

1. Project Information

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

2. Read Statistics

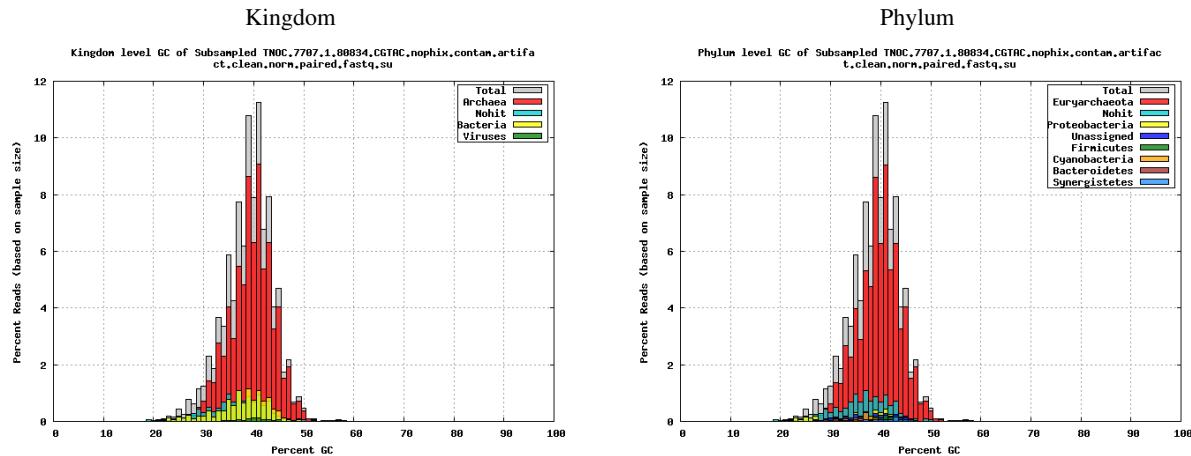
Illumina Std PE Statistics

File name	TNOC.7707.1.80834.CGTAC.nophix.contam.artifact.clean.norm.paired.fastq
Library	TNOC
Number of reads	14,976
Sequencing depth [†]	1X
Read type	2x251 bp

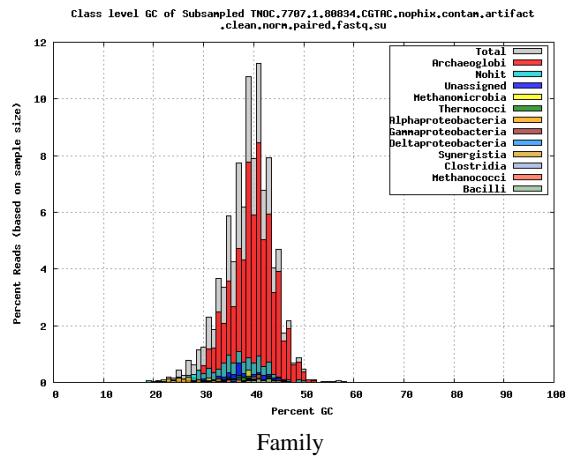
[†] A genome size of 5.0 Mbp was assumed in this calculation.

3. Read QC Results

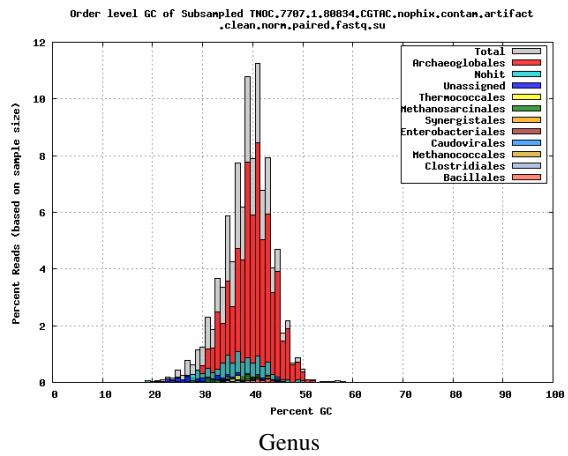
GC histogram of the reads subsampled to 10k, overlaid with GC of hits based on BLASTX, shown for different taxonomic levels.



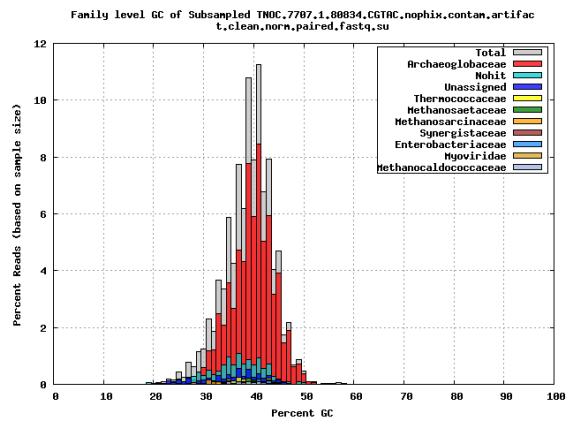
Class



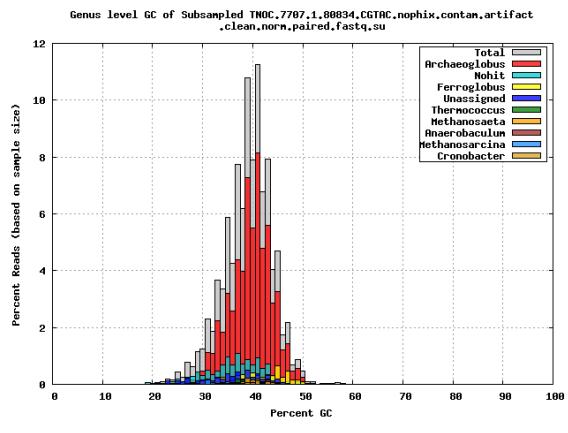
Order



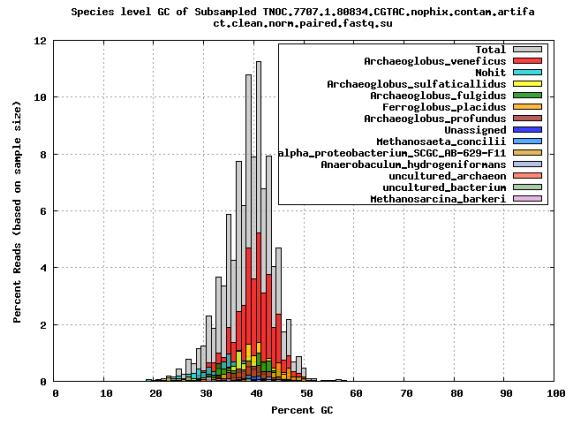
Family



Genus



Species

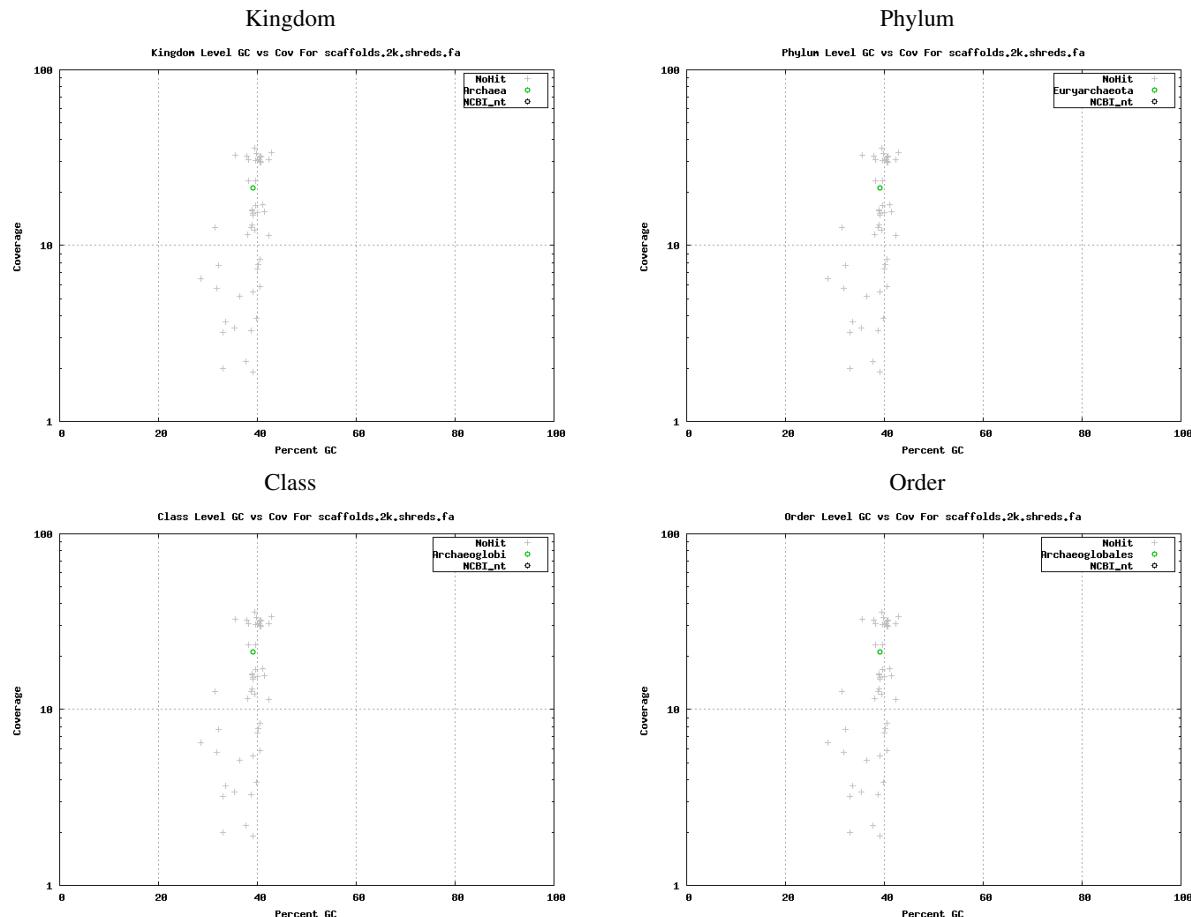


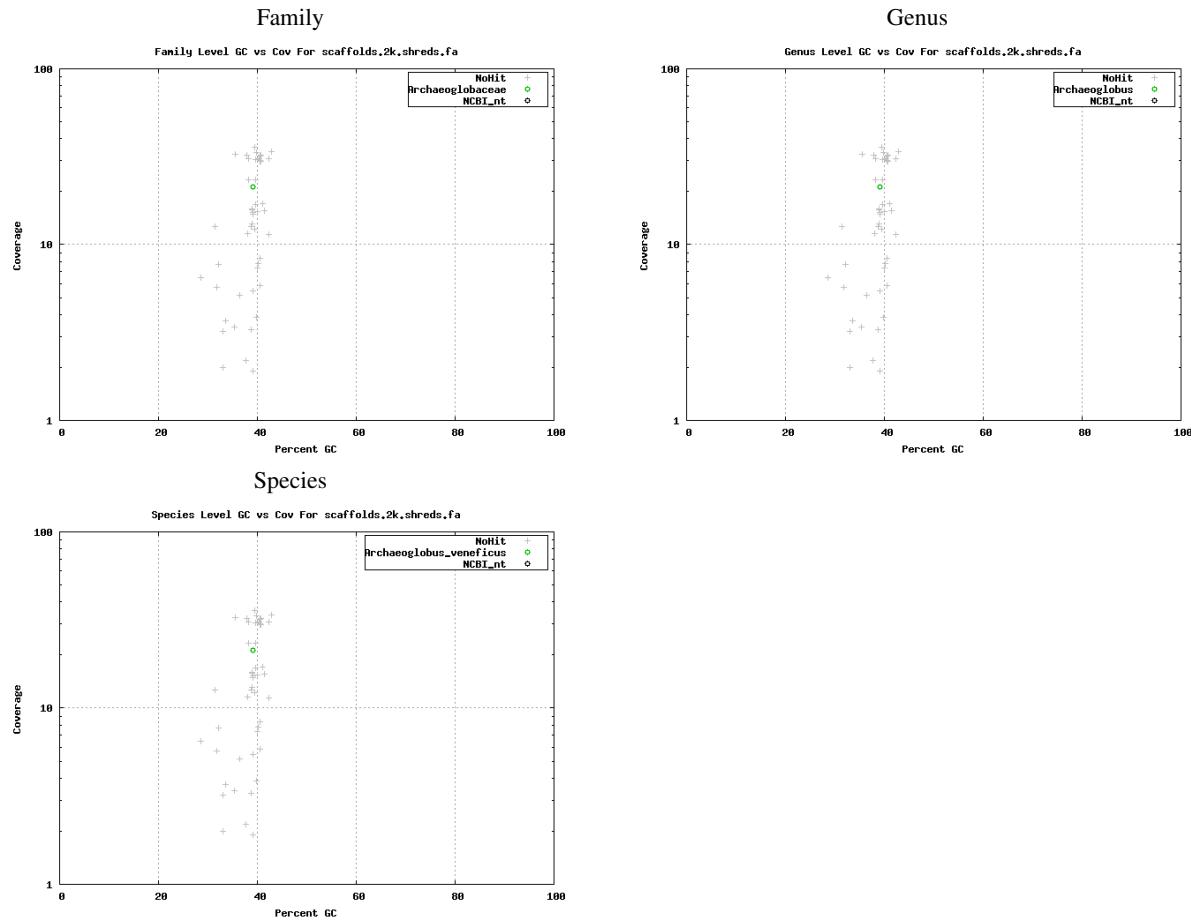
4. Assembly Statistics

Assembly method	SPAdes
Scaffold total	22
Contig total	35
Scaffold sequence length	152.4 kb
Contig sequence length	152.4 kb (0.0% gap)
Scaffold N/L50	5/9.1 kb
Contig N/L50	5/9.1 kb
Largest Contig	39.6 kb
Number of scaffolds >50 kb	0
Pct of genome in scaffolds >50 kb	0.0

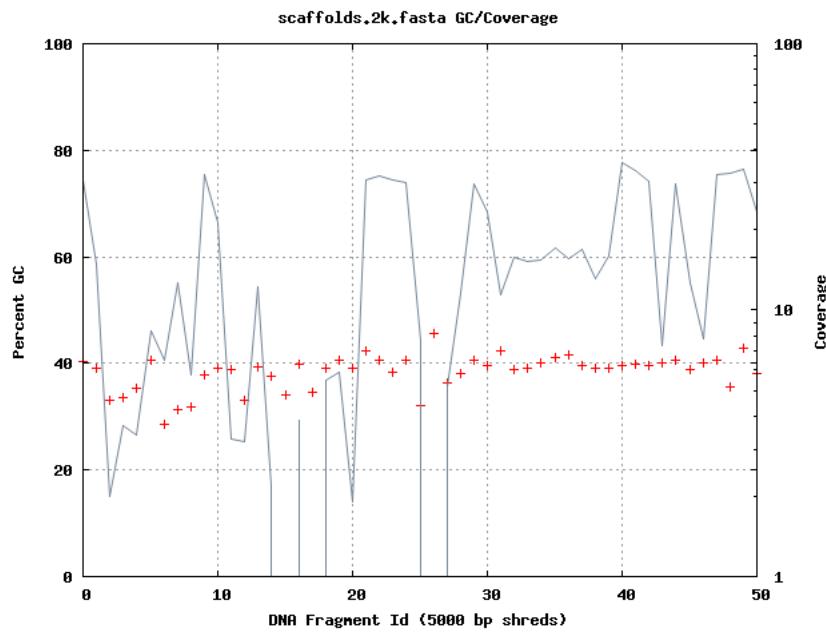
5. Assembly QC Results

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.

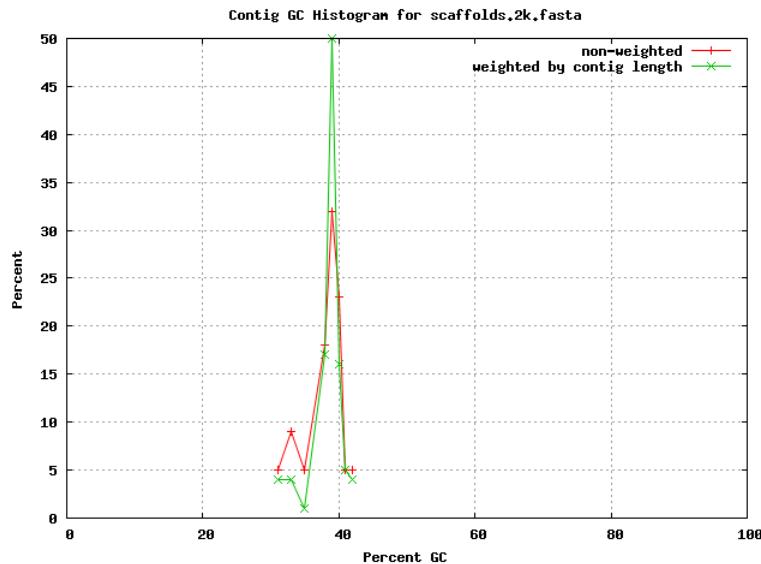




Coverage vs GC. Contigs were shredded into non-overlapping 5kbp and the GC of each shred was plotted as a point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.



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

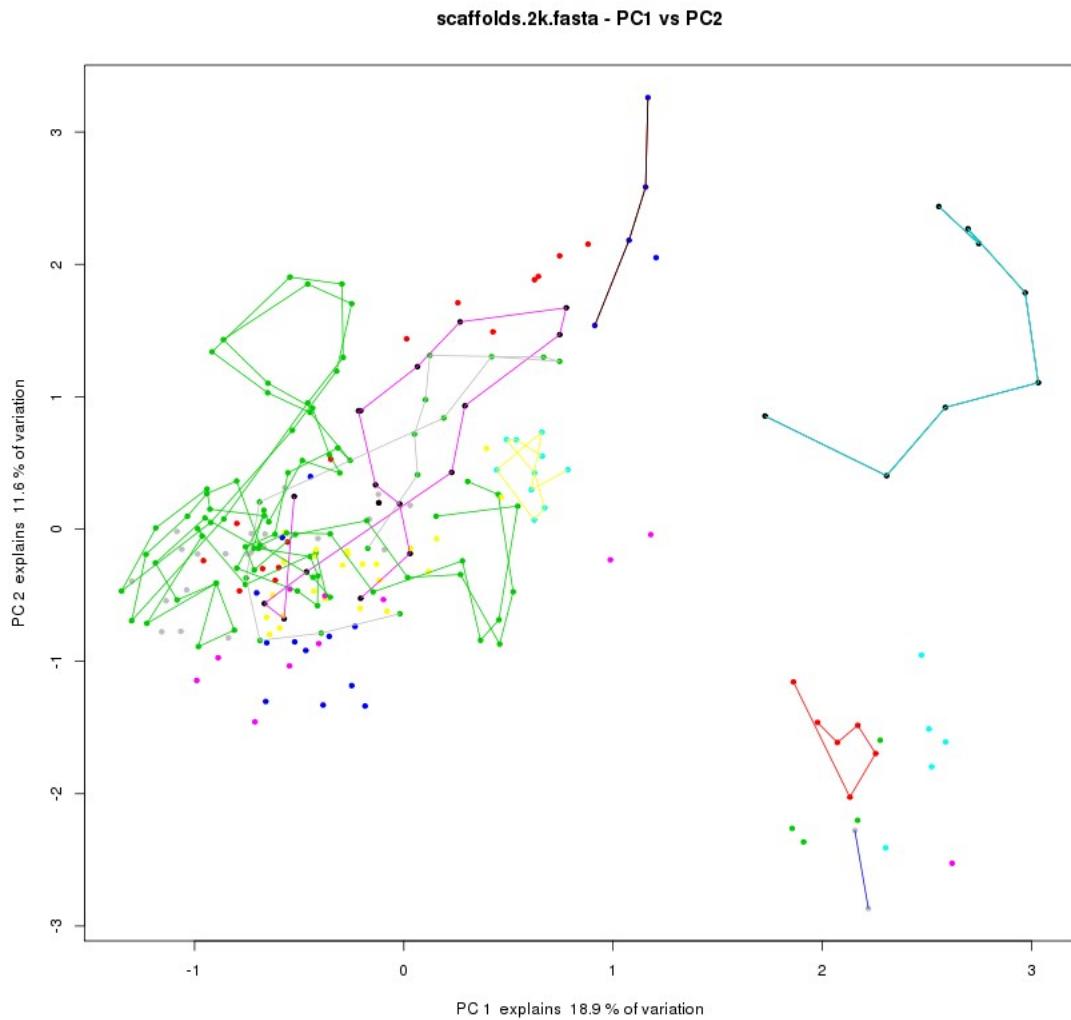


List of contigs and average percent GC, grouped in bins of 5:

Pct GC Bin	Contig Name
30	NODE_10.length_5556_cov_8.40647.ID_48332, NODE_13.length_3728_cov_5.68364.ID_48878 NODE_20.length_2115_cov_3.66311.ID_48876
35	NODE_1.length_39619_cov_26.0382.ID_49376, NODE_2.length_12963_cov_27.0814.ID_49524, NODE_4.length_9624_cov_21.6109.ID_49480, NODE_5.length_9132_cov_26.6519.ID_49526, NODE_6.length_8608_cov_24.9645.ID_49534, NODE_8.length_6754_cov_3.97179.ID_48038, NODE_11.length_4712_cov_9.14129.ID_48986, NODE_14.length_3541_cov_9.52897.ID_47936, NODE_18.length_2132_cov_11.1863.ID_49460, NODE_19.length_2132_cov_12.7289.ID_49462,

	NODE_21.length_2082 cov_3.98125 ID_48492, NODE_22.length_2078 cov_5.68463 ID_48972
40	NODE_3.length_12169 cov_19.158 ID_49136, NODE_7.length_7687 cov_26.1343 ID_49372, NODE_9.length_5738 cov_27.1848 ID_49536, NODE_12.length_3840 cov_6.34214 ID_48604, NODE_15.length_2934 cov_23.0837 ID_49538, NODE_16.length_2686 cov_24.6526 ID_49366 NODE_17.length_2570 cov_6.74712 ID_49528

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	0.8 %
archaea	4.12 %

6. Sequence Data Availability

Files can be downloaded from our JGI portal website.
<http://portal.nerc.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 14,976 reads totaling 3.8 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 35 contigs in 22 scaffolds, totalling 152.4 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, Lascasio 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, Kyprides 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, Kyprides NC. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 2009; 25:2271–2278.