MAG5 Assembly Research Report

R. Davison, J. Elmurib, N. Liu, K. Malaviya, N. Malinov, D. Podorefsky

June 3rd, 2019

INTRODUCTION

We are studying one genome of the 719 Metagenome Assembled Genomes of previously uncharacterized microorganisms originating from the goat rumen microbiome. This is in hopes that we can identify the microbe, its basic functions, and perhaps any novel genes it may encode. Studying these MAGs is crucial towards understanding the biotechnological capacities of microbes native to anaerobic environments. In herbivores, these microorganisms produce enzymes that break down plant mass into simple sugars that the host can digest. This process could be harnessed and applied at an industrial scale towards endeavors such as biofuel production. Additionally, these microbiomes hold a collection of genes that can possibly encode for antibiotic biosynthesis or the production of other valuable chemicals. In this report, a single genome from the dataset of 719 was studied in order to learn about its nature, closest relatives, and biological relevance through parsing the data using software including K-Base and CYverse.

METHODS

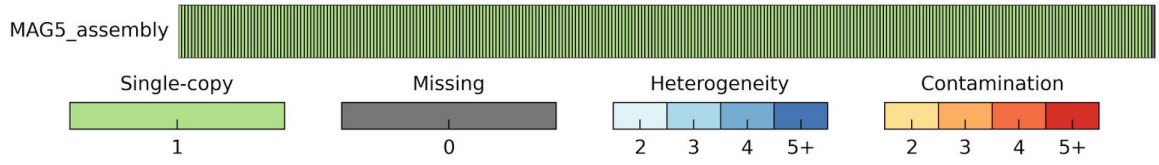
In order to understand the designated MAG, different apps and annotations tools were used in K-Base and CYVerse. On K-Base, the "CheckM" app assessed the quality of the MAG genome. The "RAST Annotate Assembly" app annotated the genome and the "Prokka" app re-annotated the genome to provide additional annotations. Using the "Build Metabolic Model" app, the annotated genome provided a roadmap for predicted metabolic pathways cataloged in the KEGG database to understand the MAG's functionality. The phylogeny was further investigated by using the "Insert Genome Into Species Tree" app which determines evolutionary relationships between organisms based on the differences in their genomic sequences. The app compares ribosomal subunit-specific clusters of orthologous genes (COGs), as these genes are present in every organism and can be easily compared among a range of organisms. Once the family of the MAG was established, the "Compare Proteomes" app juxtaposed neighboring proteomes to identify the nearest relative.

This information was then parsed on CYVerse using the app "Virsorter" to look for viral elements (complete phage and prophage) with different confidence categories in the MAG. To be specific, we used "Viromes" database as reference, which contains all bacterial and archaeal viral genomes in RefSeq and non-redundant predicted genes from viral metagenomes. Lastly, the CAZymes in the MAG were counted and analyzed utilizing the dbCAN meta server and supplied python script. The provided assembly file was uploaded to the server and three hidden Markov models variations connected to databases were used to determine the CAZymes. The python script, supplied by Tom, counted the CAZymes outputted by all of the HMM variations. Due to the HMMER database occasionally outputting a string of multiple CAZymes classes in individual elements of the spreadsheet matrix, manual counting was needed to supplement the script.

RESULTS

Pursuing the workflow described above, we formulated a comprehensive, surface level overview of the nature of the MAG.

The “CheckM” app produced the following results displayed in Figure 1, suggesting this was a high-quality MAG since it was near complete (99.367%), with no contamination.



Bin Name	Marker Lineage	# Genomes	# Markers	# Marker Sets	0	1	2	3	4	5+	Completeness	Contamination
MAG5_assembly	p__Firmicutes	100	295	158	1	294	0	0	0	0	99.367	0.0

Figure 1. “CheckM” App results displaying a 99.367% completeness and 0.0% contamination after assessing the quality of the app. These results were obtained by parsing the genomes for the number of copies, missing copies, heterogeneity, and contamination.

After constructing the metabolic model, the list of predicted pathways suggested that the MAG is a heterotroph. Numerous pathways pertaining to sugar degradation, such glycolysis/glucogenesis, pentose phosphate metabolism, galactose metabolism, fructose, and mannose metabolism, and others involving sugars are evidence to indicate that the organism was a heterotroph. While there is one identified pathway for carbon fixation in photosynthetic organisms, the sheer number of pathways associated with sugar degradation is a strong argument for the organism’s identity as a heterotroph.

Once the phylogenetic tree was built using the ribosomal subunit COGs, it suggested that the nearest relative of the MAG was in the *Selenomas* genus, *S. sputigena*.

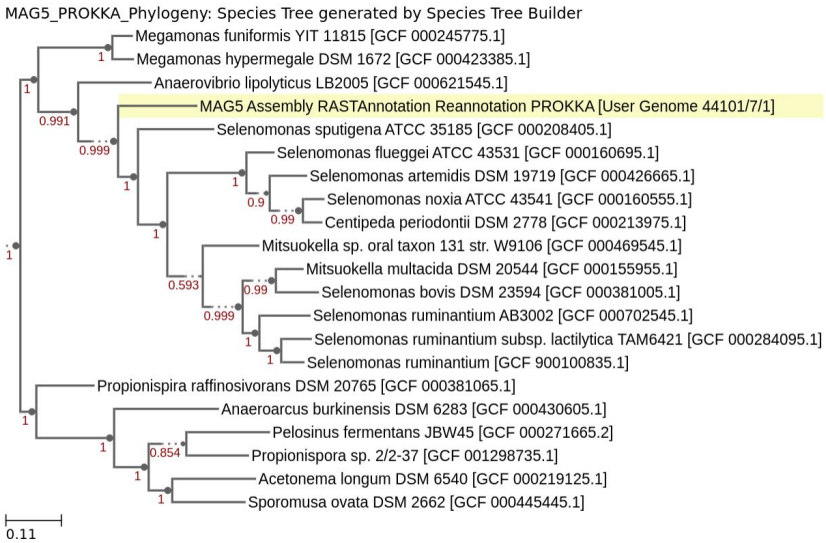


Figure 2. The phylogenetic tree produced by the “Insert Genome Into Species Tree 2.1.10” App.

After comparing the *MAG5* proteome with the *S. sputigena* proteome, K-base produced the following synteny plot (Figure 3a). The synteny plot confirmed that *S. sputigena* was not the same microbe as *MAG5*. For two identical microbes, the synteny plot should appear as an identity line (Figure 3b). The K-base produced synteny plot (Figure 3a) had hits all across the plot, suggesting the two proteomes were significantly different. Of the 2409 genes on *MAG5*, and 2268 genes on *S. sputigena,* >1600 were hits, indicating a ~70% similarity in the two proteomes.

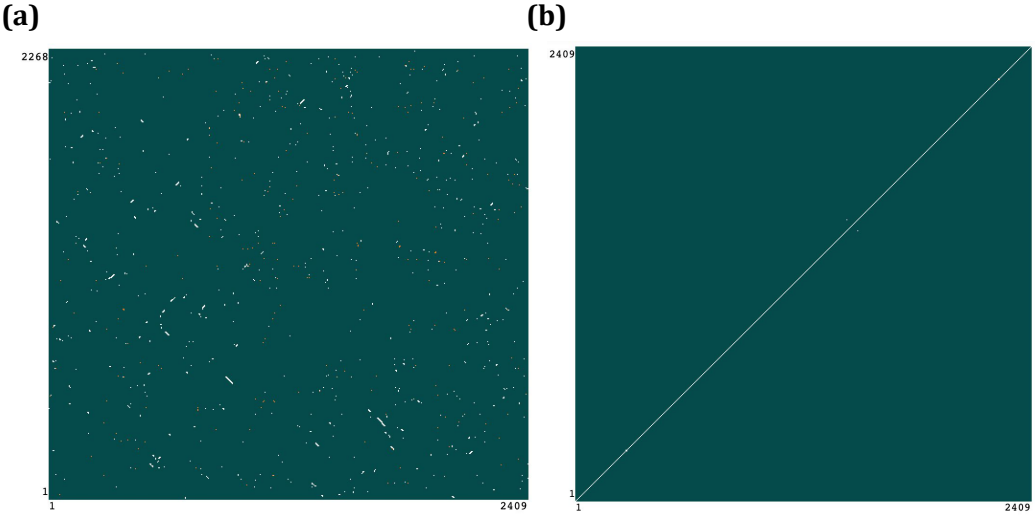


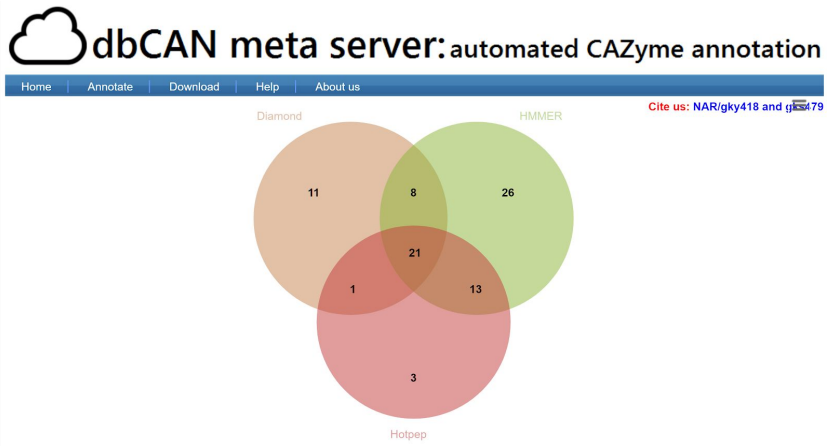
Figure 3. Synteny plots displaying the comparison of two proteomes. The K-base synteny plot represents the comparison of *S. sputigena* proteome and the MAG 5 proteome (a). The identity line represents two identical proteomes (b).

The metabolic model reconstruction provided a weak indication towards the MAG's potential for antibiotic synthesis, but nothing to fully confirm this possibility. The output of the reconstruction identified several pathways pertaining to antibiotic production: penicillin and cephalosporin biosynthesis, novobiocin biosynthesis, streptomycin biosynthesis, and several others. However, from observation of the KEGG maps associated with these pathways, only the streptomycin and novobiocin pathways displayed highlighted EC numbers. This indicates that the KBase algorithm predicts that the MAG produces a subset of enzymes required for the pathway. The metabolic model reconstruction also generated a list of predicted compounds which the MAG can produce. This list was queried against the KEGG database of known antibiotics resulting in no matches.

CAZymes are carbohydrate-active enzymes which allow the organism to do metabolic tasks, such as break down sugars or perform glycosylation. The dbCAN meta server allowed us to annotate which CAZymes were likely present in the metagenome assembly. Comparing the Diamond, HMMER, and Hotpep databases that each use hidden Markov model variations, a venn diagram (Figure 4a) was automatically constructed by dbCAN. Through further inspection of the output we found additional overlaps due to their being a string of multiple enzymes in the spreadsheet for several entries detected by HMMER, resulting in a total prediction of 26 CAZymes (Figure 4b). The main classes of enzymes, in order of count, are glucosyltransferase (20), glycoside hydrolase (4), carbohydrate esterase (1), carbohydrate binding module (1). Since CAZymes often work together with each other

and with other important genes (e.g. TFs, sugar transporters) to synergistically degrade or synthesize various highly complex carbohydrates, we also used the CGC Finder on dbCAN server to identify carbohydrate gene clusters (CGC) in *MAG5*. They define CGCs as genomic regions containing at least one CAZyme gene, one transporter/TC gene and one transcription factor/TF gene. We found 13 CGCs in total in MAG 5.

(a)



(b)

GT2	6
GT51	1
GT4	1
GT35	1
CBM48	1
GT5	1
GT83	1
GT30	1
GT19	1
CE11	1
GH84	1
GT41	5
GT28	1
GT9	1
GH13	2
GH77	1

(c)

Show 15 entries Search:

Cluster	Contig ID	# of Genes	# of CAZymes	# of TC	# of TF	Cluster Start	Cluster End
CGC1		12	1	6	0	184995	200444
CGC10		2	1	1	0	151765	154876
CGC11		3	1	1	0	12762	15952
CGC12		6	1	2	0	125774	131698
CGC13		3	1	1	0	170482	174002
CGC2		4	1	1	0	295175	298678
CGC3		2	1	1	0	454119	456221
CGC4		3	1	2	0	9284	12401
CGC5		8	3	4	0	65624	82334
CGC6		3	2	1	0	142971	148385
CGC7		12	2	6	0	240595	257338
CGC8		2	1	1	0	1667	3655
CGC9		3	1	1	0	32391	35584

Showing 1 to 13 of 13 entries First Previous 1 Next Last

Figure 4. CAZyme prediction **(a)** Venn diagram displaying the combined results from the three tools for prediction of CAZymes present in the assembly, via the dbCAN meta server. **(b)** Count of CAZymes genes in *MAG5* obtained through Python pipeline. **(c)** 13 carbohydrate gene clusters (CGCs) found in MAG 5.

Virsorter predicted only one not-so-sure prophage in our *MAG5*. It is 95,476 bp long (94,934 - 190,409) with 100 genes (gene 74 - 173) on the contig NODE_112. There are genes encoding for transcriptional regulator, ribonucleases, restriction enzymes and those involved in DNA replication and repair, RNA synthesis and chromosome partitioning (see Appendix). The existence of these genes but the lack of transposons suggests that this prophage may lose its mobility after integration into the bacterial genome. We also aligned the prophage sequence against BLAST nucleotide collection and found that there was a hit of *S. sputigena ATCC 35185* with only 2% coverage and 76.77% identity.

DISCUSSION

The phylogenetic tree indicates that the closest relative of the MAG was in the genus *Selenomonas*, which is within the phylum firmicutes which "Check M" provided us. Most published research is on cow gut microbiomes, so this led us to believe that this was a goat specific strain of *Selenomonas*. Furthermore using RNAmmer, the 5s ribosomal subunit was identified and then run through BLAST. All the top hits were in the phylum firmicutes, while three of the top 6 were within the genus *Selenomonas*.

The major class of CAZymes predicted in the assembly was glucosyltransferases, which serve several purposes. First and foremost, the GT41 class is capable of attaching saccharide moieties to target proteins via residues containing amide groups (asparagine). This glycosylation allows for stability in surface attachment of the organism [1]. In *Actinobacillus pleuropneumoniae*, a gram-negative bacterium, the GT41 family was found to hydrolyze sugars when there was no protein substrate present, allowing for UDP-Glucose to be converted into α/β glucose and UDP [2]. The GT41 class additionally includes O-linked beta-N-acetylglucosamine (O-GlcNAc) transferase, which is structurally similar to glucosyltransferases, and catalyzes the addition of O-GlcNAc to serine and threonine [3]. The GT2 is a large family with diverse sequences from animal, plant, yeast and bacteria. There are over 12 distinct GT2 functions being recognized including cellulose synthase, chitin synthase, mannosyltransferase, glucosyltransferase, galactosyltransferase, rhamnosyltransferase, etc. There are also four glycoside hydrolases (GH13, GH77, GH84), one carbohydrate esterase (CE11) and one carbohydrate binding module (CBM48) genes existing in the *MAG5* genome. These CAZyme genes indicate *MAG5* has potential for carbohydrate biosynthesis and degradation. However, due to the fact that we didn't find any cellulase or hemicellulase and CAZyme genes account for only 1% of the total number of genes in the genome, we suggest that *MAG5* may not be a dominant species for breaking down grass and other complex polysaccharides, but has a role in secondary or tertiary fermentation downstream carbohydrate metabolism.

Regards to potential antibiotic production, the fact that only several of the EC numbers were highlighted does not suggest that the MAG engages in the entirety of the pathway. One possibility is that only the subset of reactions surrounding the highlighted enzymes within the larger pathway are carried out by the MAG. Thus, the MAG produces several of these "intermediate" metabolites within the pathway, but perhaps for purposes aside from antibiotic synthesis. Searching the list of predicted compounds for these intermediate metabolites validates this speculation. Another possibility is that the MAG can perform the entirety of some of the identified antibiotic biosynthesis pathways, but they simply aren't recognized by the Kbase algorithm. However, this is highly unlikely as the quality of the metabolic model depends on the quality of the annotated genome. Given that the genome is over 99% complete with zero contamination, this speculation is ruled out.

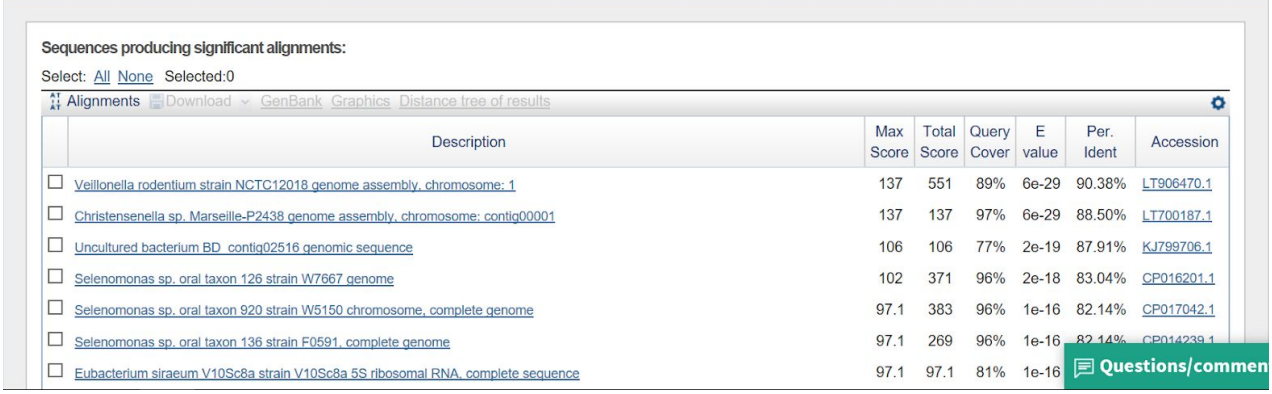
CONCLUSION

MAG5 is a near-complete genome assembly with 99% completeness and 0% contamination. The metabolic model constructed on K-Base indicates that *MAG5* is a heterotroph. Additionally the phylogenetic tree and synteny plots both suggest the MAG is of the genus *Selenomonas*. There are 26 CAZyme genes predicted by dbCAN2 that predominantly deal with sugar metabolism and glycosylation. Although the metabolic model predicted the MAG to engage in pathways pertaining to antibiotic synthesis, closer analysis proved that this possibility is very minimal to nonexistent.

REFERENCES

1. Grass, S., Buscher, A. Z., Swords, W. E., Apicella, M. A., Barenkamp, S. J., Ozchlewski, N., and Geme, J. W. S. (2003). "The Haemophilus influenzae HMW1 adhesin is glycosylated in a process that requires HMW1C and phosphoglucomutase, an enzyme involved in lipooligosaccharide biosynthesis." *Molecular Microbiology*, 48(3), 737–751.
2. Naegeli, A., Neupert, C., Fan, Y.-Y., Lin, C.-W., Poljak, K., Papini, A. M., Schwarz, F., and Aebi, M. (2014). "Molecular Analysis of an Alternative N-Glycosylation Machinery by Functional Transfer from Actinobacillus pleuropneumoniae to Escherichia coli." *Journal of Biological Chemistry*, 289(4), 2170–2179.
3. "Pfam: Family: Glyco_transf_41 (PF13844)." (n.d.). <http://pfam.xfam.org/family/Glyco_transf_41> (May 29, 2019).

APPENDIX



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☐	Veillonella rodentium strain NCTC12018 genome assembly, chromosome: 1	137	551	89%	6e-29	90.38%	LT906470.1
☐	Christensenella sp. Marseille-P2438 genome assembly, chromosome: contig00001	137	137	97%	6e-29	88.50%	LT700187.1
☐	Uncultured bacterium BD_contig02516 genomic sequence	106	106	77%	2e-19	87.91%	KJ799706.1
☐	Selenomonas sp. oral taxon 126 strain W7667 genome	102	371	96%	2e-18	83.04%	CP016201.1
☐	Selenomonas sp. oral taxon 920 strain W5150 chromosome, complete genome	97.1	383	96%	1e-16	82.14%	CP017042.1
☐	Selenomonas sp. oral taxon 136 strain F0591, complete genome	97.1	269	96%	1e-16	82.14%	CP014239.1
☐	Eubacterium siraeum V10Sc8a strain V10Sc8a 5S ribosomal RNA, complete sequence	97.1	97.1	81%	1e-16		

Questions/commen

Figure 5. RNAmmer was used to annotate the MAG 5 genome for its ribosomal RNA genes. The 5s-ribosomal RNA gene was obtained and subsequently submitted to BLAST. Three hits pertaining to the genus *Selenomonas* with strong values in each significant metric serve as evidence to our speculation that MAG 5 is most likely a member of the *Selenomonas* genus.

List of COGS:

COG0012	COG0012	Predicted GTPase, probable translation factor [Translation, ribosomal structure and biogenesis].
COG0013	AlaS	Alanyl-tRNA synthetase [Translation, ribosomal structure and biogenesis].
COG0016	PheS	Phenylalanyl-tRNA synthetase alpha subunit [Translation, ribosomal structure and biogenesis].
COG0018	ArgS	Arginyl-tRNA synthetase [Translation, ribosomal structure and biogenesis].
COG0030	KsgA	Dimethyladenosine transferase (rRNA methylation) [Translation, ribosomal structure and biogenesis].
COG0041	PurE	Phosphoribosylcarboxyaminoimidazole (NCAIR) mutase [Nucleotide transport and metabolism].
COG0046	PurL	Phosphoribosylformylglycinamidine (FGAM) synthase, synthetase domain [Nucleotide transport and metabolism].
COG0048	RpsL	Ribosomal protein S12 [Translation, ribosomal structure and biogenesis].
COG0049	RpsG	Ribosomal protein S7 [Translation, ribosomal structure and biogenesis].
COG0051	RpsJ	Ribosomal protein S10 [Translation, ribosomal structure and biogenesis].
COG0052	RpsB	Ribosomal protein S2 [Translation, ribosomal

		structure and biogenesis].
COG0072	PheT	Phenylalanyl-tRNA synthetase beta subunit [Translation, ribosomal structure and biogenesis].
COG0080	RplK	Ribosomal protein L11 [Translation, ribosomal structure and biogenesis].
COG0081	RplA	Ribosomal protein L1 [Translation, ribosomal structure and biogenesis].
COG0082	AroC	Chorismate synthase [Amino acid transport and metabolism].
COG0086	RpoC	DNA-directed RNA polymerase, beta' subunit/160 kD subunit [Transcription].
COG0087	RplC	Ribosomal protein L3 [Translation, ribosomal structure and biogenesis].
COG0088	RplD	Ribosomal protein L4 [Translation, ribosomal structure and biogenesis].
COG0089	RplW	Ribosomal protein L23 [Translation, ribosomal structure and biogenesis].
COG0090	RplB	Ribosomal protein L2 [Translation, ribosomal structure and biogenesis].
COG0091	RplV	Ribosomal protein L22 [Translation, ribosomal structure and biogenesis].
COG0092	RpsC	Ribosomal protein S3 [Translation, ribosomal structure and biogenesis].
COG0093	RplN	Ribosomal protein L14 [Translation, ribosomal structure and biogenesis].
COG0094	RplE	Ribosomal protein L5 [Translation, ribosomal structure and biogenesis].
COG0096	RpsH	Ribosomal protein S8 [Translation, ribosomal structure and biogenesis].
COG0097	RplF	Ribosomal protein L6P/L9E [Translation, ribosomal structure and biogenesis].
COG0098	RpsE	Ribosomal protein S5 [Translation, ribosomal structure and biogenesis].
COG0099	RpsM	Ribosomal protein S13 [Translation, ribosomal structure and biogenesis].
COG0100	RpsK	Ribosomal protein S11 [Translation, ribosomal structure and biogenesis].
COG0102	RplM	Ribosomal protein L13 [Translation, ribosomal structure and biogenesis].
COG0103	RpsI	Ribosomal protein S9 [Translation, ribosomal structure and biogenesis].
COG0105	Ndk	Nucleoside diphosphate kinase [Nucleotide transport and metabolism].
COG0126	Pgk	3-phosphoglycerate kinase [Carbohydrate transport

		and metabolism].
COG0127	COG0127	Xanthosine triphosphate pyrophosphatase [Nucleotide transport and metabolism].
COG0130	TruB	Pseudouridine synthase [Translation, ribosomal structure and biogenesis].
COG0150	PurM	Phosphoribosylaminoimidazole (AIR) synthetase [Nucleotide transport and metabolism].
COG0151	PurD	Phosphoribosylamine-glycine ligase [Nucleotide transport and metabolism].
COG0164	RnhB	Ribonuclease HII [DNA replication, recombination, and repair].
COG0172	SerS	Seryl-tRNA synthetase [Translation, ribosomal structure and biogenesis].
COG0185	RpsS	Ribosomal protein S19 [Translation, ribosomal structure and biogenesis].
COG0186	RpsQ	Ribosomal protein S17 [Translation, ribosomal structure and biogenesis].
COG0215	CysS	Cysteinyl-tRNA synthetase [Translation, ribosomal structure and biogenesis].
COG0244	RplJ	Ribosomal protein L10 [Translation, ribosomal structure and biogenesis].
COG0256	RplR	Ribosomal protein L18 [Translation, ribosomal structure and biogenesis].
COG0343	Tgt	Queuine/archaeosine tRNA-ribosyltransferase [Translation, ribosomal structure and biogenesis].
COG0504	PyrG	CTP synthase (UTP-ammonia lyase) [Nucleotide transport and metabolism].
COG0519	GuaA	GMP synthase, PP-ATPase domain/subunit [Nucleotide transport and metabolism].
COG0532	InfB	Translation initiation factor 2 (IF-2; GTPase) [Translation, ribosomal structure and biogenesis].
COG0533	QRI7	Metal-dependent proteases with possible chaperone activity [Posttranslational modification, protein turnover, chaperones].

List of genes in the prophage sequence:

Putative HTH-type transcriptional regulator YwnA
Ribonuclease HI
Pyrophosphatase PpaX
putative chromosome-partitioning protein ParB
Sporulation initiation inhibitor protein Soj
Membrane protein insertase YidC 2
Ribonuclease P protein component

Chromosomal replication initiator protein DnaA
DNA polymerase III subunit beta
DNA replication and repair protein RecF
DNA gyrase subunit B
Protease HtpX
Putative DNA repair helicase RadD
RNA polymerase sigma factor FliA
Transcriptional regulator LytR
Type-1 restriction enzyme R protein
putative type I restriction enzymeP M protein
Hypothetical proteins