

1. Project Information

Program	Microbial/CSP 2014
Seq Proj ID	1092845
Sequencing Project Name	Klosterneuburg WWTP active sludge C10.H0v2

2. Read Statistics

Illumina Std PE Statistics

File name	9720.1.141321.CGTTACTA-TCTTACG.anqdph.t.fastq
Library	ATSCA
Number of reads	2,069,228
Read type	2x146 bp

3. Read QC Results

The following are the results of reads screened against artifact sequences. Pairs of matching reads (>=95% ID) were removed from the dataset.

Illumina Std PE Read Filter Statistics

Description	Num Reads	Pct Reads
Input	2,069,228	100
Normalized removed	1,414,604	68.4
Total removed	1,414,604	68.4
Total remaining	654,624	31.6

The following are the results of reads screened against potential reagent and process contaminants but were not removed from the dataset.

Illumina Std PE Contamination Identification Statistics

Description	Num Reads	Pct Reads
Input	2,069,228	100
Contam identified	12	0.0

List of Contaminants Identified

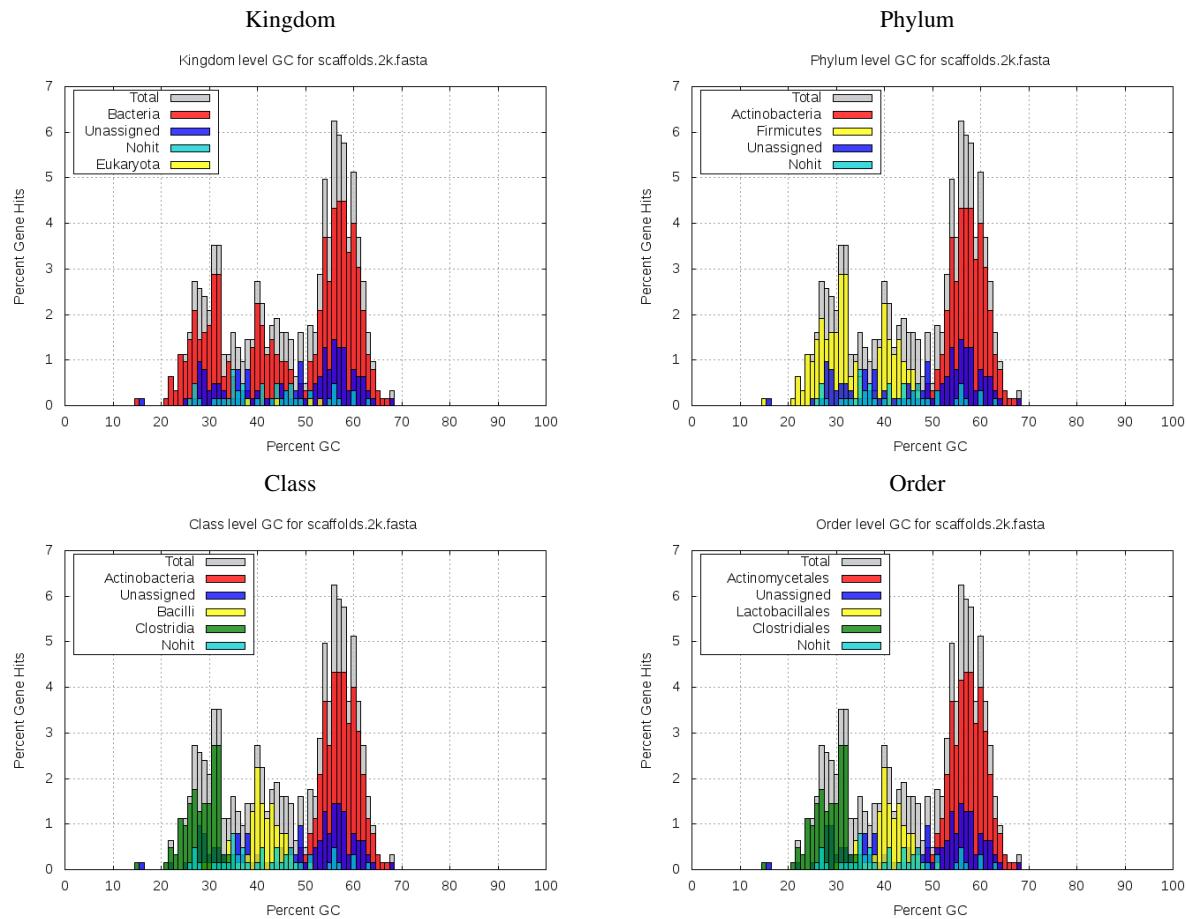
Description	Num Reads	Pct Reads
<i>Pseudomonas</i>	10	0.00
<i>Ralstonia</i>	2	0.00

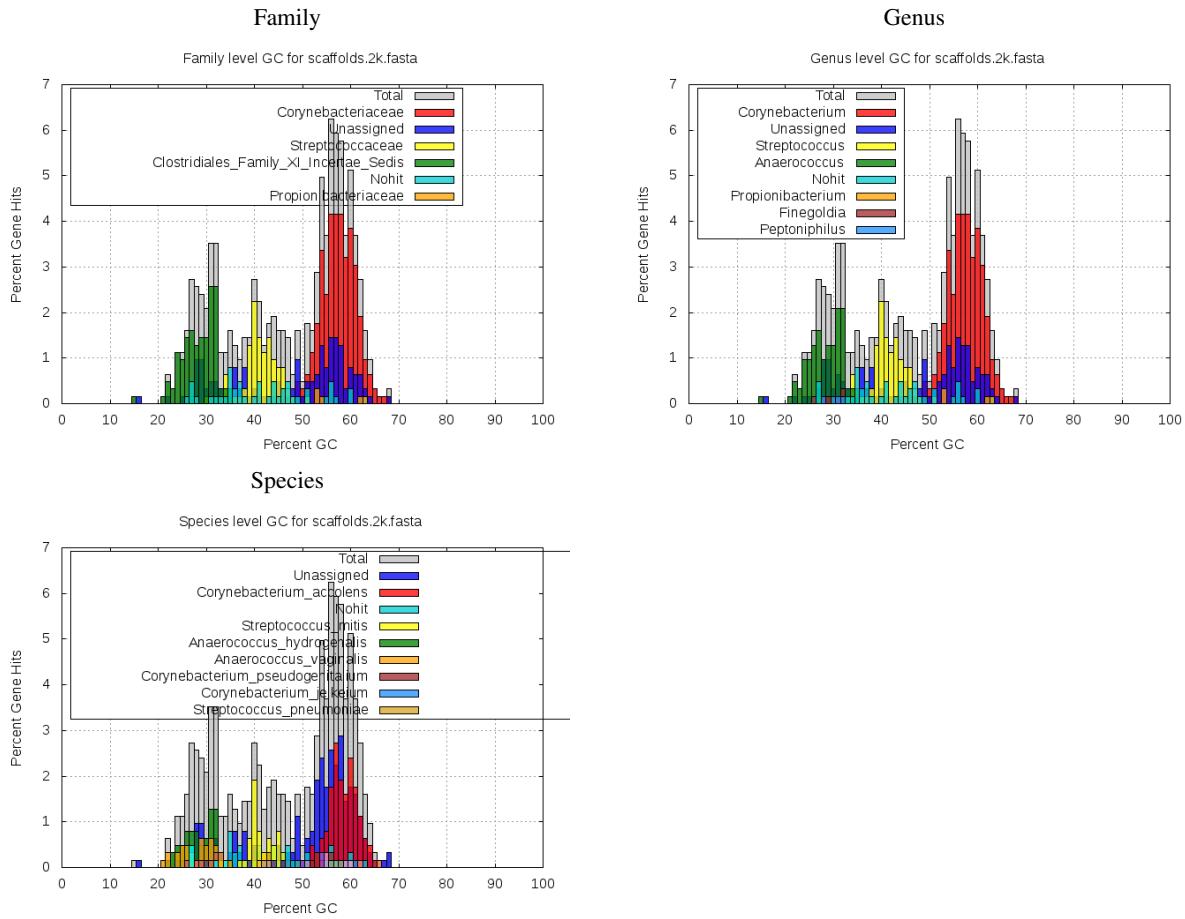
4. Assembly Statistics

Assembly method	SPAdes
Scaffold total	83
Contig total	86
Scaffold sequence length	0.609 Mb
Contig sequence length	0.609 Mb 0.010% gap
Scaffold N/L50	15/11.346 kb
Contig N/L50	15/11.346 kb
Largest Contig	54.0 kb
Number of scaffolds >50 kb	1
Pct of genome in scaffolds >50 kb	8.88
Pct of reads assembled	43.5

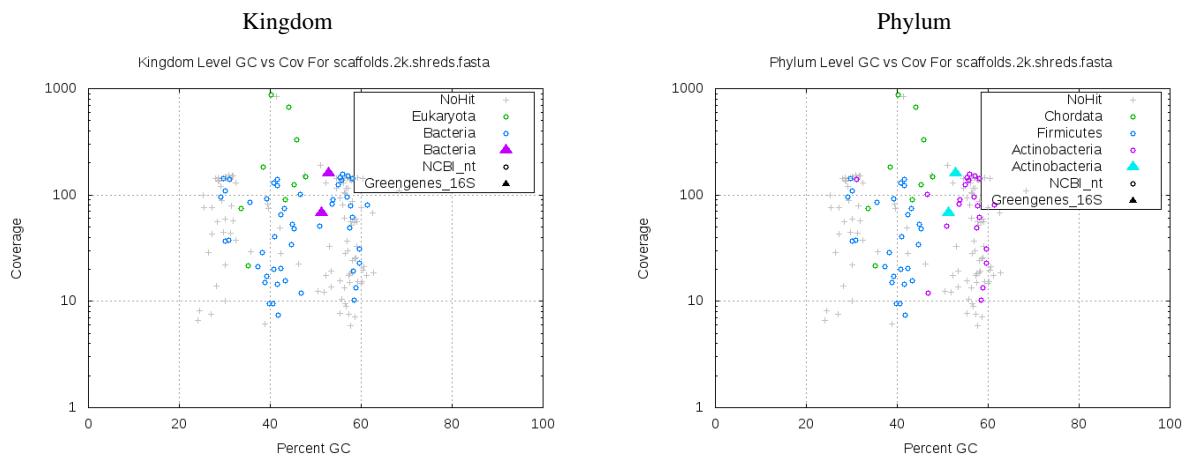
5. Assembly QC Results

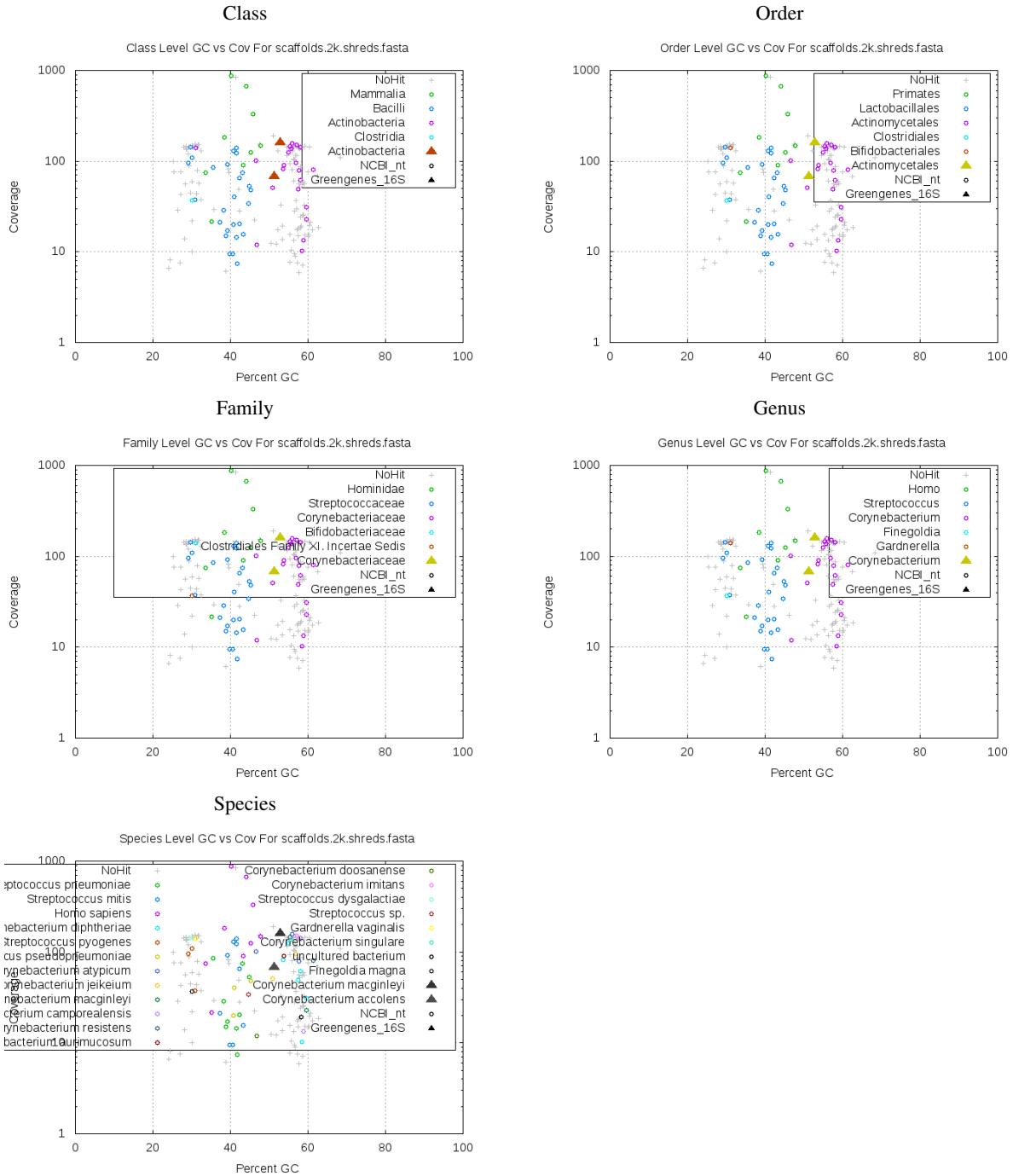
GC histogram of the predicted genes on each scaffold, overlaid with GC of hits based on LAST, shown for different taxonomic levels.



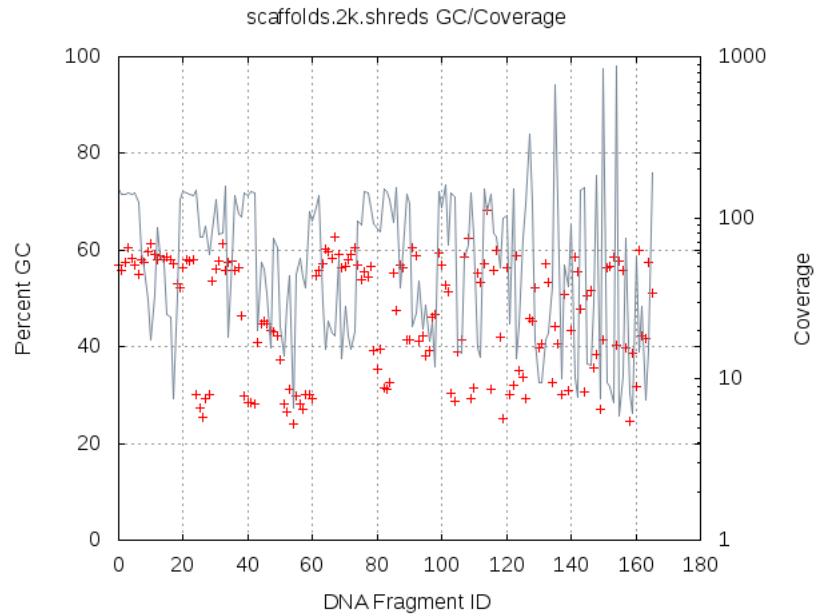


GC vs coverage based on GC of NCBI nt and Greengenes 16S rRNA gene hits to the assembly using megablast, shown for different taxonomic levels.

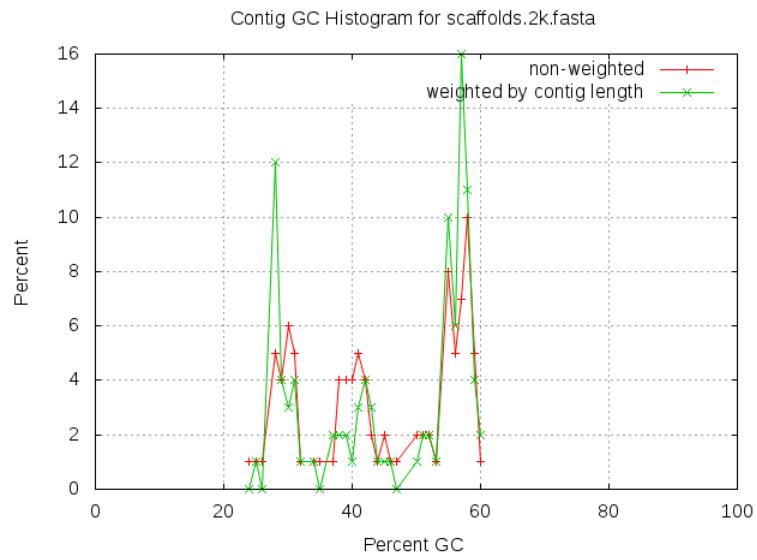




Coverage vs GC. Scaffolds were shredded into non-overlapping 5 kbp fragments and the GC content of each shred was plotted as a point, colored by scaffold id. Coverage was calculated by mapping the fragment library to the final assembly and plotted as connected points.

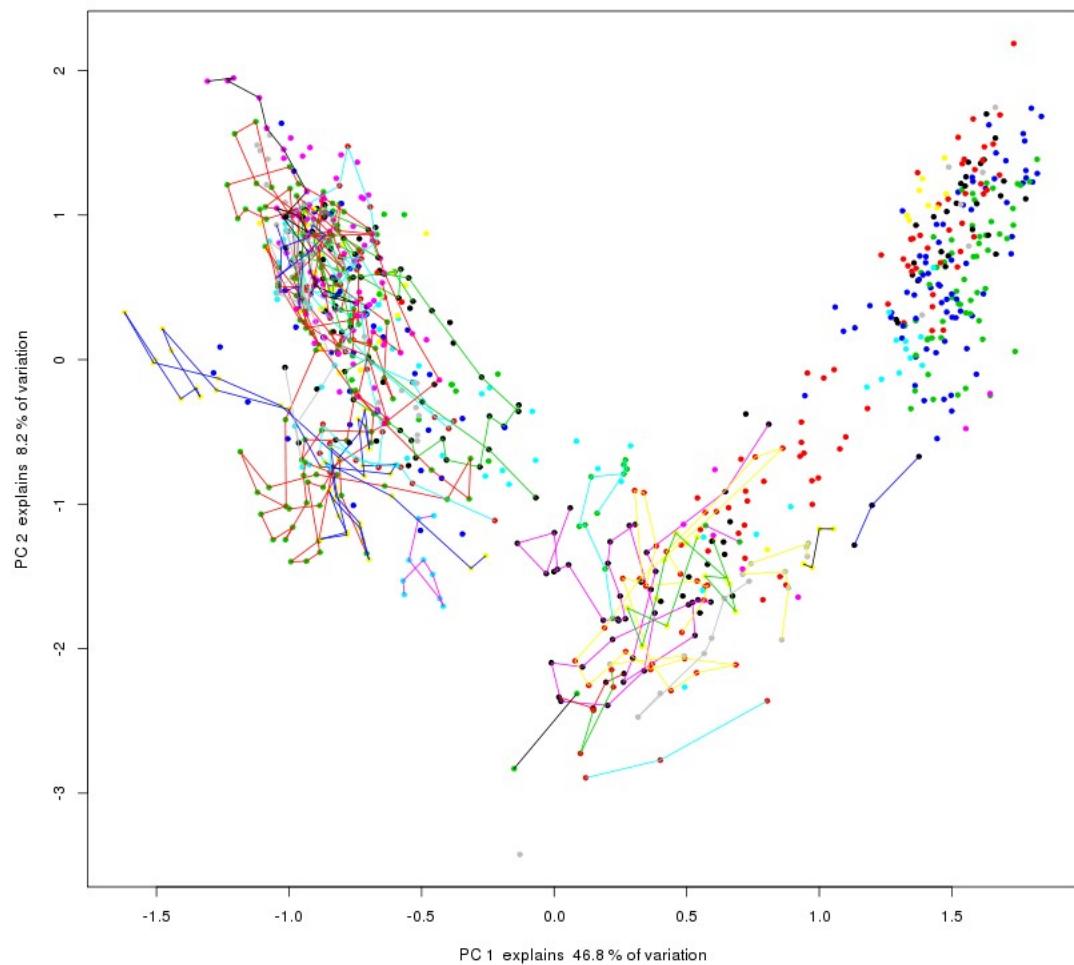


GC histogram of the scaffolds, including scaffold length weighted distribution.



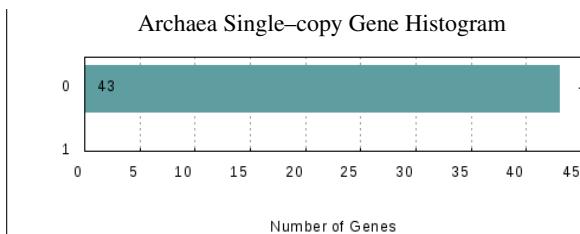
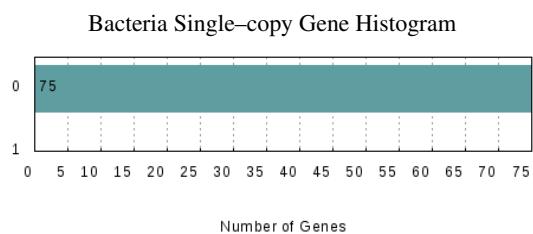
Principal component analysis of tetramer frequencies of scaffolds. Detectable variations are highlighted in color.

scaffolds.2k.fasta - PC1 vs PC2

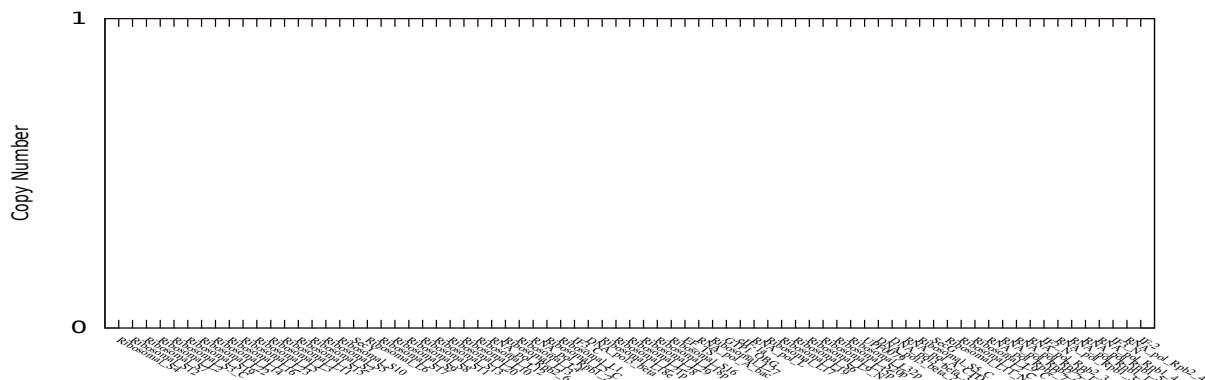


Estimated genome recovery derived from analysis of universal single-copy genes detected in final assembly.

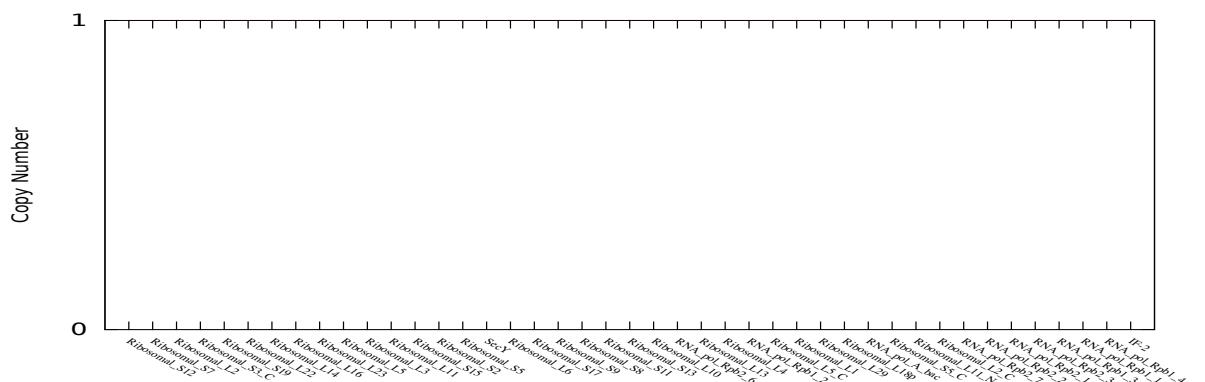
HMM	Pct Recovered
archaea	0 %
bacteria	0 %



Bacteria Single-copy Genes (N=75)



Archaea Single-copy Genes (N=43)



6. Sequence Data Availability

The following sequence fasta files can be downloaded from our JGI portal website.
<http://www.jgi.doe.gov/genome-projects>

7. Annotation Data Availability

The annotation of the assembled contigs can be found within IMG.
<http://img.jgi.doe.gov>

8. Methods

Cell-Enrichment Minimal Draft

Genome sequencing and assembly

The draft genome of *No Tax ID assigned* was generated at the DOE Joint Genome Institute (JGI) using the Illumina

technology [1]. An Illumina standard shotgun library was constructed and sequenced using the Illumina NextSeq platform which generated 2,069,228 reads totaling 302.1 Mbp. 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 [2], a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts. Furthermore, the reads were screened for human, dog and cat contaminant sequences. Reads with high k-mer coverage (>100X average k-mer depth) were normalized and error corrected to an average depth of 100X. Reads with an average k-mer depth of less than 2X were removed. Normalization was performed using BBNorm [3] and error correction was performed using Tadpole [3]. (1) Artifact filtered and normalized Illumina reads were assembled using SPAdes (version 3.6.0) [4]; (2) assembly contigs were trimmed 200 bp at the ends and discarded if length is <2 kbp or read coverage is <2X (BBMap: nodisk ambig=all maxindel=100 covstats=covstats.txt minhits=2 and filterbycoverage.sh: cov=covstats.txt mincov=2 minr=6 minp=95 minl=2000 trim=200 overwrite=true). Parameters for the SPAdes assembly were -t 16 -m 120 —sc —careful -k 25,55,95 —12. The final draft assembly contained 86 contigs in 83 scaffolds, totalling 0.609 Mbp in size and was based on 94.9 Mbp of Illumina data.

1. Bennett S. Solexa Ltd. *Pharmacogenomics*. 2004;5(4):433–8.
2. Mingkun L, Copeland A, Han J. DUK, unpublished, 2011.
3. B. Bushnell: BBTools software package, URL <http://sourceforge.net/projects/bbmap>.
4. Bankevich A, et.al, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; 19:455–77.
