

1. Project Information

Program	Microbial
JGI Sequencing Project ID	1042292
Sequencing Project Name	Nostoc sp. 996

2. Read Statistics

	Raw Reads	Filtered SubReads	Error Corrected Reads
Reads	790,275	258,626	14,724
Bases	1,742,692,286	863,525,044	101,006,414
Avg Read Length	2,205.2 +/- 2,722.6	3,338.9 +/- 2,158.9	6,860.0 +/- 3,685.3
Reads >5 kbp	85,053	35,611	10,396
Bases, reads >5 kbp	702,725,947	269,424,185	91,389,211
Avg Read Length, reads >5 kbp	8,262.2 +/- 3,411.5	7,565.8 +/- 2,604.5	8,790.8 +/- 2,400.6

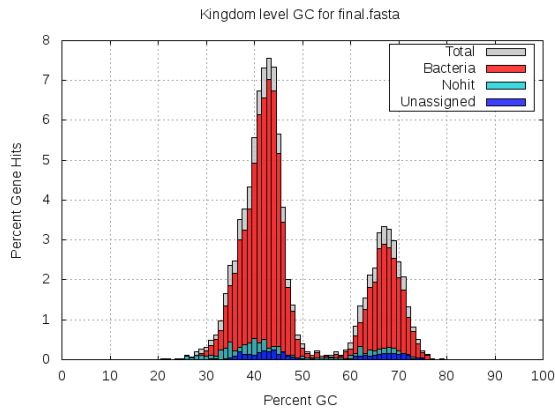
3. Assembly Statistics

Scaffold total	134
Contig total	134
Scaffold sequence length	11.004 mb
Contig sequence length	11.004 mb 0.000% gap
Scaffold N/L50	2/2.59 mb
Largest Contig	3,129.7 kbp
Number of scaffolds >50 kb	10
Pct of genome in scaffolds >50 kb	96.25%

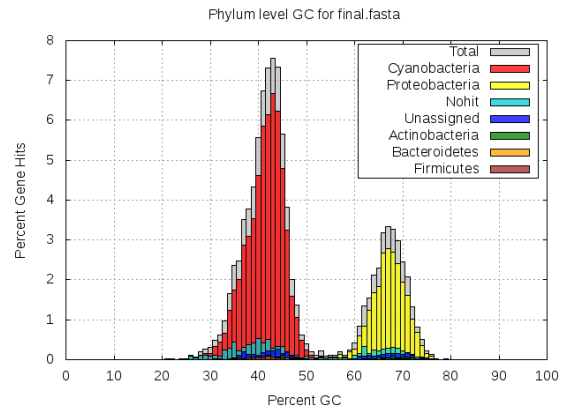
4. Assembly QC Results

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

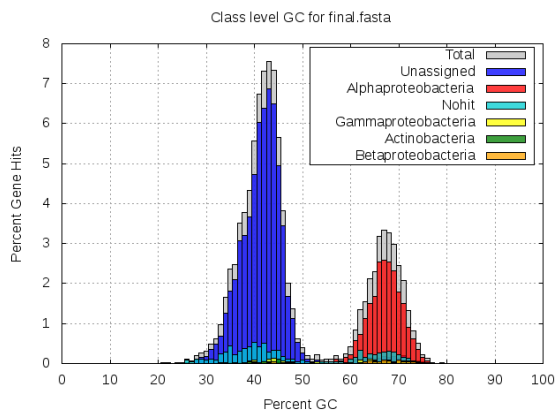
Kingdom



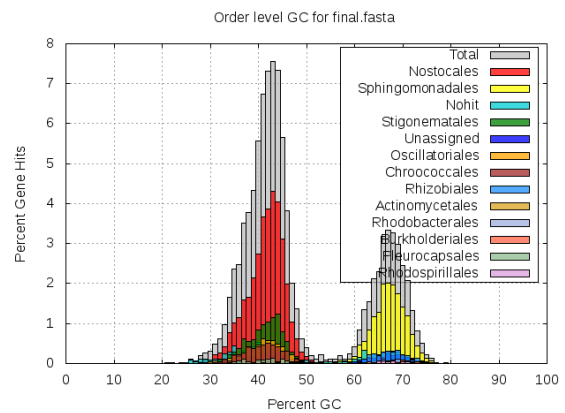
Phylum



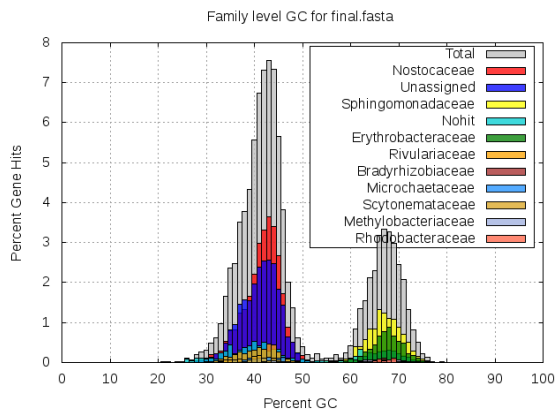
Class



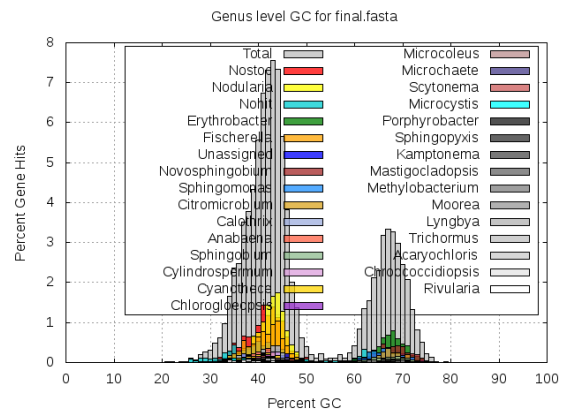
Order

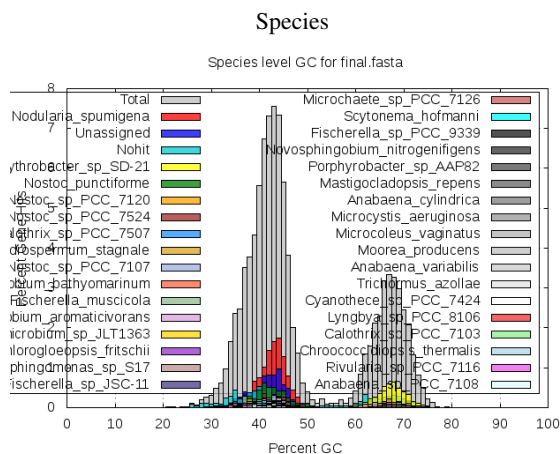


Family

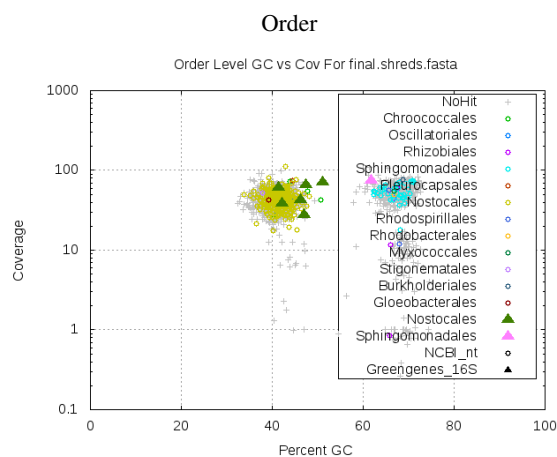
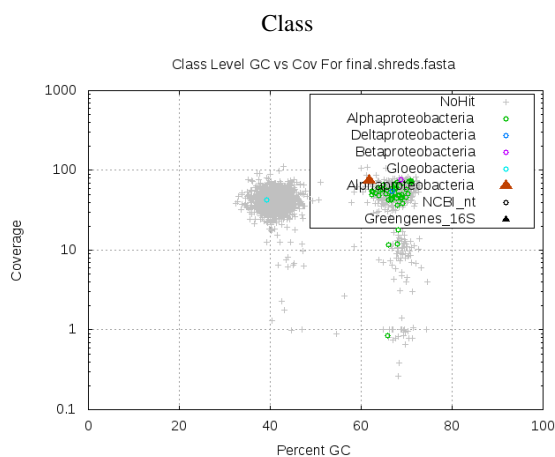
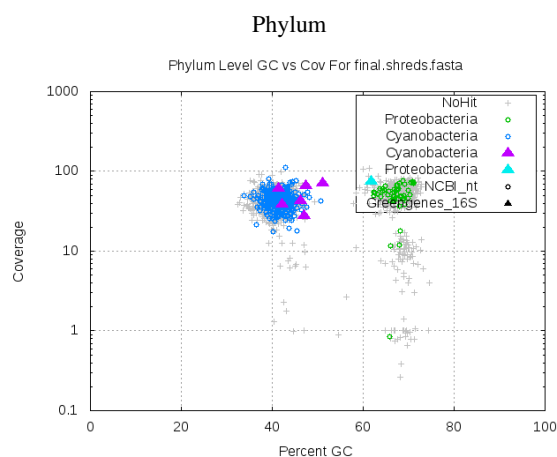
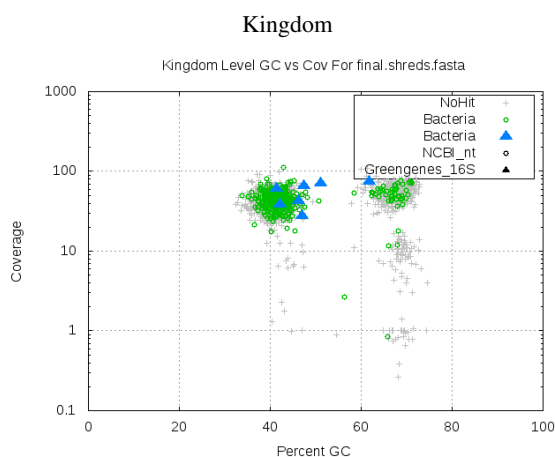


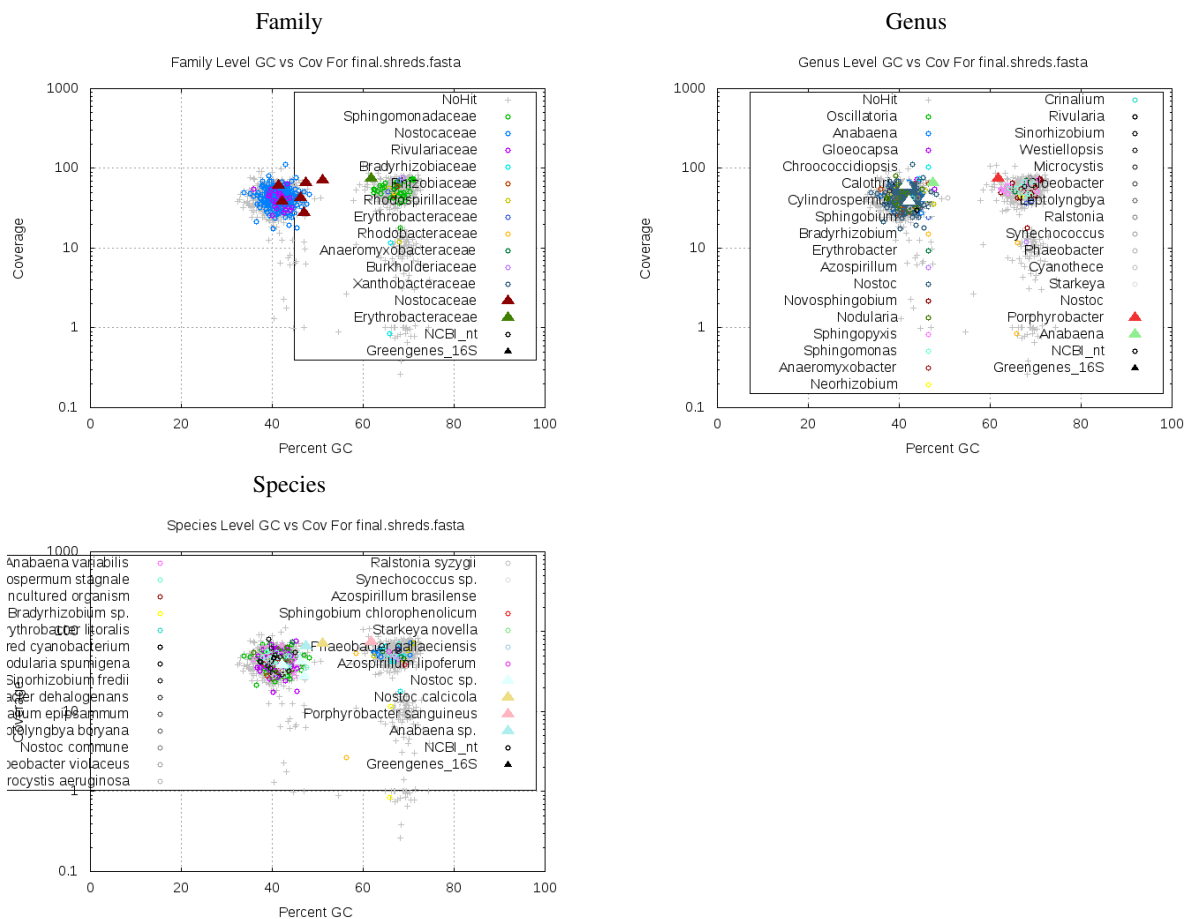
Genus



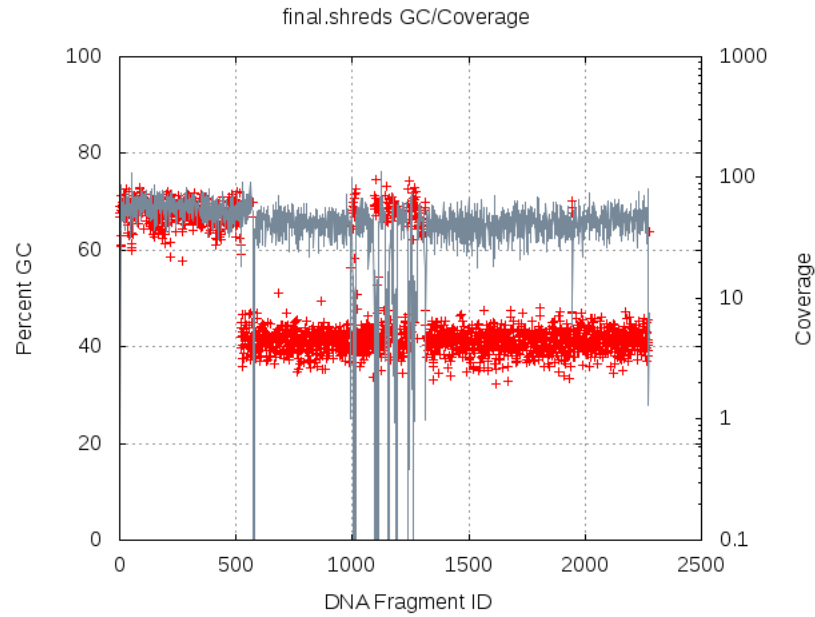


GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

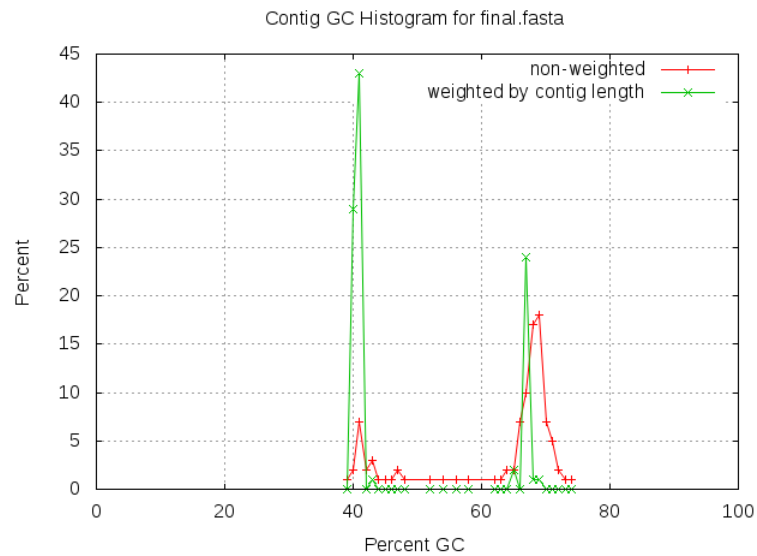




Coverage vs GC. Contigs were shredded into non-overlapping 5kbp and the GC of each shred was plotted as a point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.



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



List of contigs and average percent GC bin:

Pct GC Bin	Contig Name
35	unitig_183 quiver
40	unitig_1 quiver, unitig_2 quiver, unitig_3 quiver, unitig_4 quiver, unitig_5 quiver, unitig_6 quiver, unitig_7 quiver, unitig_9 quiver, unitig_10 quiver, unitig_11 quiver, unitig_12 quiver, unitig_13 quiver, unitig_14 quiver, unitig_19 quiver, unitig_26 quiver, unitig_34 quiver, unitig_52 quiver, unitig_59 quiver, unitig_72 quiver, unitig_85 quiver, unitig_152 quiver
45	unitig_31 quiver, unitig_62 quiver, unitig_74 quiver, unitig_134 quiver, unitig_302 quiver, unitig_304 quiver, unitig_396 quiver
50	unitig_29 quiver, unitig_167 quiver

55	unitig_191 quiver, unitig_405 quiver
60	unitig_63 quiver, unitig_65 quiver, unitig_66 quiver, unitig_73 quiver, unitig_307 quiver
65	unitig_0 quiver, unitig_8 quiver, unitig_15 quiver, unitig_16 quiver, unitig_17 quiver, unitig_20 quiver, unitig_21 quiver, unitig_22 quiver, unitig_23 quiver, unitig_24 quiver, unitig_25 quiver, unitig_27 quiver, unitig_28 quiver, unitig_32 quiver, unitig_33 quiver, unitig_35 quiver, unitig_36 quiver, unitig_37 quiver, unitig_38 quiver, unitig_42 quiver, unitig_44 quiver, unitig_45 quiver, unitig_46 quiver, unitig_47 quiver, unitig_49 quiver, unitig_50 quiver, unitig_53 quiver, unitig_54 quiver, unitig_56 quiver, unitig_57 quiver, unitig_58 quiver, unitig_60 quiver, unitig_61 quiver, unitig_67 quiver, unitig_70 quiver, unitig_71 quiver, unitig_75 quiver, unitig_76 quiver, unitig_77 quiver, unitig_78 quiver, unitig_89 quiver, unitig_93 quiver, unitig_96 quiver, unitig_120 quiver, unitig_131 quiver, unitig_133 quiver, unitig_142 quiver, unitig_146 quiver, unitig_166 quiver, unitig_180 quiver, unitig_190 quiver, unitig_192 quiver, unitig_196 quiver, unitig_232 quiver, unitig_247 quiver, unitig_256 quiver, unitig_264 quiver, unitig_265 quiver, unitig_277 quiver, unitig_279 quiver, unitig_305 quiver, unitig_322 quiver, unitig_334 quiver, unitig_348 quiver, unitig_349 quiver, unitig_365 quiver, unitig_366 quiver, unitig_387 quiver, unitig_402 quiver, unitig_411 quiver, unitig_412 quiver, unitig_413 quiver, unitig_414 quiver
70	unitig_18 quiver, unitig_30 quiver, unitig_39 quiver, unitig_40 quiver, unitig_41 quiver, unitig_43 quiver, unitig_48 quiver, unitig_51 quiver, unitig_55 quiver, unitig_64 quiver, unitig_68 quiver, unitig_69 quiver, unitig_129 quiver, unitig_160 quiver, unitig_187 quiver, unitig_195 quiver, unitig_198 quiver, unitig_205 quiver, unitig_244 quiver, unitig_271 quiver, unitig_376 quiver, unitig_400 quiver, unitig_401 quiver

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

Organism	Align Length (bp)	Pct Id	Contig Name
<i>Ralstonia pickettii</i> 12J chromosome 1, complete	199	93.97	unitig_0 quiver
<i>Vibrio vulnificus</i> YJ016 DNA, chromosome 1, complete	153	94.77	unitig_191 quiver

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

Organism	Align Length (bp)	Pct Id	Contig Name
148662 <i>Nostoc</i> sp. 'Mollenhauer 1:1-067' str. Mollenhauer 1:1-067 DQ185207.1	1,424	98.17	unitig_1 quiver
255867 <i>Anabaena</i> sp. SKSF1 EU022718.1	1,492	90.01	unitig_14 quiver
39689 <i>Porphyrobacter sanguineus</i> str. ATCC 25661 AB 062106.1	1,479	96.28	unitig_0 quiver
4715 <i>Rasbo</i> AF007948.1	1,193	97.90	unitig_191 quiver

5. Data Access

The following sequence fasta files can be downloaded from our JGI portal website.

<http://www.jgi.doe.gov/genome-projects>

The annotation of the assembled contigs can be found within IMG.

<http://img.jgi.doe.gov>

6. Methods

Isolate Improved Draft

Genome sequencing and assembly

The draft genome of *Nostoc* sp. 996 was generated at the DOE Joint Genome Institute (JGI) using the Pacific Biosciences (PacBio) sequencing technology [1]. A >10kbp Pacbio SMRTbell™ library was constructed and sequenced on the PacBio RS platform, which generated 258,626 filtered subreads totaling 863.5 Mbp. All general aspects of library construction and sequencing performed at the JGI can be found at <http://www.jgi.doe.gov>. The raw reads were assembled using HGAP (version: 2.3.0, protocol version=2.3.0 method=RS_HGAP_Assembly.3, smrtpipe.py

v1.87.139483,) [2]. The final draft assembly contained 134 contigs in 134 scaffolds, totaling 11.004 Mbp in size. The input read coverage was 51.8X.

Genome annotation

Genes were identified with Prodigal [3], followed by one round of manual curation using GenePRIMP [4] for genomes in fewer than 10 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 [5] was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA [6]. 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 [7]. Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) platform [8] developed by the Joint Genome Institute, Walnut Creek, CA, USA [9].

1. Eid John, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* 2008
 2. Chin C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013
 3. 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.
 4. 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.
 5. 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.
 6. 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.
 7. INFERNAL. Inference of RNA alignments. <http://infernal.janelia.org>.
 8. The Integrated Microbial Genomes (IMG) platform. <http://img.jgi.doe.gov>.
 9. 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.
-