



OPEN

DATA DESCRIPTOR

High-Quality Genome Assembly, Metabolome, Pangenome, and Metabolic Models of *Megasphaera hexanoica* KCCM 43214^T

Darsha Prabhakaran^{1,3}, Seongcheol Kang^{1,3}, Pranav Sasidharan Nair^{1,3}, Byeoung Seong Jeon² & Byoung-In Sang¹✉

Megasphaera hexanoica KCCM 43214^T, isolated from cow rumen, is capable of producing medium-chain carboxylic acids such as hexanoate and octanoate. In this study, we present a high-quality genome assembly, along with intracellular metabolomic profiling and pangenomic analysis. Illumina sequencing generated 2.3 Gbp from 15,293,634 reads with a GC content of 49.5%, while PacBio HiFi sequencing produced 331.5 Mbp across 45,266 reads, with an average read length of 7,323 bp and a HiFi read N50 of 8,214 bp. Hybrid assembly of short and long reads resulted in a single 2.88 Mbp contig, containing 2,835 protein-coding genes. Genome-scale metabolic models were constructed to evaluate its metabolic capabilities under specific growth conditions. Intracellular metabolomic analysis of cells grown in medium containing fructose and lactate revealed key metabolic activities associated with chain elongation. Pangenomic analysis across nine annotated genomes identified 6,721 orthologous genes using OrthoMCL, emphasizing the genetic and functional diversity within the *Megasphaera* genus. This dataset offers valuable insights into the metabolism and biotechnological potential of *M. hexanoica* KCCM 43214^T.

Background & Summary

The production of medium-chain carboxylic acids (MCCAs) has attracted significant interest as a sustainable approach to bio-based chemical synthesis. MCCAs, such as hexanoate and octanoate, serve as versatile platform chemicals, finding applications in antimicrobial formulation, cosmetic, flavor, fragrance industries, and advanced biofuels. Recent studies have focused on the conversion of organic waste materials, such as food waste and agricultural by-products, into MCCAs, aiming to reduce waste and carbon emissions while generating high value products.

Despite growing interest in microbial production of MCCAs, the understanding of the metabolic and genetic mechanisms underlying MCCA biosynthesis remains limited. *Megasphaera hexanoica* KCCM 43214^T, a bacterium isolated from the cow rumen, has been shown to utilize both fructose and lactate as carbon source for chain elongation¹. Notably, this strain is capable of producing not only hexanoate but also longer chain carboxylic acids such as octanoate^{1,2}. A key feature of *M. hexanoica* KCCM 43214^T is the presence of 9 acetate CoA transferase (ACT) genes, which are involved in the reverse β - oxidation (rBOX) pathway and exhibit broad substrate specificity towards carboxylic acids^{3,4}. During chain elongation, these ACT enzymes play a pivotal role by catalyzing the terminal step where they transfer the CoA moiety from an elongated acyl-CoA to a short-chain carboxylate, such as acetate. This crucial reaction not only releases the final MCCA product but also recycles the free CoA, thereby driving continuous and efficient cycles of the rBOX pathway. This enzymatic versatility contributes to the strains superior MCCA production capabilities. However, it remains unclear how the regulatory mechanisms and overall system-level metabolic fluxes, such as intracellular carbon routing from diverse substrates like fructose and lactate differ from those of other chain elongating bacteria. Furthermore, the

¹Department of Chemical Engineering, Hanyang University, Seoul, Republic of Korea. ²Center of Convergence Bioceramic Materials, Korea Institute of Ceramic Engineering & Technology, Cheongju-si, Chungcheongbuk-do, Republic of Korea. ³These authors contributed equally: Darsha Prabhakaran, Seongcheol Kang, Pranav Sasidharan Nair. ✉e-mail: biosang@hanyang.ac.kr

genetic distinctions and evolutionary metabolic adaptations of this strain relative to other diverse members of the *Megasphaera* genus have not been systematically characterized. To bridge this specific knowledge gap, in this study, we present a hybrid genome assembly, comprehensive functional annotations, genome scale metabolic models (GSMs), pangenomic comparison, and intracellular metabolomic profiling of *M. hexanoica* KCCM 43214^T, providing new insights into its metabolic potential and paving the way for its application in biotechnological processes. The genomic DNA (gDNA) from the active culture was sequenced using Illumina and PacBio HiFi technologies. Illumina sequencing generated high-quality paired-end reads, while PacBio sequencing produced HiFi reads suitable for long-read assembly. Together, these datasets provided reliable sequencing data for downstream genome assembly and analysis. Genome assembly of *M. hexanoica* KCCM 43214^T was performed using SPAdes^{5,6} for Illumina data, Flye⁷ for PacBio data, and Unicycler^{8,9} for hybrid integration. The SPAdes assembly yielded 2,766 genes across 24 contigs, while Flye assembly produced a single contig of 2,882,543 bp with 2,800 genes. The Unicycler hybrid assembly resulted in a single contig of 2,882,434 bp with 2,801 genes. All assemblies were annotated using Prokaryotic Genome Annotation System -v1.14.5 (PROKKA)¹⁰ and Rapid Annotations using Subsystems Technology - v1.073 (RAST)¹¹. Domain annotation was performed using multiple libraries, highlighting high functional relevance and numerous non-hypothetical genes.

Using the RAST-annotated hybrid genome, GSMs were constructed and gap-filled with both glucose rich auxomedia and strain-specific fructose and lactate media. Metabolite profiling of *M. hexanoica* KCCM 43214^T was performed using GC-TOF/MS on cultures grown in fructose and lactate supplemented media. Phylogenomic analysis utilized Insert Set of Genomes into Species Tree - v2.2.043¹². Pangenomic analysis, combining prominent strains of the *Megasphaera* genus, identified 6,721 orthologous gene from nine annotated genomes using OrthoMCL¹³. It is anticipated that the data provided here will enhance the understanding of MCCAs production mechanisms and support advancements in production process research.

Methods

Isolation and extraction of genomic DNA from *M. hexanoica* KCCM 43214^T. In this study, gDNA was extracted from actively growing liquid cultures of *M. hexanoica* KCCM 43214^T using the DNeasy UltraClean Microbial Kit (Qiagen Korea Ltd., South Korea), following the manufacturer's protocol. The concentration of the extracted gDNA was initially assessed using a NanoDrop spectrophotometer (ThermoFisher Scientific, USA) to estimate DNA concentration and evaluate sample purity prior to sending the sample to the sequencing service provider. Additional DNA quantification was performed by the sequencing service provider using fluorescence-based assays, including the Quant-iTTM PicoGreenTM dsDNA Assay Kits and dsDNA Reagents (ThermoFisher Scientific, USA) and the QuantiFluor[®] dsDNA System (Promega Corporation, USA), measured with the Victor Nivo Multimode Microplate Reader (PerkinElmer, Inc., USA). Both PicoGreen and QuantiFluor are well-established fluorescence-based dsDNA quantification reagents that operate on the same principle and deliver equivalent analytical performance, and were applied as part of the sequencing facility's standard quality control (QC) workflow. These fluorescence-based assays provide accurate quantification of double-stranded DNA for sequencing library preparation, and no significant discrepancies were observed between the measurements, confirming the reliability of the gDNA quantification data^{14,15}. The integrity and quality of gDNA were evaluated using two complementary platforms. The Agilent 2100 Bioanalyzer (Agilent Technologies, USA) was used for general DNA quality assessment, providing an initial evaluation of gDNA integrity and fragment size distribution. The Femto Pulse System (Agilent Technologies, USA), which can measure fragment sizes up to 165 kb, was used for high molecular weight (HMW) DNA verification specifically required for PacBio HiFi long-read sequencing, as it is the officially recommended platform by PacBio for sample QC prior to long-read library preparation¹⁶. The average fragment size was approximately 18,400 bp, which is well suited for both sequencing platforms used in this study. For Illumina short-read sequencing, gDNA is fragmented during library preparation to a target insert size of approximately 350 bp, making the initial fragment size of 18,400 bp entirely sufficient¹⁷. For PacBio HiFi long-read sequencing, the PacBio Microbial Library Kit v3.0 specifies a minimum average fragment size of 15 kb, and the 18,400 bp average fragment size of the *M. hexanoica* KCCM 43214^T sample exceeds this requirement, confirming its suitability for high-quality long-read library preparation. RNase treatment was performed prior to library preparation to eliminate RNA contamination. While A260/280 and A260/230 purity ratios were not separately measured, this is consistent with the standard QC protocol for the PacBio Microbial Library Kit v3.0, which specifies sample QC criteria based on average fragment size of ≥ 15 kb and total DNA amount of ≥ 3 μ g rather than purity ratios for microbial samples¹⁸. Additional quality validation was performed via agarose gel electrophoresis. As presented in (Table 1), the final gDNA concentration was 164.91 ng/ μ L in a 100 μ L volume, yielding a total of 16.491 μ g of high-quality gDNA suitable for downstream applications (Table 1).

Genome sequencing. The extracted gDNA from *M. hexanoica* KCCM 43214^T was sequenced using both Illumina short-read and PacBio HiFi long-read technologies.

Illumina short-read sequencing. For the Illumina short-read library, whole genome *de novo* sequencing was conducted using 151 bp paired-end reads. Library preparation was carried out using the TruSeq Nano DNA Kit (Illumina, USA) with an average insert size of 350 bp, following the manufacturer's protocol¹⁹. The reported average fragment size of 650 bp represents the total library fragment length, including the 350 bp target insert size and adapter sequences ligated during library preparation, which is consistent with the expected size distribution for the TruSeq Nano DNA Kit. This difference between insert size and fragment size does not reflect any quality issue and had no negative impact on sequencing quality or assembly performance. The library QC metrics indicated that the final library had a concentration of 64.64 nM (corresponding to 27.31 ng/ μ L) and an average fragment size of 650 bp, confirming high quality DNA suitable for sequencing. The library passed all quality assessments, confirming its suitability for further downstream analysis. Raw data statistics from the

gDNA Conc.	Illumina read statistics					
ng/μL	Total bases (bp)	Total reads	GC (%)	AT (%)	Q20 (%)	Q30 (%)
164.91	2,309,338,734	15,293,634	49.5	50.5	98.5	95.2
	PacBio read statistics					
	HiFi reads bases	HiFi reads	HiFi read N50	Avg. read length	Avg. read quality	
	331,500,498	45,266	8,214	7,323	Q30	

Table 1. gDNA Concentration and Sequencing statistics.

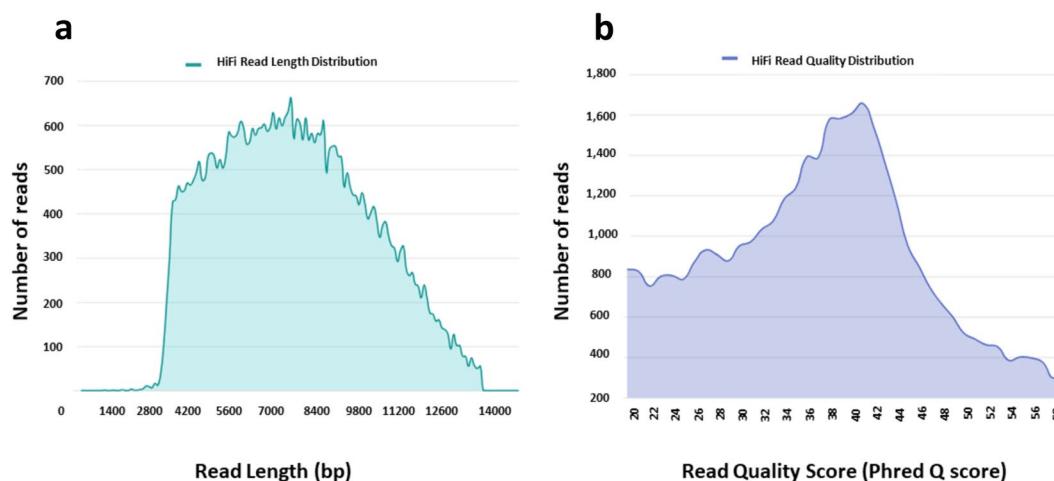


Fig. 1 Sequencing metrics: (a) distribution of HiFi read lengths and (b) quality scores.

Illumina sequencing provided a comprehensive overview of read quality, including total bases, read count, GC content, and Q-scores. Illumina sequencing of the *M. hexanoica* KCCM 43214^T generated 15,293,634 reads, resulting in approximately 2.3 Gbp of data (Table 1). The GC content was 49.5%, and the Q20 and Q30 scores were recorded at 98.5% and 95.2%, respectively, indicating excellent read quality for downstream applications.

PacBio HiFi long-read sequencing. PacBio HiFi sequencing of *M. hexanoica* KCCM 43214^T was performed using the whole-genome *de novo* workflow on the Sequel IIe platform (Pacific Biosciences, USA), with library preparation carried out using the PacBio Microbial Library Kit v3.0 (Pacific Biosciences, USA). Library QC confirmed a concentration of 23.6 ng/μL in a 15 μL volume, with an average fragment size of 6,349 bp. The guideline for the PacBio Microbial Library Kit v3.0 specifies a DNA shearing target insert length of 7–10 kb for microbial samples. It is important to note that in PacBio library QC reporting, the fragment size is reported as the DNA insert size prior to SMRTbell adapter ligation, unlike Illumina where the reported fragment size includes adapter sequences. Therefore, the reported average fragment size of 6,349 bp represents the DNA insert size prior to SMRTbell adapter ligation. The library QC of *M. hexanoica* KCCM 43214^T confirmed a main peak of 8,183 bp, which falls well within the recommended 7–10 kb range for microbial samples²⁰. Thus, the *M. hexanoica* KCCM 43214^T sample met the required criteria, confirming its suitability for HiFi library construction and that the fragment size had no negative impact on read length, yield, or data quality.

As summarized in (Table 1), HiFi read statistics revealed a total of 45,266 reads, generating approximately 331.5 Mbp of sequencing data. The N50 read length, indicating that 50% of all bases are contained in reads of this length or longer, was recorded at 8,214 bp. The distribution of HiFi read lengths and corresponding quality scores is visualized in (Fig. 1). The average read length was 7,323 bp, reflecting a high degree of continuity. Additionally, the average read quality reached Q30, signifying excellent sequencing accuracy, with an average of 20 polymerase passes per read. These metrics confirm the high quality and reliability of the long-read data, supporting its suitability for hybrid genome assembly and downstream functional analysis.

Genome assembly. Genome assembly of *M. hexanoica* KCCM 43214^T was performed using three complementary strategies: Illumina short-read assembly with SPAdes - v3.15.3^{5,6}, long-read assembly with Flye - v2.9.4⁷, and hybrid assembly with Unicycler - v0.4.8^{8,9}.

Read preprocessing and quality control. Prior to assembly, Illumina reads were quality checked with FastQC app - v0.12.1²¹ and poor sequences were trimmed with Trimmomatic - v0.36²². Out of a total input read pairs, 98.6% were retained with both reads surviving the trimming process. Additionally, 1.3% of read pairs retained only the forward read, while 0.1% retained only the reverse read. A minimal percentage of reads, 0.0% (3,021 pairs), were dropped due to insufficient quality or adapter contamination. PacBio HiFi reads were filtered using

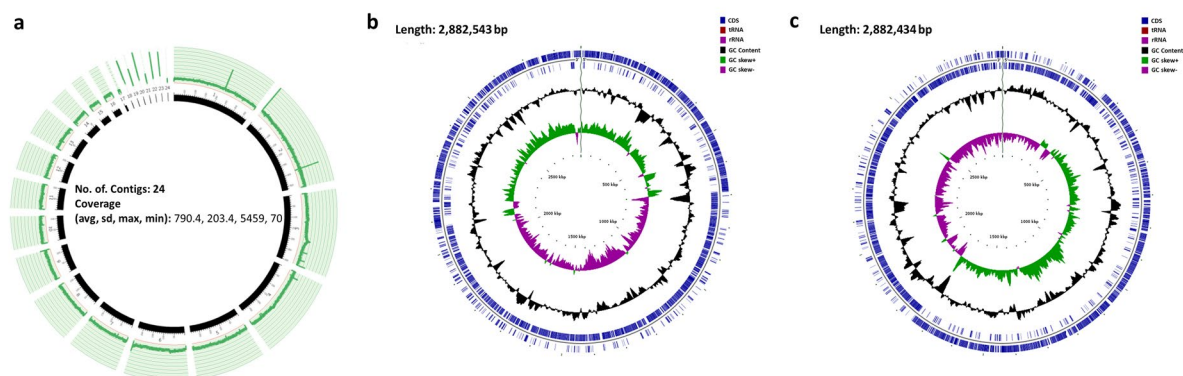


Fig. 2 Visualization and annotation of the genome assembly of *M. hexanoica* KCCM 43214^T. (a) Genome assembly using SPAdes with Illumina short reads, (b) Long-read genome assembly using Flye with PacBio HiFi data, and (c) Hybrid genome assembly using Unicycler.

Filtlong - v0.2.1²³, with the target set to 298,350,448 bp. After filtering, 298,353,003 bp of high-quality long reads were retained, optimizing the dataset for accurate long-read assembly.

Illumina assembly with SPAdes. To obtain an approximate estimate of genome size and basic assembly statistics, Illumina reads were assembled with SPAdes - v3.15.3^{5,6} and the resulting assembly was evaluated using QUAST - v4.4²⁴. From (Fig. 2a), the assembly yielded 24 contigs in total, with 21 contigs measuring over 1,000 bp, 16 contigs over 10,000 bp, and 12 contigs exceeding 100,000 bp. The largest contig reached 372,334 bp, contributing to an overall assembly length of 2,844,982 bp, with an estimated genome coverage of approximately $812\times$ based on the filtered illumina reads used for assembly. Notably, the length of contigs $\geq 100,000$ bp summed up to 2,587,535 bp, highlighting the assembly's robustness.

Assembly metrics indicate an N50 of 243,554 bp and an N75 of 160,662 bp, with L50 and L75 values of 5 and 8 contigs, respectively, reflecting the assembly's contiguity. The GC content was calculated at 48.97%, and only 94 ambiguous bases (Ns) were identified in the entire assembly, corresponding to 3.3 Ns per 100 kbp, indicating a low incidence of gaps. Gene prediction revealed 2,076 unique genes, including 1,957 genes over 300 bp, 442 genes over 1,500 bp, and 79 genes exceeding 3,000 bp.

Long-read assembly with Flye and polishing. The Flye assembler - v2.9.4⁷ produced a highly contiguous single contig of 2,882,543 bp (Fig. 2b), covering the full genome at approximately $104\times$ coverage based on quality-filtered PacBio HiFi reads used for assembly. Both the N50 and N75 values matched the total contig length, and L50/L75 were 1, confirming that the genome was fully assembled into a single piece. The GC content was 49%, and no Ns were present. Gene prediction identified a total of 2,083 unique genes, including 1,977 genes that were at least 300 bp long, 446 genes longer than 1,500 bp, and 80 genes exceeding 3,000 bp. The Flye assembly was further polished using Polypolish - v0.6.0²⁵, with Illumina reads used as the short-read input for error correction. After polishing, the contig maintained a total length of 2,882,543 bp and a GC content of 49%, confirming the assembly's robustness and accuracy.

Hybrid assembly with Unicycler. The hybrid genome assembly for *M. hexanoica* KCCM 43214^T was generated using Unicycler - v0.4.8^{8,9}, which integrates Illumina short reads and PacBio HiFi long reads. Assembly parameters were set to include a minimum contig length of 100 bp, a minimum long-read length of 100 bp, and a moderate threshold for contig bridging to balance contiguity and assembly accuracy. The resulting hybrid assembly produced a single contiguous sequence measuring 2,882,434 bp (Fig. 2c), with Illumina and PacBio HiFi reads contributing approximately $801\times$ and $104\times$ coverage, respectively, based on the filtered reads used for assembly. Both the N50 and N75 values were equal to the full contig length, and the L50 and L75 values were each 1, indicating that the genome was assembled completely in a single piece. The GC content was 49%, and no Ns were detected throughout the assembly. Gene prediction based on the Unicycler assembly identified 2,075 unique genes, with a total of 2,101 genes including partial or overlapping fragments. Among these, 1,974 genes were at least 300 bp in length, 446 genes exceeded 1,500 bp, and 80 genes were longer than 3,000 bp. This hybrid assembly effectively leverages the high base accuracy of Illumina reads and the structural continuity of long-read sequencing, resulting in a highly reliable genome for functional and comparative genomic analyses.

A comparison of genome assemblies generated by different methods is presented in (Fig. 2), highlighting the superior contiguity and completeness achieved through long-read (Fig. 2b) and hybrid assembly strategies (Fig. 2c).

Genome annotation. After genome assembly, all assemblies were annotated using PROKKA-v1.14.5¹⁰ and RAST-v1.073¹¹ simultaneously, providing complementary insights into the genomes. Detailed annotation

Domain Libraries	Hybrid				PacBio HiFi				Illumina			
	RAST		PROKKA		RAST		PROKKA		RAST		PROKKA	
	Genes	Domains	Genes	Domains	Genes	Domains	Genes	Domains	Genes	Domains	Genes	Domains
COGs	2,008	1,891	2,003	1,890	2,009	1,891	2,004	1,890	1,999	1,890	1,996	1,889
NCBI-CDD	1,352	4,606	1,349	4,601	1,352	4,606	1,349	4,601	1,348	4,605	1,345	4,600
SMART	518	221	517	221	518	221	517	221	518	221	517	221
PRK-CDD	1,571	4,568	1,568	4,545	1,571	4,568	1,568	4,545	1,572	4,568	0	0
Pfam	2,176	1,235	2,172	1,231	2,176	1,235	2,172	1,231	2,165	1,237	2,162	1,233
TIGRFAMs	826	716	826	715	826	716	825	715	821	716	821	715
NCBI PGAP	957	827	955	826	956	827	954	826	948	826	947	825

Table 2. Domain annotation summary for PROKKA and RAST in *M. hexanoica* KCCM 43214^T.

statistics generated by PROKKA and RAST for all assemblies are summarized in (Supplementary Tables 1, 2), respectively.

In the Illumina reads-based SPAdes assembly, a total of 2,766 predicted genes were identified, with 2,714 classified as protein-coding genes using PROKKA. Among these, 1,553 genes exhibited non-hypothetical functions, while 707 were assigned with EC numbers, indicating their enzymatic roles. The average protein length across this assembly was approximately 310 amino acids. While, the RAST based functional annotation resulted in 2,923 new features, including 102 non-coding features, with the output comprising 2,821 coding genes, 49 non-coding repeats, and 53 non-coding RNAs. This annotation provided a comprehensive overview of the genome, highlighting its functional capabilities.

In contrast, the PROKKA annotation of PacBio HiFi read-based Flye assembly yielded 2,800 predicted genes, with 2,723 identified as protein-coding. This annotation included 1,586 genes with non-hypothetical functions, and 707 were assigned EC numbers, reflecting their enzymatic roles. The average protein length was slightly higher at 311 amino acids. The RAST annotation of the PacBio HiFi assembly revealed 2,971 new features, which included 138 non-coding features and 2,833 coding genes, 63 non-coding repeats, and 75 non-coding RNAs. Additional feature types scanned included rRNA, tRNA, selenoproteins, pyrrolysoproteins, repeat regions, and CRISPR elements. In total, 2,971 new features were identified, which included 138 non-coding features. The output features consisted of 2,833 coding genes, 63 non-coding repeats, and 75 non-coding RNAs. The analysis provided valuable insights into the functional potential of the genome derived from PacBio sequencing.

Finally, the PROKKA annotation of Unicycler hybrid assembly resulted in the prediction of 2,801 genes, with 2,724 classified as protein-coding. Consistent with the previous assemblies, 1,586 genes exhibited non-hypothetical functions, and 707 were assigned with EC numbers. The average protein length was approximately 310 amino acids. The RAST annotation for the Unicycler assembly identified a genome sequence consisting of a single contig with 2,882,434 nucleotides. A total of 2,973 new features were identified, with 138 classified as non-coding. The output revealed 2,835 coding genes, 63 non-coding repeats, and 75 non-coding RNAs. Additional genomic features identified during RAST annotation included rRNA, tRNA, selenoproteins, pyrrolysoproteins, repeat regions, and CRISPR elements.

All three assemblies demonstrated a significant number of genes with non-hypothetical functions and EC numbers, indicating their functional relevance. The annotations from RAST offered additional insights into the genomic features and functional capabilities of each assembly. Additionally, all assemblies with both RAST and PROKKA annotation were domain annotated using the KBase Annotate Domains in a Genome app - v1.0.10, with all available domain libraries i.e., COGs - v3.19²⁶, NCBI-CDD - v3.19²⁷, SMART - v6.0^{28,29}, PRK - v6.0, Pfam - v35.0³⁰, TIGRFAMs - v15.0³¹, and NCBI PGAP - v8.0³² (refer to Table 2). Furthermore, the RAST-annotated hybrid genome assembly of *M. hexanoica* KCCM 43214^T was analyzed using Phylomarker, Bioelements, dbCAN2 HMMs match, and DRAM³³ for comprehensive functional annotations. The COG functional category distribution of the hybrid assembly is summarized in (Supplementary Table 3). The CAZyme families identified by dbCAN2 and the key metabolic features identified by DRAM are summarized in (Supplementary Tables 4, 5), respectively.

Genomic scale metabolic models. A GSMM of *M. hexanoica* KCCM 43214^T was reconstructed using CarveMe v1.5.2, a top-down automated framework that generates organism-specific metabolic models from genome sequences³⁴. A total of 2,835 protein-coding sequences derived from RAST annotation of the hybrid genome assembly were used as input for model reconstruction. CarveMe establishes gene-protein-reaction (GPR) associations through DIAMOND based homology searches against the BiGG universal reaction database^{35,36}. The initial draft model comprised 1,749 reactions, 1,135 metabolites, and 730 gene-associated reactions, representing a comprehensive whole-cell metabolic network. This draft included 220 exchange reactions, 442 transport reactions, and 296 spontaneous reactions, accounting for extracellular metabolite exchange, membrane transport processes, and non-enzymatic conversions, respectively. These reaction types were automatically included in the initial CarveMe generated draft model without manual intervention. No gapfilling was applied at this stage, and the draft model was retained as a genome-derived baseline for subsequent manual refinement. This approach was intentionally adopted to preserve transparency of the reconstruction process and to allow evaluation of the unrefined genome-derived network prior to curation.

Manual curation was subsequently performed to improve pathway completeness, eliminate non-physiological flux routes, and ensure biological relevance, with particular emphasis on central carbon metabolism and rBOX,

which are critical for MCCA acid production. A detailed list of all reactions added or modified during manual curation, along with their associated pathways and functional roles, is provided in (Supplementary Tables 6, 7). A total of 20 reactions were added and 7 reactions were modified during curation. Key reactions were introduced to complete the rBOX pathway, including thiolase (THLA), 3-hydroxyacyl-CoA dehydrogenases (HACD1, HACD3), and condensation and elongation steps (HEXCOAT and HEXCOATRANS), enabling the conversion of acetyl-CoA to hexanoate. A pathway-level assessment of model completeness, including missing elements and curation actions, is summarized in (Supplementary Table 8).

To ensure efficient carbon flux into this pathway, reactions involved in pyruvate to acetyl-CoA conversion, including pyruvate dehydrogenase (PDH) and pyruvate:ferredoxin oxidoreductase (POR5), were incorporated. Transport reactions for key metabolites, including lactate (L_LACT2), pyruvate (PYRt), acetate (ACt), and hexanoate (HXAt), were added to enable proper exchange between intracellular and extracellular compartments under simulated growth conditions. In addition, several reactions were constrained or modified to prevent non-physiological flux distributions, including glycerol-3-phosphate dehydrogenase reactions (G3PD1ir, G3PD2), lactate dehydrogenase variants (L_LACD2, L_LACD3), and methylglyoxal synthase (MGSA), thereby improving flux realism and eliminating alternative bypass routes. To better represent growth in complex media, amino acid interconversion pathways were incorporated through transaminase reactions (e.g., ALATA_L, ASPTA, VALTA, LEUTA, and ILETA), improving nitrogen assimilation, redox balance, and connectivity between carbon and nitrogen metabolism. Together, these curation steps systematically improved pathway completeness and flux realism, enabling biologically consistent simulation of hexanoate production.

Flux balance analysis (FBA) was performed under anaerobic conditions, with oxygen uptake constrained to zero. Simulations were conducted using media conditions designed to reflect experimental cultivation, including fructose and lactate based substrates³⁷. These conditions correspond to complex media due to the presence of yeast extract. In the model, yeast extract was represented through uptake of amino acids, vitamins, and essential cofactors. Carbon source uptake rates were fixed according to experimentally measured values. In the fructose condition, fructose uptake was fixed while lactate uptake was blocked, whereas in the lactate condition, lactate uptake was fixed and fructose uptake was blocked. In addition, butyrate uptake and acetate secretion rates were constrained based on experimental observations. An additional auxomedia condition was used as a permissive rich medium, in which multiple nutrients were made available without uptake constraints to evaluate model feasibility. Other nutrients, including minerals and cofactors, were left unconstrained. The detailed composition of all simulated media is provided in (Supplementary Table 9).

To investigate MCCA production, FBA simulations were specifically analyzed with respect to hexanoate production flux, while biomass production was retained as the primary objective function. The objective function for all FBA simulations was biomass flux maximization, representing the predicted specific growth rate (h^{-1}). The biomass reaction was automatically incorporated during model reconstruction using CarveMe v1.5.2, based on the universal Gram-negative biomass composition from the BiGG database, which includes standard macromolecular components such as proteins, lipids, carbohydrates, DNA, RNA, and cofactors. No species-specific biomass composition is currently available for *M. hexanoica* KCCM 43214^T; therefore, simulated growth rates should be interpreted as relative indicators of metabolic performance across conditions rather than absolute values. Biomass was retained as the primary objective function to ensure physiologically realistic flux distributions, while hexanoate production was analyzed as a key metabolic output.

The final curated model comprised 1,769 reactions and 1,147 metabolites, with 730 gene associations, ensuring all reactions remained genome-supported. Mass and charge balance of the curated model was evaluated using COBRApy v0.26.3^{38,39}. All manually curated internal reactions were individually verified and confirmed to be balanced. During this process, an initial imbalance in the BUTKr reaction was identified and resolved by correcting missing formula and charge information for the metabolite butanoyl phosphate (butp_c). Apparent imbalances observed in exchange reactions and the biomass reaction are expected due to their boundary or pseudo-reaction nature, while imbalance in the POR5 reaction reflects known limitations in formula definitions for ferredoxin-related metabolites. Details are provided in (Supplementary Table 10).

Intracellular metabolite extraction and analysis in *M. hexanoica* KCCM 43214^T. For intracellular metabolite analysis, *M. hexanoica* KCCM 43214^T was cultured in a medium containing both fructose and lactate to simulate the optimal conditions for maximal MCCA production, as previously demonstrated, and incubated at 37 °C with 150 rpm. Upon reaching the mid-exponential growth phase (18 hours post-inoculation, determined based on prior growth kinetics under identical conditions), triplicate biological samples containing 0.5 mg of cell dry weight was used for intracellular metabolome extraction. Immediately after sampling, cellular metabolism was quenched by mixing the sample with an equal volume of prechilled saline solution⁴⁰. Prechilled saline was selected to rapidly reduce metabolic activity while minimizing osmotic shock and potential metabolite leakage from the cells. The mixture was centrifuged at 12,000 rpm for 2 min under 4 °C, and the supernatant was discarded. The resulting cell pellet was resuspended in saline solution to remove extracellular metabolites, followed by centrifuging again under the same condition and subsequent removal of the supernatant. The cell pellet was dried for 2 hours at 4 °C using a centrifugal evaporator CVE-3000 (Eyela, Japan) and stored at −80 °C until further metabolome extraction. This drying step at 4 °C was introduced to minimize solvent dilution during subsequent extraction. While we acknowledge that this process may potentially compromise the stability of certain thermolabile metabolites, all biological triplicates were handled under strictly identical conditions to allow for consistent analysis. For intracellular metabolite extraction, a 1:1 (v/v) solution of acetonitrile and water, spiked with 1 μM adonitol as an internal standard, was used as the extraction solvent. To the tube containing the dried cell pellet, 100 mg of glass beads (425–600 μm) and 1 mL of extraction solvent were added. The tube was tightly sealed and vortexed vigorously for 2 min to disrupt the cell wall and extract intracellular metabolites. Following vortexing, the sample was centrifuged at 12,000 rpm for 2 min at 4 °C, and 800 μL of the resulting supernatant containing

intracellular metabolome was transferred to a new microtube. For subsequent derivatization, the metabolite solution was completely dried in a centrifugal evaporator CVE-3000 (Eyela, Japan) at 4 °C for 6 to 8 hours. The dried sample was subjected to two-step derivatization to enhance volatility and improve detection sensitivity prior to GC × GC-TOF/MS system (Pegasus BT 4D, LECO, USA) analysis. For methoxylation, 10 µL of methoxyamine hydrochloride in pyridine (20 mg mL⁻¹) was added to the sample and incubated at 1,000 rpm for 90 min at 30 °C. This was followed by silylation, achieved by adding 90 µL of N-Methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) and reacting at 1,000 rpm for 30 min at 37 °C. The derivatized samples were transferred to glass vials and analyzed using a GC × GC-TOF/MS system (Pegasus BT 4D, LECO, USA) operated in 1D GC mode for GC × GC setup. This 1D configuration was intentionally chosen over 2D mode to ensure high-confidence metabolite identification. Operating in 1D mode allowed us to utilize extensive, established 1D Retention Index (RI) libraries for precise compound verification, which was prioritized because corresponding RI databases for 2D mode are currently limited. In this configuration, although the secondary column was connected downstream of the primary column to enable GC × GC analysis, the flow modulator was not activated, and the samples were analyzed in a conventional 1D manner—sequentially passing through the primary and secondary columns before reaching the detector. A DB-5MS UI capillary column (30 m × 0.25 mm i.d., 0.25 µm film thickness; Agilent, USA) was utilized as the primary column, while an Rxi-17Sil MS column (1 m × 0.25 mm i.d., 0.25 µm film thickness; Restek, USA) was employed as the secondary column in the series. Helium was used as the carrier gas at a constant flow rate of 1 mL min⁻¹. A volume of 1 µL was injected in splitless mode, with an inlet purge time of 30 seconds. The inlet temperature was set at 250 °C. The oven temperature program began at 50 °C (held for 1 min), increased to 320 °C at a rate of 20 °C min⁻¹, and was held at 320 °C for 6.5 min. The secondary oven temperature was maintained at 5 °C above the main oven temperature throughout the analysis, and the transfer line temperature was maintained at 280 °C. Mass spectra were acquired after a solvent delay of 385 seconds, scanning a mass range of 50–500 m/z at an acquisition rate of 20 spectra per second. To establish a retention index reference, a mixture of C7–C30 saturated alkanes was analyzed under the same conditions. Additionally, various standard compounds, including amino acids, sugars, and organic acids, were derivatized and analyzed to build an in-house spectral library.

Using Non-Target Deconvolution algorithm (NTD) in ChromaTOF software – v5.51.06⁴¹, overlapped peaks were deconvoluted and the peaks were identified by comparing the mass spectra with NIST library and in-house library database. To strictly exclude analytical impurities and prevent false-positive identifications in our non-targeted approach, the minimum signal-to-noise (S/N) ratio and minimum similarity for matching were set to 300 and 500, respectively. An alkane scale retention index method was applied to calculate the retention index of each identified peak. The processed data were then exported to mzML format.

Phylogenomics and pangenomics. The hybrid assembly of *M. hexanoica* KCCM 43214^T was used for phylogenomic analysis using Insert Set of Genomes into Species Tree - v2.2.043 tool in KBase¹². This approach reconstructs species trees using 49 conserved universal marker genes derived from Clusters of Orthologous Groups (COG) families, primarily consisting of ribosomal proteins, translation-related factors, and essential housekeeping genes. The marker genes from each genome were inserted into curated multiple sequence alignments that were trimmed using GBLOCKS to remove poorly aligned regions, and the resulting alignments were concatenated for phylogenetic reconstruction. The species tree was inferred using FastTree2 (v2.1.10) employing an approximate maximum likelihood algorithm under the Whelan and Goldman Model (WAG) + Category Model (CAT) model of evolution, with Shimodaira Hasegawa like local support values calculated for each node. The resulting phylogeny positioned *M. hexanoica* KCCM 43214^T close to a previously deposited 454-based assembly. To quantitatively support this relationship, average nucleotide identity (ANI) was calculated between the two genomes, revealing an ANI value of 99.991%, confirming their extremely close genomic relationship. Pangenomic analysis was performed across popular type strains belonging to the *Megasphaera* genus to identify core and accessory genes, with particular focus on genomic features unique to *M. hexanoica* KCCM 43214^T that may underlie its distinct chain elongation capacity (Fig. 3). The strains selected for pangenomics are listed below, i.e., *Megasphaera vaginalis* DSM 111203^{T42}, *Megasphaera sueciensis* DSM 17042^{T43}, *Megasphaera stantonii* DSM 106750^{T44}, *Megasphaera elsdenii* DSM 20460^{T45}, *Megasphaera micronuciformis* F0359⁴⁶, *M. hexanoica* KCCM 43214^{T47}, *Megasphaera massiliensis* DSM 26228^{T48}, *Megasphaera paucivorans* DSM16981^{T49}, and *Megasphaera cerevisiae* DSM 20462^{T50}. All genomes were retrieved from the NCBI database and annotated using the RAST -v1.073 toolkit¹¹. Additionally, domain-level annotations were conducted using major domain libraries, including COGs²⁶, CDD²⁷, SMART^{28,29}, and PRK models³⁰ from the NCBI Conserved Domain Database (CDD v3.19), analyzed with RPS-BLAST (v2.13.0) using an E-value cutoff of 1e⁻⁴. In parallel, domain identification was also performed using Pfam (v35.0)³⁰, TIGRFAMs (v15.0)³¹ from the J. Craig Venter Institute, and NCBIfam (PGAP v8.0)³² from NCBI, employing HMMER (v3.3.2) with model-specific trusted cutoffs.

All nine RAST and Domain annotated genomes were utilized for building pangenome with Build Pangenome with OrthoMCL – v2.0¹³ implemented in KBase under default settings, resulting in 6,721 orthologous genes. Orthologous groups were identified through all-against-all BLASTP similarity searches followed by Markov clustering using default parameters, including a BLASTP E-value cutoff of 1 × 10⁻⁵, a minimum sequence identity threshold of 30%, and an inflation parameter of 1.5. Paralogous sequences were retained and grouped according to sequence similarity as part of the standard OrthoMCL workflow. All genomes were re-annotated using the RAST toolkit prior to pangenome analysis to ensure consistent gene prediction and functional annotation across strains, thereby minimizing bias in ortholog identification and in the estimation of core and accessory gene sets. Genome completeness and contamination were evaluated using CheckM, and all genomes showed high completeness with contamination levels within acceptable limits for publicly available reference genomes (Supplementary Table 11), confirming that all genomes are of sufficient quality and suitable for pangenome analysis. Core, accessory, and unique gene families were functionally annotated using COG and Pfam databases, and comparatively interpreted to distinguish conserved and variable functional features across

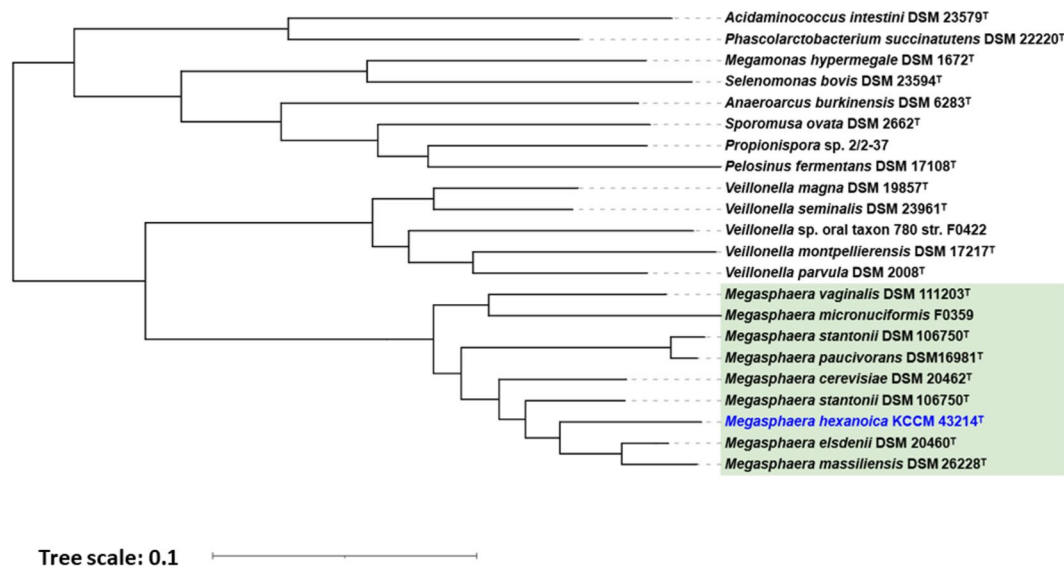


Fig. 3 Phylogenetic Tree Construction for *M. hexanoica* KCCM 43214^T Using Insert Set of Genomes into SpeciesTree App v2.2.0. The tree was generated with 49 core COG genes, including a selection of closely related genomes from the public database, and visualized with FastTree2.

Genome	Genes	Homologs genes	Singletons genes	Families genes
<i>Megasphaera vaginalis</i> DSM 111203 ^T	2,327	1,883	444	1,820
<i>Megasphaera sueciensis</i> DSM 17042 ^T	2,937	2,489	448	2,281
<i>Megasphaera stantonii</i> DSM 106750 ^T	2,568	2,144	424	1,978
<i>Megasphaera elsdenii</i> DSM 20460 ^T	2,224	2,007	217	1,914
<i>Megasphaera micronuciformis</i> F0359	1,692	1,512	180	1,460
<i>M. hexanoica</i> KCCM 43214 ^T	2,835	2,433	482	2,283
<i>Megasphaera massiliensis</i> DSM 26228 ^T	2,503	2,113	390	2,139
<i>Megasphaera paucivorans</i> DSM16981 ^T	2,915	2,433	482	1,992
<i>Megasphaera cerevisiae</i> DSM 20462 ^T	3,122	2,534	588	2,337

Table 3. Pangenome analysis of *Megasphaera* strains: gene count, homologs, singletons, and homolog families.

Megasphaera genomes. This analysis revealed a total of 22,721 protein-coding genes across all genomes, organized into 6,721 gene families. Of these, 3,409 were homologous gene families, defined as clusters of genes sharing significant sequence similarity across two or more genomes, encompassing 19,409 genes. The remaining 3,312 were classified as singleton families, representing genes unique to single genome with no detectable homologs elsewhere in the dataset. Each genome specific genes count, homologs genes count, singleton gene count, and homolog families are represented in (Table 3).

In (Fig. 4), the pangenome was visualized using the Pangenome Circle Plot – v1.2.0, with our hybrid assembly of *M. hexanoica* KCCM 43214^T set as the reference genome. The feature sets including singletons, core genomes, and non-core genome elements were extracted and were subsequently analyzed for functional annotation using the COGs - v3.19²⁶, and Pfam - v35.0³⁰ domain libraries.

Data Records

The data represented in this study has been deposited in NCBI under the BioProject number PRJNA287738. The accession number for both the Illumina and PacBio HiFi raw reads are SRP536548⁵¹, and the GenBank accession number for the genome assemblies are listed as follows: Illumina SPAdes (JBIEKR 010000001-JBIEKR010000024)⁵², PacBio Flye (GCA_043949995.1)⁵³, and Hybrid Unicyc ler assembly (GCF_043520715.1)⁴⁷. The functional annotation of the genome assembly, GSMMs with related data, GC-MS data, and pangenomics data have been deposited in Figshare³⁴.

Technical Validation

All the raw data processing steps, software, and parameters used in this study are described in the “Materials and Methods” section. The quality scores for the genome reads were assessed using FastQC v0.12.1²¹. The quality of the genome assemblies was validated using QUAST-v4.4^{24,55} and Compare Assembled Contig Distributions – v1.1.2⁵⁶.

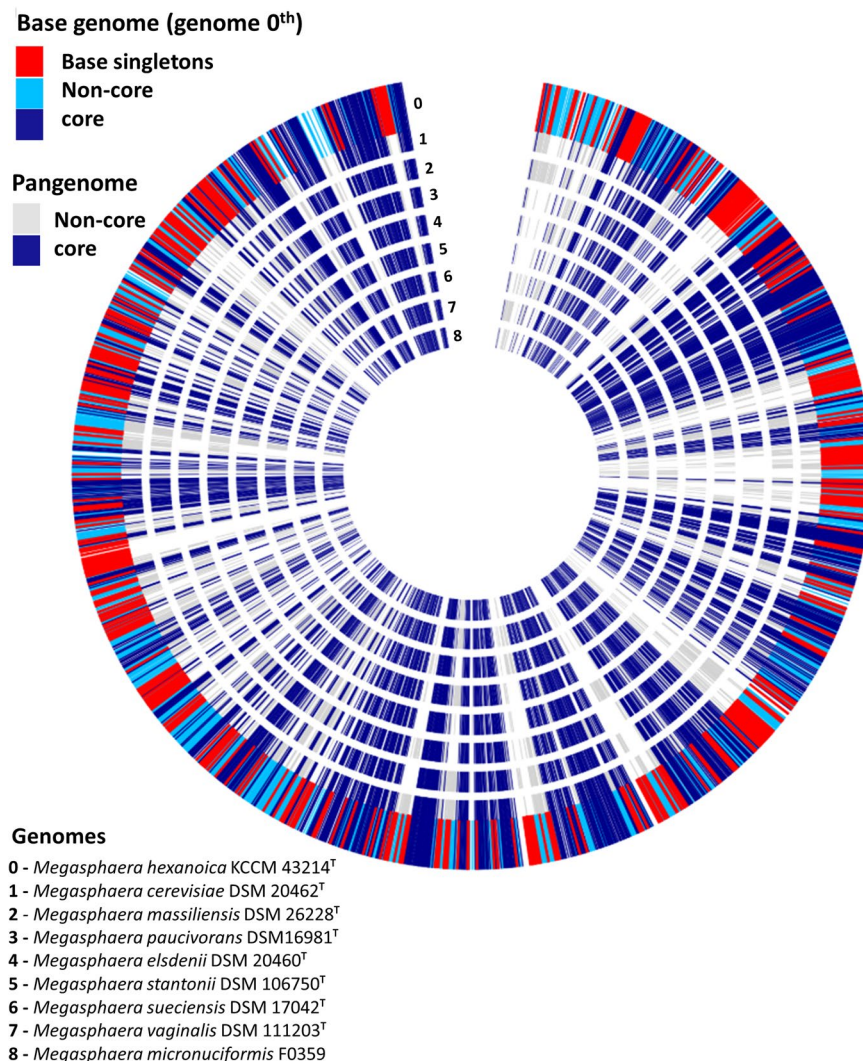


Fig. 4 Pangenome Circle Plot of *Megasphaera* Species. This figure displays a pangenome circle plot for the *Megasphaera* species, encompassing ten genomes namely, *M. hexanoica* KCCM 43214^T (Genome 0), *Megasphaera cerevisiae* DSM 20462^T (Genome 1), *Megasphaera massiliensis* DSM 26228^T (Genome 2), *Megasphaera paucivorans* DSM16981^T (Genome 3), *Megasphaera elsdenii* DSM 20460^T (Genome 4), *Megasphaera stantonii* DSM 106750^T (Genome 5), *Megasphaera sueciensis* DSM 17042^T (Genome 6), *Megasphaera vaginalis* DSM 111203^T (Genome 7), and *Megasphaera micronuciformis* F0359 (Genome 8). The plot categorizes the genes into core, singleton, and non-core pangenomes. Core genes are present in all the analyzed genomes. Singletons are unique to individual genomes, and non-core pangenomes represent genes found in some but not all genomes. The functional profiles for these gene categories are accessible on Figshare⁵⁴.

To further validate the assemblies, sequencing reads were mapped back to their respective assemblies and evaluated using Qualimap v2.2.1⁵⁷. Illumina short reads showed a high mapping rate of 98.89%, with a mean coverage of $784.82 \times$ based on mapped reads. PacBio HiFi reads showed a mapping rate of 100.0%, with a mean coverage of $114.85 \times$ across the Flye and hybrid assemblies. Together, these results confirm strong agreement between the raw sequencing data and all assembled genomes.

Genome completeness and contamination were assessed using CheckM⁵⁸. The assembled genome exhibited 100% completeness and 0.2% contamination, as shown in (Fig. 5a), confirming the high quality and purity of the genome assembly. Genome completeness was further assessed using BUSCO v5 against the bacteria_odb10 dataset of 124 genes, evaluated in both genome and protein modes across all three assemblies including Flye, Unicycler, and SPAdes⁵⁹. As shown in (Fig. 5b), all assemblies yielded identical scores of 122 complete single-copy BUSCOs, representing 98.4% completeness, with zero duplicated, two fragmented, and zero missing genes in both modes. The consistent results across all assemblers and modes confirm that the genome is highly complete and free of contamination or mis-assembly artifacts. The absence of duplicated BUSCOs is consistent with a pure haploid bacterial isolate, and the two fragmented BUSCOs likely reflect minor sequence divergence from the reference profiles rather than assembly gaps, as they persist even in the closed single-contig assemblies produced by Flye and Unicycler.

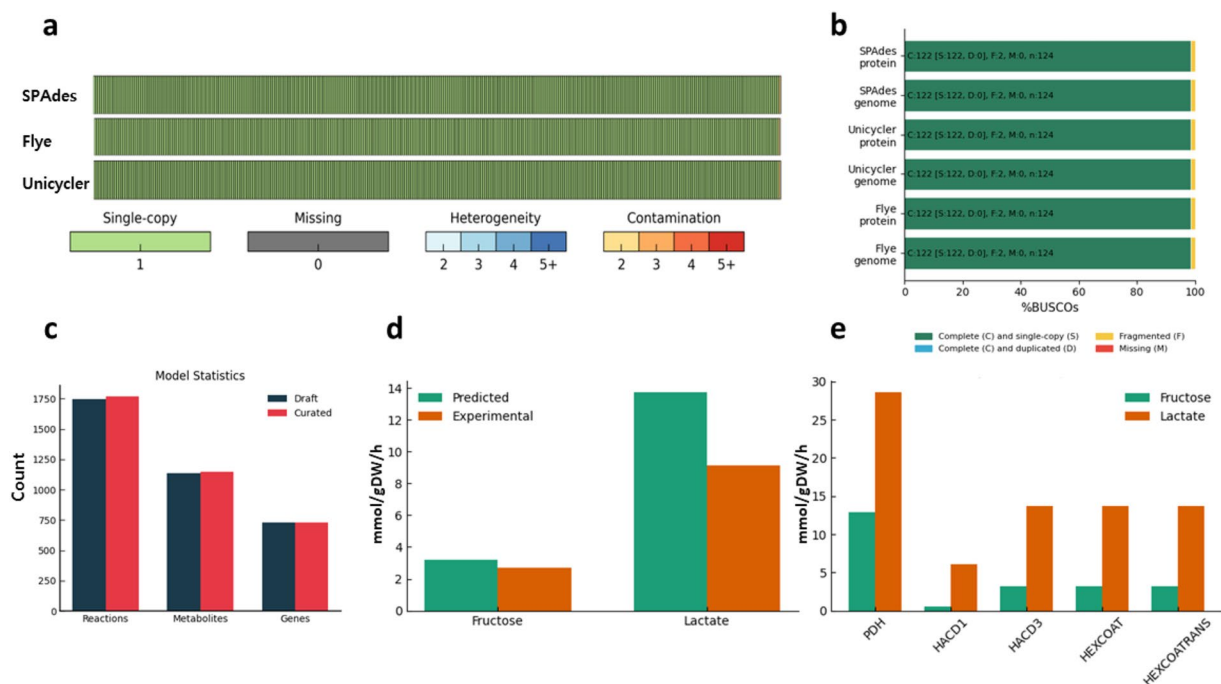


Fig. 5 Genome quality assessment and GSMM summary and validation of hexanoate production in *M. hexanoica* KCCM 43214^T. **(a)** CheckM assessment showing genome completeness and contamination. **(b)** BUSCO assessment of genome completeness across all three assemblies including Flye, Unicycler, and SPAdes evaluated in both genome and protein modes against the bacteria_odb10 dataset. **(c)** Model statistics before and after manual curation. **(d)** Comparison of predicted and experimental hexanoate production under fructose and lactate conditions. **(e)** Flux comparison of key rBOX pathway reactions, showing increased pathway activity under lactate conditions.

To provide comparative context for the genome annotation, the coding gene count, non-coding RNA content, and CRISPR elements of *M. hexanoica* KCCM 43214^T were compared with all eight *Megasphaera* type strains included in the pangenome analysis, all annotated using the same RAST pipeline for consistency, as summarized in (Supplementary Table 12)^{42–50}. The coding gene count of *M. hexanoica* KCCM 43214^T is 2,835, which falls within the range observed across the genus of 1,692 to 3,122. The total non-coding RNA content is 150, which includes tRNA and rRNA, and is consistent with the genus range of 51 to 150, comparable to *M. stantonii* DSM 106750^T, which also has 150 non-coding RNA elements. No CRISPR arrays were detected in *M. hexanoica* KCCM 43214^T, which is consistent with *M. vaginalis* DSM 111203^T and *M. paucivorans* DSM16981^T, indicating that the absence of CRISPR elements is not uncommon within the *Megasphaera* genus. These comparisons confirm that the annotation results of *M. hexanoica* KCCM 43214^T are typical for the genus and support the reproducibility and reliability of the genome annotation.

GSMMs checked and validated using FBA analysis and Memote, respectively³⁹. The curated model achieved a Memote score of 86%, indicating high stoichiometric consistency, annotation completeness, and network connectivity. A summary of model statistics, predicted versus experimental hexanoate production, and key flux differences in the rBOX pathway is presented in (Fig. 5c–e). Model predictions were evaluated under experimentally relevant fructose and lactate-based conditions. Experimental results showed that hexanoate production was higher under lactate conditions, reaching 9.126 mmol/gDW/h, compared to 2.711 mmol/gDW/h under fructose conditions. The model successfully reproduced this substrate-dependent trend, predicting hexanoate production of 13.74 mmol/gDW/h under lactate and 3.22 mmol/gDW/h under fructose, as illustrated in (Fig. 5d).

This agreement confirms that the model captures the metabolic shift toward enhanced rBOX under lactate feeding. Consistent with this, flux through key rBOX reactions, including HACD1, HACD3, HEXCOAT, and HEXCOATRANS, was higher under lactate conditions, supporting increased chain elongation and hexanoate production, as shown in (Fig. 5e).

Data availability

All sequencing datasets have been deposited at NCBI under BioProject PRJNA287738. Raw reads (both Illumina and PacBio HiFi) are available in the Sequence Read Archive under SRP536548⁵¹. Genome assemblies are available as: Illumina SPAdes WGS accession JBIEKR010000001–JBIEKR010000024⁵²; PacBio Flye GenBank assembly accession GCA_043949995.1⁵³; and hybrid Unicycler RefSeq assembly accession GCF_043520715.1⁴⁷. The derived resources including functional annotations of the assemblies, GSMMs, and associated inputs, and pangenome datasets are deposited on Figshare⁵⁴. The intracellular metabolomics data generated in this study have been deposited in the MetaboLights repository under the accession number MTBLS14019⁶⁰.

Code availability

All versions of third-party software and scripts used in this study are described and referenced accordingly in the Materials and Methods section.

Received: 20 August 2025; Accepted: 15 May 2026;

Published online: 01 June 2026

References

- Jeon, B. S., Kim, S. & Sang, B.-I. *Megasphaera hexanoica* sp. nov., a medium-chain carboxylic acid-producing bacterium isolated from a cow rumen. *International Journal of Systematic and Evolutionary Microbiology* **67**, 2114–2120 (2017).
- Kim, H., Jeon, B. S. & Sang, B.-I. An efficient new process for the selective production of odd-chain carboxylic acids by simple carbon elongation using *Megasphaera hexanoica*. *Scientific reports* **9**, 11999 (2019).
- Kim, S. G. *et al.* Optimization of hexanoic acid production in recombinant *Escherichia coli* by precise flux rebalancing. *Bioresource technology* **247**, 1253–1257 (2018).
- Park, H., Jeon, B. S. & Sang, B.-I. Efficient, simple production of corresponding alcohols from supplemented C 2–C 8 carboxylic acids in *Escherichia coli* using acyl-CoA transferase from *Megasphaera hexanoica*. *Biotechnology and Bioengineering* **25**, 599–606 (2020).
- Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology* **19**, 455–477 (2012).
- Prijbelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes de novo assembler. *Current protocols in bioinformatics* **70**, e102 (2020).
- Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* **37**, 540–546 (2019).
- Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology* **13**, e1005595 (2017).
- Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial genomics* **3**, e000132 (2017).
- Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
- Brettin, T. *et al.* RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports* **5**, 1–6 (2015).
- Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490 (2010).
- Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13**, 2178–2189 (2003).
- Dragan, A. *et al.* Characterization of PicoGreen interaction with dsDNA and the origin of its fluorescence enhancement upon binding. *Biophysical journal* **99**, 3010–3019 (2010).
- Promega Corporation. QuantiFluor dsDNA System Technical Manual. *Promega* <https://www.promega.com> (2020).
- PacificBiosciences. HiFi whole genome sequencing using the Agilent Femto Pulse system PacBio <https://www.pacb.com/wp-content/uploads/HiFi-WGS-sequencing-using-Agilent-Femto-Pulse-system.pdf> (2021).
- Illumina. TruSeq Nano DNA Library Prep Kit. I. *Illumina* <https://www.illumina.com> (2019).
- PacificBiosciences. Procedure checklist: Preparing whole genome and metagenome libraries using SMRTbell Prep Kit 3.0. *PacBio* <https://www.pacb.com> (2022).
- TruSeq, R. TruSeqTM RNA and DNA Sample Preparation Kits v2. *CHIP-SEQ RIBO-SEQ*, 133.
- Lang, D. *et al.* Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *Gigascience* **9**, gaa123 (2020).
- Bioinformatics group at the Babraham Institute. FastQC source, version 0.12.1. *Babraham Bioinformatics* <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2023).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- WickRyan (2017).
- Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
- Wick, R. R. & Holt, K. E. Polypolish: short-read polishing of long-read bacterial genome assemblies. *PLoS computational biology* **18**, e1009802 (2022).
- Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
- Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic acids research* **45**, D200–D203 (2017).
- Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic acids research* **46**, D493–D496 (2018).
- Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic acids research* **43**, D257–D260 (2015).
- El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic acids research* **47**, D427–D432 (2019).
- Haft, D. H. *et al.* TIGRFAMs and genome properties in 2013. *Nucleic acids research* **41**, D387–D395 (2012).
- Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic acids research* **44**, 6614–6624 (2016).
- Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic acids research* **48**, 8883–8900 (2020).
- Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research* **46**, 7542–7553 (2018).
- Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59–60 (2015).
- King, Z. A. *et al.* BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research* **44**, D515–D522 (2016).
- Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nature biotechnology* **28**, 245–248 (2010).
- Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: constraints-based reconstruction and analysis for python. *BMC systems biology* **7**, 74 (2013).
- Lieven, C. *et al.* MEMOTE for standardized genome-scale metabolic model testing. *Nature biotechnology* **38**, 272–276 (2020).
- Sake, C. L., Newman, D. M. & Boyle, N. R. Evaluation of quenching methods for metabolite recovery in photoautotrophic *Synechococcus* sp. PCC 7002. *Biotechnology Progress* **36**, e3015 (2020).
- Gaida, M., Franchina, F. A., Stefanuto, P.-H. & Focant, J.-F. Top-down approach to retention time prediction in comprehensive two-dimensional gas chromatography–mass spectrometry. *Analytical chemistry* **94**, 17081–17089 (2022).
- Srinivasan, S., Beamer, M. A. & Fiedler, T. L. *GenBank* https://identifiers.org/insdc.gca:GCA_000478965.1 (2021).

43. GenBank https://identifiers.org/insdc.gca:GCA_050219315.1 (2025).
44. GenBank https://identifiers.org/insdc.gca:GCA_003030445.1 (2018).
45. GenBank https://identifiers.org/insdc.gca:GCF_003010495.1 (2018).
46. GenBank https://identifiers.org/insdc.gca:GCF_000165735.1 (2010).
47. Sang, B.-I. & Prabhakaran, D. GenBank https://identifiers.org/insdc.gca:GCA_043520715.1 (2024).
48. GenBank https://identifiers.org/insdc.gca:GCF_000455225.1 (2013).
49. GenBank https://identifiers.org/insdc.gca:GCA_900103535.1 (2016).
50. GenBank https://identifiers.org/insdc.gca:GCA_001045675.1 (2015).
51. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRP536548> (2025).
52. Sang, B.-I. & Prabhakaran, D. GenBank https://identifiers.org/insdc.gca:GCF_043498135.1 (2024).
53. Sang, B.-I. & Prabhakaran, D. GenBank https://identifiers.org/insdc.gca:GCA_001045675.1 (2024).
54. Prabhakaran, D., Kang, S., Nair, P. S., Jeon, B. & Sang, B. I. High-Quality Genome Assembly, Metabolome, Pangenome, and Metabolic Models of *Megasphaera hexanoica* KCCM 43214^T. *figshare* <https://doi.org/10.6084/m9.figshare.29925416> (2025).
55. Mikhchenko, A., Valin, G., Pribelski, A., Saveliev, V. & Gurevich, A. Icarus: visualizer for de novo assembly evaluation. *Bioinformatics* **32**, 3321–3323 (2016).
56. Arkin, A. P. *et al.* KBase: the United States department of energy systems biology knowledgebase. *Nature biotechnology* **36**, 566–569 (2018).
57. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
58. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043–1055 (2015).
59. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular biology and evolution* **38**, 4647–4654 (2021).
60. Prabhakaran, D., Kang, S., Nair, P. S., Jeon, B. & Sang, B. I. MetaboLights MTBLS14019. <https://www.ebi.ac.uk/metabolights/MTBLS14019> (2026).

Acknowledgements

This work was supported by Korea Environment Industry & Technology Institute (KEITI) through Development of Demonstration technology for energy conversion using waste resources Program, funded by Korea Ministry of Environment (MOE) (2022003480001).

Author contributions

D.P. and S.K. prepared the data and wrote the draft under the supervision of B.I., P.N. wrote and reviewed the manuscript, B.S. Isolated the strain and reviewed the manuscript, B.I. supervised the study and revised the draft, All authors reviewed and contributed to the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-026-07473-z>.

Correspondence and requests for materials should be addressed to B.-I.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026