BIOINFORMATICS ANALYSIS SERVICES

GENOME ASSEMBLY AND ANNOTATION
FULL TECHNICAL REPORT

1S34

Identification of genes associated with resistance, transport, efflux, detoxification, and biotransformation of cadmium (Cd), zinc (Zn), gold (Au), lead (Pb), and iron (Fe).

1. Introduction

1.1 Background

Heavy metals are naturally occurring elements that can exert toxic effects on microbial growth and cellular functions, particularly at elevated concentrations. However, certain microorganisms possess specialized genetic mechanisms that enable them to tolerate, resist, and transform these metals through processes such as transport, efflux, sequestration, and detoxification. These mechanisms are mediated by specific genes and biological pathways that contribute to microbial survival under metal-induced stress. Understanding these genetic determinants is essential for characterising microbial heavy metal resistance and biotransformation capabilities, particularly in relation to cadmium (Cd), zinc (Zn), gold (Au), lead (Pb), and iron (Fe).

1.2 Objectives

The primary objective of this analysis was to identify and characterise genes associated with the resistance and biotransformation of cadmium (Cd), zinc (Zn), gold (Au), lead (Pb), and iron (Fe). The analysis specifically aimed to identify genes encoding metal transporters, efflux systems, detoxification mechanisms, and metal-transforming enzymes, as well as to determine their functional roles and associated biological pathways.

1.3 Sample Information

Expected species	*Vibrio alginolyticus*
Assembly-based species identification	***Lysinibacillus irui***
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: *IS34\tools.txt*

3. Results

3.1 Sequencing Data Quality

Sequencing generated high-quality paired-end reads, with **12.15 million reads** and **1.81 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, while the duplication rate was relatively low (0.23%). 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	12.222814 M	12.153604 M
Total bases	1.833422 G	1.805101 G
Q20 bases	1.825966 G (99.593319%)	1.799700 G (99.700795%)
Q30 bases	1.800560 G (98.207603%)	1.778258 G (98.512930%)
Q40 bases	605.750277 M (33.039324%)	600.319487 M (33.256835%)
GC content	36.517457%	36.464690%

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	148bp, 148bp
Duplication rate	0.230307%
Insert size peak	265

Filepath: *IS34\01_QC\IS34_fastp.html*

3.2 Genome Assembly

Genome assembly was performed using **SPAdes**, resulting in a total assembly length of **8,527,563 bp**, comprising **1,339 contigs**. The assembly exhibited relatively low contiguity, indicating a fragmented genome assembly.

Table 3. *Assembly statistics.*

Metric	Value
Assembler used	SPAdes
Total assembly length (bp)	8527563
Number of contigs	1339
Largest contig (bp)	89272
N50	8492
L50	248
GC	35.06%
N's per 100 kbp	0

Filepath: *IS34\01_QC\IS34_quast\report.tsv*

3.3 Assembly Completeness (CheckM2)

The CheckM2 analysis indicated a genome completeness of **89.06%**, suggesting relatively **high genome completeness**. However, the contamination level and coding density was relatively **high**, indicating that further assessment of genome quality may be required.

Table 4. *CheckM2 result.*

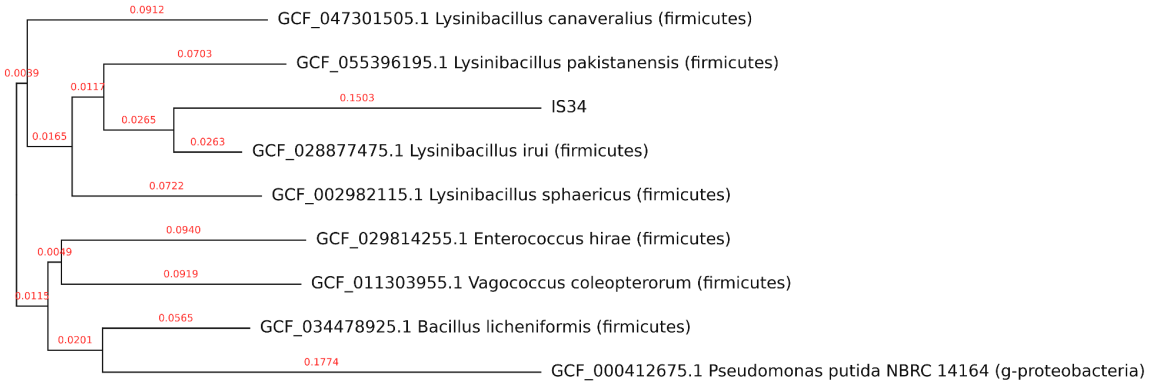
Parameter	Value
Completeness	89.06%
Contamination	43.43%
Coding density	85.30%

Filepath: *IS34\02_Assembly\IS34_checkm2_results\quality_report.tsv*

3.4 Phylogenomic Placement

The phylogenetic analysis revealed that **IS34** clustered closely with ***Lysinibacillus irui***, indicating a relatively close evolutionary relationship based on the analysed sequence. The isolate was positioned within the *Lysinibacillus* group, alongside *L. pakistanensis* and *L. sphaericus*, and was more distantly related to other members of the Firmicutes. *Pseudomonas putida*, representing the γ-proteobacteria group, formed a more distantly related lineage, highlighting the contrast phylogenetic separation

Figure 1. *Phylogenomic placement*



Filepath: *IS34\05_Comparison\tree_taxon.nwk*
IS34\05_Comparison\tree_taxon.png

3.5 Structural Annotation

Genome annotation identified **8,497 genomic features**, consisting mainly of CDSs, followed by tRNA, rRNA, and tmRNA genes. The total length of the annotated features was approximately **6.80 Mb**.

Table 5. *Structural annotation summary statistics.*

Feature	Count / Value
Total transcript length (bp)	6797059
Transcripts	8497
Total CDS length (bp)	6792825
CDS	8452
• Minimum	90
• Maximum	15054
• Mean	803.6944
• Median	699
tRNA	35
rRNA	9
tmRNA	1

Filepath: *IS34\03_Annotation\IS34_prokka\IS34.tsv*
IS34\03_Annotation\IS34_prokka\IS34.gbk
IS34\03_Annotation\IS34_prokka\IS34.fna
IS34\03_Annotation\IS34_prokka\IS34.ffn
IS34\03_Annotation\IS34_prokka\IS34.faa
RAST result: *IS34\03_Annotation\1S34.gbk*
IS34\03_Annotation\Bacillus_IS34_1386.6045.xls

3.6 Functional Annotation: Metal Resistance

Biosynthetic Gene Clusters Analysis

Biosynthetic gene cluster (BGC) analysis identified **15 potential clusters** belonging to various types, including NRPS, terpene, RiPP-like, and siderophore-associated clusters. Among these, Regions **591.1** and **733.1** showed high similarity to known petrobactin and bacillibactin biosynthetic clusters, respectively, suggesting potential roles in iron acquisition. Other regions showed low similarity to known clusters or had **no significant** matches, indicating potential diversity in the secondary metabolites produced by the strain.

Table 6.1. *Biosynthetic Gene Clusters (BGCs).*

Region	Predicted BGC Type	Most Similar Known Cluster	Similarity Level
591.1	NI-siderophore	Petrobactin	High
733.1	NRP-metallophore / NRPS	Bacillibactin	High
27.1	NRPS	Desmamide A/B/C	Low
1107.1	NI-siderophore	Petrobactin	Low
1134.1	NRPS	Bacillibactin	Low

Filepath: *IS34\03_Annotation\IS34_Antismash\index.html*

Homology-based Functional Analysis

Homology-based functional analysis identified potential resistance-related genes and pathways associated with cadmium (Cd), zinc (Zn), and lead (Pb), alongside iron (Fe) transport and homeostasis. The identified proteins were primarily associated with metal transport, efflux, sensing, and intracellular regulation, with some showing potential involvement in responses to multiple metals. However, **no genes or proteins specifically associated with gold (Au) resistance were identified**.

Table 6.2. *Cadmium (Cd) resistance-related genes.*

qseqid	Matched GO Terms / Keywords	Protein Name
FJLGGECO_08251	P-type cadmium transporter activity [GO:0008551]	Cd(2+)-exporting ATPase (EC 7.2.2.21)
FJLGGECO_07680	Cadmium resistance	Metalloregulator ArsR/SmtB family transcription factor
FJLGGECO_07253; FJLGGECO_03173	monoatomic cation transmembrane transporter activity [GO:0008324]	Cobalt-zinc-cadmium resistance protein czcD
FJLGGECO_06808	Cadmium resistance	Transcriptional regulator
FJLGGECO_03596	cadmium ion transmembrane transporter activity [GO:0015086]	Divalent metal cation transporter MntH
FJLGGECO_03397; FJLGGECO_03365	cadmium ion sensor activity [GO:0097063]; response to cadmium ion [GO:0046686]	Transcriptional regulator, ArsR family
FJLGGECO_02813	ATPase-coupled monoatomic cation transmembrane transporter activity [GO:0019829]	Cadmium-translocating P-type ATPase
FJLGGECO_02324	cadmium ion binding [GO:0046870]; stress response to cadmium ion [GO:1990170]	ClpC ATPase
FJLGGECO_00633	response to cadmium ion [GO:0046686]	Lactoylglutathione lyase (EC 4.4.1.5)

Table 6.3. *Zinc (Zn) resistance-related genes.*

qseqid	Matched GO Terms / Keywords	Protein Name
FJLGGECO_05168	response to zinc ion [GO:0010043]	High-affinity zinc uptake system membrane protein znuB
FJLGGECO_04307	response to zinc ion [GO:0010043]	Manganese transport system membrane protein MntC

Table 6.4. *Lead (Pb) resistance-related genes.*

qseqid	Matched GO Terms / Keywords	Protein Name
FJLGGECO_03397; FJLGGECO_03365	lead ion binding [GO:0032791]; response to lead ion [GO:0010288]	Transcriptional regulator, ArsR family

Table 6.5. *Iron (Fe) resistance-related genes.*

qseqid	Matched GO Terms / Keywords	Protein Name
FJLGGECO_07991; FJLGGECO_05838	cellular response to iron ion [GO:0071281]	Iron(III) dicitrate transport ATP-binding protein fecE
FJLGGECO_05776; FJLGGECO_04754	cellular response to iron ion [GO:0071281]	High-affinity heme uptake system protein IsdE (Iron-regulated surface determinant protein E) (Staphylococcal iron-regulated protein F)
FJLGGECO_05774; FJLGGECO_00864	cellular response to iron ion [GO:0071281]	Ferrichrome transport ATP-binding protein fhuC
FJLGGECO_05739; FJLGGECO_04295	intracellular iron ion homeostasis [GO:0006879]; iron ion transport [GO:0006826]	Ferritin (EC 1.16.3.2)
FJLGGECO_04307	cellular response to iron ion [GO:0071281]	Manganese transport system membrane protein MntC
FJLGGECO_03629	cellular response to iron ion [GO:0071281]	ABC transporter ATP-binding protein

Filepath: *IS34\04_Protein\IS34_upimapi\Heavy metal resistance-related genes*

4. Insights

Homology-based functional analysis further revealed that isolate IS34 possesses a substantial set of genes potentially associated with **arsenic resistance**. These genes are primarily involved in arsenate reduction and arsenite efflux through several transporter systems, including ArsB, ACR3, and arsenical pump proteins, alongside the ArsD metallochaperone and ArsR transcriptional regulator. Collectively, these findings suggest that IS34 has potential arsenic detoxification mechanisms involving the conversion of arsenate to arsenite followed by its export from the cell.

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

qseqid	Matched GO Terms / Keywords	Protein Name
FJLGGECO_08091; FJLGGECO_02064	arsenite transmembrane transporter activity [GO:0015105]	Arsenical pump family protein
FJLGGECO_08082; FJLGGECO_02424	arsenite transmembrane transporter activity [GO:0015105]	Arsenical pump membrane protein
FJLGGECO_07730	arsenite transmembrane transporter activity [GO:0015105]	ArsB/NhaD family transporter
FJLGGECO_06284	ATPase-coupled arsenite transmembrane transporter activity [GO:0015446]	arsenite-transporting ATPase (EC 7.3.2.7)
FJLGGECO_05643	ATPase-coupled arsenite transmembrane transporter activity [GO:0015446]	Arsenical pump-driving ATPase
FJLGGECO_05642	response to arsenic-containing substance [GO:0046685]	Arsenite efflux transporter metallochaperone ArsD
FJLGGECO_05641; FJLGGECO_04302; FJLGGECO_00635	arsenate reductase (thioredoxin) activity [GO:0030612]; response to arsenic-containing substance [GO:0046685]	Arsenate reductase (EC 1.20.4.4)
FJLGGECO_05640; FJLGGECO_04301	arsenite transmembrane transporter activity [GO:0015105]	ACR3 family arsenite efflux transporter
FJLGGECO_04887	arsenite transmembrane transporter activity [GO:0015105]	Sodium/bile acid symporter family protein
FJLGGECO_04877	arsenite transmembrane transporter activity [GO:0015105]	Arsenic resistance protein
FJLGGECO_04543	DNA-binding transcription factor activity [GO:0003700]	Arsenical resistance operon transcriptional regulator ArsR
FJLGGECO_02090	arsenite transmembrane transporter activity [GO:0015105]	Arsenic transporter
FJLGGECO_00634	arsenite transmembrane transporter activity [GO:0015105]; arsenite transport [GO:0015700]	Arsenical-resistance protein ACR3

Filepath: *IS34\04_Protein\IS34_upimapi\Heavy metal resistance-related genes*

5. Conclusion

Genome analysis of isolate IS34 produced an 8.53 Mb draft assembly comprising 1,339 contigs. Phylogenomic analysis placed the isolate close to *Lysinibacillus irui*, rather than the expected *Vibrio alginolyticus*. Functional analysis identified genes potentially associated with cadmium (Cd), zinc (Zn), and lead (Pb) resistance, as well as iron (Fe) transport and homeostasis. Several arsenic (As) resistance-related genes, including arsenate reductases, arsenite efflux transporters, and regulatory proteins, were also identified, suggesting potential arsenic detoxification through arsenate reduction and arsenite export. No genes specifically associated with gold (Au) resistance were detected. Overall, these findings suggest the potential of IS34 for multi-metal resistance, particularly arsenic detoxification, although the predicted functions require experimental validation.

Note: *The genome assembly may be inflated, as indicated by the high estimated contamination level (43.43%).*