**Construction of accurate and concise phylogenomic trees from *de novo* metagenome assemblies**

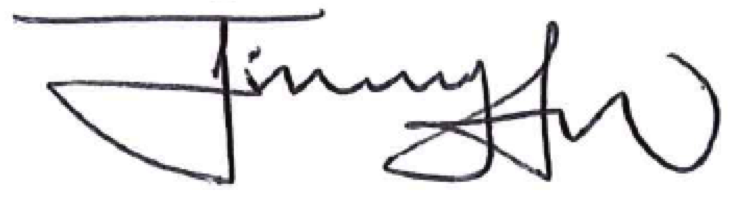
Henry Becker

[hhbecker@gwmail.gwu.edu](mailto:hhbecker@gwmail.gwu.edu)

Thesis submitted for Special Honors in Biology

April 19, 2020

This thesis has been formatted for submission to Frontiers in Microbiology

*signature indicating approval* 

Dr. Jimmy Saw

Assistant Professor

Department of Biological Sciences

George Washington University

[jsaw@gwu.edu](mailto:jsaw@gwu.edu)

\_\_\_\_\_\_\_\_

# Abstract

In order to understand life, microbes must be studied. Microbes were the very first life forms on earth, some 3.8 billion years ago. Microbes contributed to the oxygenation of earth's atmosphere allowing more complex eukaryotic organisms like humans to evolve. Microbes are undoubtedly the most abundant form of life on earth. They are also the most diverse, both metabolically and phylogenetically (Oren, 2004). Additionally, microbes determine the outer boundaries of life sustaining environments on earth. Whether the environmental parameter is temperature, acidity, salinity, altitude, or ocean depth, microbes inhabit the extremes of each parameter (Merino, 2019). It is clear that in order to understand life on earth, we must understand microbes. This manuscript describes a metagenomics research project that can be split into two main parts. Part one is the de novo assembly of metagenome assembled genomes from publicly available metagenomic samples collected at Yellowstone National Park and downloaded from the NCBI Short Read Archive. In the second part of this project, two methods for tree building with metagenome assembled genomes are introduced and compared. This project is motivated by the need for current metagenomic workflows and tree building approaches with metagenome assembled genomes. All software developed for this project will be posted on Github for public use.

Part 1:

A Metagenomics Workflow

# History of DNA sequencing

To understand the significance and limitations of current sequencing capabilities it is important to understand the history of sequencing technology. All genomic research is possible because of sequencing technology while simultaneously being limited by it. DNA sequencing technology (methods to determine the order of base pairs of sampled DNA) has come a long way since the very first base pairs were sequenced in the late 1960's and early 1970's (Shendure, 2019). A continual stream of optimizations and advancements has increased the speed, accuracy, and decreased the cost of DNA sequencing. Sanger sequencing is seen as the first major development in DNA sequencing and is considered "first generation" sequencing technology. Sanger sequencing relies on electrophoresis and a single well within which a single strand of DNA is elongated and observed (Shendure et al., 2019). This single strand electrophoretic method was ultimately responsible for the sequencing of the human genome project. Sanger sequencing is known as "low-throughput" sequencing technology (Slatko et al., 2018).

High-throughput sequencing (also known as massively parallel sequencing or next generation sequencing) revolutionized DNA sequencing technology in the late 90’s and early 2000’s. High-throughput sequencing technology encompasses many different methods and strategies that share a characteristic improvement over Sanger sequencing in their ability to sequence hundreds of millions of DNA strands in one reaction chamber at the same time (making them rapid and cost effective). The ability to sequence many strands in parallel greatly improves the speed and efficiency of sequencing. High throughput sequencing allowed for the creation of much larger genetic datasets which allowed for more comprehensive insights into genomics (Churko et al., 2013).

Third generation sequencing technology is currently available and has a promising future. Third generation technology differs from second generation technology by producing longer read lengths. A read is the inferred DNA sequence that is determined from the fragmented DNA library being sequenced. While Second generation technology produces reads of 100-500 base pairs long, third generation technology (in specific Pacific Biosciences platforms) can produce read lengths of 1500 base pairs or longer. Longer read lengths make de novo assembly of reads more accurate. Though third generation technology is very promising, second generation sequencing methods are responsible for the birth and current growth of microbial metagenomics (Slatko et al., 2018).

# Metagenomic sequence data

A metagenome is defined as the theoretical collection of all genomes from members in a microbial community from a specific environment (Escobar-Zepeda et al., 2015). A metagenome theoretically contains all genetic material from a given environmental sample. Metagenomics refers to the application of sequencing and computational techniques to analyse the totality of metagenome samples. Importantly, environmental samples were difficult to sequence until recently. DNA sequencing traditionally required culture dependent library preparation. Microbiologists would attempt to culture organisms from the environment on agar plates. This method was restrictive because almost 99% of bacteria living in a given natural environment cannot be cultured with traditional methods Staley and Konopka, “Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats.”. Culture dependent sequencing does not provide an accurate snapshot of microbial life present in the environment because the microbes that grow on conventional media are often not the most abundant in natural habitats. Culture dependent methods of DNA library preparation greatly restricted our view of the full scope of natural microbial diversity.

The first methods used to access uncultivated microbes involved the cloning of environmental DNA in bacterial hosts using recombinant DNA technology. This was followed by the direct random shotgun shotgun sequencing of environmental DNA. DNA library preparation that does not require cloning vectors or bacterial hosts simplifies the library preparation and reduces DNA contamination from other organisms that are not part of the metagenome.” (Thomas et al., 2012; Wang et al., 2000; Gillespie et al., 2002). If an environment can be sampled and the DNA of microbes in the sample can be separated and purified without needing to culture any of the organisms in a lab, much more of the existing microbial diversity is reachable. In summary, the combination of culture independent DNA library preparation and high throughput sequencing gave microbiologists access to previously unreachable genetic material which spawned the field of metagenomics (Vaz-Moreira et al., 2011; Felczykowska et al., 2015).

Use of metagenomic samples can be broken into two main categories: metataxonomics and metagenomics. This project focuses on metagenomics but an understanding of metataxonomics is included for clarification. (Breitwieser et al., 2017).

## Metataxonomics

Metataxonomics (also called metaprofiling, metabarcoding, or marker gene sequencing,) is the targeted sequencing of specific marker genes in order to provide a snapshot of the diversity present in a sample. Ribosomal RNA genes are highly conserved across many taxa making them a useful reference point from which to classify and differentiate microbial species. Metataxonomics targets the rRNA genes using amplicon sequencing. Specifically, the 16S (prokaryotic) or 18S (eukaryotic) rRNA gene or of the ITS internal transcribed spacer (the DNA sequence between the small and large ribosomal subunit sequences) of fungi are targeted. Metataxonomics provides rRNA based phylogenies of the diversity in a sample. This method is useful in detecting the presence of certain species but is limited in use otherwise. Because there is no gene conserved across prokaryotes and eukaryotes, both domains cannot be targeted in the same sample (prokaryotes are often chosen over eukaryotes). Additionally, the functional capabilities of the detected organisms is not known if only the rRNA gene is sequenced (Piwosz et al., 2020). Because viruses rely on host ribosomes and do not contain rRNA coding regions, viruses cannot be identified with this method. Despite the limitations of targeting rRNA genes, public databases have far more species coverage of rRNA sequences than complete genomes. Metataxonomic data is also cheaper to sequence and requires less computational resources to analyze.

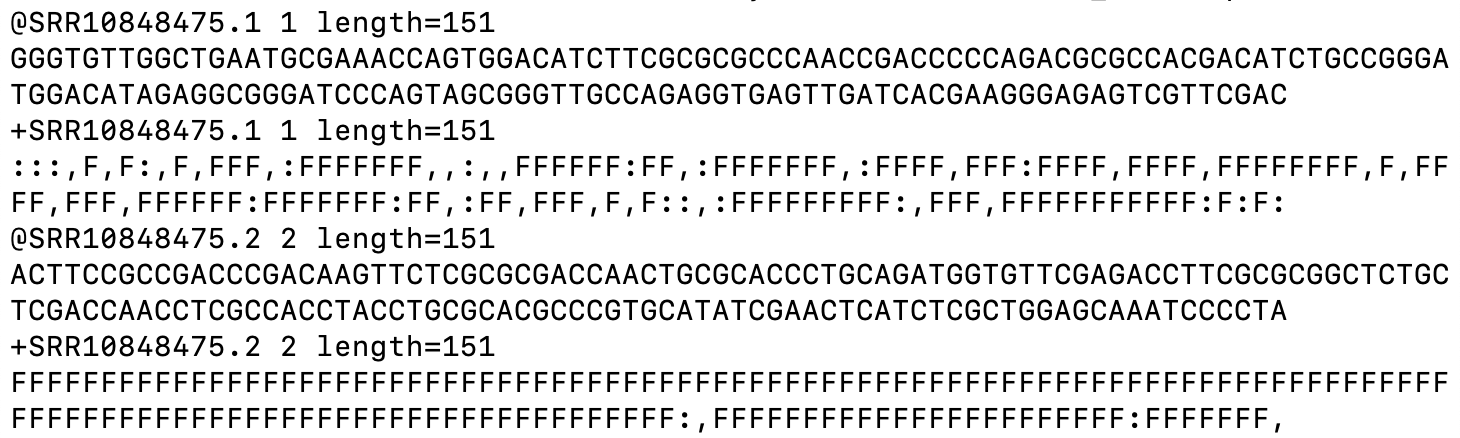
## Metagenomics

While metataxonomics selects and amplifies certain target genes for sequencing, metagenomics attempts to sequence all genetic material present in a prepared sample. This sequencing technique is known as Whole Metagenome Shotgun sequencing (WMGS). This method can theoretically detect and classify bacteria, archaea, viruses, and eukaryotes in the same sample which metataxonomics is unable to do. Additionally, draft genomes can be reassembled from WMGS data. With large portions of genomes reassembled, it is possible to assess the metabolic capabilities of organisms in the original sample. One downside to WMGS sequencing is the massive amount of data produced. The more data produced, the more computational resources and memory required to process and store that data. Metagenomics is revolutionary in that it provides access to the genetic material of all members of diverse microbial communities at high resolution without culturing or targeting of specific gene regions (Laudadio et al., 2019). Metagenomics is capable of answering the two key questions microbial ecologists tend to ask: “Who is out there?” and “What are they doing?” (Laudadio et al., 2019).

# Metagenomic sample collection and sequencing

The environmental DNA in a metagenomic sample should be representative of all organisms present. Therefore, sensitive sampling strategies must be strictly followed to obtain quality samples (Felczykowska et al., 2015). The basic steps of WMGS sample preparation include cell separation through centrifugation, cell lysis, DNA separation and fragmentation. The DNA is then tagged with adapters to prepare it for sequencing. Different sample environments (water vs soil) require different methodologies but storage time should be kept at a minimum across sample types. The specifics of different sample methodologies are outside the scope of this paper but it should be noted that sample processing significantly affects the sequencing quality and depth and overall usefulness of the data. Using data without knowledge of the specific sample processing methods is subject to scrutiny. (Felczykowska et al., 2015)

As stated previously, whole metagenome shotgun sequencing is non targeted sequencing of all DNA in an environmental sample. Illumina next generation technology is widely popular for metagenomic sequencing. The Illumina platform determines the order of nitrogenous bases for millions of DNA strands as they are elongated with fluorescently tagged bases inside the machine. This information is reported in text strings called a read. WMGS produces hundreds of millions of reads in paired files. Reads for Illumina WMGS sequencing range from 150 to 250 base pairs which equates to hundreds of millions of text strings composed of 150 to 250 As,Ts, Cs, and Gs. The files are known as read pairs and represent each end of each DNA strand that was sequenced (Escobar-Zepeda et al. 2015).



[Figure 1] This image shows the FASTQ DNA sequence file format. As you can see, the file stores a string of letters representing the bases of DNA. Underneath each sequence is a quality score. Each base is assigned a quality score as it is sequenced by the illumina machine. Two sequence records are shown in this figure.

# How is genetic data stored?

There are several common formats for storing sequence data. FASTA, FASTQ, and SAM are the most widely used formats. Each format has its advantages and all three are usually used at different points of high throughput sequencing analysis. All three formats are text based ways to store sequence data and metadata.

As more and more sequencing data was produced, more storage was needed. The International Nucleotide Sequence Database Collaboration (INSDC) is an international public repository of sequencing data that began in the 1980's. The INSDC coordinates and data sharing between the National Center for Biotechnology Information (NCBI), the DNA Databank of Japan (DDBJ), and the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) (Cochrane 2011).

The National Center for Biotechnology Information (NCBI) is a US government funded institute under the purview of the National Institutes of Health. NCBI has separate databases for each type of sequencing data. The NCBI Short Read Archive (SRA) stores raw unprocessed short reads from environmental and metagenomic samples from all branches of life. The NCBI SRA is the largest publicly available repository for data of this kind.

DNA short read datasets submitted to the Short Read Archive are organized in a data container identified by an SRA accession number. The SRA accession number is not a data object itself but rather a common identifier (virtual container) that groups together 5 independent data objects of different types. Underneath each SRA reference number is an SRP, SRX, SRR, SRS, and SRZ accession number.

The Joint Genome Institute (JGI) is responsible for many sequencing projects. Typically, JGI provides researchers with a grace period before JGI sequenced data becomes public. JGI data is then automatically piped to the corresponding NCBI databases for public access. JGI recently changed their policy regarding data transfer to NCBI after a data leak was uncovered. Interestingly, the data used in this project may be part of this data leak. A letter from the Joint Genome Institute director explaining the data leak and policy change is included in the supplementary material.

# Octopus and Mushroom Springs short read data

The sequence data used in this project were produced by the US Department of Energy Joint Genome Institute (JGI) http://www.jgi.doe.gov/ in collaboration with the user community. The data was downloaded from the NCBI SRA but was provided by JGI meaning the data is subject to the JGI data utilization policies. This data is not able to be published without consent from the principal investigator: Dr. Devaki Bhaya.

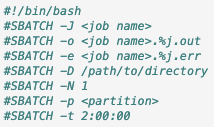
Roughly 500 metagenomic samples were sequenced in November, 2018 as part of a project investigating Cyanophages and Cyanobacteria in the microbial mat communities of two hot springs in Yellowstone National Park. This data is publicly available but is not included in any publication to date. Two of these samples were selected and downloaded for use in this metagenomics workflow. The first sample corresponds to NCBI accession SRR10848475 and was collected at Octopus Spring, Yellowstone National Park. The second sample corresponds to NCBI accession SRR8863378 and was collected at Mushroom Spring, Yellowstone National Park. Both samples were sequenced using the Illumina NovaSeq platform.

These two springs and the thick mats of microbial life that exist around them have been the subject of study for almost 50 years. Most research interest has focused on the diversity and metabolic function of the phototrophic community members including Cyanobacteria as was the case with the project responsible for the sequencing of the samples used in this project (Thiel et al., 2016)

# High Performance Computing

Next generations sequencing projects are able to produce very large genetic datasets. As datasets become larger, more robust computational resources are required to gain insights from them (Escobar-Zepeda et al., 2015). The data downloaded for use in this project (short read files) was over 60 gigabytes per sample. The amount of data and the computational requirements of the tools of each of these steps overwhelm a conventional laptop computer. Pegasus, GW’s High Performance Computing cluster was used to complete this project.

Computing power comes down to the processor (a small chip that performs basic operations based on the instructions it is fed). The laws of physics place certain constraints on the improvements that can be made of individual processors (the speed of light and the smallest possible wire diameter for example). A computer architecture strategy to overcome the limitations of individual processors is to connect multiple processors together so the load of instructions is shared between them. This is known as parallel computing. While most laptop computers today contain multi core processors, a computing cluster is a group of powerful multiprocessor “nodes” that can be recruited to accomplish tasks faster than any individual laptop computer could. Pegasus consists of 210 of these computing nodes connected via a high speed network. Pegasus uses the SLURM resource manager which is an environment and setup for organizing and prioritizing all of the jobs that different users submit to the cluster.



[Figure 2] This image shows the syntax of an SBATCH script. This SBATCH syntax is used to provide the SLURM resource manager necessary information for job scheduling and node allocation.

-J = job name that will appear in queue

-o = name of output file

-e = name of error file

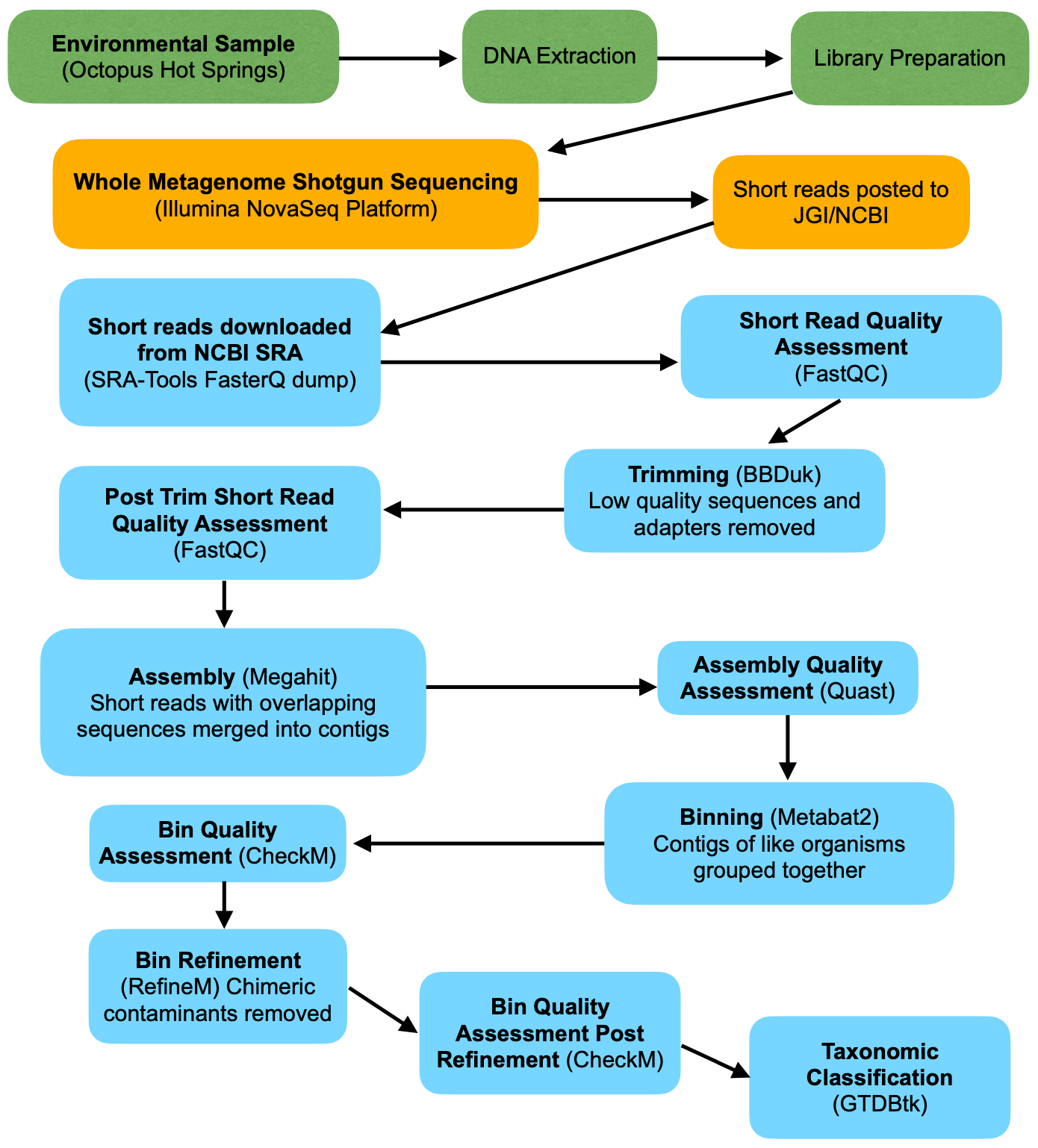
-D = path to the relevant directory within the Pegasus file system

-N = the number of nodes requested

-p = the partition (type of node) requested (defq, HighMem, debug)

-t = the maximum run time allowed before the job is cancelled

# Metagenomics Workflow Overview



[Figure 3] This figure provides a useful overview of the metagenomics workflow that will be explained below. The environmental sample is prepared for sequencing, the sequencer produces short reads, and the analyses begin. First the quality of short reads is assessed and low quality sequences are removed. Next, the short sequences are “assembled” into longer sequences called contigs. Contigs predicted to belong to the same species are then grouped or “binned” together. Lastly, the completeness and purity of the bins is assessed and contaminants are removed.

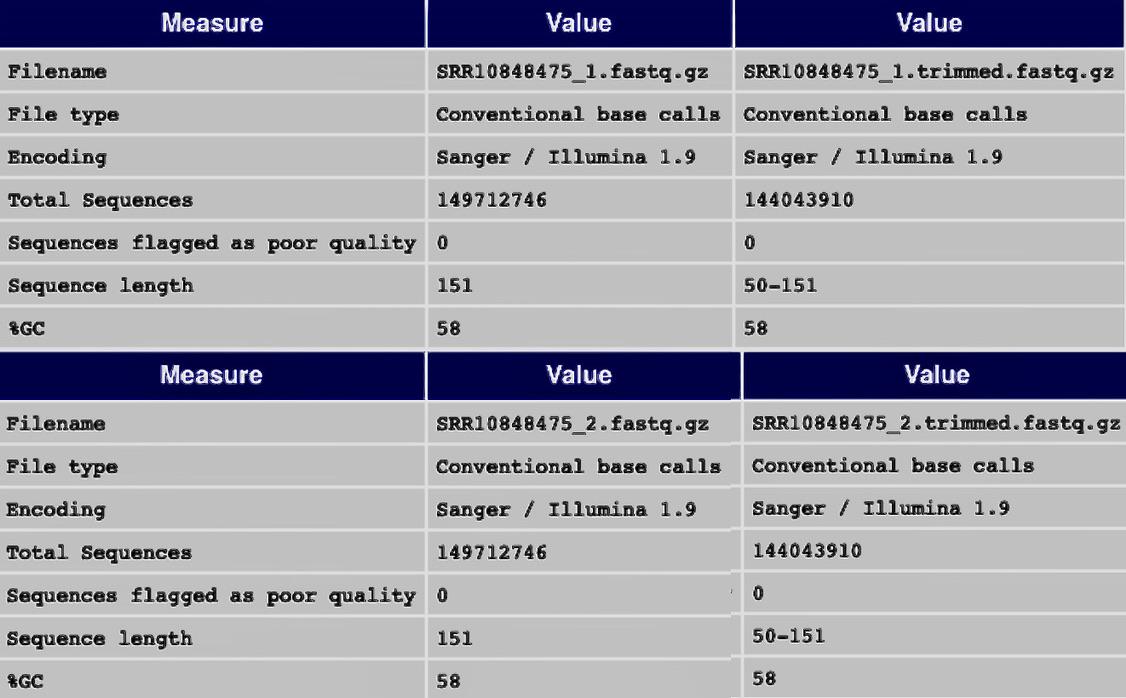
Although metagenomics requires expensive computational resources, all of the software used in this project is free and open source. Each step of the workflow contains an explanation of the reasoning for that step, the general approach tools use to accomplish the step, and information on the specific tool used. The results of the step are posted and analyzed for the first sample (SRR10848475). The SBATCH scripts written for each step can be found in the supplementary materials.

# Quality Assessment and Trimming

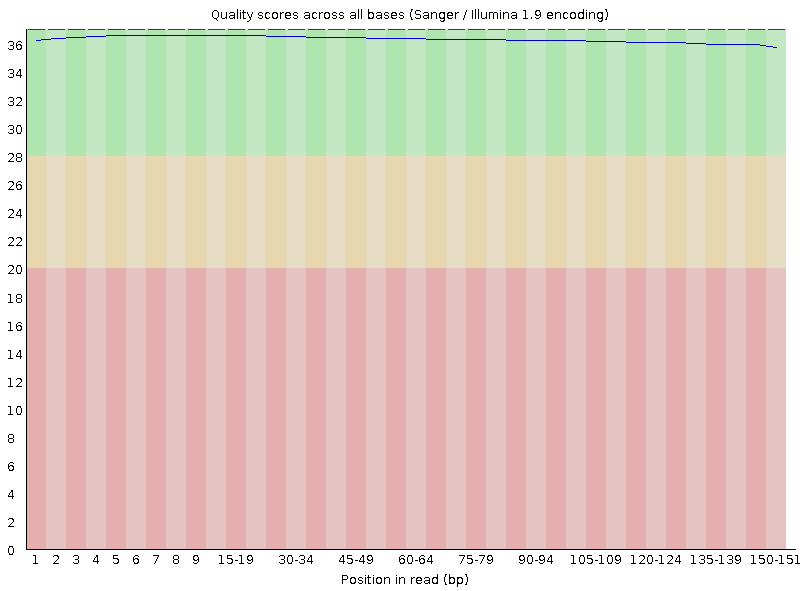
A common first step in any metagenomic analysis workflow is to assess the quality of the data. As this data was retrieved from a public repository there is no guarantee it will be of suitable quality for further analysis. During the sequencing process, bases are assigned a quality score which is included in the FASTQ file format. There are a variety of tools that will provide quality summaries using the FASTQ quality scores. FastQC from the Babraham Bioinformatics institute is a widely used and effective tool. This tool can be found at the Babraham Bioinformatics Institute webpage. Once an initial quality assessment has been run the data is passed through a trimming tool that will remove sequences and bases of low quality. Trimming tools can also remove adapter contaminants. Adapters are short segments of DNA sequencing platforms use in the sequencing process. Illumina NGS platforms use adapters to connect the sample DNA onto the flow cell plates to prepare the DNA for bridge amplification and then sequencing. These adapter sequences are not part of the actual metagenomic sample and should be removed before the workflow proceeds (Pérez et al., 2020).

BBTools is a collection of tools written by scientists at the Joint Genome Institute (JGI). Within the BBTools suite is a tool called BBDuk, which is used to filter and trim adapter sequences from the short reads. BBDuk uses a k-mer approach to identify adapters (Duk stands for Decontamination Using K-mers). A K-mer is a sequence of ‘k’ base pairs in length. The tool searches the short reads against k-mers made from a database of adapter sequences. The k-mer length can be specified and intuitively, longer the k-mer lengths increase the specificity of the search for adapters. Trimming tools can also be used to remove DNA of known eukaryotic origin if the target is prokaryotic microbial communities (a human microbiome sample for example). Quality assessment is usually carried out by a tool like FastQC before and after trimming to determine how much trimming improved the sequence quality. If adapter sequences and other contaminants are present in the original short reads and are removed by the trimming tool there should be a clear increase in quality from the pre to post trimming quality assessment.

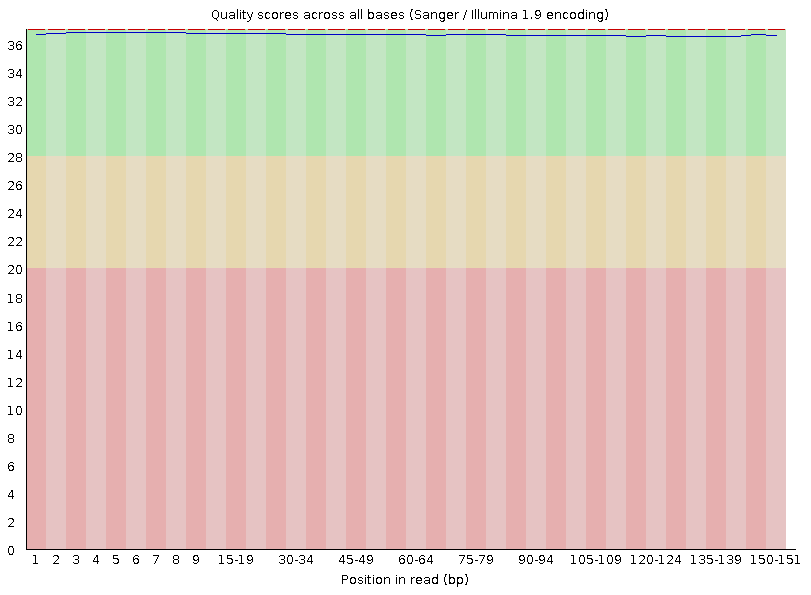
## Quality Assessment and Trimming Results



[Figure 4] This table shows the FastQC results before and after trimming for both read pairs of the first sample (SRR10848475). The leftmost “Value” columns show the sequence information before trimming and the rightmost “Value” columns show the sequence information after trimming. Notice zero sequences were flagged as poor quality before trimming. This signifies the reduction in total sequences is due to the trimming tool removing adapters. The trimming process removed almost 6 million reads and decreased the length of reads but 144 million reads remain.



[Figure 5] This bar graph is produced by the FastQC tool and represents the average sequence quality as the position in the read increases (before trimming (SRR 10848475)). Average sequence quality starts to drop as read position increases. This is due to the nature of Illumina sequencing technology. Fluorescent molecules are tagged onto bases as a way to distinguish which bases are being added by the polymerase. The fluorescence signal fades towards the later cycles of sequencing. As the signal fades it becomes more difficult to confidently assign correct bases and therefore the instrument records lower read qualities.



[Figure 6] This is the post trimming graph of average sequence quality as position along read increases (SRR10848475). The line is slightly straighter signifying the trimming tool shortened certain lower quality reads.

The NCBI SRA contains mostly quality checked sequences and therefore trimming is not always necessary. In this case, because the short reads were already of good quality, trimming was beneficial but not necessary. FastQC and BBDuk SBATCH scripts are included in the supplementary material.

# Metagenome Assembly

Assembly is a key part of this workflow. All subsequent steps, results, and analyses depend on the assembly. (Laudadio et al, 2019) As previously discussed, at this point in the workflow we have paired files of ‘short reads,’ (roughly 150 base pairs in length) that have been checked for quality and removed of contaminants. These short reads represent millions of fragments of what were originally complete organism genomes in the environmental sample.

The goal of metagenome assembly is to piece together short reads to create longer fragments called contigs. If the assembly is performed correctly, the contigs are longer pieces of the genomes that were originally sampled from the environment before they were fragmented and sequenced. Assembly is a valuable step because longer sequence fragments are easier to group into species bins, easier to annotate for gene functions, and easier to search and match against sequence databases. Essentially, assembly improves the effectiveness of all downstream analyses. (Escobar-Zepeda et al., 2015)

The problem of sequence assembly is similar to piecing a book back together after its pages have been shredded and the shredded pieces have been mixed together. This is a fascinatingly difficult problem to solve. There are two main strategies commonly used to solve the assembly problem. The first is comparative assembly which uses a reference genome as a backbone and roadmap for reconstruction of a similar genome that exists within the short reads. The second strategy is known as *de novo* assembly. *De novo* assembly does not use a reference genome and relies on the information that can be gathered statistically from the short reads themselves.

Piecing together DNA fragments is a difficult problem to solve whether dealing with DNA from a single species or from a metagenomic sample containing thousands of species. However, assembling genomes from a single species is considered less difficult because there is inevitably more sequencing depth and more even sequencing coverage than there typically would be when sequencing many species. During single genome sequencing, the same region of a gene is sequenced many times starting and ending at slightly different places. The assembler uses the overlap of the unique pieces to recreate the original sequence. Metagenomic short reads typically have shallow depth and uneven coverage. If a genome region does not have multiple unique fragments for the assembler to compare, it becomes much more difficult to recreate that genome region. Uneven depth and coverage makes repeats and highly similar sequences notoriously difficult for assemblers to recreate from metagenomic samples (Breitwieser et al., 2017).

Despite these challenges, metagenomic assembly often is able to recreate quality contigs which makes binning, annotation, taxonomic assignment, and all other downstream analyses more accurate and effective (Breitwieser et al., 2017).

Advancements in sequencing and computational techniques to overcome the problem of uneven coverage have increasingly improved our ability to assemble genomes from more and more diverse microbial samples. (Parks et al., 2017; Laudadio et al., 2019)

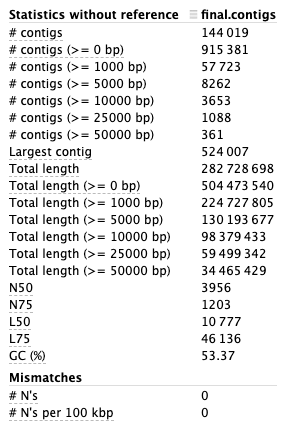
Metagenome assembly tools employ cutting edge algorithms relating to graph theory (A branch of discrete mathematics with many applications in computer science). The most used computational paradigms for genome assembly are the “greedy” approach, overlap– layout–consensus (OLC), or De Bruijn graph methods (Pérez et al., 2020).

The tool used in this workflow is called Megahit (Li et al., 2016). Megahit uses the most popular DeBruijin graph assembly strategy. The de Brujin method uses de Brujin graph representations of k-mers derived from the short reads. As previously mentioned, k-mers are sequence fragments of length k. The algorithm traverses the graphs to efficiently search for k-mers that are shared among many reads in order to find areas of overlap (Pérez et al., 2020). The megahit algorithm attempts to find the best assembly possibility by iteratively testing multiple k-mer lengths.

A complete understanding of the megahit algorithms used is outside the scope of this project. It is important to note the algorithms used are critical to the legitimacy and value of the results. For that reason, algorithm development for metagenomic assembly is an area of ongoing research. To read more about de Bruijin graph assemblers and the Megahit algorithm read these papers (Li et al., 2016)(Compeau et al., 2011).

## Assembly Results

As noted by the creators of the Quast tool, no genome assembly strategy is perfect. The Quast tool provides a quality assessment of the assembly results so that researchers can see the effects of parameter changes. (Gurevich et al., 2013).



[Figure 7] These Quast results show the distribution of contigs by length for the first run (SRR10848475). Longer contigs are more valuable but much less common because of the difficulties of assembly discussed above. If the results are poor, the short reads can be reassembled with different parameters. This assembly produced roughly 140 thousand contigs from 140 million short reads. The short reads were roughly 150 base pairs in length while the contigs reach 500 thousand base pairs in length. Clearly, a genome fragment 500 thousand base pairs long is much easier to analyze than any number of 150 base pair fragments. Herein lies the value of assembly. SBATCH scripts for both Megahit and Quast are included in the supplementary materials.

# Metagenomic Binning

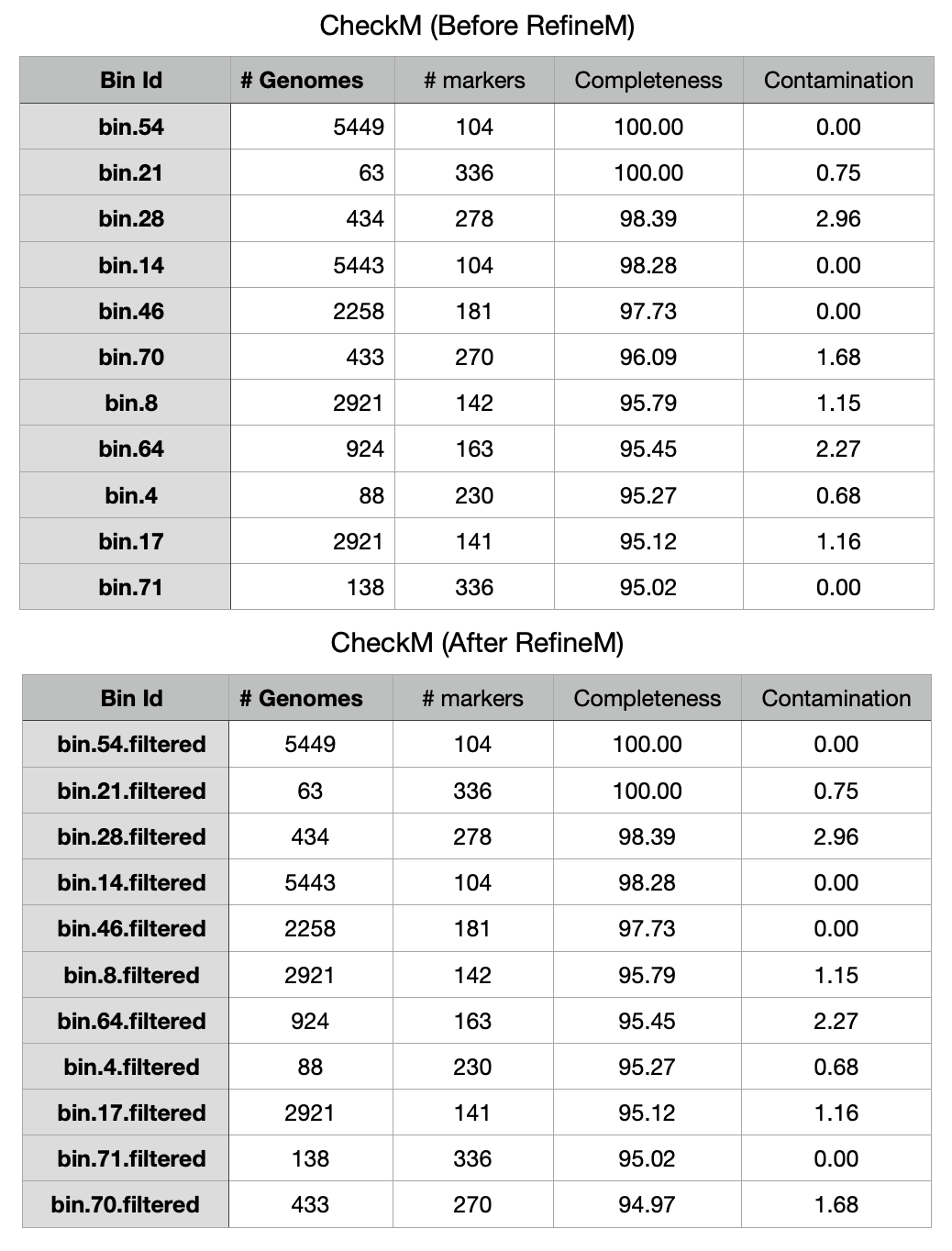
In this workflow, metagenome assembly is followed by binning to produce ‘bins,’ also known as Metagenome Assembled Genomes (MAGs), or draft genomes. Binning produces several to several hundred bins depending on the sample and the success of the assembly. Each bin is a collection of contigs the binning tool predicted to be from the same species. Binning essentially produces population level genomes because it is exceedingly difficult to distinguish contigs belonging to two slightly different strains of the same species. These population level bins are usually incomplete and may be chimeric (contigs from multiple strains or species in a single bin) because of the difficulties of metagenomic *de novo* assembly discussed previously. Still, incomplete draft genome bins can provide much more information than short reads in taxonomic assignment, functional analysis and other analysis. Binning tools are capable of clustering contigs from rare, low abundance, and previously uncultivated bacteria. This opens the scope of scientific inquiry into previously uninterpretable genetic material (Breitwieser, Lu, and Salzberg, “A Review of Methods and Databases for Metagenomic Classification and Assembly.”) There are two main binning methods: taxonomy-dependent methods (also called supervised methods) and taxonomy independent methods (also called unsupervised methods) (Pérez et al., 2020). Binning algorithms can use taxonomic information from a reference database (taxonomy-dependent or supervised binning), or they can cluster sequences using statistical properties and/or contig coverage (unsupervised binning). Many current methods use a combination of these features (Breitwieser et al., 2017).

Marker Gene Classification is a common supervised binning strategy. Contigs in the data are searched against a database of single copy conserved genes that have been curated based on their ability to distinguish certain species. The contigs are compared to these marker genes using fast and computationally inexpensive aligners (relative to comparing entire genomes). Amino acid classification works similarly to marker gene classification except amino acid sequences are used in place of marker genes. Because of the redundancy in codons during the translation process, amino acid sequences are conserved at much greater evolutionary distances than DNA sequences. Mutations in DNA do not necessarily mutate the amino acid sequence. This improves the sensitivity of the contigs classification but is more computationally intensive because sequences must be translated to amino acids (Breitwieser et al., 2017).

The binning tool used in this workflow is called Metabat2 (Kang et al., 2019). Metabat uses weighted graphs to group contigs. A complete explanation of the algorithm is out of the scope of this project. Because of the inevitable imperfections and contaminants in draft genome bins, quality assessment of these bins by dedicated software is another important step of the workflow. (Parks et al., 2018) CheckM and RefineM are two tools developed to provide quality assessment and refinement of bins (Parks et al., 2015).

## Binning Results

Starting from roughly 140 million short reads (150 base pairs in length) this workflow produced 72 bins of varying completeness.

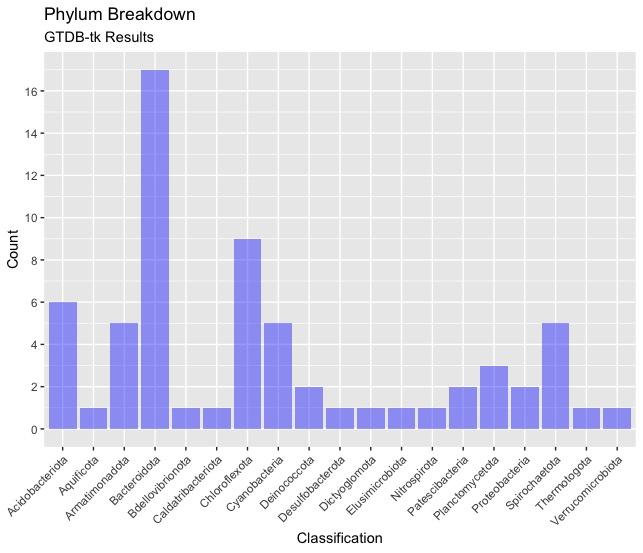


[Figure 8] These tables represent a truncated output from the tool CheckM before and after use of the tool RefineM for the first run data (SRR10848475) (Parks et al., 2015)(Parks et al., 2018) The table is sorted in descending order by genome completeness and shows the first eleven bins. The RefineM tool, removed contaminants from bin 70 which reduced its completeness and moved it from the sixth most complete to the eleventh most complete bin. Most of the bins were not reduced in completeness by the RefineM tool. SBATCH scripts for Metabat2, CheckM, and RefineM are included in the supplementary material.

## Taxonomic Classification

Once metagenome assembled genomes have been constructed and refined, taxonomic classification is the next logical step. In order to classify newly created MAGs, all or part of the MAG must be searched against a database of already classified genomes.

The NCBI microbial taxonomy is largely based on the ribosomal RNA gene. Metagenome Assembled Genomes may have very fragmented rRNA regions or the gene may be missing altogether. Therefore, marker gene and amino acid sequence classification methods are more useful for MAGs. A new taxonomy based on amino acid sequence phylogeny was proposed in 2018 which accommodates classification of MAGs very well (Parks et al., 2018) To build this taxonomic system, the researchers retrieved over 80 thousand publicly available microbial genomes and derived marker amino acid sequences to distinguish between species. The final database (referred to as a concatenated protein reference database) is known as the Genome Taxonomy Database (GTDB). Along with this newly proposed taxonomy, the researchers created a command line tool for efficient classification of MAGs based on the GTDB Taxonomy. (Chaumeil et al., 2020) The tool locates gene regions within the MAGs compatible with amino acid sequences in the GTDB database to place the MAGs within the GTDB taxonomy reference tree. An SBATCH script for running the GTDB-tk is included in the supplementary material.



[Figure 9] This histogram shows the distribution of phyla classified in the first run bins. Phylum Bacteriodita is most abundant in this sample (GTDB taxonomy). Bacterioidota are rod shaped bacteria very widely distributed in many different environments.

Part 2:

Two Methods for tree building with Metagenome Assembled Genomes

At this point in the workflow the initial 144 million short reads have become 65 population level bins classified to varying levels of taxonomic specificity. GTDB-tk is able to classify most bins to the phylum level but as taxonomic specificity increases, classification becomes less and less likely. Depending on the questions being asked of the data, a certain clade may be of interest and the researcher may want a deeper understanding of the bins belonging to that clade. The need to gain more information about a specific bin or group of bins is common in metagenomic research. Phylogenetic trees provide value information about the relative position of a bin amongst the other sequences or genomes used to construct the tree. Once a bin is placed amongst other genomes, similar genomes can be downloaded and used for further analyses. The second part of this project is the development of two methods for tree building to gain more insight on new metagenome assembled genomes produced by the workflow above.

The Phylum *Aquificae* (NCBI taxid200783) was chosen as a test case because the relatively small number of assemblies present in public databases cuts down on computational time and resources. Interestingly, the GTDB taxonomy has renamed this phylum to Aquificota (Parks et al., 2018).

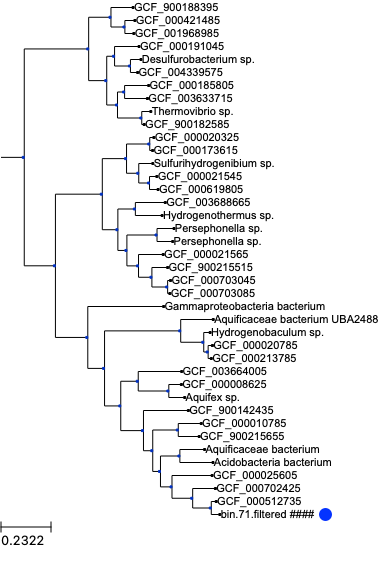
# Method 1: GTDB-tk Tree modification

The GTDB-toolkit classifies bacterial MAGs by placing them among the complete GTDB Bacterial Reference tree. The first method of tree building is actually a modification of the GTDB bacterial reference tree. The first problem with this tree is its size. This tree has 60 thousand nodes and 30 thousand leaves. A tree this large is cumbersome to work with.



[Figure 10] This image shows the complete GTDB Domain Bacteria reference tree produced in this workflow and visualized using the ITOL tree viewer (Letunic and Bork, 2019). This tree has roughly 60 thousand nodes and 30 thousand leaves. The pink lines represent branches, the yellow dashes represent evolutionary distance markers, and the blue text encircling the tree are the names of each node.

A command line tool written in Python was developed to detach only the clade of interest from the greater GTDB reference tree. The tool requires two inputs, a tree data structure in Newick format, and a clade of interest. The tool first loads the tree and immediately roots the tree using a midpoint outgroup. The tree leaves are then iteratively searched for the input clade. If the clade name is not located the program will print troubleshooting information and self terminate. If the clade name is located a subtree is detached. Next, the program downloads the relevant NCBI assembly metadata columns and cleans the accession number column to match with the GTDB-tk leaf name format. The program prints out a subtree in Newick format and a png before and after species names have been swapped in case the accession numbers in the leaf names are needed. The tool also tags bins to make them stand out against other leaves. If the resulting tree is too big, the tool can be modified to produce a text file necessary for the input into the Treemmer trimming tool (Menardo et al., 2018). This tool was developed using the Pandas and Ete3 Python Libraries and requires the Os, Numpy, and Sys libraries as dependencies. Convenience is the biggest advantage of this tool. Once the GTDB-tk output is obtained, this tool can run on a laptop computer in under a minute. The majority of the runtime is imposed by downloading the NCBI metadata table.



[Figure 11] This tree shows the output of the tool with the GTDB reference tree and the MAG belonging to phylum Aquificota obtained from the metagenomic samples as inputs. A blue dot has been added next to the bin of interest. With this tree, the researcher can get a sense of how novel the genome is and the surrounding species can be located and downloaded for further comparison and analyses.

This tree was created with the CommandLineTool.py and bacterial reference ree in the same directory using the following terminal (Unix Shell) command: “python CLI.py gtdbtk.bac120.classify.tree Aquificota”.

# Method 2: De Novo Reference Tree

Though the GTDB bacterial reference tree is valuable and convenient, it has drawbacks. The GTDB reference tree is constructed from a curated database which is updated biannually. Given the constant influx of sequencing data to public databases, a researcher may benefit greatly from having up to date genomes in their tree, as opposed to genomes in a months old database. Additionally, as the GTDB is curated, use of the GTDB reference tree means acceptance of the quality thresholds and other filtering measures that database uses. Lastly, the GTDB reference tree is not a fact but a hypothesis. For full autonomy over the tree building process and the contents of the tree, a researcher must build their own tree from scratch. This second method is a newly developed de novo tree building workflow specifically designed for placement of metagenome assembled genomes. This method employs computationally intensive tools and therefore requires the use of a computing cluster. The method has three main steps. The first step is to download all assemblies of a clade of interest. The second step is to dereplicate these assemblies so only high quality assemblies of significant difference from each other remain. The final step is to add the bins of interest and build a phylogeny.

## Pyani download

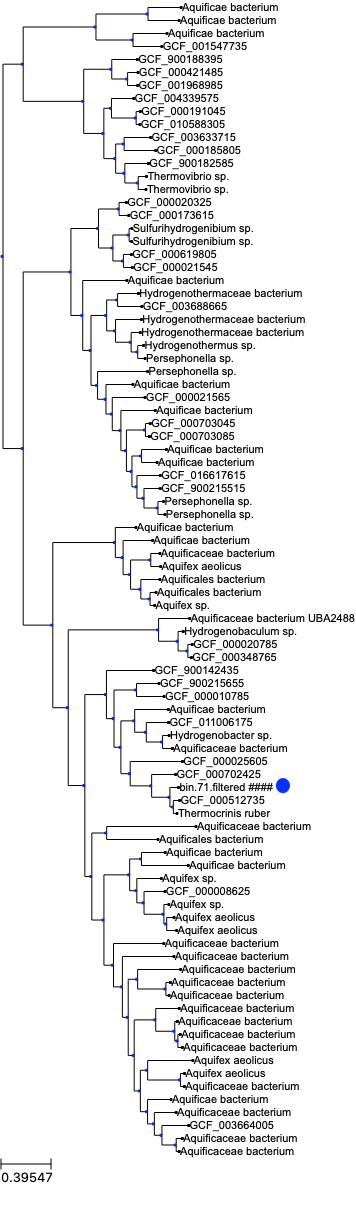
NCBI provides convenient access to all available assemblies of any given clade. The tool Pyani which was originally written to perform FastANI comparison between genomes contains a genome downloading functionality (Pritchard et al., 2015). The Pyani downloading script takes a taxon ID as input and will rapidly download all assemblies below that taxon ID. The SBATCH script used to download all 147 Aquificacea assemblies (TaxID 200783) is provided in the supplementary material.

## dRep dereplication

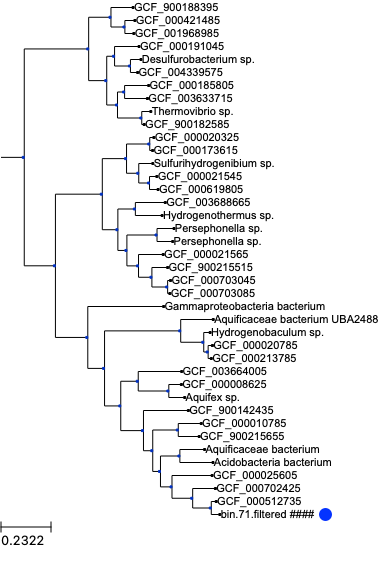
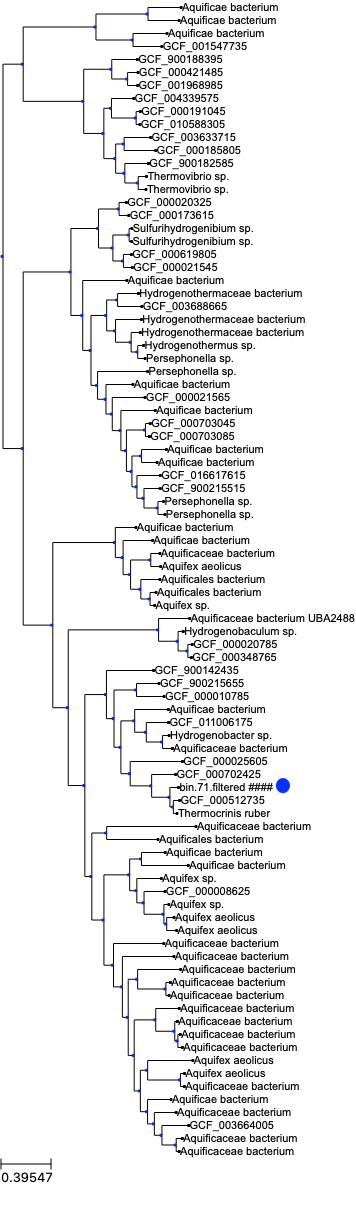
The dRep tool was developed to select the highest quality and least redundant genomes from a group (Olm et al., 2017). The dRep algorithm impressively has a linear time complexity as the number of input genomes increases. First genomes are clustered for similarity using a computationally efficient program called mash. Next, genomes within the clusters are compared using the very computationally intensive gANI program to create secondary clusters of nearly identical genomes. By combining low and high power comparison tools dRep accomplishes accurate comparisons with less runtime than using gANI alone. The secondary clusters are then assessed for quality using CheckM and the lower quality genomes are removed. dRep provides a range of options for cluster similarity thresholds and quality thresholds (Olm et al., “DRep.”). The SBATCH script used to dereplicate the 179 Aquificacea assemblies can be found in the supplementary material. With the default parameters on dRep removed 91 of the 179 assemblies leaving 88 of the most representative, highest quality Aquificacea assemblies for tree building.

## Phylophlan 3.0 tree building

PhyloPhlan is one of the most popular tree building tools. The tool was originally produced in 2013 but has since been rewritten in the 2020 release of PhyloPhlan 3.0 (Asnicar et al., 2020) PhyloPhlan is able to build phylogenies at high and low taxonomic resolutions by using clade specific maximally informative marker sequences. Notably, PhyloPhlan was developed for a similar purpose as the GTDB-toolkit and has the ability to classify newly assembled MAGs but has similar drawbacks to the GTDB-tk discussed above. The Aquficacea bin identified by the GTDB-tk (bin 71) was added to the group of genomes remaining after dereplication by dRep. This collection of genomes was then run on PhyloPhlan using the low diversity parameter. The default configuration file was used. The SBATCH script used to run PhyloPhlan to create the Aquificacea tree seen below is located in the supplementary materials.



[Figure 12] This is the resulting tree of the de novo assembly of the Aquificacea phylum. A blue dot has been added next to the bin of interest. This tree provides a more robust and up to date view of how bin 71 fits among the known Aquificacea diversity.



[Figure 13] This shows the trees produced by both methods adjacent to each other. In comparison, the *de novo* tree shows a much more thorough representation of phylum Aquificacea.

# Discussion and Conclusion

Metagenomic analysis of whole metagenome shotgun sequencing has replaced cultivation based methods of studying microbial communities (Kang et al., 2019). As more and more metagenomic data is sampled and sequenced, more workflows and computational techniques will be developed to squeeze more insight from the data. The metagenomic de novo assembly workflow and tree building methods provided in this project add to and update the existing workflows and methods in current use.

There are two improvements to consider if implementing the assembly pipeline. First is the use of the MetaSpades assembler (Nurk et al., 2017). The MetaSpades assembler is more computationally intensive than the Megahit assembler and therefore has a much longer runtime but tends to produce more accurate contigs of high depth areas within a sample. Megahit is often better able to assemble low depth areas (Breitwieser et al., 2017). A second improvement could be the use of a dereplication tool as part of the binning process. The DAS tool takes metagenomic assembly and binning results into consideration and produces potentially more complete, dereplicated bins (Sieber et al., 2018).

The de novo reference tree building method could be substantially improved in efficiency by downloading reference genomes from NCBI rather than assemblies. Reference genomes are non redundant. By downloading assemblies, the dRep tool will inevitably need to remove genomes that would not be present if only reference genomes were downloaded. It is certainly possible to download reference genomes only but no existing tool was located and writing a script to do so was outside the scope of this project. The next most important improvement to the de novo tree building method is the conversion of individual bash scripts to a consolidated pipeline using a workflow language such as Nextflow. NextFlow acts as a wrapper for individual SBATCH scripts and could be written to manage job submission so the user would simply run the NextFlow script instead of needing to submit 3 individual SBATCH scripts.

# Supplementary material and test data

The supplementary material includes

-SBATCH scripts for each step of the metagenome assembly pipeline

-SBATCH scripts for the de novo tree building pipeline

-the python command line tool described in Part 2 Method 1

-a GTDB-tk reference tree to test the python tool

-a letter from the JGI director describing the change in policy regarding data transfers to the NCBI SRA

### **Bibliography**

Asnicar, F., Thomas, A. M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nature Communications* 11, 2500. doi:10.1038/s41467-020-16366-7.

Baker, M., and Buyya, R. Cluster Computing at a Glance. 45.

Breitwieser, F. P., Lu, J., and Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 20, 1125–1136. doi:10.1093/bib/bbx120.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36, 1925–1927. doi:10.1093/bioinformatics/btz848.

Churko, J. M., Mantalas, G. L., Snyder, M. P., and Wu, J. C. (2013). Overview of High Throughput Sequencing Technologies to Elucidate Molecular Pathways in Cardiovascular Diseases. *Circ Res* 112. doi:10.1161/CIRCRESAHA.113.300939.

Cochrane, G., Karsch-Mizrachi, I., and Nakamura, Y. (2011). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* 39, D15–D18. doi:10.1093/nar/gkq1150.

Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). Why are de Bruijn graphs useful for genome assembly? *Nat Biotechnol* 29, 987–991. doi:10.1038/nbt.2023.

Escobar-Zepeda, A., Vera-Ponce de León, A., and Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Front Genet* 6. doi:10.3389/fgene.2015.00348.

Felczykowska, A., Krajewska, A., Zielińska, S., and Łoś, J. M. (2015). Sampling, metadata and DNA extraction - important steps in metagenomic studies. *Acta Biochim Pol* 62, 151–160. doi:10.18388/abp.2014\_916.

Frontiers | Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context | Microbiology Available at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00780/full [Accessed April 19, 2021].

Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., et al. (2002). Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol* 68, 4301–4306. doi:10.1128/aem.68.9.4301-4306.2002.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi:10.1093/bioinformatics/btt086.

Jones, M. B., Highlander, S. K., Anderson, E. L., Li, W., Dayrit, M., Klitgord, N., et al. (2015). Library preparation methodology can influence genomic and functional predictions in human microbiome research. *PNAS* 112, 14024–14029. doi:10.1073/pnas.1519288112.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., et al. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7. doi:10.7717/peerj.7359.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* 47, W256–W259. doi:10.1093/nar/gkz239.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., et al. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102, 3–11. doi:10.1016/j.ymeth.2016.02.020.

Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S. M., Rutaihwa, L. K., et al. (2018). Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* 19, 164. doi:10.1186/s12859-018-2164-8.

Merino, N., Aronson, H. S., Bojanova, D. P., Feyhl-Buska, J., Wong, M. L., Zhang, S., et al. (2019). Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context. *Front. Microbiol.* 10. doi:10.3389/fmicb.2019.00780.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 27, 824–834. doi:10.1101/gr.213959.116.

Olm, M. R., Brown, C. T., Brooks, B., and Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal* 11, 2864–2868. doi:10.1038/ismej.2017.126.

Oren, A. (2004). Prokaryote diversity and taxonomy: current status and future challenges. *Philos Trans R Soc Lond B Biol Sci* 359, 623–638. doi:10.1098/rstb.2003.1458.

Pangenomics Analysis Reveals Diversification of Enzyme Families and Niche Specialization in Globally Abundant SAR202 Bacteria | mBio Available at: https://mbio.asm.org/content/11/1/e02975-19 [Accessed April 18, 2021].

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* 36, 996–1004. doi:10.1038/nbt.4229.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25, 1043–1055. doi:10.1101/gr.186072.114.

Paul, F., Otte, J., Schmitt, I., and Dal Grande, F. (2018). Comparing Sanger sequencing and high-throughput metabarcoding for inferring photobiont diversity in lichens. *Scientific Reports* 8, 8624. doi:10.1038/s41598-018-26947-8.

Pérez, A. E., Gomez, L., and Buchrieser, C. Metagenomic approaches in microbial ecology: an update on whole-g­ enome and marker gene sequencing analyses. *Microbial Genomics*, 22.

Piwosz, K., Shabarova, T., Pernthaler, J., Posch, T., Šimek, K., Porcal, P., et al. (2020). Bacterial and Eukaryotic Small-Subunit Amplicon Data Do Not Provide a Quantitative Picture of Microbial Communities, but They Are Reliable in the Context of Ecological Interpretations. *mSphere* 5. doi:10.1128/mSphere.00052-20.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., and Toth, I. K. (2015). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods* 8, 12–24. doi:10.1039/C5AY02550H.

Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life | Nature Microbiology Available at: https://www.nature.com/articles/s41564-017-0012-7 [Accessed April 12, 2021].

Segata, N., Börnigen, D., Morgan, X. C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 4, 2304. doi:10.1038/ncomms3304.

Shakya, M., Lo, C.-C., and Chain, P. S. G. (2019). Advances and Challenges in Metatranscriptomic Analysis. *Front. Genet.* 10. doi:10.3389/fgene.2019.00904.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2017). DNA sequencing at 40: past, present and future. *Nature* 550, 1–9. doi:http://dx.doi.org.proxygw.wrlc.org/10.1038/nature24286.

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., et al. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* 3, 836–843. doi:10.1038/s41564-018-0171-1.

Singer, E., Andreopoulos, B., Bowers, R. M., Lee, J., Deshpande, S., Chiniquy, J., et al. (2016). Next generation sequencing data of a defined microbial mock community. *Sci Data* 3. doi:10.1038/sdata.2016.81.

Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of Next Generation Sequencing Technologies. *Curr Protoc Mol Biol* 122, e59. doi:10.1002/cpmb.59.

Staley, J. T., and Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* 39, 321–346. doi:10.1146/annurev.mi.39.100185.001541.

Thiel, V., Wood, J. M., Olsen, M. T., Tank, M., Klatt, C. G., Ward, D. M., et al. (2016). The Dark Side of the Mushroom Spring Microbial Mat: Life in the Shadow of Chlorophototrophs. I. Microbial Diversity Based on 16S rRNA Gene Amplicons and Metagenomic Sequencing. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.00919.

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

Vaz-Moreira, I., Egas, C., Nunes, O. C., and Manaia, C. M. (2011). Culture-dependent and culture-independent diversity surveys target different bacteria: a case study in a freshwater sample. *Antonie van Leeuwenhoek* 100, 245–257. doi:10.1007/s10482-011-9583-0.

Wang, G. Y., Graziani, E., Waters, B., Pan, W., Li, X., McDermott, J., et al. (2000). Novel natural products from soil DNA libraries in a streptomycete host. *Org Lett* 2, 2401–2404. doi:10.1021/ol005860z.