

1. Project Information

Program	Microbial/BRC-BESC 2014
Sequencing Project ID	1176337
Sequencing Project Name	Shinella zoogloeoides JH0039

2. Read Statistics

Metric	Raw Reads	Filtered Subreads	Error Corrected Reads
Reads	534,808	146,523	8,641
Bases	1,301,160,754	877,424,928	87,578,516
Average Read Length	2,432.9 $\pm$ 4,002.1	5,988.3 $\pm$ 3,855.8	10,135.2 $\pm$ 4,492.1
Reads >5kbp	97,055	74,035	7,000
Bases, reads >5 kbp	960,929,749	676,017,858	84,973,668
Avg Read Length, reads >5kbp	9,900.9 $\pm$ 3,748.0	9,131.1 $\pm$ 2,824.3	12,139.1 $\pm$ 1,840.7

3. Assembly Statistics

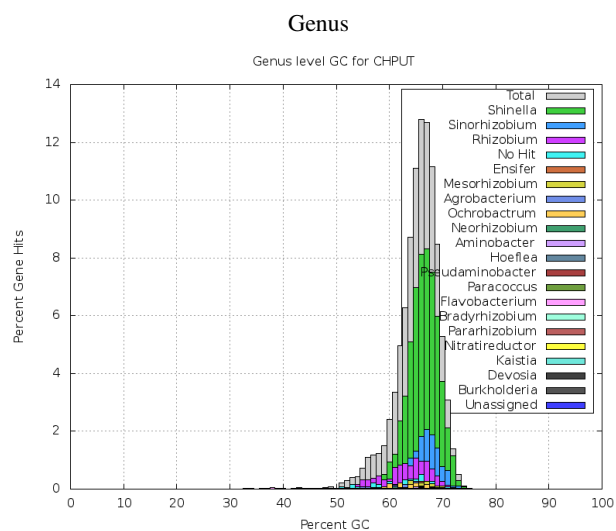
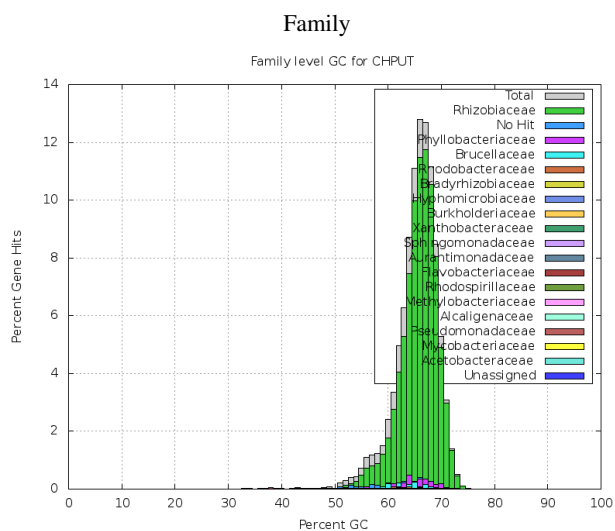
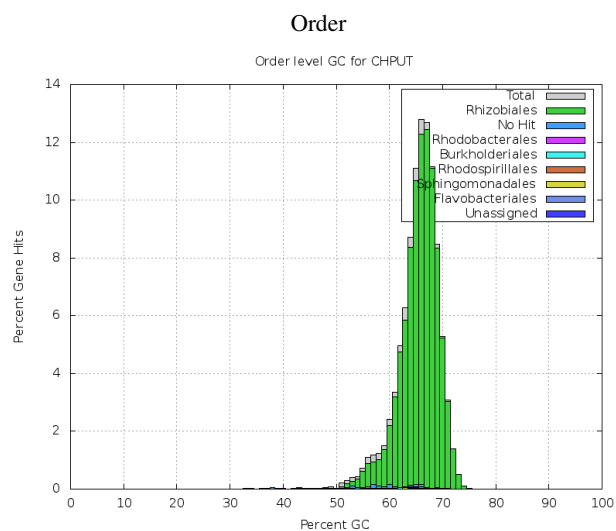
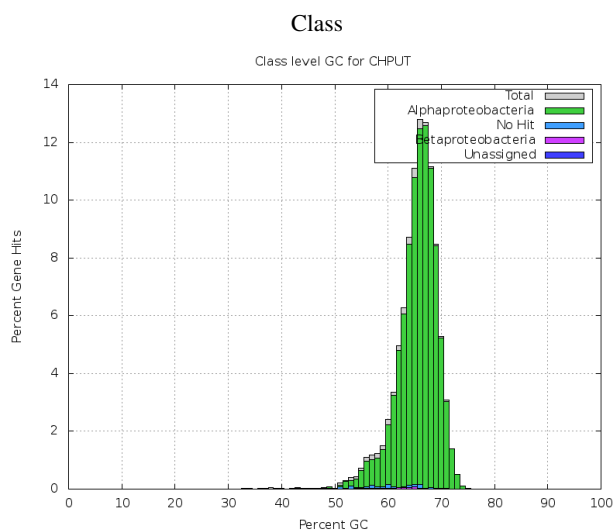
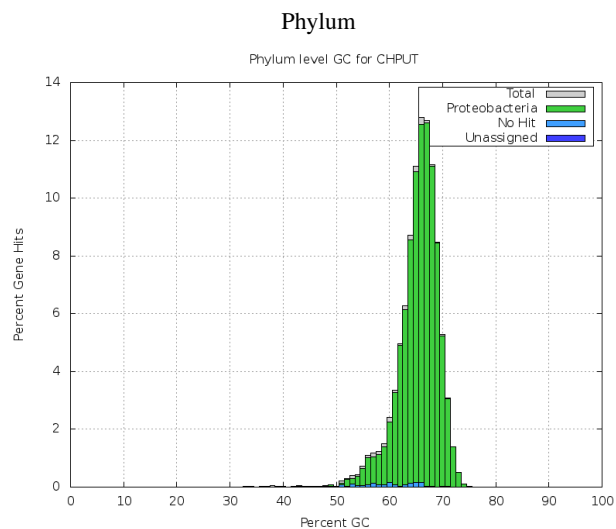
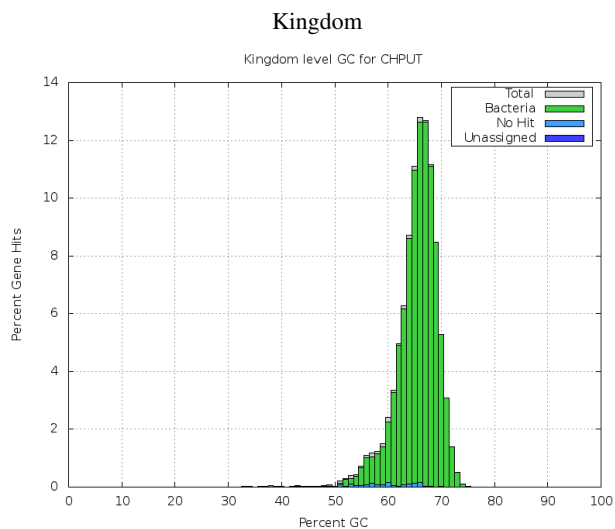
The filtered subreads were assembled using PacBio's HGAP assembler.

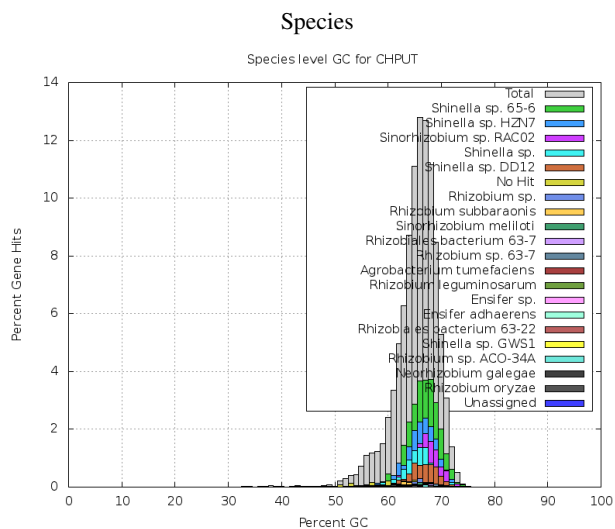
Assembly version: smrtanalysis/2.3.0_p5, HGAP 3

Scaffold total	8
Contig total	8
Scaffold sequence length (bp)	5,680,044
Contig sequence length (bp)	5,680,044
Scaffold N/L50	1/3519860
Contig N/L50	1/3519860
Largest Contig (bp)	3,519,860
Number of scaffolds >50 kb	5
Percent of genome in scaffolds >50 kb	99.89%
Percent of reads assembled	82.42%

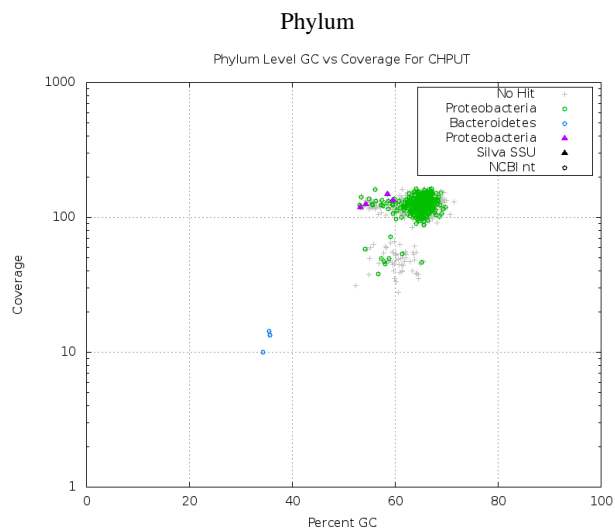
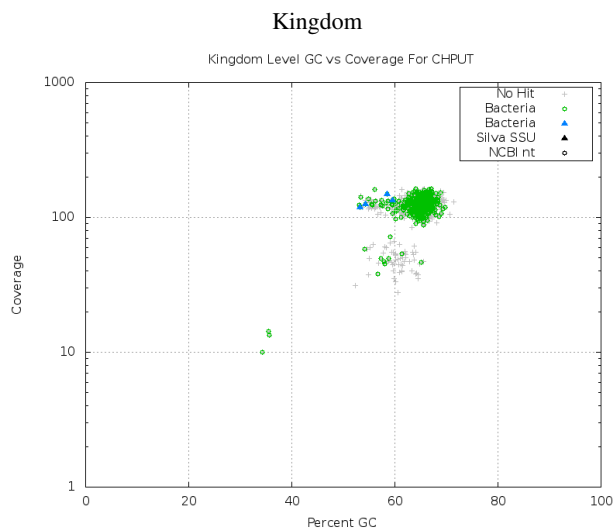
4. Assembly QC Results

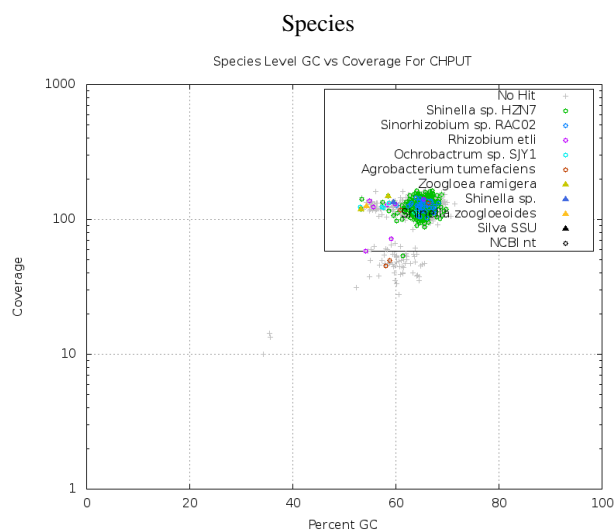
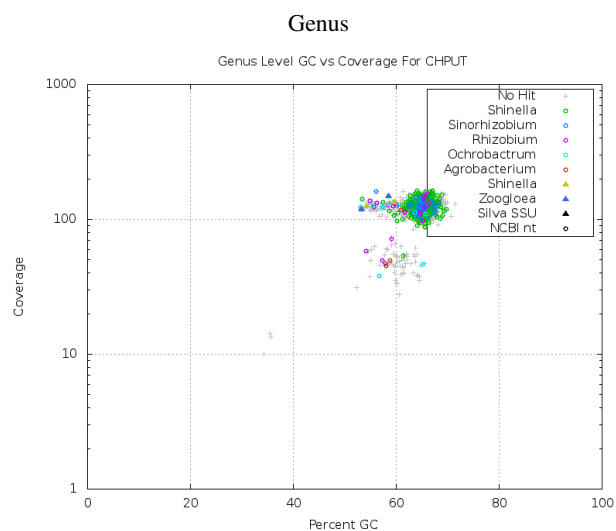
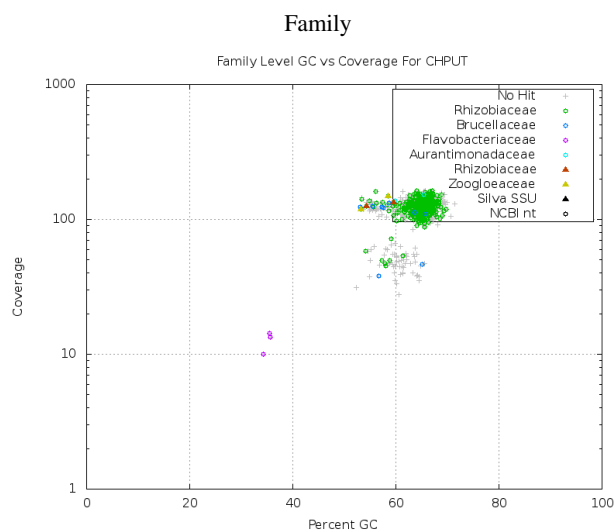
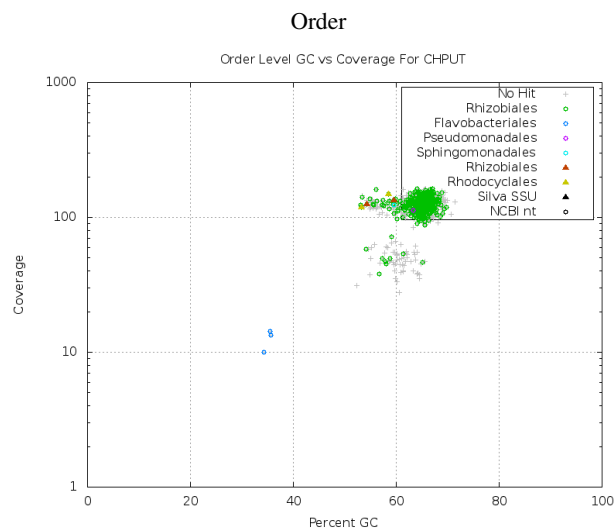
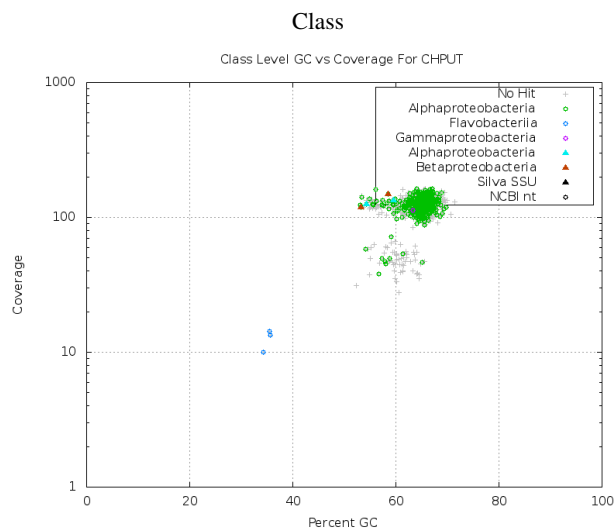
Prodigal was used to predict cds on each scaffold and the output protein sequences were aligned to NCBI nr using LAST. Taxonomic information was extracted from the alignments and used to color-code scaffold GC content histograms.



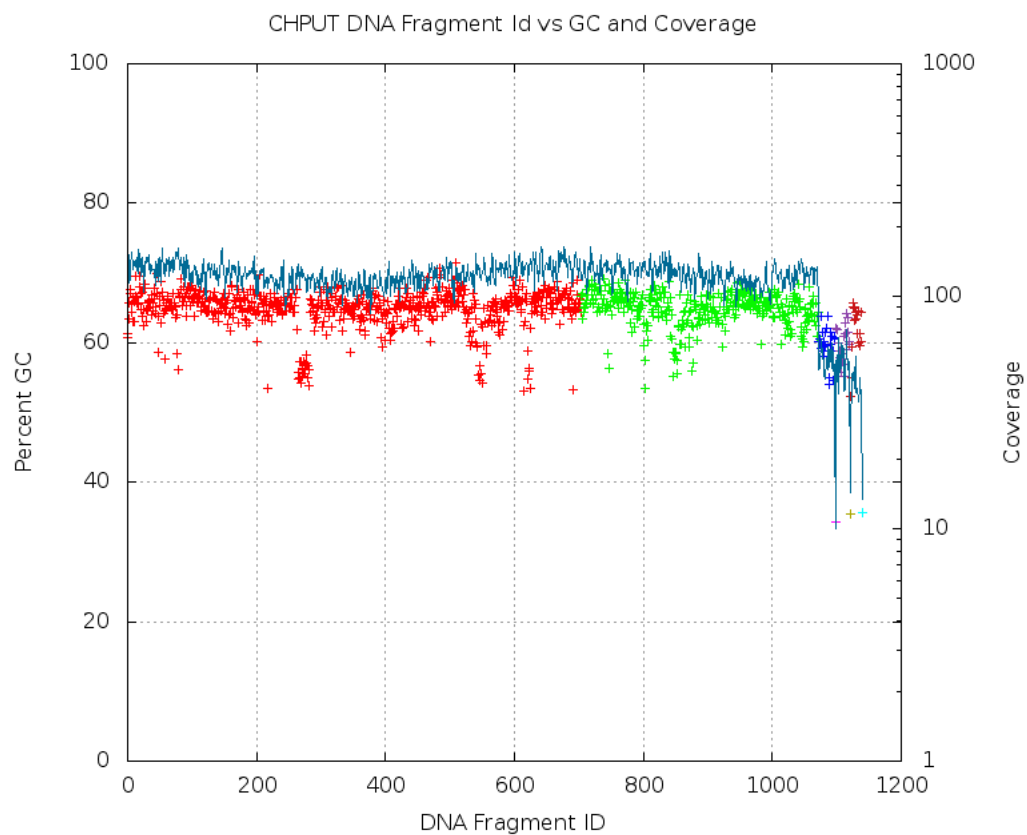


GC versus coverage of assembled scaffolds, overlaid with GreenGenes 16s rRNA gene hits and NCBI nt megablast hits shown for different taxonomic levels.

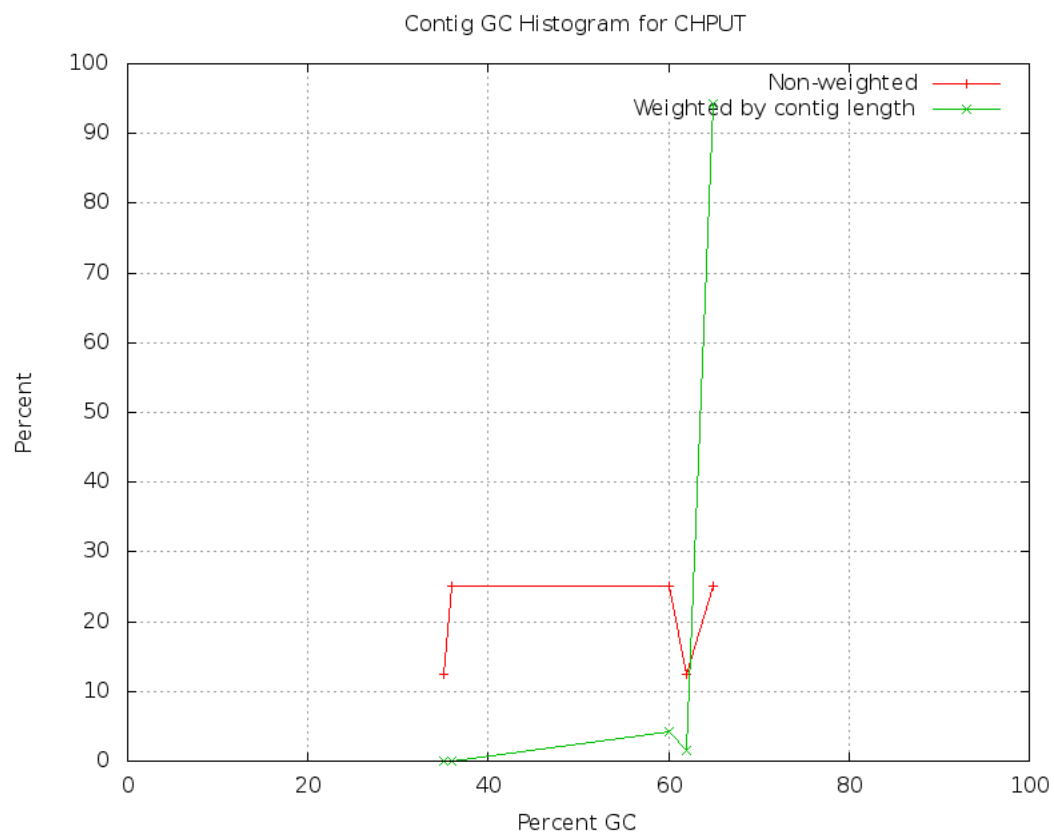




Coverage and GC information. Scaffolds were shredded into non-overlapping 5 kbp fragments and the GC content of each shred was plotted as a data point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.



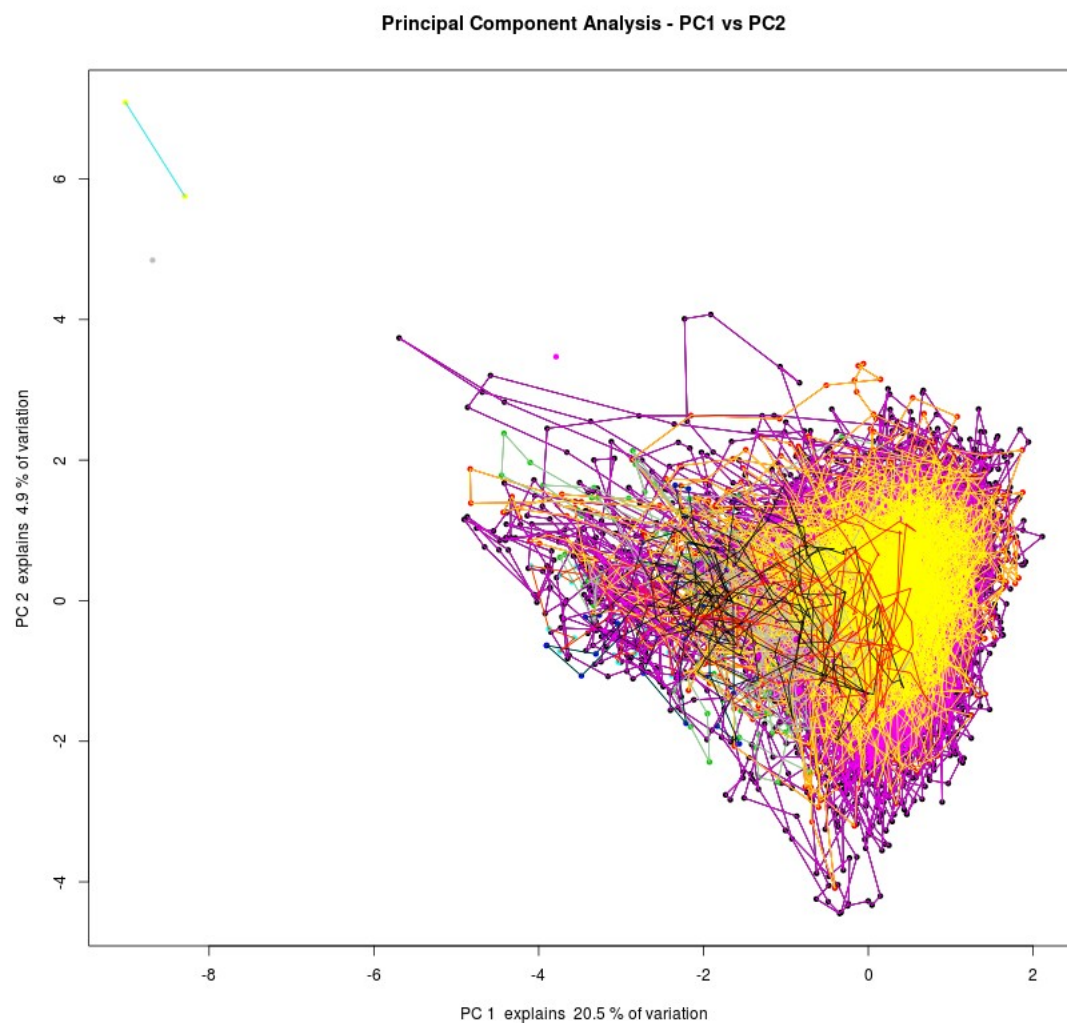
GC histogram of the scaffolds, including scaffold length weighted distribution.



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

No hits to the reference.

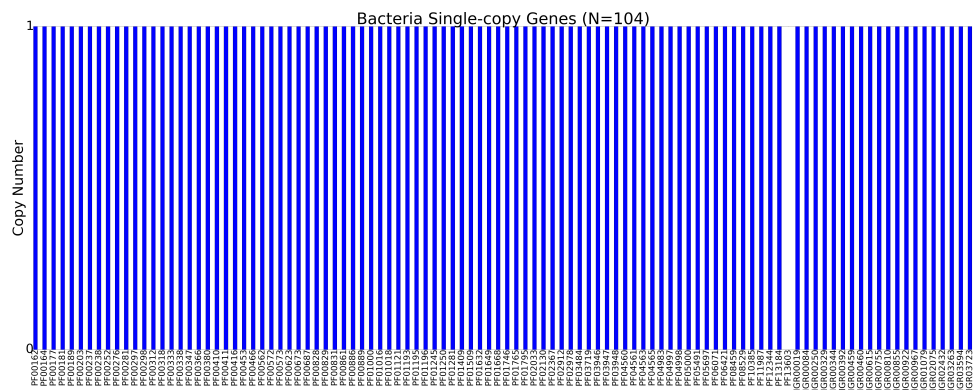
Tetramer frequencies are calculated over 5kb sliding windows of all scaffolds, followed by principal component analysis. Plots of the first two principal components are colored by scaffold.



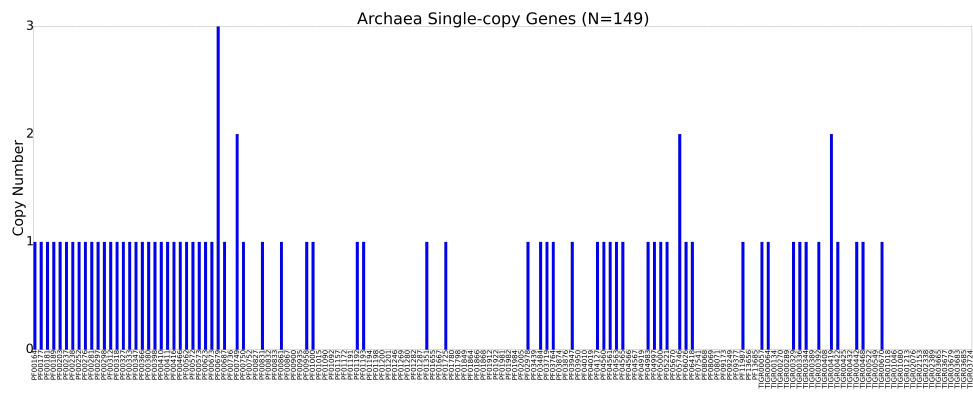
Estimated genome recovery derived from analysis of universal single-copy genes detected in final assembly using CheckM.

HMM	Found Genes	Total Genes	Percent Recovered
Archaea	70	149	46.98%
Bacteria	103	104	99.04%
Lineage Workflow	429	430	99.77%

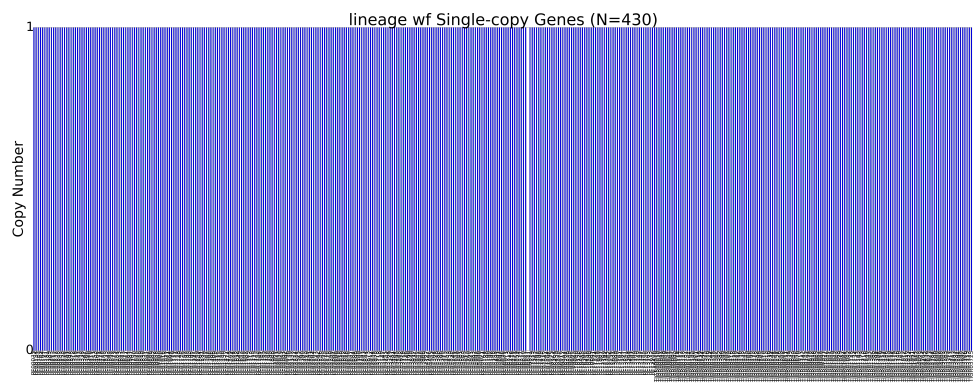
Bacteria Single-copy Gene Histogram

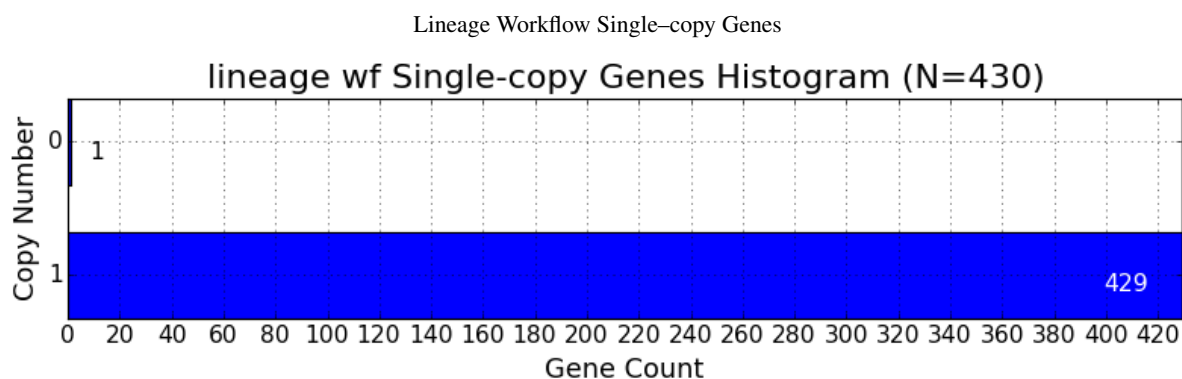
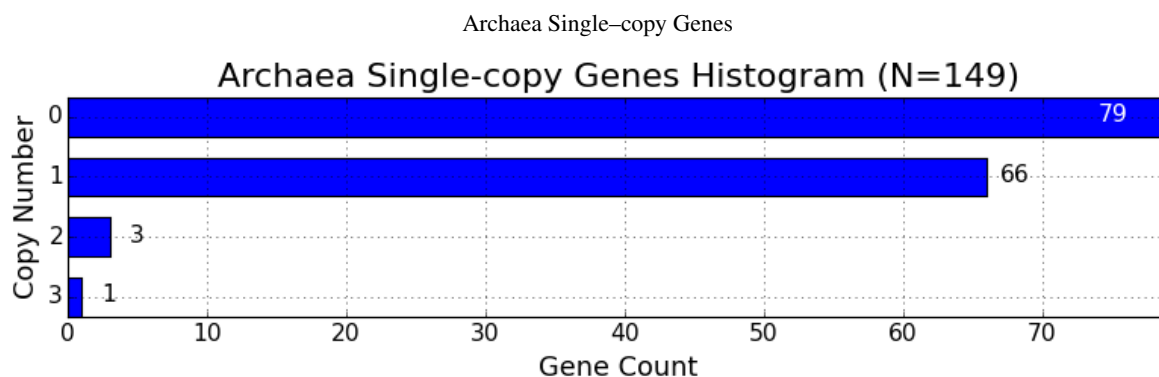
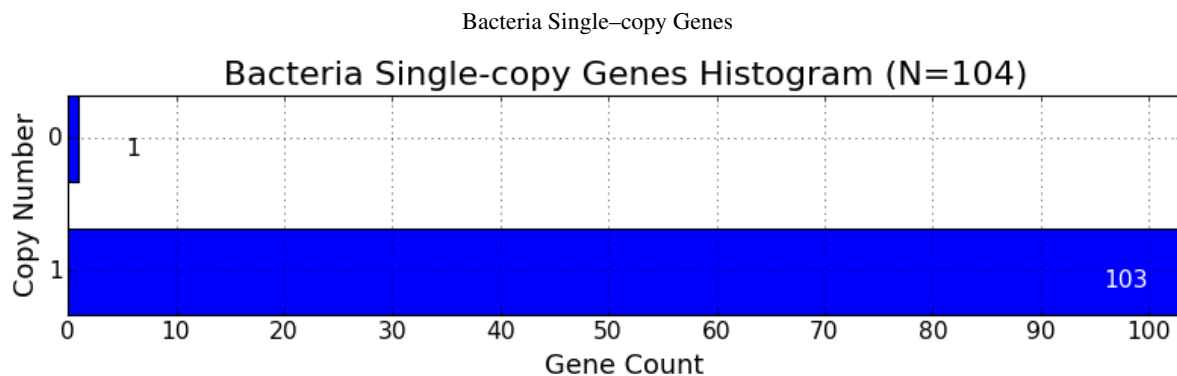


Archaea Single-copy Gene Histogram



Lineage Workflow Single-copy Gene Histogram





5. Sequence Data Availability

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

6. Methods

Isolate Improved Draft

Genome Sequencing and Assembly

The draft genome of *Shinella zoogloeoides* was generated at the DOE Joint Genome Institute (JGI) using the Pacific Biosciences (PacBio) sequencing technology [1]. A >10kbp Pacbio SMRTbell™ library was constructed and sequenced on the PacBio RS2 platform, which generated 146,523 filtered subreads totaling 877,424,928 bp. All general aspects of library construction and sequencing performed at the JGI can be found at <http://www.jgi.doe.gov>. The raw reads were assembled using HGAP (smrtanalysis/2.3.0_p5, HGAP 3) [2]. The final draft assembly contained 8 contigs in 8 scaffolds, totaling 5,680,044 bp in size. The input read coverage was 119.7X.

1. Eid John, et al. RealDNA Sequencing from Single Polymerase Molecules. Science 2008
2. Chin C, et al. Nonhybrid, finished microbial genome assemblies from longread SMRT sequencing data. Nat Methods 2013

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 #1717.
