

1. Project Information

Program	Microbial/BRC–BESC 2014
Sequencing Project ID	1176338
Sequencing Project Name	Pseudomonas sp. JH0040

2. Read Statistics

Metric	Raw Reads	Filtered Subreads	Error Corrected Reads
Reads	207,688	100,798	11,334
Bases	1,168,533,286	640,110,113	80,211,166
Average Read Length	5,626.4 ± 6,419.9	6,350.4 ± 4,233.1	7,077.0 ± 5,740.7
Reads >5kbp	82,457	52,181	5,961
Bases, reads >5 kbp	1,002,656,021	502,485,452	72,672,249
Avg Read Length, reads >5kbp	12,159.7 ± 5,439.2	9,629.7 ± 3,307.6	12,191.3 ± 2,534.2

3. Assembly Statistics

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

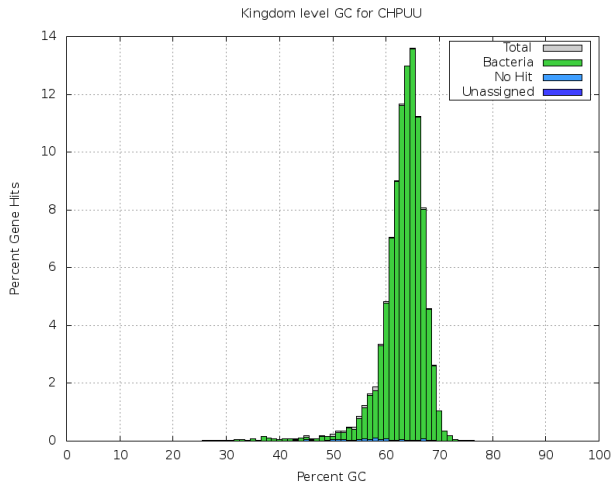
Assembly version: smrtanalysis/2.3.0_p5, HGAP 3

Scaffold total	374
Contig total	374
Scaffold sequence length (bp)	6,225,173
Contig sequence length (bp)	6,225,173
Scaffold N/L50	1/4437513
Contig N/L50	1/4437513
Largest Contig (bp)	4,437,513
Number of scaffolds >50 kb	1
Percent of genome in scaffolds >50 kb	71.28%
Percent of reads assembled	69.65%

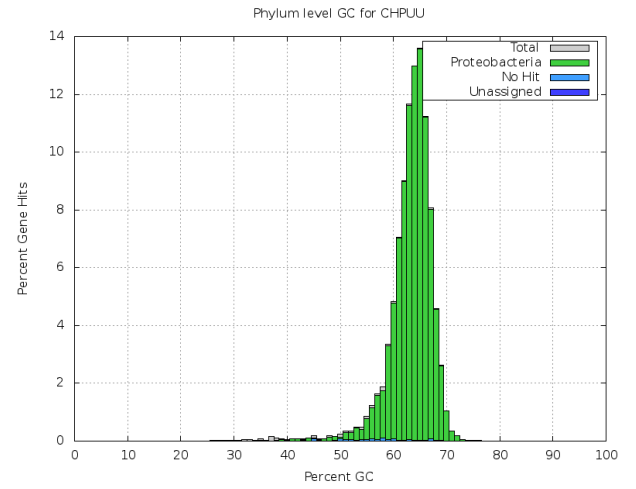
4. Assembly QC Results

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.

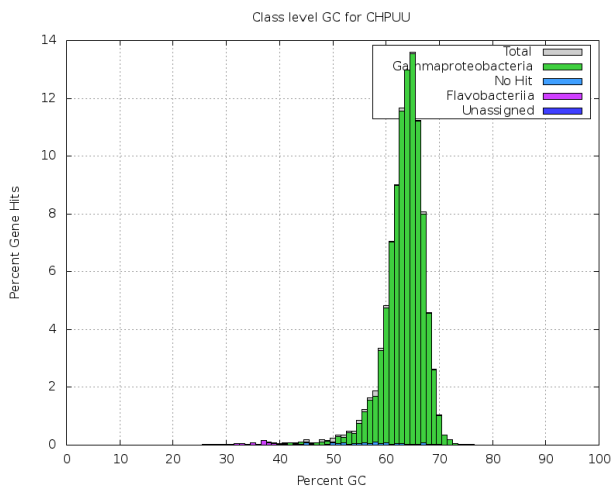
Kingdom



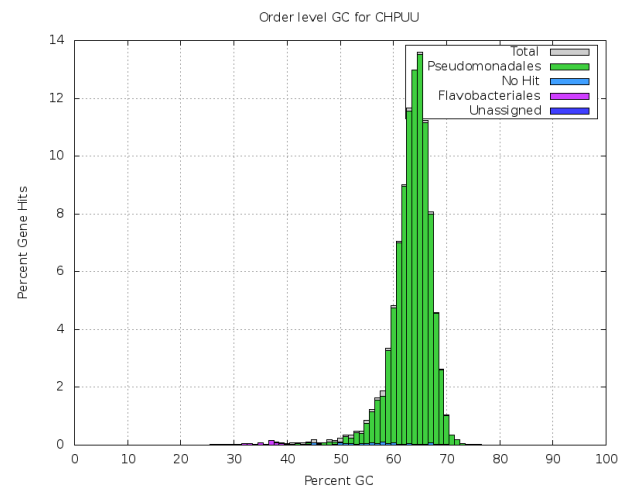
Phylum



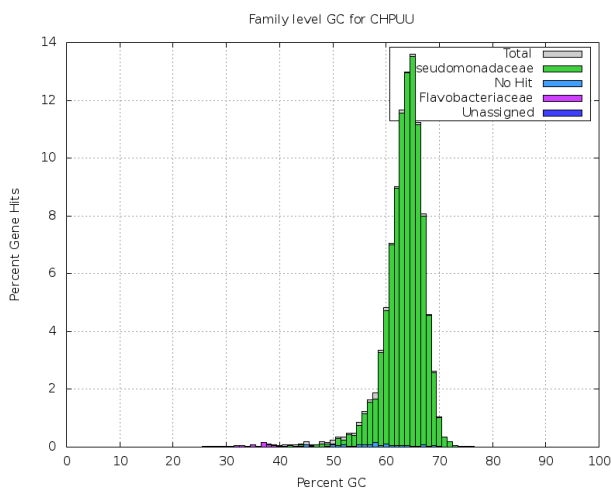
Class



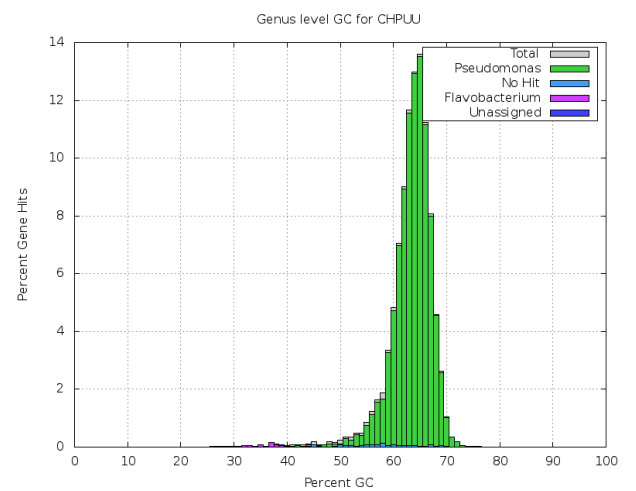
Order

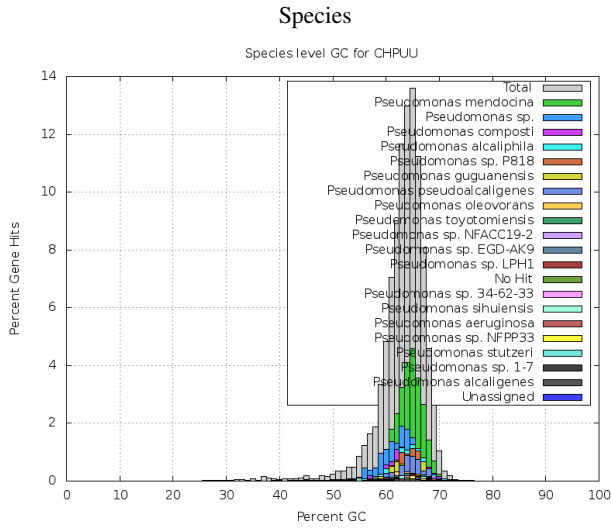


Family

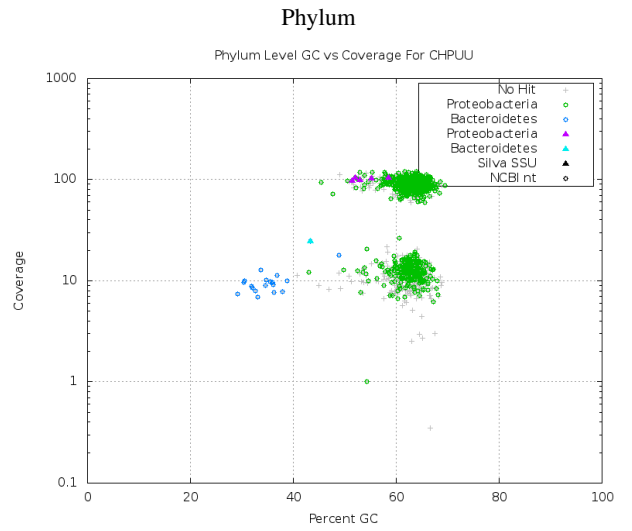
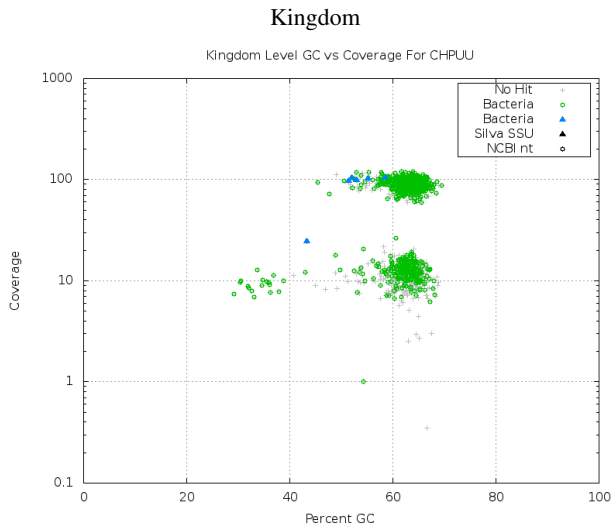


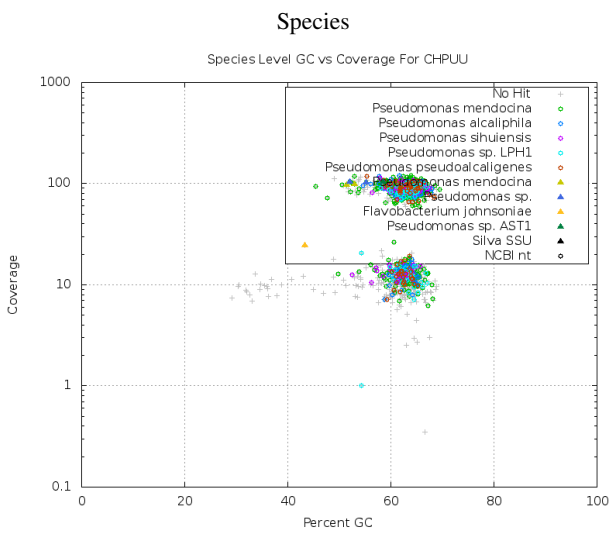
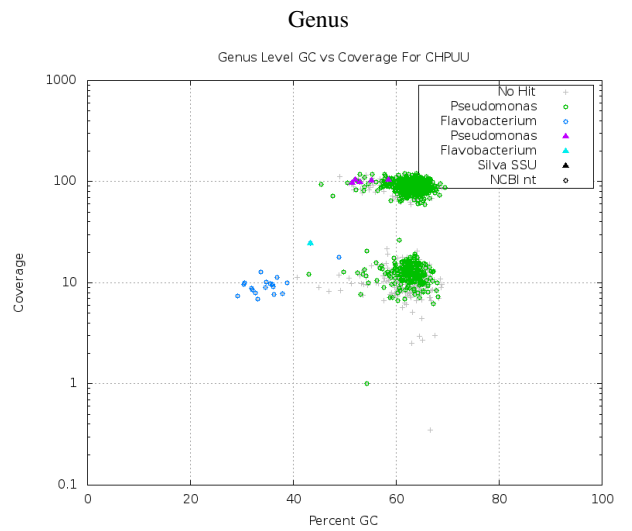
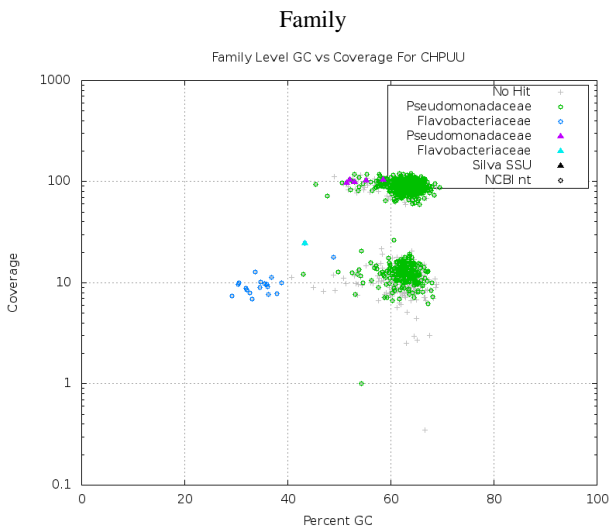
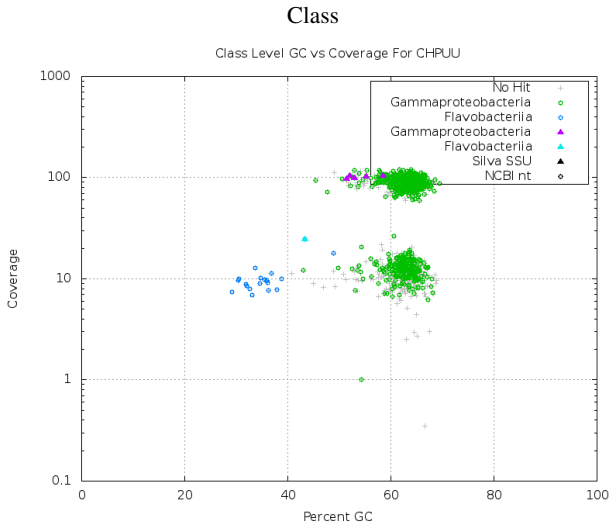
Genus



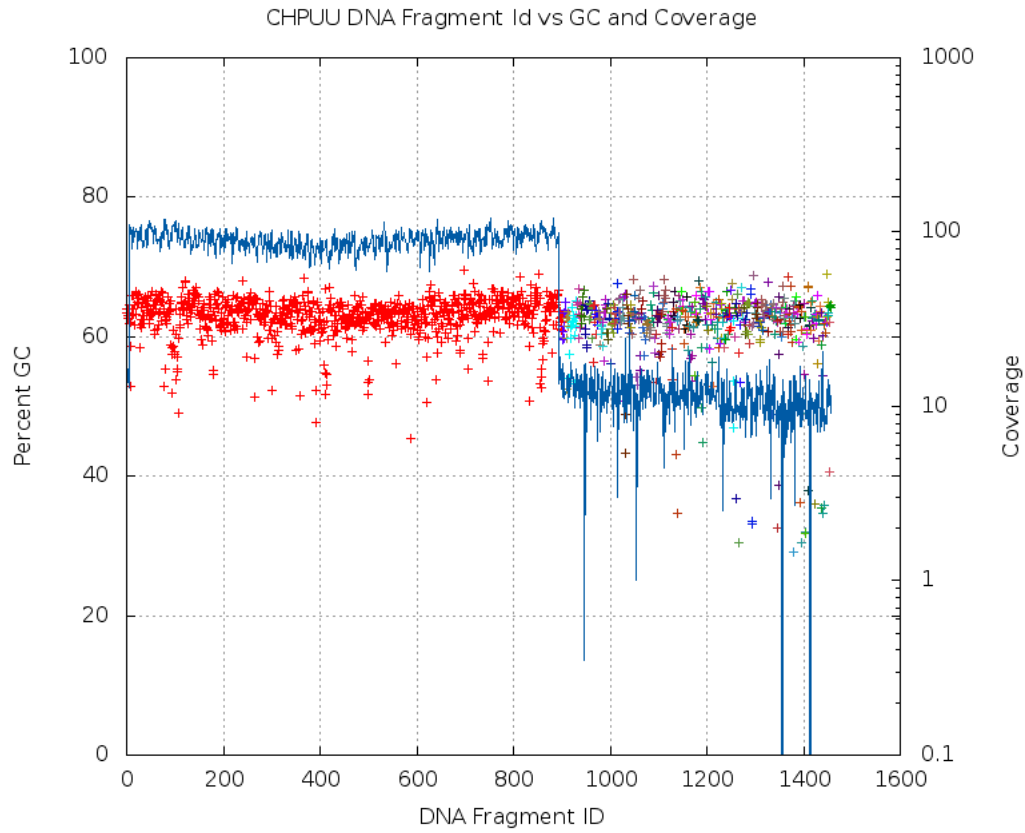


GC versus coverage of assembled scaffolds, overlaid with GreenGenes 16s rRNA gene hits and NCBI nt megablast hits shown for different taxonomic levels.

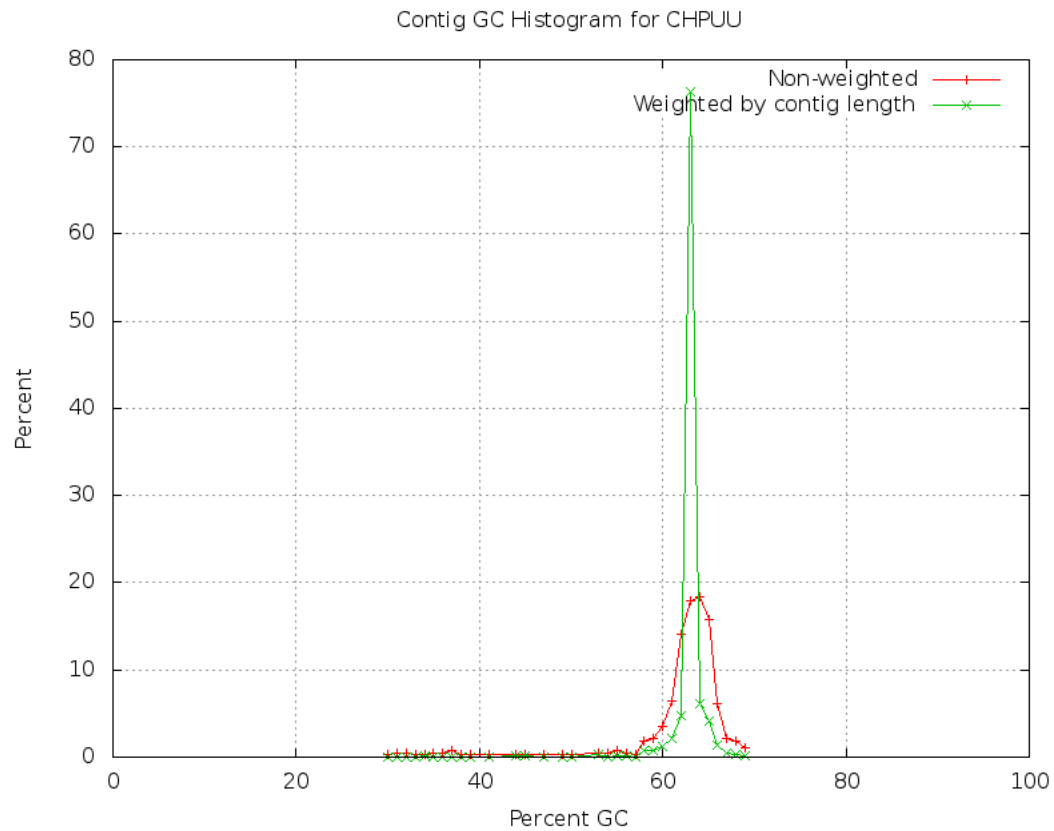




Coverage and GC information. Scaffolds were shredded into non-overlapping 5 kbp fragments and the GC content of each shred was plotted as a data point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.



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

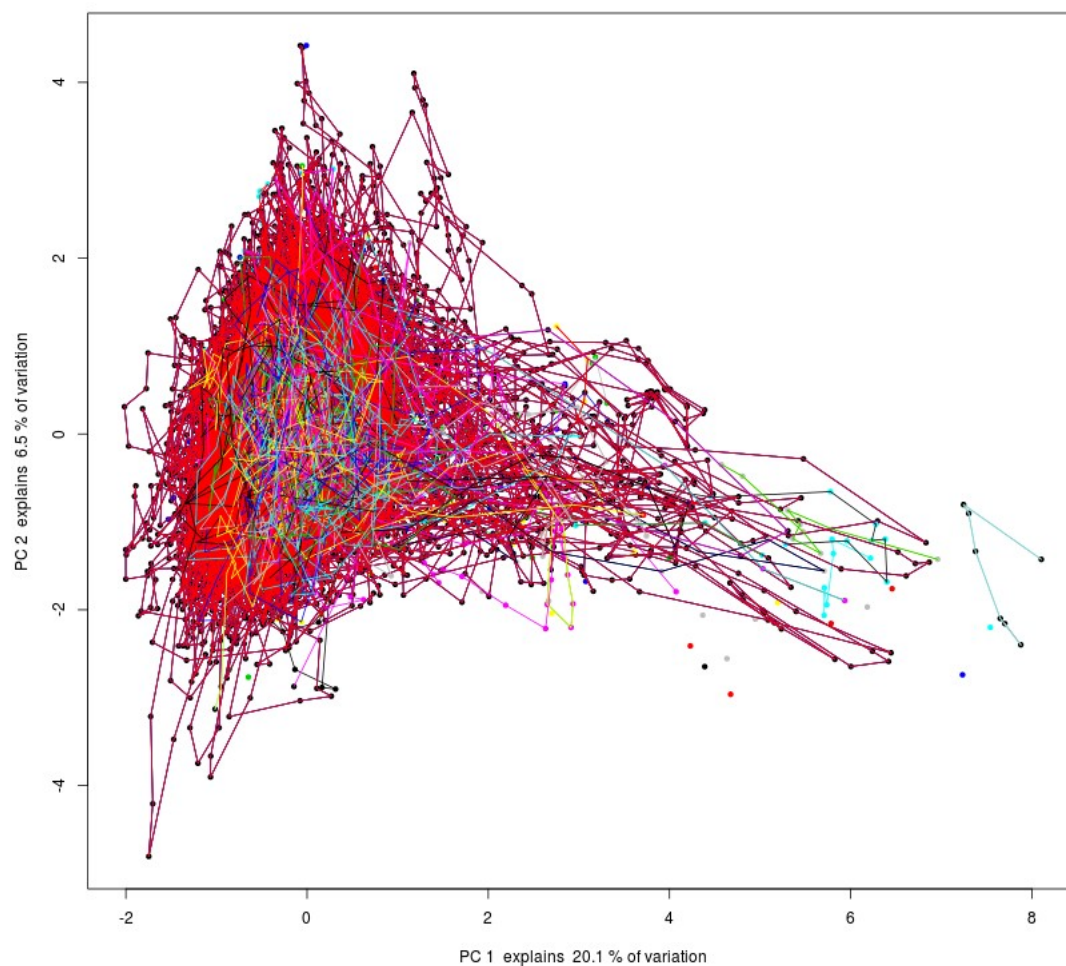


List of the top scaffold megablast hits against 16s ribosomal genes using the GreenGenes database.

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.

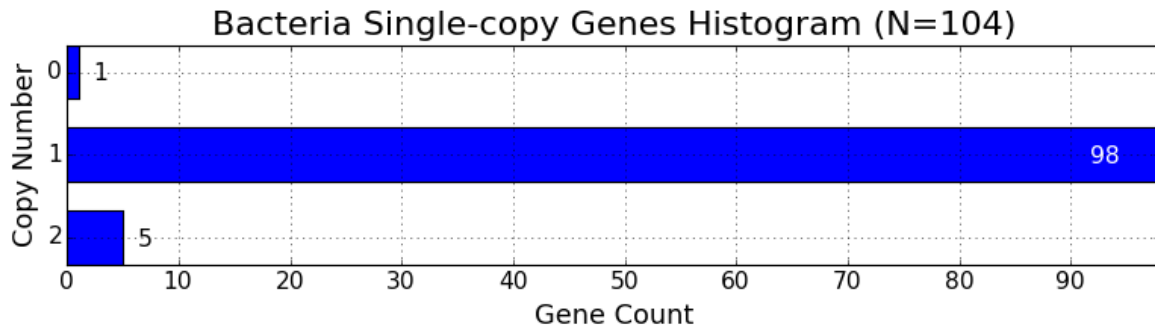
Principal Component Analysis - PC1 vs PC2



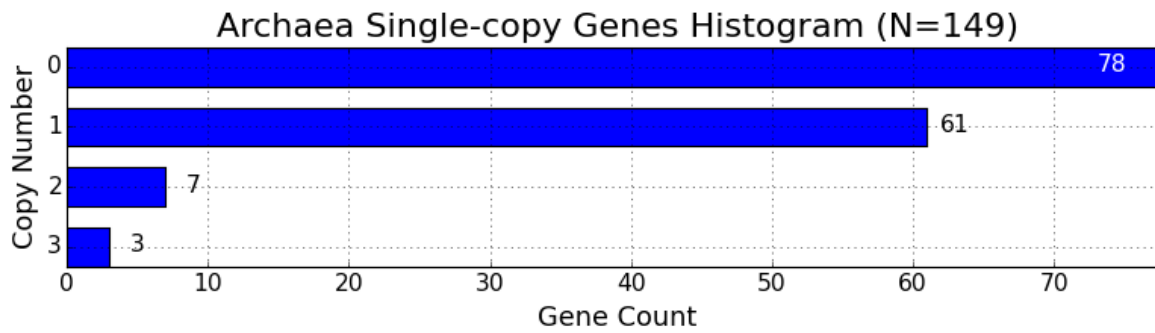
Estimated genome recovery derived from analysis of universal single-copy genes detected in final assembly using CheckM.

HMM	Found Genes	Total Genes	Percent Recovered
Archaea	71	149	47.65%
Bacteria	103	104	99.04%
Lineage Workflow	805	813	99.02%

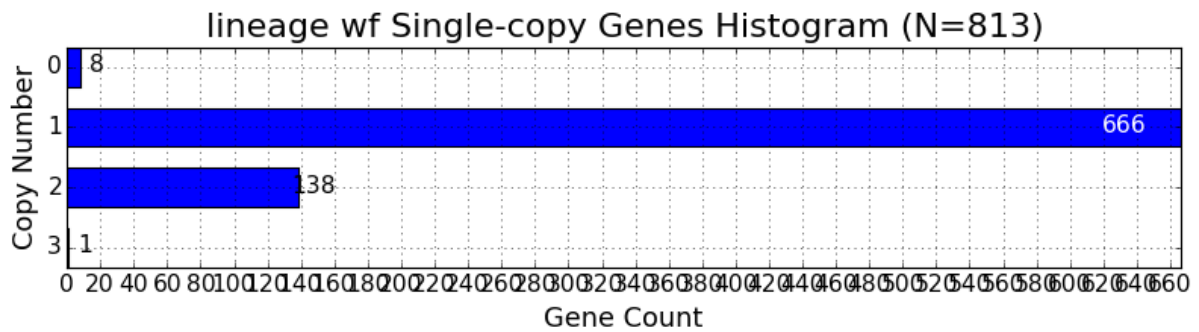
Bacteria Single-copy Genes



Archaea Single-copy Genes



Lineage Workflow Single-copy Genes



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 *Pseudomonas sp.* 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 100,798 filtered subreads totaling 640,110,113 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 374 contigs in 374 scaffolds, totaling 6,225,173 bp in size. The input read coverage was 59.4X.

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.
