Whole Genome Sequencing Analysis Report

Research Use Only (RUO)

29 November, 2023

Eurofins Order ID: **NG-12345**
Sample Name: **sample1**
Sample Received on: **28 July, 2023**
Sample Analyzed on: **29 November, 2023**
Technology Used: **Oxford Nanopore Technology (ONT) Sequencing**
Pipeline: **Whole Genome Sequencing Analysis Pipeline v1.0**
Report Version: **v1.0**

Eurofins proprietary Nanopore data analysis pipeline is used to analyze samples sequenced with Oxford Nanopore Technologies sequencers, which utilize a third-generation sequencing technology capable of real-time long-read sequencing of DNA. The technology feeds a single-stranded DNA molecule through a protein nanopore and measures changes in electrical current as the DNA passes through. The resulting reads are then subjected to quality filtering, assembly, annotation, and quality checks using the Nanopore data analysis pipeline developed by Eurofins.

/ Results

/// Whole Genome Assembly

The following table shows the overall assembly statistics:

Total assembly length	No. of contigs	Largest contig	GC %	Total reads	Total bases (Mb)	Coverage depth
4630718	1	4630718	50.81	24983	114.96	24

- Total assembly length: Size of the assembled genome (all contigs) in basepairs (bp).
- No. of contigs: Number of contigs in the assembly.
- Largest contig: Size of the largest contig.
- GC %: GC content percentage of the assembled genome.
- Total reads: Number of sequence-cleaned reads used for assembly.
- Total bases (Mb): Number of sequenced-cleaned bases used for the assembly.
- Coverage depth: Average coverage depth of the assembled genome.

/// Quality Check

Assembly completeness check

Genome assembly completeness is assessed using a machine learning algorithm trained on published high quality reference genome marker gene sets as a data model. Assembly completeness is an estimated completeness based on a prediction model using neural networks, and assembly contamination is an estimated contamination based on a prediction model using gradient boost.

Assembly Completeness: 100 %
Contamination: 0.12 %

Species Determination
Based on sequence homology (>90% identity) in a compiled hash database derived from the published high quality reference genomes, the likely species is predicted from the assembled genome.

Best species match: Escherichia coli
Sequence Identity: 100 %

/// Assembly Annotation Table

The summary of the Annotations are shown in the table below:

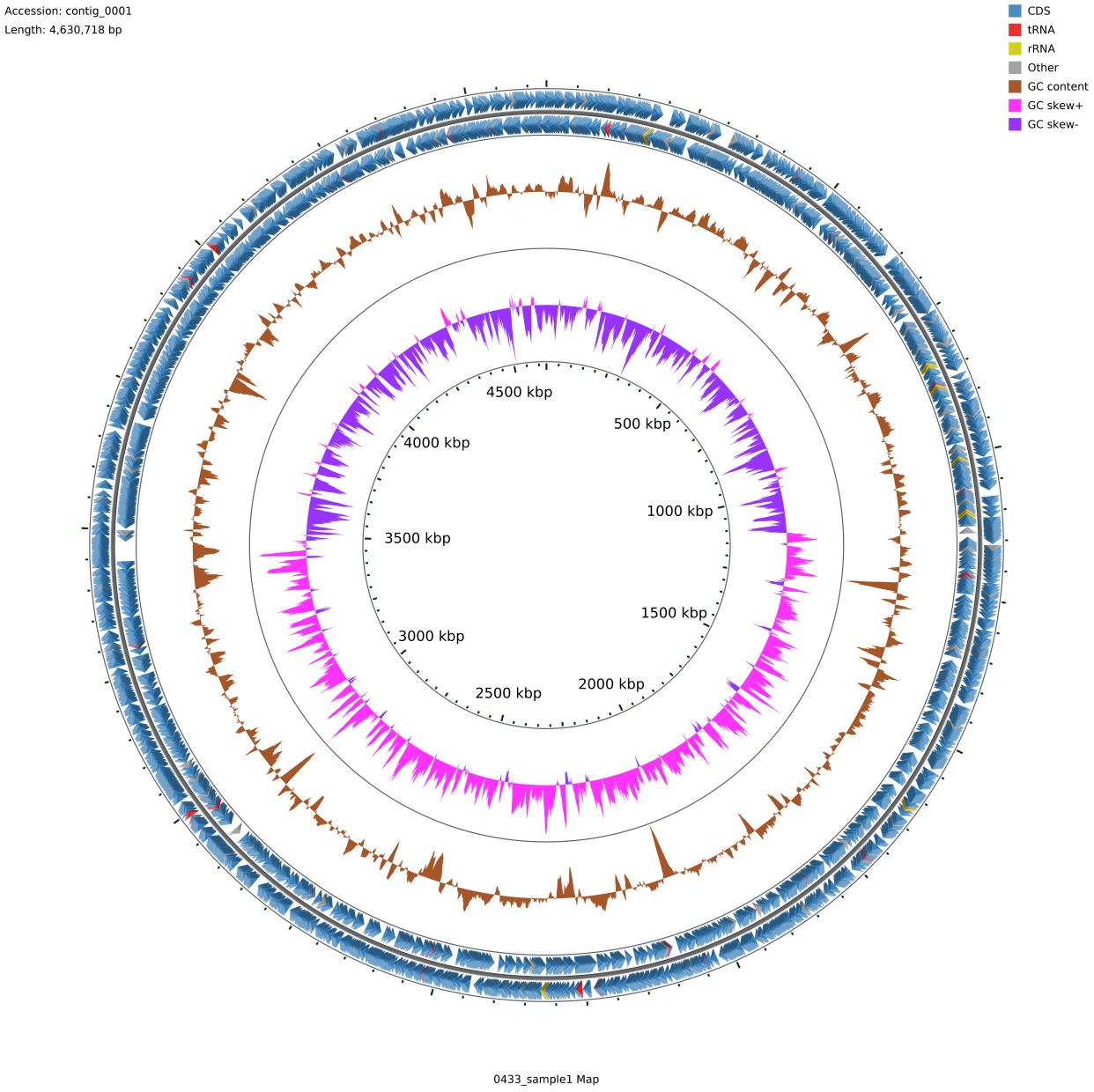
tRNAs	rRNAs	ncRNAs	CDSs	pseudogenes
88	22	230	4292	5

- tRNA: transfer RNA
- rRNA: ribosomal RNA
- ncRNA: non-coding RNA
- CDS: Coding sequences
- Pseudogenes: Non-functional coding sequences.

Details of the annotated features are listed in the table below. The full description of the features, including nucleotide sequence, can be found in the deliverables.

/// Assembly Map

A circular assembly genome map visually depicts the structure and arrangement of an assembled genome sequence, offering an overview of an organism's genetic information. It emphasizes key features such as genes, regulatory elements, and significant genomic landmarks. The outer circle denotes genomic coordinates, marked with indicators for specific DNA locations. Genes are typically annotated on the map, specifying their positions and transcriptional directions. Additionally, the map displays information about GC content and GC skew along the genome, providing insights into nucleotide composition.



0433_sample1 Map

- The two outer rings depict the coding regions on the forward and reverse strands along with the tRNA ,rRNA and others features as indicated in the legend.
- GC content: This track displays the deviation from the average of the entire genome.
- GC skew: The GC skew gives hints on a replicon's replication bubble and hence, on the completeness of the assembly.

/// Purity Check

To check the assembled genome purity, the sequenced reads are mapped against the assembled genome and the variants (SNPs, insertions & deletions) are determined. Variants detected with at least 0.3 minor allele frequency(MAF) and >30x read support are shown below

/// Deliverables

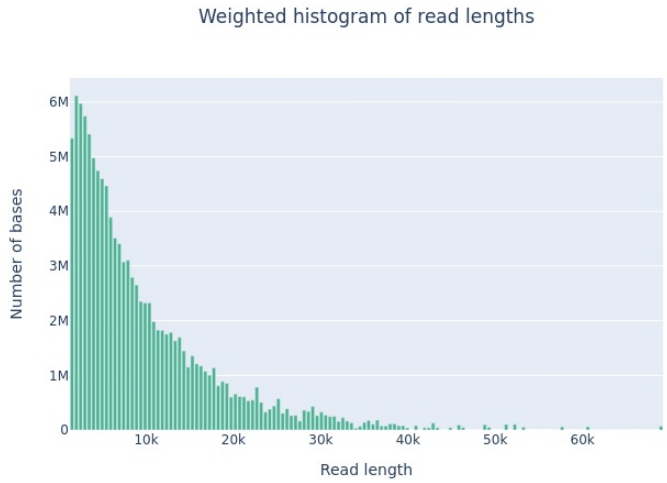
The **ORDERID.SAMPLE.WGS_analysis.zip** archive contains the following files:

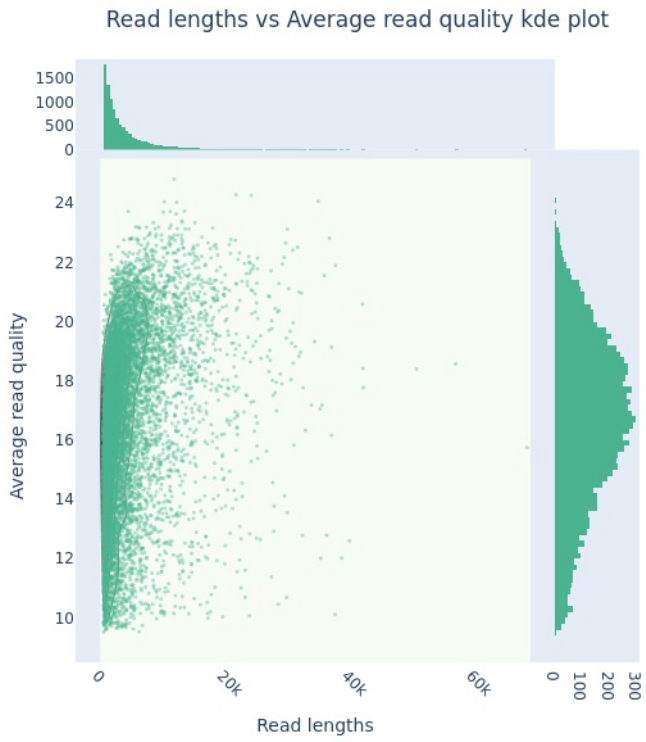
1. **SAMPLE.WGS_Analysis_Report.html**: This is the analysis report.
2. **ORDERID.SAMPLE.rawdata.fastq.gz**: Raw Sequencing data.
3. **SAMPLE.assembly.fasta**: Assembled genome in FASTA sequence.
4. **SAMPLE.circular_map.png**: Circular genome plot.
5. **Annotation/SAMPLE.assembly_annotations.gbk**: Annotated genome sequence of all contigs in GENBANK* format.
6. **Annotation/SAMPLE.assembly_annotations.gff3**: Annotated genome features in GFF3 format.
7. **Annotation/SAMPLE.assembly_annotations.csv**: Annotated feature table, including the nucleotide sequence of each feature.
8. **Annotation/SAMPLE.assembly_annotations.ffn**: Nucleotide sequences of all gene sequences in FASTA format.
9. **Annotation/SAMPLE.assembly_annotations.faa**: Amino acid sequences of all annotated protein sequences in FASTA format.

*(Use this file to import the genome sequence and features in software like SnapGene, Geneious Prime, Benchling, SeqBuilder Pro etc.)

Sequencing Reads QC

Distribution of read lengths from sequenced data is shown in the following histograms. The weighted histogram displays the number of sequenced bases (bp) on the y-axis and the read length on the x-axis. Each bar in the histogram represents a range of read lengths, and the height of the bar indicates the total number of bases (bp) falling within that range. This results in a weighted plot by the number of nucleotides per bin, as longer reads carry more weight in the histogram.
The Length vs Quality Scatter kernel density estimate (kde) plot is a bivariate plot of read length against base call quality.
Read length histograms can be used to assess the quality of sequencing data, as the distribution of read lengths can indicate the presence of contaminants or biases in the sequencing process.





REMARKS

Whole genome shotgun bacterial sequencing using nanopore technology has some limitations. One limitation is the relatively high error rate associated with long-read nanopore sequencing in comparison to short read sequencing technologies, which can lead to errors in the assembled sequence. Additionally, nanopore sequencing can be sensitive to sequencing errors, particularly in homopolymer regions, which can affect the accuracy of the sequencing data. The quality of input DNA is a very important factor that can influence the accuracy of generated sequence data. Any impurities in the DNA sample can significantly affect the accuracy of the sequencing data, which may result in failure of genome assembly reconstruction.

DISCLAIMER

The results presented and delivered are generated by following best practices available for nanopore sequencing of bacterial genomes. Before interpretation of the results, customers are advised to inspect the results thoroughly and consider the technological and bioinformatical limitations

carefully. This report and the provided deliverables are for research use only (RUO). Please see the interpretation guide https://eurofinsgenomics.eu/en/next-generation-sequencing/applications/bacterial-genome-sequencing-on-ont/bgs-data-interpretation/.