uncultured archaeon

Single–cell Assembly QC Report 04/07/2014

1. Project Information

Program	Microbial/CSP 2012
PMO Project	0
Seq Proj ID	1027082
Sequencing Project Name	NAG2 archaeon OSP418 JGI 000156CP–G16
JGI Project ID	0

2. Read Statistics

Illumina Std PE Statistics

File name	7666.6.80850.GTAGAG.fastq
Library	TGSU
Number of reads	24,950,880
Sequencing depth †	749X
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	24,950,880	100
Contam removed	262	0.0
Artifact removed	806,276	3.2
Total removed	4,950,880	19.8
Total remaining	20,000,000	80.2

List of Contaminants Removed

Description	Num Reads	Pct Reads
gi|357579535|Canis_lupus_familiaris_chr20	146	0.00
human_chr2	96	0.00
gi|357579577|Canis_lupus_familiaris_chr3	94	0.00
gi|357579571|Canis_lupus_familiaris_chr5	12	0.00
human_chr10	4	0.00
human_chr14	2	0.00
human_chr7	2	0.00

human_chrX	2	0.00
human_chr16	2	0.00
human_chr3	2	0.00
human_chr6	2	0.00
human_chr20	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

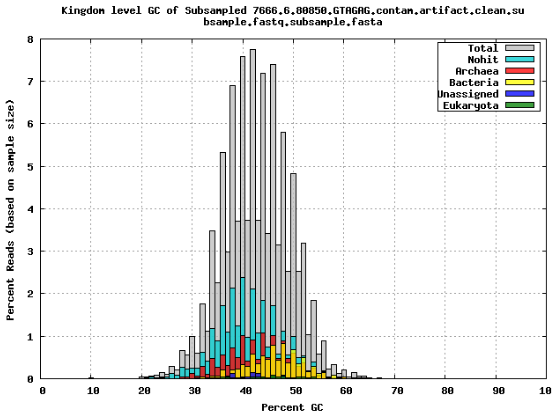
Description	Num Reads	Pct Reads
Input	24,950,880	100
Contam identified	4	0.0

List of Contaminants Identified

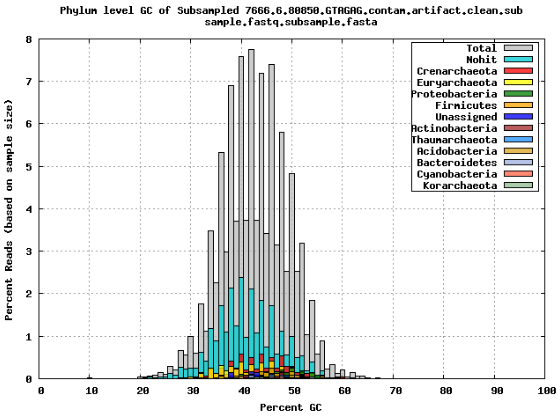
Description	Num Reads	Pct Reads
Klebsiella	2	0.00
Pseudomonas	2	0.00

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

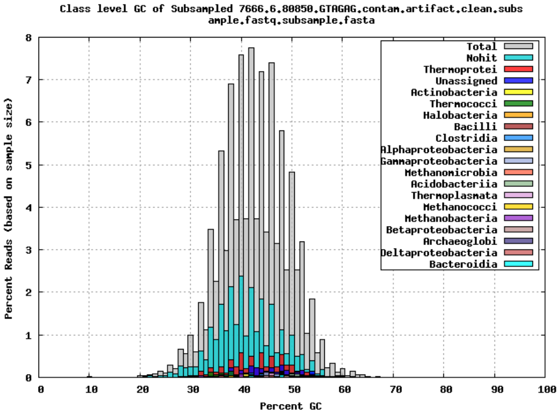
Kingdom



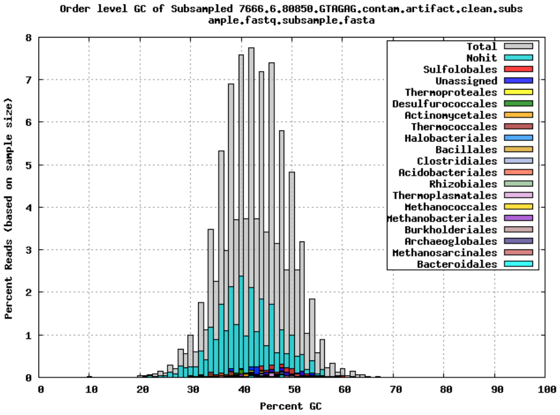
Phylum



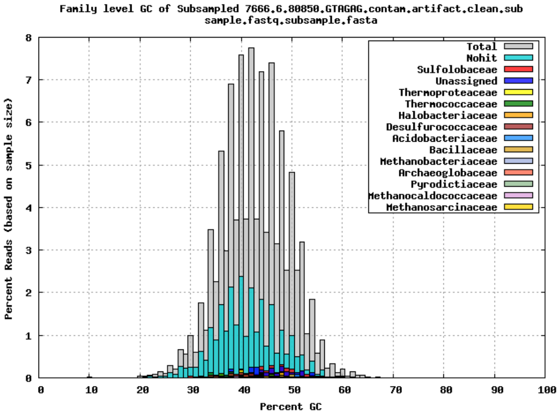
Class



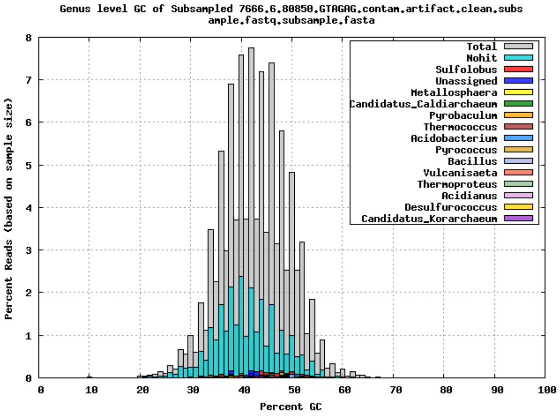
Order



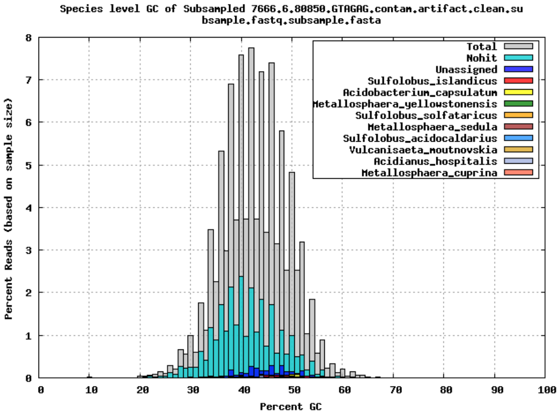
Family



Genus



Species



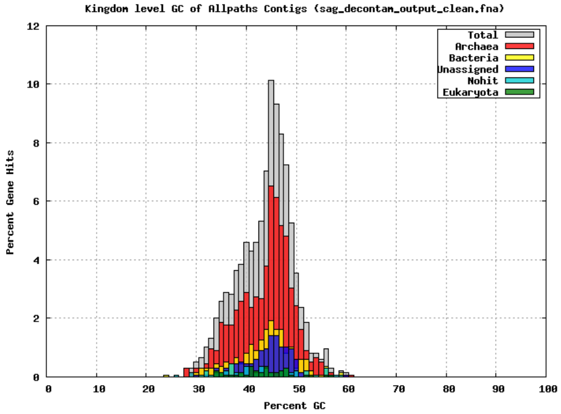
4. Assembly Statistics

Assembly method	SPAdes with auto decontamination
Scaffold total	99
Contig total	99
Scaffold sequence length	1.3 Mb
Contig sequence length	1.3 Mb (0.0% gap)
Scaffold N/L50	29/14.2 kb
Contig N/L50	29/14.2 kb
Largest Contig	48.4 kb
Number of scaffolds >50 kb	0
Pct of genome in scaffolds >50 kb	0.0
Pct of reads asssembled (raw)	87.3
Pct of reads asssembled (decontam)	67.7

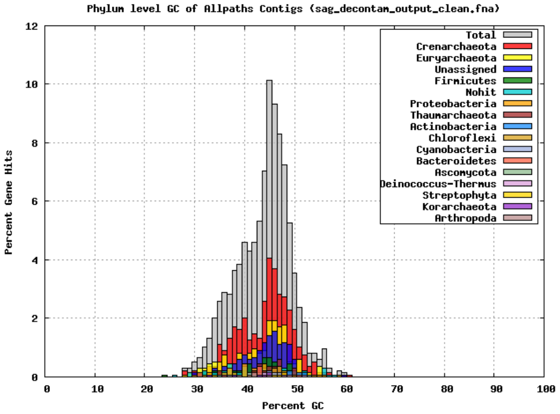
5. Assembly QC Results

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

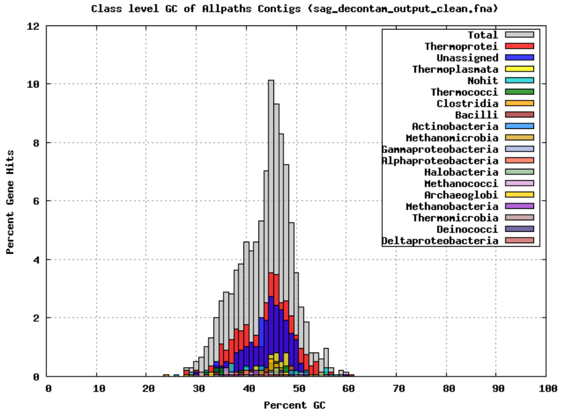
Kingdom



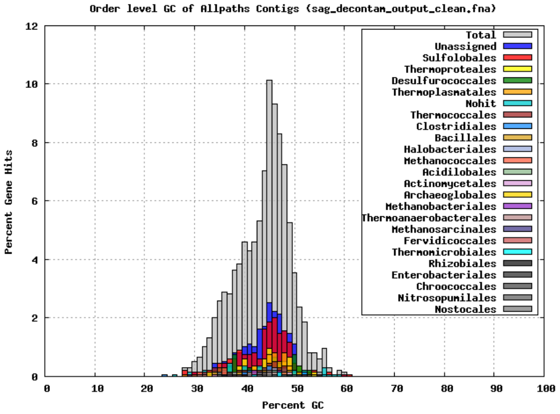
Phylum



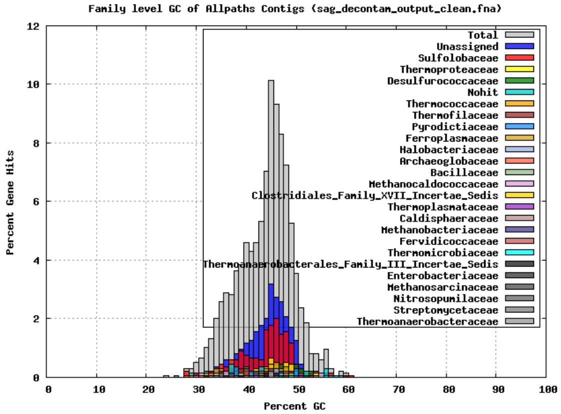
Class



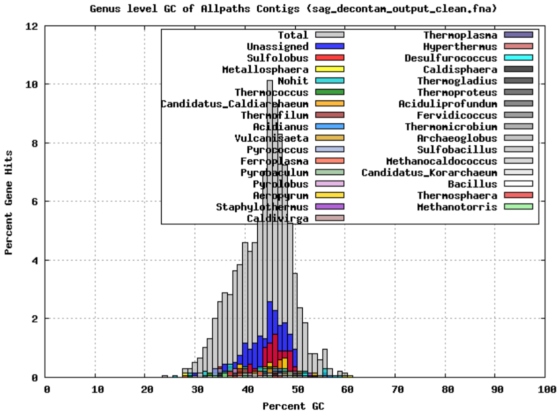
Order



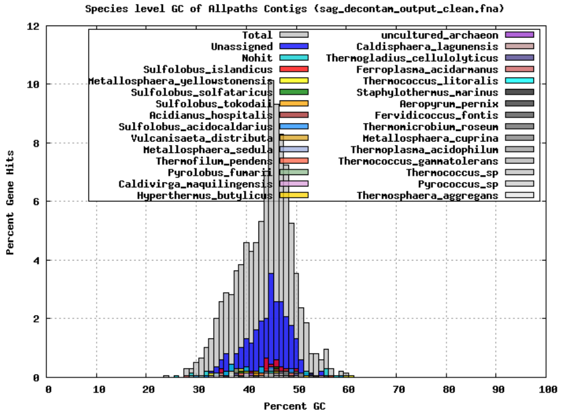
Family



Genus

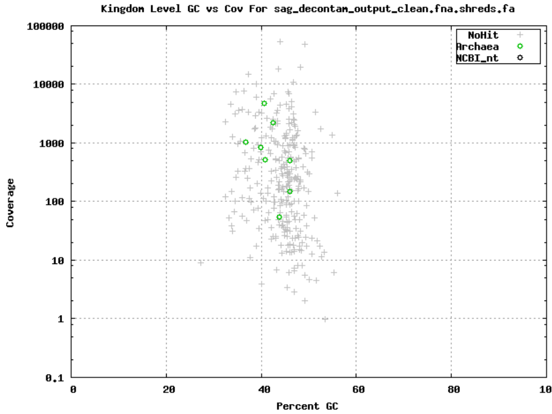


Species

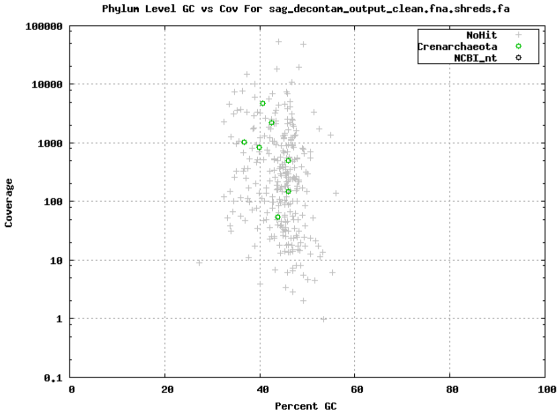


GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

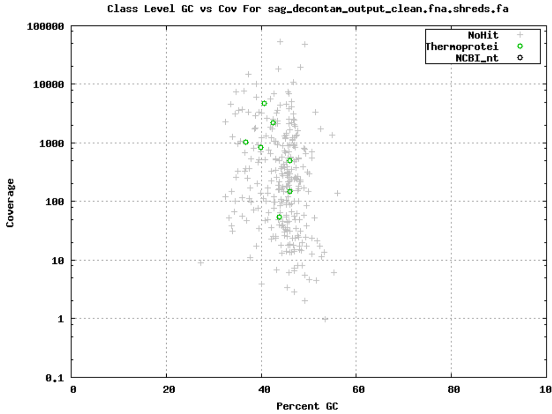
Kingdom



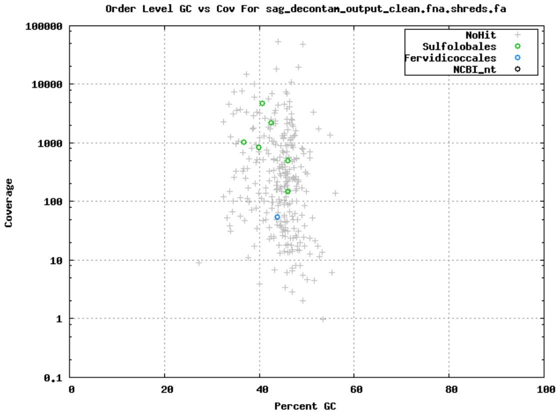
Phylum



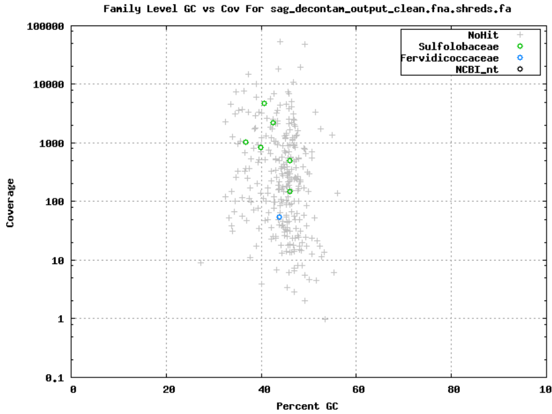
Class



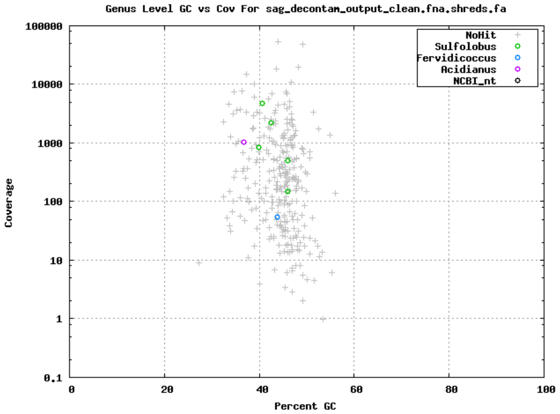
Order



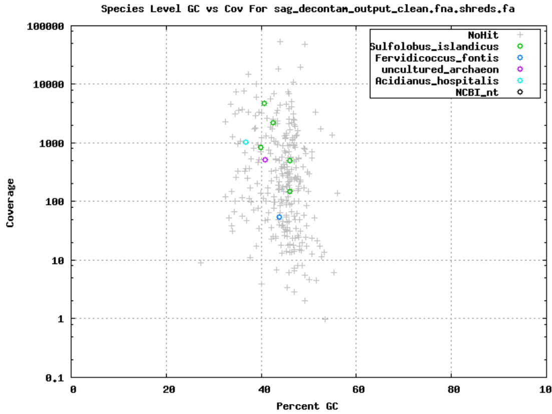
Family



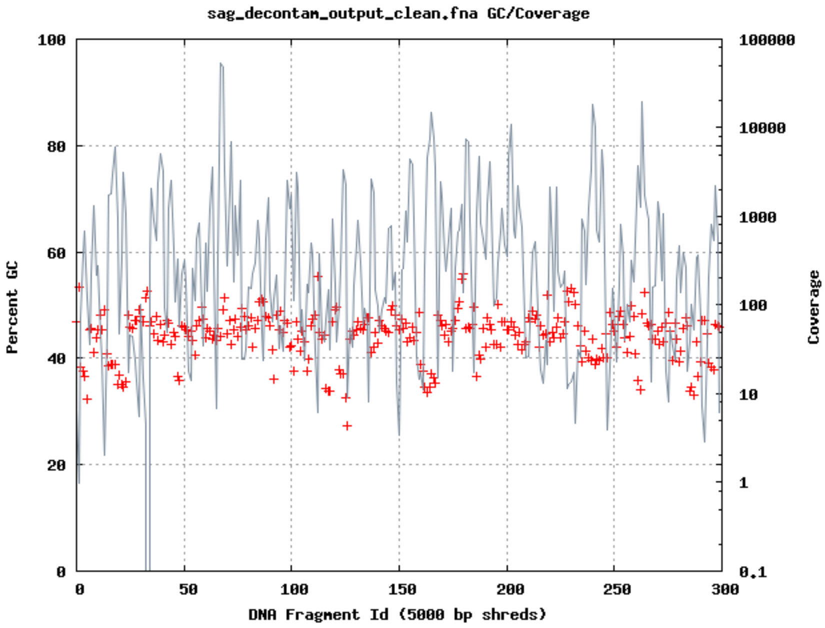
Genus



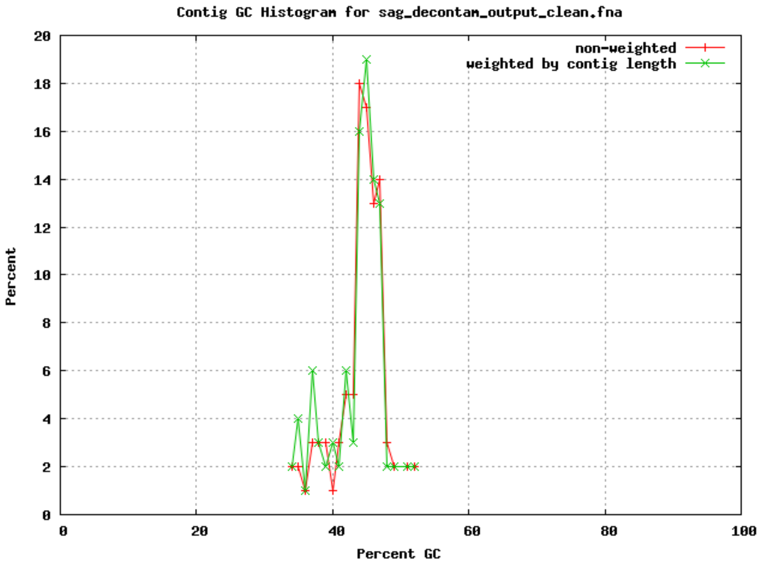
Species



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.

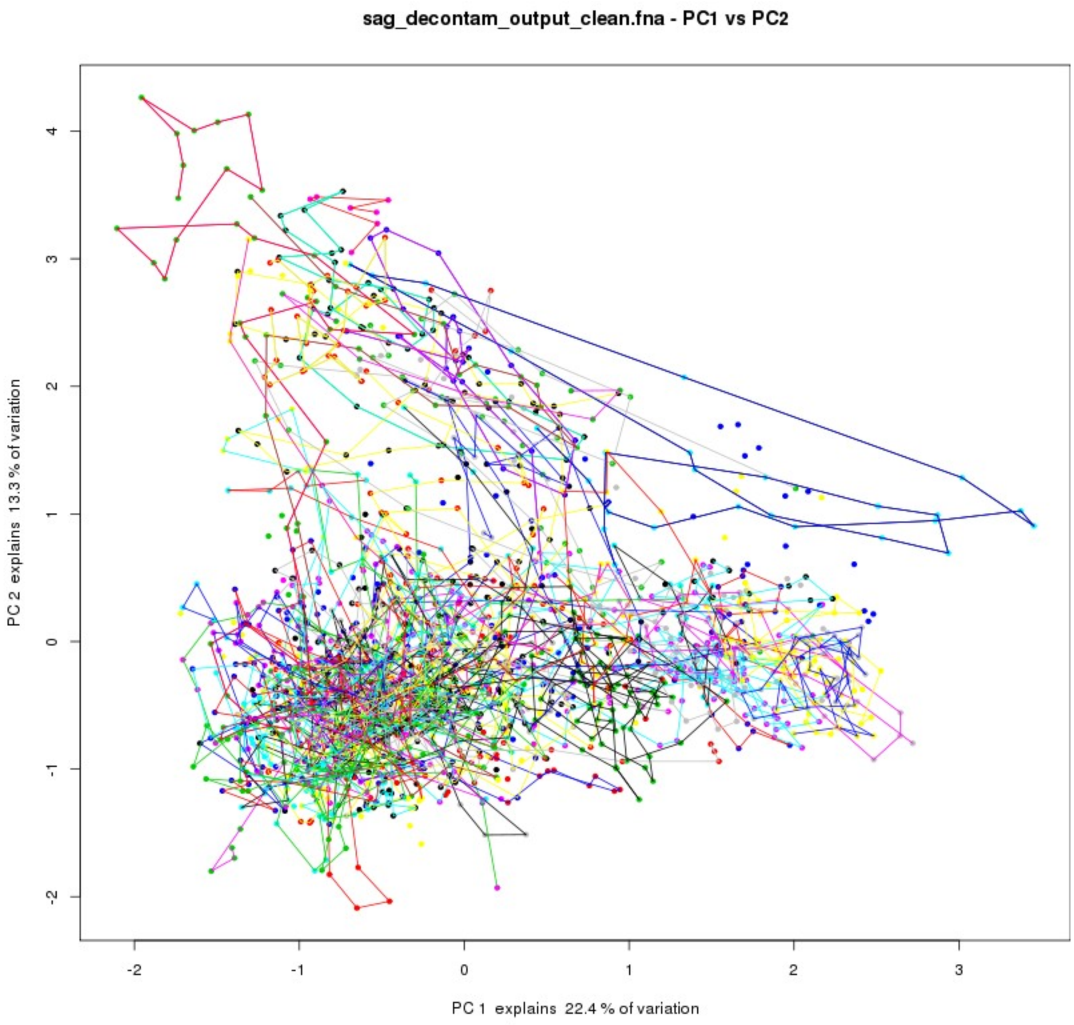


List of contigs and average percent GC, grouped in bins of 5:

Pct GC Bin	Contig Name
30	NODE_50_length_11953_cov_64.834_ID_103, NODE_51_length_11570_cov_46.4097_ID_105
35	NODE_1_length_48388_cov_1260.61_ID_1, NODE_7_length_25249_cov_4745.46_ID_15, NODE_14_length_20857_cov_983.589_ID_31, NODE_26_length_15259_cov_252.293_ID_55, NODE_27_length_14976_cov_1564.85_ID_57, NODE_37_length_13548_cov_1299.02_ID_77, NODE_44_length_13176_cov_352.993_ID_91, NODE_52_length_10852_cov_223.65_ID_107, NODE_56_length_9607_cov_108.349_ID_119, NODE_59_length_9092_cov_10.0214_ID_125, NODE_72_length_7081_cov_168.249_ID_159, NODE_101_length_4753_cov_23.1882_ID_221

40	NODE_3_length_39247_cov_3632.16_ID_5, NODE_5_length_27364_cov_2410.82_ID_9, NODE_10_length_23717_cov_199.924_ID_23, NODE_16_length_19700_cov_944.365_ID_35, NODE_17_length_18883_cov_1033.13_ID_37, NODE_18_length_17340_cov_81.6797_ID_39, NODE_20_length_17203_cov_148.877_ID_43, NODE_29_length_14470_cov_110.999_ID_61, NODE_30_length_14223_cov_107.524_ID_63, NODE_41_length_13273_cov_289.163_ID_85, NODE_43_length_13225_cov_450.171_ID_89, NODE_46_length_13007_cov_454.413_ID_95, NODE_47_length_12919_cov_674.695_ID_97, NODE_53_length_9943_cov_420.608_ID_113, NODE_57_length_9580_cov_221.933_ID_121, NODE_58_length_9214_cov_193.345_ID_123, NODE_60_length_9082_cov_130.348_ID_127, NODE_64_length_7999_cov_320.139_ID_147, NODE_65_length_7997_cov_110.828_ID_149, NODE_66_length_7937_cov_19.1456_ID_151, NODE_67_length_7802_cov_61.8645_ID_139, NODE_68_length_7739_cov_44.7738_ID_153, NODE_70_length_7169_cov_91.6503_ID_155, NODE_74_length_6976_cov_1366.36_ID_165, NODE_76_length_6931_cov_913.73_ID_169, NODE_77_length_6917_cov_392.362_ID_171, NODE_80_length_6492_cov_1687.21_ID_181, NODE_86_length_5852_cov_100.628_ID_193, NODE_87_length_5846_cov_11.9118_ID_195, NODE_90_length_5275_cov_214.787_ID_201, NODE_96_length_4986_cov_19.1929_ID_213, NODE_102_length_4732_cov_175.328_ID_223
45	NODE_2_length_42579_cov_1154.93_ID_3, NODE_4_length_30881_cov_1188.87_ID_7, NODE_6_length_25632_cov_162.377_ID_11, NODE_8_length_24152_cov_14480.5_ID_19, NODE_11_length_22659_cov_2280.18_ID_25, NODE_12_length_22470_cov_287.178_ID_27, NODE_13_length_22191_cov_206.643_ID_29, NODE_15_length_20432_cov_93.9323_ID_33, NODE_19_length_17203_cov_916.723_ID_41, NODE_21_length_16636_cov_536.379_ID_45, NODE_22_length_15784_cov_343.077_ID_47, NODE_23_length_15670_cov_17.7896_ID_49, NODE_24_length_15668_cov_207.137_ID_51, NODE_28_length_14875_cov_2088.75_ID_59, NODE_31_length_14180_cov_172.831_ID_65, NODE_32_length_14117_cov_200.176_ID_67, NODE_33_length_14019_cov_557.217_ID_69, NODE_34_length_14000_cov_266.161_ID_71, NODE_35_length_13605_cov_61.1829_ID_73, NODE_36_length_13569_cov_365.087_ID_75, NODE_38_length_13543_cov_1623.87_ID_79, NODE_40_length_13343_cov_55.4165_ID_83, NODE_42_length_13235_cov_628.645_ID_87, NODE_45_length_13056_cov_724.316_ID_93, NODE_48_length_12906_cov_44.4407_ID_99, NODE_49_length_11999_cov_172.444_ID_101, NODE_54_length_9847_cov_139.583_ID_115, NODE_55_length_9699_cov_125.457_ID_117, NODE_61_length_9011_cov_73.1975_ID_129, NODE_62_length_8846_cov_621.613_ID_135, NODE_63_length_8444_cov_1017.3_ID_137, NODE_69_length_7467_cov_18.2406_ID_145, NODE_71_length_7084_cov_110.232_ID_157, NODE_73_length_7018_cov_33.6348_ID_109, NODE_75_length_6942_cov_10.257_ID_167, NODE_78_length_6900_cov_198.049_ID_173, NODE_79_length_6682_cov_11.8743_ID_175, NODE_82_length_6154_cov_1047.56_ID_185, NODE_83_length_5901_cov_22.8141_ID_187, NODE_84_length_5884_cov_58.8535_ID_189, NODE_85_length_5866_cov_765.863_ID_191, NODE_89_length_5451_cov_26.4863_ID_199, NODE_91_length_5266_cov_4.81328_ID_203, NODE_92_length_5242_cov_12838.4_ID_205, NODE_93_length_5216_cov_9.27262_ID_207, NODE_95_length_5116_cov_4.17961_ID_211, NODE_98_length_4963_cov_36.0092_ID_217, NODE_99_length_4892_cov_9.8987_ID_219 NODE_100_length_4768_cov_9.43539_ID_131
50	NODE_9_length_24025_cov_437.031_ID_21, NODE_39_length_13473_cov_8.48562_ID_81, NODE_88_length_5606_cov_10.7062_ID_197, NODE_94_length_5137_cov_2.97029_ID_209

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	22.38 %
archaea	59.67 %

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 24,950,880 reads totaling 3,742.6 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 99 contigs in 99 scaffolds, totalling 1.3 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, 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.