

1. Project Information

Program	Microbial/CSP 2012
PMO Project	0
Seq Proj ID	1027046
Sequencing Project Name	Thaumarchaeota archaeon JGI 000156CP-J13
JGI Project ID	0

2. Read Statistics

Illumina Std PE Statistics

File name	7667.5.80864.TGACCA.fastq
Library	TGPY
Number of reads	24,895,556
Sequencing depth [†]	747X
Read type	2x150 bp

[†] A genome size of 5.0 Mbp was assumed in this calculation.

3. Read QC Results

The following are the results of reads screened against contaminants. Pairs of matching reads were removed from the dataset.

Illumina Std PE Read Filter Statistics

Description	Num Reads	Pct Reads
Input	24,895,556	100
Contam removed	80	0.0
Artifact removed	273,350	1.1
Total removed	4,895,556	19.7
Total remaining	20,000,000	80.3

List of Contaminants Removed

Description	Num Reads	Pct Reads
gi 357579577 Canis.lupus_familiaris_chr3	36	0.00
human_chr2	30	0.00
gi 357579535 Canis.lupus_familiaris_chr20	16	0.00
human_chr1	10	0.00
gi 357579505 Canis.lupus_familiaris_chromosome_X	4	0.00
gi 357579571 Canis.lupus_familiaris_chr5	4	0.00
gi 362110638 Felis.catus_breed_Abyssinian_chrD2	2	0.00

gi 357579515 Canis_lupus_familiaris_chr32	2	0.00
human_chr22	2	0.00
gi 357579522 Canis_lupus_familiaris_chr28	2	0.00
human_chr7	2	0.00
human_chr10	2	0.00
human_chr4	2	0.00

The following are the results of reads screened against potential reagent and process contaminants but were not removed from the dataset.

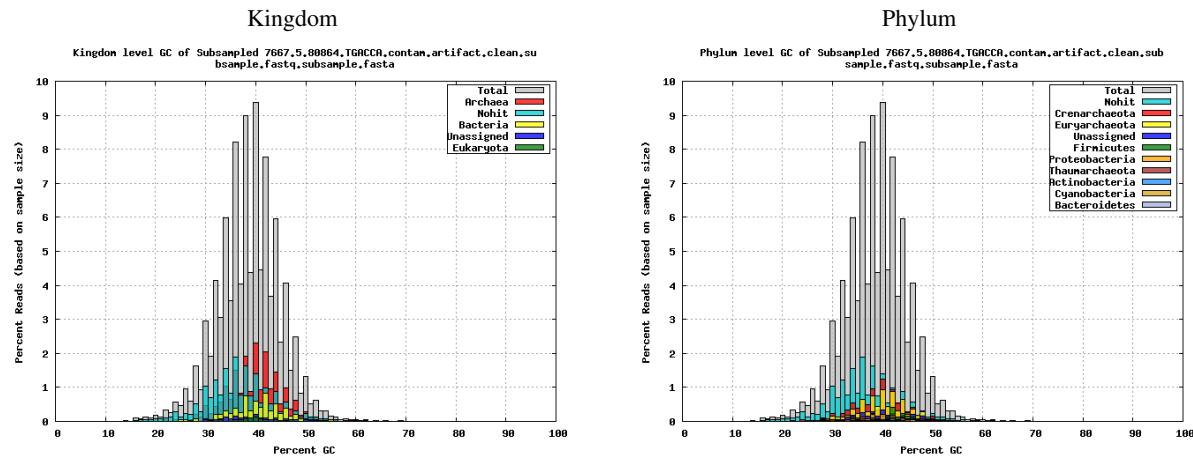
Illumina Std PE Contamination Identification Statistics

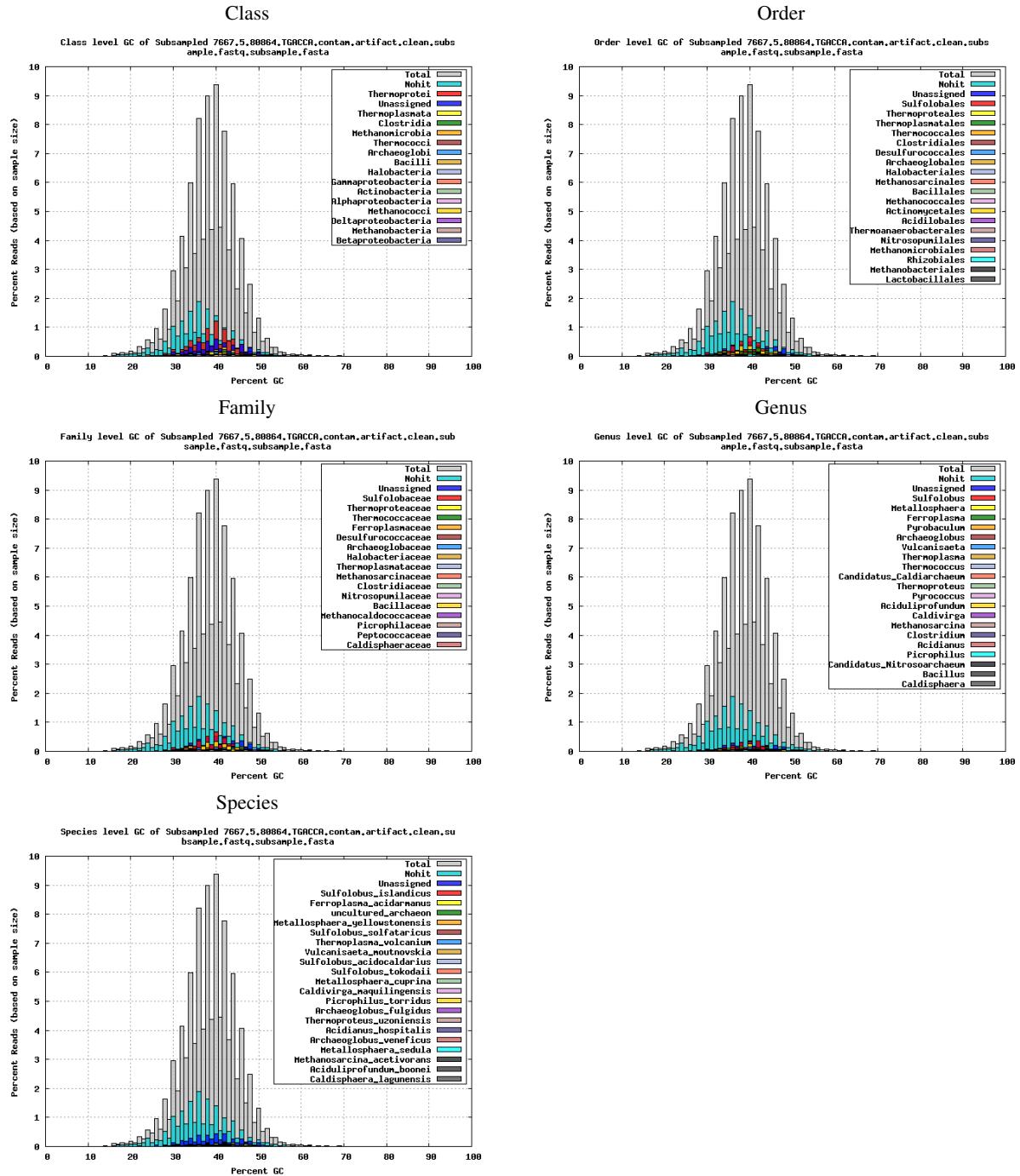
Description	Num Reads	Pct Reads
Input	24,895,556	100
Contam identified	14	0.0

List of Contaminants Identified

Description	Num Reads	Pct Reads
<i>Pseudomonas</i>	10	0.00
<i>Delftia</i>	2	0.00
<i>Shigella</i>	2	0.00

GC histogram of the reads subsampled to 10k, overlaid with GC of hits based on BLASTX, shown for different taxonomic levels.



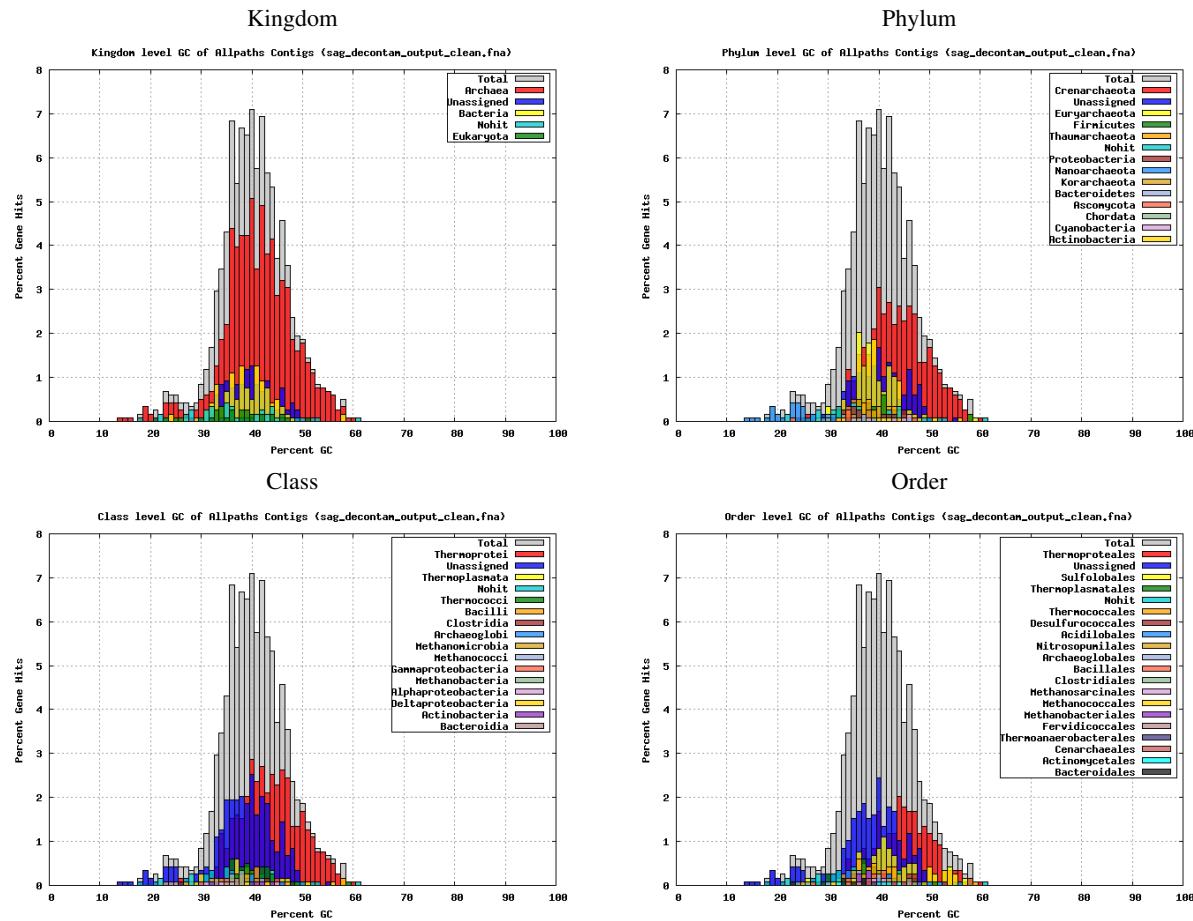


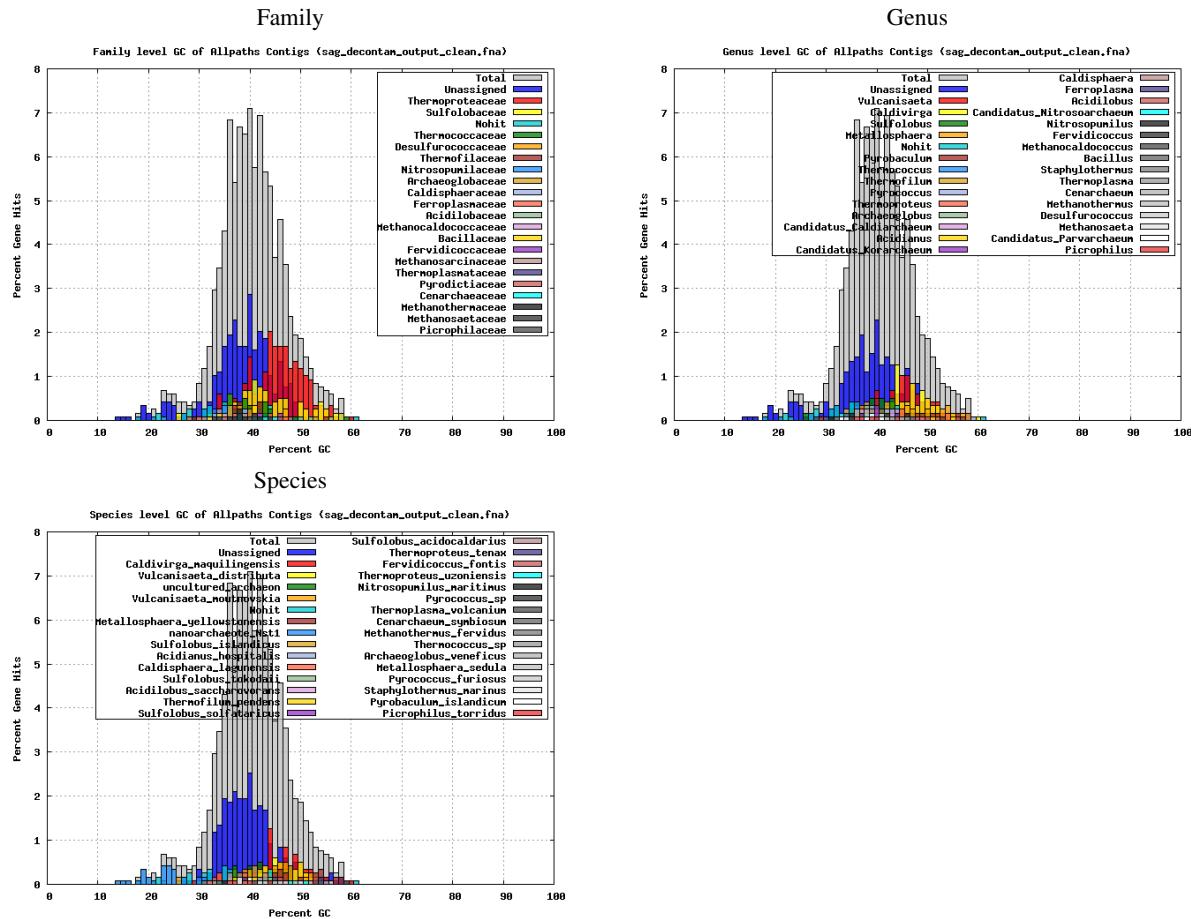
4. Assembly Statistics

Assembly method	SPAdes with auto decontamination
Scaffold total	101
Contig total	101
Scaffold sequence length	1.1 Mb
Contig sequence length	1.1 Mb (0.0% gap)
Scaffold N/L50	31/11.5 kb
Contig N/L50	31/11.5 kb
Largest Contig	34.1 kb
Number of scaffolds >50 kb	0
Pct of genome in scaffolds >50 kb	0.0
Pct of reads assembled (raw)	75.0
Pct of reads assembled (decontam)	62.2

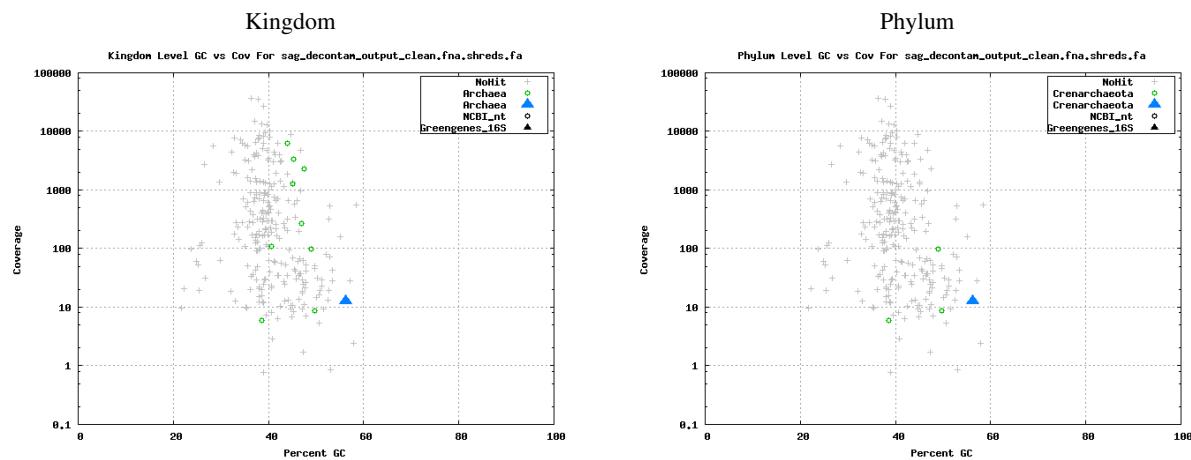
5. Assembly QC Results

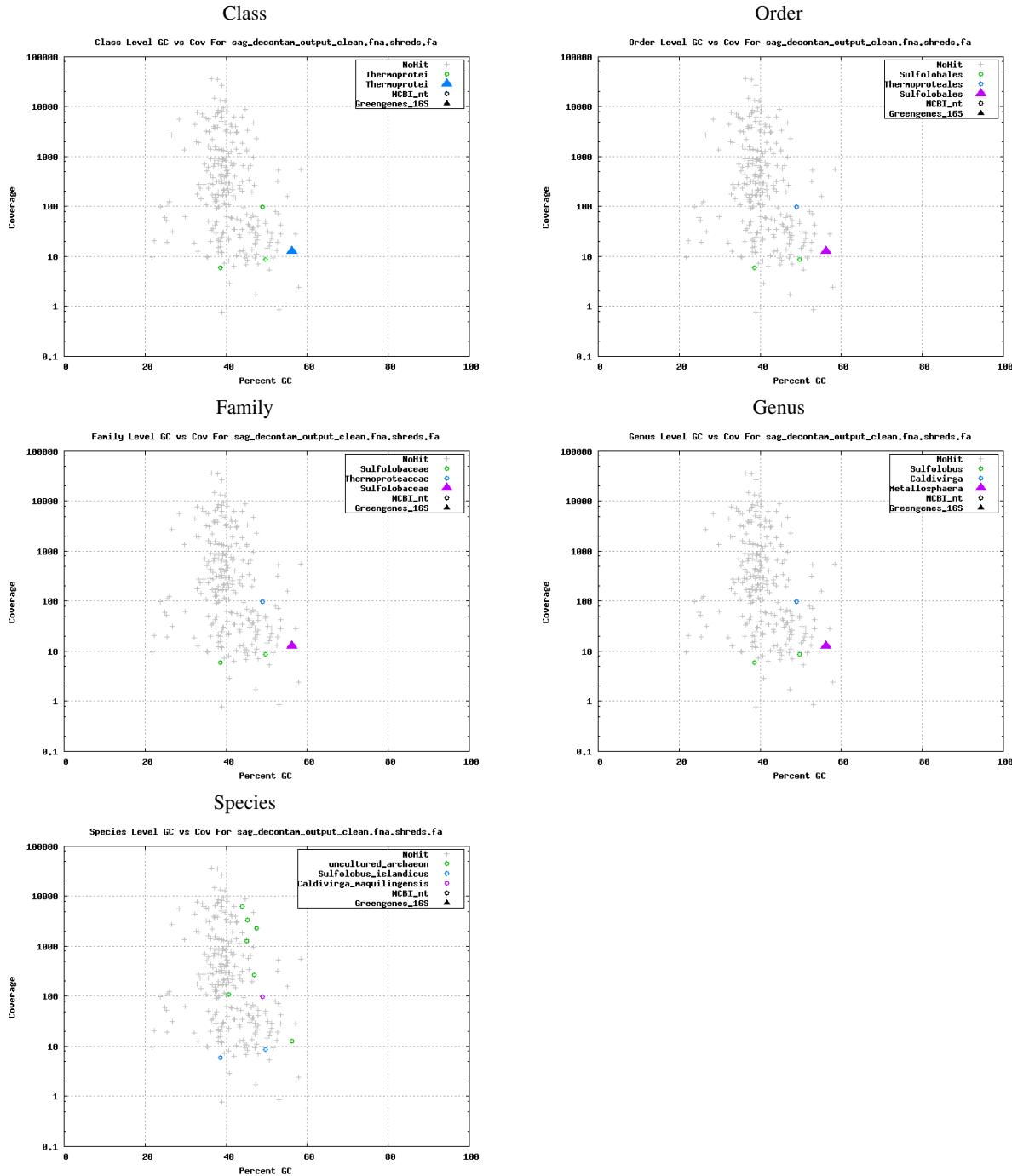
GC histogram of the predicted genes on each contig, overlaid with GC of hits based on BLASTP, shown for different taxonomic levels.



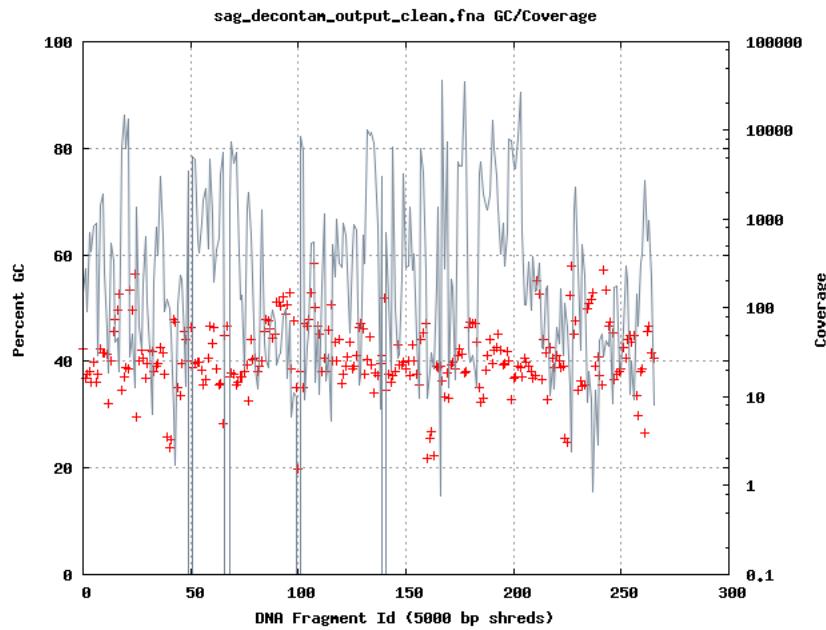


GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

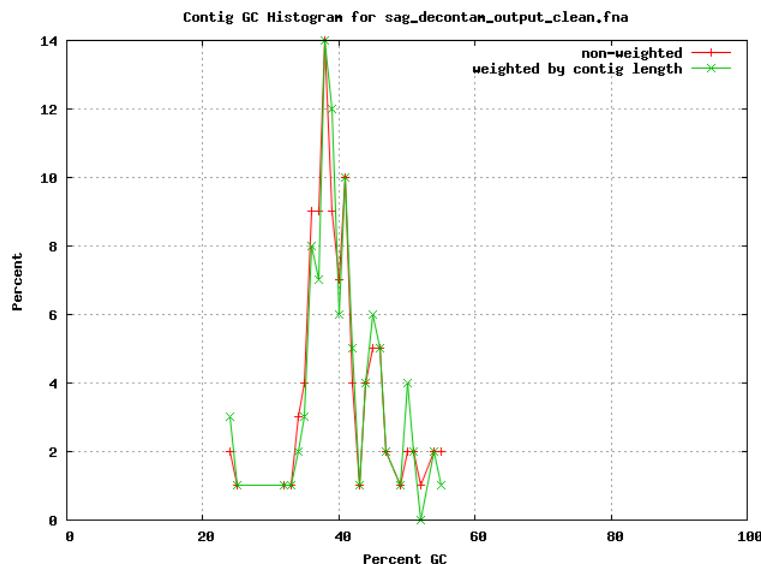




Coverage vs GC. Contigs were shredded into non-overlapping 5kbp and the GC of each shred was plotted as a point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.



GC histogram of the contigs, including contig length weighted distribution.



List of contigs and average percent GC, grouped in bins of 5:

Pct GC Bin	Contig Name
20	NODE_10.length_19158 cov_13.7202.ID_19, NODE_28.length_11889 cov_69.6179.ID_57
25	NODE_42.length_9811 cov_57.8797.ID_85
30	NODE_32.length_11396 cov_2443.64.ID_65, NODE_33.length_10993 cov_122.775.ID_67, NODE_72.length_6703 cov_134.943.ID_149, NODE_79.length_6393 cov_2815.86.ID_163 NODE_101.length_5091 cov_6.78693.ID_209

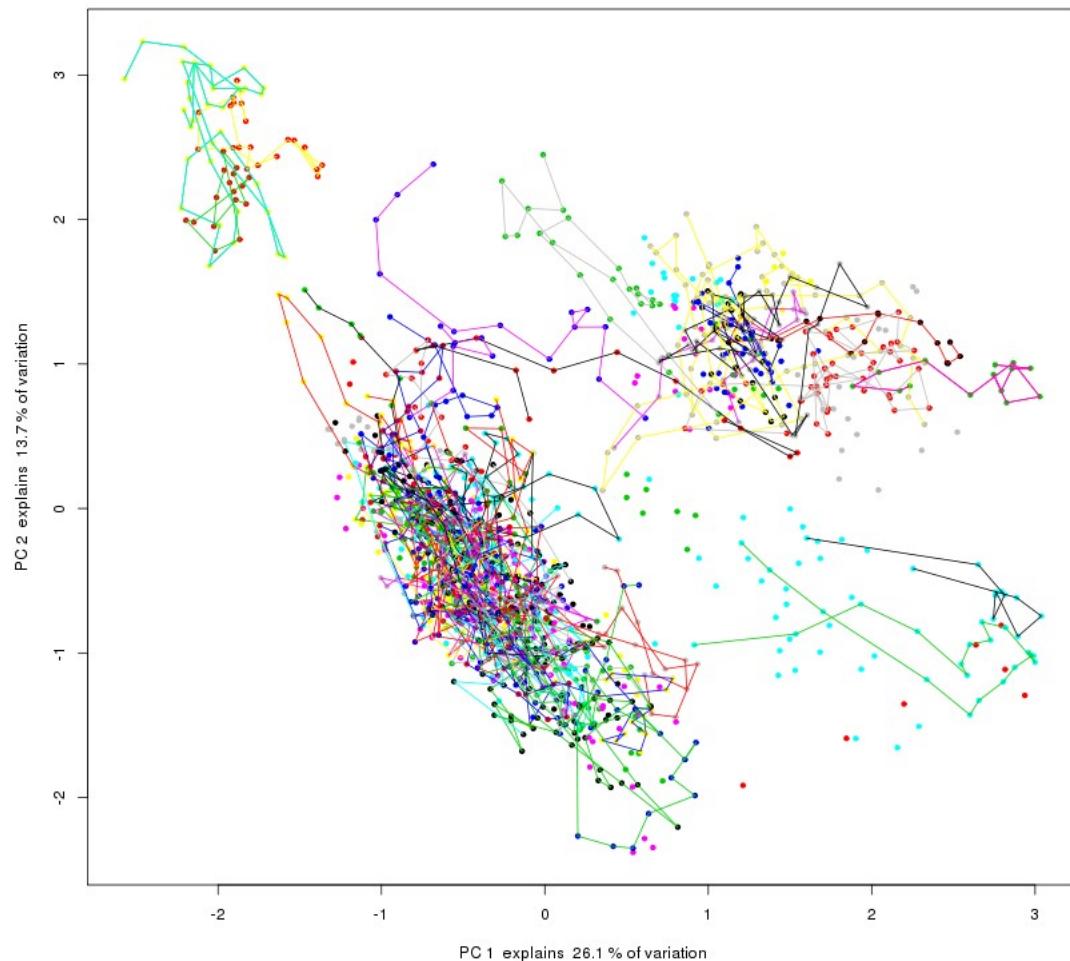
35	NODE_2.length_27437_cov_598.137.ID_3, NODE_3.length_26241_cov_2614.64.ID_5, NODE_5.length_24570_cov_793.923.ID_9, NODE_7.length_21087_cov_476.738.ID_13, NODE_8.length_20189_cov_510.344.ID_15, NODE_9.length_19696_cov_750.864.ID_17, NODE_17.length_14810_cov_310.416.ID_33, NODE_18.length_14170_cov_2455.66.ID_35, NODE_20.length_13791_cov_463.265.ID_43, NODE_24.length_12651_cov_239.31.ID_49, NODE_27.length_11890_cov_109.997.ID_55, NODE_35.length_10575_cov_1528.8.ID_71, NODE_36.length_10292_cov_315.117.ID_73, NODE_38.length_10077_cov_512.891.ID_77, NODE_39.length_10000_cov_11395.1.ID_79, NODE_40.length_9836_cov_285.522.ID_81, NODE_43.length_9516_cov_4801.47.ID_87, NODE_44.length_9388_cov_158.644.ID_89, NODE_45.length_9297_cov_164.668.ID_91, NODE_47.length_9056_cov_155.92.ID_95, NODE_49.length_8842_cov_31.581.ID_99, NODE_52.length_8717_cov_15902.5.ID_105, NODE_54.length_8606_cov_90.3096.ID_113, NODE_55.length_8391_cov_276.1.ID_115, NODE_57.length_8267_cov_522.658.ID_119, NODE_58.length_8224_cov_205.128.ID_121, NODE_59.length_8176_cov_28.5578.ID_123, NODE_63.length_7772_cov_1070.53.ID_131, NODE_66.length_7256_cov_14.4402.ID_137, NODE_69.length_6794_cov_1624.17.ID_143, NODE_71.length_6717_cov_6086.59.ID_147, NODE_73.length_6660_cov_5702.61.ID_151, NODE_75.length_6653_cov_679.533.ID_155, NODE_78.length_6523_cov_1031.53.ID_161, NODE_81.length_6261_cov_28.5989.ID_171, NODE_83.length_6138_cov_24.2477.ID_107, NODE_84.length_5977_cov_3059.95.ID_175, NODE_85.length_5938_cov_300.524.ID_177, NODE_86.length_5845_cov_82.0611.ID_179, NODE_87.length_5825_cov_4205.02.ID_181, NODE_91.length_5308_cov_7.80735.ID_189, NODE_95.length_5227_cov_69.1019.ID_197, NODE_97.length_5168_cov_417.104.ID_201, NODE_103.length_5027_cov_23572.8.ID_217 NODE_104.length_4994_cov_502.849.ID_219
40	NODE_1.length_34122_cov_2600.42.ID_1, NODE_12.length_17767_cov_171.843.ID_23, NODE_13.length_16024_cov_2312.52.ID_25, NODE_14.length_15419_cov_6489.68.ID_27, NODE_15.length_15171_cov_582.694.ID_29, NODE_21.length_13385_cov_10.6239.ID_39, NODE_29.length_11886_cov_975.971.ID_59, NODE_30.length_11563_cov_11.5433.ID_61, NODE_31.length_11484_cov_2863.85.ID_63, NODE_37.length_10159_cov_139.644.ID_75, NODE_46.length_9158_cov_19.9255.ID_93, NODE_53.length_8632_cov_238.816.ID_111, NODE_56.length_8339_cov_274.325.ID_117, NODE_61.length_8072_cov_320.615.ID_127, NODE_64.length_7673_cov_5.36939.ID_133, NODE_65.length_7529_cov_370.779.ID_135, NODE_68.length_6911_cov_474.414.ID_141, NODE_70.length_6749_cov_77.0257.ID_145, NODE_74.length_6660_cov_37.417.ID_153, NODE_76.length_6574_cov_35.3829.ID_157, NODE_82.length_6176_cov_20.1085.ID_173, NODE_90.length_5637_cov_164.042.ID_187, NODE_99.length_5112_cov_2088.85.ID_205, NODE_100.length_5108_cov_2268.81.ID_207, NODE_102.length_5045_cov_4.54168.ID_211, NODE_105.length_4864_cov_710.911.ID_221
45	NODE_6.length_21111_cov_17.3559.ID_11, NODE_11.length_19011_cov_41.1444.ID_21, NODE_19.length_13997_cov_1424.6.ID_37, NODE_22.length_12902_cov_20.6803.ID_45, NODE_25.length_12541_cov_2653.47.ID_51, NODE_26.length_12215_cov_26.5859.ID_53, NODE_34.length_10980_cov_31.0981.ID_69, NODE_48.length_9003_cov_20.5872.ID_97, NODE_51.length_8747_cov_28.4755.ID_103, NODE_60.length_8101_cov_1151.61.ID_125, NODE_77.length_6560_cov_9.80953.ID_159, NODE_88.length_5817_cov_18.1498.ID_183 NODE_93.length_5264_cov_4.53081.ID_193
50	NODE_4.length_24762_cov_34.7596.ID_7, NODE_16.length_15083_cov_8.32792.ID_31, NODE_23.length_12657_cov_298.186.ID_47, NODE_41.length_9811_cov_26.3295.ID_83, NODE_62.length_8040_cov_24.602.ID_129, NODE_67.length_7246_cov_231.383.ID_139 NODE_94.length_5237_cov_8.70822.ID_195
55	NODE_50.length_8755_cov_23.1709.ID_101, NODE_96.length_5216_cov_497.823.ID_199

List of the top contig megablast hits against 16S ribosomal RNA genes.

Organism	Align Length (bp)	Pct Id	Contig Name
981_Metallocphaera_sp._str._II_AF167083.1.I.1448	158	91.14	NODE_96.length_5216_cov_497.823.ID_199

Principal component analysis of tetramer frequencies of contigs. Detectable variations are highlighted in color.

sag_decontam_output_clean.fna - PC1 vs PC2



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

HMM	Pct Recovered
bacteria	38.37 %
archaea	77.5 %

6. Sequence Data Availability

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

Filename	Description
sag_decontam_output_clean.fna	SPAdes with auto decontamination

7. Annotation Data Availability

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

8. Methods

Single Cell Minimal Draft

Genome sequencing and assembly

The draft genome of was generated at the DOE Joint genome Institute (JGI) using the Illumina technology [1]. An Illumina std shotgun library was constructed and sequenced using the Illumina HiSeq 2000 platform which generated 24,895,556 reads totaling 3,734.3 Mb. All general aspects of library construction and sequencing performed at the JGI can be found at <http://www.jgi.doe.gov>. All raw Illumina sequence data was passed through DUK, a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts [2]. Following steps were then performed for assembly: (1) artifact filtered Illumina reads were assembled using SPAdes [3] (version 3.0.0), (3) Parameters for assembly steps were `-t 16 -m 120 --sc --careful --12`. The final draft assembly contained 101 contigs in 101 scaffolds, totalling 1.1 Mb in size. The final assembly was based on 3,000.0 Mb of Illumina data. Based on a presumed genome size of 5.0 Mb, the average input read coverage used for the assembly was 600.0X.

Genome annotation

Genes were identified using Prodigal [4], followed by a round of manual curation using GenePRIMP [5] for finished genomes and Draft genomes in fewer than 20 scaffolds. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGRFam, Pfam, KEGG, COG, and InterPro databases. The tRNAscanSE tool [6] was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA [7]. Other non-coding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching the genome for the corresponding Rfam profiles using INFERNAL [8]. Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) platform [9] developed by the Joint Genome Institute, Walnut Creek, CA, USA [10].

1. Bennett S. Solexa Ltd. *Pharmacogenomics*. 2004;5(4):433–8.
2. Mingkun L, Copeland A, Han J. DUK, unpublished, 2011.
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.
4. Hyatt D, Chen GL, Lascasio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; 11:119.
5. Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* 2010; 7:455–457.
6. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997; 25:955–964.
7. Pruesse E, Quast C, Knittel, Fuchs B, Ludwig W, Peplies J, Glckner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nuc Acids Res* 2007; 35: 2188–7196.
8. INFERNAL. Inference of RNA alignments. <http://infernal.janelia.org>.
9. The Integrated Microbial Genomes (IMG) platform. <http://www.ncbi.nlm.nih.gov/pubmed/24165883>
10. Markowitz VM, Mavromatis K, Ivanova NN, Chen IMA, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 2009; 25:2271–2278.