BIOINFORMATICS ANALYSIS SERVICES

GENOME ASSEMBLY AND ANNOTATION
FULL TECHNICAL REPORT

A6

Identification of genes potentially associated with polyethylene terephthalate (PET) degradation

1. Introduction

1.1 Background

Polyethylene terephthalate (PET) is a widely used synthetic polymer commonly found in plastic packaging, bottles, and textiles. Due to its chemical stability and resistance to natural degradation, PET can persist in the environment and contribute to plastic pollution. However, certain microorganisms possess enzymatic systems capable of breaking down PET into smaller chemical compounds through the action of PET-degrading enzymes, such as PET hydrolases and esterases. Understanding the genetic determinants underlying these processes is important for identifying the PET degradation potential of microorganisms and exploring their possible applications in plastic waste biodegradation and bioremediation.

1.2 Objectives

The primary objective of this analysis was to identify genes potentially associated with polyethylene terephthalate (PET) degradation. The analysis focused on identifying genes encoding potential PET-degrading enzymes, such as PET hydrolases, cutinases, and esterases, as well as genes involved in the metabolism of PET degradation products and their associated metabolic pathways.

1.3 Sample Information

Expected species	*Aeromonas piscicola*
Assembly-based species identification	***Paenibacillus hebaui***
Sequencing platform / instrument	MGI - BGI

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 De Novo Genome Assembly

De novo genome assembly was performed using **SPAdes v3.15+**, which employs a multi-sized de Bruijn graph algorithm to reconstruct genomic sequences from overlapping k-mers. The resulting assembly consisted of contiguous sequences (contigs) representing the assembled genome.

2.4 Assembly Quality Assessment

Assembly quality was evaluated using **QUAST v5.2+**. Key metrics, including total assembly length, number of contigs, N50, and GC content, were examined to assess assembly continuity and overall quality.

2.5 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 assembled genome. The identified 16S rRNA sequence was used to support taxonomic characterisation.

2.6 Structural Genome Annotation

Structural genome annotation was performed using **Prokka v1.14+** to identify and annotate genomic features, including protein-coding sequences (CDSs), transfer RNA (tRNA), and ribosomal RNA (rRNA). The pipeline integrates multiple tools, including Prodigal, for gene prediction and functional feature assignment.

2.7 Biosynthetic Gene Cluster Analysis

Biosynthetic gene cluster (BGC) analysis was conducted using **antiSMASH v7.0+** to identify and characterise genomic regions associated with secondary metabolite biosynthesis. The analysis provided insights into the potential metabolic capabilities of the assembled genome.

2.8 Data Processing and Analysis

Sequence processing and alignment file management were performed using **SAMtools v1.17+**, where applicable. **Python 3** was used for downstream data processing, analysis, tabulation, visualisation, and bioinformatics workflow automation.

Filepath: *A6\tools.txt*

3. Results

3.1 Sequencing Data Quality

Sequencing generated high-quality paired-end reads, with **10.03 million reads** and **1.49 Gb of bases** retained after quality control and filtering. Base quality remained consistently high, with Q20 and Q30 proportions exceeding 99% and 98%, respectively. GC content remained stable following filtering (64.02%), while the duplication rate was relatively low (0.10%). The insert size peak was observed at 265 bp. Overall, the sequencing results demonstrated good data quality and consistency, providing a reliable basis for subsequent genome assembly and downstream analyses.

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

Parameter	Before Filtering	After Filtering
Total reads	10.069252 M	10.027968 M
Total bases	1.510388 G	1.487361 G
Q20 bases	1.502084 G (99.450189%)	1.481040 G (99.574983%)
Q30 bases	1.476063 G (97.727444%)	1.458451 G (98.056294%)
Q40 bases	882.829736 M (58.450534%)	876.297299 M (58.916236%)
GC content	63.979620%	64.021130%

Table 2. *Sequencing platform summary.*

Parameter	Value
Sequencing strategy	Paired-end (150 cycles + 150 cycles)
Mean read length – before filtering	150bp, 150bp
Mean read length – after filtering	149bp, 147bp
Duplication rate	0.101279%
Insert size peak	265

Filepath: *A6\01_QC\A6_fastp.html*

3.2 Genome Assembly

Genome assembly was performed using **SPAdes**, resulting in a total assembly length of **6,293,536 bp**, comprising **276 contigs**. The assembly exhibited relatively good contiguity, indicating a less fragmented genome assembly.

Table 3. *Assembly statistics.*

Metric	Value
Assembler used	SPAdes
Total assembly length (bp)	6293536
Number of contigs	276
Largest contig (bp)	488493
N50	150448
L50	12
GC	62.84%
N's per 100 kbp	0

Filepath: *A6\01_QC\A6_quast\report.tsv*

3.3 Assembly Completeness (CheckM2)

The CheckM2 analysis indicated **100.0%** genome completeness, suggesting that the genome was fully recovered based on the assessed marker genes. However, the contamination level was 6.95%, indicating the presence of potential contaminating sequences that may require further assessment.

Table 4. *CheckM2 result.*

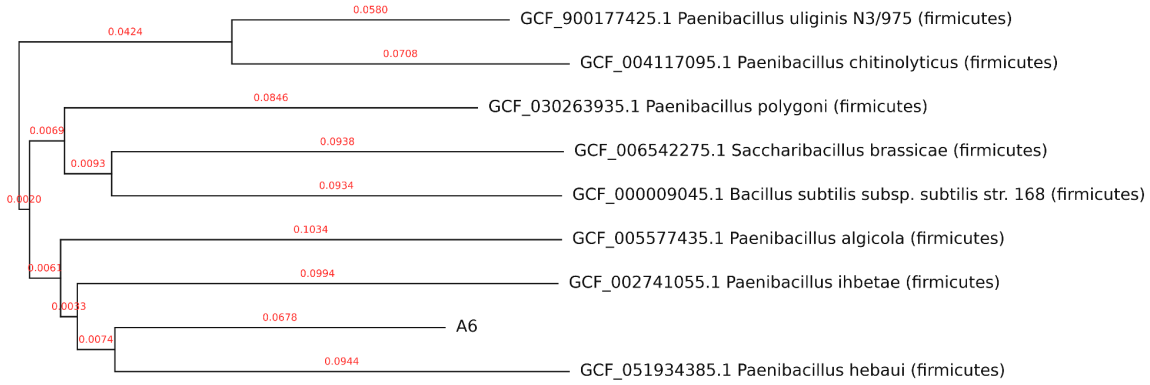
Parameter	Value
Completeness	100.0%
Contamination	6.95%
Coding density	88.6%

Filepath: *A6\02_Assembly\A6_checkm2_results\quality_report.tsv*

3.4 Phylogenomic Placement

The phylogenetic analysis revealed that **A6** clustered closely with ***Paenibacillus hebaui***, indicating a relatively close evolutionary relationship based on the analysed sequence. The isolate was positioned within the *Paenibacillus* group, alongside *P. ihbetae* and was more distantly related to other members of the Firmicutes. *Saccharibacillus brassicae* and *Bacillus subtilis* formed more distantly related lineages, highlighting the contrast phylogenetic separation.

Figure 1. *Phylogenomic placement*



Filepath: *A6\05_Comparison\tree_taxon.nwk*
A6\05_Comparison\tree_taxon.png

3.5 Structural Annotation

Genome annotation identified **5,663 genomic features**, consisting mainly of CDSs, followed by tRNA, rRNA, and tmRNA genes. The total length of the annotated features was approximately **5.41 Mb**.

Table 5. *Structural annotation summary statistics.*

Feature	Count / Value
Total transcript length (bp)	5411587
Transcripts	5663
Total CDS length (bp)	5399481
CDS	5575
• Minimum	90
• Maximum	15414
• Mean	968.5168
• Median	822
tRNA	78
rRNA	9
tmRNA	1

Filepath: *A6\03_Annotation\A6_prokka\A6.tsv*
A6\03_Annotation\A6_prokka\A6.gbk
A6\03_Annotation\A6_prokka\A6.fna
A6\03_Annotation\A6_prokka\A6.ffn
A6\03_Annotation\A6_prokka\A6.faa
RAST result: *A6\03_Annotation\rast_annotated_genome_A6.gbk*
A6\03_Annotation\Paenibacillus_A4_44249.864.xls

3.6 Functional Annotation: PET Degradation

Functional analysis revealed **no genes encoding established PET-degrading enzymes**, such as PETase or MHETase, in the genome. However, several proteins emerged as potential candidates for further investigation, including a cell-surface-associated esterase (**GOHIFCNO_01480**), a PHA depolymerase (**GOHIFCNO_03710**), and three dienelactone hydrolase family proteins (**GOHIFCNO_05075**, **GOHIFCNO_03487**, and **GOHIFCNO_02017**). Notably, recent studies have demonstrated PET-hydrolysing activity in certain dienelactone hydrolases, making these proteins particularly interesting targets for sequence comparison and functional characterisation. Nevertheless, their PET-degrading activity remains unconfirmed and requires experimental validation.

Table 6. *Potential PET degradation-related genes.*

qseqid	Matched GO Terms	Protein Name	Explanation
GOHIFCNO_01480	outer membrane [GO:0019867]; lipase activity [GO:0016298]	Autotransporter domain-containing esterase	Its predicted cell-surface localisation and esterase activity make it a plausible candidate for PET hydrolysis, as surface-associated esterases have been experimentally shown to degrade PET. However, PET activity has not been demonstrated for this specific protein. Reference: Communications Biology 2023
GOHIFCNO_03710	monoacylglycerol lipase activity [GO:0047372]	Poly(3-hydroxyalkanoate) depolymerase	PHA depolymerases contain polyester-binding domains that may support PET surface interactions. A PHA-derived binding module has enhanced PET binding by LCC, suggesting potential relevance to PET degradation screening, although direct PET hydrolysis remains unconfirmed. Reference: Biotechnology Journal 2023
GOHIFCNO_05075	hydrolase activity [GO:0016787]	Dienelactone hydrolase family protein	Recent biochemical evidence shows that certain dienelactone hydrolases (DLHs) can hydrolyse PET and its intermediates (BHET and MHET). This makes the protein a particularly interesting candidate for sequence-level comparison with these enzymes. Reference: Applied and Environmental Microbiology 2026
GOHIFCNO_03487		Dienelactone hydrolase	
GOHIFCNO_02017	hydrolase activity [GO:0016787]	Dienelactone hydrolase family protein (EC 3.1.-.-)	

Filepath: *A6\04_Protein\A6_upimapi\UPIMAPI_results.tsv*
A6\04_Protein\A6_upimapi\functional_result\Potential PET degradation-related genes.tsv
A6\04_Protein\A6_upimapi\functional_result\Sequences\Potential PET degradation-related genes.ffn
A6\04_Protein\A6_upimapi\functional_result\Sequences\Potential PET degradation-related genes.faa

4. Insights

Although the PET degradation potential of isolate A6 requires further investigation, the identification of genes potentially encoding **beta-xylanases** and **beta-glucosidases** suggests its broader potential for processing lignocellulosic agricultural residues, such as oil palm empty fruit bunches (EFB), coconut husks, and cocoa pod husks. These enzymes may contribute to the breakdown of plant biomass and the release of fermentable sugars for biotechnological applications, such as bioethanol production.

Table 7. *Arsenic (As) resistance-related genes.*

qseqid	Matched GO Terms	Protein Name
GOHIFCNO_05657	endo-1,4-beta-xylanase activity [GO:0031176]; carbohydrate catabolic process [GO:0016052]	Beta-xylanase (EC 3.2.1.8)
GOHIFCNO_05530	endo-1,4-beta-xylanase activity [GO:0031176]; xylan catabolic process [GO:0045493]	Endo-1,4-beta-xylanase (EC 3.2.1.8)
GOHIFCNO_05603; GOHIFCNO_05395	beta-glucosidase activity [GO:0008422]; carbohydrate catabolic process [GO:0016052]	Amygdalase (Cellobiase) (Gentiobiase)
GOHIFCNO_02063	beta-glucosidase activity [GO:0008422]; glucan catabolic process [GO:0009251]	Glycoside hydrolase family 5 protein
GOHIFCNO_01641	beta-glucosidase activity [GO:0008422]; glucan catabolic process [GO:0009251]	Periplasmic beta-glucosidase (EC 3.2.1.21) (Beta-D-glucoside glucohydrolase) (Cellobiase) (Gentiobiase)

Filepath: *A6\04_Protein\A6_upimapi\functional_result\Cellulose- and xylan-degrading genes.tsv*
A6\04_Protein\A6_upimapi\functional_result\Sequences\Cellulose- and xylan-degrading genes.ffn
A6\04_Protein\A6_upimapi\functional_result\Sequences\Cellulose- and xylan-degrading genes.faa

5. Conclusion

Genome analysis of isolate A6 produced a 6.29 Mb draft assembly comprising 276 contigs. Phylogenomic analysis placed the isolate close to *Paenibacillus hebaui*, rather than the expected *Aeromonas piscicola*. No established PET-degrading enzymes, such as PETase or MHETase, were identified. However, a cell-surface-associated esterase, a PHA depolymerase, and three dienelactone hydrolase family proteins were identified as potential candidates for further investigation. Additionally, predicted beta-xylanase and beta-glucosidase genes indicate potential for processing lignocellulosic agricultural residues. These findings highlight the biotechnological potential of A6, although its PET-degrading and biomass-conversion capabilities require experimental validation.

Note: *The absence of identified PET-degrading enzymes in this genome-based analysis does not necessarily rule out PET degradation activity. Such activity may still be demonstrated experimentally, as bioinformatics predictions rely on sequence similarity and previously reported patterns, potentially missing novel or uncharacterised enzymes.*