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

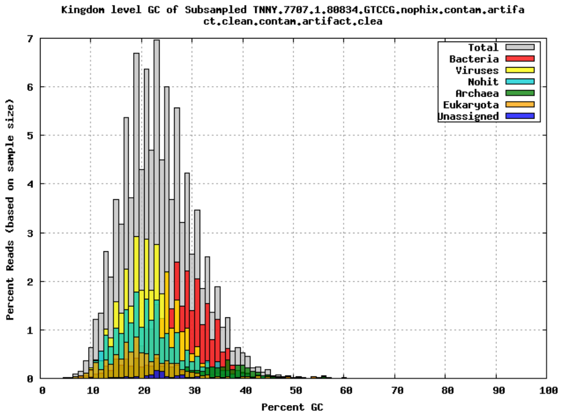
File name	TNNY.7707.1.80834.GTCCG.nophix.contam.artifact.clean.contam.artifact.clean.norm.paired.fa
Library	TNNY
Number of reads	38,982
Sequencing depth [†]	2X
Read type	2x251 bp

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

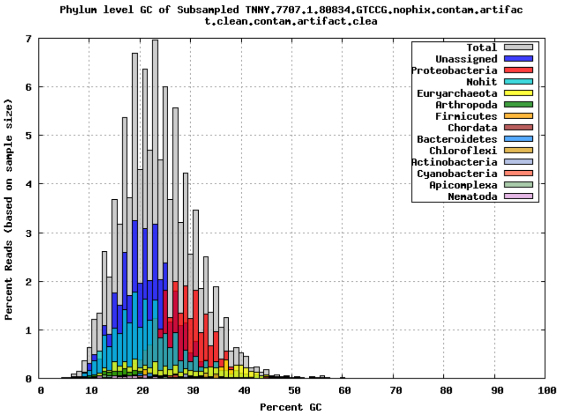
3. Read QC Results

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

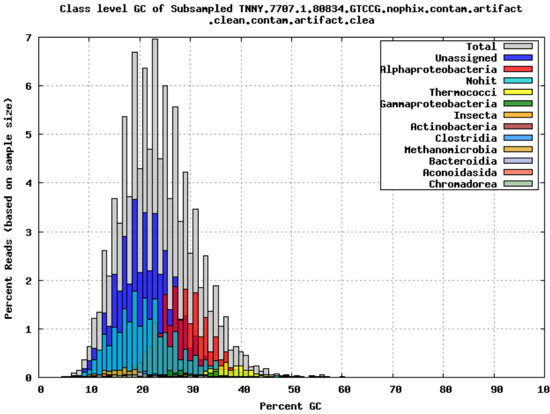
Kingdom



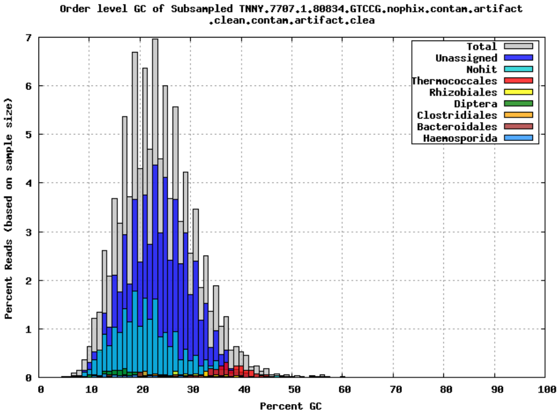
Phylum



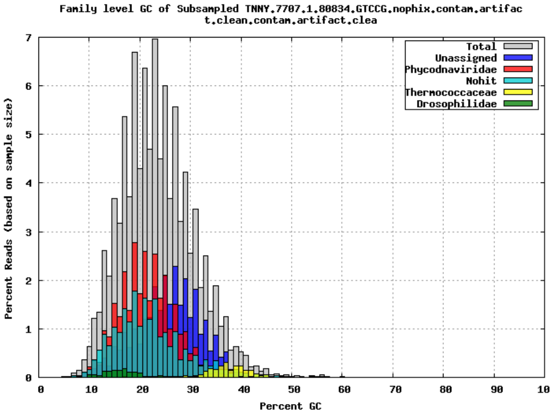
Class



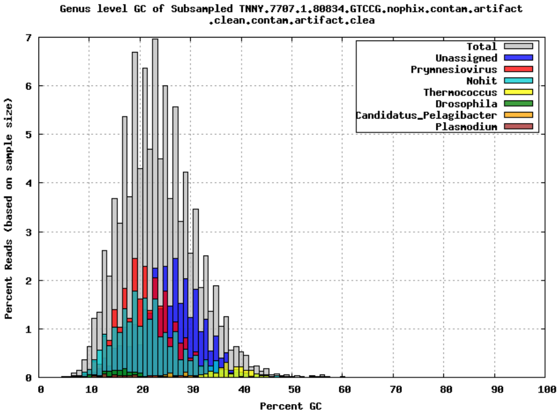
Order



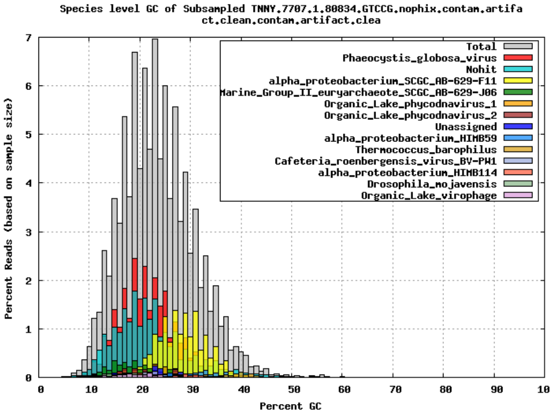
Family



Genus



Species



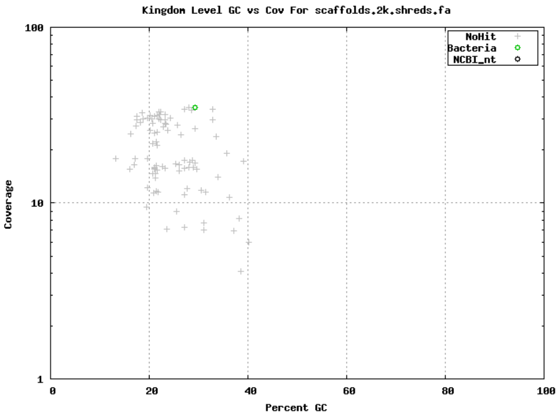
4. Assembly Statistics

Assembly method	SPAdes
Scaffold total	22
Contig total	30
Scaffold sequence length	385.2 kb
Contig sequence length	385.0 kb (0.0% gap)
Scaffold N/L50	4/44.9 kb
Contig N/L50	6/26.4 kb
Largest Contig	55.5 kb
Number of scaffolds >50 kb	2
Pct of genome in scaffolds >50 kb	28.2

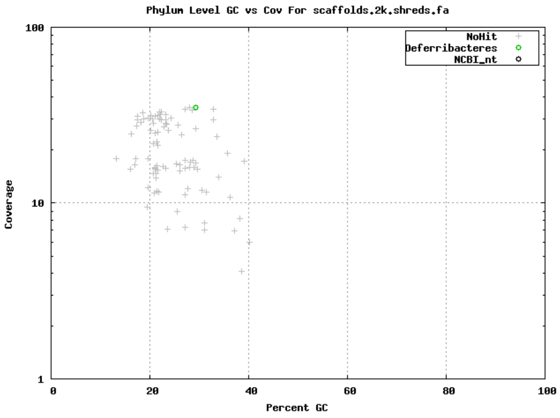
5. Assembly QC Results

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.

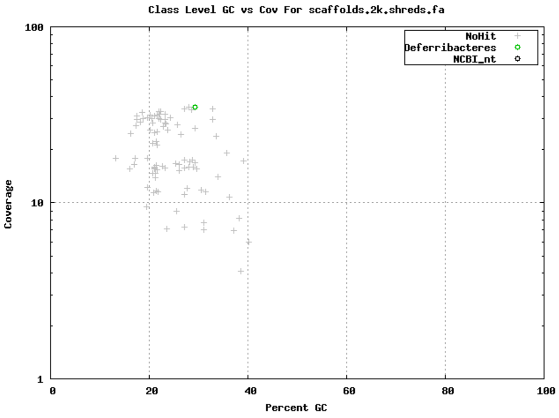
Kingdom



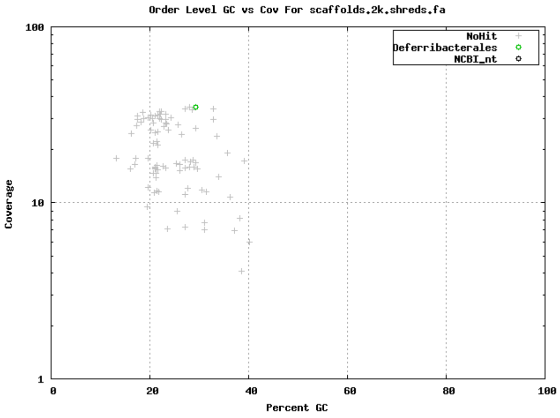
Phylum

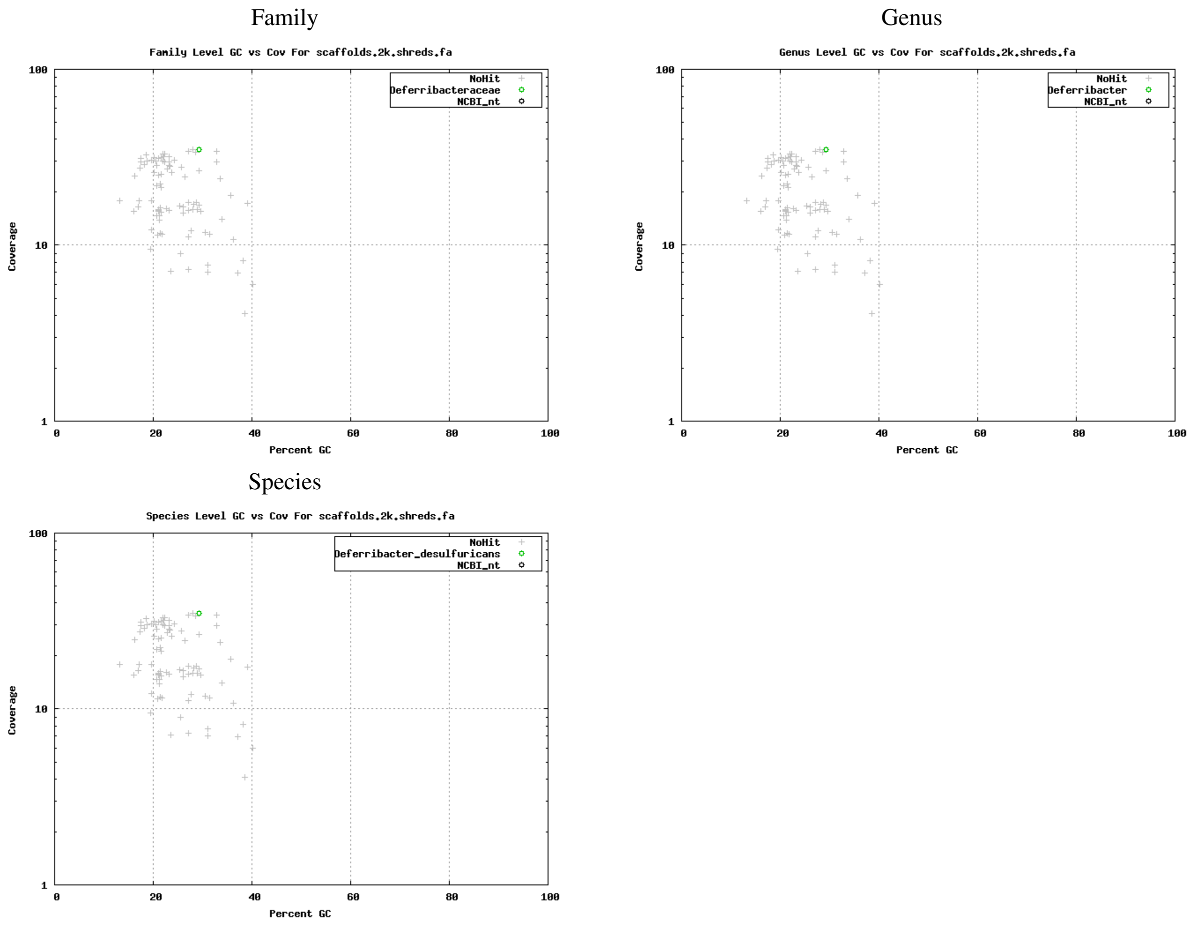


Class



Order



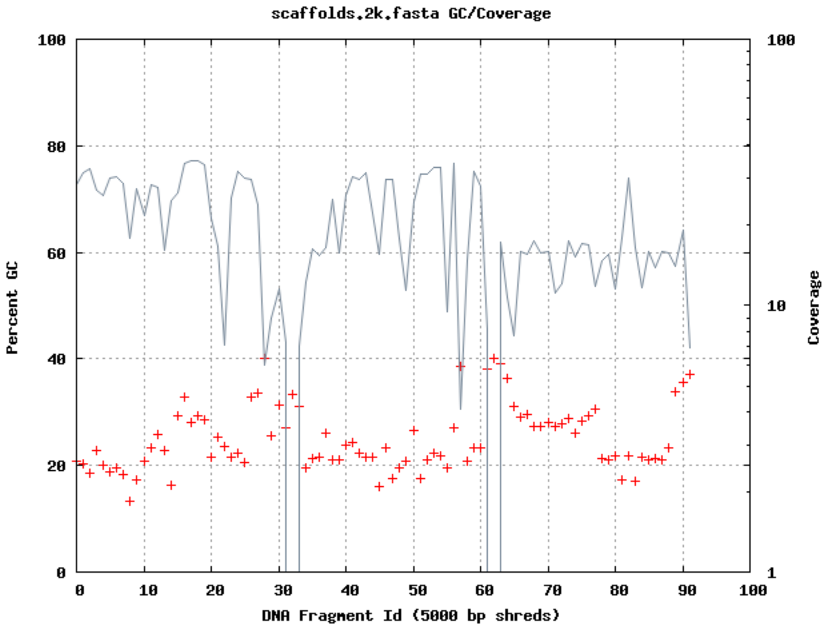


Family

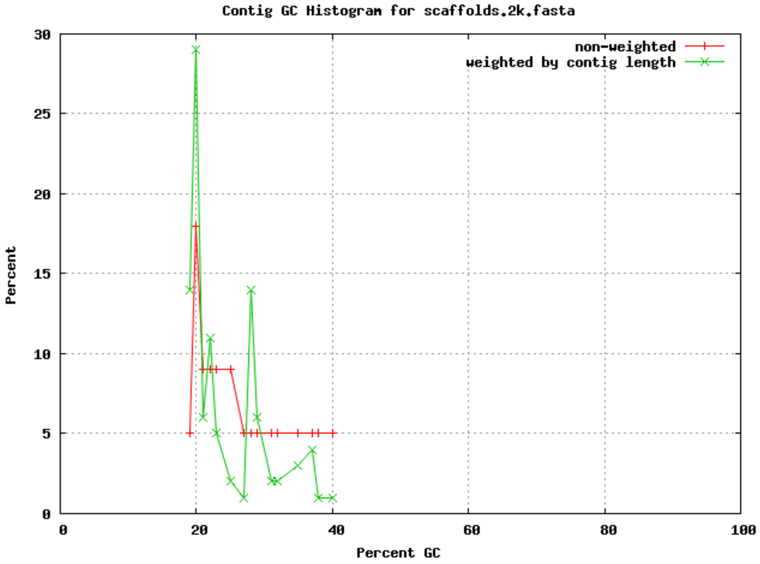
Genus

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.

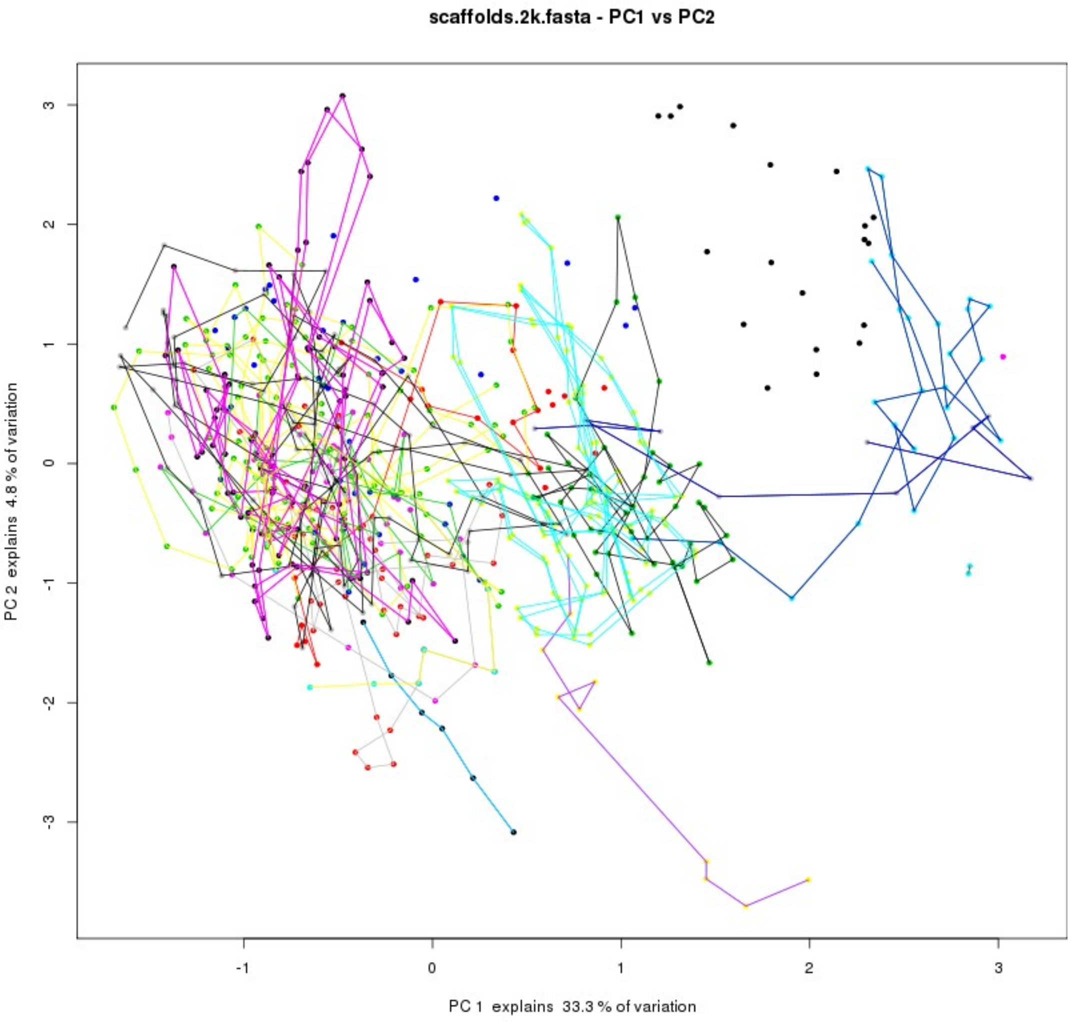


List of contigs and average percent GC, grouped in bins of 5:

Pct GC Bin	Contig Name
15	NODE_1_length_55506_cov_23.5006_ID_103587
20	NODE_3_length_49083_cov_21.2718_ID_109853, NODE_4_length_44946_cov_25.3923_ID_109431, NODE_5_length_28158_cov_21.854_ID_109843, NODE_7_length_14539_cov_20.4643_ID_109829, NODE_9_length_14247_cov_23.9434_ID_108787, NODE_11_length_10461_cov_15.7316_ID_109835, NODE_12_length_9228_cov_15.2517_ID_109849, NODE_13_length_8728_cov_11.4689_ID_109055, NODE_14_length_8448_cov_13.7143_ID_109845, NODE_15_length_7808_cov_16.4565_ID_109447

25	NODE_2_length_53019_cov_26.6767_ID_109861, NODE_6_length_24231_cov_26.5084_ID_109393, NODE_18_length_4789_cov_7.18399_ID_108445, NODE_19_length_4707_cov_27.7562_ID_108859 NODE_20_length_3578_cov_27.1646_ID_109463
30	NODE_16_length_6770_cov_7.95473_ID_108695, NODE_17_length_6650_cov_22.8849_ID_109385
35	NODE_8_length_14273_cov_9.95745_ID_107615, NODE_10_length_11178_cov_12.7956_ID_108411 NODE_21_length_2574_cov_3.36681_ID_103401
40	NODE_22_length_2242_cov_4.60082_ID_101525

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	5.6 %
archaea	2.74 %

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 38,982 reads totaling 9.8 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 30 contigs in 22 scaffolds, totalling 385.0 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.