Regulatory Submission report

# Bioinformatics analysis of bacterial *Microbacterium arborensens spp. -* For EFSA submission of WGS analysis of microorganisms intentionally used in the food chain

The used bacterial strain is claimed to be *Microbacterium arborensens JCMM 5884* (DSM20754). This bacterium produces *xylose isomerase*, which is the enzyme of interest. The organism did not undergo genetic modifications and was used as obtained from the original manufacturer.

* **Session 2.1 - Microorganism and nucleic acid extraction**

BGI contractor generated the Whole Genome Sequencing of the raw reads, and generated the QC reads.

Project ID: F24A040008288\_BACfcruR

Product name: filter - Bacterial whole-genome resequencing

Sample size: 1

Library type: Short-Insert Library

Sequencing Platform: DNBSEQ

Sequencing read Length: PE150

Clean fastq phred quality score encoding: Phred+33

**Experimental Procedure**: The library construction method and sequencing process are carried out according to the following steps: DNBSEQ Short-read library preparation： 1. DNA fragmentation 2. Size selection 3. End repair and "A" tailing 4. Adapter ligation 5. Product size selection 6. PCR reaction 7. Library QC.

* **Session 2.2 - Sequencing and data quality control**

After sequencing, the raw reads were filtered. Data filtering includes removing adapter sequences, contamination and low-quality reads from raw reads.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sample Name | **Total Reads** | **Total Bases generated** | **Read Length** | **Q20(%)** | **Q30(%)** | **GC(%)** |
| 树状微杆菌A | 8,136,166 | 2,440,849,800 | PE150 | 93.78 | 82.46 | 62.35 |

Table 1: Statistics of the clean raw reads data.

The quality of data was examined after filtering. The distribution of base percentage and qualities along reads after data filtering. In the figure 1 (left), x-axis represents base position along reads, y-axis represents base percentage at the position; each color represents a type of nucleotide. Under normal conditions, the sample does not have AT/GC separation. It is normal to see fluctuations in the first several bp positions, which is caused by random primer and the instability of enzyme-substrate binding at the beginning of the sequencing reaction. In the right figure, x-axis represents base position along reads, y-axis represents base quality; each dot represents the base quality of the corresponding position along reads, color intensity reflects the number of nucleotides, a more intense color along a quality value indicates a higher proportion of this quality in the sequencing data.

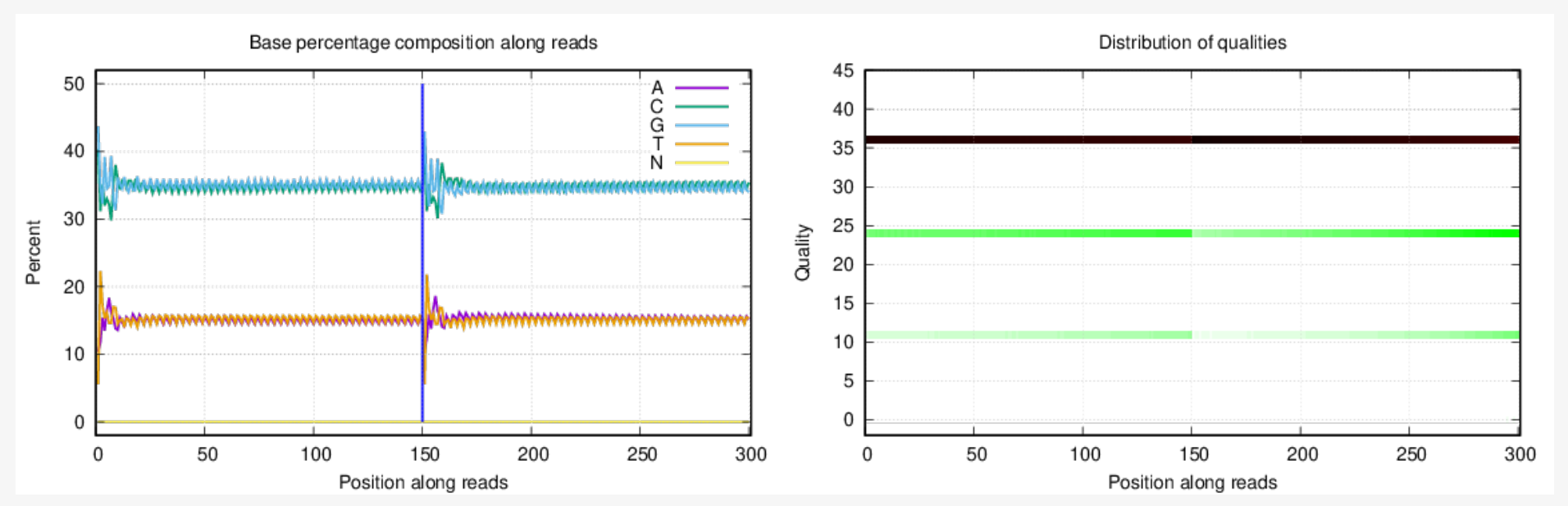


Figure 1: Distribution of base percentage and qualities of sample 树状微杆菌A.

**Bioinformatic Analysis Workflow of the sequenced reads:**

**Parameters for Data Filtering:** Raw data with adapter sequences or low-quality sequences was filtered. We first went through a series of data processing to remove contamination and obtain valid data. This step was completed by SOAPnuke (1) developed by BGI. SOAPnuke software filter parameters: "-n 0.001 -l 20 -q 0.4 --adaMis 3 --rmdup --minReadLen 150", steps of filtering:

1. Filter adapter: if the sequencing read matches 50.0% or more of the adapter sequence (maximum 3 base mismatches are allowed), remove the entire read;
2. Filter read length: if the length of the sequencing read is less than 150 bp, discard the entire read;
3. Remove N: if the N content in the sequencing read accounts for 0.1% or more of the entire read, discard the entire read;
4. Filter low-quality data: if the bases with a quality value of less than 20 in the sequencing read account for 40.0% or more of the entire read, discard the entire read;
5. Obtain Clean reads: the output read quality value system is set to Phred+33.

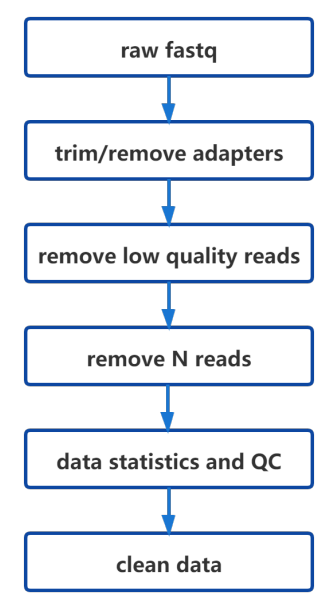


Figure 2: Bioinformatic analysis workflow for QC of the sample reads.

* **2.3 - De novo assembly and annotation**

The assembler software used was SPAdes (version: 3.15.3) (2); detailed parameters used:

**spades.py -1 reads\_R1.fq.gz -2 reads\_R2.fq.gz -o spades\_output –careful -t 8 -m 16**

The SPAdes assembler program automatically detects the preferred k-mer length for the assembly based on the read length, if not specified by the user. SPAdes assembler used 21, 33, 55, 77 k-mers for the assembly. SPAdes automatically polishes the reads based on read quality scores, and map them back to the final assembly to generate an improved per contig average coverage and correct assembly errors (-careful mode on).

After assembly, post-assembly processing was carried out as following: Mapping of the raw reads onto the generated contigs to determine the correctness of the assembly and discard contigs that has coverage lower than on average 24.23 bp and shorter than 5,256 base pairs. The final polished assembly results are presented on Table 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | N50 | N50 Index | Number of Contigs | Average Length |
| Assembly\_Microbacterium\_arborescens\_spp.fasta | 676,389 | 2 | 9 | 364,314 |

|  |  |  |
| --- | --- | --- |
| Contig | Length | Coverage |
| Contig\_1 | 1,475,171 | 354.83 |
| Contig\_2 | 676,389 | 354.80 |
| Contig\_3 | 491,278 | 352.00 |
| Contig\_4 | 275,063 | 365.95 |
| Contig\_5 | 178,176 | 341.84 |
| Contig\_6 | 134,275 | 361.18 |
| Contig\_7 | 24,482 | 367.29 |
| Contig\_8 | 18,736 | 24.23 |
| Contig\_9 | 5,256 | 1,097.37 |

Table 2: Whole Genome Assembly statistics of the complete assembly generated by SPAdes, and per contig detailed information for length and average coverage of reads.

**Comparison of the SPAdes assembly obtained against published assemblies for *M.* *arborescens* and genome size expectation for the species*:***

The expected genome size of *Microbacterium arborescens* strain DSM 20754, based on the whole genome assembly (WGA) of a previously published genome assembly (NCBI accession GCF\_003339645.1) is 3.4 Mb. However, this assembly, published in 2018, was classified as low-quality by NCBI due to contamination and is therefore not considered representative of the type strain. The assembly consisted of 125 contigs, with an N50 of 1 Mb, a GC content of 69%, and a genome coverage of 440x. Due to the identified contamination issues, this assembly was not used as a reference in the current study.

The total length of the whole genome assembly (WGA) obtained for this sample *Microbacterium arborescens* strain spp. is a total of 3,278,826 base pairs. This is in close agreement with the previously published genome assembly for the same strain (GCF\_003339645.1). The difference between the two values is approximately 121,174 base pairs, which represents a 3.6% difference. This falls within the acceptable ±20% range of expected genome size variation for the species, suggesting that the assembly is consistent with the expected genome size.

To assess the genomic differences between strains, the sample reads were mapped to the full chromosome genome assembly of reference strain *M.* *arborescens* SSF12 (RefSeq NCBI accession CP128474), which is assembled to chromosome level and curated for contaminants by NCBI RefSeq. The sample reads mapping, and the sequencing depth was computed with the software Geneious Prime version 2024.0.5 (3). The following parameters were used for mapping:

**Medium sensitivity/Fast; Map Multiple best matches Randomly; Only map paired reads which map nearby; Allow Gaps Maximum per Read 5%; Maximum Gap Size 10; Minimum Overlap 90; Minimum Overlap Identity 95%; Word Length 50; Index Word Length 13; Maximum Mismatches Per Read 5%**

The mapping analysis revealed several large gaps (Figure 3), or regions of zero mapped reads, indicating significant genomic differences between the DSM 20754 and SSF12 strains.

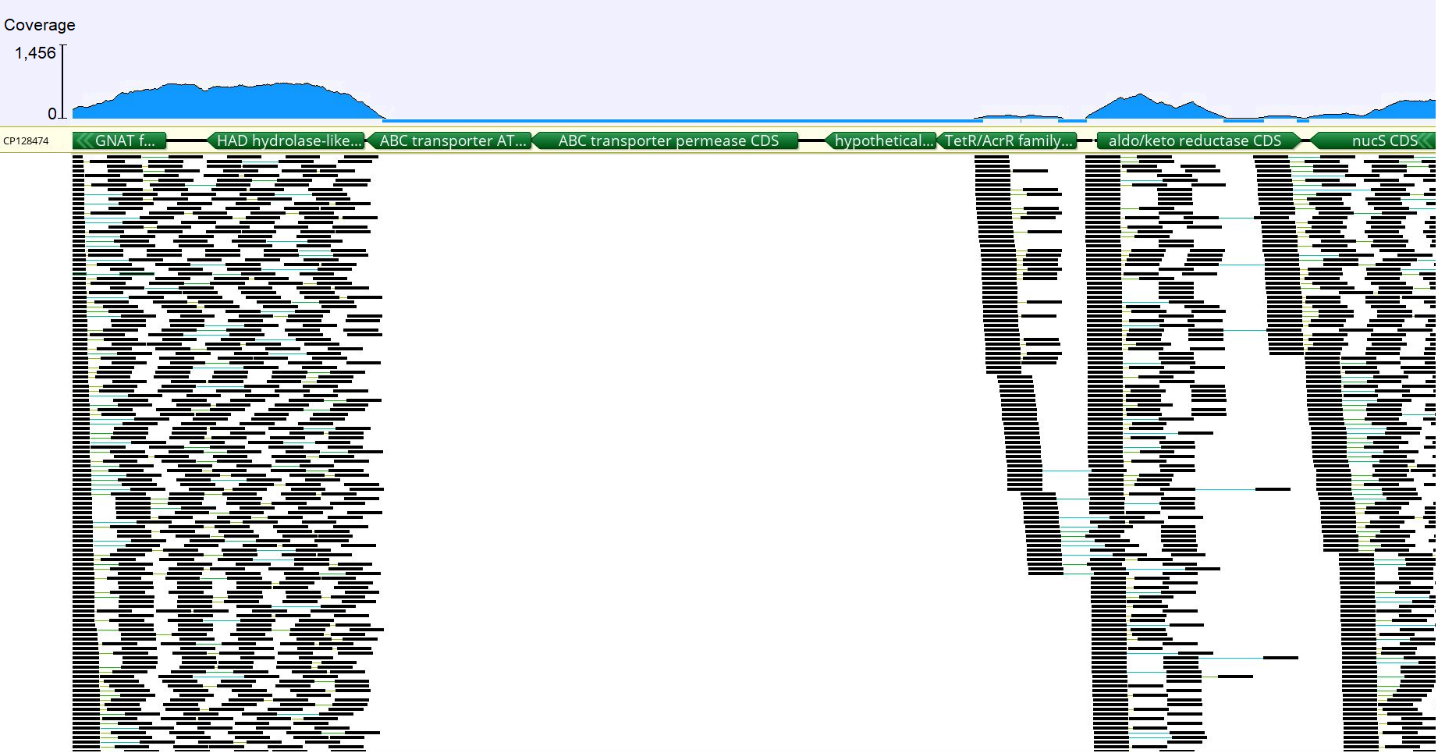


Figure 3. Genome coverage graph of the mapping of the sample reads to *M.* *arborescens* strain SSF12 showing a zoomed region of an example of a large coverage gap. The genome sequence is shown schematically below the per-base coverage, indicated in light blue. The genome annotation obtained from NCBI is shown in the middle panel part, with gene annotations in dark green. When zero reads mapped to the genome, a light blue line is plotted instead, below the zero reads mark.

**Conserved genes analysis:**

The BUSCO (Benchmarking Universal Single-Copy Orthologs) (4) analysis was performed to assess the completeness of the genome assembly. BUSCO version 5.8.0 was used with the **'micrococcales\_odb10' lineage dataset** (Creation date: 2024-01-08), containing 234 genomes and 537 BUSCOs. The analysis was conducted in **'prok\_genome\_min' mode** with the gene predictor Prodigal. Dependencies used include **hmmsearch** (version 3.4), **miniprot\_index** (version 0.13-r248), and **miniprot\_align** (version 0.13-r248). Python version used was 3.10.15. The analysis was executed in Galaxy version 5.8.0, an online platform for bioinformatics tools (5).

The BUSCO analysis results are summarized in the table 3. The gene annotation GFF file generated by BUSCO is available in the submission folder. The analysis found a total of 537 BUSCO groups, with 514 complete and single-copy BUSCOs, and 23 missing BUSCOs. The assembly was highly complete, with no fragmented BUSCOs detected.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistic** | **Complete (C)** | **Complete and Single-Copy (S)** | **Complete and Duplicated (D)** | **Fragmented (F)** | **Missing (M)** |
| **BUSCOs** | **95.7%** | **95.7%** | **0.0%** | **0.0%** | **4.3%** |

Table 3: BUSCO Results Summary.

The assembly consists of 9 scaffolds and 9 contigs, with a total length of 3,278,826 base pairs. The scaffold N50 is 676 KB, and the contig N50 is also 676 KB. The assembly has 0% gaps.

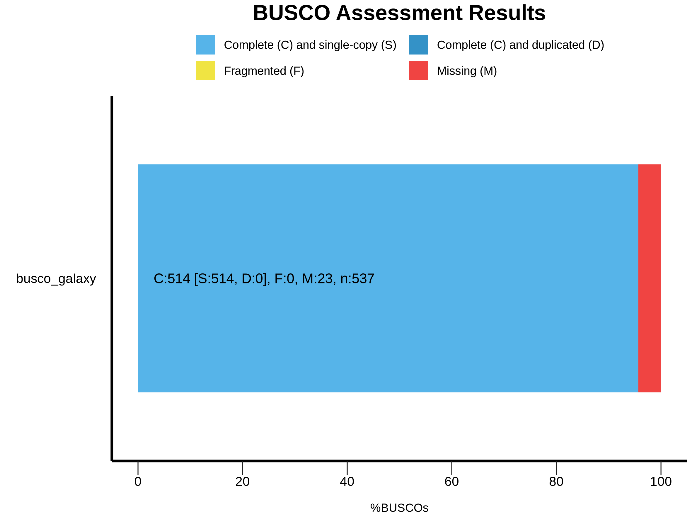


Figure 5: BUSCO assessment results with 95.7% overall gene completeness for the sample SPAdes WGA.

**Gene prediction and annotation:**

For the annotation of the identified coding sequences (CDS) gene prediction and annotation were performed using Bakta (version 1.9.4) (6). Coding sequence (CDS) prediction was conducted with the option **Prodigal**, utilizing the training file specified by Bakta. The AMRFinderPlus database (V3.12-2024-05-02.2) was used for antimicrobial resistance and virulence factor identification. The translation table applied was 11 (**Bacterial, Archaeal, and Plant Plastid Code**). A total of 1,436 CDS were identified in the assembly. Additionally, the predicted open reading frame (ORF) encoding the *xylose isomerase* gene (1,188 nucleotides, 395 amino acids) in **contig\_3 (NODE\_3\_length\_491278\_cov\_352.009914)** was manually curated and validated using the genome viewer tool in Geneious Prime (version 2024.0.5) (3).

**2.5 - Use of whole genome sequence-based data for the characterization of the microorganism**

**2.5.1 - Identification of the microorganism:** The assembled 16s rRNA sequence of the sample strain was compared to the published 16s rRNA sequence of the parental strain *Microbacterium arborescens* DSM 20754 (NCBI accession NR\_029265). All multiple sequence alignments were generated with MAFFT aligner version v7.45 (7). Parameters used for all the alignment were:

**Algorithm L-INS-i Scoring\_matrix 200PAM k=2 Gap\_open\_penalty 1.53 Offset\_value 0.123 Automatically\_determine\_sequence\_direction yes**

The result of the sequence alignment of the sequence obtained for this sample and the published 16S rRNA is shown on figure 4. The percent of identity with the compared reference genome obtained was 99.9%.



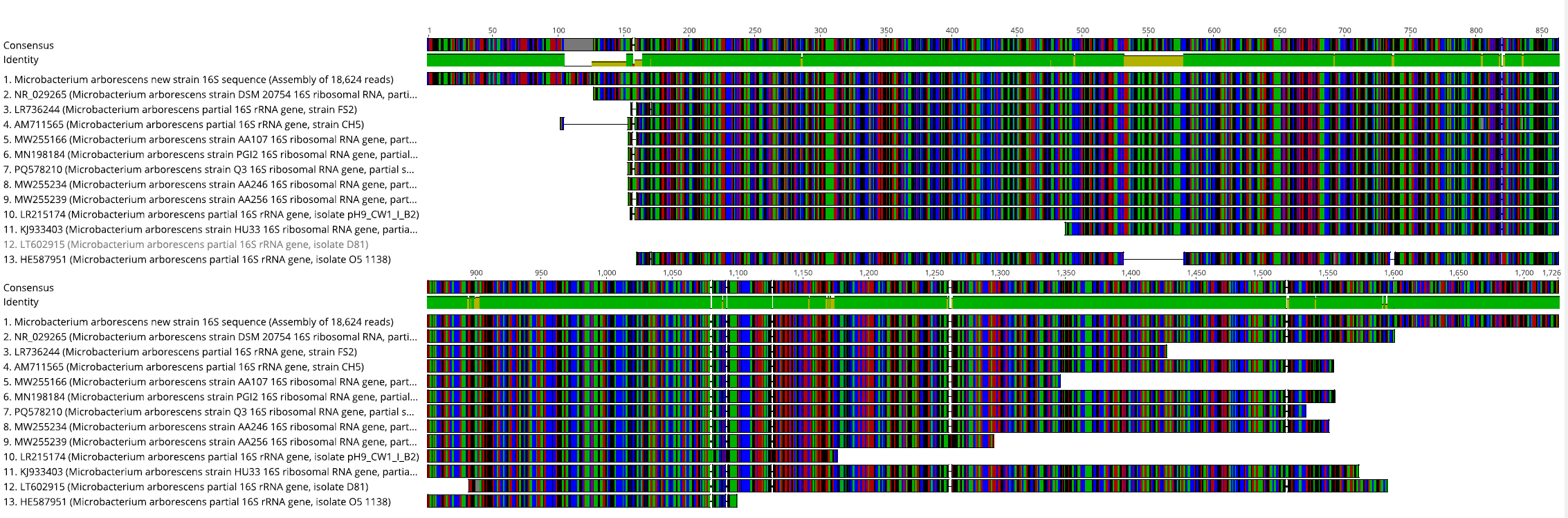
Figure 4: Full 16S rRNA sequence alignment, against the parental strain. Identity graph is shown as a colorful line on top of the MSA where green indicates 100% identity.

The average nucleotide identity (ANI) analysis (8) result showed 100% of identity of our obtained strain 16S rRNA and the published sequence of the parental strain *Microbacterium arborescens* DSM 20754 (NCBI accession NR\_029265):

|  |  |
| --- | --- |
| Metric | Value |
| OrthoANIu value (%) | 100.00 |
| Genome A length (bp) | 1,020 |
| Genome B length (bp) | 1,020 |
| Average aligned length (bp) | 894 |
| Genome A coverage (%) | 87.65 |
| Genome B coverage (%) | 87.65 |

Table 3: The average nucleotide identity (ANI) analysis based on the web tool OrthoANIu (8) for the 16S rRNA sequence alignment against the parental strain.

For the multiple sequence alignment (MSA) against multiple strains of this bacteria species, the published 16S rRNA sequences of the following strains were used: *Microbacterium arborescens* strain DSM 20754 (NR\_029265), strain FS2 (LR736244), strain CH5 (AM711565), strain AA107 (MW255166), strain PGI2 (MN198184), strain Q3 (PQ578210), strain AA246 (MW255234), strain AA256 (MW255239), isolate pH9\_CW1\_I\_B2 (LR215174), strain HU33 (KJ933403), isolate D81 (LT602915), and isolate O5 1138 (HE587951). These sequences were selected to represent a broad range of genetic diversity within the species, enabling a comprehensive analysis of the 16S rRNA gene region across different *M. arborescens* strains. The multiple sequence alignment (MSA) of the 16S rRNA sequence of 12 published strains available in NCBI is shown on figure 5.

 Figure 5: Multiple sequence alignment of the sample strain 16S rRNA. Identity graph is shown as a colorful line on top of the MSA where green indicates 100% identity, yellow indicates at least 30% and lower than 100%, and red indicates lower than 30% identity.

**Whole genome average nucleotide identity (ANI) analysis:**

To assess the genomic similarity between the sample *M. arborescens* WGA and the reference genome strain *M. arborescens* DSM 20754 (GCA\_003339645.1), an OrthoANIu (8) comparison was performed. The analysis revealed a total ANI value of 99.97%, indicating that the assembly closely matches the reference genome. These results suggest a high degree of genomic similarity between the sample assembly and the reference genome, with a well-aligned region that represents over 70% of the total genome length for both genomes. Below is a summary of the key metrics from the comparison (table 4):

|  |  |
| --- | --- |
| Metric | Value |
| OrthoANIu value (%) | 99.97 |
| Genome A length (bp) | 3,275,220 |
| Genome B length (bp) | 3,389,460 |
| Average aligned length (bp) | 2,346,256 |
| Genome A coverage (%) | 71.64 |
| Genome B coverage (%) | 69.22 |

Table 4: The average nucleotide identity (ANI) analysis based on the web tool OrthoANIu (8) for the SPAdes polished assembly genome of the sample (A) against the parental strain (B) - *M. arborescens* DSM 20754 (GCA\_003339645.1).

**Xylose isomerase sequence comparison:**

The sequence of the enzyme of interest, xylose isomerase, from the sequenced strain was obtained from the WGA and compared to the available published sequences of *Microbacterium arborescens strains* DSM 20754 (MDQ1218072), *ND21* (ASM166277v1), *RCB1* (ASM333966v1), and *SSF12* (CP128474). Multiple sequence alignment (MSA) of the published protein sequences and the newly generated protein sequence of the sample strain was generated to identify sequence similarities and potential variations between the strains. The final protein alignment is depicted in figure 6.

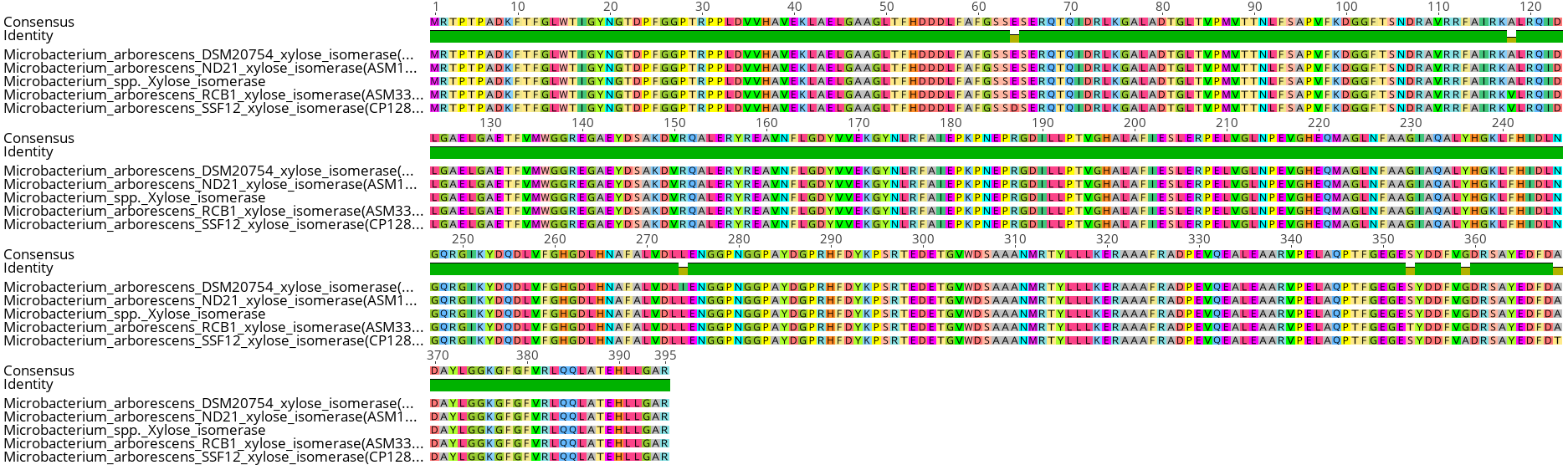


Figure 6: Multiple sequence alignment of the sample xylose isomerase sequence and four other published proteins from closely related strains. Identity graph is shown as a colorful line on top of the MSA where green indicates 100% identity, yellow indicates at least 30% and lower than 100%, and red indicates lower than 30% identity.

The protein sequence of thexylose isomerase from the sequenced strain was compared to the available published sequence of *Microbacterium arborescens strains* DSM 20754 (MDQ1218072) and showed a 99.75% identity, and the only amino acid difference was at position 274, from leucine (L) 🡪 isoleucine (I). The newly obtained protein sequence:

**>Microbacterium\_** **arborescens\_spp.\_xylose\_isomerase**

**MRTPTPADKFTFGLWTIGYNGTDPFGGPTRPPLDVVHAVEKLAELGAAGLTFHDDDLFAFGSSESERQTQIDRLKGALADTGLTVPMVTTNLFSAPVFKDGGFTSNDRAVRRFAIRKALRQIDLGAELGAETFVMWGGREGAEYDSAKDVRQALERYREAVNFLGDYVVEKGYNLRFAIEPKPNEPRGDILLPTVGHALAFIESLERPELVGLNPEVGHEQMAGLNFAAGIAQALYHGKLFHIDLNGQRGIKYDQDLVFGHGDLHNAFALVDLLENGGPNGGPAYDGPRHFDYKPSRTEDETGVWDSAAANMRTYLLLKERAAAFRADPEVQEALEAARVPELAQPTFGEGESYDDFVGDRSAYEDFDADAYLGGKGFGFVRLQQLATEHLLGAR**

This amino acid change from leucine to isoleucine in the xylose isomerase protein sequence is considered a conservative amino acid replacement. This means that the substitution involves two amino acids with similar properties, which is less likely to significantly affect the protein's function or structure (9). Leucine and isoleucine are both hydrophobic, aliphatic amino acids with similar structures. They differ only in the position of a methyl group in their side chains.

**2.5.1 - Identification of genes and/or genetic elements of concern:**

**Antimicrobial resistance**

Antimicrobial resistance (AMR) gene screening was conducted on the SPADes assembled genome of sample *Microbacterium arborescens* spp. using ABRicate (version 1.0.1) (10), in Galaxy version 5.8.0 (5). The analysis was performed against four AMR reference databases: **NCBI Bacterial Antimicrobial Resistance Reference Gene Database** (5386 sequences, December 2024) (11), **CARD** (2631 sequences, December 2024) (12), **ARG-ANNOT** (2223 sequences, December 2024) (13),and **ResFinder** (3077 sequences, December 2024) (14). The screening parameters were set to a minimum DNA identity threshold of 80% and a minimum coverage threshold of 70%. No AMR genes were identified in the genome when compared to any of the four databases. These results suggest that *Microbacterium arborescens* spp. does not harbor known acquired AMR genes within the detection limits of the selected databases and thresholds.

**Toxigenicity, pathogenicity and antimicrobial production**

The identification of virulence factors was conducted using ABRicate (version 1.0.1) (10) in Galaxy version 5.8.0 (5) with the **VFDB** (Virulence Factor Database) (2,597 sequences, December 2024) (15). The analysis was performed on the SPADes assembled genome of this sample *Microbacterium arborescens* spp. using a minimum identity threshold of 80% and a minimum coverage threshold of 70%. No known virulence factor genes were detected in the sample WGA, suggesting the absence of characterized virulence determinants within the reference database at the applied thresholds. These results indicate that *M. arborescens* may lack recognized virulence-associated genes.

**References:**

1. Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., Wang, J., Yang, H., Fang, L., & Chen, Q. (2018). SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience, 7*(1), 1-6. <https://doi.org/10.1093/gigascience/gix120>
2. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics, 70*(1), e102. <https://doi.org/10.1002/cpbi.102>
3. Kearse M, et al. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647-1649. Biomatters. (2024). <https://www.geneious.com>
4. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
5. Grüning, B., Leek, J. T., Baker, D., Batut, B., Bretaudeau, A., & Afgan, E., et al. (2024). The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research.* <https://doi.org/10.1093/nar/gkae410>
6. Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., & Goesmann, A. (2021). *Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification*. Microbial Genomics**, 7**(11), e000685. <https://doi.org/10.1099/mgen.0.000685>
7. Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
8. Yoon, S. H., Ha, S. M., Lim, J. M., Kwon, S.J. & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek.* 110:1281–1286. OrthoANIu tool: <https://www.ezbiocloud.net/tools/ani>
9. Stark, T. L., & Liberles, D. A. (2021). Characterizing amino acid substitution with complete linkage of sites on a lineage. Genome Biology and Evolution, 13(10), evab225. <https://doi.org/10.1093/gbe/evab225>
10. Seemann, T. (2016). **ABRicate: mass screening of contigs for antibiotic resistance genes**. GitHub. <https://github.com/tseemann/abricate>
11. NCBI National Database of Antibiotic-Resistant Organisms (NDARO). Retrieved from <https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/>
12. Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., ... & McArthur, A. G. (2017). Comprehensive Antibiotic Resistance Database. Nucleic Acids Research, 45(D1), D566-D573. <https://doi.org/10.1093/nar/gkw1004>
13. Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., & Rolain, J. M. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrobial Agents and Chemotherapy, 58(1), 212-220. <https://doi.org/10.1128/AAC.01310-13>
14. Bortolaia, V., Kaas, R. S., Ruppe, E., et al. (2020). ResFinder 4.0 for predictions of antimicrobial resistance in whole genome sequencing data. *Journal of Antimicrobial Chemotherapy, 75*(12), 3411–3414. <https://doi.org/10.1093/jac/dkaa345>
15. Liu, B., Zheng, D., Zhou, S., Chen, L., & Yang, J. (2022). VFDB 2022: A general classification scheme for bacterial virulence factors. *Nucleic Acids Research, 50*(D1), D912–D917. <https://doi.org/10.1093/nar/gkab1107>