1. Project Information

Program	Microbial/BRC–CBI 2018
Sequencing Project ID	1189086
Sequencing Project Name	Caloramator sp. E03

2. Read Statistics

Metric	Raw Reads	Filtered Subreads	Error Corrected Reads
Reads	449,293	284,615	11,044
Bases	1,422,368,585	664,812,488	43,103,006
Average Read Length	$3,165.8 \pm 2,854.5$	$2,335.8 \pm 1,384.3$	$3,902.8 \pm 2,205.9$
Reads >5kbp	73,243	12,990	3,199
Bases, reads >5 kbp	603,028,689	89,422,961	20,920,803
Avg Read Length, reads >5kbp	$8,233.3 \pm 3,491.7$	$6,884.0 \pm 1,980.3$	$6,539.8 \pm 1,646.9$

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)	3,024,286
Contig sequence length (bp)	3,024,286
Scaffold N/L50	2/731679
Contig N/L50	2/731679
Largest Contig (bp)	901,806
Number of scaffolds >50 kb	8
Percent of genome in scaffolds >50 kb	100.00%
Percent of reads assembled	80.76%

4. Assembly Quality Assessment

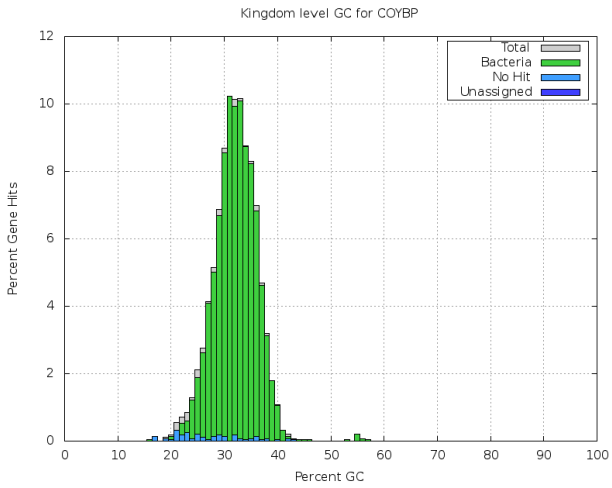
The quality of the final assembly was determined by using tRNAscan-SE to count tRNAs, barrnap to determine the presence of the 5S, 16S and 23S genes and checkm to determine completeness and contamination.

5S	Yes
16S	Yes
23S	Yes
tRNA Count	24
Completeness	99.19%
Contamination	3.23%
Quality	High Quality

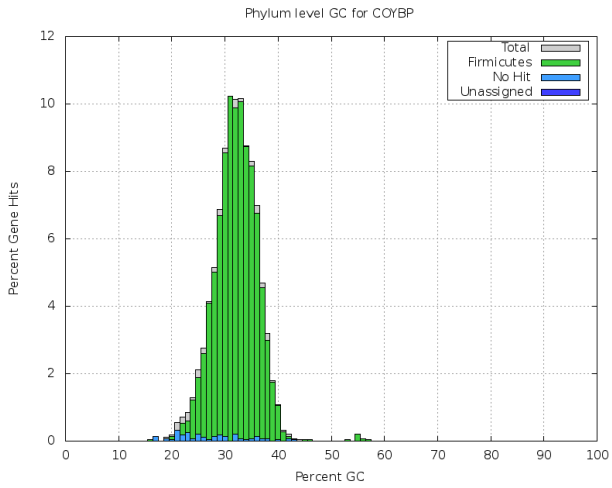
5. Assembly QC Results

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.

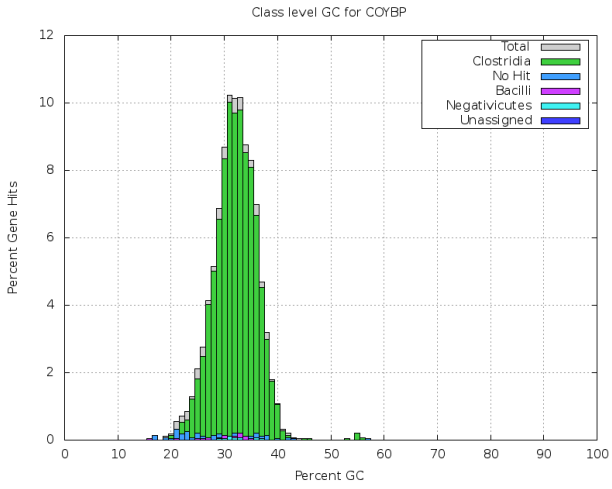
Kingdom



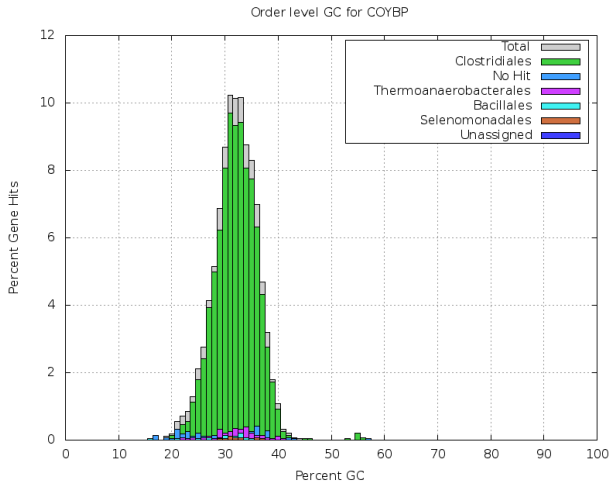
Phylum



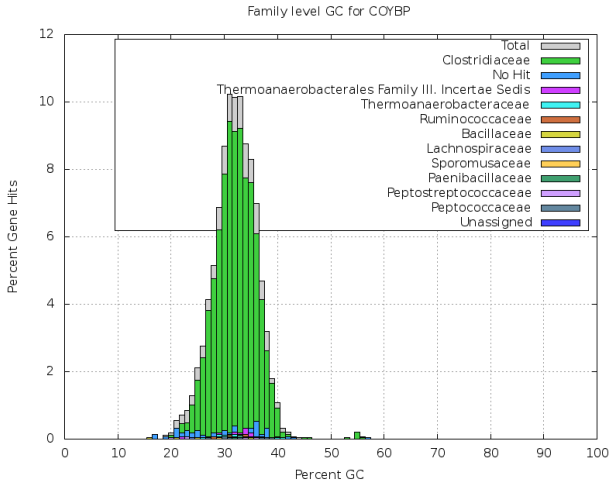
Class



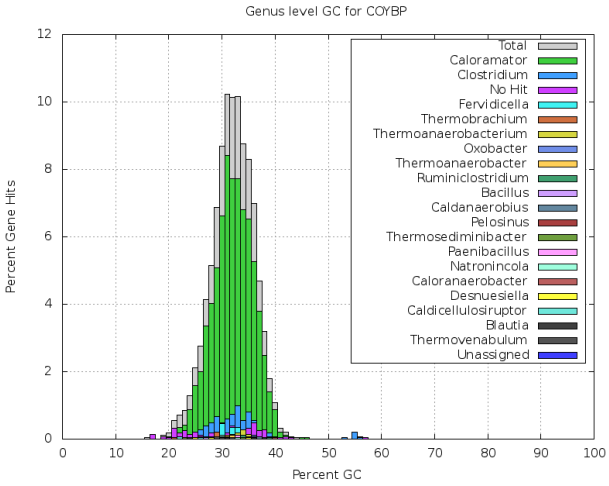
Order



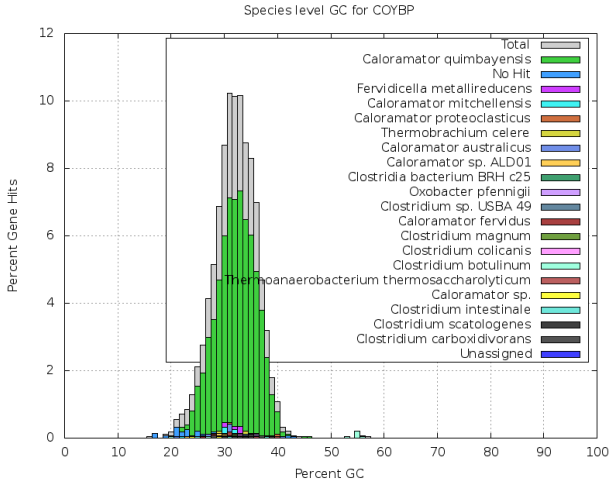
Family



Genus

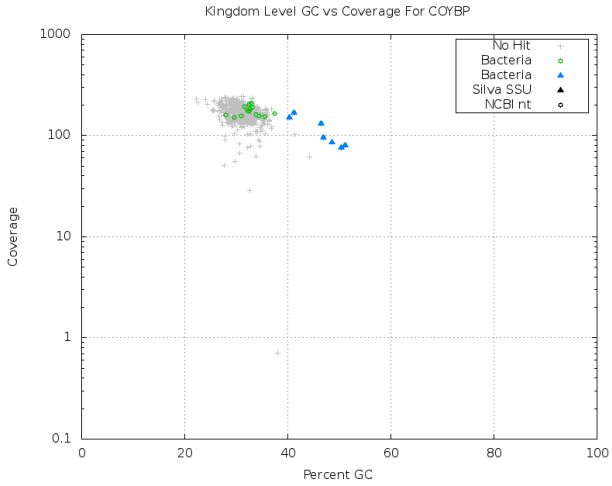


Species

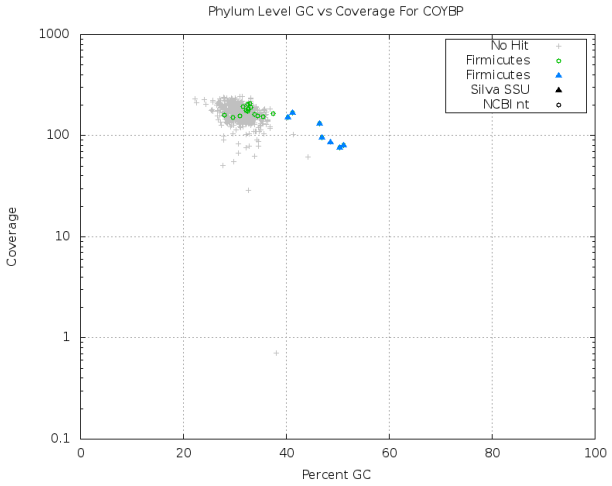


GC verses coverage of assembled scaffolds, overlaid with Silva SSU gene hits and NCBI nt megablast hits shown for different taxonomic levels.

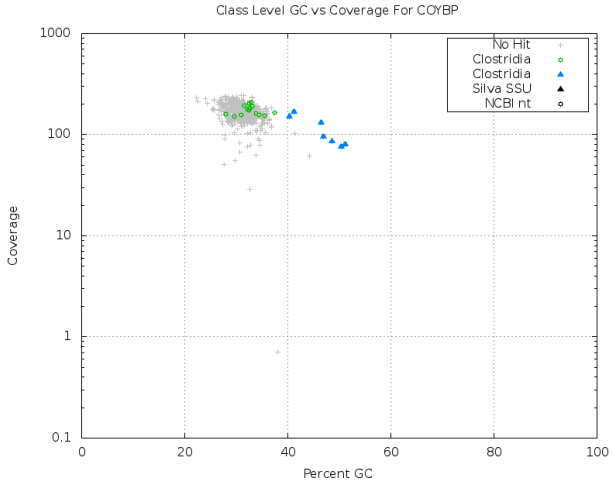
Kingdom



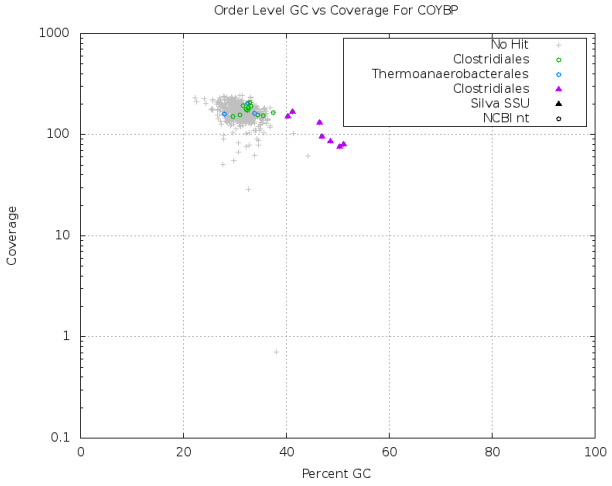
Phylum



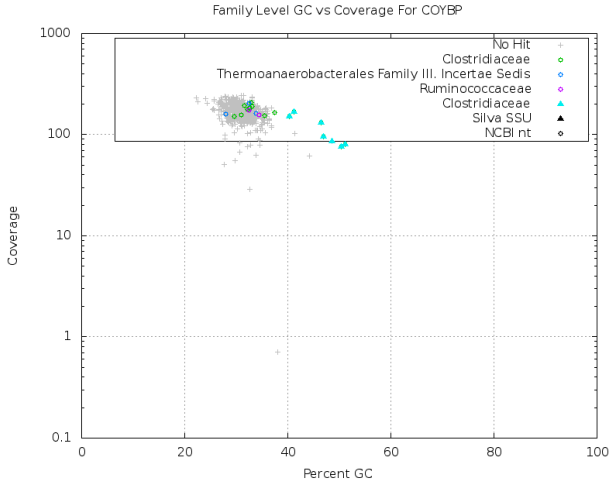
Class



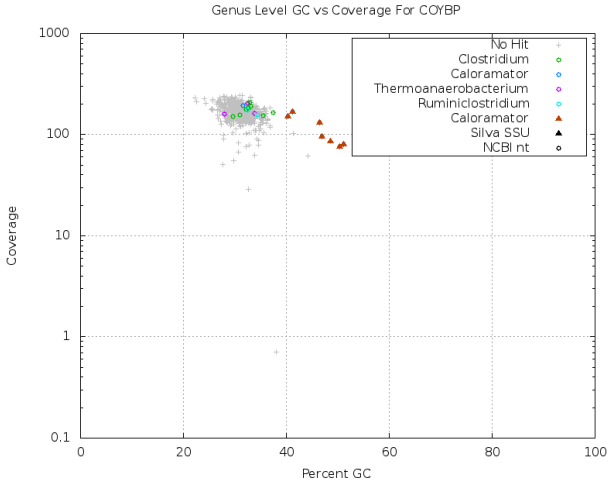
Order



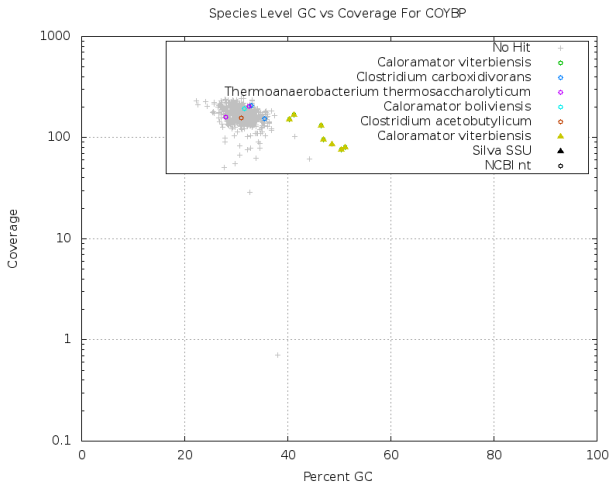
Family



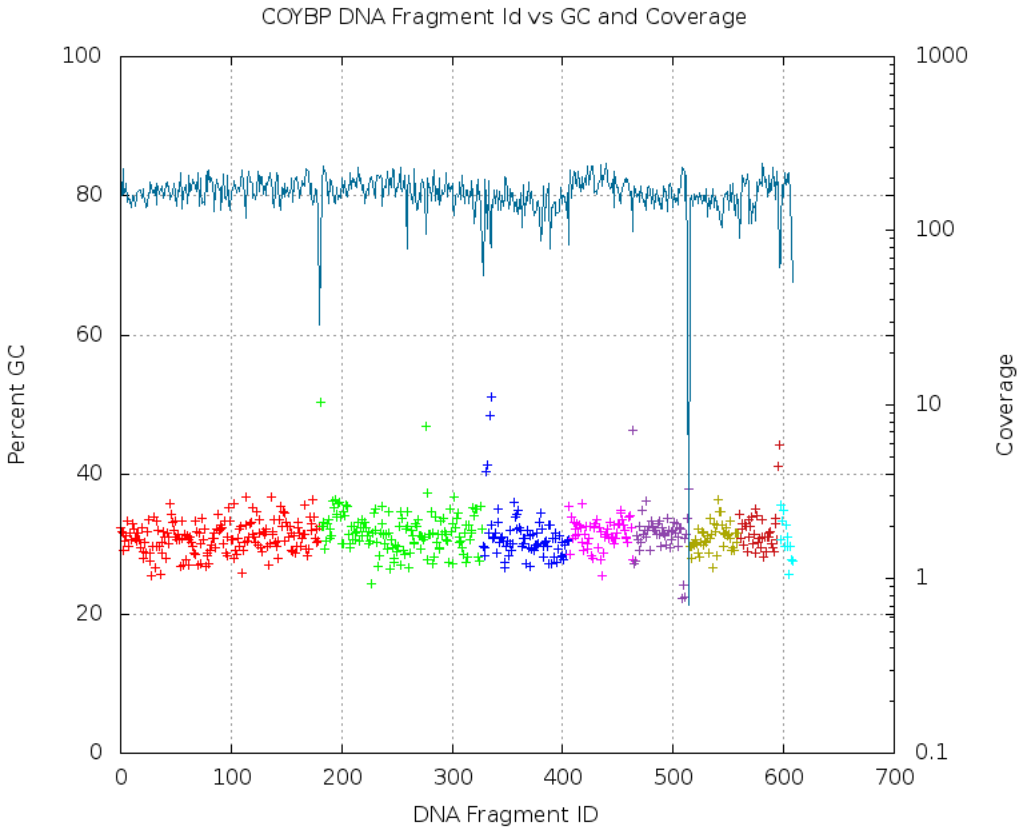
Genus



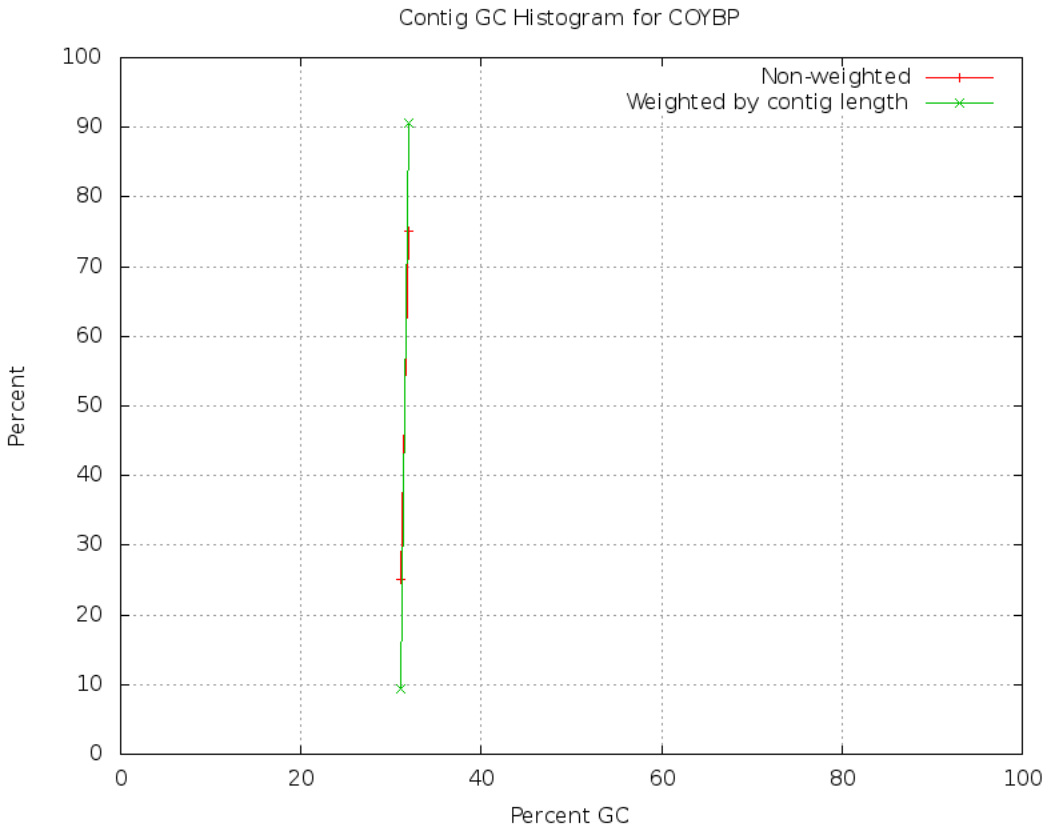
Species



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;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Caloramator;Caloramator viterbiensis;
Contig Name: COYBP_unitig_35|arrow
Align Length: 1,484 bp
Percent Id: 99.39%

Organism: N/ABacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Caloramator;Caloramator viterbiensis;
Contig Name: COYBP_unitig_29|arrow
Align Length: 1,484 bp
Percent Id: 99.33%

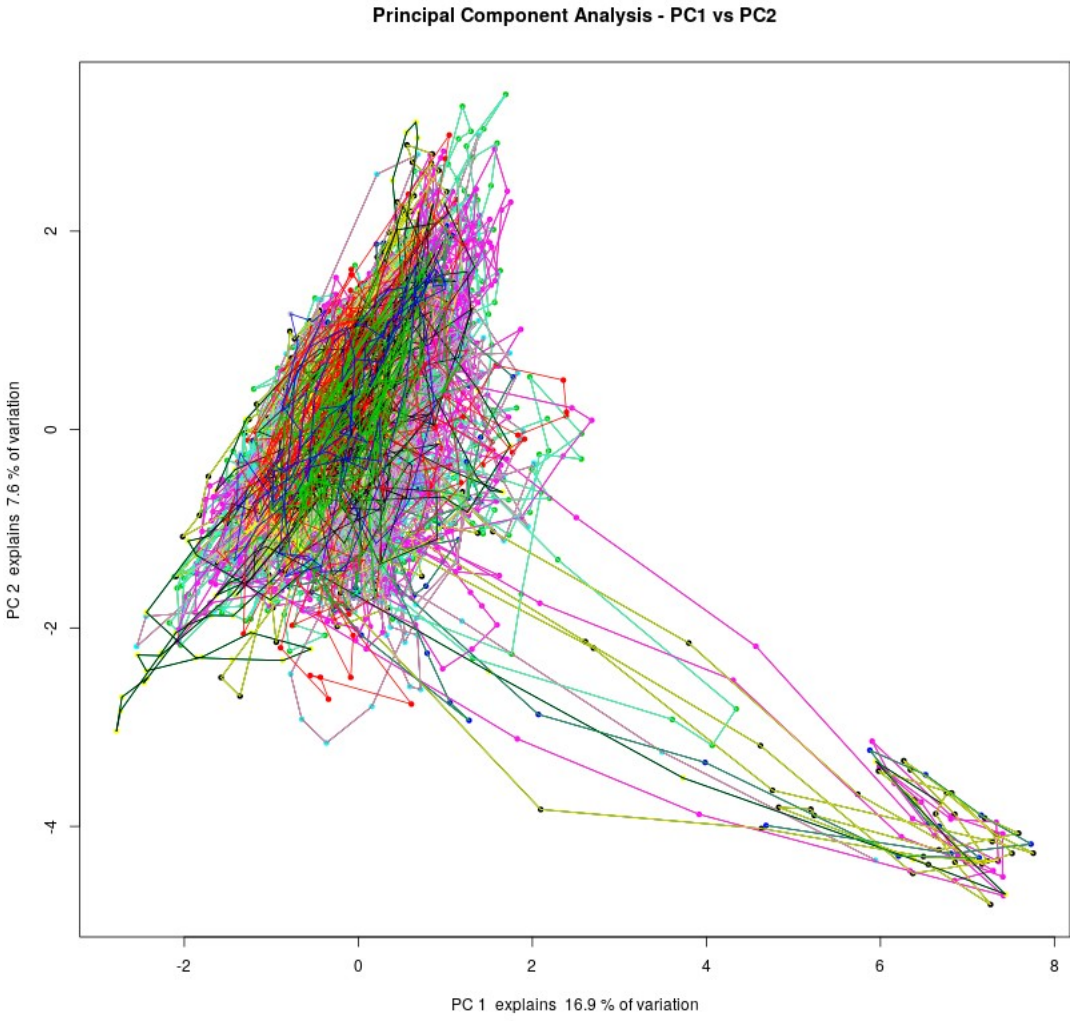
Organism: N/ABacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Caloramator;Caloramator viterbiensis;
Contig Name: COYBP_unitig_6|arrow
Align Length: 1,415 bp
Percent Id: 99.22%

Organism: N/ABacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Caloramator;Caloramator viterbiensis;

Contig Name: COYBP_unitig_5|arrow
Align Length: 1,486 bp
Percent Id: 99.12%

Organism: N/ABacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus pneumoniae;
Contig Name: COYBP_unitig_36|arrow
Align Length: 224 bp
Percent Id: 91.52%

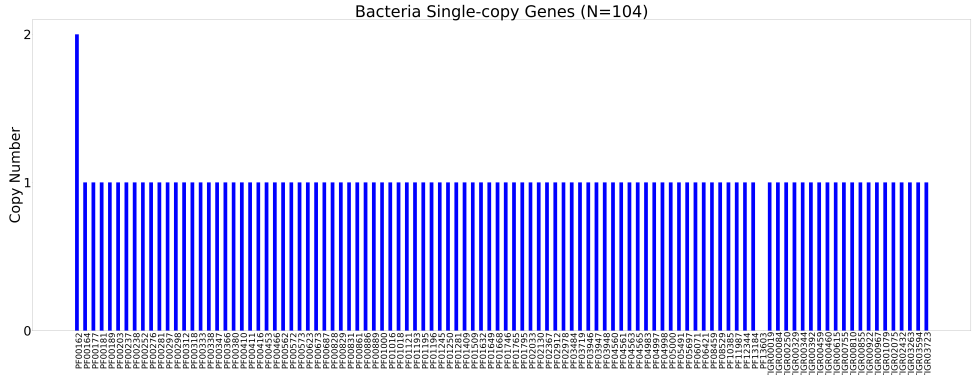
Tetramer frequencies are calculated over 5kb sliding windows of all scaffolds, followed by principal component analysis. Plots of the first two principal components are colored by scaffold.



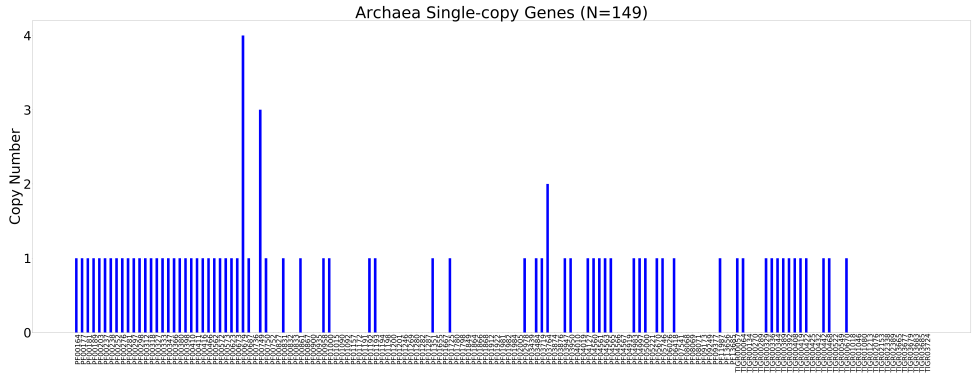
Estimated genome recovery derived from analysis of universal single-copy genes detected in final assembly using CheckM.

HMM	Found Genes	Total Genes	Percent Recovered
Archaea	72	149	48.32%
Bacteria	103	104	99.04%
Lineage Workflow	222	223	99.55%

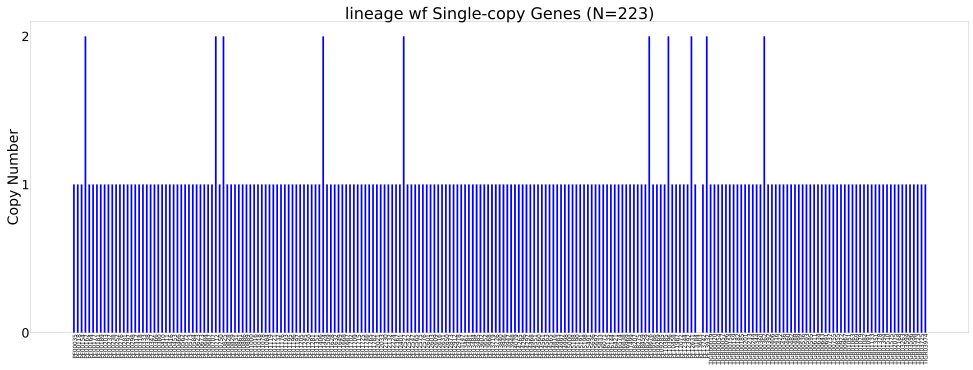
Bacteria Single–copy Gene Histogram



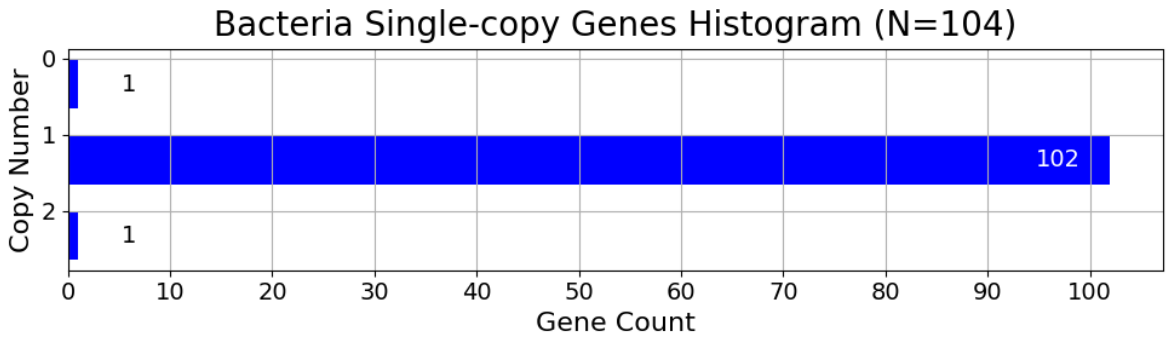
Archaea Single–copy Gene Histogram



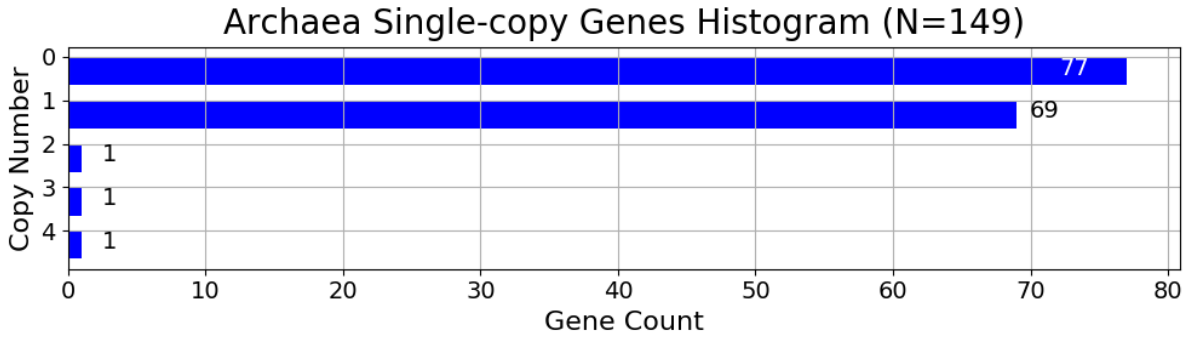
Lineage Workflow Single–copy Gene Histogram



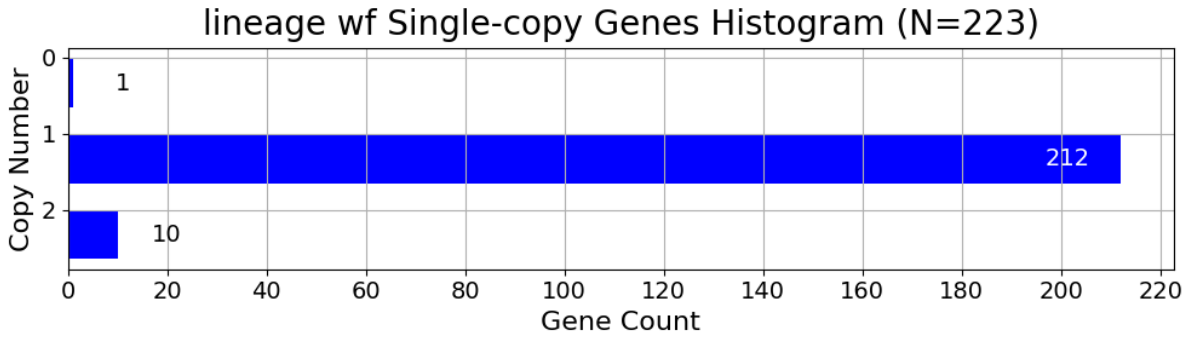
Bacteria Single–copy Genes



Archaea Single–copy Genes



Lineage Workflow Single–copy Genes



6. Sequence Data Availability

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

7. Methods

Isolate Improved Draft

Genome Sequencing and Assembly
The draft genome of *Caloramator sp.* was generated at the DOE Joint Genome Institute (JGI) using the Pacific Biosciences (PacBio) sequencing technology [1]. A >10kpb Pacbio SMRTbell™ library was constructed and sequenced on the PacBio RS2 platform, which generated 284,615 filtered subreads totaling 664,812,488 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 3,024,286 bp in size. The input read coverage was 167.1X.

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 #503761.