**BIOINFORMATICS ANALYSIS SERVICES**

# GENOME ASSEMBLY AND ANNOTATION
# FULL TECHNICAL REPORT

## A4

*Identification of genes and metabolic pathways potentially associated with tea fermentation, metabolite production, and the transformation or bioconversion of tea-associated compounds.*

# 1. Introduction

## 1.1 Background

Tea fermentation is a complex microbial process involving the transformation of various chemical compounds naturally present in tea leaves. Microorganisms contribute to this process through diverse metabolic activities, including the production of enzymes, degradation of specific substrates, and bioconversion of tea-associated compounds. These activities may influence the formation of metabolites and other chemical constituents that contribute to the biochemical characteristics of fermented tea. Understanding the genetic determinants and metabolic pathways underlying these processes is essential for characterising the potential role of microorganisms in tea fermentation, particularly in relation to metabolite production and the transformation of tea-associated compounds.

## 1.2 Objectives

The primary objective of this analysis was to identify and characterise genes and metabolic pathways potentially involved in tea fermentation. The analysis specifically aimed to identify genes encoding enzymes associated with substrate degradation, metabolite production, and the transformation of tea-associated compounds, as well as to investigate their potential metabolic products, functional roles, and contributions to the fermentation process.

## 1.3 Sample Information

| | |
|---|---|
| Expected species | *Lysinibacillus fusiformis* |
| Assembly-based species identification | ***Bacillus subtilis*** and ***Pseudomonas entomophila*** |
| 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 Taxonomic Classification and Read Extraction

**Kraken2** was used to classify sequencing reads against a reference taxonomic database based on k-mer matching and Lowest Common Ancestor (LCA) assignment. **KrakenTools** was subsequently used to extract reads assigned to specific Taxonomic IDs, enabling taxon-specific downstream analyses, including targeted genome assembly.

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

## 2.7 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.

## 2.8 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.

## 2.9 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:*** *A4\tools_for_contaminad.txt*

# 3. Results

## 3.1 Sequencing Data Quality

Sequencing quality control (QC) revealed evidence that the **original sample was not derived from a single bacterial isolate but likely contained a mixed bacterial population**, as indicated by the recovery of two distinct genome datasets, designated **A4_B** and **A4_P**. For A4_B, **3.82 million reads** and **566.93 Mb of bases** were retained after quality control and filtering. For A4_P, **1.49 million reads** and **219.89 Mb of bases** were retained. Overall, the sequencing data demonstrated good quality and consistency based on the assessed quality parameters, despite evidence of a mixed bacterial population, enabling the reconstruction of two separate genome assemblies for subsequent genome characterisation and downstream analyses.

***Table 1.1.*** *A4_B sequencing data overview, before and after filtering.*

| Parameter | Before Filtering | After Filtering |
|---|---|---|
| Total reads | 3.838454 M | 3.818516 M |
| Total bases | 575.768100 M | 566.933101 M |
| Q20 bases | 573.287576 M (99.569180%) | 565.096950 M (99.676126%) |
| Q30 bases | 564.919279 M (98.115766%) | 557.970163 M (98.419048%) |
| Q40 bases | 229.578648 M (39.873457%) | 227.652077 M (40.155016%) |
| GC content | 43.990941% | 43.956921% |

***Table 1.2.*** *A4_B 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 | 0.093579% |
| Insert size peak | 266 |

***Filepath:*** *A4\A4_Bacillus\01_QC\A4_bacillus_fastp.html*

***Table 2.1.*** *A4_P sequencing data overview, before and after filtering.*

| Parameter | Before Filtering | After Filtering |
|---|---|---|
| Total reads | 1.493434 M | 1.486846 M |
| Total bases | 224.015100 M | 219.888782 M |
| Q20 bases | 222.774259 M (99.446090%) | 218.953248 M (99.574542%) |
| Q30 bases | 218.909421 M (97.720833%) | 215.618991 M (98.058204%) |
| Q40 bases | 131.444748 M (58.676736%) | 130.142529 M (59.185615%) |
| GC content | 64.240832% | 64.314590% |

***Table 2.2.*** *A4_P 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, 147bp |
| Duplication rate | 0.051291% |
| Insert size peak | 264 |

***Filepath:*** *A4\A4_Pseudomonas\01_QC\A4_pseudomonas_fastp.html*

## 3.2 Genome Assembly

Genome assembly was performed using SPAdes, resulting in total assembly lengths of **4,083,968 bp for A4_B** and **4,353,139 bp for A4_P**, comprising 92 contigs and 578 contigs, respectively. The **A4_B assembly exhibited relatively better contiguity**, with a larger N50 value and fewer contigs, whereas the **A4_P assembly was substantially more fragmented**, as indicated by its lower N50 value and higher contig count. Both assemblies showed no ambiguous bases (Ns), indicating complete base resolution across the assembled contigs.

***Table 3.*** *Assembly statistics.*

| Metric | A4_B | A4_P |
|---|---|---|
| Assembler used | SPAdes | SPAdes |
| Total assembly length (bp) | 4083968 | 4353139 |
| Number of contigs | 92 | 578 |
| Largest contig (bp) | 333127 | 57936 |
| N50 | 130207 | 11099 |
| L50 | 11 | 119 |
| GC | 43.49% | 64.96 |
| N's per 100 kbp | 0 | 0 |

***Filepath:*** *A4\A4_Bacillus\01_QC\A4_bacillus_quast\report.html*
*A4\A4_Pseudomonas\01_QC\A4_pseudomonas_quast\report.html*

## 3.3 Assembly Completeness (CheckM2)

The CheckM2 analysis indicated 100% genome completeness for A4_B and 91.78% for A4_P, suggesting high genome completeness for both assemblies. Contamination levels were very low (0.06% for A4_B and 0.0% for A4_P), indicating minimal evidence of contaminating sequences.

***Table 4.*** *CheckM2 result.*

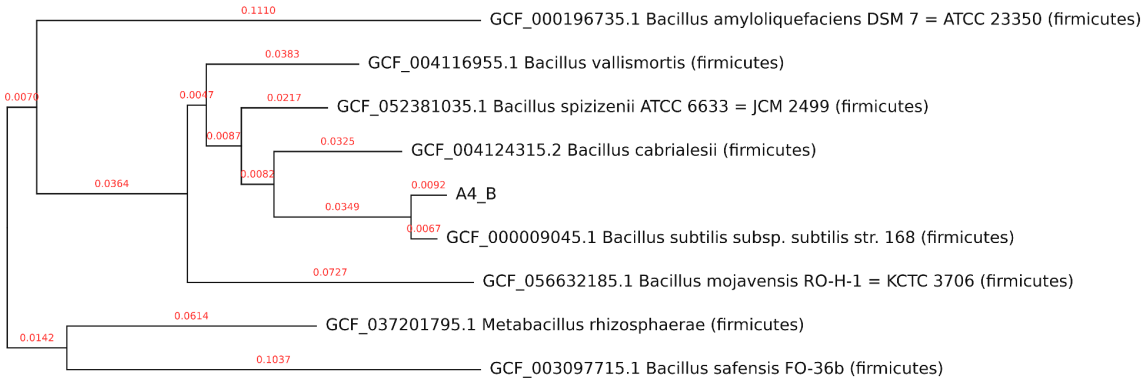
| Parameter | A4_B | A4_P |
|---|---|---|
| Completeness | 100% | 91.78% |
| Contamination | 0.06% | 0.0% |
| Coding density | 88.4% | 89.5% |

***Filepath:*** *A4\A4_Pseudomonas\02_Assembly\A4_pseudomonas_checkm2_results\quality_report.tsv*
*A4\A4_Bacillus\02_Assembly\A4_bacillus_checkm2_results\quality_report.tsv*

## 3.4 Phylogenomic Placement

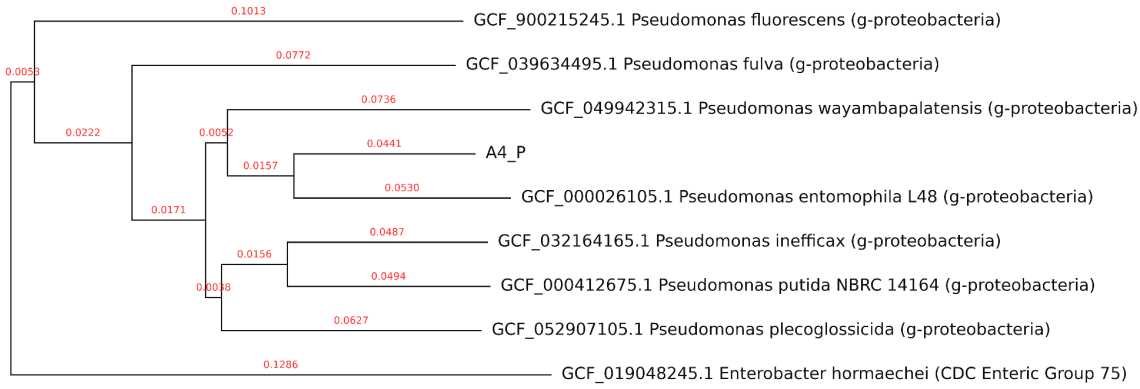
The phylogenetic analysis revealed that **A4_B** clustered closely with ***Bacillus subtilis subsp. subtilis str. 168***, indicating a relatively close evolutionary relationship based on the analysed sequences. The isolate was positioned within the *Bacillus* group, alongside *B. cabrialesii*, *B. spizizenii,* and *B. vallismortis*, while other members of the group formed more distantly related lineages. Meanwhile, **A4_P** clustered closely with ***Pseudomonas entomophila L48***, indicating a relatively close phylogenetic relationship. It was positioned within a group containing *P. wayambapalatensis*, while other *Pseudomonas* species occupied separate branches, reflecting their evolutionary relationships within the γ-proteobacteria group.

***Figure 1.1.*** *Phylogenomic placement of A4_B*




***Filepath:*** *A4\A4_Bacillus\05_Comparison\tree_taxon.nwk*
*A4\A4_Bacillus\05_Comparison\plot_tree_output\tree_taxon.png*

***Figure 1.2.*** *Phylogenomic placement of A4_P*




***Filepath:*** *A4\A4_Pseudomonas\05_Comparison\tree_taxon.nwk*
*A4\A4_Pseudomonas\05_Comparison\plot_tree_output\tree_taxon.png*

## 3.5 Structural Annotation

Genome annotation identified 4,357 transcripts in A4_B and 4,004 transcripts in A4_P, consisting predominantly of 4,300 and 3,942 CDSs, respectively, followed by tRNA, rRNA, and tmRNA genes. The total annotated CDS lengths were approximately 3.59 Mb for A4_B and 3.65 Mb for A4_P, with mean CDS lengths of 835 bp and 926 bp, respectively.

***Table 5.*** *Structural annotation summary statistics.*

| Feature | A4_B | A4_P |
|---|---|---|
| Total transcript length (bp) | 3599196 | 3655399 |
| Transcripts | 4357 | 4004 |
| Total CDS length (bp) | 3591618 | 3650028 |
| CDS | 4300 | 3942 |
| • Minimum (bp) | 90 | 90 |
| • Maximum (bp) | 10764 | 7977 |
| • Mean (bp) | 835.2600 | 925.9330 |
| • Median (bp) | 729 | 813 |
| tRNA | 49 | 59 |
| rRNA | 7 | 2 |
| tmRNA | 1 | 1 |

***Filepath:*** *A4\A4_Bacillus\03_Annotation\A4_bacillus_prokka*
*A4\A4_Pseudomonas\03_Annotation\A4_pseudomonas_prokka*
***RAST result:***
*A4\A4_Bacillus\03_Annotation\rast_annotated_genome_A4_Bacillus.gbk*
*A4\A4_Bacillus\03_Annotation\Bacillus_A4_1386.6061.xls*
*A4\A4_Pseudomonas\03_Annotation\rast_annotated_genome_A4_Pseudomonas.gbk*
*A4\A4_Pseudomonas\03_Annotation\Pseudomonas_A4_286.12654.xls*

## 3.6 Functional Annotation: Tea Flavor Development

The transformation of polyphenols and volatile compounds represents important biochemical processes contributing to flavour and aroma development during tea fermentation. Functional analysis identified genes potentially involved in these processes in both A4_B and A4_P. In A4_B, a gene encoding **laccase** was identified, suggesting potential involvement in polyphenol oxidation and transformation. Additionally, several genes associated with volatile compound metabolism were identified in both isolates, suggesting potential roles in the production and transformation of aroma-related compounds. These findings indicate that both isolates may contribute to the biochemical transformations underlying flavour and aroma development during tea fermentation, although their actual roles require further experimental validation.

***Table 6.1.*** *A4_B polyphenol transformation-related genes.*

| qseqid | Matched GO Terms | Protein Name |
|---|---|---|
| OLFBKBNO_02406 | hydroquinone:oxygen oxidoreductase activity [GO:0052716] | Laccase (EC 1.10.3.2) (Bilirubin oxidase) (EC 1.3.3.5) (Spore coat protein A) |

***Table 6.2.*** *A4_B volatile compound transformation-related genes.*

| qseqid | Matched GO Terms | Protein Name |
|---|---|---|
| OLFBKBNO_03952 | (R,R)-butanediol dehydrogenase activity [GO:0000721] | (R,R)-butanediol dehydrogenase (EC 1.1.1.4) (2,3-BD dehydrogenase) (2,3-butanediol dehydrogenase, R-alcohol forming, (R)-and (S)-acetoin-specific) |
| OLFBKBNO_03544; OLFBKBNO_00100 | carboxylesterase activity [GO:0106435] | Carboxylesterase (EC 3.1.1.1) |
| OLFBKBNO_03047 | alcohol dehydrogenase (NADP+) activity [GO:0008106] | Probable formaldehyde dehydrogenase AdhA (EC 1.1.1.-) |
| OLFBKBNO_02645 | carboxylesterase activity [GO:0106435] | Carboxylesterase YbfK (EC 3.1.1.1) |
| OLFBKBNO_01444 | formaldehyde dehydrogenase (NAD+) activity [GO:0018467] | Glutathione-independent formaldehyde dehydrogenase (FALDH) (FDH) (EC 1.2.1.46) |
| OLFBKBNO_00909 | alcohol dehydrogenase (NADP+) activity [GO:0008106]; butanol dehydrogenase (NAD+) activity [GO:1990362]; methylglyoxal reductase (NADPH) (acetol producing) activity [GO:1990002]; 1-butanol biosynthetic process [GO:0071271] | Probable NADH-dependent butanol dehydrogenase 1 (EC 1.1.1.-) |
| OLFBKBNO_00908 | alcohol dehydrogenase (NADP+) activity [GO:0008106]; butanol dehydrogenase (NAD+) activity [GO:1990362]; methylglyoxal reductase (NADPH) (acetol producing) activity [GO:1990002]; 1-butanol biosynthetic process [GO:0071271] | Probable NADH-dependent butanol dehydrogenase 2 (EC 1.1.1.-) |

***Filepath:*** *A4\A4_Bacillus\04_Protein\A4_bacillus_upimapi\UPIMAPI_results.tsv*
*A4\A4_Bacillus\04_Protein\A4_bacillus_upimapi\functional_analysis*

***Table 6.3.*** *A4_P volatile compound transformation-related genes.*

| qseqid | Matched GO Terms | Protein Name |
|---|---|---|
| LHDGGAMP_02603 | alcohol dehydrogenase (NAD+) activity [GO:0004022]; L-threonine 3-dehydrogenase activity [GO:0008743] | L-threonine dehydrogenase (EC 1.1.1.103) |
| LHDGGAMP_01894 | Formaldehyde dehydrogenase, glutathione-independent (EC 1.2.1.46) | formaldehyde dehydrogenase (NAD+) activity [GO:0018467] |

***Filepath:*** *A4\A4_Pseudomonas\04_Protein\A4_pseudo_upimapi\UPIMAPI_results.tsv*
*A4\A4_Pseudomonas\04_Protein\A4_pseudo_upimapi\functional_analysis*

## 4. Insights

Polysaccharide degradation is an important biochemical process during tea fermentation, particularly in dry fermentation processes involving diverse microbial communities, including bacteria and fungi such as *Aspergillus*. The breakdown of complex polysaccharides from tea leaves and fungal cell walls provides simpler carbohydrates that can serve as substrates for microbial metabolism and support subsequent biochemical transformations contributing to flavour development. Functional analysis identified genes potentially involved in these degradation processes in both A4_B and A4_P, suggesting that the isolates may indirectly contribute to flavour development by facilitating substrate breakdown and supporting microbial metabolic activities during tea fermentation. However, their actual contributions to these processes require further experimental validation.

***Table 7.1.*** *A4_B polysaccharide degradation-related genes.*

| qseqid | Matched GO Terms | Protein Name |
|---|---|---|
| OLFBKBNO_04341 | xylan 1,4-beta-xylosidase activity [GO:0009044] | Xylan 1,4-beta-xylosidase (EC 3.2.1.37) |
| OLFBKBNO_04231 | xylan 1,4-beta-xylosidase activity [GO:0009044] | Beta-xylosidase (EC 3.2.1.37) (1,4-beta-D-xylan xylohydrolase) (Xylan 1,4-beta-xylosidase) |
| OLFBKBNO_03608 | licheninase activity [GO:0042972] | Beta-glucanase (EC 3.2.1.73) (1,3-1,4-beta-D-glucan 4-glucanohydrolase) (Endo-beta-1,3-1,4 glucanase) (Lichenase) |
| OLFBKBNO_03501 | fructokinase activity [GO:0008865] | Putative fructokinase (EC 2.7.1.4) (Glucomannan utilization protein E) |
| OLFBKBNO_03499 | 6-phospho-beta-glucosidase activity [GO:0008706] | 6-phospho-beta-glucosidase GmuD (EC 3.2.1.86) (Aryl-phospho-beta-D-glucosidase BglD) (Glucomannan utilization protein D) |
| OLFBKBNO_03486 | oligosaccharide catabolic process [GO:0009313] | Putative sporulation-specific glycosylase YdhD (EC 3.2.-.-) |
| OLFBKBNO_02532 | alpha-amylase activity [GO:0004556]; oligosaccharide catabolic process [GO:0009313] | Trehalose-6-phosphate hydrolase (EC 3.2.1.93) (Alpha,alpha-phosphotrehalase) (Phospho-alpha-(1-1)-glucosidase) |
| OLFBKBNO_02513 | pectate lyase activity [GO:0030570] | Pectate lyase |
| OLFBKBNO_01660; OLFBKBNO_01659 | polysaccharide catabolic process [GO:0000272] | Endoglucanase (EC 3.2.1.4) |

| qseqid | Matched GO Terms | Protein Name |
|---|---|---|
| OLFBKBNO_01657; OLFBKBNO_01656 | glucuronoarabinoxylan endo-1,4-beta-xylanase activity [GO:0033940]; xylan catabolic process [GO:0045493] | Glucuronoxylanase XynC (EC 3.2.1.136) (Endoxylanase XynC) (Glucuronoxylan xylanohydrolase) |
| OLFBKBNO_01655 | alpha-L-arabinofuranosidase activity [GO:0046556] | Alpha-L-arabinofuranosidase (EC 3.2.1.55) |
| OLFBKBNO_01603 | pectate lyase activity [GO:0030570] | Polysaccharide lyase family 1 protein |
| OLFBKBNO_01173 | alpha-amylase activity [GO:0004556]; oligo-1,6-glucosidase activity [GO:0004574]; oligosaccharide catabolic process [GO:0009313] | Probable oligo-1,6-glucosidase 2 (EC 3.2.1.10) (Oligosaccharide alpha-1,6-glucosidase 2) (Sucrase-isomaltase 2) (Isomaltase 2) |
| OLFBKBNO_01139 | acetylxylan esterase activity [GO:0046555] | Cephalosporin-C deacetylase (EC 3.1.1.41) (Acetylxylan esterase) (EC 3.1.1.72) |
| OLFBKBNO_00899 | alpha-amylase activity [GO:0004556]; oligo-1,6-glucosidase activity [GO:0004574]; oligosaccharide catabolic process [GO:0009313] | Probable oligo-1,6-glucosidase 3 (EC 3.2.1.10) (Oligosaccharide alpha-1,6-glucosidase 3) (Sucrase-isomaltase 3) (Isomaltase 3) |
| OLFBKBNO_00191 | alpha-amylase activity [GO:0004556]; oligo-1,6-glucosidase activity [GO:0004574]; oligosaccharide catabolic process [GO:0009313] | Oligo-1,6-glucosidase 1 (EC 3.2.1.10) (Dextrin 6-alpha-D-glucanohydrolase) (Oligosaccharide alpha-1,6-glucosidase 1) (Sucrase-isomaltase 1) (Isomaltase 1) |

***Filepath:*** *A4\A4_Bacillus\04_Protein\A4_bacillus_upimapi\UPIMAPI_results.tsv*
*A4\A4_Bacillus\04_Protein\A4_bacillus_upimapi\functional_analysis*

***Table 7.2.*** *A4_P polysaccharide degradation-related genes.*

| qseqid | Matched GO Terms | Protein Name |
|---|---|---|
| LHDGGAMP_02346; LHDGGAMP_00802 | chitinase activity [GO:0004568]; cell wall macromolecule catabolic process [GO:0016998]; chitin catabolic process [GO:0006032] | Glycoside hydrolase family 19 protein |

***Filepath:*** *A4\A4_Pseudomonas\04_Protein\A4_pseudo_upimapi\UPIMAPI_results.tsv*
*A4\A4_Pseudomonas\04_Protein\A4_pseudo_upimapi\functional_analysis*

## 5. Conclusion

The bioinformatics analysis revealed that the original A4 sample likely contained a mixed bacterial population, represented by two distinct genome datasets phylogenetically related to *Bacillus* and *Pseudomonas*. Both assemblies showed high completeness and low contamination, providing a basis for further functional characterisation. Functional analysis identified genes potentially associated with polyphenol transformation, volatile compound metabolism, and polysaccharide degradation, suggesting that both bacterial groups may contribute to substrate breakdown and biochemical transformations during tea fermentation. These findings provide initial insights into their potential roles in flavour and aroma development, although further experimental studies are required to confirm gene functions and their actual contributions to the fermentation process.

***Note:*** *The original sample likely contained a mixed bacterial population, resulting in two distinct genome datasets (A4_B and A4_P).*