

1. Project Information

| | |
|---------------------------|------------------|
| Program | Microbial |
| JGI Sequencing Project ID | 1022154 |
| Sequencing Project Name | Pantoea sp. 9140 |

2. Read Statistics

| | Raw Reads | Filtered SubReads | Error Corrected Reads |
|-------------------------------|-----------------|-------------------|-----------------------|
| Reads | 706,972 | 165,984 | 11,728 |
| Bases | 1,846,269,929 | 604,280,908 | 70,238,265 |
| Avg Read Length | 2,611 +/- 3,023 | 3,640 +/- 2,857 | 5,988 +/- 3,703 |
| Reads >5 kbp | 107,686 | 38,654 | 6,902 |
| Bases, reads >5 kbp | 924,553,134 | 305,883,256 | 59,804,344 |
| Avg Read Length, reads >5 kbp | 8,585 +/- 3,247 | 7,913 +/- 2,577 | 8,664 +/- 2,119 |

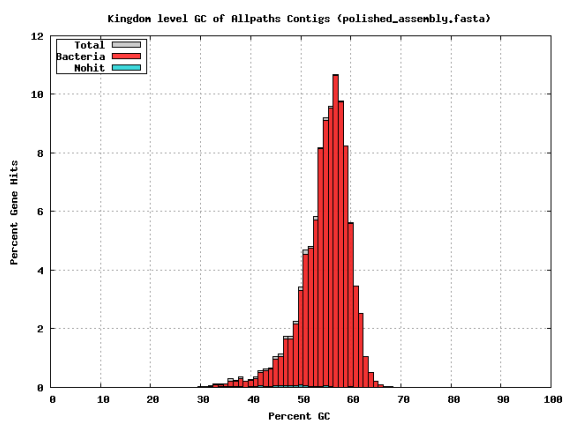
3. Assembly Statistics

| | |
|-----------------------------------|--------------------|
| Scaffold total | 3 |
| Contig total | 3 |
| Scaffold sequence length | 3 |
| Contig sequence length | 4.8 MB (0.0% gap) |
| Scaffold N/L50 | 1/4.2 |
| Largest Contig | 4,167.1 |
| Number of scaffolds >50 kb | 3 |
| Pct of genome in scaffolds >50 kb | 100.0 |

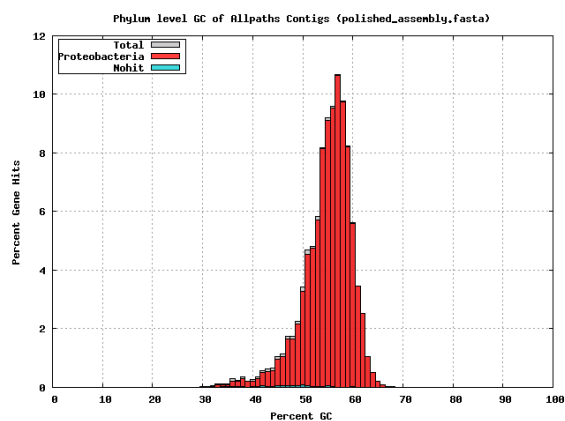
4. Assembly QC Results

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

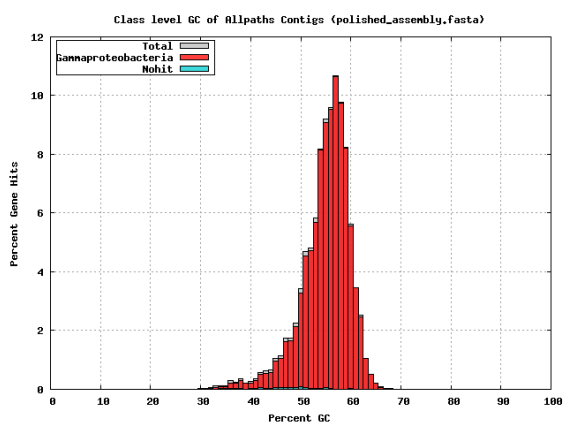
Kingdom



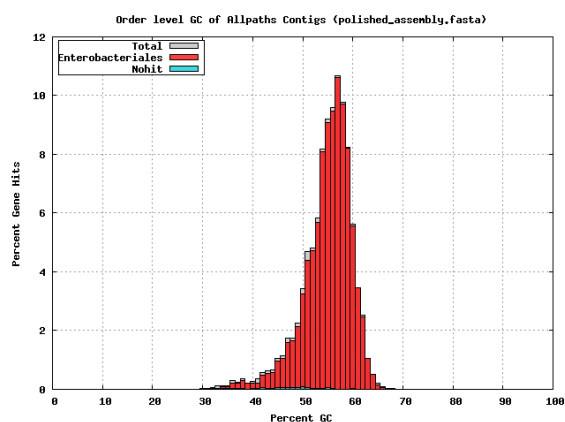
Phylum



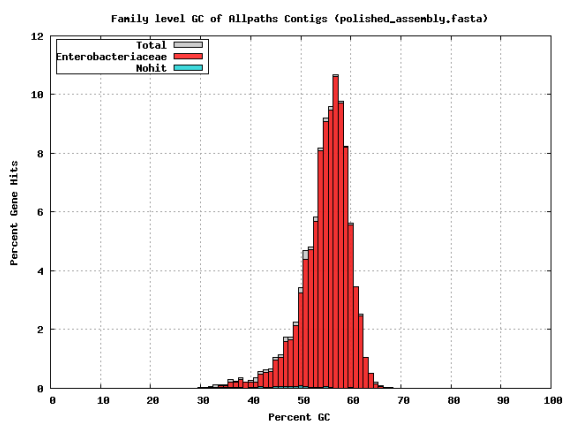
Class



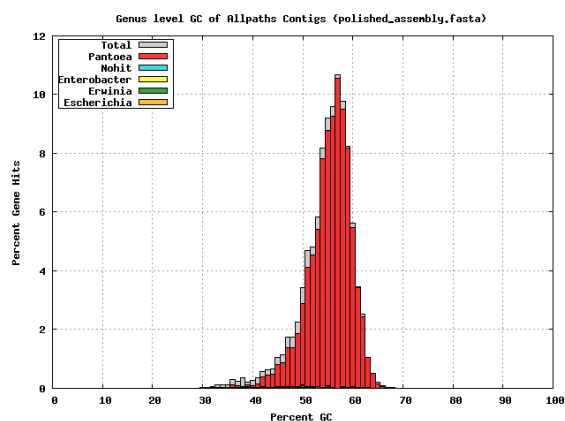
Order



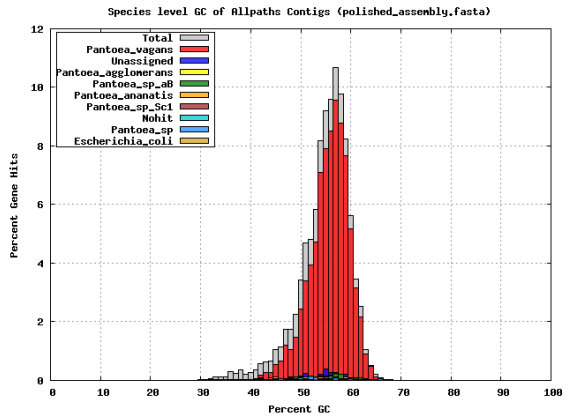
Family



Genus

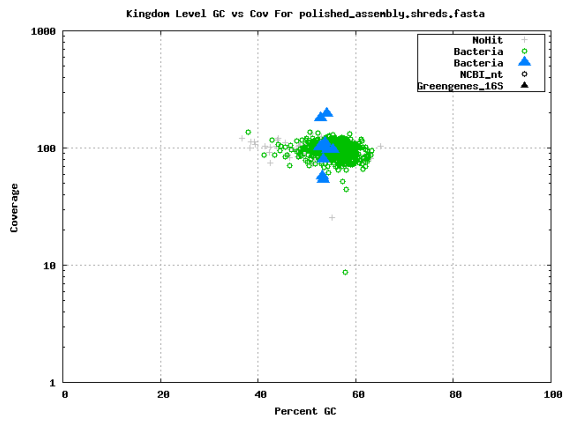


Species

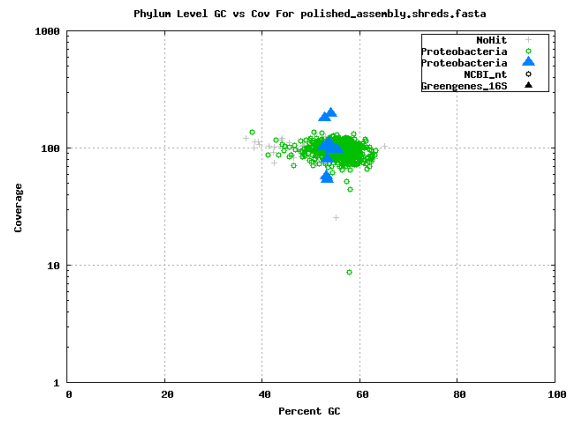


GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

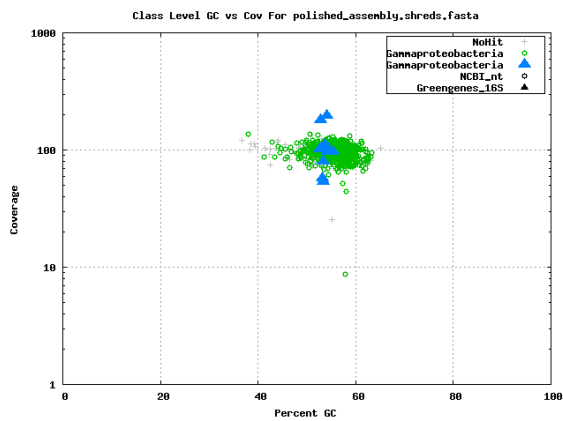
Kingdom



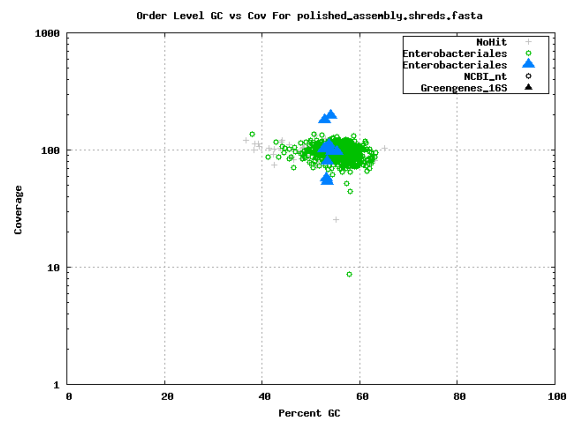
Phylum

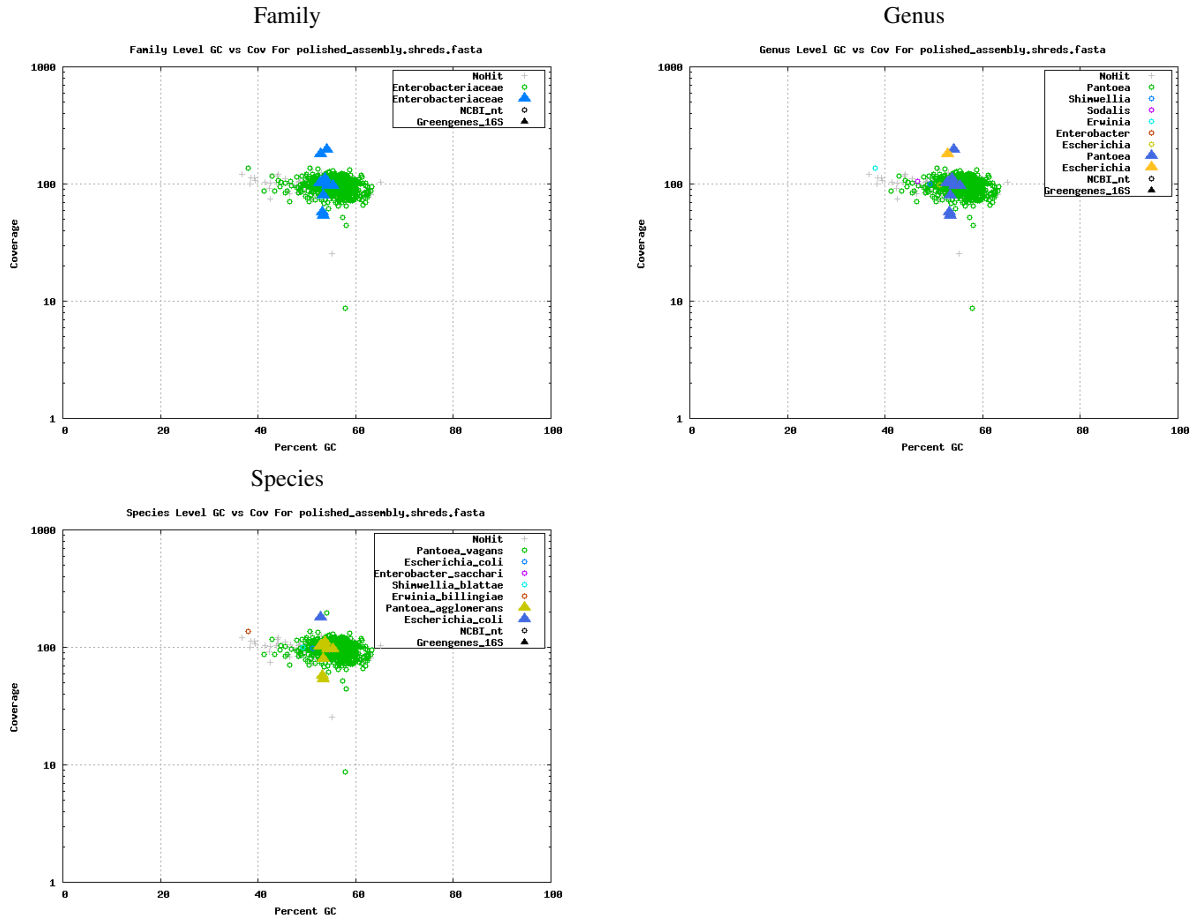


Class

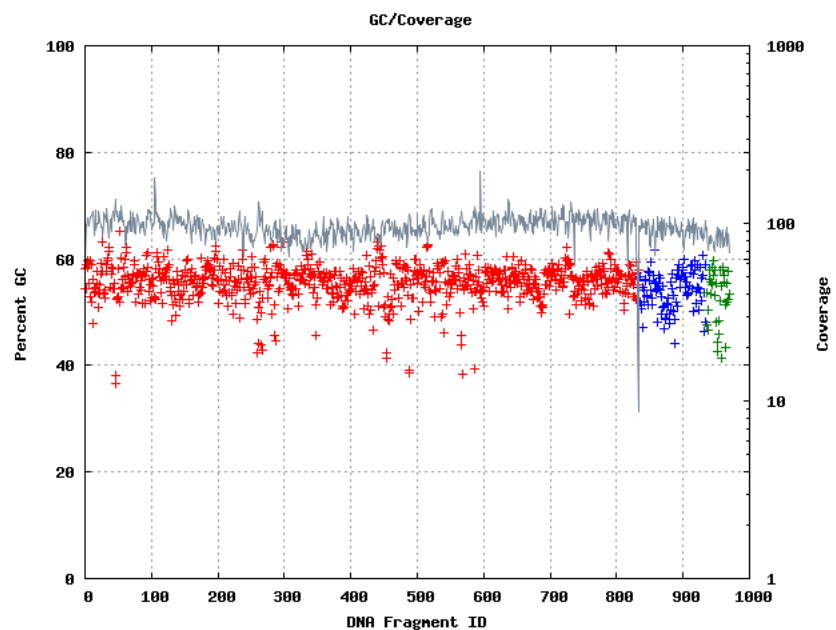


Order

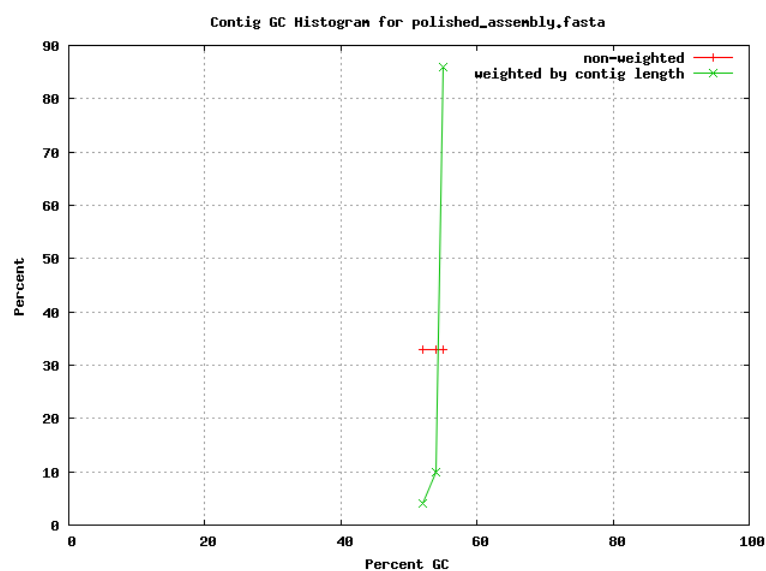




Coverage vs GC. Contigs were shredded into non-overlapping 5kbp and the GC of each shred was plotted as a point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.



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



List of contigs and average percent GC bin:

| Pct GC Bin | Contig Name |
|------------|---|
| 50 | scf7180000000007 quiver_dupTrim=8177, scf7180000000008 quiver_dupTrim=6745 |
| 55 | scf7180000000006 quiver |

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

| Organism | Align Length (bp) | Pct Id | Contig Name |
|---|-------------------|--------|------------------------|
| 267618_Pantoea_agglomerans_EU598802.1_1..1532 | 1,533.0 | 99.0 | scf718000000006 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 Minimal Draft

Genome sequencing and assembly

The draft genome of *Pantoea sp. 9140* was generated at the DOE Joint genome Institute (JGI) using the Pacific Biosciences (PacBio) sequencing technology [1]. A Pacbio SMRTbell™ library was constructed and sequenced on the PacBio RS platform, which generated 165,984 filtered subreads totaling 604.3 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: 2.2.0.p1) [2]. The final draft assembly contained 3 contigs in 3 scaffolds, totalling 4.8 Mbp in size. The input read coverage was 90.3X.

Genome annotation

Genes were identified using Prodigal [3], followed by a round of manual curation using GenePRIMP [4] for finished genomes and Draft 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.