1. Project Information

Program	Microbial/CSP 2015
Sequencing Project ID	1231330
Sequencing Project Name	Sphingobium yanoikuyae WW5

2. Library Information

File name	12828.3.289146.ATGACGTC-GACGTCAT.filter-ISO.fastq.gz
Library	GCPHY
Number of reads	28,601,758
Read type	2x151 bp

3. Read QC Results

All raw Illumina sequence data was filtered per SOP 1061 to remove known process artifacts and contaminants using BBTools. The following are the result of screening reads against a database of typical library construction artifact sequences using BBTools. Data was subsampled to 10,000,000 reads.

Illumina Std PE Read Filter Statistics

Description	Read Count	Reads Removed	Percent Reads Removed
Input	28,601,758	0	0.00%
Subsampled	10,000,000	18,601,758	65.04%
Total remaining	10,000,000		

The following are the result of screening reads against potential reagent and process contaminants using BBTools (>=95% ID) but were not removed from the dataset.

Illumina Std PE Contamination

No contaminates found in the input data set.

4. Assembly Statistics

The subsampled reads were assembled using SPAdes.
Assembly version: spades/3.13.0
Assembly parameters: —phred–offset 33 —cov–cutoff auto –t 16 –m 64 —careful –k 25,55,95

Scaffold total	136
Contig total	141
Scaffold sequence length (bp)	5,527,661
Contig sequence length (bp)	5,527,265
Scaffold N/L50	11/152855
Contig N/L50	11/152855
Largest Contig (bp)	399,272
Number of scaffolds >50 kb	29
Percent of genome in scaffolds >50 kb	78.97%
Percent of reads assembled	98.69%

5. Assembly Quality Assessment

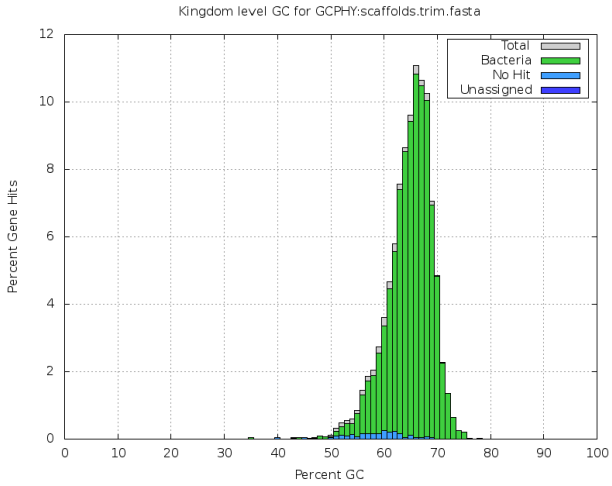
The quality of the final assembly was determined by using tRNAscan-SE to count tRNAs, barrnap to determine the presence of the 5S, 16S and 23S genes and checkm to determine completeness and contamination.

5S	Yes
16S	Yes
23S	Yes
tRNA Count	23
Completeness	99.59%
Contamination	1.43%

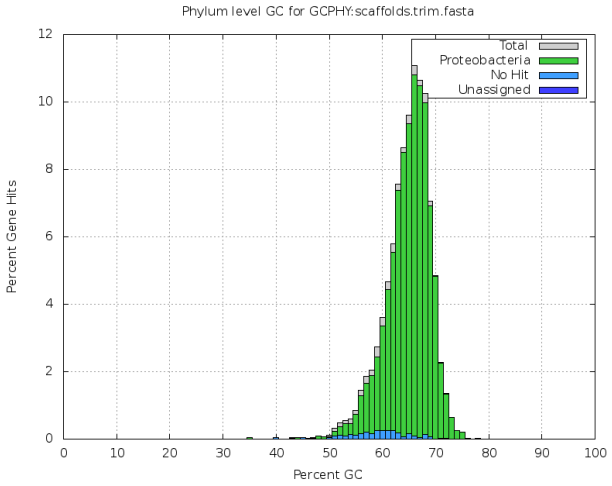
6. Assembly QC Results

Prodigal was used to predict cds on each scaffold and the output protein sequences were aligned to NCBI nr using LAST. Taxonomic information was extracted from the alignments and used to color-code scaffold GC content histograms.

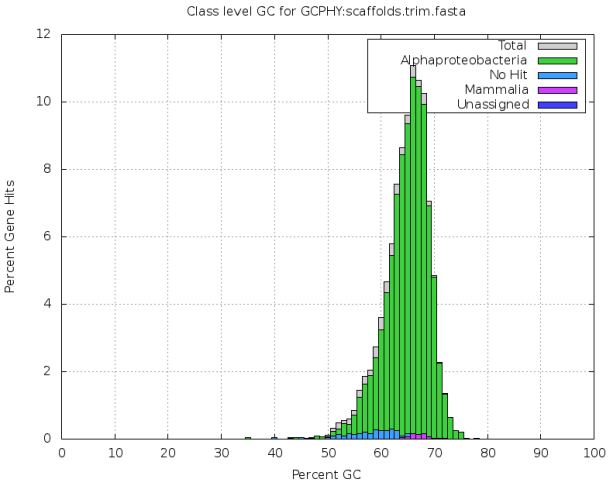
Kingdom



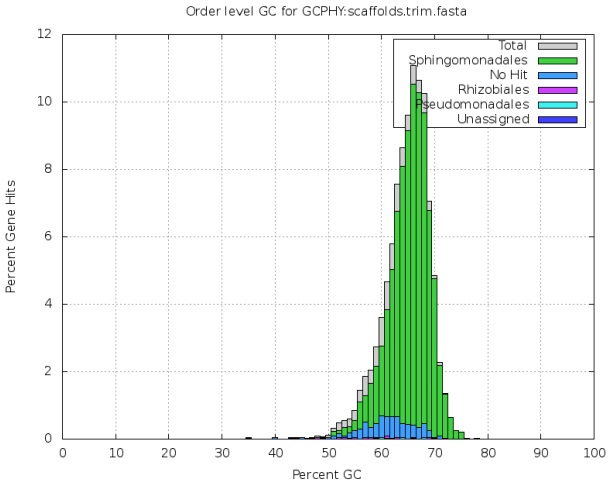
Phylum



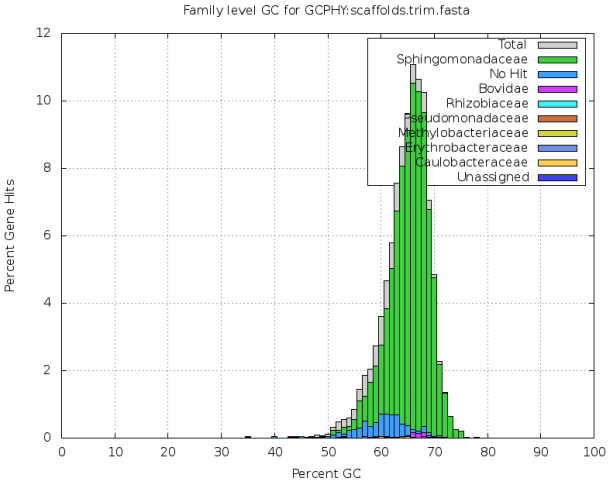
Class



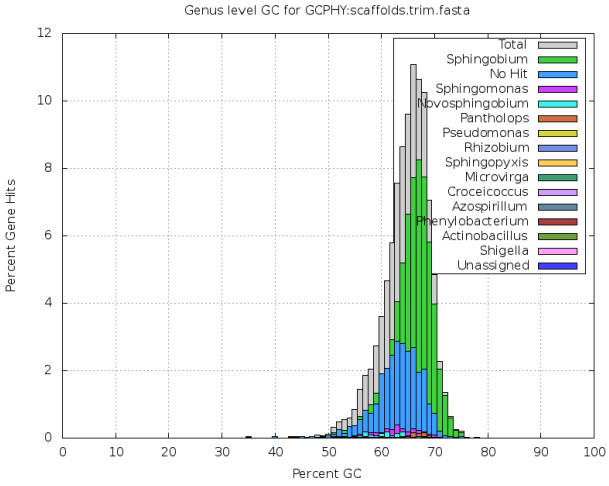
Order



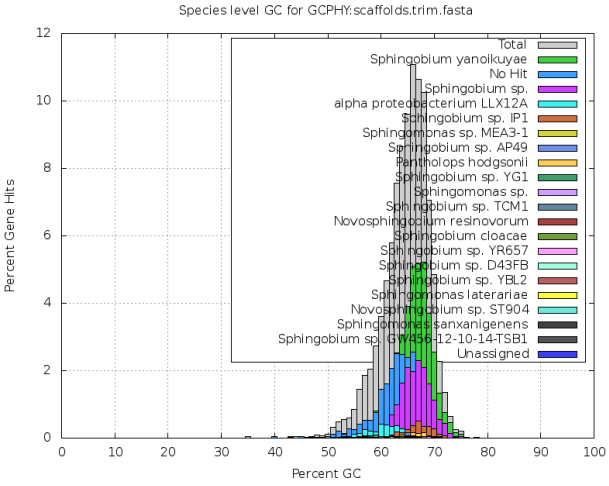
Family



Genus

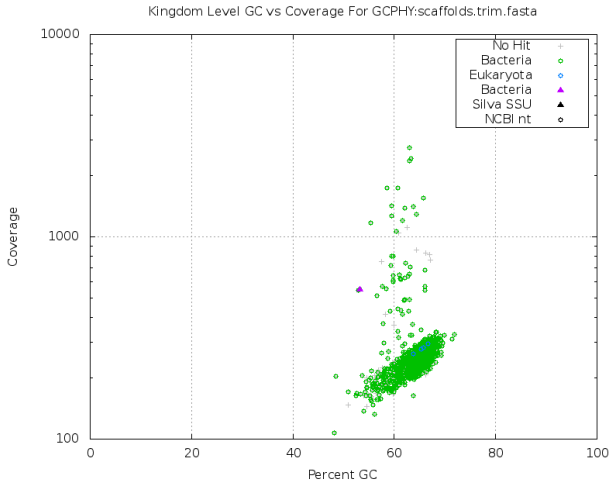


Species

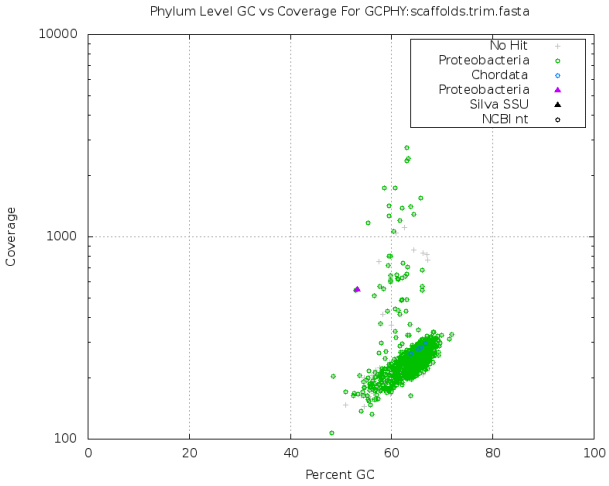


GC verses coverage of assembled scaffolds, overlaid with Silva SSU gene hits and NCBI nt megablast hits shown for different taxonomic levels.

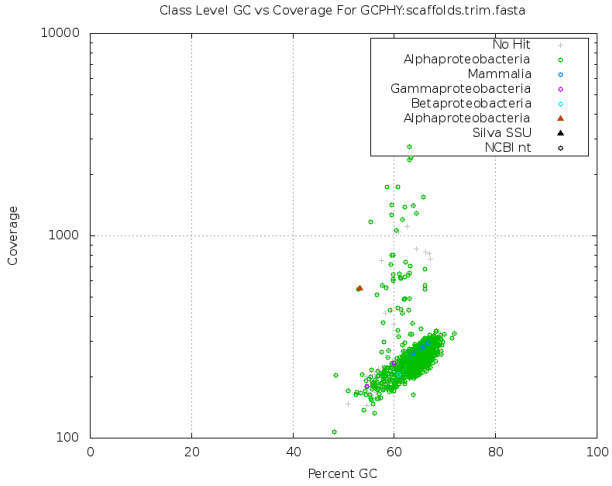
Kingdom



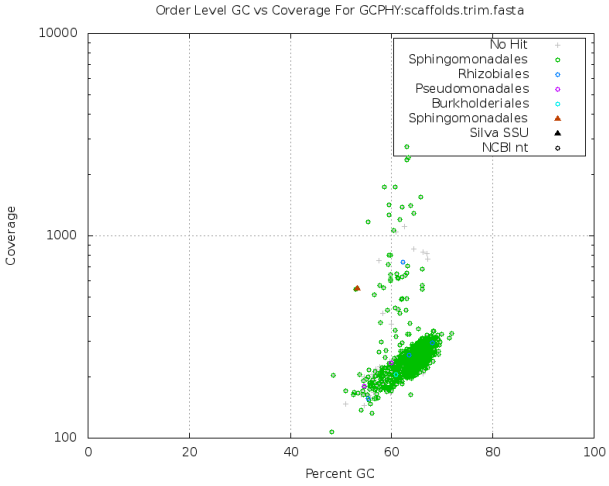
Phylum



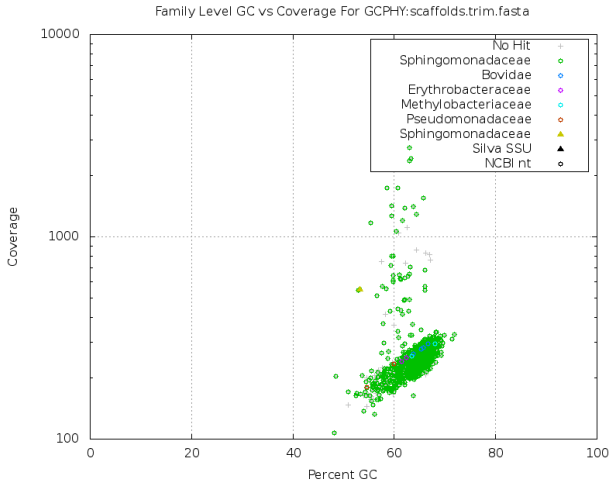
Class



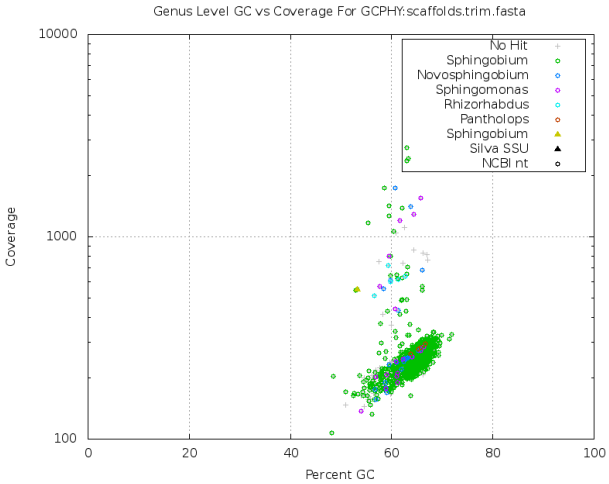
Order



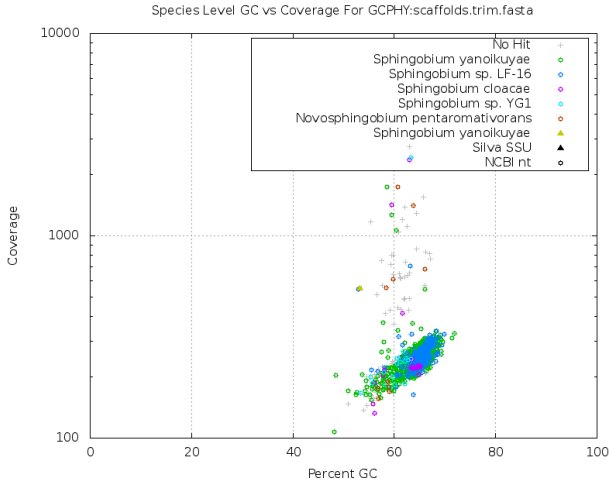
Family



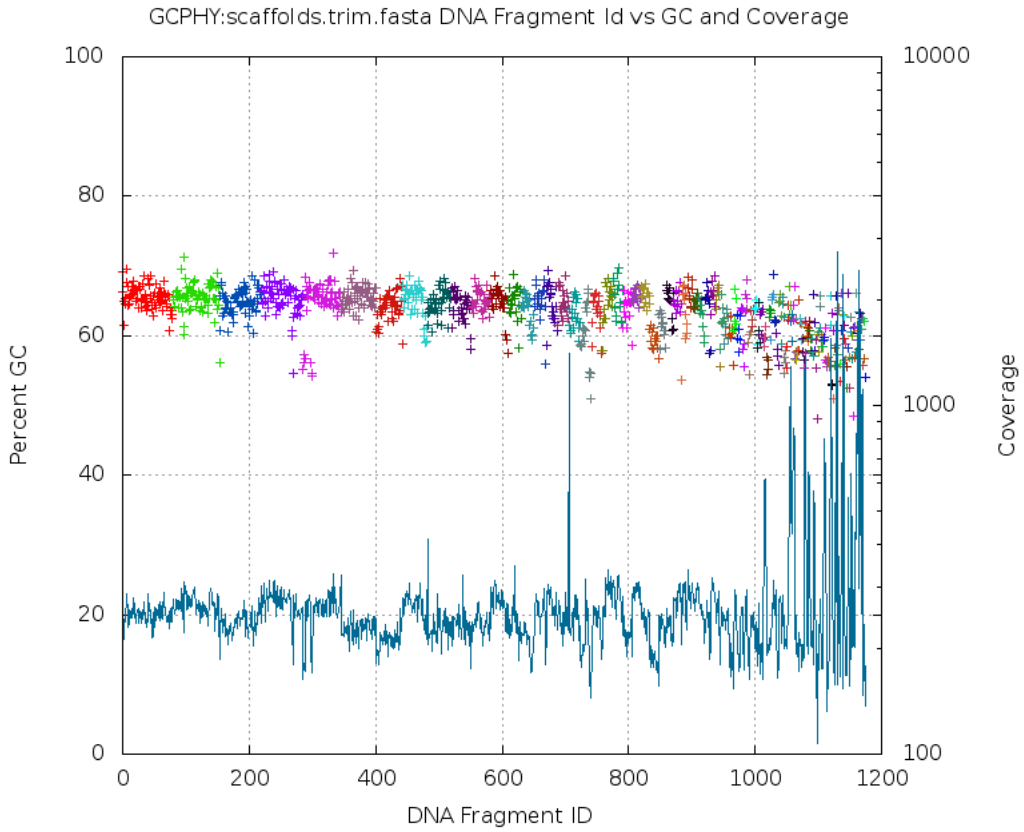
Genus



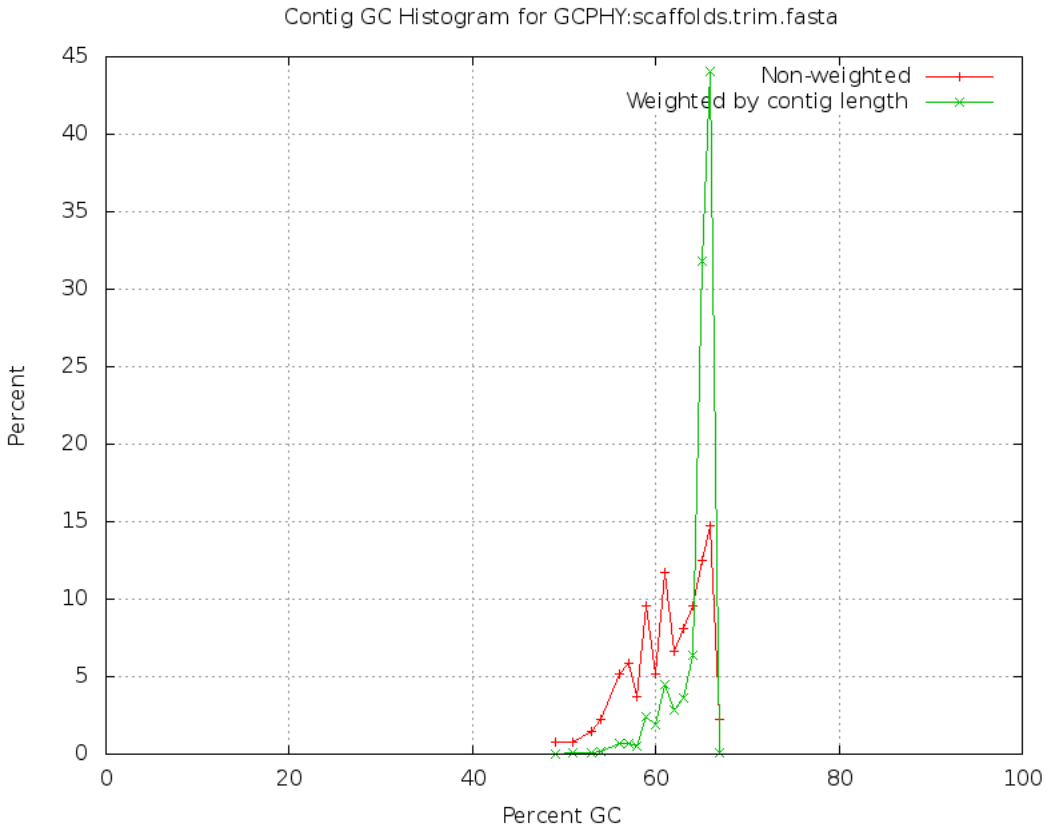
Species



Coverage and GC information. Scaffolds were shredded into non-overlapping 5 kbp fragments and the GC content of each shred was plotted as a data point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.



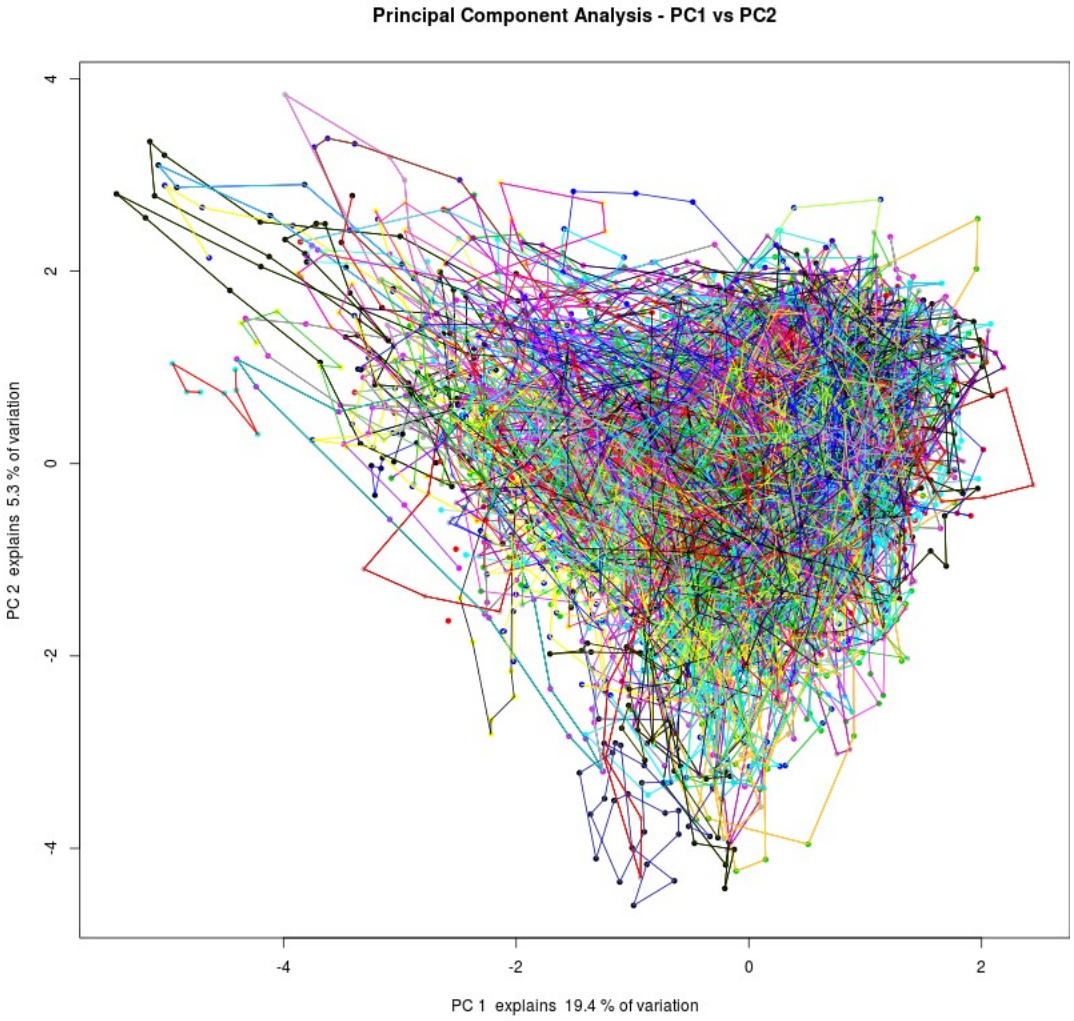
GC histogram of the scaffolds, including scaffold length weighted distribution.



List of the top scaffold megablast hits against 16s ribosomal genes using the Silva SSU database.

Organism: N/ABacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingobium;Sphingobium yanoikuyae;
Contig Name: GCPHY_NODE_85_length_5435_cov_199.085955
Align Length: 1,467 bp
Percent Id: 99.93%

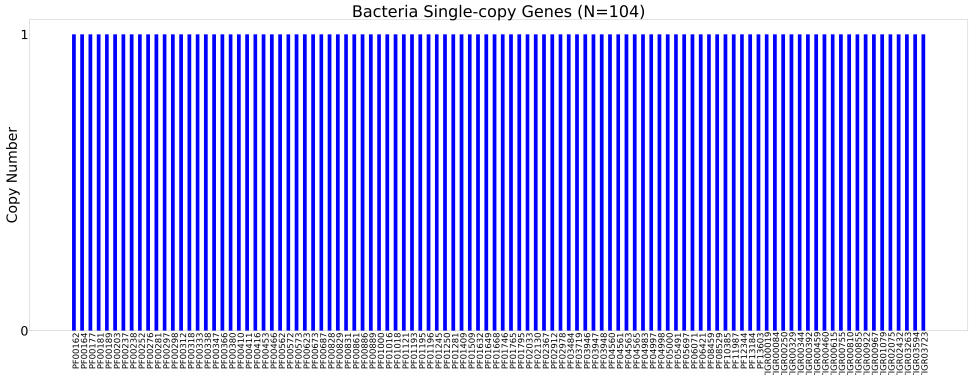
Tetramer frequencies are calculated over 5kb sliding windows of all scaffolds, followed by principal component analysis. Plots of the first two principal components are colored by scaffold.



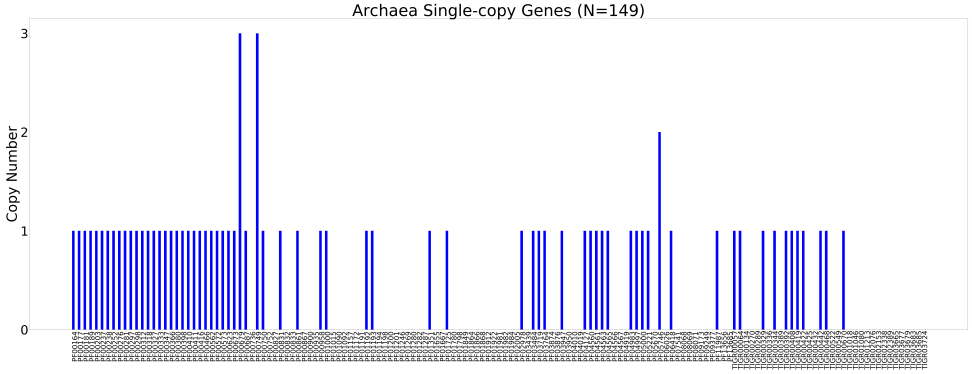
Estimated genome recovery derived from analysis of universal single-copy genes detected in final assembly using CheckM.

HMM	Found Genes	Total Genes	Percent Recovered
Archaea	69	149	46.31%
Bacteria	104	104	100.00%
Lineage Workflow	567	569	99.65%

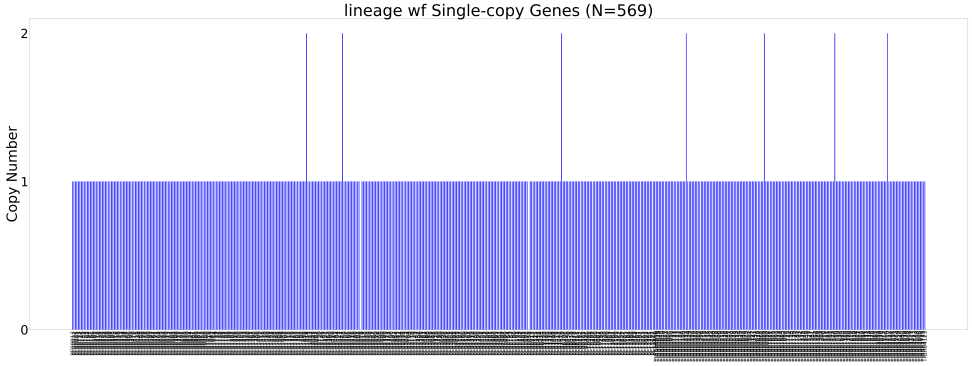
Bacteria Single–copy Gene Histogram



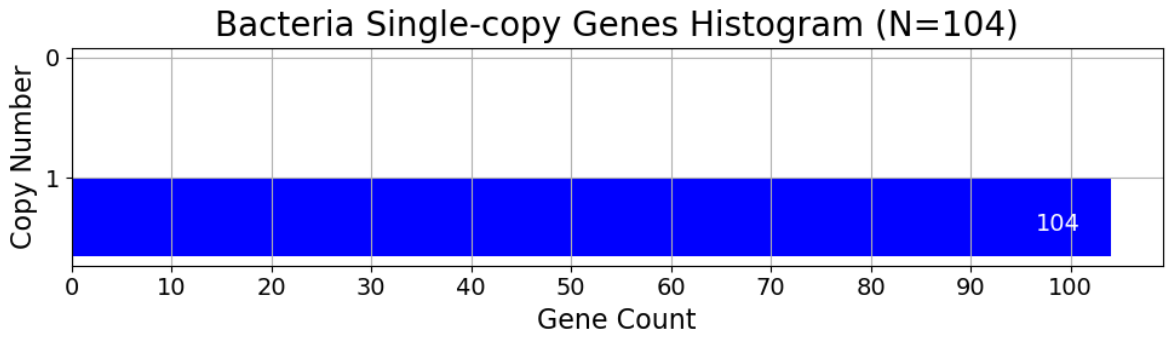
Archaea Single–copy Gene Histogram



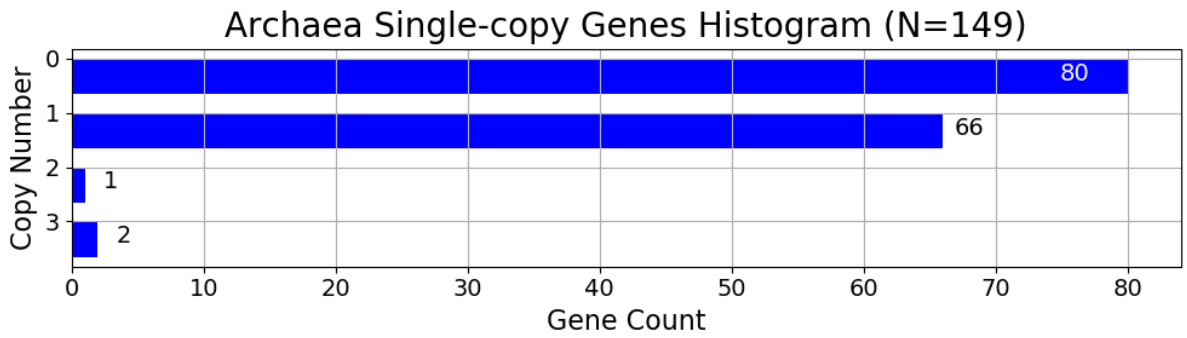
Lineage Workflow Single–copy Gene Histogram



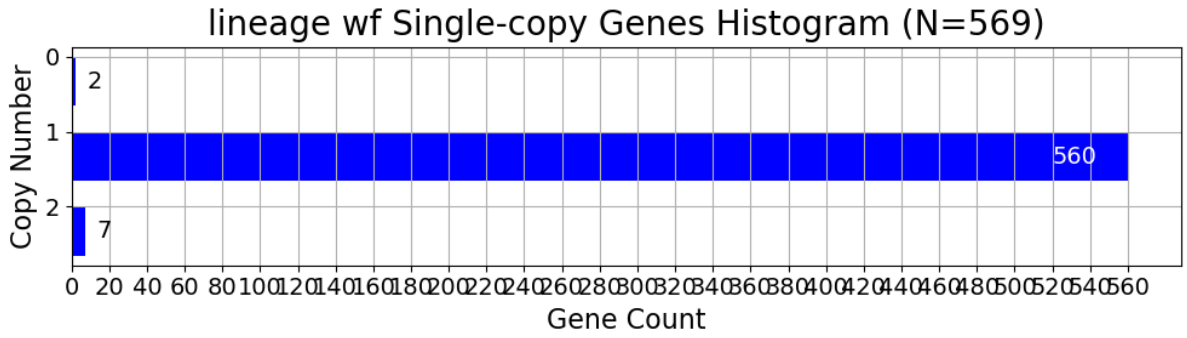
Bacteria Single–copy Genes



Archaea Single–copy Genes



Lineage Workflow Single–copy Genes



7. Sequence Data Availability

The sequence fasta files can be downloaded from our JGI portal website.
http://www.jgi.doe.gov/genome-projects

8. Annotation Data Availability

The annotation of the assembled contigs can be found within IMG.
http://img.jgi.doe.gov

9. Methods

Isolate Minimal Draft

Genome Sequencing and Assembly
The draft genome of *None* was generated at the DOE Joint Genome Institute (JGI) using Illumina technology [1]. An Illumina standard shotgun library was constructed and sequenced using the Illumina NovaSeq platform which generated 28,910,906 reads totaling 4,365,546,806 bp. Raw Illumina sequence was quality filtered using BBTools [2] per SOP 1061. The following steps were then performed for assembly: (1) artifact filtered and normalized Illumina reads were assembled using SPAdes (version v3.13.0; —phred–offset 33 —cov–cutoff auto –t 16 –m 64 —careful –k 25,55,95) [3]; (2) contigs were discarded if the length was <1kbp (BBTools reformat.sh: minlength). The final draft assembly contained 141 contigs in 136 scaffolds, totaling 5,527,661 bp in size. The final assembly was based on 1,498,704,144 bp of Illumina data with a mapped coverage of 273.8X.

1. Bennett S. Solexa Ltd. Pharmacogenomics. 2004;5(4):433–8.
2. B. Bushnell: BBTools software package (version 38.39), URL http://bbtools.jgi.doe.gov.
3. Bankevich A, et.al, SPAdes: a new genome assembly algorithm and its applications to single–cell sequencing. J Comput Biol 2012; 19:455–77.

DOE Auspice Statement for Publication
The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231.
The data was generated for JGI Proposal #1673.