1. Project Information

Program	Microbial
PMO Project	
JGI Project ID	1031158
Sequencing Project Name	uncultured virus JFR_U1362B AD–236_F14

2. Read Statistics

Illumina Std PE Statistics

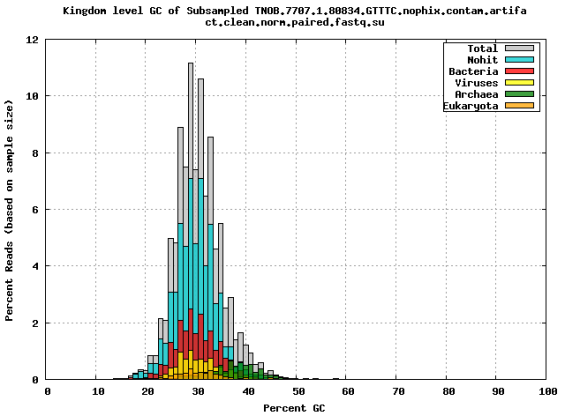
File name	TNOB.7707.1.80834.GTTTC.nophix.contam.artifact.clean.norm.paired.fastq
Library	TNOB
Number of reads	20,426
Sequencing depth †	1X
Read type	2x251 bp

† A genome size of 5.0 Mbp was assumed in this calculation.

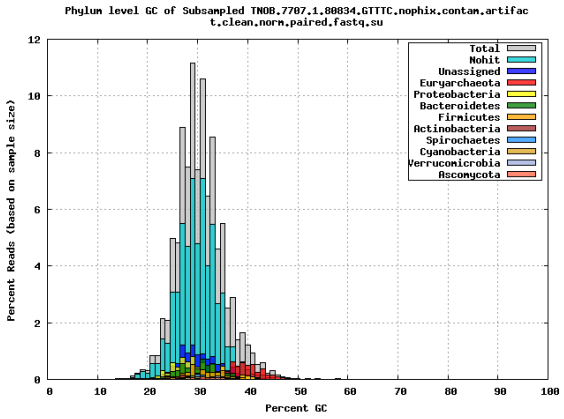
3. Read QC Results

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

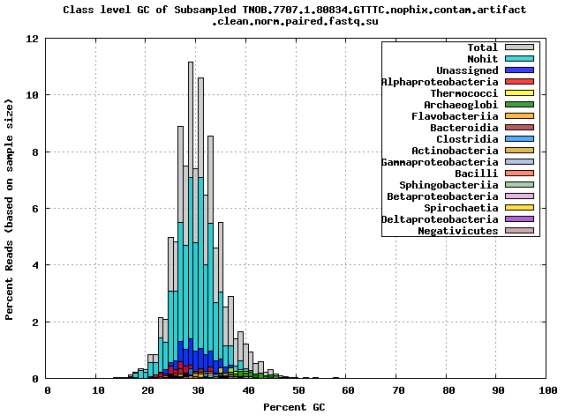
Kingdom



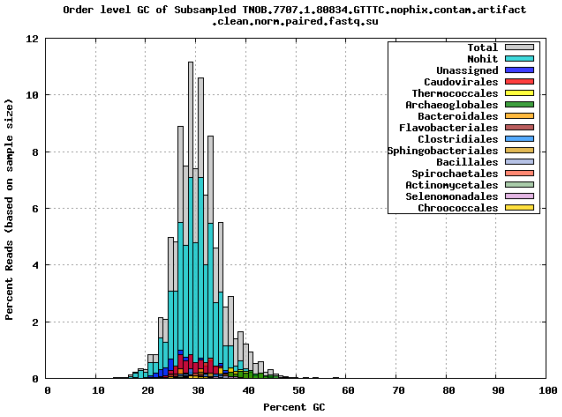
Phylum



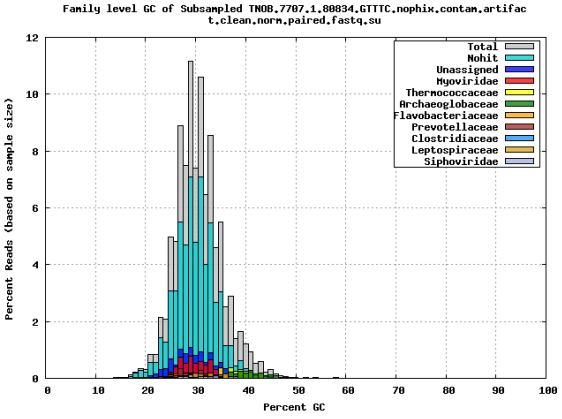
Class



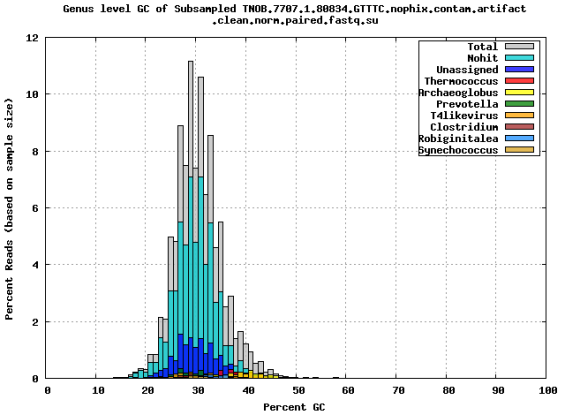
Order



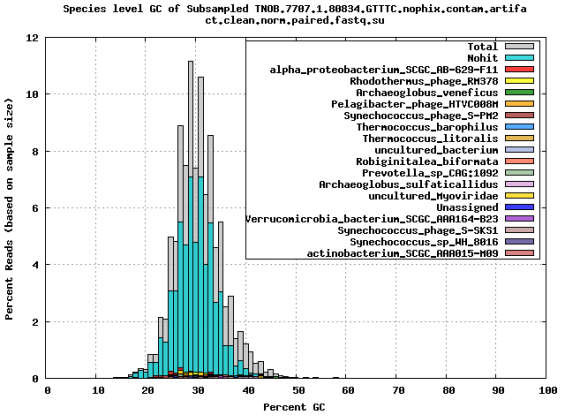
Family



Genus



Species



4. Assembly Statistics

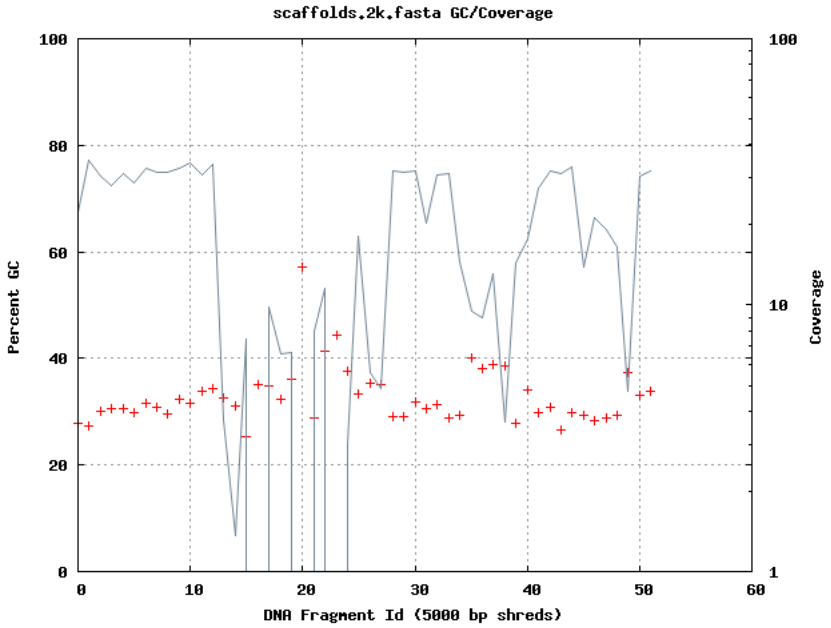
Assembly method	SPAdes
Scaffold total	17
Contig total	25
Scaffold sequence length	180.9 kb
Contig sequence length	180.9 kb (0.0% gap)
Scaffold N/L50	3/18.5 kb
Contig N/L50	3/18.5 kb
Largest Contig	58.7 kb
Number of scaffolds >50 kb	1
Pct of genome in scaffolds >50 kb	32.4

5. Assembly QC Results

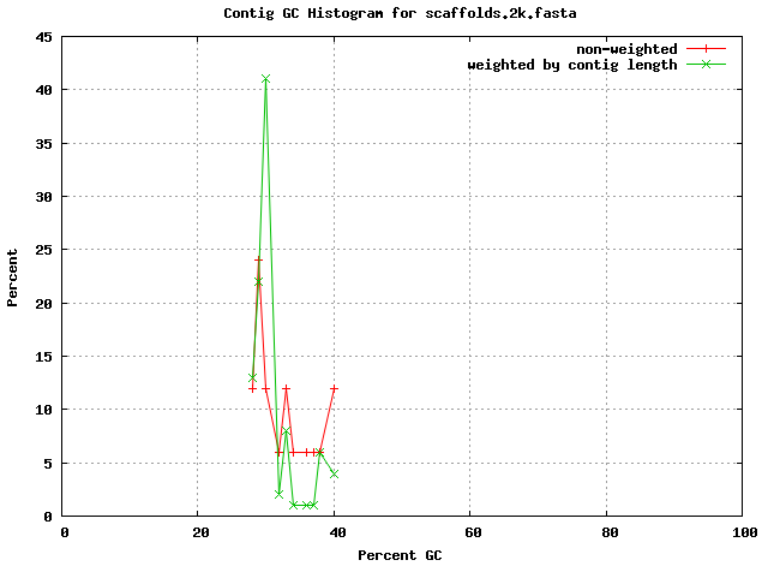
GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

No hits found.

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.



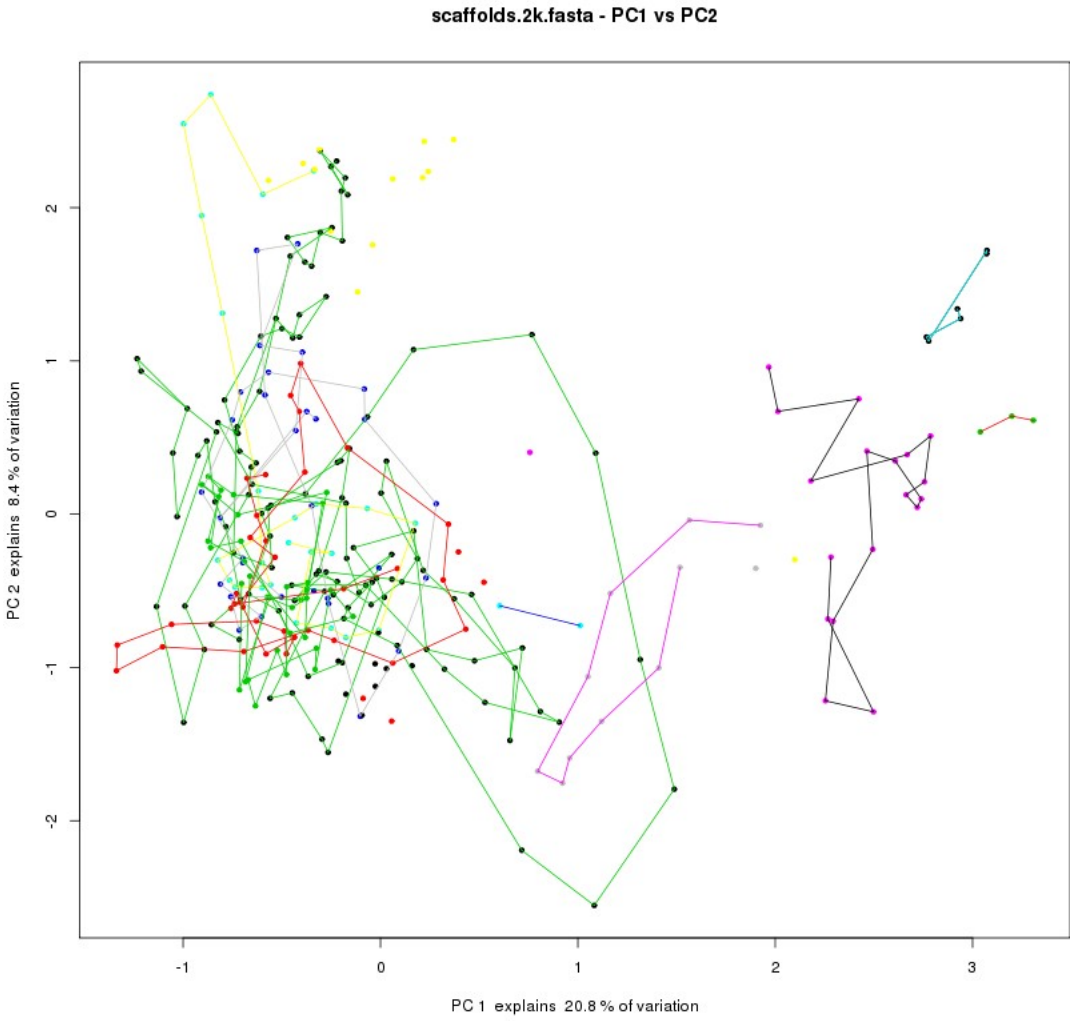
GC histogram of the contigs, including contig length weighted distribution.



List of contigs and average percent GC, grouped in bins of 5:

Pct GC Bin	Contig Name
25	NODE_2_length_18697_cov_25.7939_ID_61000, NODE_3_length_18535_cov_25.1926_ID_60632, NODE_5_length_13969_cov_26.7681_ID_60992, NODE_9_length_5741_cov_12.4063_ID_61132, NODE_11_length_3525_cov_6.64179_ID_60924, NODE_13_length_3172_cov_7.10748_ID_60720
30	NODE_1_length_58659_cov_25.4981_ID_61138, NODE_4_length_15591_cov_22.2047_ID_60904, NODE_7_length_7773_cov_26.5924_ID_61054, NODE_8_length_6916_cov_11.0686_ID_61004, NODE_14_length_2760_cov_5.08466_ID_60526, NODE_15_length_2363_cov_27.831_ID_60994
35	NODE_6_length_10754_cov_8.08702_ID_60674, NODE_16_length_2302_cov_3.8251_ID_60048 NODE_17_length_2155_cov_5.5281_ID_60866
40	NODE_10_length_4521_cov_8.34684_ID_61126, NODE_12_length_3460_cov_7.3674_ID_61060

Principal component analysis of tetramer frequencies of contigs. Detectable variations are highlighted in color.



Estimated genome recovery derived from analysis of universal single-copy genes detected in final assembly.

HMM	Pct Recovered
bacteria	0 %
archaea	0.69 %

6. Sequence Data Availability

Files can be downloaded from our JGI portal website.
http://portal.nersc.gov/microbial/assembly/GAA-691

Filename	Description
contigs.2k.fasta	SPAdes

7. 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 20,426 reads totaling 5.1 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 2.4.0), (3) Parameters for assembly steps were –t 8 –m 120 —sc —careful —12. The final draft assembly contained 25 contigs in 17 scaffolds, totalling 180.9 Kb in size. The final assembly was based on of Illumina data. Based on a presumed genome size of 5.0 Mb, the average input read coverage used for the assembly was X.

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, Lacascio 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.