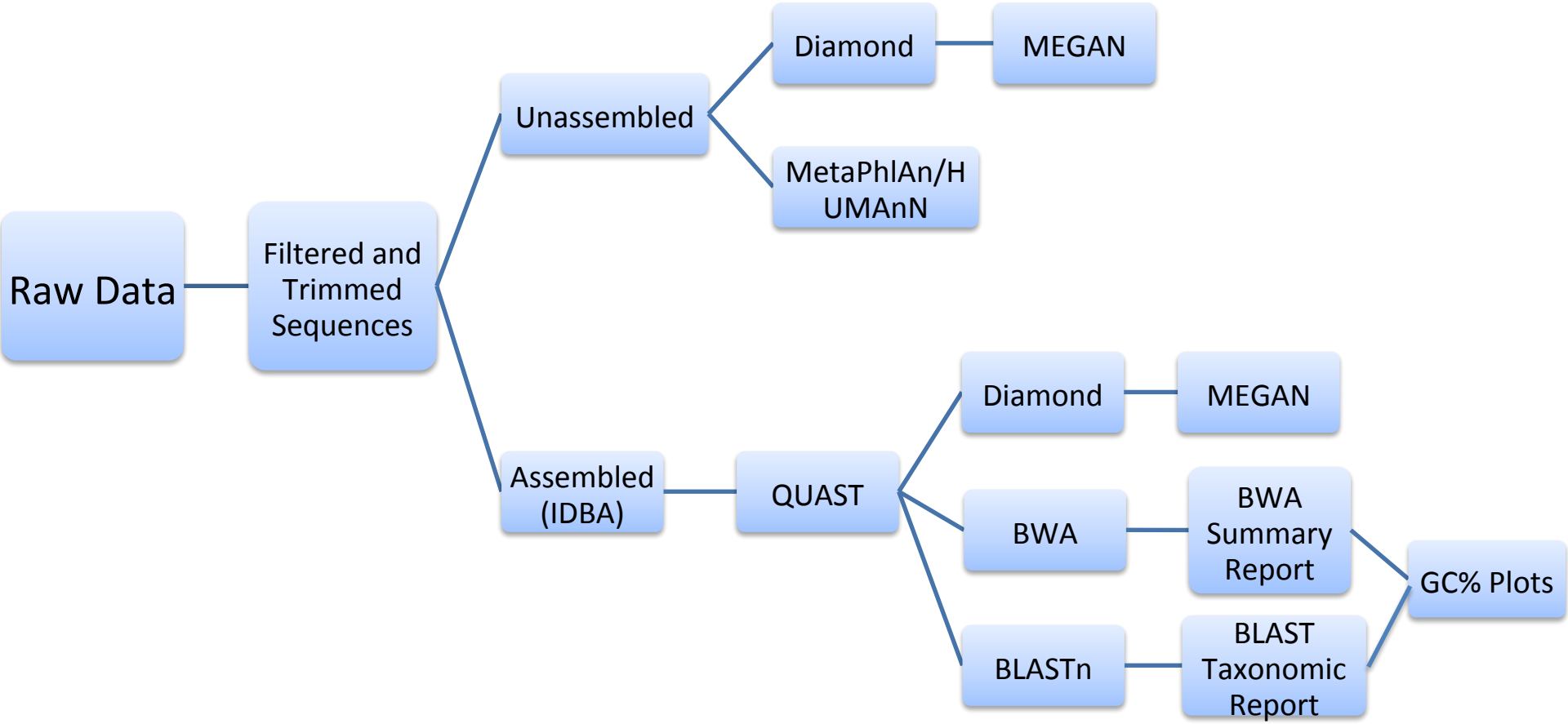
3. **Pathway coverage file**

Pathway coverage provides an alternative description of the presence (1) and absence (0) of pathways in a community, independent of their quantitative abundance.

HUMAnN2 Path Abundance



Metagenomic Analysis Pipeline



Assembled Sequences Pipeline

IDBA

- Assemble paired end sequences (forward and reverse)
  - iterative de Brujn graph assembly
- IDBA modified for longer read length

```
fq2fa --merge path/to/read1.fastq path/to/read2.fastq \
    path/to/output.fasta
idba -r path/to/input.fasta -o /path/to/output/folder/ \
     --num_threads 16
```

QUAST

- Quality assessment of assembly
- n50 - length that the collection of all contigs of that length or longer accounts for at least 50% of the assembly

Sample	n50
1	36261
2	23255
3	16459
4	24970
5	18541

Burrows Wheeler Alignment (BWA)

- Map sequences against a reference genome
- Given the paired-end sequences and the scaffold file, it outputs aligned sequences

```
bwa mem $INPUT $SEQS1 $SEQ2 > $OUT/pair5_samfile
```

Summary Report of BWA

- sam_len_cov_gc_insert.pl used

```
sam_len_cov_gc_insert.pl  -s samfile -f PL005W/scaffold.fa
```

- Generated:
  - Length: length of scaffold
  - Read-depth coverage: number of reads aligning back to reference genome (scaffold file)
  - GC Content: % of scaffold containing GC

BLASTn

- Scaffolds blasted against NCBI non-redundant nucleotide (nt) database to find best species hit
  - megablast option used

```
blastn -task megablast -query scaffold.fa -db nt -max_target_seqs 1 -outfmt 6 -num_threads 4
  > sample5_assembly.megablast.nt
```

BLAST Taxonomic Report

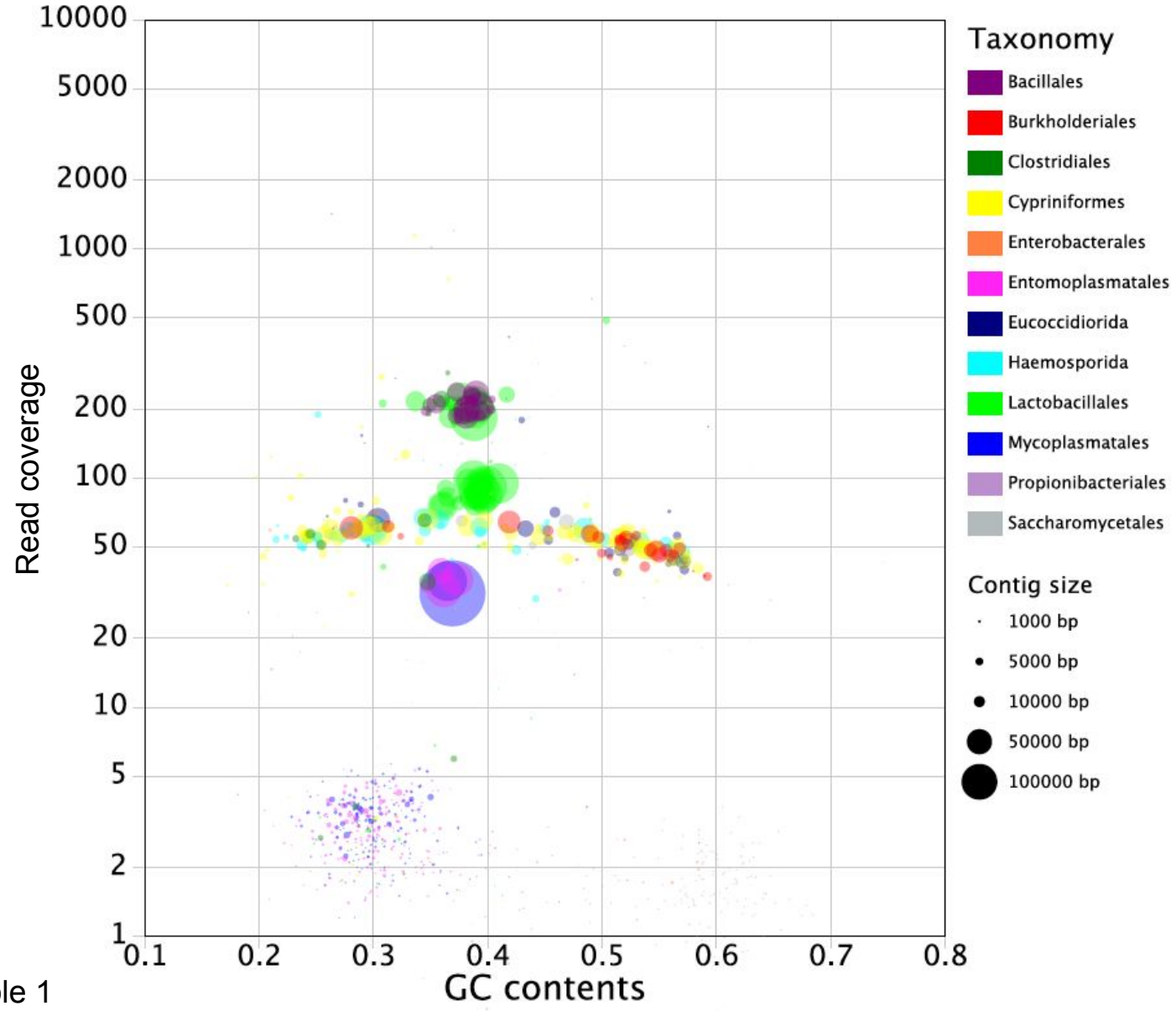
BLASTn only gives taxid - need the actual taxonomic classification

```
blast_taxonomy_report.pl \
-b PL005W/scaffold.fa.megablast.nt \
-nodes tax/nodes.dmp \
-names tax/names.dmp \
-gi_taxid_file tax/gi_taxid_nucl.dmp.gz \
-t genus=1 -t order=1 family=1 -t superfamily=1 -t kingdom=1 \
> PL005W_scaffold_annotate2.megablast.nt.taxon
```

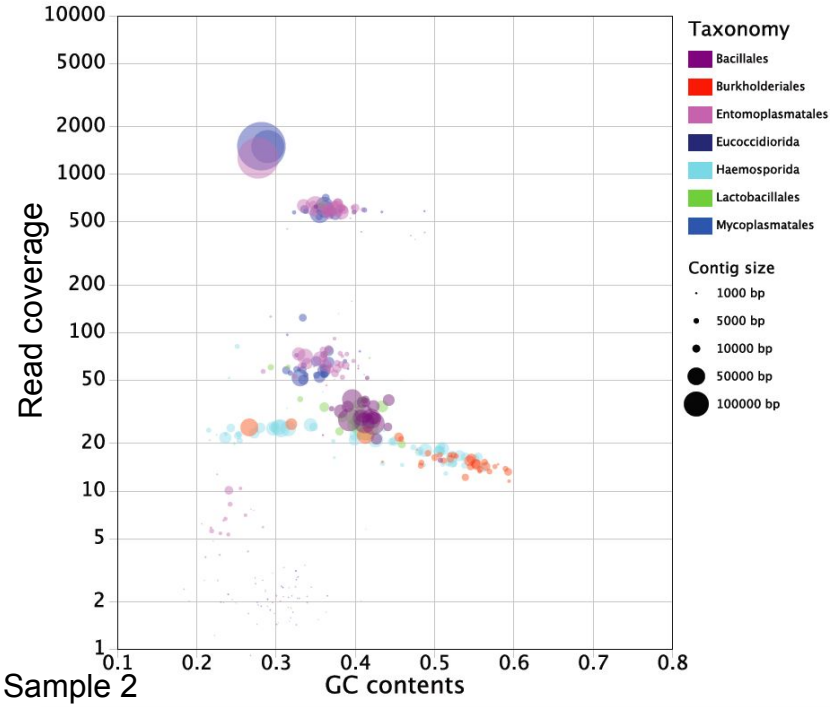
GC Plots

- Used to look at abundance of assembled sequences
- Generated using scaffold length, read-depth coverage, and GC content

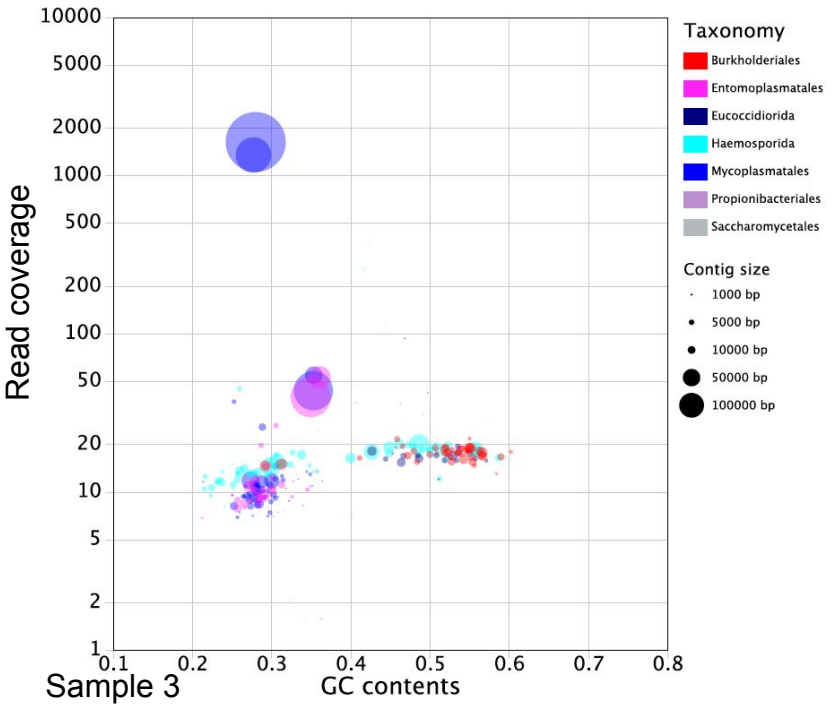
GC Plots



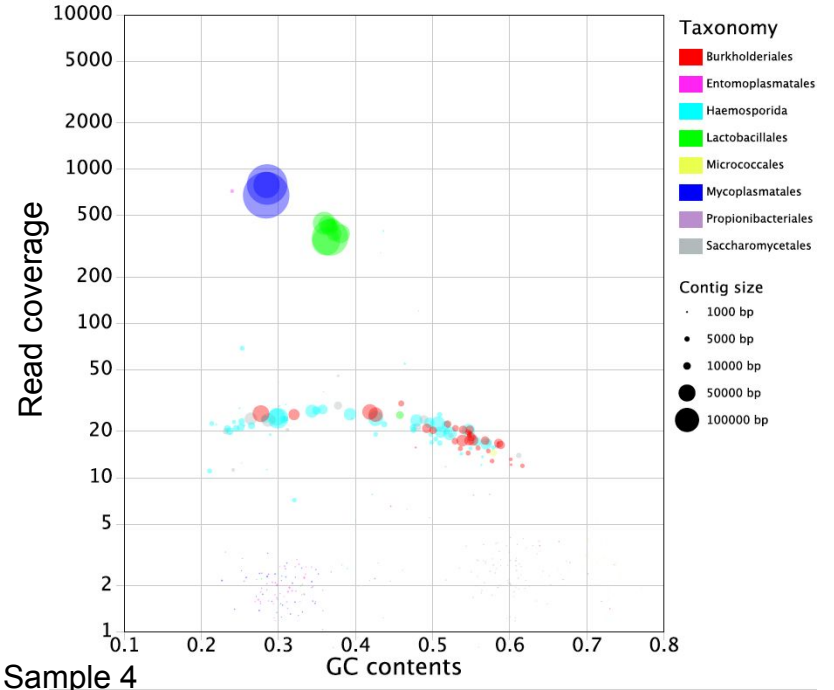
Sample 1



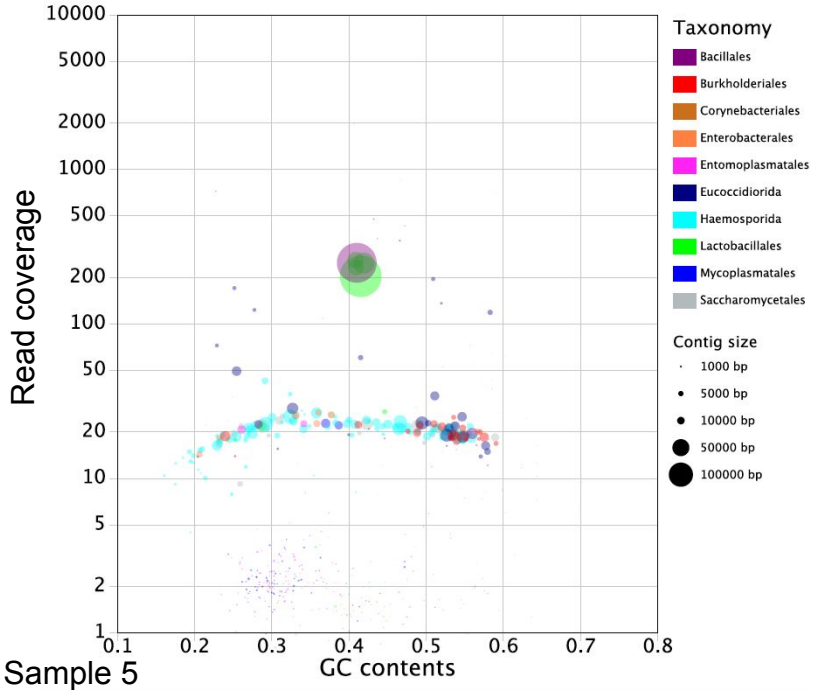
Sample 2



Sample 3

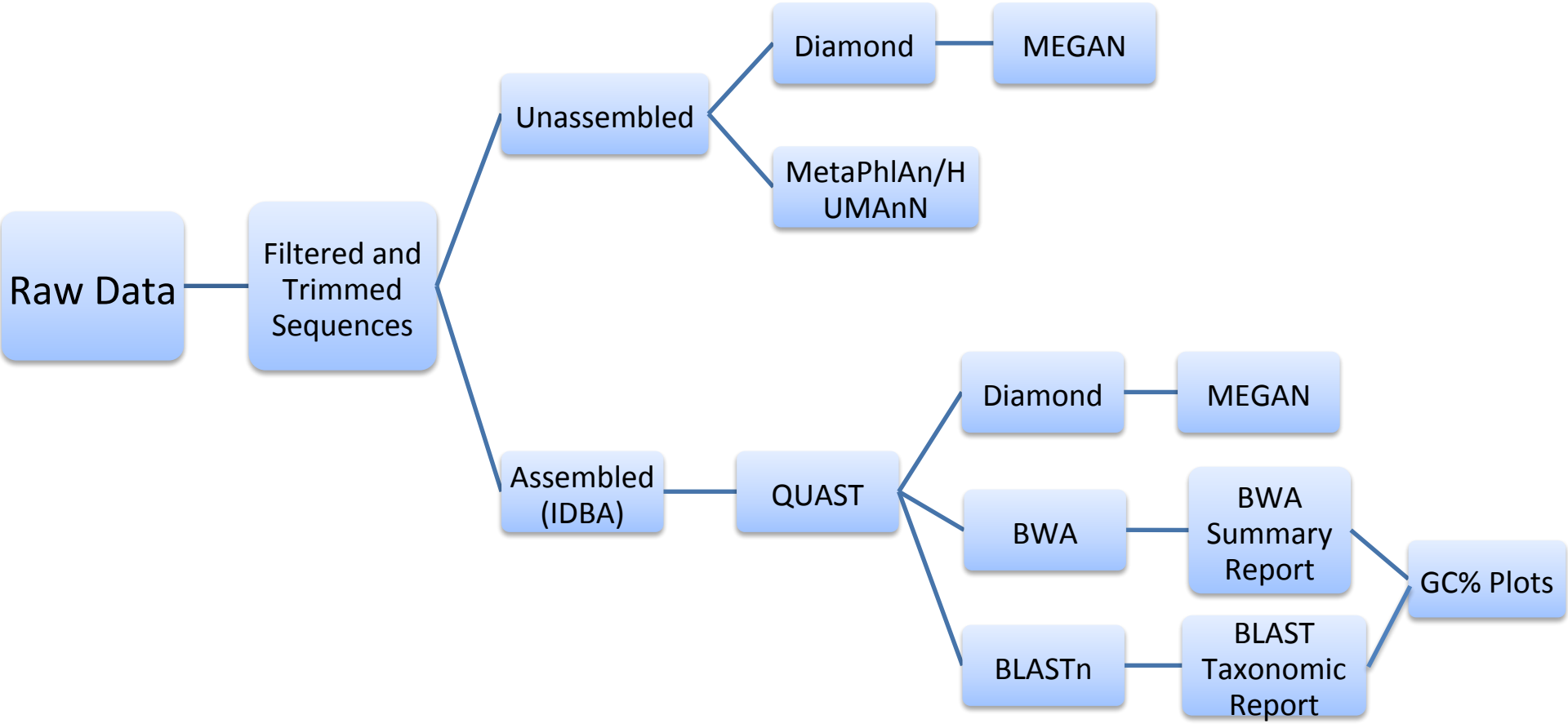


Sample 4



Sample 5

Metagenomic Analysis Pipeline



Diamond + MEGAN (Unassembled and Assembled)

Diamond

- Aligns DNA sequences to a protein reference database (NCBI-NR)
- Up to 20,000 times faster than BLASTX
- Used to align filtered sequences

```
diamond blastx -d /path/to/nr.dmnd \
 -q /path/to/input.fastq -a /path/to/output \
 -v --block-size 18.0 --index-chunks 2
```

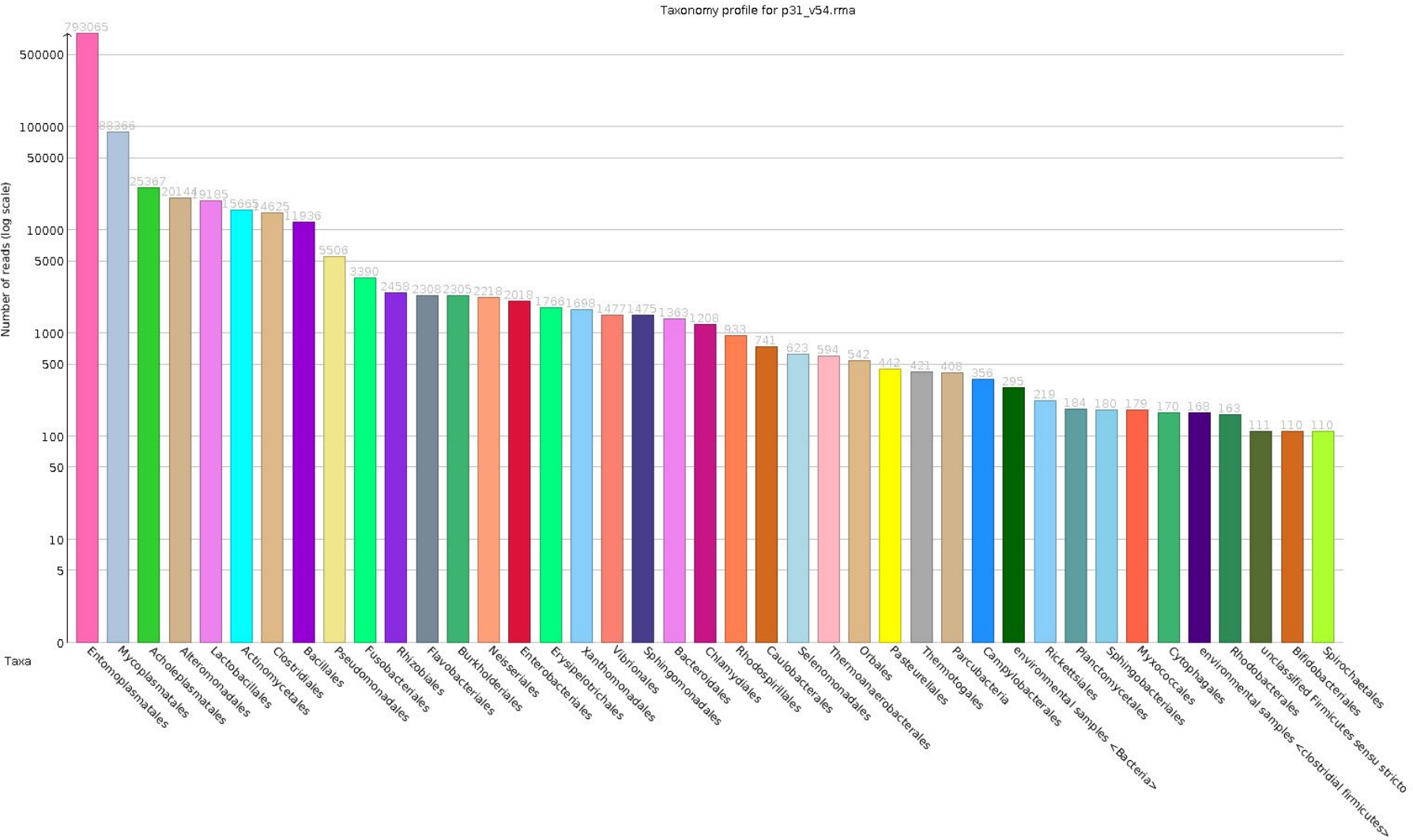
MEGAN

- GUI allows for easy manipulation
- Gives taxonomic and functional analysis
  - Just focused on taxonomic analysis
  - Can also give COG, KEGG, and SEED for functional annotation

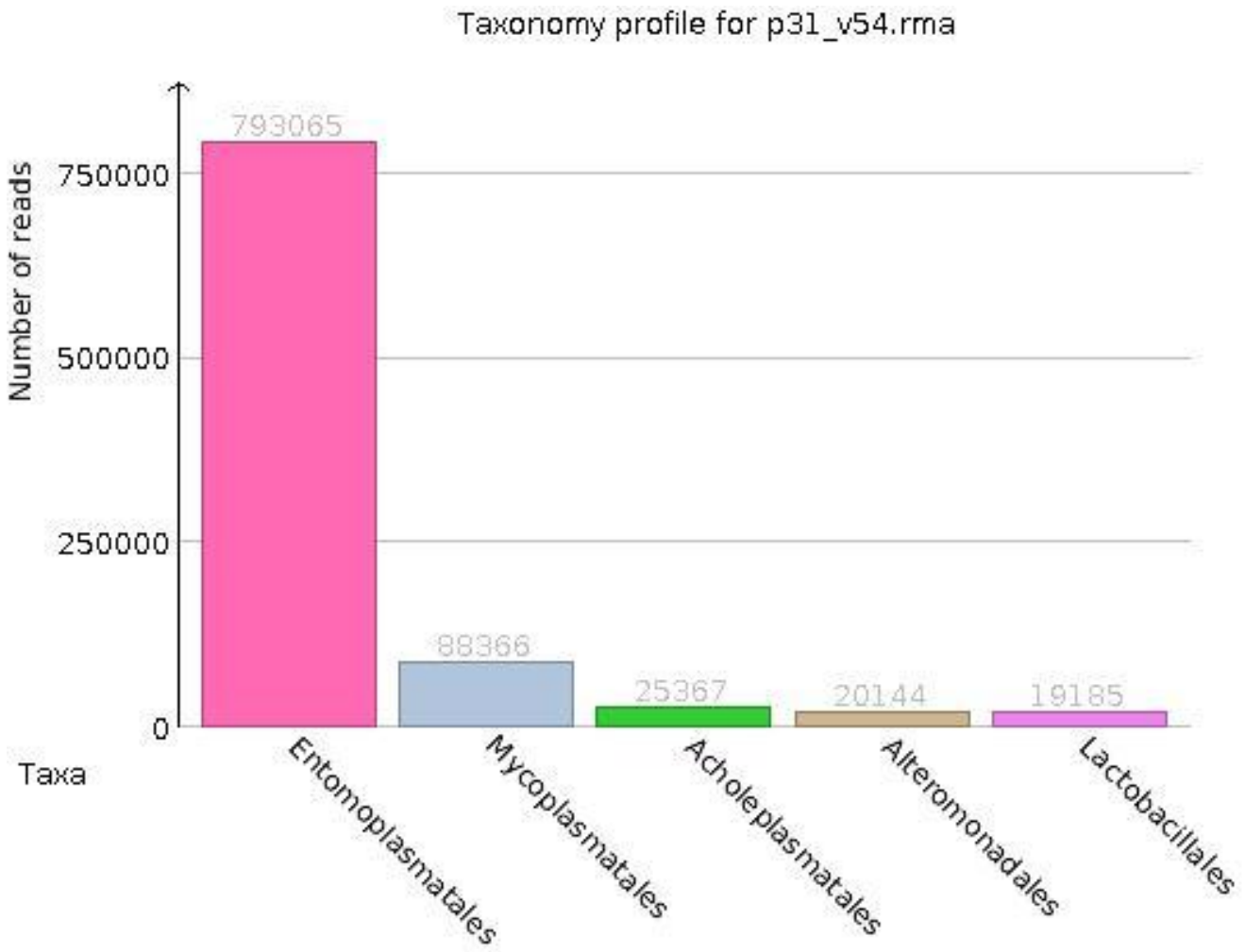
Diamond - MEGAN 5 Workaround

- Diamond marks BLAST results with NCBI Accession Number, whereas MEGAN reads Taxonomy ID
- Lookup from accession to taxon ID is hardware demanding; Proteus server does not support MEGAN to run with GUI
- A script is written for looking up taxon ID and marking diamond output, while taking advantage of Proteus

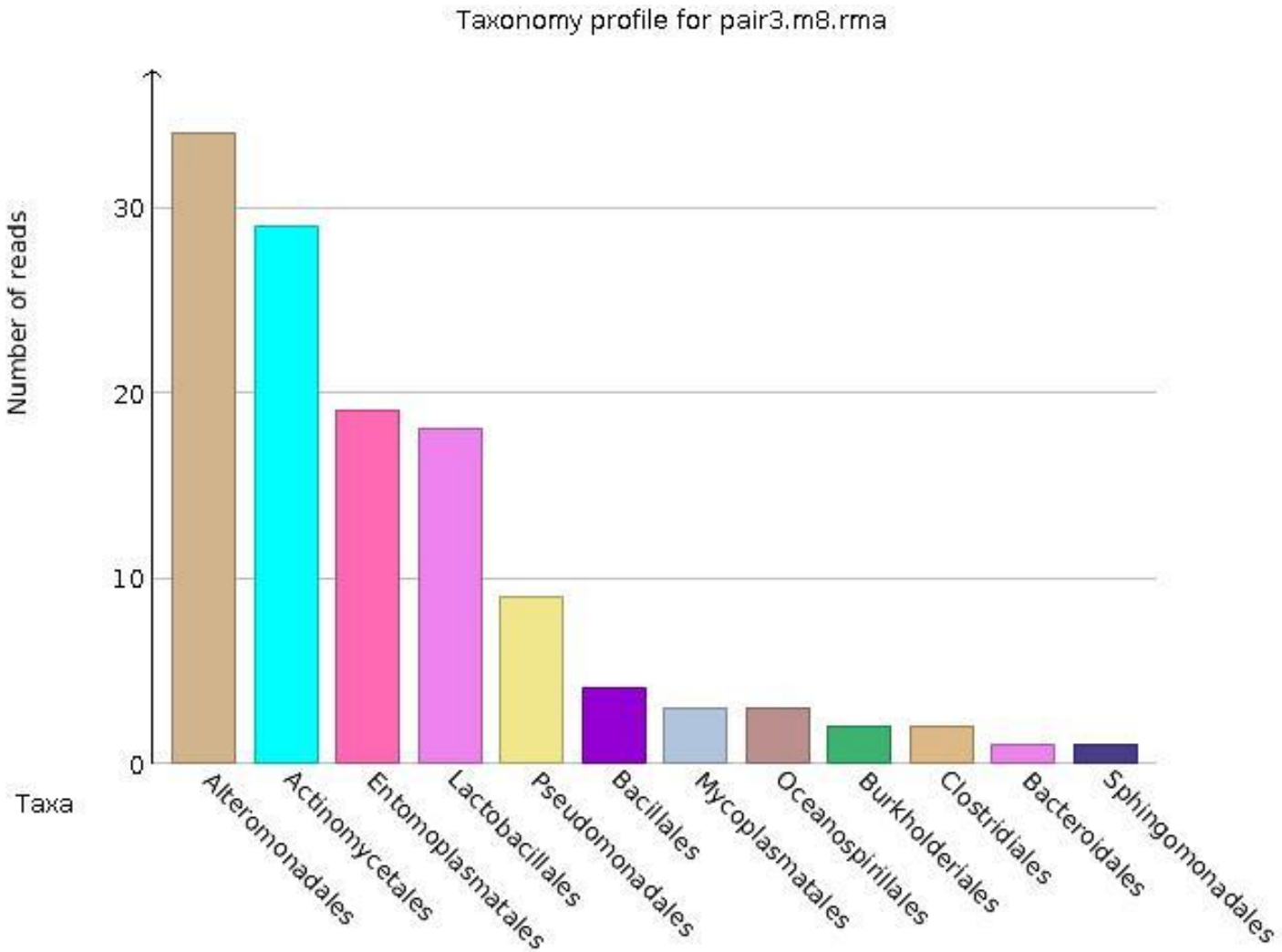
MEGAN Taxonomic Analysis Unassembled



MEGAN Taxonomic Analysis Unassembled for top 5 bacteria



MEGAN Taxonomic Analysis Assembled



Conclusion

- Similar results between assembled taxonomy analyzed using Diamond and MEGAN5 vs BLASTn and BWA
- Unassembled MetaPhlAn2 vs unassembled DIAMOND and MEGAN5 gave vastly different results
  - MetaPhlAn2 is more conservative
- Unassembled taxonomic analysis differs from assembled taxonomic analysis
- Not a wide diversity of bacteria in army ant guts

References

Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methé B, Schloss PD, Gevers D, Mitreva M, Huttenhower C. "Metabolic reconstruction for metagenomic data and its application to the human microbiome." PLoS Comput Biol. 2012 Jun;8(6):e1002358

Benjamin Buchfink, Chao Xie & Daniel H. Huson, Fast and Sensitive Protein Alignment using DIAMOND, Nature Methods, 12, 59–60 (2015) doi:10.1038/nmeth.3176.

Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Sujai Kumar, Martin Jones, Georgios Koutsovoulos, Michael Clarke, Mark Blaxter*

Funaro, C. F., D. J. C. Kronauer, C. S. Moreau, B. Goldman-Huertas, N. E. Pierce, and J. A. Russell. "Army Ants Harbor a Host-Specific Clade of Entomoplasmatales Bacteria." *Applied and Environmental Microbiology* 77.1 (2010): 346-50. Web.

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. http://doi.org/10.3389/fgene.2013.00237

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]

MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower & Nicola Segata*. Nature Methods 12, 902–903 (2015)

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2*, 3. http://doi.org/10.1186/2042-5783-2-3