1. Project Information

Program	Microbial
JGI Sequencing Project ID	1022151
Sequencing Project Name	Massilia sp. 9096

2. Read Statistics

	Raw Reads	Filtered SubReads	Error Corrected Reads
Reads	760,086	243,206	17,014
Bases	2,153,768,640	780,891,696	72,621,304
Avg Read Length	2,833 +/ 2,902	3,210 +/ 2,257	4,268 +/ 2,631
Reads >5 kbp	130,547	40,913	6,233
Bases, reads >5 kbp	1,048,132,185	292,638,751	44,609,684
Avg Read Length, reads >5 kbp	8,028 +/ 2,992	7,152 +/ 2,167	7,157 +/ 1,551

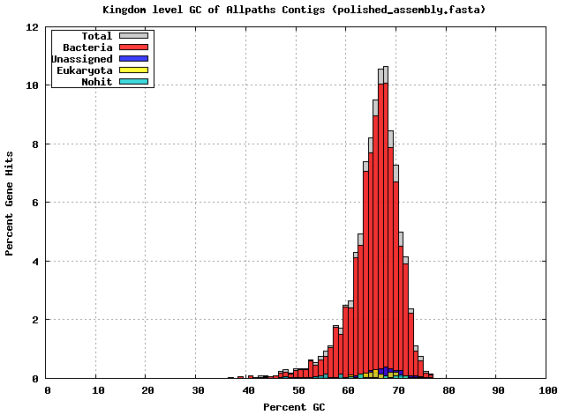
3. Assembly Statistics

Scaffold total	2
Contig total	2
Scaffold sequence length	2
Contig sequence length	5.7 MB (0.0% gap)
Scaffold N/L50	1/5.6
Largest Contig	5,554.4
Number of scaffolds >50 kb	2
Pct of genome in scaffolds >50 kb	100.0

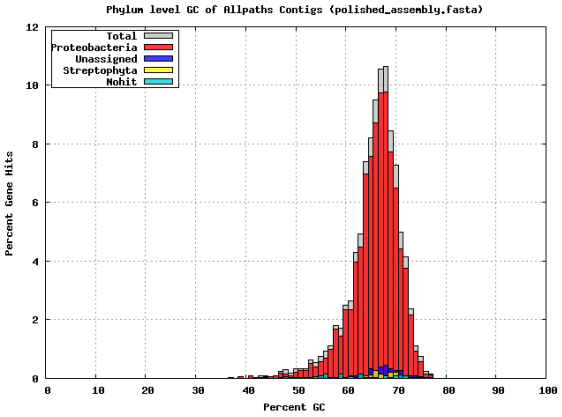
4. Assembly QC Results

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

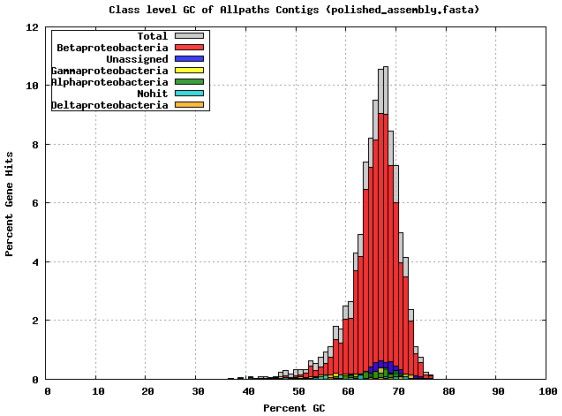
Kingdom



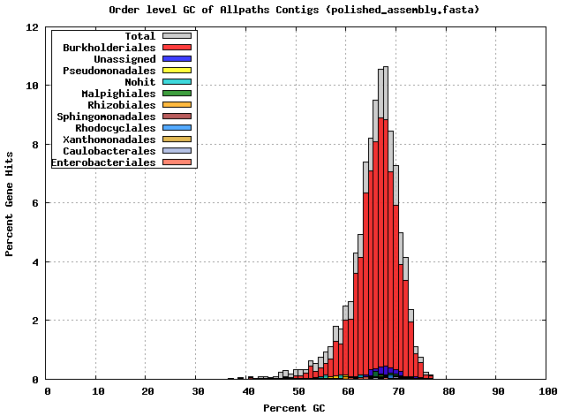
Phylum



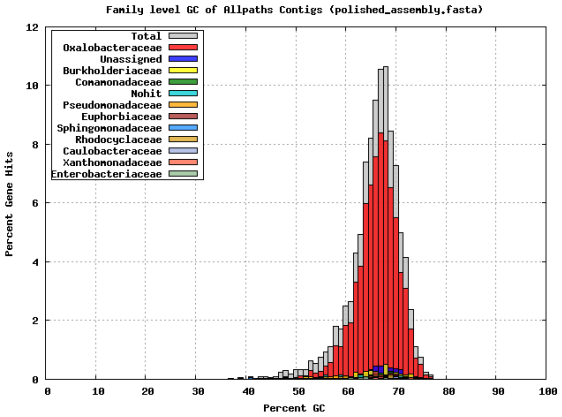
Class



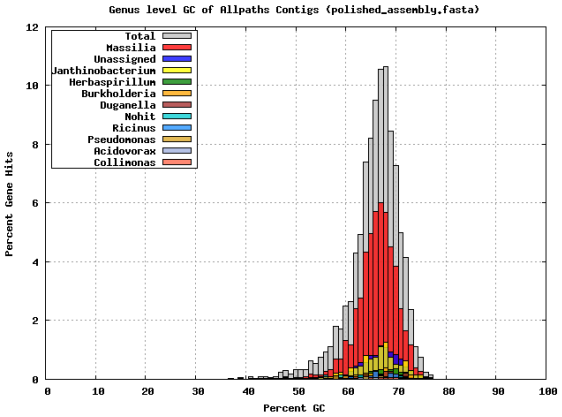
Order



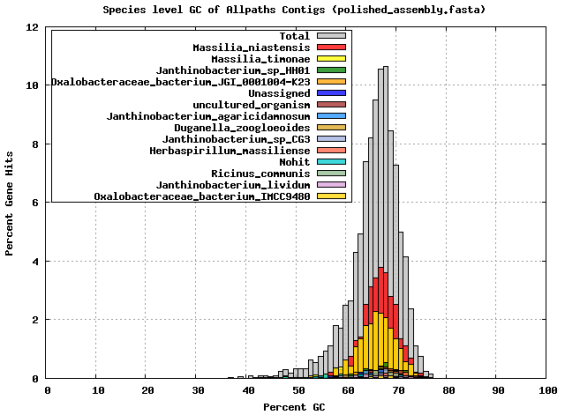
Family



Genus

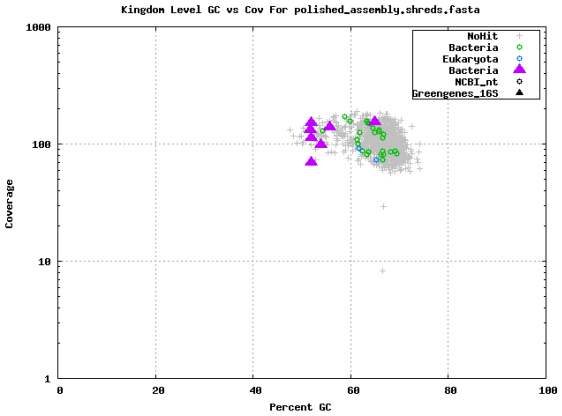


Species

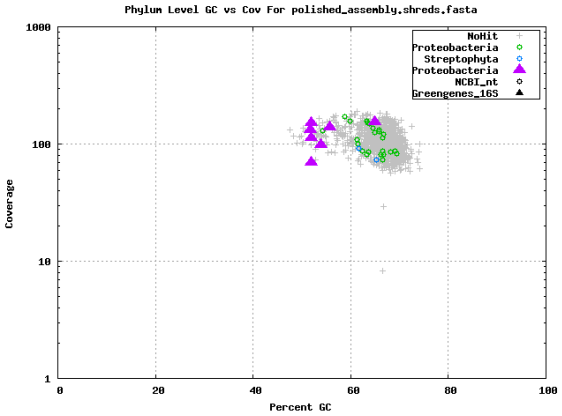


GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

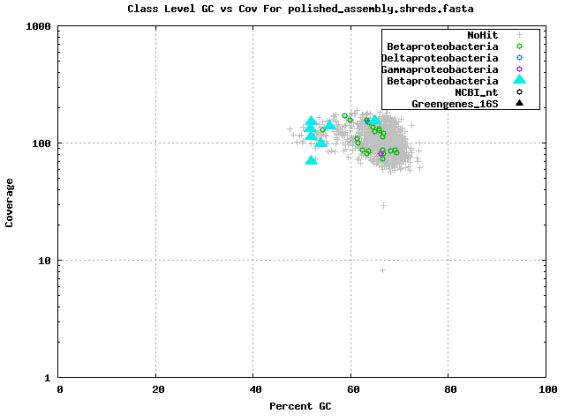
Kingdom



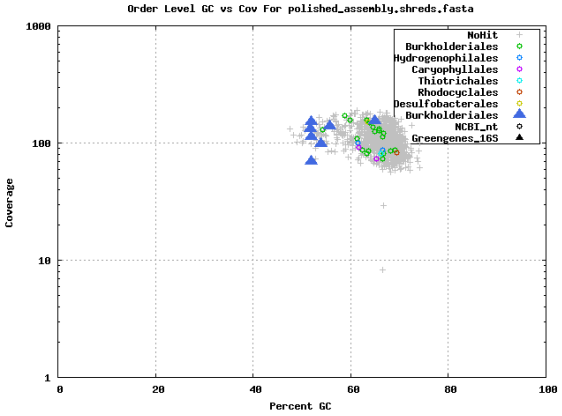
Phylum



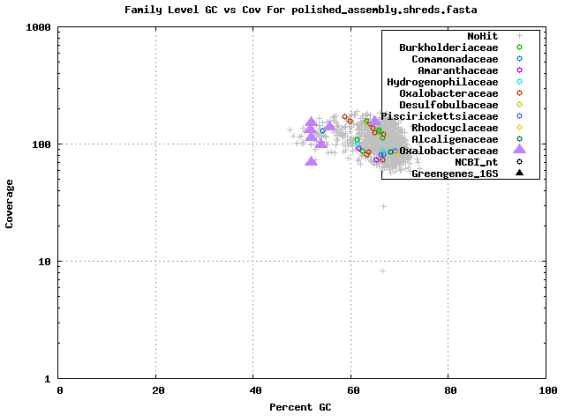
Class



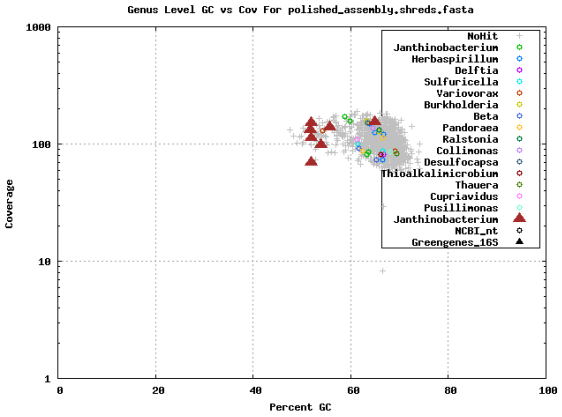
Order



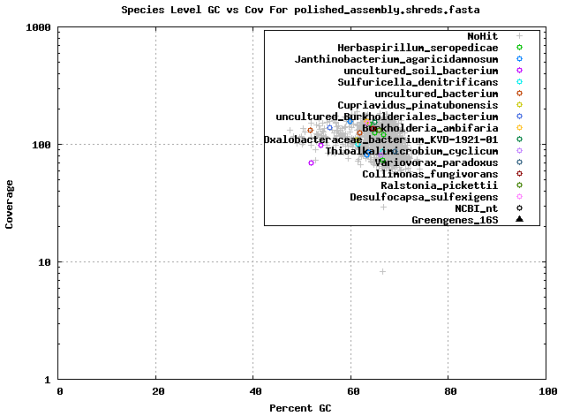
Family



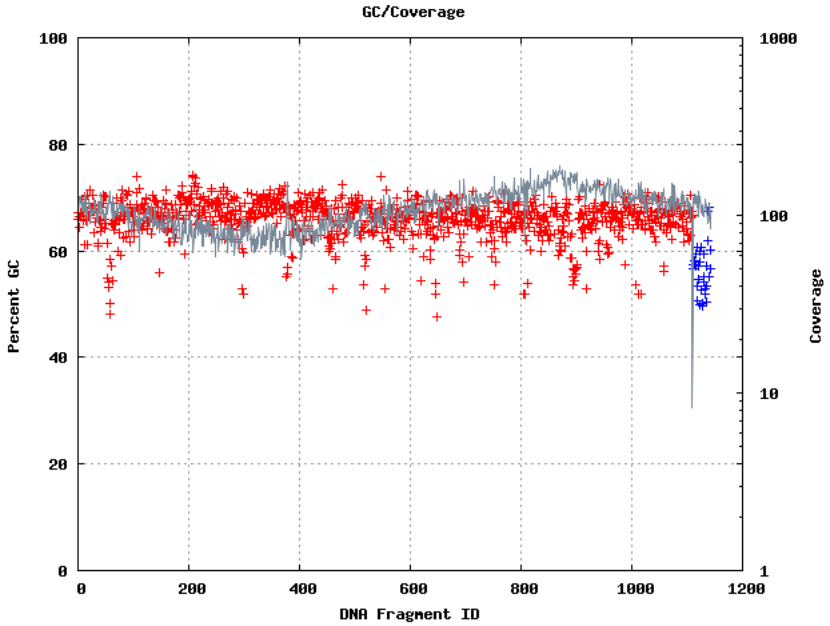
Genus



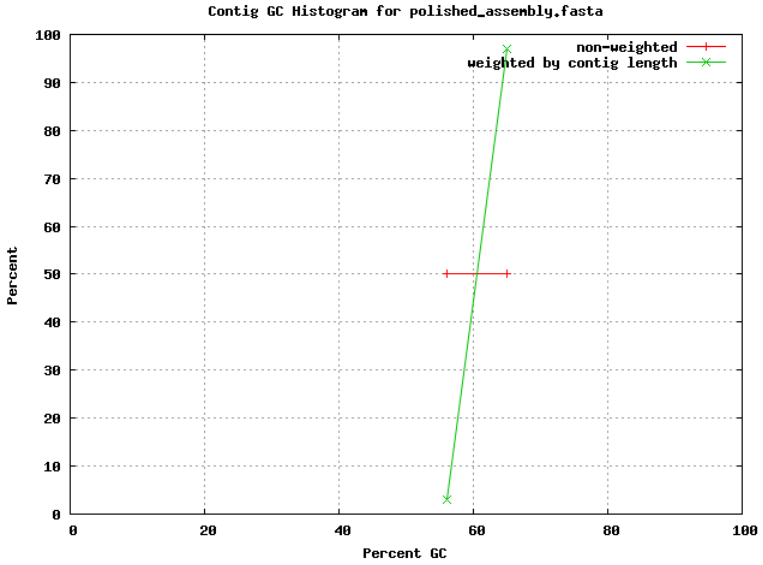
Species



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
55	scf7180000000013|quiver_dupTrim=5373
65	scf7180000000010|quiver

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

Organism	Align Length (bp)	Pct Id	Contig Name
140398_Janthinobacterium_sp._str._TibetIIK43_DQ177 470.1_2..1426	1,430.0	95.0	scf7180000000013|quiver_dupTrim=5373
162184_Janthinobacterium_sp._str._A113_AB252072.1_ 1..1522	1,525.0	98.0	scf7180000000010|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 *Massilia sp. 9096* 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 243,206 filtered subreads totaling 780.9 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 2 contigs in 2 scaffolds, totalling 5.7 Mbp in size. The input read coverage was 120.9X.

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.