uncultured Thioalkalivibrio sp.

Single–cell Assembly QC Report   04/09/2014

1. Project Information

Program	Microbial/CSP 2012
PMO Project	0
Seq Proj ID	1027130
Sequencing Project Name	Thioalkalivibrio sp. HL7711_P3F6 JGI 000155CP–E08
JGI Project ID	0

2. Read Statistics

Illumina Std PE Statistics

File name	7667.6.80858.CAGGCG.fastq
Library	TNGB
Number of reads	29,548,988
Sequencing depth †	886X
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	29,548,988	100
Contam removed	1100	0.0
Artifact removed	778,344	2.6
Total removed	9,548,988	32.3
Total remaining	20,000,000	67.7

List of Contaminants Removed

Description	Num Reads	Pct Reads
human_chr11	456	0.00
human_chr6	266	0.00
gi|357579577|Canis_lupus_familiaris_chr3	232	0.00
human_chr2	228	0.00
gi|357579535|Canis_lupus_familiaris_chr20	98	0.00
gi|357579571|Canis_lupus_familiaris_chr5	24	0.00
human_chr15	14	0.00

gi|357579564|Canis_lupus_familiaris_chr7	4	0.00
human_chr1	2	0.00
gi|357579585|Canis_lupus_familiaris_chr1	2	0.00
human_chr14	2	0.00
gi|357579553|Canis_lupus_familiaris_chr10	2	0.00
gi|357579584|Canis_lupus_familiaris_chr1	2	0.00
human_chr8	2	0.00
gi|357579517|Canis_lupus_familiaris_chr31	2	0.00
gi|357579551|Canis_lupus_familiaris_chr11	2	0.00
gi|357579511|Canis_lupus_familiaris_chr34	2	0.00

The following are the results of reads screened against potential reagent and process contaminants but were not removed from the dataset.
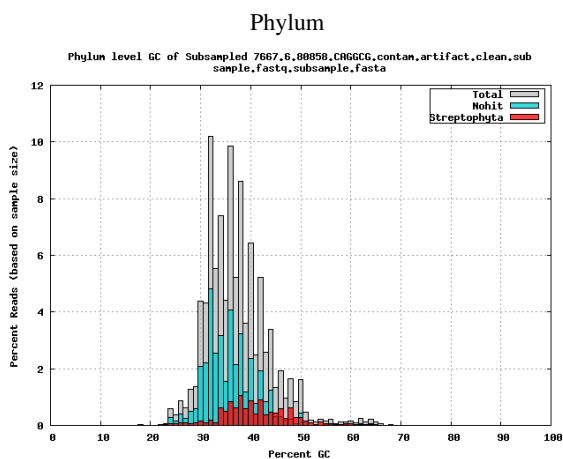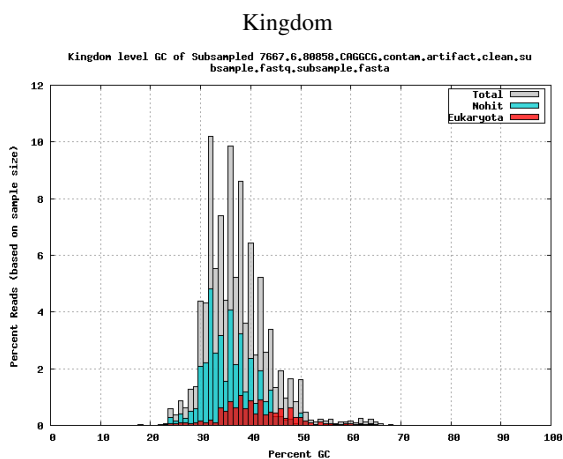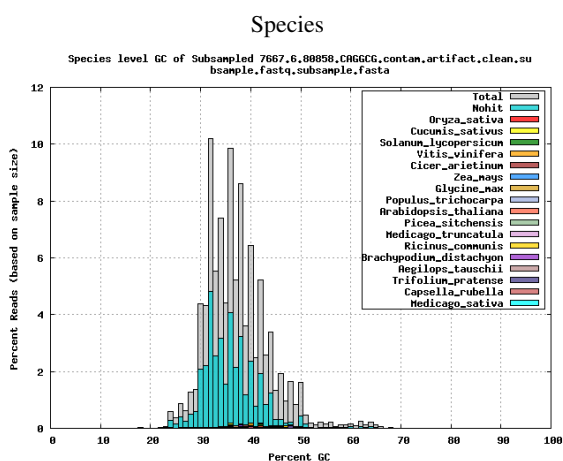
Illumina Std PE Contamination Identification Statistics

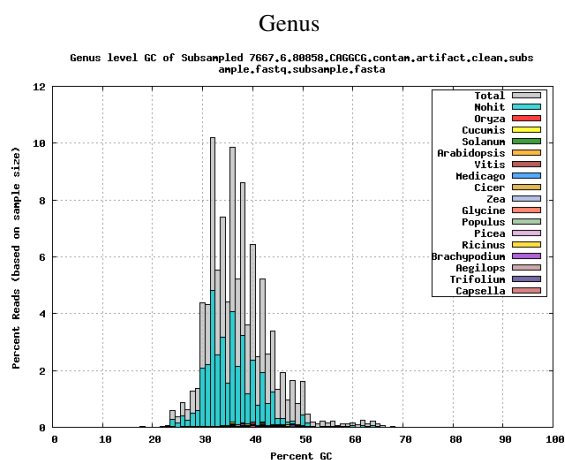| Description	Num Reads	Pct Reads
Input	29,548,988	100
Contam identified	20	0.0

List of Contaminants Identified

Description	Num Reads	Pct Reads
Pseudomonas	16	0.00
Escherichia	4	0.00

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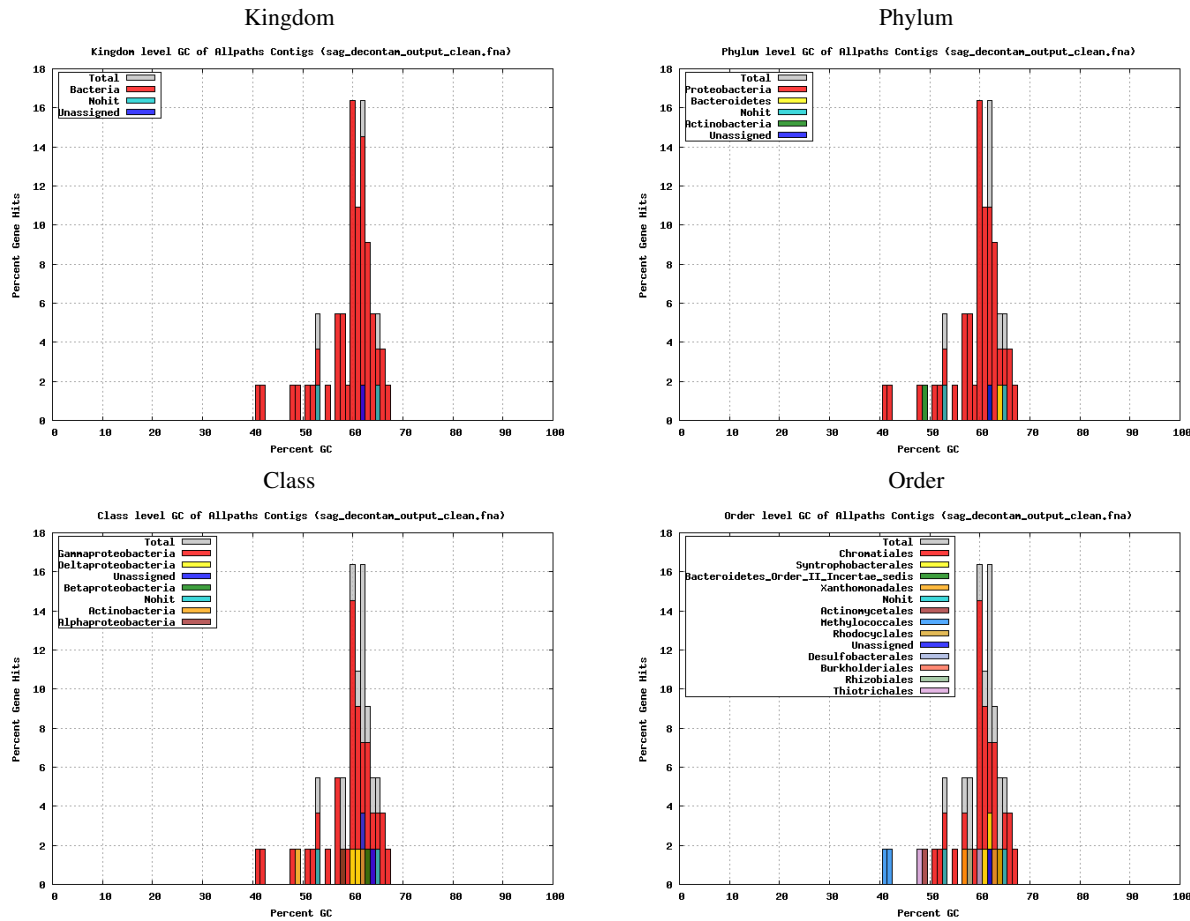
Kingdom

Phylum

Class

Class level GC of Subsampled 7667.6.80858.CAGGCG.contam.artifact.clean.subsample.fastq.subsample.fasta

Order

Order level GC of Subsampled 7667.6.80858.CAGGCG.contam.artifact.clean.subsample.fastq.subsample.fasta

Family

Family level GC of Subsampled 7667.6.80858.CAGGCG.contam.artifact.clean.subsample.fastq.subsample.fasta

Genus

Genus level GC of Subsampled 7667.6.80858.CAGGCG.contam.artifact.clean.subsample.fastq.subsample.fasta

Species

Species level GC of Subsampled 7667.6.80858.CAGGCG.contam.artifact.clean.subsample.fastq.subsample.fasta

4. Assembly Statistics

Assembly method	SPAdes with auto decontamination
Scaffold total	7
Contig total	7
Scaffold sequence length	49.9 kb
Contig sequence length	49.9 kb (0.0% gap)
Scaffold N/L50	2/12.6 kb
Contig N/L50	2/12.6 kb
Largest Contig	14.9 kb
Number of scaffolds >50 kb	0
Pct of genome in scaffolds >50 kb	0.0
Pct of reads asssembled (raw)	0.2
Pct of reads asssembled (decontam)	0.1

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.
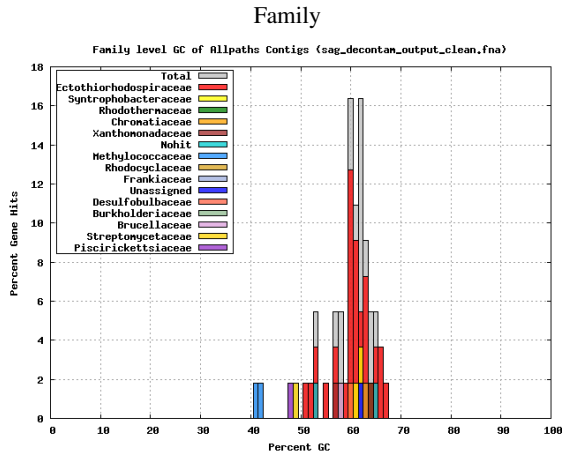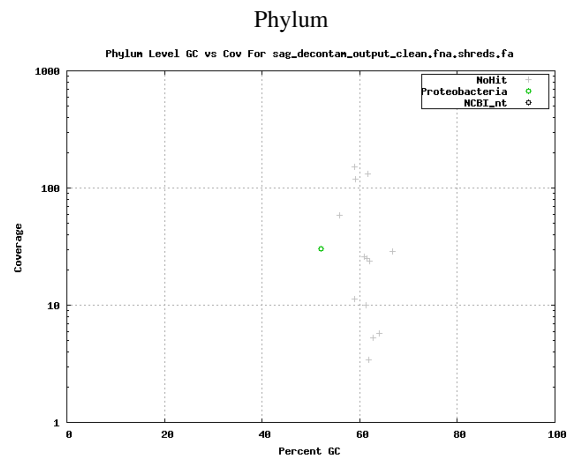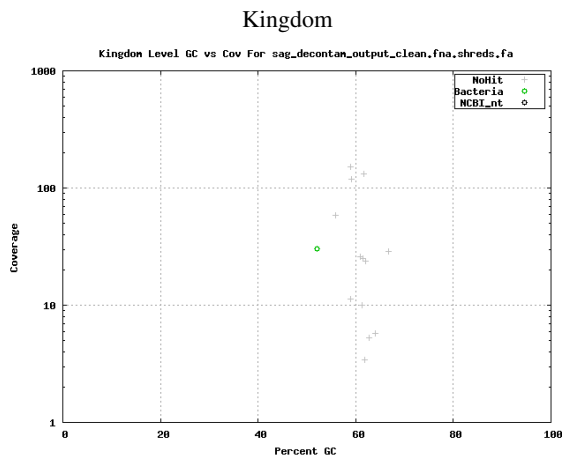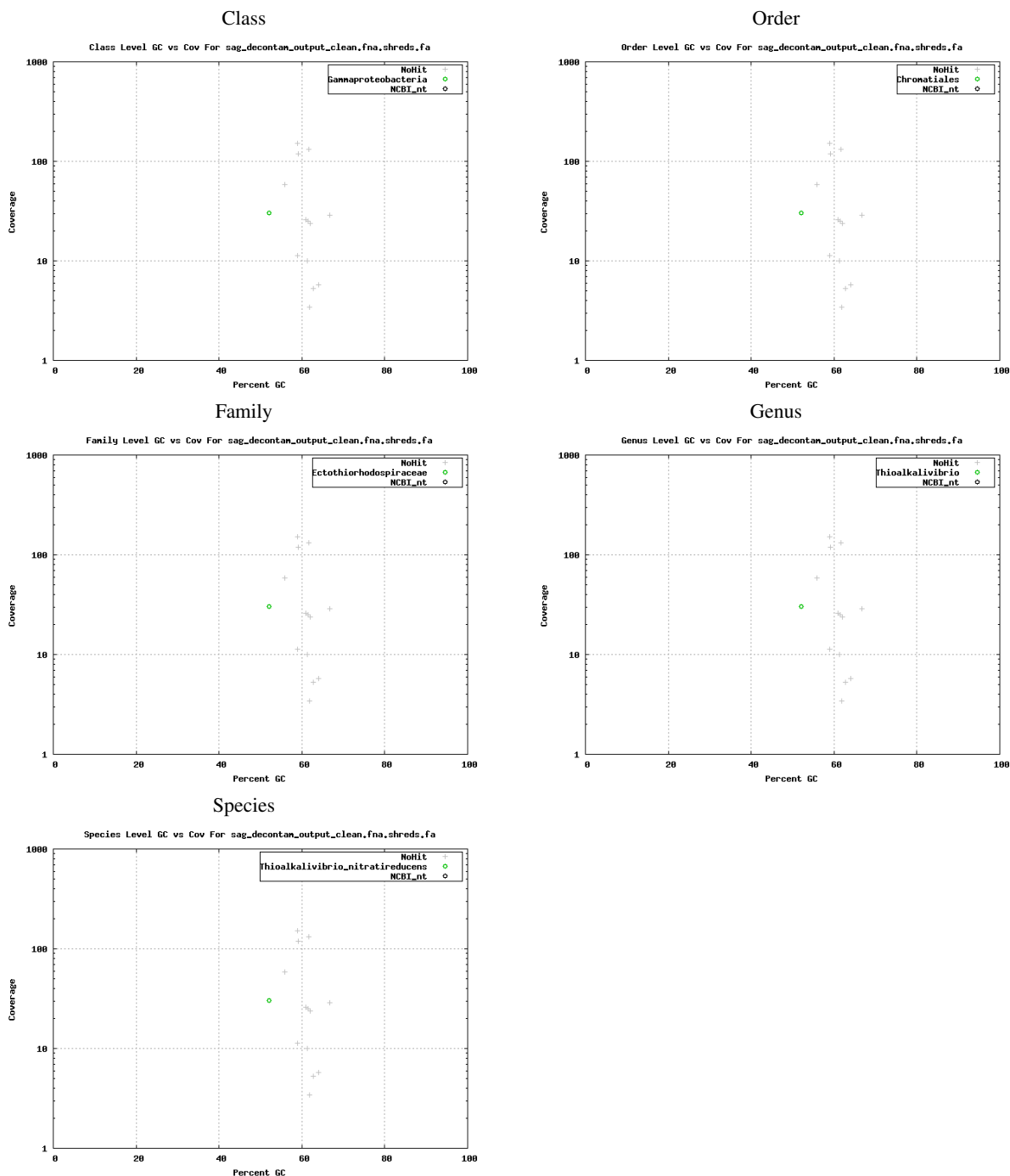
Kingdom

Phylum

Class

Order

4

Family

Family level GC of Allpaths Contigs (sag_decontam_output_clean.fna)

Genus

Genus level GC of Allpaths Contigs (sag_decontam_output_clean.fna)

Species

Species level GC of Allpaths Contigs (sag_decontam_output_clean.fna)
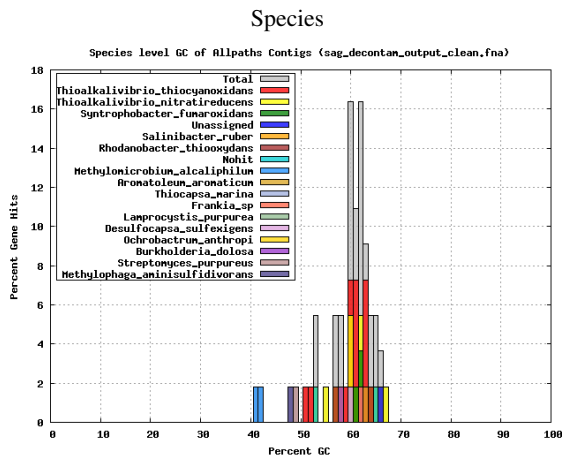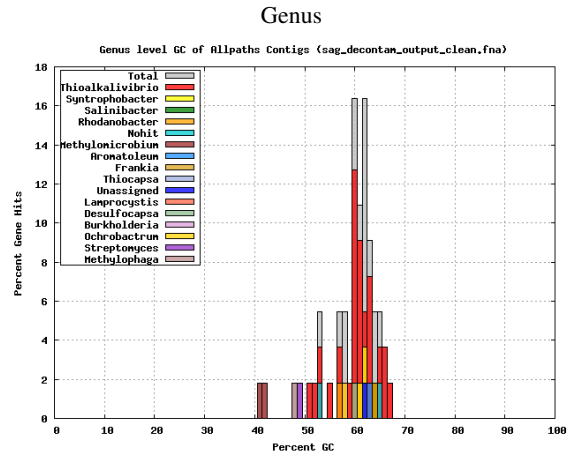
GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

Kingdom

Kingdom Level GC vs Cov For sag_decontam_output_clean.fna.shreds.fa

Phylum

Phylum Level GC vs Cov For sag_decontam_output_clean.fna.shreds.fa

Class

Class Level GC vs Cov For sag_decontam_output_clean.fna.shreds.fa

Order

Order Level GC vs Cov For sag_decontam_output_clean.fna.shreds.fa

Family

Family Level GC vs Cov For sag_decontam_output_clean.fna.shreds.fa

Genus

Genus Level GC vs Cov For sag_decontam_output_clean.fna.shreds.fa

Species

Species Level GC vs Cov For sag_decontam_output_clean.fna.shreds.fa

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 asssembly and plotted as connected points.
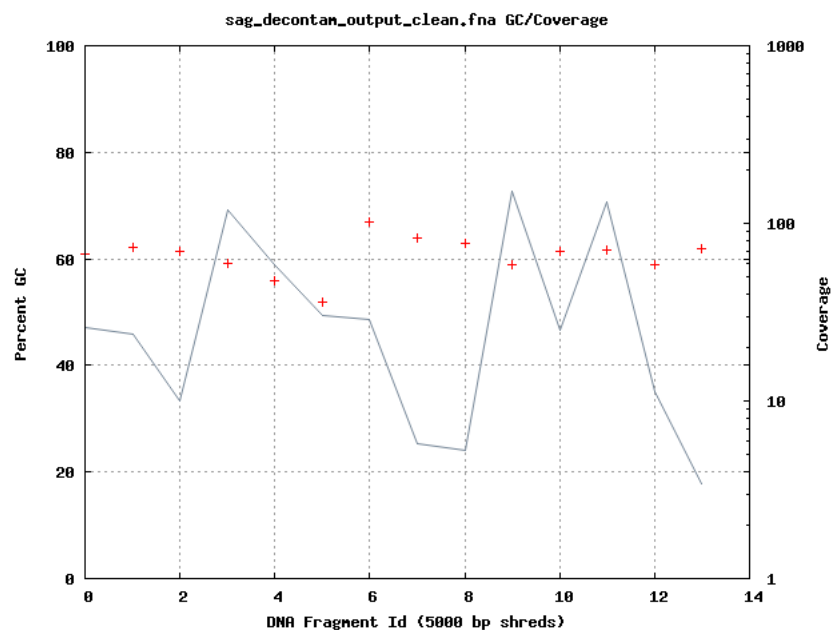
GC histogram of the contigs, including contig length weighted distribution.
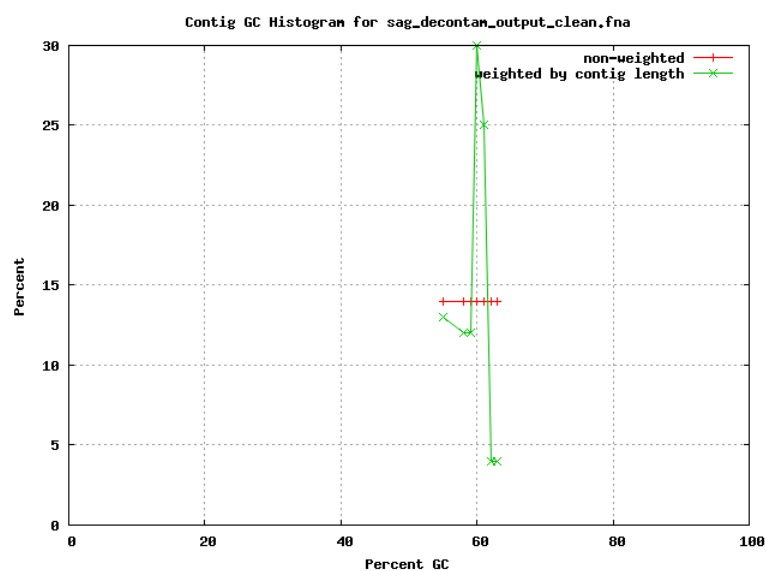
List of contigs and average percent GC, grouped in bins of 5:

Pct GC Bin	Contig Name
55	NODE_5_length_6498_cov_20.1603_ID_19, NODE_7_length_5836_cov_76.0796_ID_13 NODE_8_length_5834_cov_6.71223_ID_23
60	NODE_1_length_14872_cov_67.5662_ID_1, NODE_2_length_12596_cov_14.3749_ID_3, NODE_37_length_2178_cov_3.57183_ID_79, NODE_38_length_2042_cov_3.76548_ID_81

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

Insufficent number of contigs with length >= 2000 bp. Minimum number required is 10 to run tetramer analysis.

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

HMM	Pct Recovered
bacteria	10.39 %
archaea	9.6 %

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 Availiability

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 29,548,988 reads totaling 4,432.3 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 7 contigs in 7 scaffolds, totalling 49.9 Kb 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, Lacascio 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.