03/18/2017

1. Read Statistics

	Raw Reads	Filtered SubReads	Error Corrected Reads
Reads	345,350	208,399	12,453
Bases	1,877,993,310	825,666,766	111,997,439
Avg Read Length	5,437.9 +/- 5,349.8	3,962.0 +/- 3,128.3	8,993.6 +/- 5,173.0
Reads >5 kbp	126,288	50,315	9,041
Bases, reads >5 kbp	1,369,436,074	425,495,810	105,168,062
Avg Read Length, reads >5 kbp	10,843.8 +/- 5,390.6	8,456.6 +/- 3,207.0	11,632.3 +/- 3,282.7

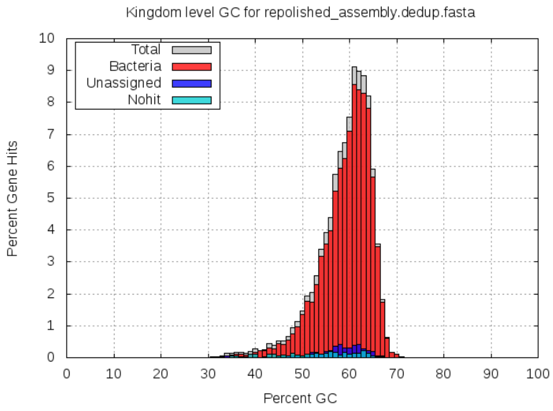
2. Assembly Statistics

Scaffold total	5
Contig total	5
Scaffold sequence length	10.862 mb
Contig sequence length	10.862 mb 0.000% gap
Scaffold N/L50	1/5.915 mb
Largest Contig	5,914.9 kbp
Number of scaffolds >50 kb	4
Pct of genome in scaffolds >50 kb	99.95%

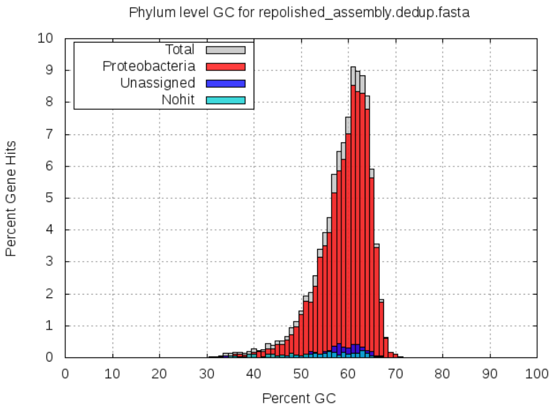
3. Assembly QC Results

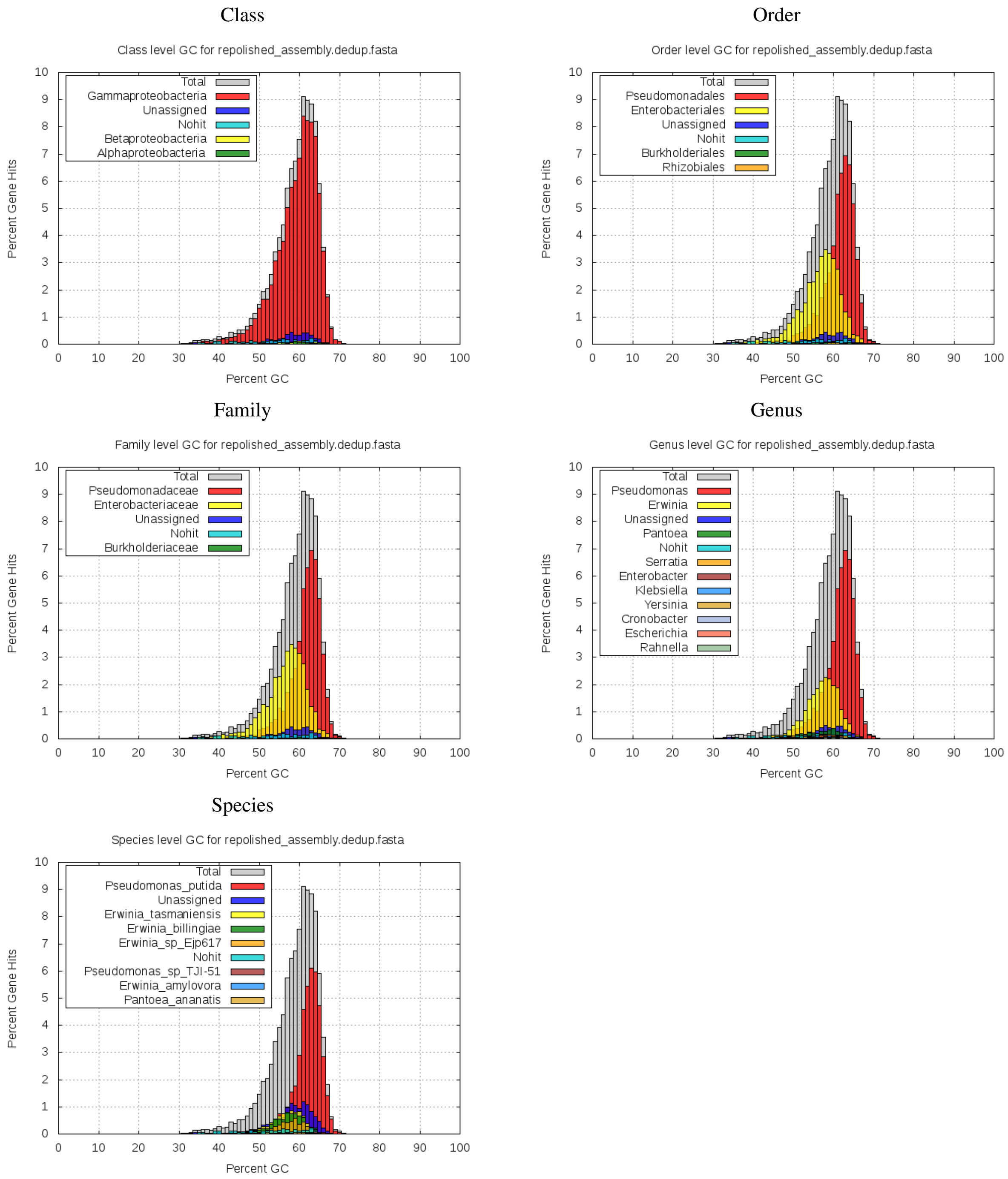
GC histogram of the predicted genes on each contig, overlaid with GC of hits based on LAST, shown for different taxonomic levels.

Kingdom



Phylum





Class

Order

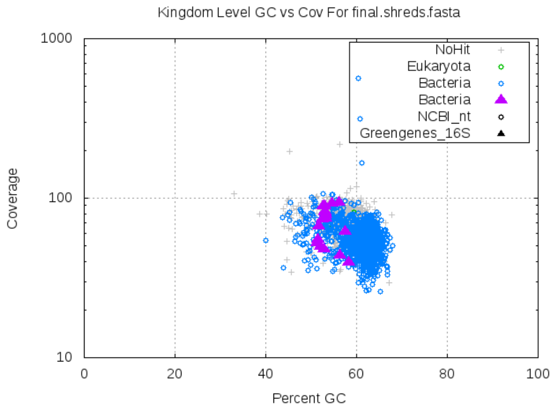
Family

Genus

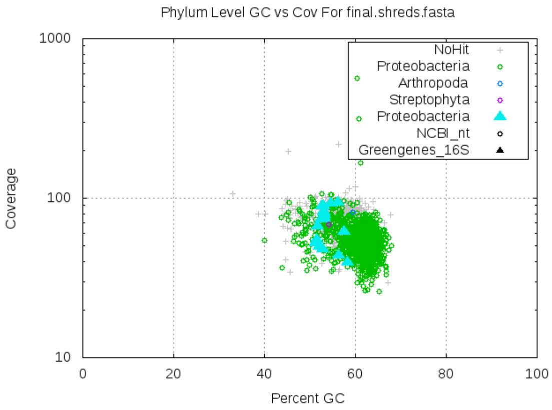
Species

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.

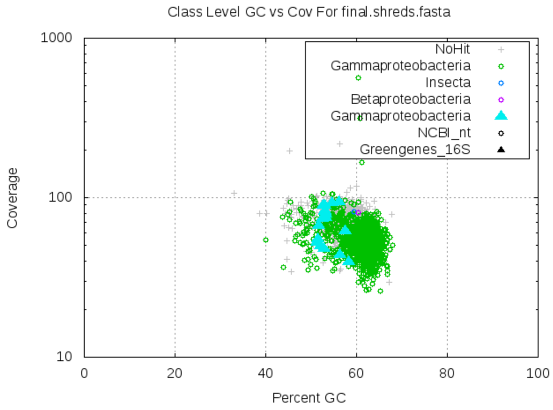
Kingdom



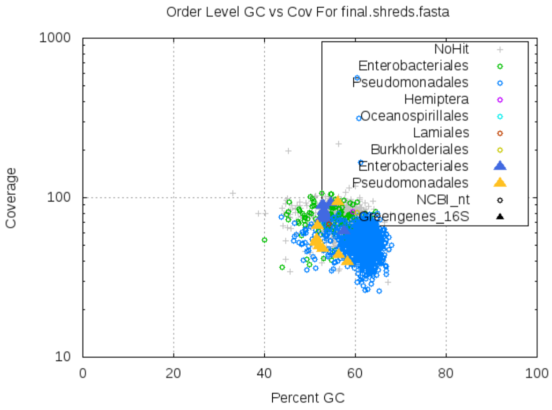
Phylum



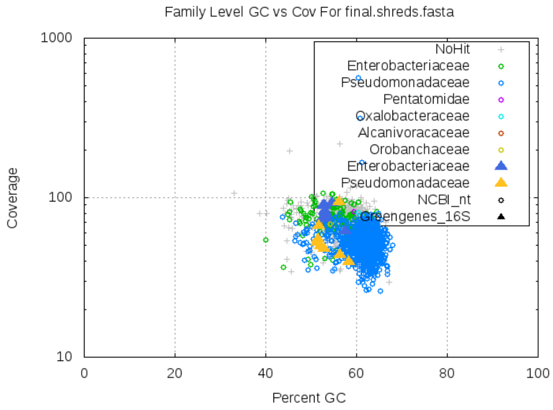
Class



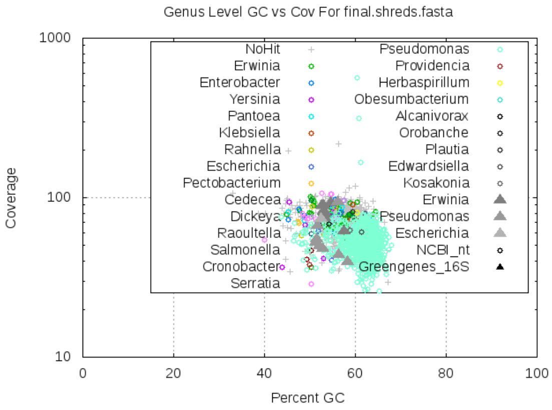
Order



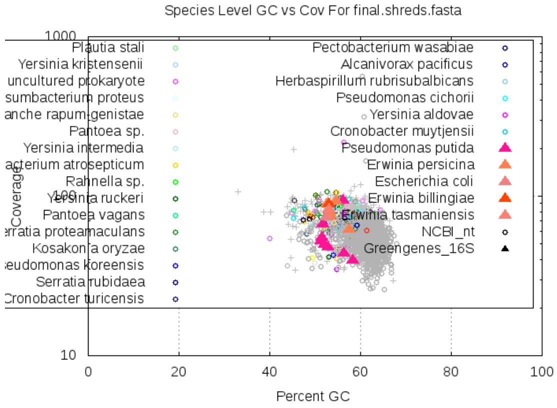
Family



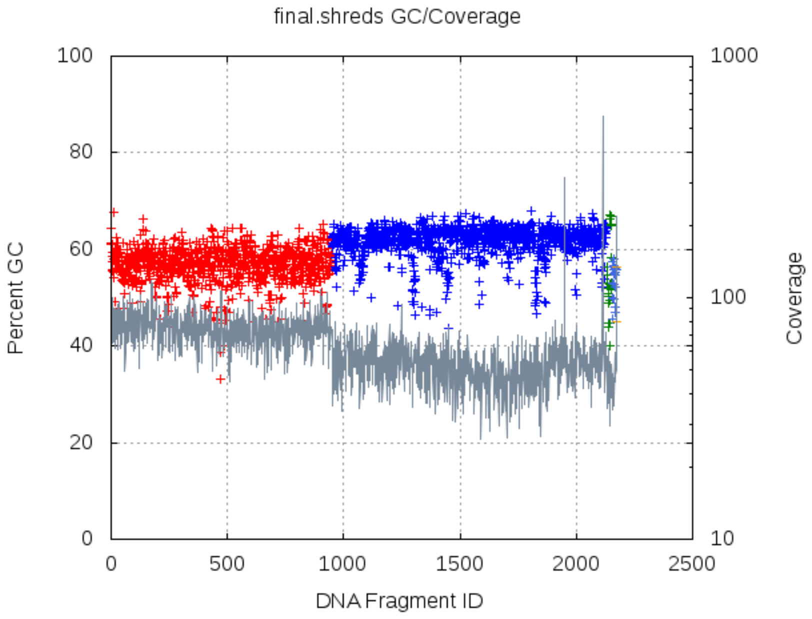
Genus



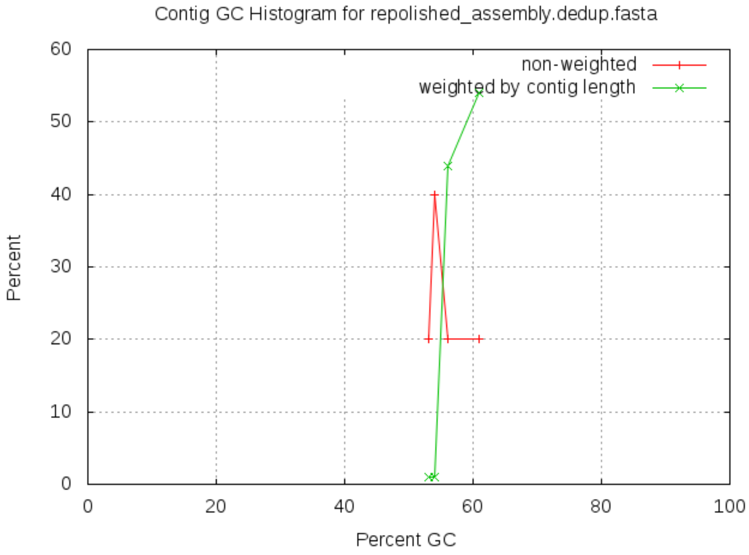
Species



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 bin:

Pct GC Bin	Contig Name
50	unitig_2|quiver, unitig_3|quiver, unitig_6|quiver
55	unitig_0|quiver
60	unitig_1|quiver

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

Organism	Align Length (bp)	Pct Id	Contig Name
Escherichia coli str. K–12 substr. DH10B, complet e	13,799	91.08	unitig_0|quiver
Pseudomonas putida KT2440 chromosome, complete	1,103,222	99.98	unitig_1|quiver

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

Organism	Align Length (bp)	Pct Id	Contig Name
263564 Pseudomonas putida str. GB–1 NC_010322.1	1,538	100.00	unitig_1|quiver
9822 Erwinia persicina str. LMG 2691 AJ001190.1	1,534	99.41	unitig_0|quiver

4. Methods

Isolate Improved Draft

Genome sequencing and assembly

The draft genome of was generated at the DOE Joint Genome Institute (JGI) using the Pacific Biosciences (PacBio) sequencing technology [1]. A >10kpb Pacbio SMRTbell™ library was constructed and sequenced on the PacBio RS platform, which generated 208,399 filtered subreads totaling 825.7 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_p5,protocol version=2.3.0 method=RS_HGAP_Assembly.3,smrtpipe.py v1.87.139483,) [2]. The final draft assembly contained 5 contigs in 5 scaffolds, totaling 10.862 Mbp in size. The input read coverage was 104.3X.

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.