BIOINFORMATICS ANALYSIS SERVICES

GENOME ASSEMBLY AND ANNOTATION
FULL TECHNICAL REPORT

73-IFO-13448

Analysis of bacterial whole-genome sequencing (WGS) data for isolate verification

1. Introduction

1.1 Background

Whole-genome sequencing (WGS) provides comprehensive genomic information that can be used for bacterial characterisation and taxonomic identification. Accurate species identification is essential for establishing the genomic identity of bacterial isolates. In this project, the client's original identification was *Clostridium acetobutylicum*, whereas the initial bioinformatics analysis indicated *Enterobacter hormaechei*. Therefore, further genomic investigation is required to clarify the taxonomic identity of the isolate.

1.2 Objectives

The primary objective of this analysis was to determine the taxonomic identity of the bacterial isolate based on its whole-genome sequencing (WGS) data. The analysis focused on evaluating its genomic characteristics and taxonomic relationships to establish whether the data support the client's original identification as *Clostridium acetobutylicum* or indicate a different bacterial species.

1.3 Sample Information

Expected species	*Clostridium acetobutylicum*
Assembly-based species identification	***Clostridium acetobutylicum***
Sequencing platform / instrument	

2. Tools and Methods

2.1 Sequencing Quality Assessment

Raw sequencing reads were assessed using **FastQC v0.12+** to evaluate sequencing quality, Phred scores, GC content, sequence length distribution, and potential adapter contamination.

2.2 Quality Control and Read Filtering

Read quality control and filtering were performed using **fastp v0.23+**. Low-quality bases, adapter sequences, and excessively short reads were removed to improve read quality prior to genome assembly.

2.3 Reference-Based Read Alignment

Quality-controlled sequencing reads were aligned against a reference genome using **BWA-MEM v0.7.17**. The alignment algorithm employs the Burrows-Wheeler Transform (BWT) and backward search principles to efficiently map sequencing reads to corresponding genomic positions. The resulting alignments were used for downstream consensus genome generation.

2.4 Alignment Processing

Read alignment files were processed using **SAMtools v1.17+**, which was used to convert, sort, and manage sequence alignment data in SAM and BAM formats. These processed alignment files provided the input for subsequent variant analysis and consensus sequence generation.

2.5 Consensus Genome Generation

Consensus genome generation was performed using **BCFtools v1.17+** to identify nucleotide variations from aligned sequencing reads and generate a consensus sequence. Genotype likelihoods and sequencing evidence were evaluated to distinguish potential genomic variants from sequencing errors, producing a representative consensus genome for downstream analysis.

2.6 Ribosomal RNA Gene Identification

Ribosomal RNA genes were identified using **Barrnap v0.9+**, which utilises Hidden Markov Models (HMMs) to detect conserved 5S, 16S, and 23S rRNA sequences within the consensus genome. The identified 16S rRNA sequence was used to support the taxonomic characterisation of the bacterial isolate.

2.7 Data Processing and Analysis

Python 3 was used for downstream data processing, analysis, tabulation, visualisation, and bioinformatics workflow automation.

Filepath: *73-IFO-13448\reference_guided_assembly_tools.txt*

3. Results

3.1 Sequencing Data Quality

Sequencing generated high-quality paired-end reads, with **11.73 million reads** and **1.73 Gb of bases** retained after quality control and filtering. Base quality remained consistently high, with Q20 and Q30 proportions reaching 99.70% and 98.53%, respectively, while Q40 bases accounted for 41.71% of the total. GC content remained stable following filtering (44.80%), indicating consistent nucleotide composition. The duplication rate was relatively low (0.089%), and the insert size peak was observed at 263 bp. The mean read length after filtering was 147 bp for both read pairs. Overall, the sequencing results demonstrated high data quality and consistency, providing a reliable basis for subsequent reference-based alignment and downstream genomic analyses.

Table 1. *Sequencing data overview, before and after filtering.*

Parameter	Before Filtering	After Filtering
Total reads	11.760822 M	11.728848 M
Total bases	1.745282 G	1.731848 G
Q20 bases	1.738629 G (99.618817%)	1.726663 G (99.700602%)
Q30 bases	1.715779 G (98.309570%)	1.706466 G (98.534350%)
Q40 bases	723.984704 M (41.482395%)	722.396038 M (41.712428%)
GC content	44.789117%	44.801726%

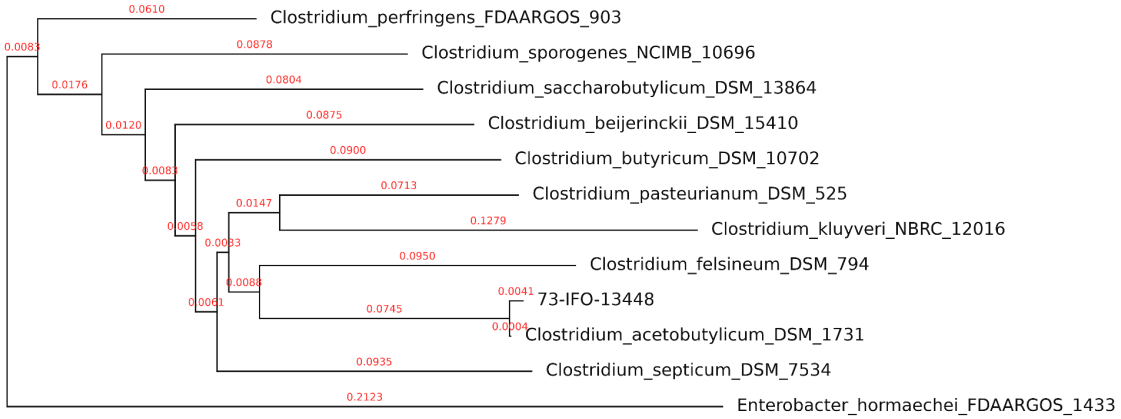
Table 2. *Sequencing platform summary.*

Parameter	Value
Sequencing strategy	Paired-end (150 cycles + 150 cycles)
Mean read length – before filtering	148bp, 148bp
Mean read length – after filtering	147bp, 147bp
Duplication rate	0.088548%
Insert size peak	263

Filepath: *73-IFO-13448\01_QC\Clostridium_fastp.html*

3.2 Phylogenomic Placement and Genome Comparison

Figure 1. *Phylogenomic placement*



The phylogenetic analysis revealed that isolate **73-IFO-13448** clustered closely with ***Clostridium acetobutylicum* DSM 1731**, indicating a very close evolutionary relationship based on the analysed sequence. In contrast, *Enterobacter hormaechei* FDAARGOS 1433 was positioned in a separate lineage. Overall, these findings indicate a closer phylogenetic relationship between isolate 73-IFO-13448 and *C. acetobutylicum* than with *E. hormaechei*.

Filepath: *73-IFO-13448\05_Comparison\tree_taxon.txt*
73-IFO-13448\05_Comparison\tree_taxon.png

Table 3. *Structural annotation summary statistics.*

Feature	Count / Value
Genome size (bp)	
• 73-IFO-13448	4145598
• *Clostridium acetobutylicum* DSM 1731 (*Reference*)	4145581
Alignment_Length (bp)	4145618
Average Nucleotide Identity (ANI)	99.56%
SNVs (bp)	17254
Insertions (bp)	15
Deletions (bp)	9

Filepath: *73-IFO-13448\05_Comparison\outputparse.csv*

3.3 Structural Annotation

Genome annotation identified **3,994 genomic features**, consisting mainly of CDSs, followed by tRNA, rRNA, and tmRNA genes. The total length of the annotated features was approximately **3.57 Mb**.

Table 4. *Structural annotation summary statistics.*

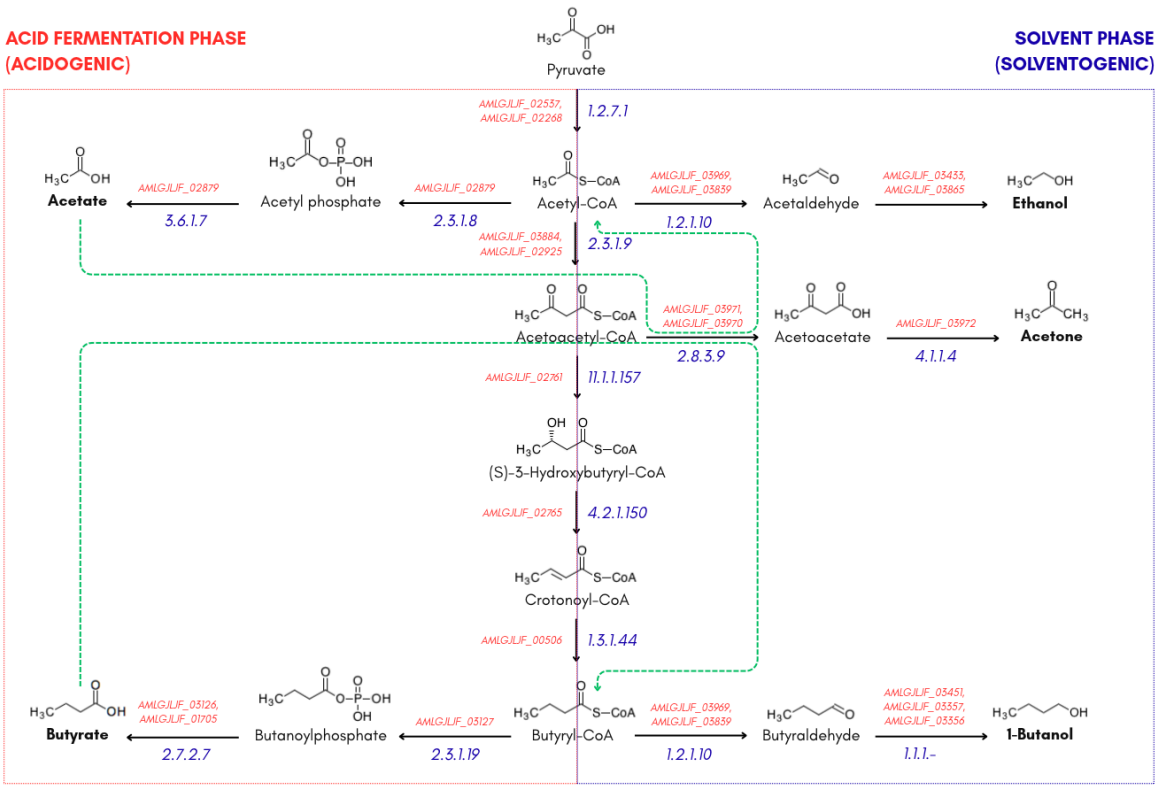
Feature	Count / Value
Total transcript length (bp)	3631753
Transcripts	3994
Total CDS length (bp)	3575913
CDS	3887
• Minimum (bp)	90
• Maximum (bp)	8613
• Mean (bp)	919.9673
• Median (bp)	789
tRNA	73
rRNA	33
tmRNA	1

Filepath: *73-IFO-13448\03_Annotation\Clostridium_prokka\Clostridium.tsv*
73-IFO-13448\03_Annotation\Clostridium_prokka\Clostridium.fna
73-IFO-13448\03_Annotation\Clostridium_prokka\Clostridium.ffn
73-IFO-13448\03_Annotation\Clostridium_prokka\Clostridium.faa

4. Insights

Clostridium acetobutylicum is widely recognised for its potential in biofuel and industrial solvent production through **Acetone–Butanol–Ethanol (ABE) fermentation**. This process involves a metabolic shift from the acidogenic phase, characterised by organic acid production, to the solventogenic phase, during which these acids are reassimilated and converted into acetone, butanol, and ethanol. Consistent with this characteristic metabolism, the genomic analysis identified several genes potentially associated with ABE fermentation, suggesting that 73-IFO-13448 may possess the genetic potential for solvent production and biofuel applications. Further functional validation is required to confirm this metabolic capability.

Figure 2. *Acetone-Butanol-Ethanol (ABE) fermentation pathway*



Filepath: *73-IFO-13448\04_Protein\Clostridium_uniprot\Acetone–Butanol–Ethanol (ABE) Fermentation.png*

Table 6. *Acetone-Butanol-Ethanol (ABE) fermentation-related genes.*

qseqid	EC Number	Protein Names
AMLGJLJF_02537, AMLGJLJF_02268	1.2.7.1	pyruvate synthase
AMLGJLJF_03884, AMLGJLJF_02925	2.3.1.9	acetyl-CoA C-acetyltransferase
AMLGJLJF_02761	1.1.1.157	3-hydroxybutyryl-CoA dehydrogenase
AMLGJLJF_02765	4.2.1.150	3-hydroxybutyryl-CoA dehydratase
AMLGJLJF_00506	1.3.1.44	trans-2-enoyl-CoA reductase (NAD+)
AMLGJLJF_03969, AMLGJLJF_03839	1.2.1.10	acetaldehyde dehydrogenase

qseqid	EC Number	Protein Names
AMLGJLJF_03451, AMLGJLJF_03357, AMLGJLJF_03356	1.1.1.-	NADH-dependent butanol dehydrogenase
AMLGJLJF_03127	2.3.1.19	phosphate butyryltransferase
AMLGJLJF_03126, AMLGJLJF_01705	2.7.2.7	butyrate kinase
AMLGJLJF_01786	2.3.1.8	phosphate acetyltransferase
AMLGJLJF_02879	3.6.1.7	acylphosphatase
AMLGJLJF_03433, AMLGJLJF_03865		alcohol dehydrogenase
AMLGJLJF_03971, AMLGJLJF_03970	2.8.3.9	butyrate-acetoacetate CoA-transferase
AMLGJLJF_03972	4.1.1.4	acetoacetate decarboxylase

Filepath: *73-IFO-13448\04_Protein\Clostridium_uniprot\UPIMAPI_results.tsv*
73-IFO-13448\04_Protein\Clostridium_uniprot\functional_result

5. Conclusion

The genomic analysis of isolate 73-IFO-13448 supports its identification as *Clostridium acetobutylicum*, rather than the initially suggested *Enterobacter hormaechei*. Phylogenetic analysis showed close clustering with *C. acetobutylicum* DSM 1731, further supported by a high average nucleotide identity (ANI) of 99.56%. Functional analysis also identified genes potentially associated with Acetone-Butanol-Ethanol (ABE) fermentation, suggesting potential for solvent production and biofuel applications. Overall, the genomic evidence strongly supports the client's original species identification as *C. acetobutylicum*.

Note: -