uncultured Roseibacterium sp.

Single–cell Assembly QC Report 04/16/2014

1. Project Information

Program	Microbial/CSP 2012
PMO Project	0
Seq Proj ID	1027172
Sequencing Project Name	Roseibacterium sp. HL7711_P4H5 JGI 000147CP–J17
JGI Project ID	0

2. Read Statistics

Illumina Std PE Statistics

File name	7667.7.80862.TCGAAG.fastq
Library	TNHA
Number of reads	24,902,674
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	24,902,674	100
Contam removed	230	0.0
Artifact removed	411,992	1.7
Total removed	4,902,674	19.7
Total remaining	20,000,000	80.3

List of Contaminants Removed

Description	Num Reads	Pct Reads
gi|357579577|Canis_lupus_familiaris_chr3	196	0.00
human_chr2	194	0.00
gi|357579535|Canis_lupus_familiaris_chr20	16	0.00
gi|357579571|Canis_lupus_familiaris_chr5	8	0.00
human_chr5	4	0.00
human_chr13	2	0.00
human_chr14	2	0.00

human_chr10	2	0.00
human_chr7	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

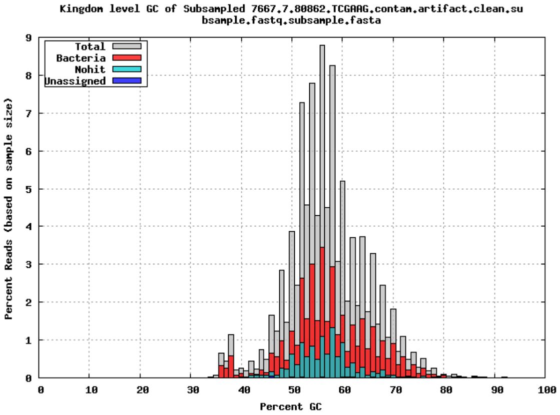
Description	Num Reads	Pct Reads
Input	24,902,674	100
Contam identified	10	0.0

List of Contaminants Identified

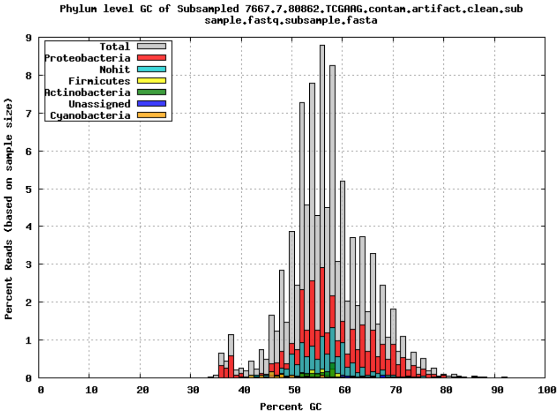
Description	Num Reads	Pct Reads
Delftia	8	0.00
Pseudomonas	2	0.00

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

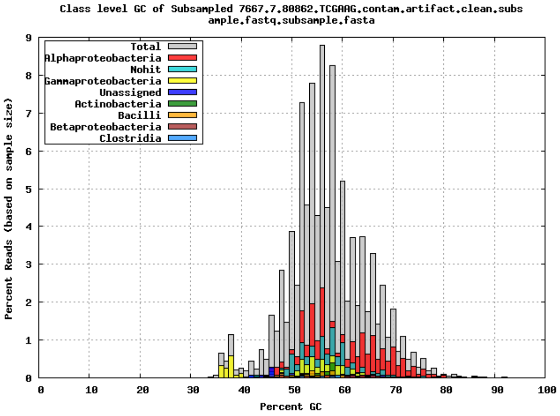
Kingdom



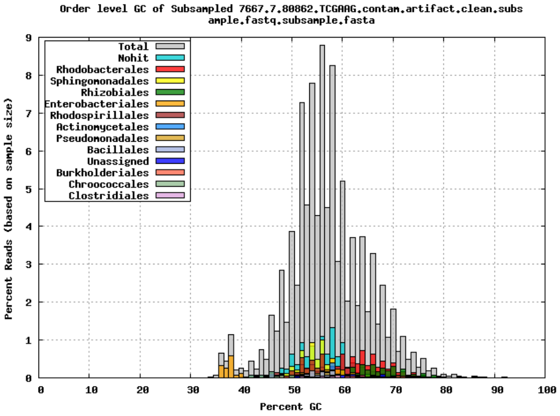
Phylum



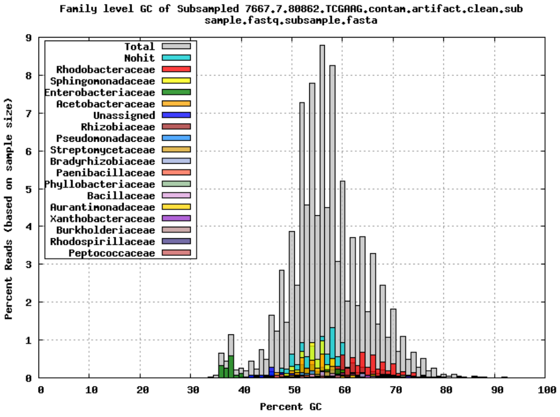
Class



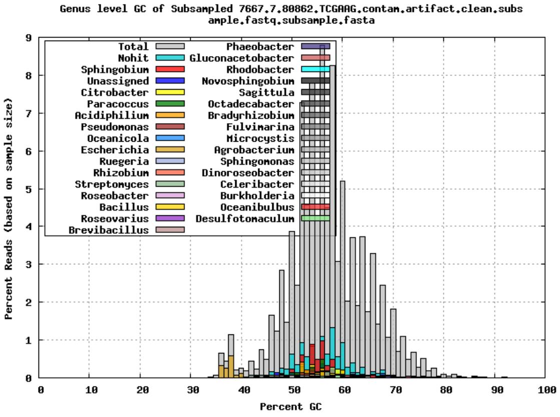
Order



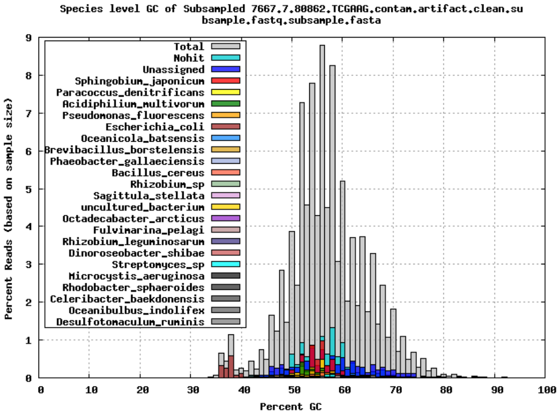
Family



Genus



Species



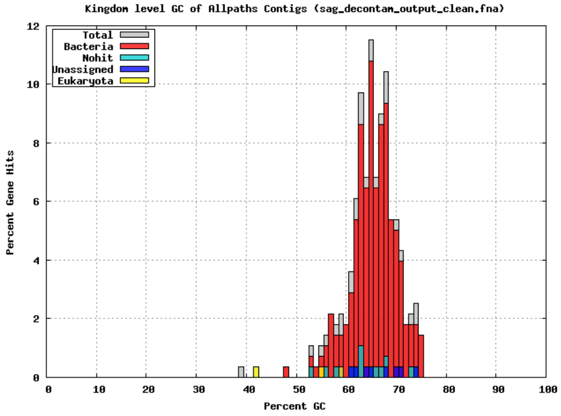
4. Assembly Statistics

Assembly method	SPAdes with auto decontamination
Scaffold total	26
Contig total	26
Scaffold sequence length	259.5 kb
Contig sequence length	259.5 kb (0.0% gap)
Scaffold N/L50	6/18.9 kb
Contig N/L50	6/18.9 kb
Largest Contig	31.5 kb
Number of scaffolds >50 kb	0
Pct of genome in scaffolds >50 kb	0.0
Pct of reads asssembled (raw)	69.4
Pct of reads asssembled (decontam)	2.0

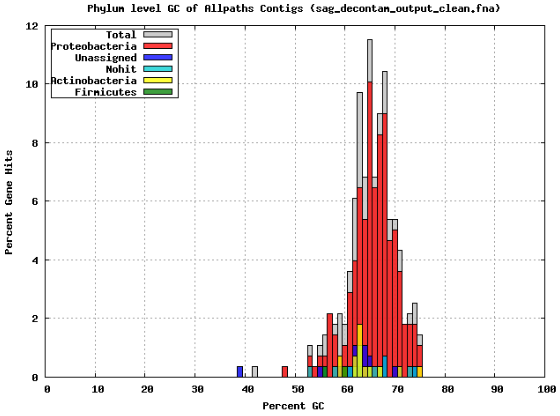
5. Assembly QC Results

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

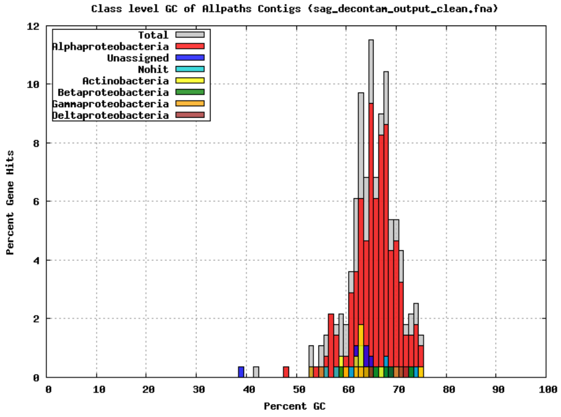
Kingdom



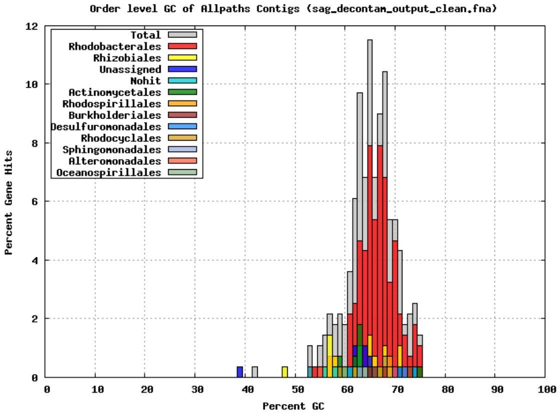
Phylum



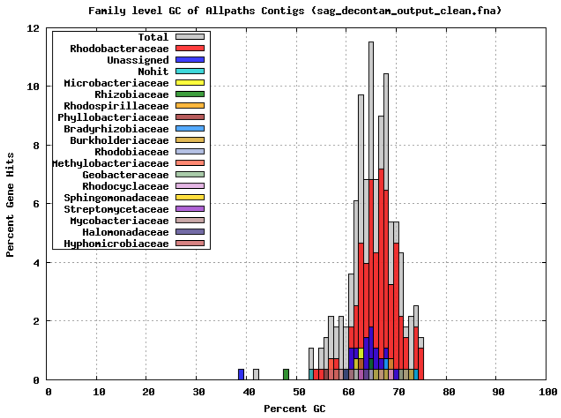
Class



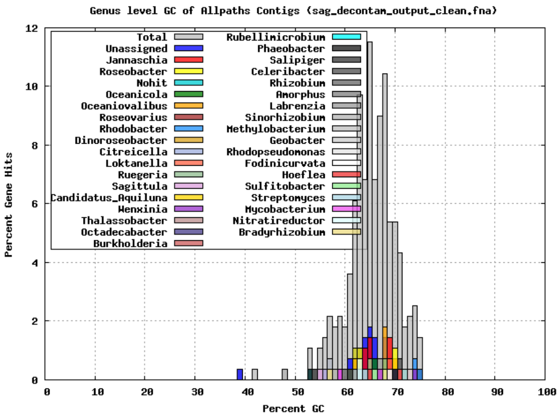
Order



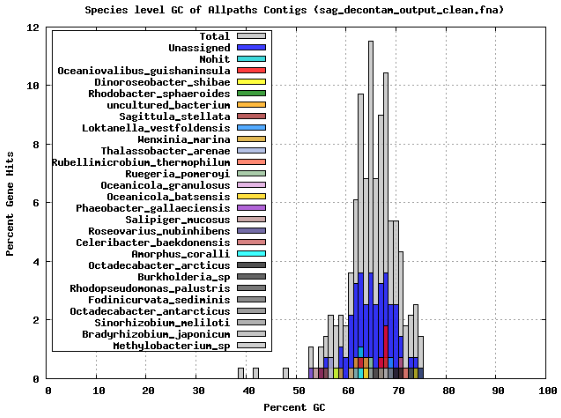
Family



Genus

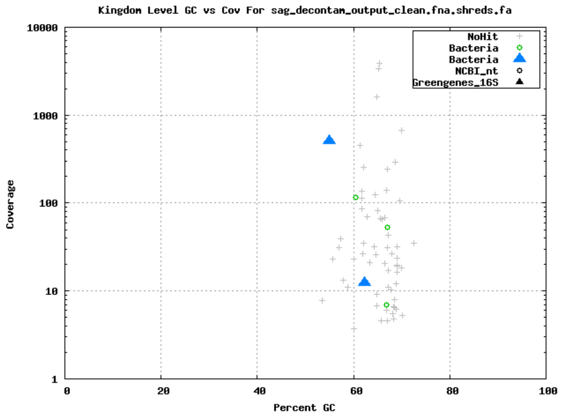


Species

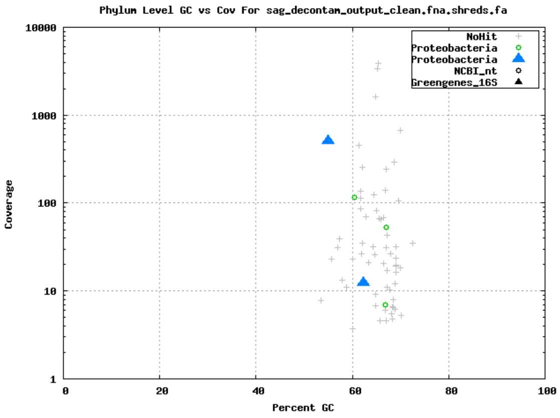


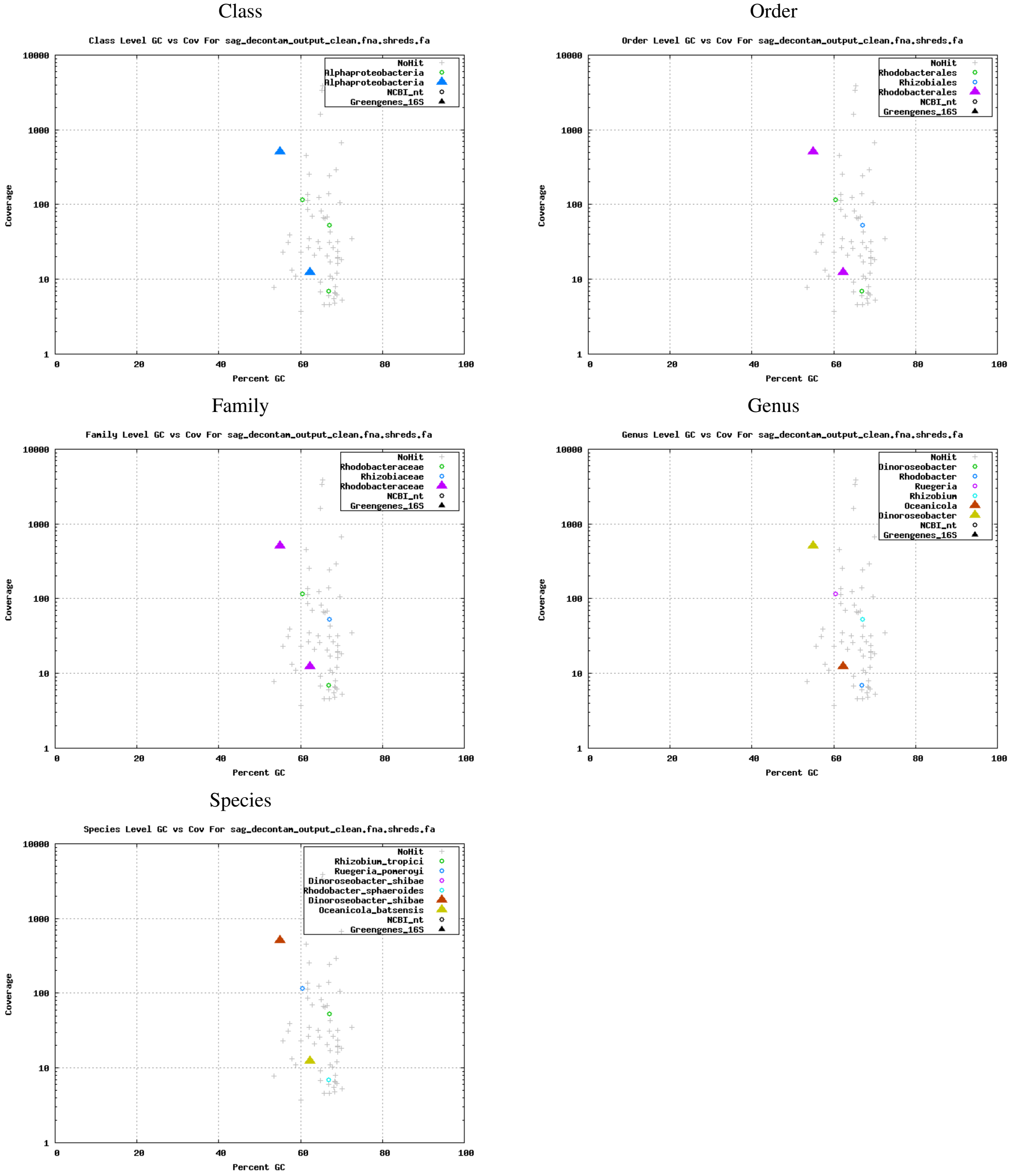
GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

Kingdom



Phylum




Class

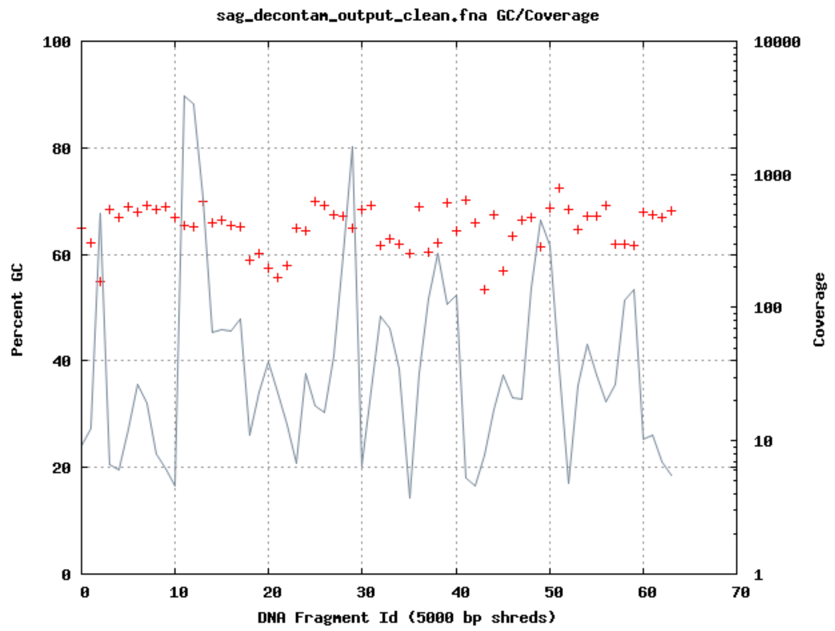
Order

Family

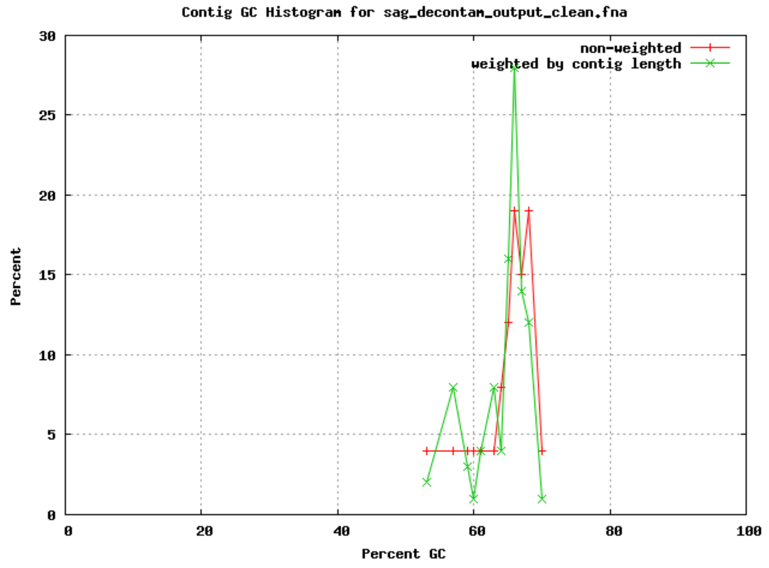
Genus

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
50	NODE_65_length_4303_cov_5.37429_ID_133
55	NODE_13_length_20530_cov_15.9907_ID_17, NODE_46_length_7278_cov_127.929_ID_95
60	NODE_16_length_19739_cov_35.6381_ID_33, NODE_32_length_10525_cov_50.6621_ID_67, NODE_43_length_7900_cov_13.5449_ID_89, NODE_72_length_3764_cov_5.70154_ID_145 NODE_101_length_2302_cov_2.78416_ID_203

65	NODE_5_length_31501_cov_81.4429_ID_9, NODE_11_length_21042_cov_88.0741_ID_25, NODE_17_length_19687_cov_151.513_ID_35, NODE_19_length_18855_cov_45.5955_ID_43, NODE_20_length_18692_cov_21.456_ID_45, NODE_23_length_17086_cov_11.2826_ID_49, NODE_27_length_14597_cov_1805.56_ID_57, NODE_44_length_7778_cov_6.7198_ID_91, NODE_55_length_5776_cov_4.32145_ID_113, NODE_58_length_5369_cov_12.7631_ID_119, NODE_64_length_4460_cov_3.93439_ID_131, NODE_73_length_3695_cov_3.84451_ID_147, NODE_84_length_3229_cov_2.89382_ID_169, NODE_97_length_2464_cov_4.18016_ID_195, NODE_99_length_2419_cov_3.8253_ID_199, NODE_106_length_2112_cov_2.84103_ID_213 NODE_109_length_2087_cov_2.84646_ID_219
70	NODE_102_length_2270_cov_3.25779_ID_205

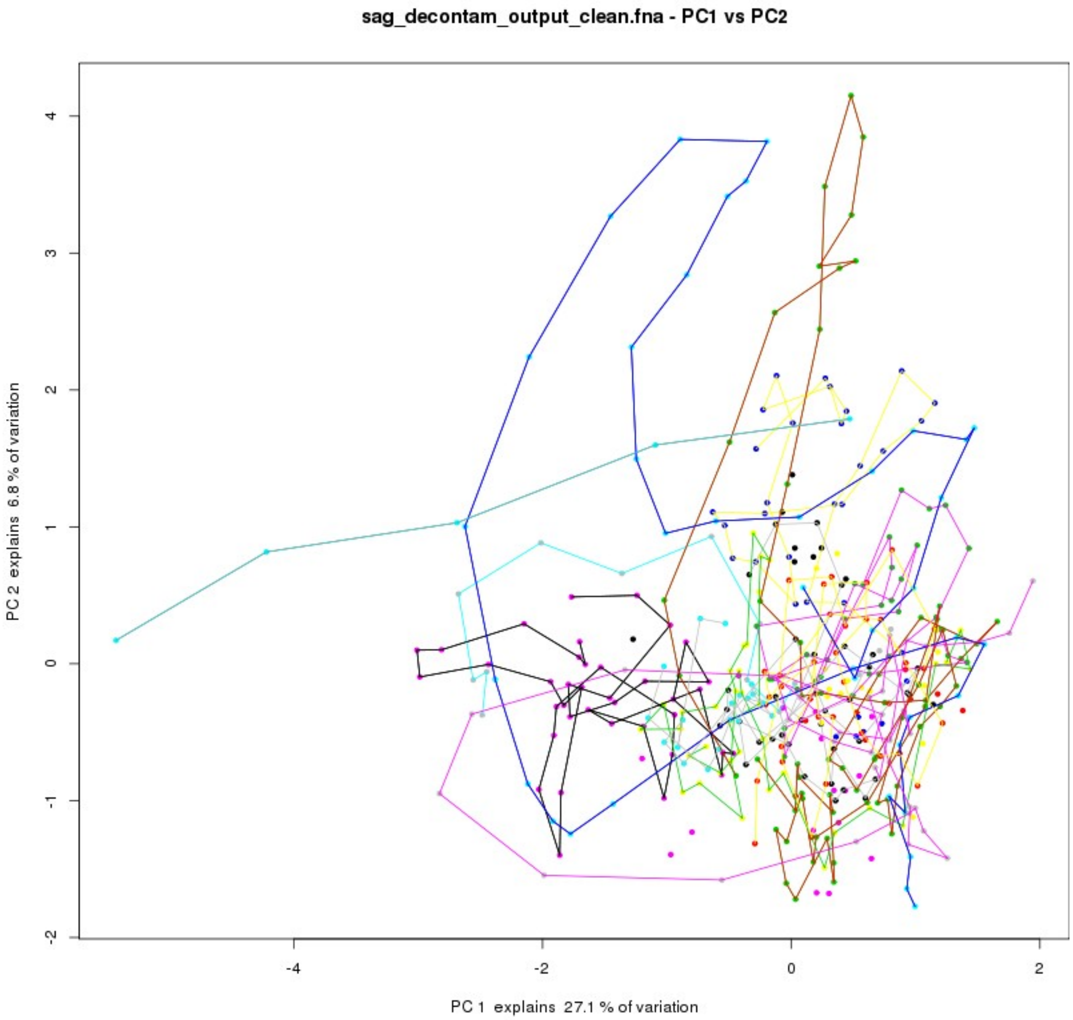
List of the top contig megablast hits against potential reagent and process contaminants.

Organism	Align Length (bp)	Pct Id	Contig Name
Ralstonia_solanacearum_str_PSI07_chromosome_complete_genome	198	91.92	NODE_46_length_7278_cov_127.929_ID_95

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

Organism	Align Length (bp)	Pct Id	Contig Name
250733_Dinoroseobacter_shibae_str_DFL_12_NC_009952_1_373473_374942	1,471	94.77	NODE_46_length_7278_cov_127.929_ID_95

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	3.2 %
archaea	3.43 %

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 24,902,674 reads totaling 3,735.4 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 26 contigs in 26 scaffolds, totalling 259.5 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.