

## 1. Project Information

Program	Microbial/CSP 2012
PMO Project	0
Seq Proj ID	1027175
Sequencing Project Name	Roseibacterium sp. HL7711_P4H5 JGI 000147CP-I17
JGI Project ID	0

## 2. Read Statistics

### Illumina Std PE Statistics

File name	7667.7.80862.TCGGCA.fastq
Library	TNHB
Number of reads	30,134,390
Sequencing depth <sup>†</sup>	904X
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	30,134,390	100
Contam removed	232	0.0
Artifact removed	1,090,028	3.6
Total removed	10,134,390	33.6
Total remaining	20,000,000	66.4

### List of Contaminants Removed

Description	Num Reads	Pct Reads
gi 357579577 Canis.lupus_familiaris_chr3	204	0.00
human_chr2	194	0.00
gi 357579535 Canis.lupus_familiaris_chr20	16	0.00
gi 357579522 Canis.lupus_familiaris_chr28	4	0.00
gi 357579571 Canis.lupus_familiaris_chr5	4	0.00
human_chr21	2	0.00
human_chr9	2	0.00

human_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

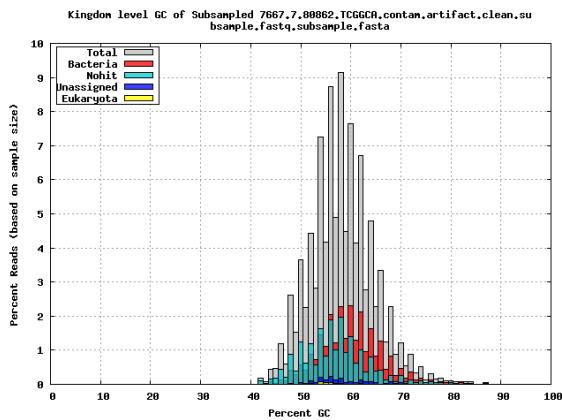
Description	Num Reads	Pct Reads
Input	30,134,390	100
Contam identified	4	0.0

#### List of Contaminants Identified

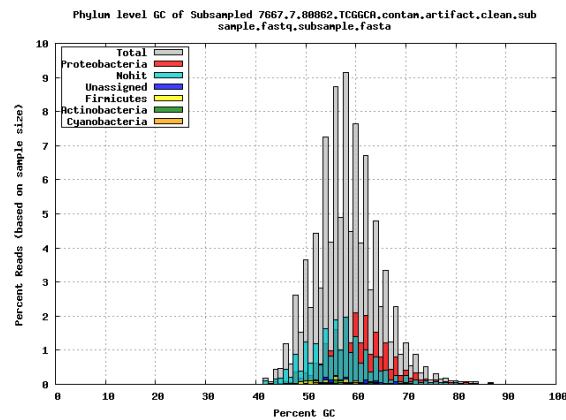
Description	Num Reads	Pct Reads
<i>Cupriavidus</i>	2	0.00
<i>Pseudomonas</i>	2	0.00

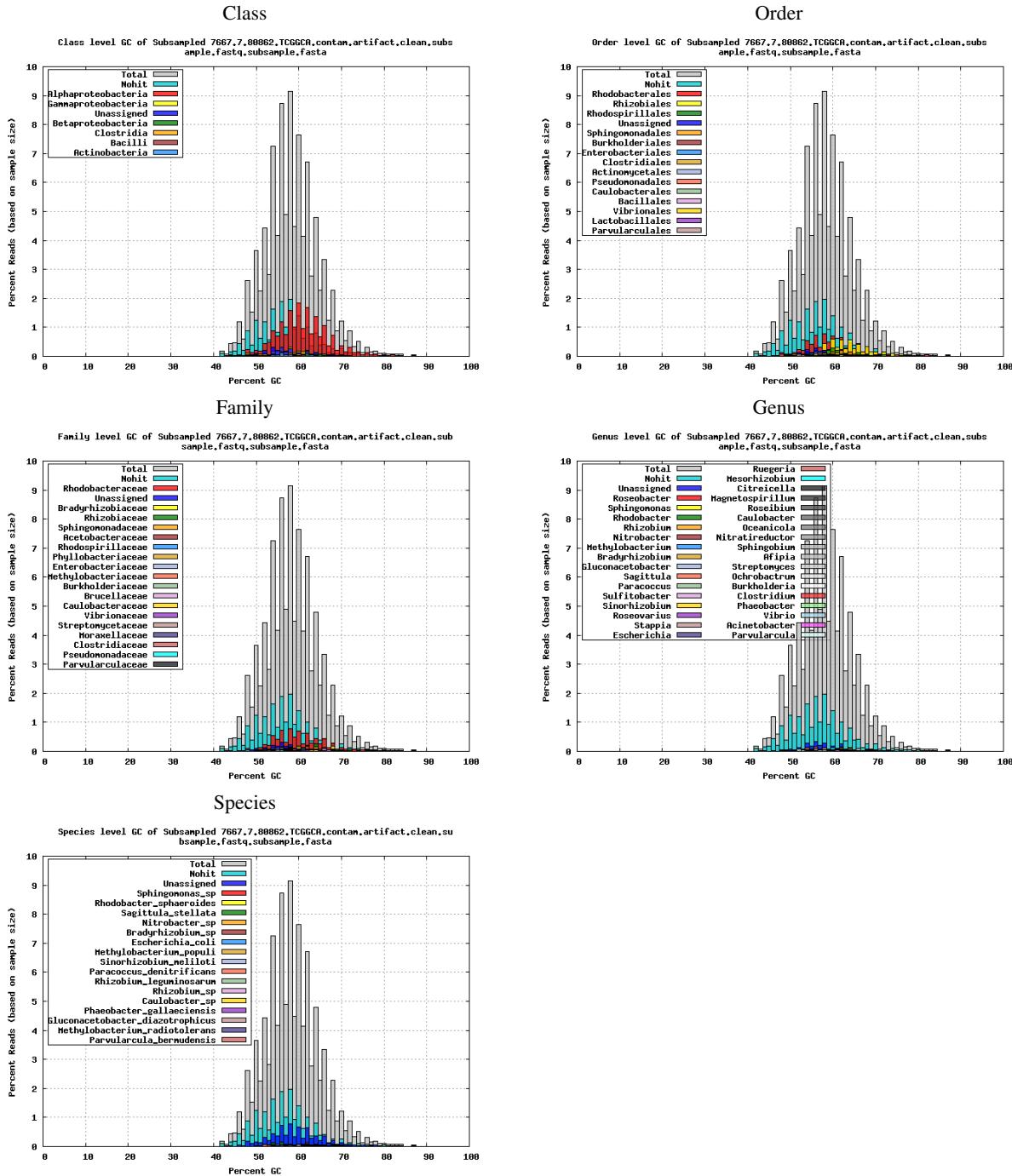
GC histogram of the reads subsampled to 10k, overlaid with GC of hits based on BLASTX, shown for different taxonomic levels.

#### Kingdom



#### Phylum



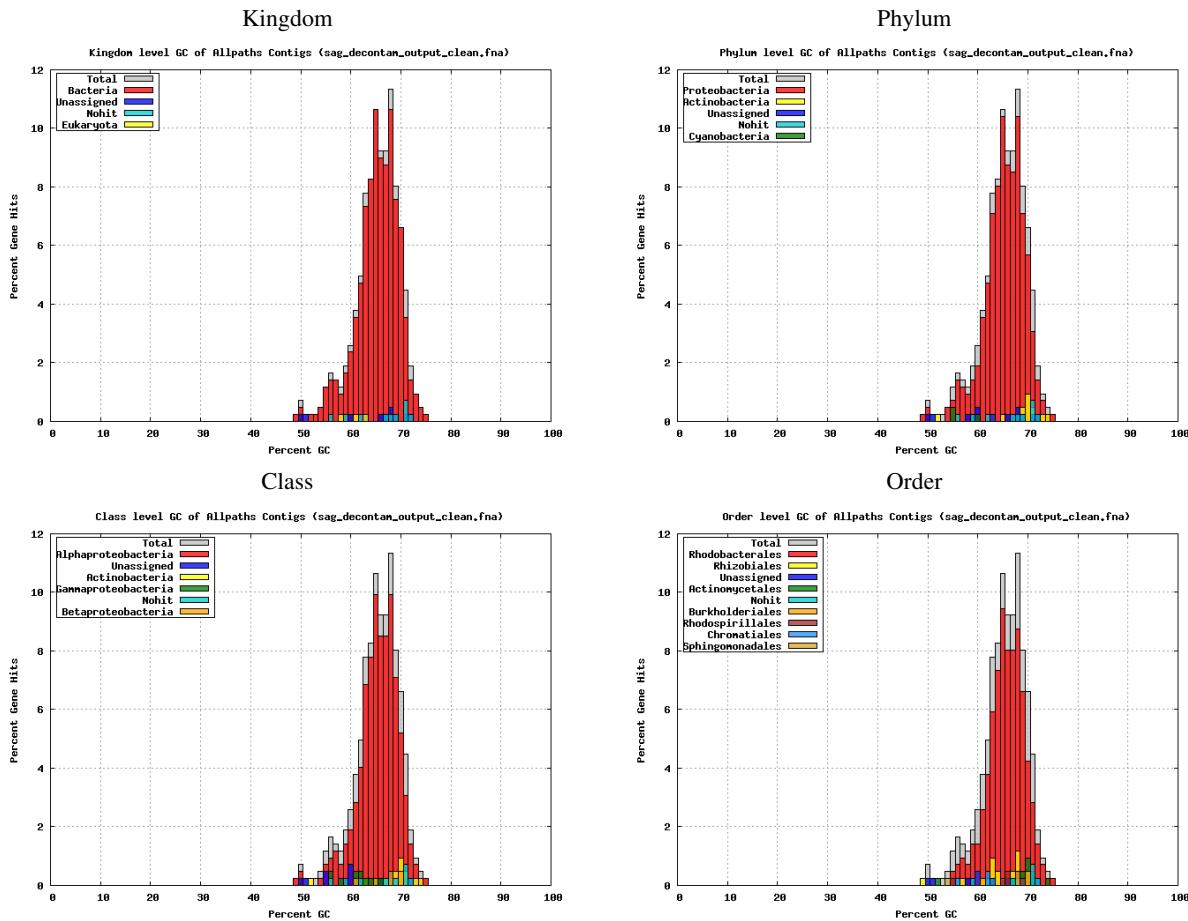


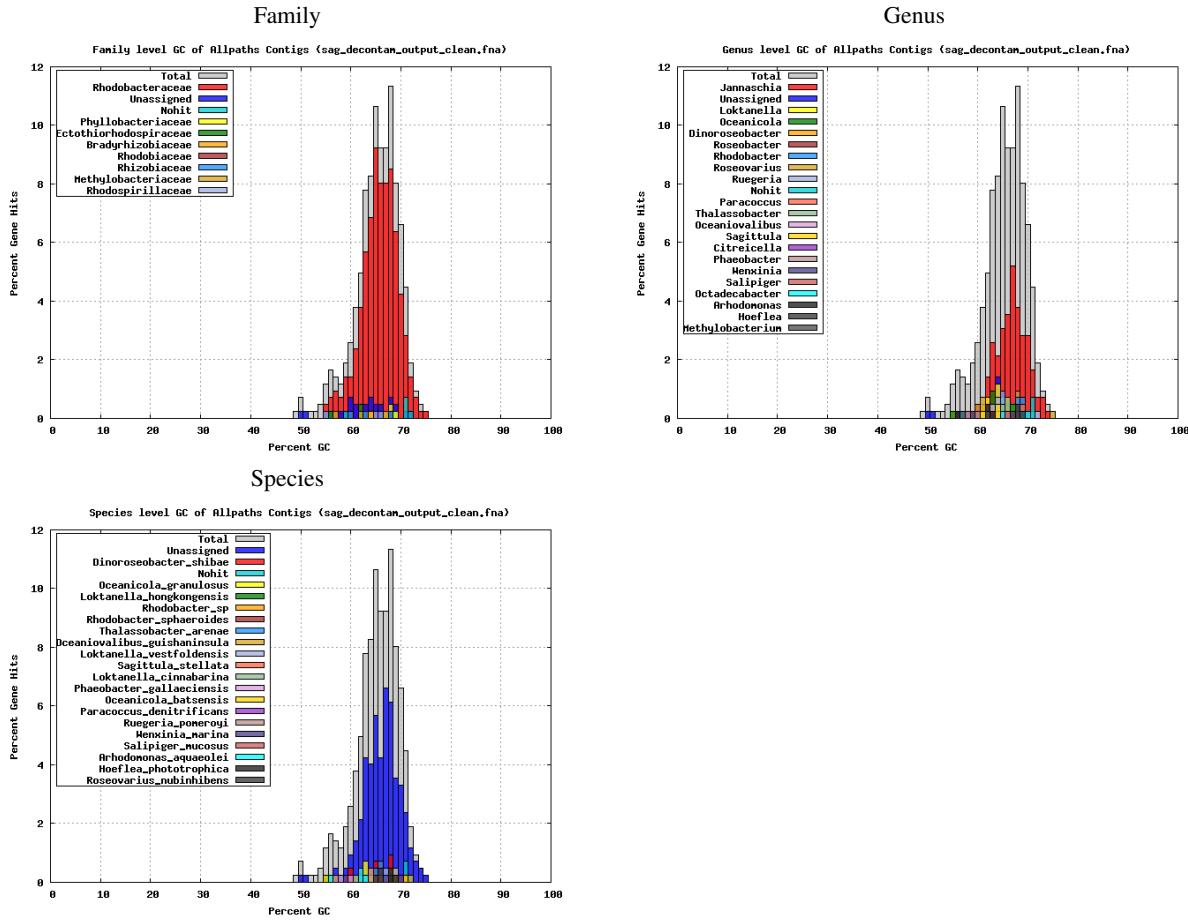
## 4. Assembly Statistics

Assembly method	SPAdes with auto decontamination
Scaffold total	50
Contig total	50
Scaffold sequence length	394.4 kb
Contig sequence length	394.4 kb ( 0.0% gap)
Scaffold N/L50	10/11.4 kb
Contig N/L50	10/11.4 kb
Largest Contig	35.5 kb
Number of scaffolds >50 kb	0
Pct of genome in scaffolds >50 kb	0.0
Pct of reads assembled (raw)	74.0
Pct of reads assembled (decontam)	8.8

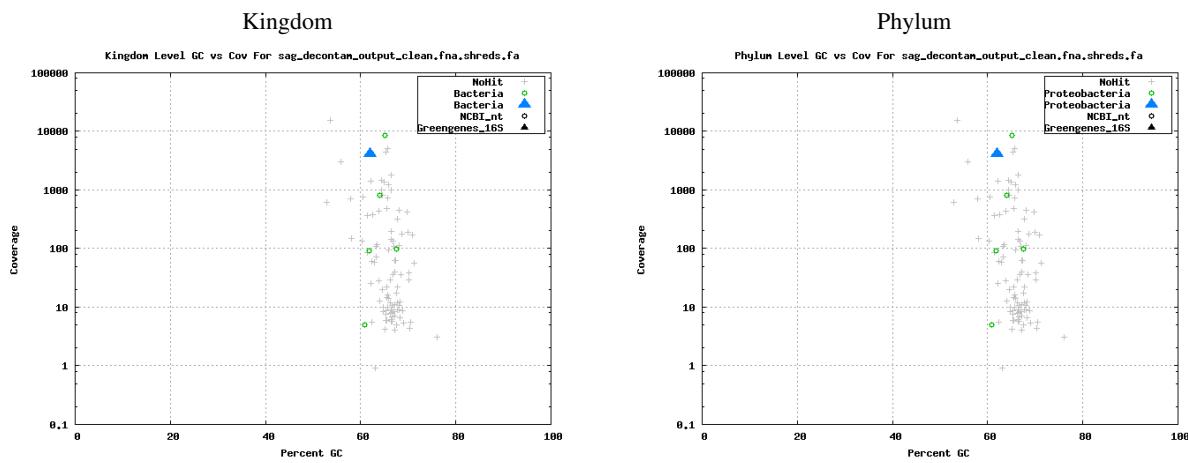
## 5. Assembly QC Results

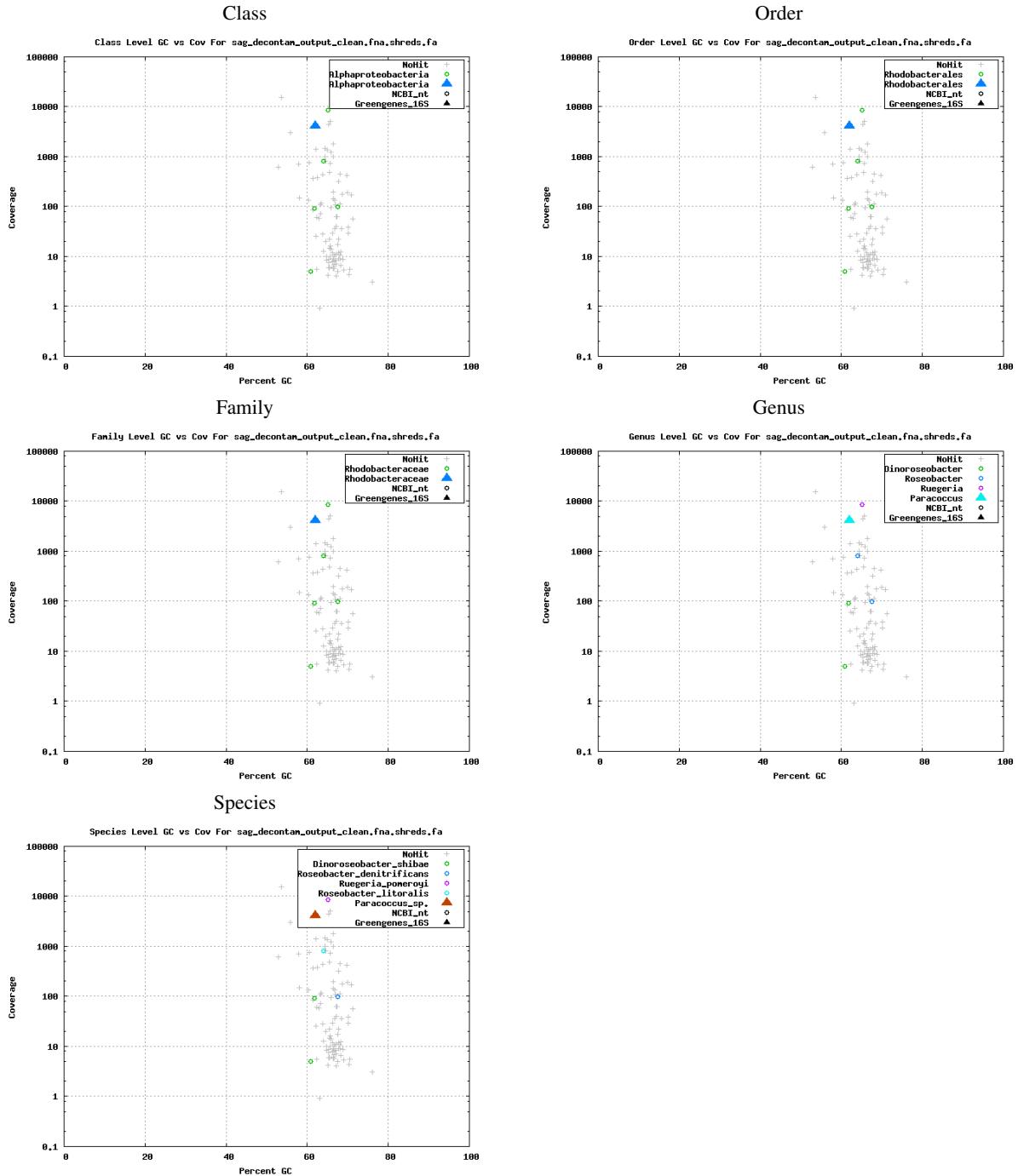
GC histogram of the predicted genes on each contig, overlaid with GC of hits based on BLASTP, shown for different taxonomic levels.



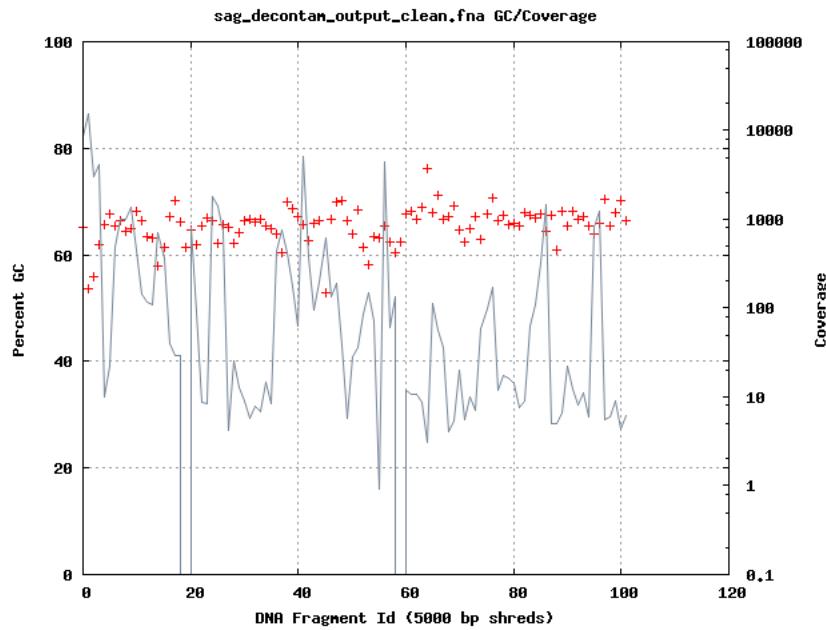


GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

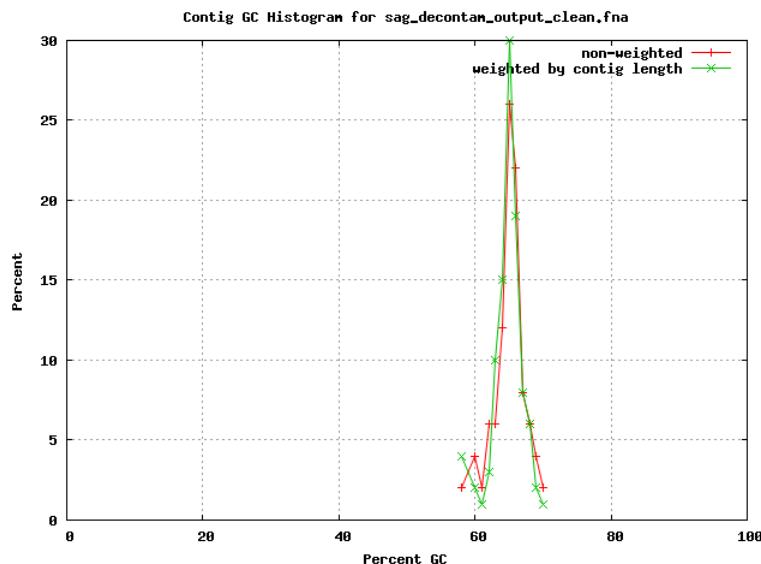




Coverage vs GC. Contigs were shredded into non-overlapping 5kbp and the GC of each shred was plotted as a point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.



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
55	NODE_10.length_14608 cov_6100.22.ID_19
60	NODE_3.length_25222 cov_48.2302.ID_5, NODE_4.length_25052 cov_154.32.ID_7, NODE_18.length_9782 cov_678.733.ID_35, NODE_21.length_9519 cov_162.692.ID_41, NODE_22.length_8913 cov_13.1384.ID_43, NODE_31.length_7339 cov_1045.29.ID_61, NODE_35.length_6415 cov_194.858.ID_69, NODE_42.length_5120 cov_95.4788.ID_83, NODE_45.length_4881 cov_925.746.ID_89, NODE_53.length_4109 cov_2748.83.ID_105, NODE_67.length_3626 cov_526.823.ID_133, NODE_69.length_3541 cov_267.457.ID_137, NODE_90.length_2690 cov_13.8398.ID_181, NODE_102.length_2155 cov_3.28619.ID_205

	NODE.106.length.2054_cov.3.58379.ID.213
65	NODE_1.length.35519_cov.420.89.ID.1, NODE_2.length.27308_cov.531.542.ID.3, NODE_6.length.20380_cov.6.57181.ID.11, NODE_7.length.19549_cov.301.656.ID.13, NODE_9.length.15454_cov.92.1083.ID.17, NODE_12.length.13428_cov.9.64825.ID.23, NODE_16.length.11354_cov.60.6445.ID.31, NODE_20.length.9611_cov.5.67089.ID.39, NODE_27.length.7896_cov.5.97832.ID.53, NODE_28.length.7875_cov.8.15742.ID.55, NODE_34.length.6861_cov.12.6674.ID.67, NODE_38.length.6061_cov.4.03397.ID.75, NODE_39.length.5885_cov.111.332.ID.77, NODE_40.length.5872_cov.5.35499.ID.79, NODE_41.length.5619_cov.5.48472.ID.81, NODE_44.length.4988_cov.6.15852.ID.87, NODE_49.length.4329_cov.4.64577.ID.97, NODE_52.length.4170_cov.3.86416.ID.103, NODE_56.length.3980_cov.14.6479.ID.111, NODE_58.length.3847_cov.84.2484.ID.115, NODE_68.length.3662_cov.9.07291.ID.135, NODE_76.length.3097_cov.4.71302.ID.155, NODE_79.length.2976_cov.3.25026.ID.161, NODE_80.length.2966_cov.5.95465.ID.163, NODE_81.length.2935_cov.476.837.ID.165, NODE_86.length.2768_cov.2.66126.ID.175, NODE_88.length.2736_cov.1363.3.ID.177, NODE_92.length.2478_cov.4.33966.ID.185, NODE_93.length.2435_cov.3.4395.ID.187, NODE_94.length.2424_cov.6.1068.ID.189, NODE_99.length.2290_cov.3.74452.ID.199, NODE_103.length.2130_cov.3.90265.ID.207 NODE_104.length.2123_cov.3.39603.ID.209
70	NODE.48.length.4400_cov.3.48446.ID.95

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

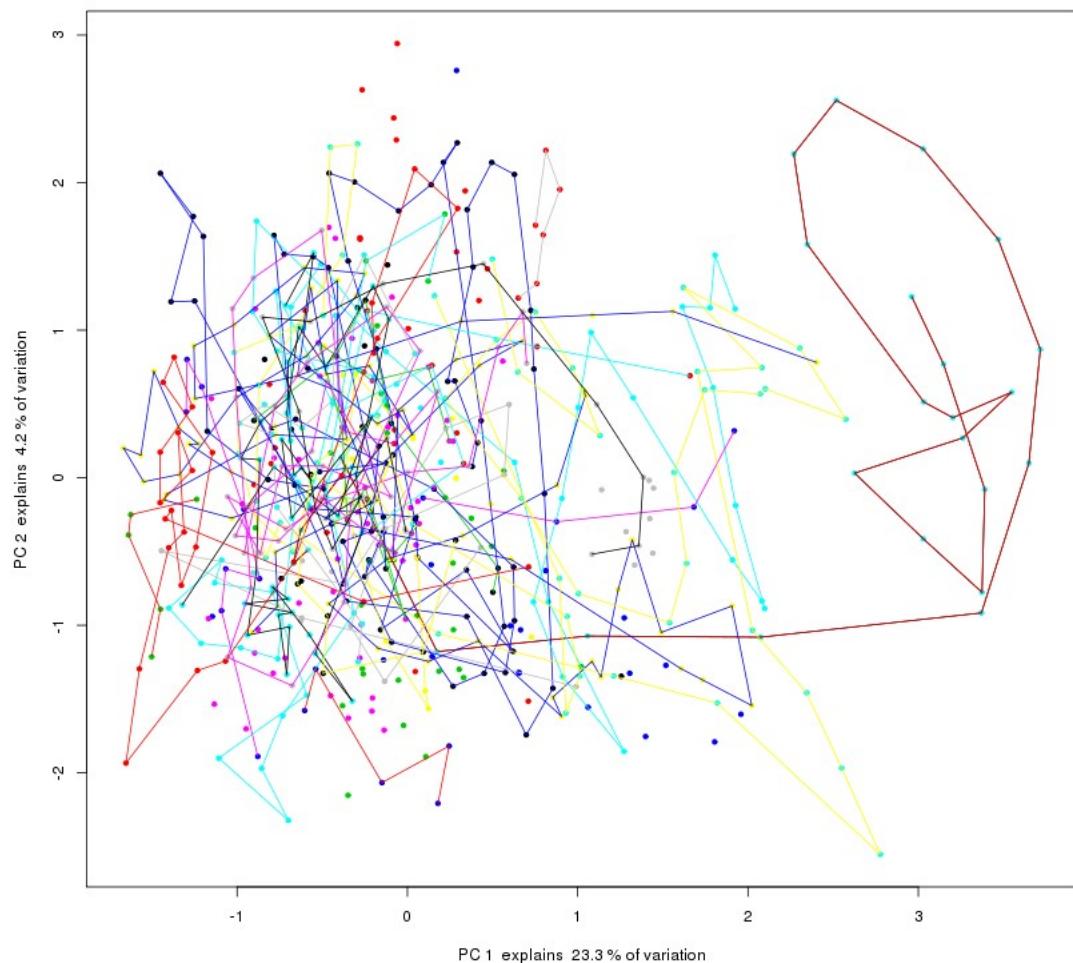
Organism	Align Length (bp)	Pct Id	Contig Name
<i>Ralstonia_solanacearum_str_PSI07_chromosome_complete_genome</i>	198	91.92	NODE_53.length.4109_cov.2748.83.ID.105

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

Organism	Align Length (bp)	Pct Id	Contig Name
<i>88947_Paracoccus_sp_OII8_AY312056_1_879_2345</i>	410	95.85	NODE_53.length.4109_cov.2748.83.ID.105

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	7.19 %
archaea	5.49 %

## 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 30,134,390 reads totaling 4,520.2 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 50 contigs in 50 scaffolds, totalling 394.4 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, 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.