uncultured virus JFR_U1362B AD–236_F14

Single–cell Assembly QC Report 02/14/2014

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

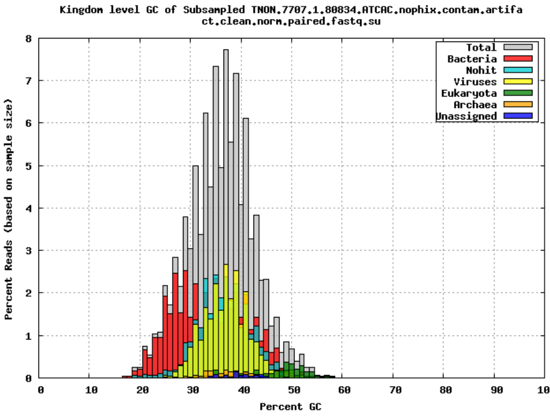
File name	TNON.7707.1.80834.ATCAC.nophix.contam.artifact.clean.norm.paired.fastq
Library	TNON
Number of reads	39,870
Sequencing depth †	2X
Read type	2x251 bp

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

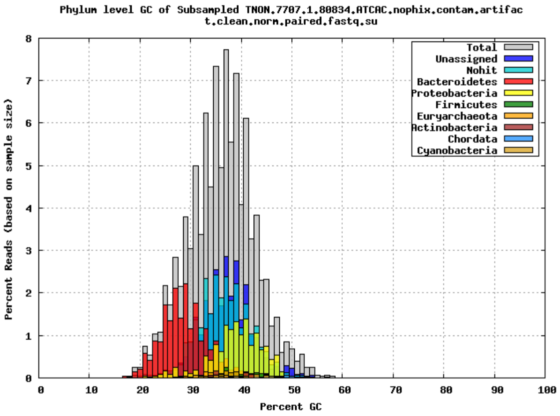
3. Read QC Results

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

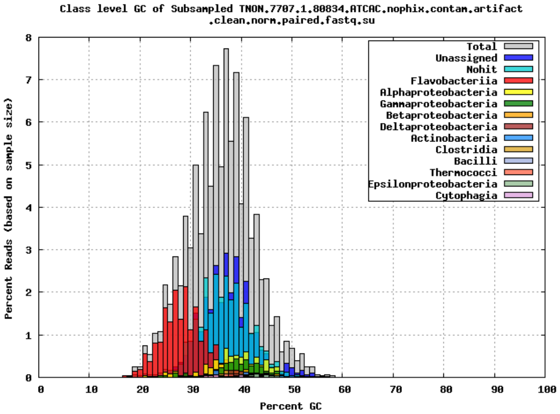
Kingdom



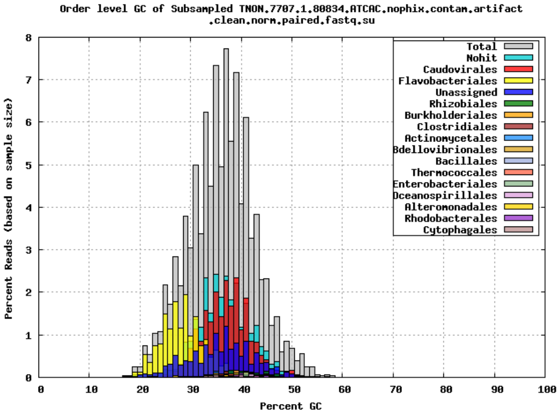
Phylum



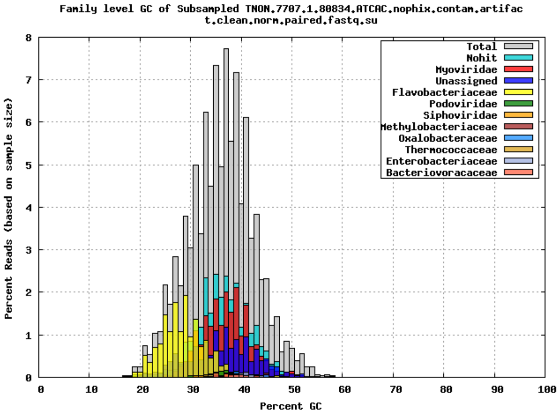
Class



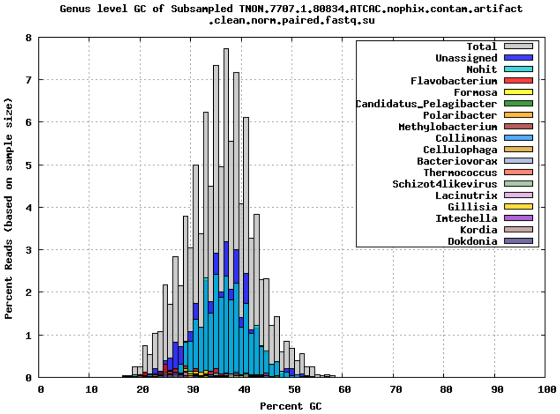
Order



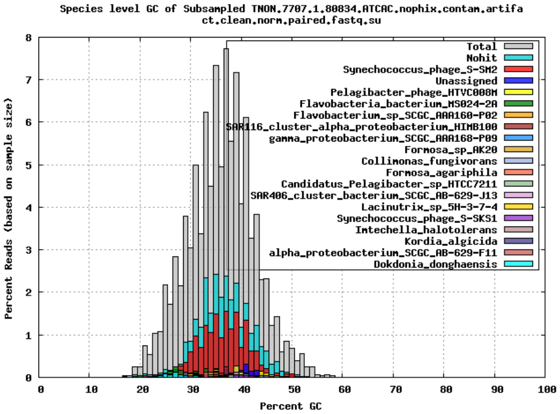
Family



Genus



Species



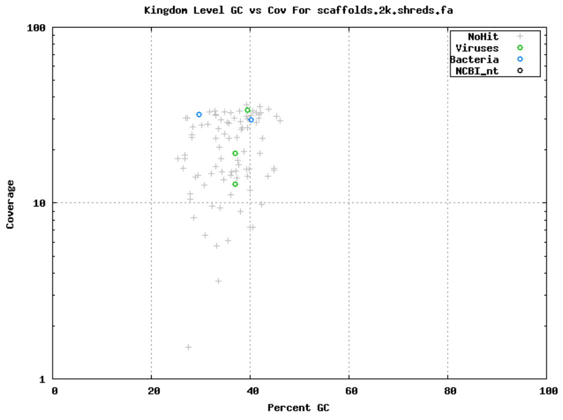
4. Assembly Statistics

Assembly method	SPAdes
Scaffold total	36
Contig total	38
Scaffold sequence length	376.4 kb
Contig sequence length	376.4 kb (0.0% gap)
Scaffold N/L50	9/16.9 kb
Contig N/L50	9/16.9 kb
Largest Contig	30.9 kb
Number of scaffolds >50 kb	0
Pct of genome in scaffolds >50 kb	0.0

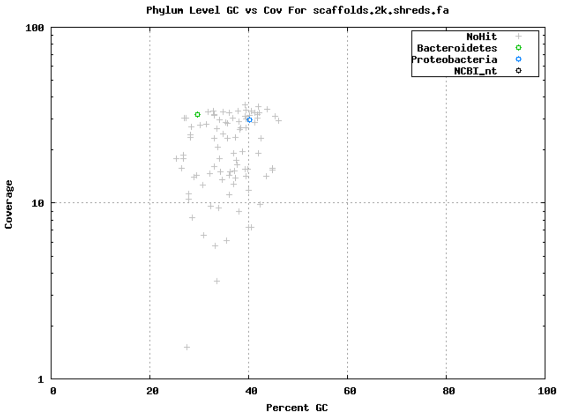
5. Assembly QC Results

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.

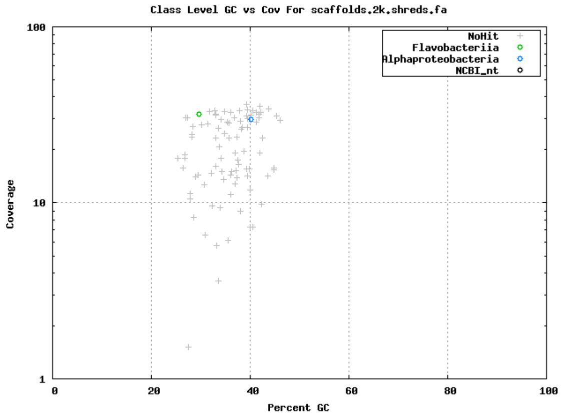
Kingdom



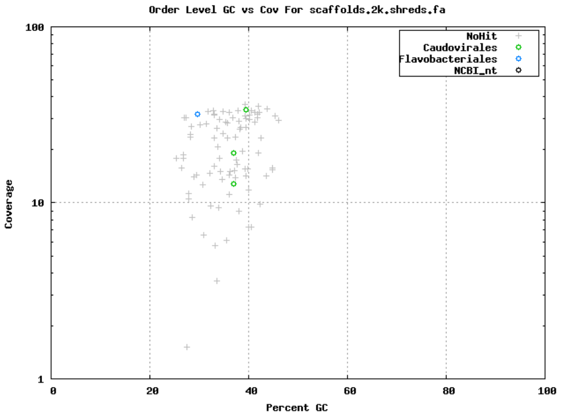
Phylum

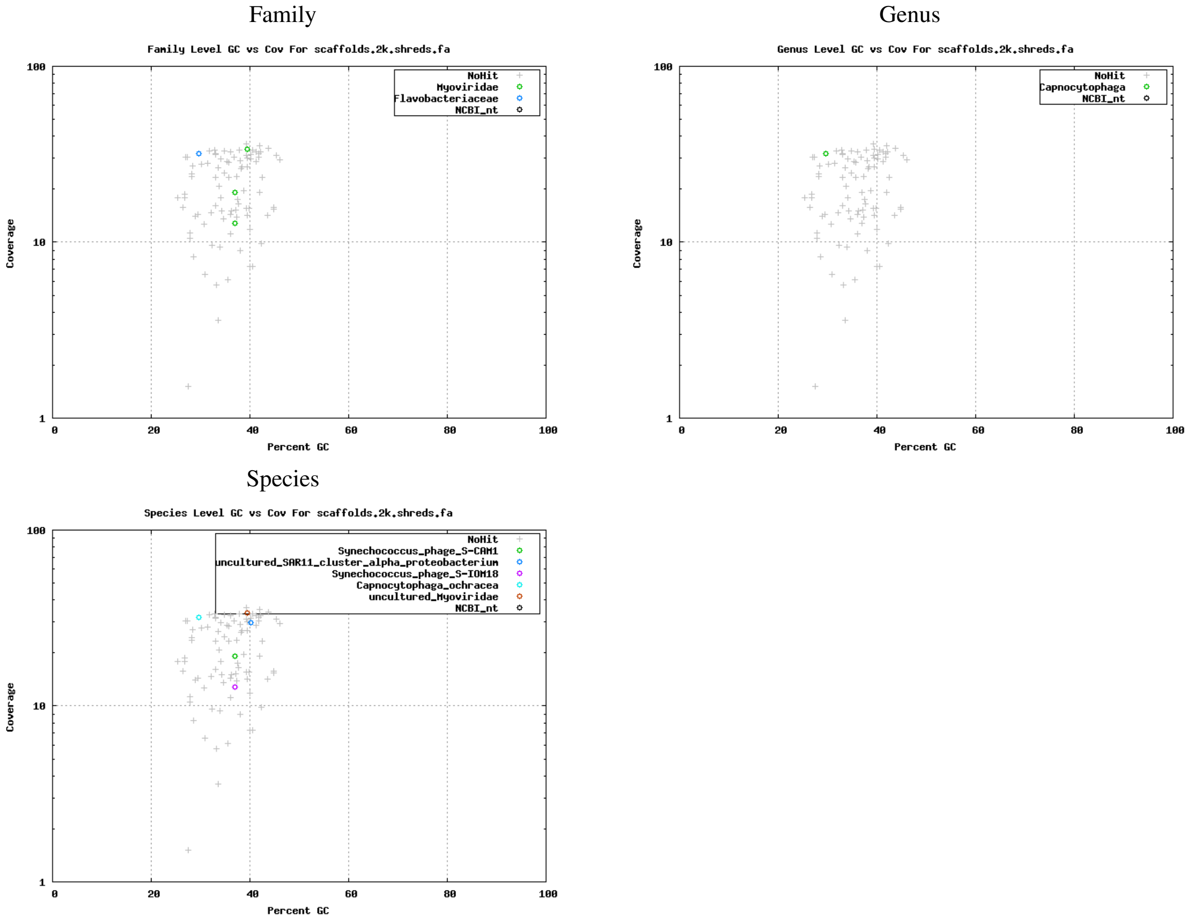


Class



Order



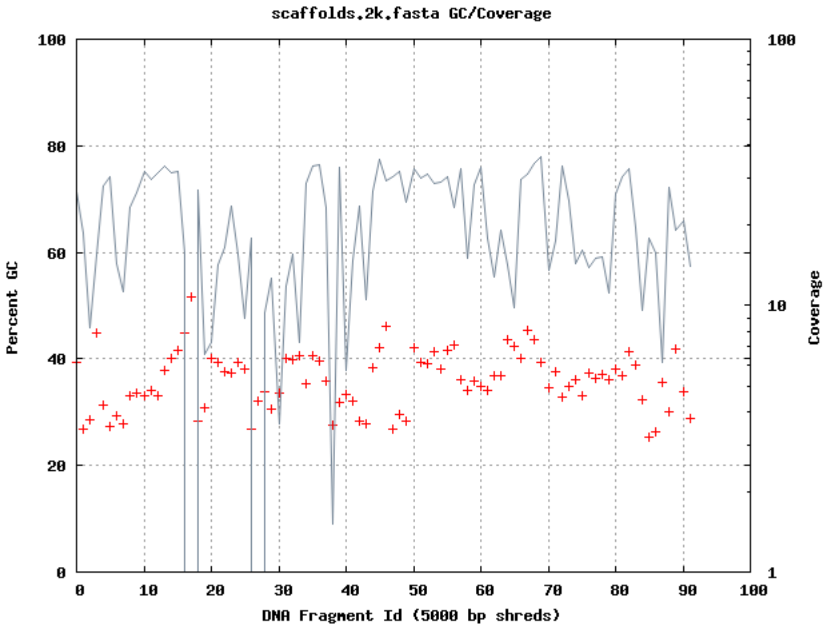


Family

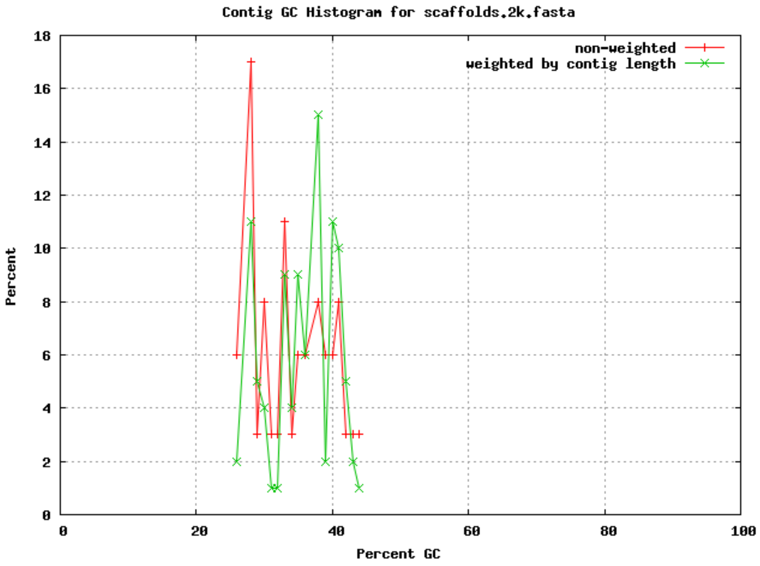
Genus

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.

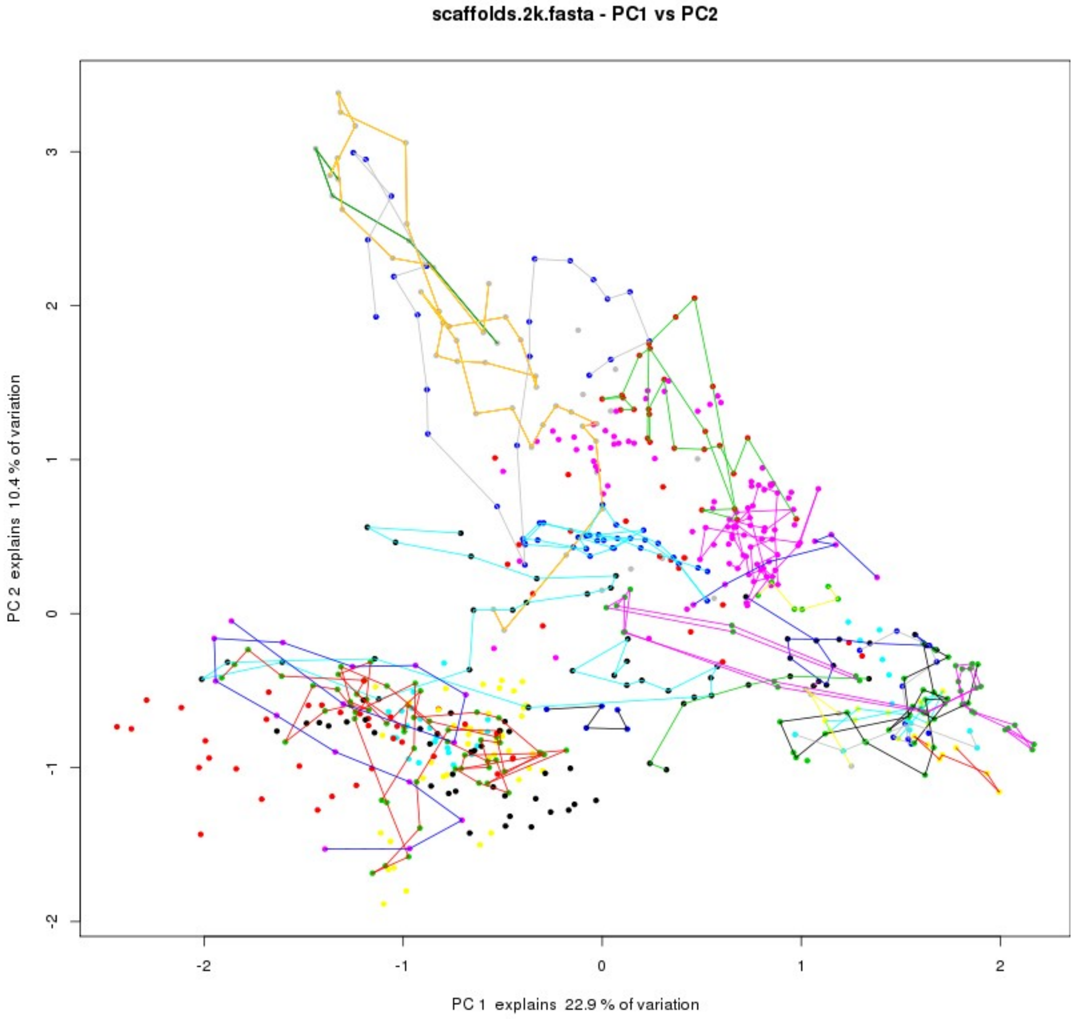


List of contigs and average percent GC, grouped in bins of 5:

Pct GC Bin	Contig Name
25	NODE_8_length_18021_cov_22.0241_ID_120431, NODE_12_length_14630_cov_24.5598_ID_120232, NODE_16_length_8407_cov_19.6982_ID_120247, NODE_20_length_6302_cov_18.9307_ID_120137, NODE_21_length_5886_cov_18.3365_ID_120693, NODE_23_length_5268_cov_18.4523_ID_119959, NODE_29_length_4180_cov_11.9775_ID_120307, NODE_33_length_2551_cov_6.41426_ID_117215 NODE_35_length_2266_cov_16.1443_ID_120573
30	NODE_3_length_22900_cov_23.4759_ID_119733, NODE_11_length_14634_cov_21.6911_ID_120701, NODE_18_length_7488_cov_24.4145_ID_120321, NODE_22_length_5445_cov_9.46271_ID_120685,

	NODE_25_length_4944_cov_7.0045_ID_119671, NODE_26_length_4660_cov_17.2321_ID_120712, NODE_27_length_4614_cov_11.7359_ID_120267, NODE_31_length_3582_cov_32.6422_ID_120105, NODE_32_length_2574_cov_28.3581_ID_120325, NODE_34_length_2478_cov_4.43541_ID_119673
35	NODE_2_length_26576_cov_25.4936_ID_120425, NODE_4_length_21435_cov_26.6528_ID_120251, NODE_6_length_19469_cov_22.7138_ID_118945, NODE_9_length_16865_cov_24.4964_ID_120717, NODE_10_length_16491_cov_25.9853_ID_118631, NODE_17_length_8191_cov_20.2942_ID_120703, NODE_19_length_6597_cov_13.5167_ID_93717, NODE_28_length_4252_cov_5.47057_ID_116923 NODE_36_length_2215_cov_22.7204_ID_120215
40	NODE_1_length_30856_cov_25.2158_ID_120183, NODE_5_length_20031_cov_26.0875_ID_120375, NODE_7_length_19354_cov_26.4856_ID_120327, NODE_13_length_13297_cov_25.218_ID_120413, NODE_14_length_11748_cov_19.7808_ID_119929, NODE_15_length_9112_cov_9.87844_ID_120001, NODE_24_length_4990_cov_0.987437_ID_120714, NODE_30_length_4062_cov_15.9613_ID_119223

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	4.8 %
archaea	0 %

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 39,870 reads totaling 10.0 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 38 contigs in 36 scaffolds, totalling 376.4 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, 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.