

## 1. Project Information

Program	Microbial/CSP 2012
PMO Project	0
Seq Proj ID	1027163
Sequencing Project Name	Rhodospirillaceae bacterium HL7711_P1E10 JGI 000149CP-D12
JGI Project ID	0

## 2. Read Statistics

### Illumina Std PE Statistics

File name	7667.7.80862.TATAAT.fastq
Library	TNGX
Number of reads	26,380,356
Sequencing depth <sup>†</sup>	791X
Read type	2x150 bp

<sup>†</sup> 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	26,380,356	100
Contam removed	162	0.0
Artifact removed	1,165,976	4.4
Total removed	6,380,356	24.2
Total remaining	20,000,000	75.8

### List of Contaminants Removed

Description	Num Reads	Pct Reads
gi 357579577 Canis.lupus_familiaris_chr3	112	0.00
human_chr2	108	0.00
gi 357579535 Canis.lupus_familiaris_chr20	22	0.00
gi 357579571 Canis.lupus_familiaris_chr5	10	0.00
human_chr1	4	0.00
human_chr17	4	0.00
human_chr5	4	0.00

human_chr11	4	0.00
human_chr13	2	0.00
gi 362110625  <i>Felis_catus_breed_Abyssinian</i> _chrE3	2	0.00
human_chr14	2	0.00
human_chr7	2	0.00
human_chr3	2	0.00
human_chrY	2	0.00
human_chr4	2	0.00
gi 357579570  <i>Canis_lupus_familiaris</i> _chr6	2	0.00

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

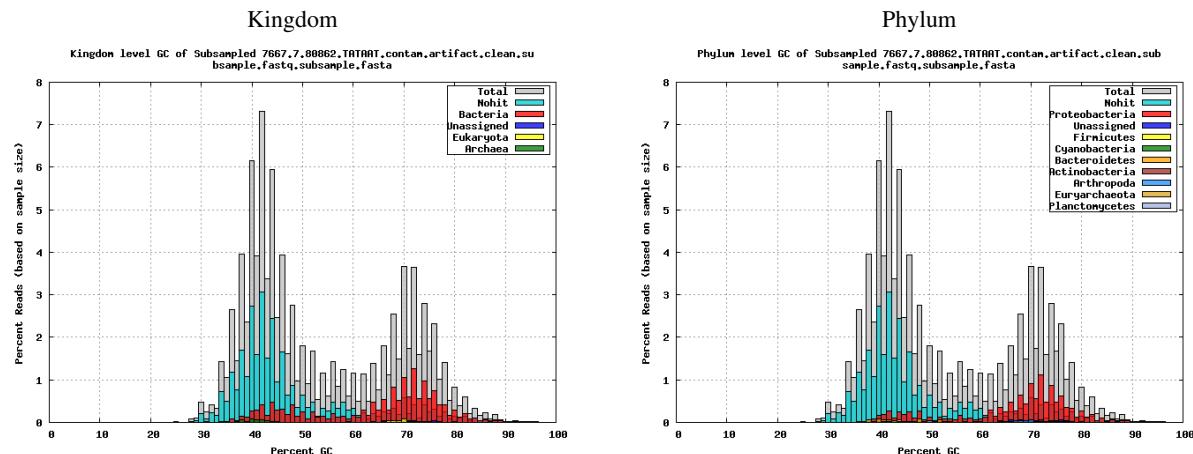
#### Illumina Std PE Contamination Identification Statistics

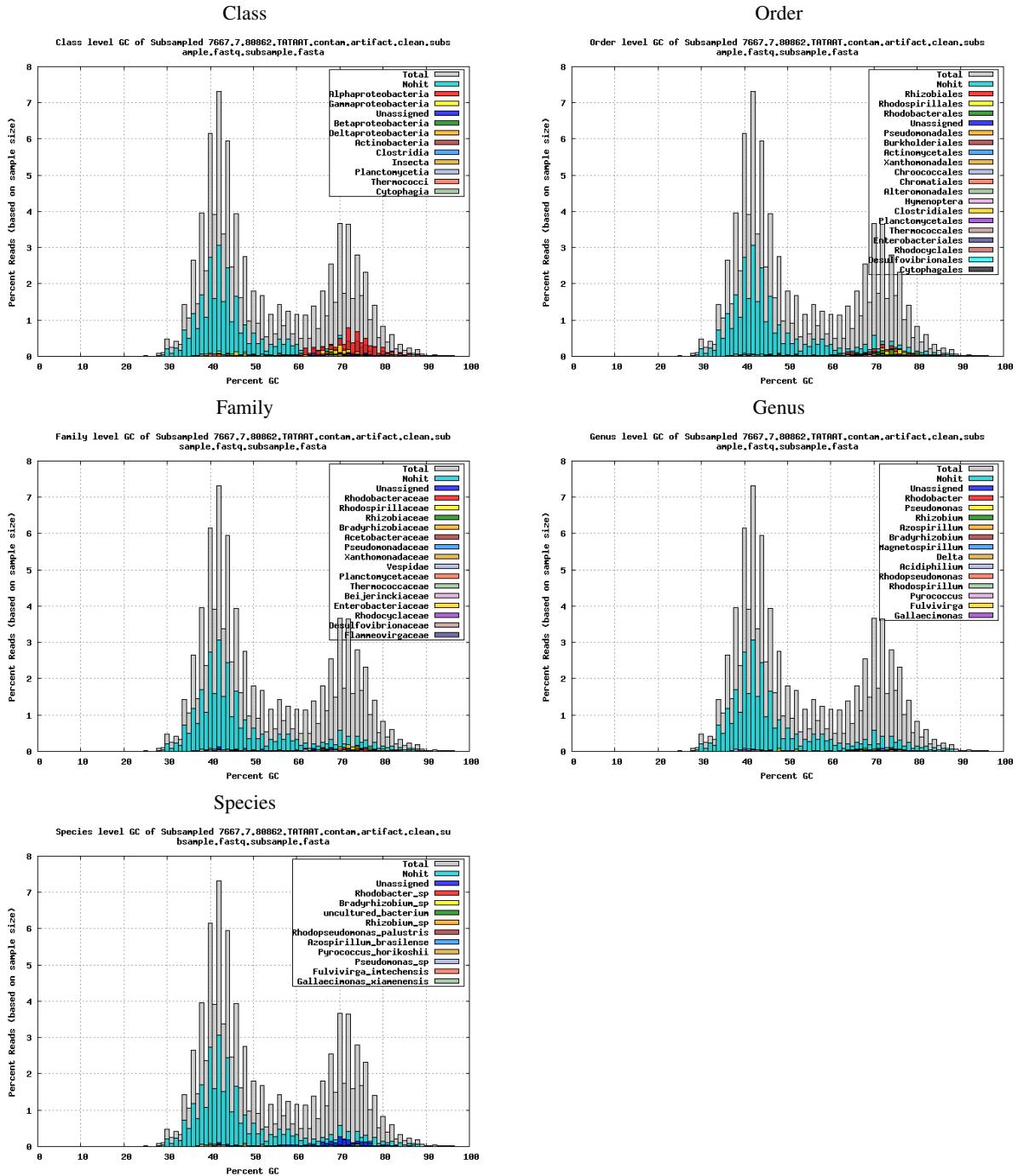
Description	Num Reads	Pct Reads
Input	26,380,356	100
Contam identified	10	0.0

#### List of Contaminants Identified

Description	Num Reads	Pct Reads
<i>Pseudomonas</i>	4	0.00
<i>Escherichia</i>	2	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.



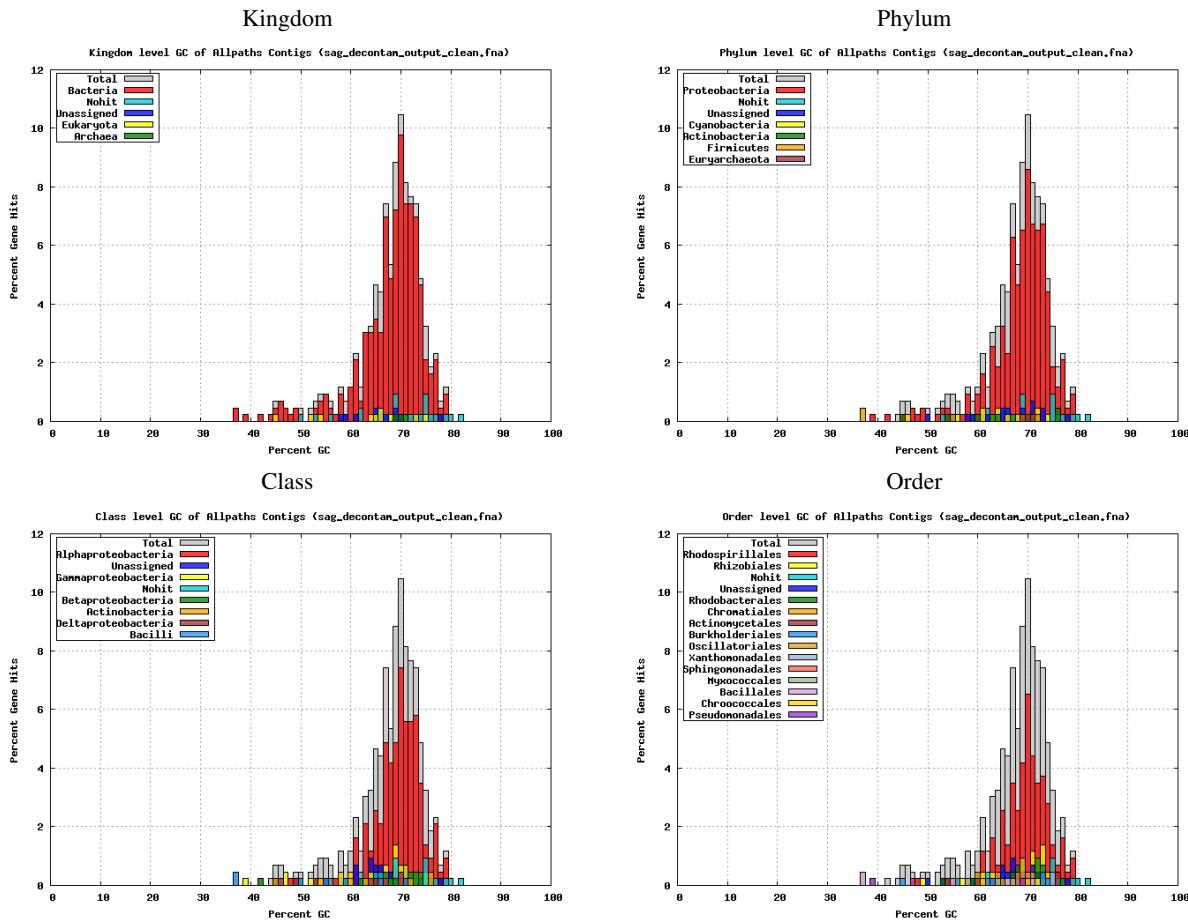


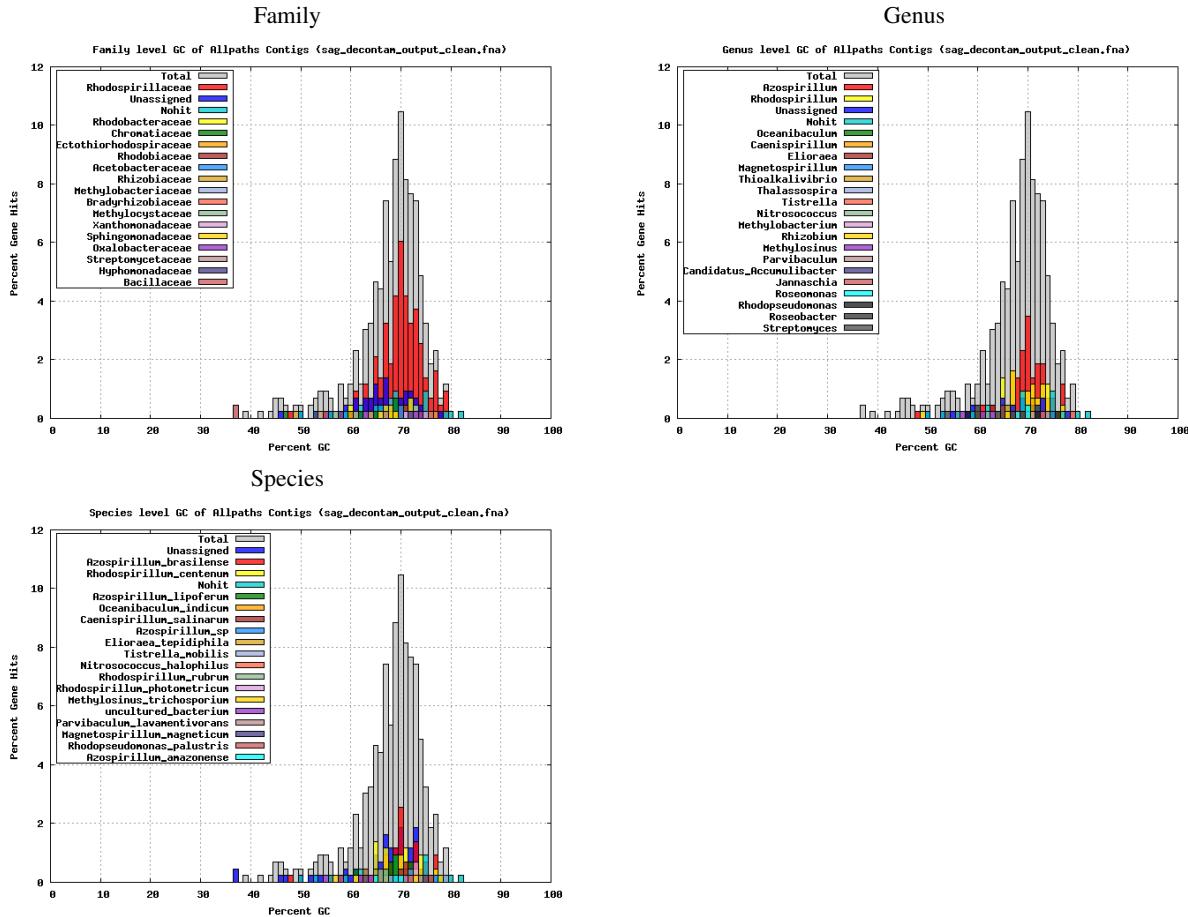
## 4. Assembly Statistics

Assembly method	SPAdes with auto decontamination
Scaffold total	49
Contig total	49
Scaffold sequence length	441.2 kb
Contig sequence length	441.2 kb ( 0.0% gap)
Scaffold N/L50	10/12.5 kb
Contig N/L50	10/12.5 kb
Largest Contig	35.6 kb
Number of scaffolds >50 kb	0
Pct of genome in scaffolds >50 kb	0.0
Pct of reads assembled (raw)	28.4
Pct of reads assembled (decontam)	22.8

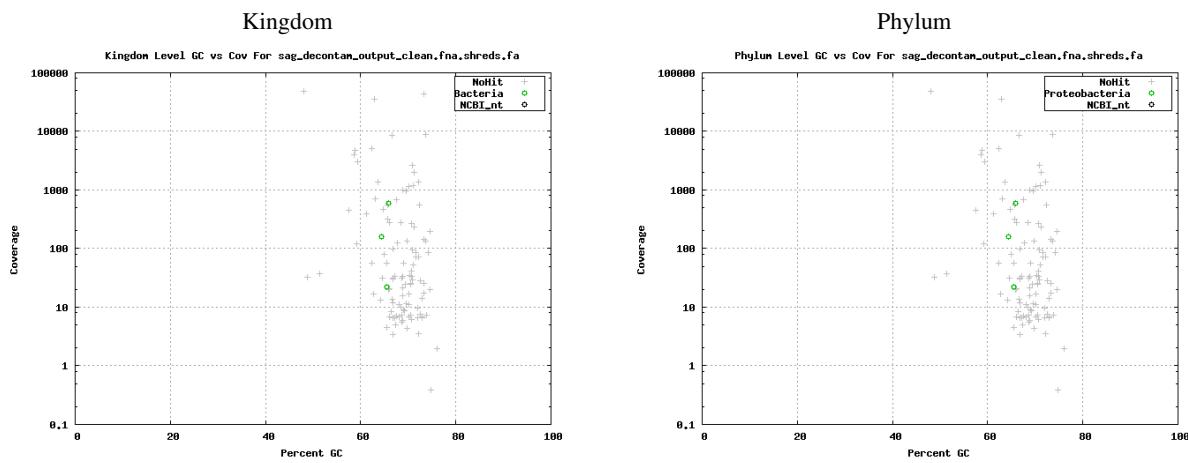
## 5. Assembly QC Results

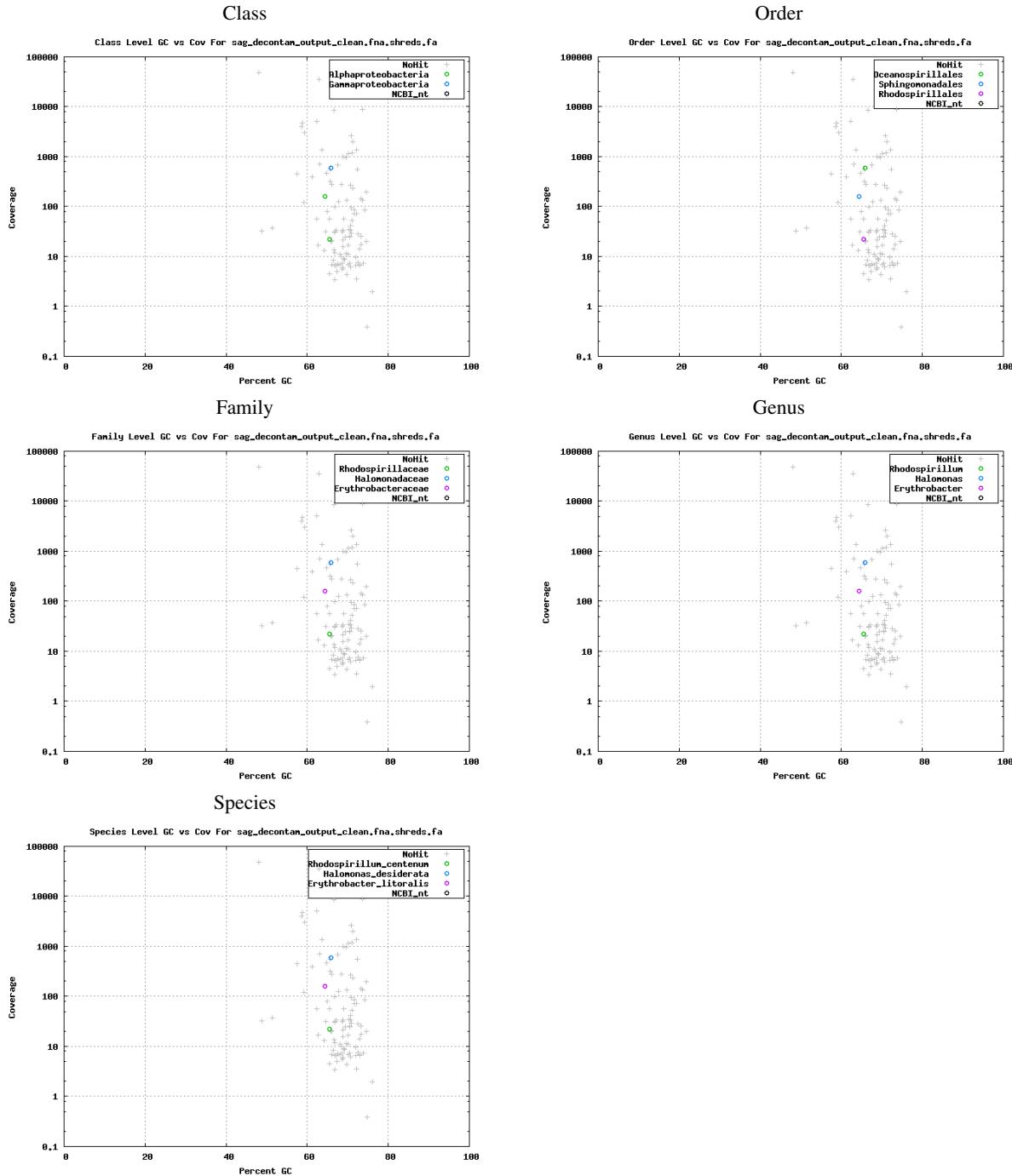
GC histogram of the predicted genes on each contig, overlaid with GC of hits based on BLASTP, shown for different taxonomic levels.



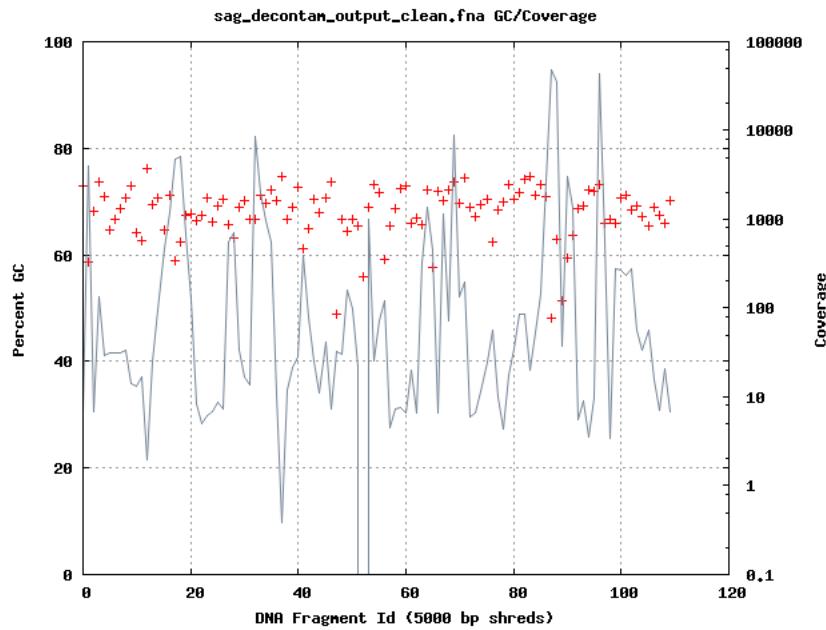


GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

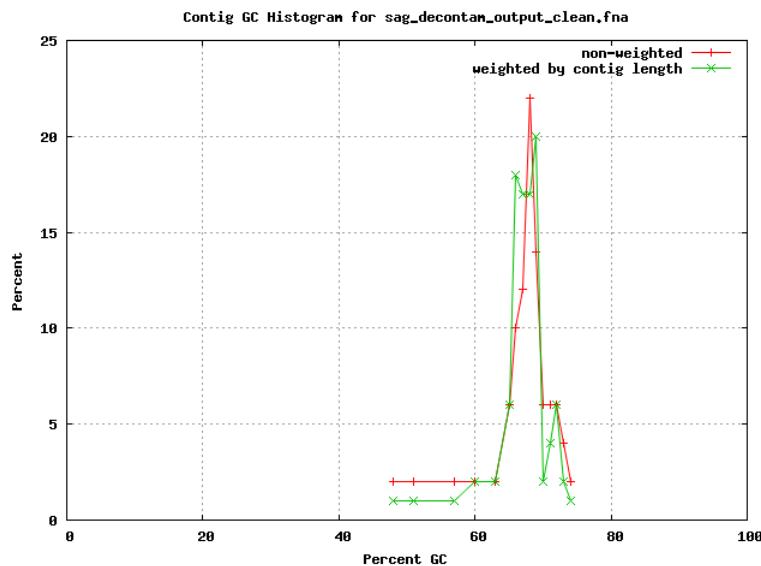




Coverage vs GC. Contigs were shredded into non-overlapping 5kbp and the GC of each shred was plotted as a point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.



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

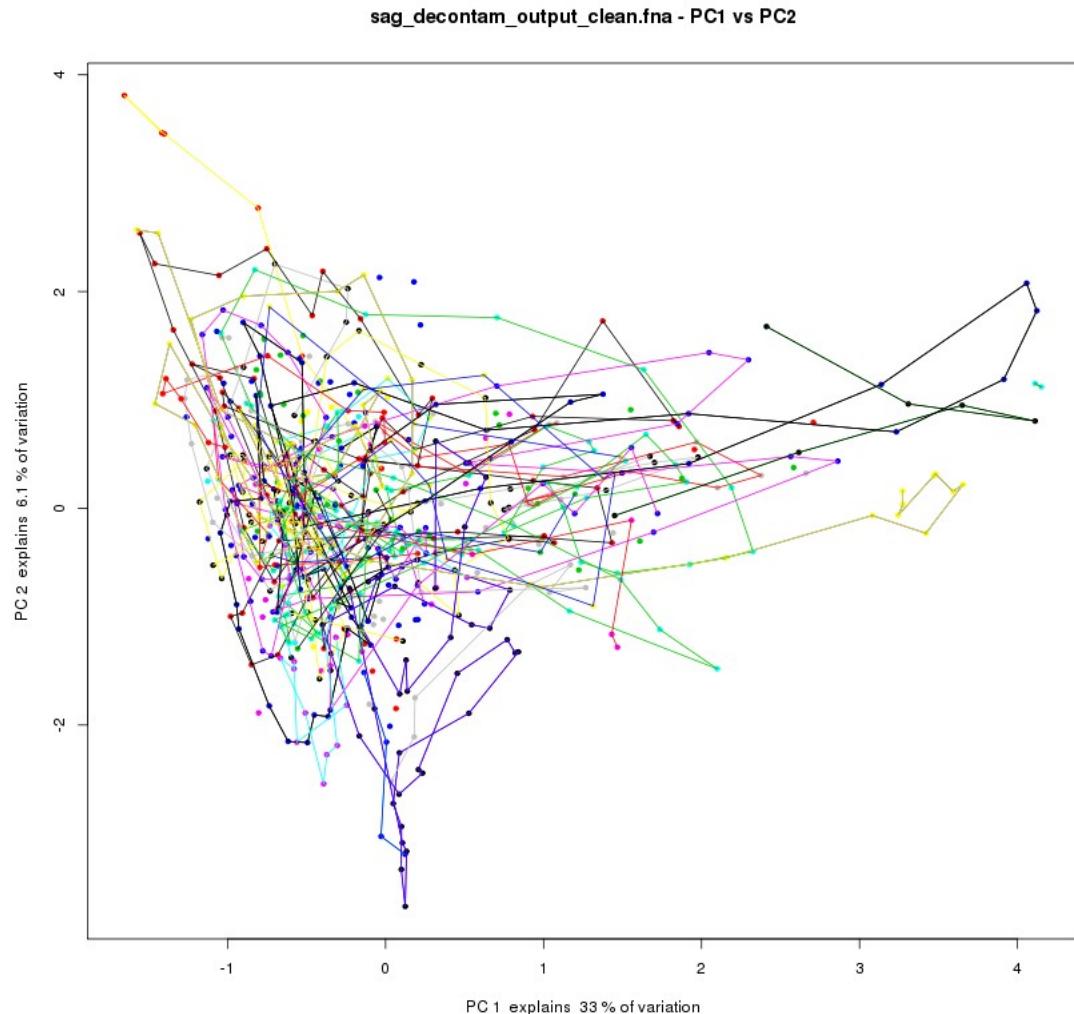


List of contigs and average percent GC, grouped in bins of 5:

Pct GC Bin	Contig Name
45	NODE_111.length_2623.cov_23.1153.ID_235
50	NODE_102.length_2792.cov_38.5049.ID_217
55	NODE_63.length_4837.cov_300.544.ID_139
60	NODE_20.length_10212.cov_9.7005.ID_45, NODE_35.length_7522.cov_1618.86.ID_73
65	NODE_2.length_35617.cov_1123.96.ID_3, NODE_3.length_34971.cov_29.5369.ID_5, NODE_4.length_31555.cov_58.608.ID_11, NODE_5.length_25967.cov_7541.82.ID_9,

	NODE_6.length_20617_cov_208.298.ID_13, NODE_7.length_20025_cov_50.1779.ID_15, NODE_9.length_18720_cov_2032.35.ID_19, NODE_12.length_12781_cov_6.83883.ID_25, NODE_14.length_12549_cov_40.4123.ID_29, NODE_15.length_12476_cov_170.795.ID_35, NODE_16.length_11748_cov_15.3657.ID_37, NODE_19.length_10531_cov_47.2872.ID_43, NODE_22.length_9586_cov_1704.44.ID_53, NODE_25.length_9001_cov_498.227.ID_59, NODE_29.length_8452_cov_166.16.ID_65, NODE_31.length_8031_cov_3.98533.ID_67, NODE_34.length_7701_cov_5.25569.ID_7, NODE_37.length_7234_cov_4.58532.ID_77, NODE_42.length_6731_cov_10.6637.ID_91, NODE_56.length_5108_cov_376.338.ID_125, NODE_61.length_4963_cov_6.12993.ID_135, NODE_65.length_4684_cov_4.23374.ID_143, NODE_81.length_3753_cov_4.36723.ID_179, NODE_82.length_3718_cov_5.5976.ID_109, NODE_83.length_3680_cov_5.42841.ID_181, NODE_87.length_3464_cov_3.07744.ID_189, NODE_101.length_2725_cov_666.999.ID_215, NODE_109.length_2632_cov_2.79511.ID_231, NODE_119.length_2502_cov_3.81079.ID_251, NODE_128.length_2308_cov_2.7723.ID_267, NODE_142.length_2122_cov_3.38413.ID_293, NODE_144.length_2094_cov_5.93477.ID_297
70	NODE_10.length_18184_cov_33.8928.ID_21, NODE_23.length_9392_cov_2909.56.ID_55, NODE_55.length_5256_cov_6.60527.ID_123, NODE_64.length_4664_cov_123.991.ID_141, NODE_70.length_4067_cov_2.20588.ID_155, NODE_71.length_4052_cov_4.04929.ID_157, NODE_76.length_3885_cov_4.76684.ID_167, NODE_79.length_3862_cov_43.7339.ID_171, NODE_85.length_3636_cov_4.36945.ID_185, NODE_92.length_3192_cov_25.935.ID_199, NODE_97.length_2985_cov_26248.6.ID_209, NODE_147.length_2029_cov_788.86.ID_303

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



---

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

HMM	Pct Recovered
bacteria	6.39 %
archaea	0 %

## 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 26,380,356 reads totaling 3,957.1 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 49 contigs in 49 scaffolds, totalling 441.2 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.