School of Biological Sciences

ASSESSMENT COVER SHEET AND TEMPLATE

Section A – to be completed by the student

Student Number	**2228202**		
Programme	MSc Bioinformatics		
Unit Name	Genome Biology and Genomics	Unit Code:	BIOLM0030
Assessment name	Genome Biology and Genomics Coursework 2023		
Word Count	**1193**		
Do you give permission for you work to be used anonymously in examples given to students in the future? (type Yes or No)			

By submitting this assignment cover sheet, I confirm that I understand and agree with the following statements:

- 'I have not committed plagiarism, cheated or otherwise committed academic misconduct as defined in the University's Assessment Regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)
- 'I have not submitted this piece, in part or in its entirety, for assessment in another unit assignment (including at other institutions) as outlined in section 4 of the University's Assessment regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)
- I understand that this piece will be scrutinised by anti-plagiarism software and that I may incur penalties if I am found to have committed plagiarism, as outlined in sections 3 of the University's Examination Regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)

1a) Evaluate the quality of raw sequencing data.

The FastQC analysis confirms that the sequences are Illumina short read due to no variation in the sequence length distribution (Table 1.0) and the drop in base quality above the 210bp threshold (Figure 1.0)

Filename	coursework2023_R1.fastq	coursework2023_R1_trimmed.fastq.gz
File type	**Conventional base calls**	**Conventional base calls**
Encoding	Sanger / Illumina 1.9	Sanger / Illumina 1.9
Total Sequences	200001	199906
Sequences flagged as poor quality	0	0
Sequence length	301	18-301
%GC	41	41

Table 1.0 – Fastqc basic statistics on raw and trimmed sequencing data

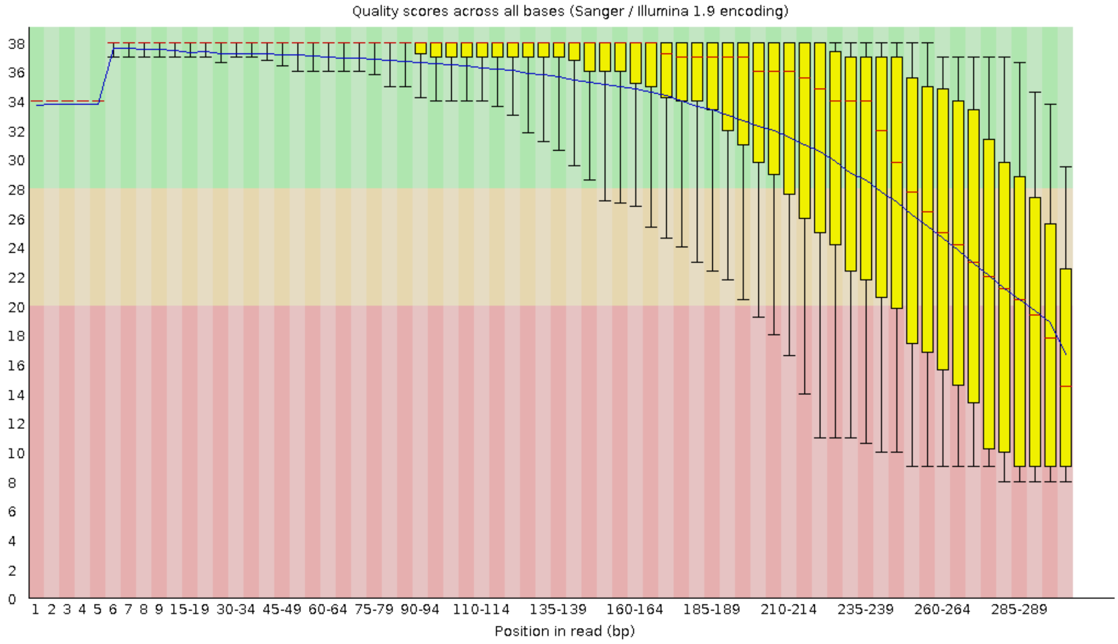

Figure 1.0 – FastQC per base sequence quality. coursework2023_R1_fastqc.zip

There is a high PHRED score for the untrimmed sequence files with 90,000 being above 32 indicating 99.9% base call accuracy (Figure 2.0).

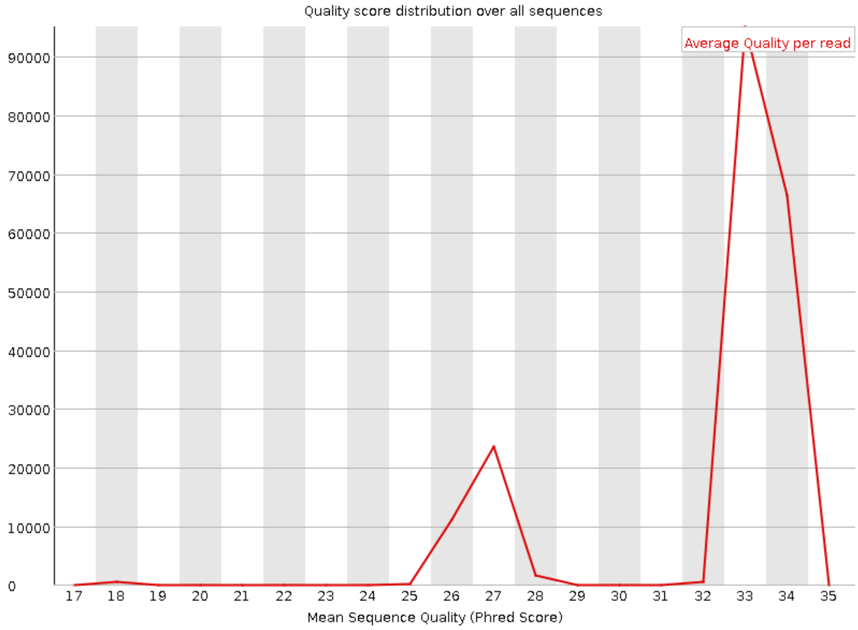

Figure 2.0 – Per sequence quality score of raw sequences.

1b) Trimming and assembly contig statistics

Statistics without reference	final.contigs
# contigs	606
# contigs (>= 1000 bp)	305
# contigs (>= 50000 bp)	38
Largest contig	291886
Total length	6226606
Total length (>= 1000 bp)	6020893
Total length (>= 50000 bp)	4638750
N50	115654

Table 2.0 – Quast contig analysis of the MEGAHIT assembly.

The total contig length of this assembly is 6,226,606. The N50 value is 115,654 bp which is the smallest contig that covers half of the assembly, providing a partial indication of contiguous assembly. Mean coverage x15 was calculated from a bam file by read mapping back to the assembled contigs (Supplementary 5).

Table 3 – Quast reference genome statistics

There is a high fraction of alignment in this assembly with *C.P.syntrophicum*.

Genome statistics	final.contigs
Genome fraction (%)	99.189
Duplication ratio	1.002
Largest alignment	291886
Total aligned length	4402517
NGA50	145153

BUSCO completeness is at 90% indicating a good level of gene content, which provides an indication of completeness alongside to mean coverage.

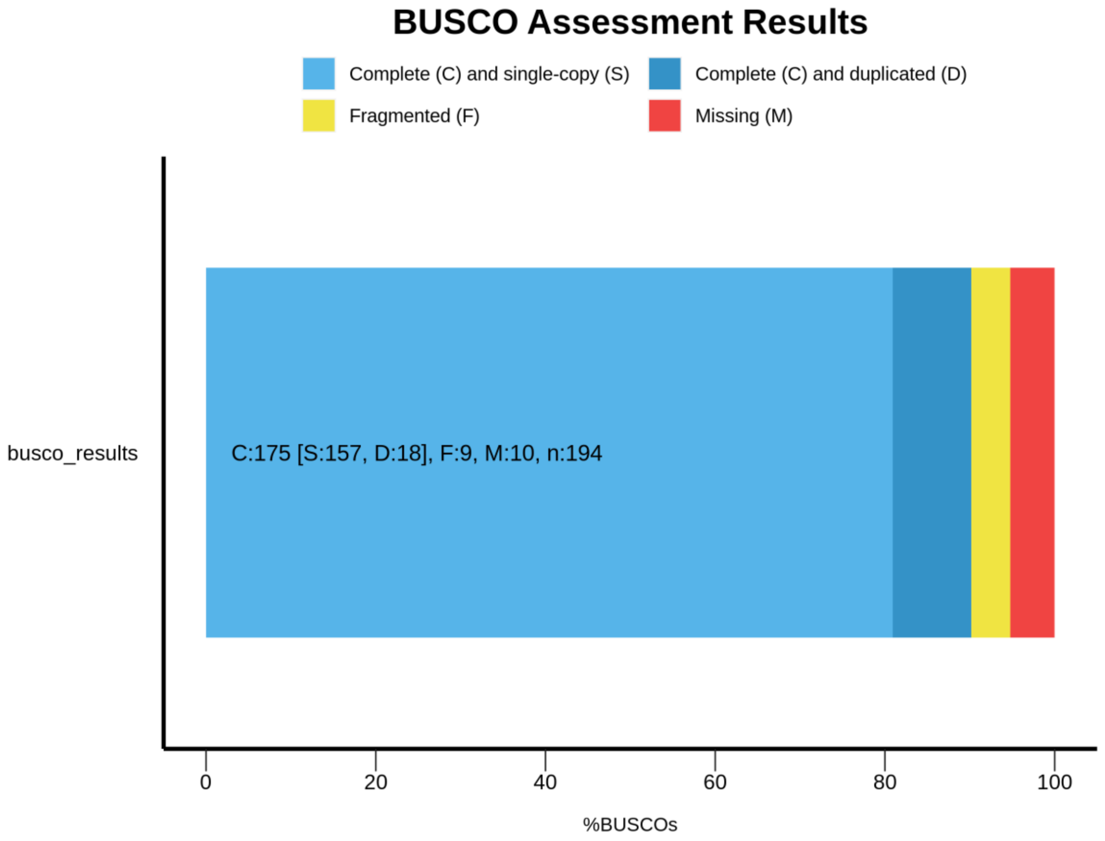

Figure 2.1 – BUSCO Completeness of Megahit assembly

C) Methods of analysis (295 / 300 words)

Data Summary

Data was obtained from an online repository hosted by the University of Bristol. The pool of genetic material is likely to be prokaryotic DNA sequences generated by Illumina paired short reads.

File	**Description**
coursework2023_R1.fastq	FASTQ file with assumed prokaryotic DNA sequences.
coursework2023_R2.fastq	FASTQ file with assumed prokaryotic DNA sequences.
GCF_008000775.1_genomic.fna.gz	Genome of Candidatus Prometheoarchaeum syntrophicum

Table 3.1 – Data sources

Quality checking with FASTQC

FASTQC was used to assess the quality of the two fastq files (Andrews, 2010), providing information on sequence length distribution and phred scores. FASTQC was used before and after trimming.

Pre-processing of sequences

Trimmomatic (Bolger et al., 2014) was used to remove adaptor sequences and low quality bases using a newly created bash script looping through files and renaming outputs (Supplementary 1). Custom Biopython scripts were used to filter and analyse the fasta/fastqc files (Supplementary 2).

De-novo genome assembly using MEGAHIT

The software MEGAHIT was used for a fast parallel assembly utilizing de bruijn graphs, to create a de novo genome assembly and contig outputs (Li et al., 2015).

Similarity searches

Blast was used to search sequences to determine the type of organisms in the sample (Altschul et al., 1990). This helped with defining the BUSCO lineage parameters below and species identification.

Assessing genome assembly quality and completeness

Quast was used to provide basic assembly statistics such as the number of contigs and N50 against a reference (Gurevich et al., 2013). The default parameters of contigs <500bp were removed from the results. BWA was used to align trimmed reads back to the assembly. The samtools depth function generated mean coverage of the aligned assembly. BUSCO was used to provide a quantitative estimation of genome assembly completeness via the expected gene content of the prokaryote assembly. All the Prokka and Ghost Koala results were collated in excel (Kanehisa et al., 2016), outputs were cleaned and joined in R for grouping sequences by taxonomy (Supplementary 3).

2) How many organisms were sequenced in this sample? (198 /200 words, 10%)

The blastn resulted in archaeal identification justifying using Prokka to annotate the assembly and predict microbial genes (Seemann, 2014). Prokka amino acid outputs were used for further searches detecting the species *Candidatus Prometheoarchaeum syntrophicum* and *Methanogenium cariaci* (Table 4*)*.

Table 4 – Summary of blastP results

Similarity searches from Prokka protein prediction.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Acc. Len	Accession
PAS domain-containing protein	Candidatus Prometheoarchaeum syntrophicum	2076	2076	100%	0	100	1023	WP_147661239.1
ATP-dependent protease LonB	Methanogenium cariaci	1275	1275	100%	0	100	631	WP_062399716.1
ATP-dependent protease LonB	Methanogenium marinum	1172	1172	99%	0	93.95	675	WP_274924851.1
hypothetical protein	Candidatus Prometheoarchaeum syntrophicum	729	729	100%	0	100	362	WP_147662554.1
hypothetical protein	Candidatus Prometheoarchaeum syntrophicum	1503	1503	100%	0	100	751	WP_147663367.1

Ghost Koala provided an improved taxonomic estimation, confirming a sample with an Archaeal species richness in the proposed phyla Lokiarchaeota and Euryarchaeota. The genera Candidatus Prometheoarchaeum (N = 3941) and Methanogenium (N = 1563) were the highest proportion of species detected. Genera such as Methanofollis and Methanolacinia were observed in small quantities also from Euryarchaeota. There are small quantities of the bacterial genus Actinomycetota (1.88%) and species Streptomyces (16/6489) present but the low quantities are inconclusive.

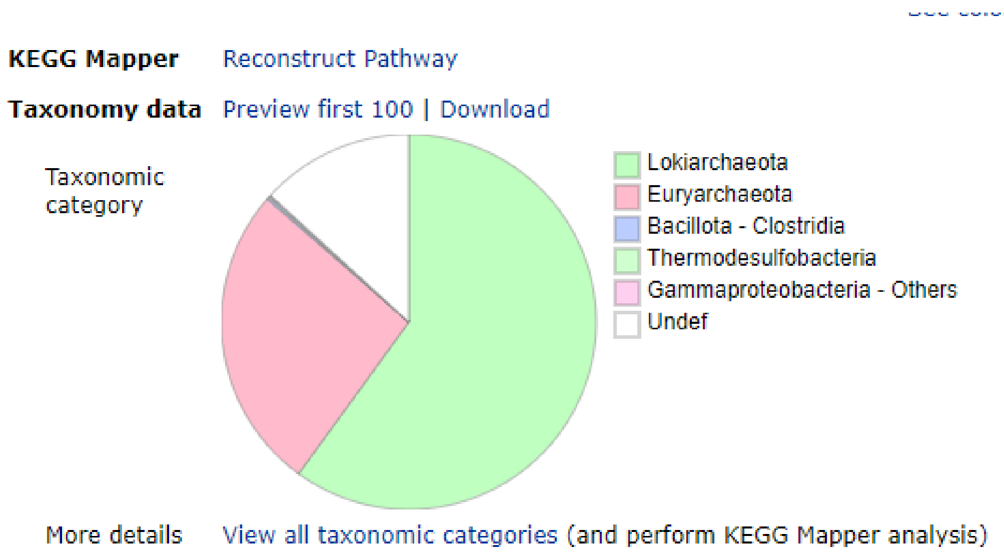

Figure 3 – Taxonomic analysis of sequence data

Table 5- Subset of ghost Koala taxonomic results.

Taxonomy	Ghost count	Proportion
Archaea	**5944**	**91.6%**
Lokiarchaeota	**3945**	60.8%
Candidatus Prometheoarchaeum	3941	60.7%
Candidatus Lokiarchaeum	4	0.1%
Euryarchaeota	**1990**	**30.7%**
Methanogenium	1563	24.1%
Methanofollis	73	1.1%
Methanolacinia	72	1.1%
Methanoplanus	70	1.1%
Bacteria	**545**	**8.40%**
Actinomycetota	**122**	**1.88%**
Total	**6489**	

To conclude, this is a metagenomic sample with the two primary organisms estimated to be *Candidatus Prometheoarchaeum syntrophicum* and *Methanogenium cariaci*. However, there are four smaller quantities of organisms in different phyla such as Euyarchaeota and Actinomycetota. The co-culture may be contaminated with bacterial strains but it is not possible to determine the species from the data provided.

3) Propose and justify, with evidence from your analyses, a hypothesis for the core energy metabolism of each of the predominant community members (280 Up to 300 words, 30%).

A hypothesis is that *Candidatus Prometheoarchaeum syntrophicum* (CP-S1) has a syntrophic relationship with *Methanogenium,* utilizing the latter's amino acid and methane production as a core metabolite for growth. GhostKoala and Prokka results infer the metabolisms of these prominent community members (Figure 3). The Asgard group has interesting eukaryotic protein coding regions with a range of physiological properties. Studies have shown that CP-S1 is largely anaerobic and undergoes syntrophic amino acid utilization with its co-culture partner *Methanogenium.* It has been demonstrated that it produces both formate and hydrogen from methane and CO2 substrates depending on the type of partner (Imachi et al., 2020). Identification in the analysis of genes encoding for enzymes such as formate dehydrogenase corresponds with this paper (Table 5.1). Interestingly, CP-S1 can grow syntrophically with methane producing bacteria when replaced in vitro further supporting its dependence on other microbes (Imachi et al., 2011). Additionally, CP-S1 is likely to switch between syntrophic oxidation and hydrolysis of the amino acid intermediates such as 2-oxoacid.

Methanogenium is a strictly anaerobic methanogen which uses substrates such as CO2 and hydrogen as substrates to produce methane. The Prokka and Ghost koala results have provided functional gene annotations for a group of enzymes called methyl-coenzyme M reductases (MCRs) in *Methanogenium* (Table 5.1). MCRs are central to anaerobic methane metabolism, providing the final catalysation step in methanogenesis and the first step in the anaerobic oxidation of methane. The enzymes also exhibit novel post-translational modifications assumed to be important in metabolic enzyme function (Chen et al., 2020). These findings have required the evolution of methanogenesis to be revisited and re-examined.

Table 5.1 – Structural, functional and taxonomic links by Prokka and GhostKoala.

Prokka locus_tag	ftype	genus	COG	KEGG_annotation
BJMNOHND_00887	CDS	Methanogenium	COG4058	mcrA; methyl-coenzyme M reductase alpha subunit [EC:2.8.4.1]
BJMNOHND_04214	CDS	Candidatus Prometheoarchaeum	NA	fdhB; formate dehydrogenase (coenzyme F420) beta subunit [EC:1.17.98.3 1.8.98.6]
BJMNOHND_02154	CDS	Candidatus Prometheoarchaeum	NA	mvhA, vhuA, vhcA; F420-non-reducing hydrogenase large subunit [EC:1.12.99.- 1.8.98.5]

4) Isolation of co-culture in vitro. (223 / 300 words)

The metagenomic sample is very difficult to grow in isolation because the natural deep sea sediment environment is not easily replicated in vitro. There is significant microbial diversity within deep sea sediment ecosystems including Lokiarchaeota and Euryarchaeota phyla. The anaerobic conditions result in syntrophic amino acid utilisation and symbiotic metabolic relationships that are difficult to control experimentally. The slow growth rate and low cell yields of Lokiarchaeota are problematic and require advanced bioreactors with a continuous methane supply which is not widely accessible. Repeated sub-culturing is also required for successful enrichment over a long period of time, eventually leading to appropriate isolation of the co-culture (Imachi et al., 2020). During this process it is important to remove any competitive bacterial strains that can produce compounds that reduce the growth rate of the desired species. As demonstrated by Imachi et al. (2020) a 12 year bioreactor enrichment study was required to obtain a pure non-bacterial co-culture of the deep sea sediment targets *Methanogenium* and *Candidatus Prometheoarchaeum syntrophicum*. Within this time frame there were 7 years of in-vitro enrichment for this co-culture to be successful. During the isolation process it is important to carry out quantitative DNA analysis such as quantitative PCR (qPCR) to monitor microbial growth. Evidently, the complex and time consuming methods required to isolate this co-culture is challenging compared to other microbes.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215**,** 403-410.

Andrews, S. 2010. *Fastqc: A quality control tool for high throughput sequence data* [Online]. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc [Accessed 16/03/2023].

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

Chen, H., Gan, Q. & Fan, C. 2020. Methyl-coenzyme m reductase and its post-translational modifications. *Frontiers in Microbiology,* 11.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. 2013. Quast: Quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-1075.

Imachi, H., Aoi, K., Tasumi, E., Saito, Y., Yamanaka, Y., Saito, Y., Yamaguchi, T., Tomaru, H., Takeuchi, R. & Morono, Y. 2011. Cultivation of methanogenic community from subseafloor sediments using a continuous-flow bioreactor. *The ISME journal,* 5**,** 1913-1925.

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., Matsui, Y., Miyazaki, M., Murata, K., Saito, Y., Sakai, S., Song, C., Tasumi, E., Yamanaka, Y., Yamaguchi, T., Kamagata, Y., Tamaki, H. & Takai, K. 2020. Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature,* 577**,** 519-525.

Kanehisa, M., Sato, Y. & Morishima, K. 2016. Blastkoala and ghostkoala: Kegg tools for functional characterization of genome and metagenome sequences. *J Mol Biol,* 428**,** 726-731.

Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. 2015. Megahit: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics,* 31**,** 1674-1676.

Seemann, T. 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics,* 30**,** 2068-2069.