International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Hybrid Whole-Genome Assembly Using Oxford Nanopore and Illumina Platforms for Functional Genomics of *Streptococcus equi* subsp. *zooepidemicus* MTCC3523 in Optimized Hyaluronic Acid Production

Fatemeh Khatami Toosi [1], Maryam Naseroleslami [1] and Seyedeh Zoha Tabatabaei [2,*]

[1] *Department of Cellular and Molecular Biology, Faculty of Advanced Science and Technology, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran.*
[2] *Cardiogenetic Research Center, Rajaei Cardiovascular Institute, Tehran, Iran.*

International Journal of Science and Research Archive, 2025, 14(03), 565-577

Publication history: Received on 29 January 2025; revised on 07 March 2025; accepted on 10 March 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.3.0655

Abstract

Recent advancements in sequencing technologies have transformed bacterial genomics, enabling high-resolution genome characterization. Whole-genome sequencing (WGS) plays a crucial role in bacterial taxonomy, functional genomics, and evolutionary studies. Hybrid sequencing approaches that combine short-read and long-read technologies have significantly improved genome assembly and annotation, particularly for complex and repetitive genomes.

This study compares Illumina and Oxford Nanopore sequencing in the WGS of *Streptococcus equi* subsp. *zooepidemicus* MTCC 3523. Illumina sequencing, known for its high accuracy, produces short reads, while Nanopore sequencing generates long reads but has higher error rates. By integrating both platforms, a hybrid assembly approach enhances sequencing accuracy, gene prediction, and structural variant detection.

Hybrid assembly yielded a 2.1 Mb genome with 11 contigs, improving contiguity and completeness. Gene prediction and functional annotation identified 2,008 coding sequences, which were mapped to key metabolic pathways. Comparative genome analysis showed ~100% similarity with the reference strain *S. equi* subsp. *zooepidemicus* NCTC 7023, with six major genomic rearrangements detected.

Additionally, this study provides insights into the genetic regulation of hyaluronic acid (HA) biosynthesis, a key biomedical and industrial product. Precise localization of the hyaluronidase gene offers new possibilities for genetic modifications to enhance HA yield.

The findings emphasize the complementary strengths of Illumina and Nanopore sequencing in bacterial genomics. This research demonstrates the effectiveness of hybrid approaches in microbial genome assembly and has important implications for strain optimization in industrial HA production

Keywords: *Streptococcus equi* subsp. *zooepidemicus* MTCC 3523; Hyaluronic Acid; Oxford Nanopore; Illumina; Whole-genome sequencing

1. Introduction

Advancements in sequencing technologies have revolutionized the field of genomics, enabling precise and comprehensive characterization of bacterial genomes. Whole-genome sequencing (WGS) serves as a pivotal tool for understanding bacterial taxonomy, functional genomics, and evolutionary relationships(1). Hybrid sequencing

* Corresponding author: Seyedeh Zoha Tabatabaei

approaches, which integrate short-read and long-read sequencing technologies, have emerged as a transformative strategy for genome assembly and annotation, particularly for organisms with complex, repetitive, or previously unsequenced genomes(2).

Oxford Nanopore and Illumina sequencing are among the most widely employed platforms in hybrid genome assembly. Illumina sequencing, a second-generation technology, offers high-throughput short reads with exceptional accuracy, making it ideal for detecting single-nucleotide polymorphisms (SNPs), small insertions, and deletions. However, its short read lengths limit its capability to resolve repetitive regions and structural variations(3). In contrast, Oxford Nanopore sequencing, a fourth-generation technology, provides real-time long-read sequencing, enabling the resolution of complex genomic regions, structural variants, and complete plasmid structures. Despite its advantages, Nanopore sequencing has a higher intrinsic error rate, necessitating hybrid approaches that combine the accuracy of short reads with the extended read lengths of long-read platforms(4).

By leveraging the strengths of both technologies, hybrid de novo assembly facilitates the reconstruction of contiguous bacterial genomes with superior accuracy and completeness. This approach is particularly invaluable for sequencing bacterial species with unknown or highly variable genomic structures, allowing researchers to generate near-complete assemblies that faithfully represent the true genetic architecture of an organism(5).

One of the most extensively studied bacterial strains for industrial applications is *Streptococcus equi* subsp. *zooepidemicus* MTCC 3523. This strain is recognized for its exceptional ability to produce high-molecular-weight hyaluronic acid (HA), a polysaccharide with significant applications in medicine, cosmetics, and biotechnology(6). HA plays a crucial role in tissue hydration, wound healing, and joint lubrication, making it a key ingredient in dermatological treatments, osteoarthritis therapy, ophthalmic surgery, and advanced drug delivery systems(7). Large-size HA, characterized by its increased molecular weight, is particularly valued for its enhanced viscoelastic properties, prolonged retention time in biological tissues, and superior performance in medical and aesthetic applications. The molecular weight of HA significantly influences its biological functions, with high-molecular-weight HA exhibiting anti-inflammatory, immunomodulatory, and protective extracellular matrix properties(8).

Despite its industrial importance, HA production in *S. equi* subsp. *zooepidemicus* is hindered by the activity of hyaluronidase, an enzyme that catalyzes the depolymerization of HA into lower-molecular-weight fragments. Hyaluronidase activity not only reduces HA yield but also diminishes the quality of the final product by generating low-viscosity and biologically less stable HA fractions. To optimize HA production for industrial applications, it is essential to eliminate or suppress hyaluronidase activity, ensuring the retention of high-molecular-weight HA and maximizing product stability and efficacy(9).

The regulation of hyaluronidase expression and activity remains a major challenge in microbial HA production(10). Genetic engineering approaches aimed at knocking out or downregulating the hyaluronidase gene require precise knowledge of its genomic location and regulatory elements. However, previous studies have reported variability in the genomic organization of *S. equi* subsp. *zooepidemicus* strains, necessitating strain-specific genomic analysis to pinpoint the exact position of the hyaluronidase gene(11).

The objective of this study is to employ hybrid sequencing technologies to accurately determine the genomic location of the hyaluronidase gene in *S. equi* subsp. *zooepidemicus* MTCC 3523. By precisely mapping this gene, targeted mutagenesis strategies can be designed to inactivate hyaluronidase expression, thereby enhancing HA yield and promoting the production of high-molecular-weight HA. In addition to gene knockout strategies, this study aims to provide a comprehensive comparative genomic analysis of *S. equi* subsp. *zooepidemicus* MTCC 3523, assessing its phylogenetic relationships, genomic stability, and metabolic potential for industrial-scale HA synthesis. This research will contribute to optimizing microbial HA production, offering new insights into genetic modifications that can improve HA biosynthesis and establishing a foundation for future advancements in biopharmaceutical and cosmetic applications.

2. Material and methods

2.1. DNA Extraction and Quality Control

Genomic DNA from bacterial cells was extracted using the Qiagen DNeasy Blood and Tissue Kit (Cat No. 69506). The bacterial cell pellet was resuspended in lysozyme (10 mg/mL; Cat No. L6876) and incubated at 37°C for 30 minutes. Cell lysis was performed using AL buffer and Proteinase K at 56°C for 2 hours. RNase A treatment (50 mg/mL; Cat No. 2101076) was carried out at 65°C for 30 minutes. The lysate was mixed thoroughly with half the volume of absolute ethanol and loaded onto a DNeasy Mini Spin Column placed in a 2 mL collection tube. Following centrifugation at 8,000

rpm, the flow-through was discarded. Subsequent wash steps were conducted according to the manufacturer's protocol. DNA was eluted with 10 mM Tris-HCl (pH 8.0). The concentration and purity of genomic DNA were assessed using a Nanodrop Spectrophotometer (Thermo Scientific 2000) and quantified with the Qubit dsDNA HS Assay Kit (Cat No. Q32854).

2.2. Sanger Sequencing

To confirm bacterial strain purity, 16S rRNA gene amplification was performed. PCR amplification was carried out using approximately 30 ng of genomic DNA as a template, 16S rDNA primers (27F and 1492R), and Takara Ex Taq Polymerase (Cat No. RR001A) in a 25 µL reaction volume. The amplification of a ~1.5 kb fragment was confirmed via agarose gel electrophoresis. Purified PCR products were sequenced using Sanger sequencing. BLAST analysis identified the sample *MTCC3523* as *Streptococcus sp.* 16S ribosomal RNA, with a sequence identity match of 99.77%.

2.3. Illumina Library Preparation and Sequencing

Library preparation was performed using the QIASeq FX DNA Library Preparation Kit (Cat No. 180475), following the manufacturer's instructions. Approximately 50 ng of Qubit-quantified DNA was enzymatically fragmented, end-repaired, and A-tailed in a one-tube reaction using the FX Enzyme Mix. Adapter ligation was carried out using index-incorporated Illumina adapters to generate sequencing libraries. The libraries were enriched through six cycles of indexing PCR (initial denaturation at 98°C for 2 min, followed by 6 cycles of 98°C for 20 sec, 60°C for 30 sec, and 72°C for 30 sec, with a final extension at 72°C for 1 min). The amplified libraries were purified using Cytiva beads (Magbio, #AC-60050), quantified with a Qubit fluorometer (Thermo Fisher Scientific, MA, USA), and assessed for fragment size distribution using an Agilent 2200 TapeStation. The libraries were paired-end sequenced on an Illumina NovaSeq X Plus sequencer (Illumina, San Diego, USA) for 150 cycles.

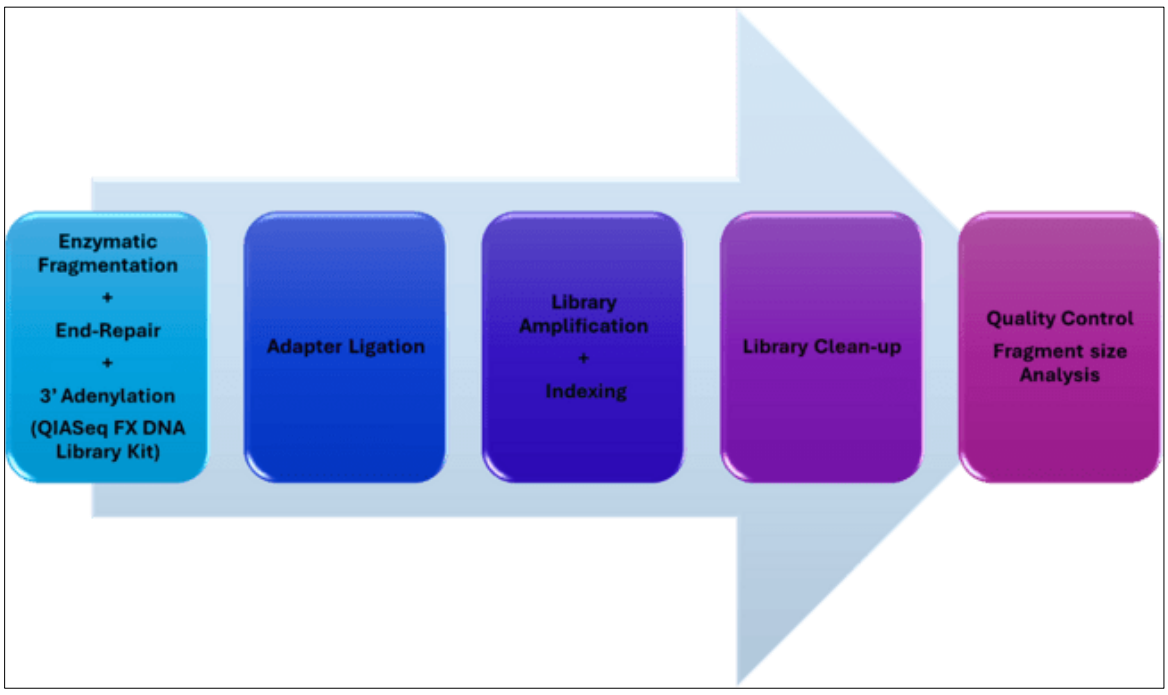

Figure 1 Workflow diagram illustrating the QIASeq FX DNA library preparation process. The protocol begins with enzymatic fragmentation, end-repair, and 3' adenylation of DNA using the QIASeq FX DNA Library Kit. This is followed by adapter ligation, which enables sequencing compatibility. The next step involves library amplification and indexing to ensure sample traceability. A subsequent library clean-up step removes unwanted fragments and impurities. Finally, a quality check and fragment size analysis are performed to verify library integrity before proceeding with sequencing. This streamlined approach ensures high-quality DNA libraries suitable for next-generation sequencing applications

2.4. Nanopore Library Preparation and Sequencing

Genomic DNA was end-polished and A-tailed using the NEBNext Ultra II End Repair Kit (New England Biolabs, MA, USA). The end-prepared DNA was barcoded using the Blunt TA Ligase Master Mix (M0367L), followed by equimolar pooling

of barcoded samples. Sequencing adapters (SQK-LSK114.96) were ligated to double-stranded DNA fragments using NEB Quick T4 DNA Ligase (New England Biolabs, MA, USA). After purification with AMPure XP beads, the final library was quantified using Qubit and subsequently sequenced on the Nanopore PromethION system (PromethION P24, Oxford Nanopore Technologies, UK) using a PromethION Flow Cell (FLO-PRO114M).

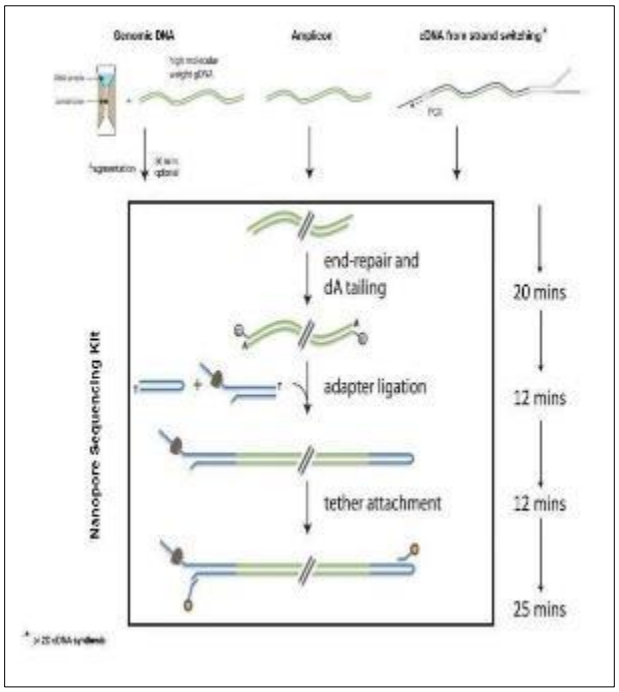

Figure 2 Schematic representation of the Nanopore sequencing library preparation workflow. The process begins with high-molecular-weight genomic DNA, amplicons, or cDNA generated via strand-switching PCR. The DNA undergoes optional fragmentation (30 min), followed by end-repair and dA-tailing (20 min). Adapter ligation is then performed (12 min), allowing for the attachment of sequencing-compatible adapters. Subsequently, tether attachment occurs (12 min), ensuring efficient loading onto the Nanopore sequencing flow cell. The final step involves sequencing-ready molecules being loaded onto the platform for real-time long-read sequencing (25 min). This workflow enables high-throughput sequencing of native or amplified DNA and RNA molecules.(Reprinted with permission from 26)

The bacterial strain *Streptococcus equi* subsp. *zooepidemicus* MTCC 3523 was utilized for sequencing analysis. The extracted DNA sample was processed using a barcoded sequencing approach to ensure accurate tracking and high-quality sequencing output. The component sequence assigned to this sample was NB01 CACAAAGACACCGACAACTTTCTT, which played a crucial role in identifying the sample during sequencing and data processing.

The DNA quantification was carried out using a Qubit fluorometer, which measured a Qubit yield of 1.23 ng and a total DNA yield of 24.6 ng, indicating sufficient DNA concentration for downstream sequencing applications. To facilitate indexing and multiplex sequencing, unique dual-index barcode sequences were assigned to this sample. Barcode 1 index sequence was UDI042 TGTGGAACCG, while Barcode 2 index sequence was UDI042 CATTGACTCT. These barcodes ensured precise demultiplexing and accurate identification of sequencing reads associated with *S. equi* subsp. *zooepidemicus* MTCC 3523, contributing to robust and high-fidelity genome assembly.

2.5. Data Analysis

A total of ~1.7 million paired-end Illumina sequencing reads and ~0.3 GB of Nanopore long-read data were generated. Illumina raw reads underwent adapter trimming and quality filtering using TrimGalore (v0.4.01)(12), while Nanopore reads were processed using Porechop (v0.2.32)(13) for adapter removal. Hybrid genome assembly was performed using Unicycler (v0.4.83)(14), integrating both short- and long-read data. The final draft assembly resulted in an ~2.1 Mb genome comprising 11 contigs.

Gene and protein predictions were conducted using Prokka (v1.14)(15), and functional annotations were assigned using DIAMOND BlastP(16) against the UniProt bacterial database. Metabolic pathways were identified using KAAS (KEGG Automatic Annotation Server)(17). DNA-DNA hybridization-based phylogenetic analysis was performed using the Type (Strain) Genome Server (TYGS)(18). Circular genome comparison with the reference strain *Streptococcus equi* subsp. *zooepidemicus* NCTC 7023 was carried out using BRIG (v0.95)(19), and genome synteny analysis was performed using Mauve (v2.4.0)(20). A detailed bioinformatics workflow is provided in Figure 3.

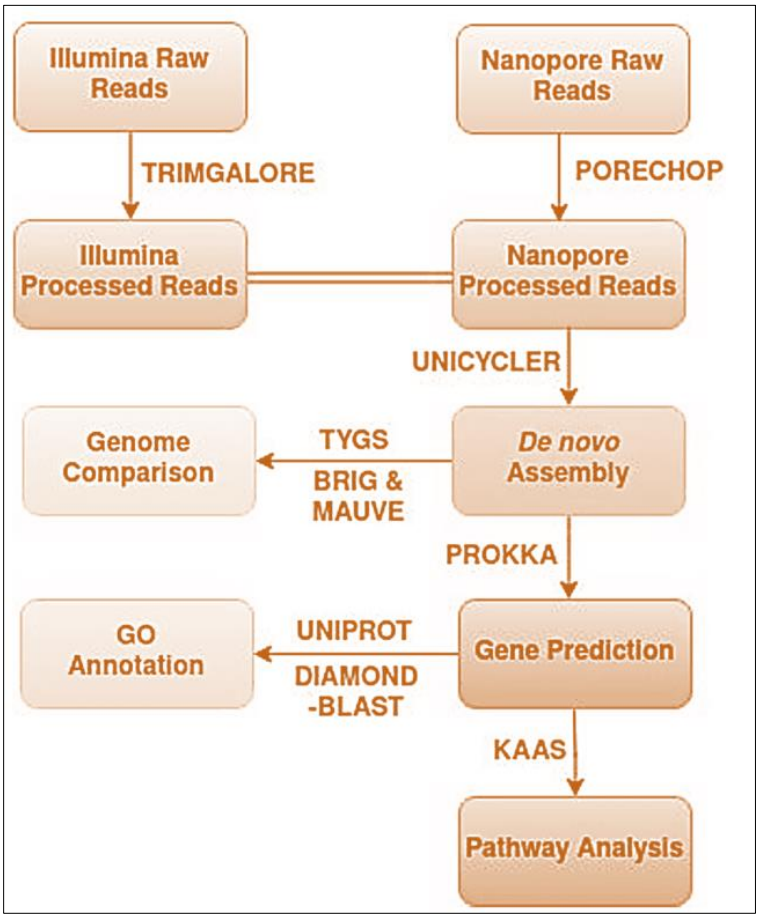

Figure 3 Bioinformatics workflow for genome assembly, annotation, and functional analysis of *Streptococcus equi* subsp. *zooepidemicus* MTCC3523. Illumina and Nanopore raw reads are processed using TrimGalore and Porechop, respectively. The processed reads are merged and assembled using Unicycler for de novo genome assembly. Gene prediction is performed with Prokka, followed by functional annotation using UNIPROT and DIAMOND-BLAST for Gene Ontology (GO) annotation. KAAS is used for pathway analysis, while genome comparison is conducted using TYGS, BRIG, and MAUVE

3. Results

3.1. DNA Quality Control

Quality assessment confirmed optimal DNA yield and concentration for Illumina and Nanopore library preparation. For the quality control and sequencing assessment of *S. equi* subsp. *zooepidemicus* MTCC 3523, several key metrics were evaluated. The DNA concentration, as measured by Nanodrop, was 15.5 ng/µL, with an absorbance ratio of 1.67 for 260/280 nm and 0.5 for 260/230 nm, indicating moderate purity. The total DNA volume used for sequencing was 25 µL, yielding 387.5 ng of extracted DNA.

Further quantification using Qubit fluorometry showed a concentration of 4.52 ng/µL, with a total volume of 25 µL, resulting in a final Qubit-based DNA yield of 113 ng. The quality control (QC) assessment classified the purity as sub-optimal, but the overall DNA yield was deemed admissible for sequencing applications.

For Illumina sequencing, the total number of reads generated was 1,703,972, out of which 1,669,374 reads were successfully processed after quality filtering. This resulted in a high sequencing coverage of 232X, ensuring robust genome assembly and downstream analysis. These results confirm the adequacy of the extracted DNA for whole-genome sequencing and support high-confidence data interpretation.

3.2. Illumina Library QC and Data Demultiplexing

The sequencing library exhibited an average fragment size of 451 bp, with sufficient concentration for sequencing. Data from the sequencing run was demultiplexed using Bcl2Fastq (v2.20), and FastQ files were generated based on unique dual barcodes. The sequencing quality was assessed using Commander software (Figure 4).

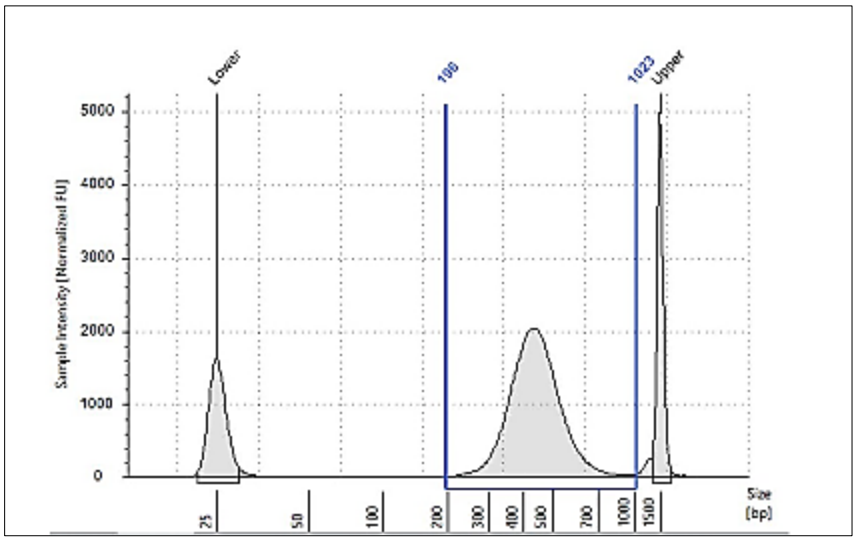

Figure 4 TapeStation electropherogram of *Streptococcus equi* subsp*. zooepidemicus* MTCC3523. The profile displays DNA fragment sizes ranging from 196 bp to 1023 bp, with an average fragment size of 451 bp. The DNA concentration is measured at 21.0 ng/µL, with a regional molarity of 74.7 nmol/L, accounting for 96.43% of the total DNA content. The lower and upper markers indicate size standards used for accurate fragment analysis

3.3. Primary Data Analysis

Approximately 1.7 million paired-end Illumina reads and 0.3 GB of Nanopore long reads were obtained. Illumina raw reads were processed using TrimGalore for quality filtering (Phred score <30, length <20 bp), while Nanopore reads were trimmed using Porechop.

3.4. Secondary Data Analysis

3.4.1. De Novo Hybrid Assembly

Hybrid genome assembly was performed using Unicycler, integrating both Illumina and Nanopore reads. The assembled draft genome contained 11 contigs with a total size of ~2.1 Mb.

For the sequencing analysis of *S. equi* subsp. *zooepidemicus* MTCC 3523, raw reads were generated to facilitate comprehensive genome assembly and annotation. A total of 836,174 reads were obtained, contributing to a total base count of 0.34 GB. The sequencing yielded an average read length of 401.1 base pairs, with a mean quality score of 12, ensuring a high proportion of reliable sequences. Notably, 99.90% of the reads had a quality score exceeding Q7, indicating minimal sequencing errors.

The sequencing coverage was estimated at 152.47X, providing sufficient depth for an accurate genome reconstruction. The assembly process resulted in the generation of 11 contigs, with the largest contig spanning 1.6 MB and the smallest contig measuring 119 base pairs. The total length of assembled contigs amounted to 2.1 MB, forming a near-complete draft genome of the bacterium.

An essential quality metric of the assembly, the N50 value, was recorded at 1.6 MB, signifying a well-assembled genome with long contiguous sequences. No non-ATGC characters were detected in the assembled genome, ensuring the

integrity and completeness of the sequence data. Furthermore, the assembly included five contigs exceeding 10 Kbp, with one contig surpassing 1 Mbp, highlighting the robustness of the sequencing strategy in capturing large genomic fragments.

These sequencing results provide a strong foundation for further genomic analysis, including functional annotation, metabolic pathway reconstruction, and comparative genomics with closely related strains.

3.4.2. Gene Prediction and GO Annotation

Gene prediction was performed using Prokka, identifying 2,008 predicted proteins. Gene ontology (GO) annotation was carried out using DIAMOND BlastP against the UniProt bacterial database, considering proteins with a minimum identity threshold of 30% as significant hits (Figure 5).

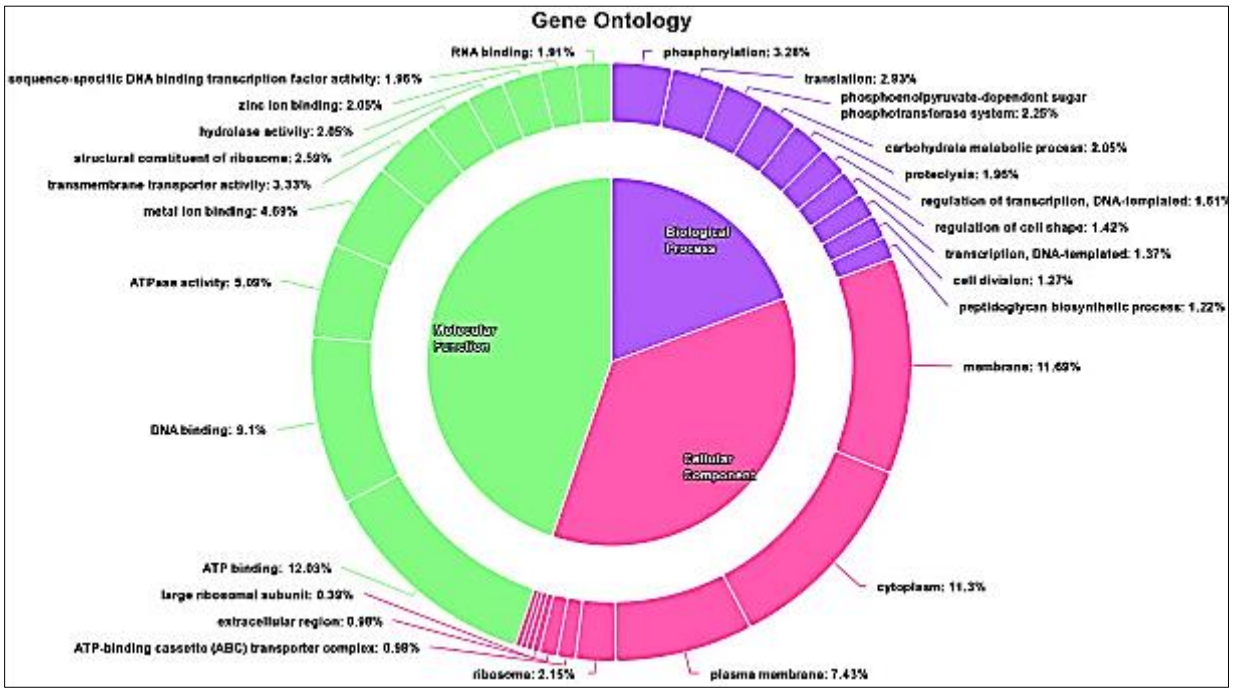

Figure 5 Gene Ontology (GO) classification of annotated genes in *Streptococcus equi* subsp*. zooepidemicus* MTCC3523. The circular diagram represents the distribution of genes across three major GO categories: Molecular Function (green), Biological Process (purple), and Cellular Component (pink). Key molecular functions include ATP binding (12.03%) and DNA binding (9.1%), while significant biological processes include phosphorylation (3.28%) and translation (2.93%). The cellular component category highlights the predominance of membrane-associated (11.69%) and cytoplasmic (11.3%) genes, indicating their functional significance in this strain

3.4.3. Pathway Analysis

Pathway annotation was conducted using KAAS. The predicted genes were mapped to KEGG pathways, with reference datasets including *Streptococcus equi* subsp. *zooepidemicus* MGCS10565 and H70. The top 10 KEGG pathway entries are shown in Figure 6.

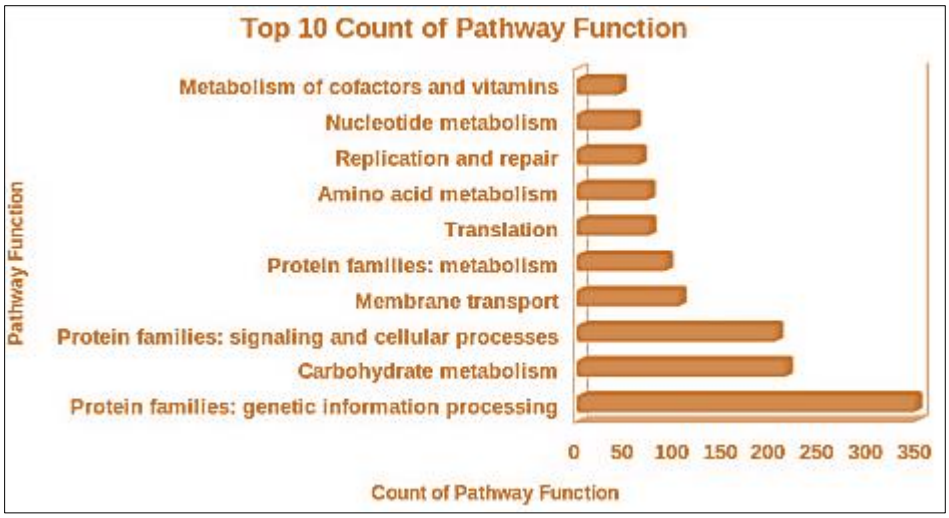

Figure 6 Top 10 pathway functions based on gene annotation in *Streptococcus equi* subsp*. zooepidemicus* MTCC3523. The bar chart displays the count of pathway functions categorized into major metabolic and cellular processes. The most represented category is "Protein families: genetic information processing," followed by "Carbohydrate metabolism" and "Protein families: signaling and cellular processes," indicating the key biological pathways active in this strain

3.4.4. DNA-DNA Hybridization Analysis

Phylogenetic analysis was conducted using TYGS. Whole-genome comparisons indicated that sample *MTCC3523 i*s closely related to *Streptococcus equi* subsp. *zooepidemicus* NCTC 7023, with a DNA-DNA hybridization (dDDH) sequence identity of 100%. The whole-genome phylogenetic tree is shown in Figure 7.

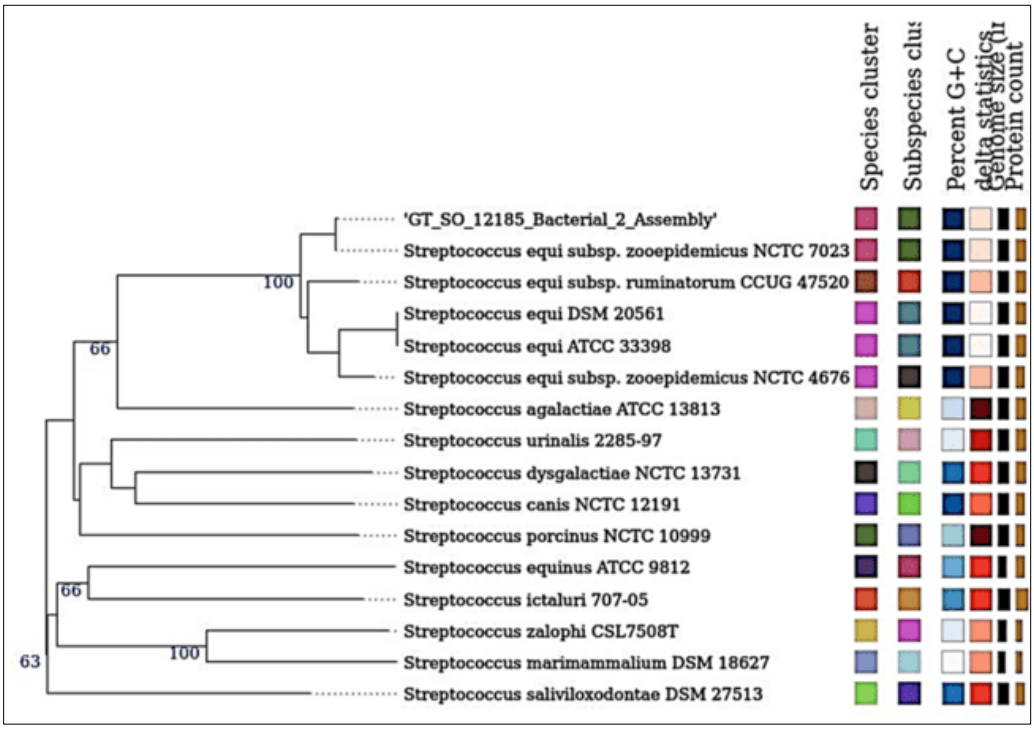

Figure 7 Phylogenetic tree of Streptococcus species, including *Streptococcus equi* subsp*. zooepidemicus* and related taxa. The tree was constructed based on genomic sequence data, with bootstrap values indicated at key nodes. The annotation on the right providesinformation on species clusters, subspecies classification, GC content, and protein count. The placement of MTCC3523 within the S. equi cluster suggests its phylogenetic relationship to other strains

3.4.5. Comparative Genome Analysis

Circular genome comparisons were performed using BRIG, highlighting sequence similarities with *S. equi* subsp. *zooepidemicus* NCTC 7023. Synteny analysis using Mauve identified potential genomic rearrangements or recombination events (Figures 8a and 8b).

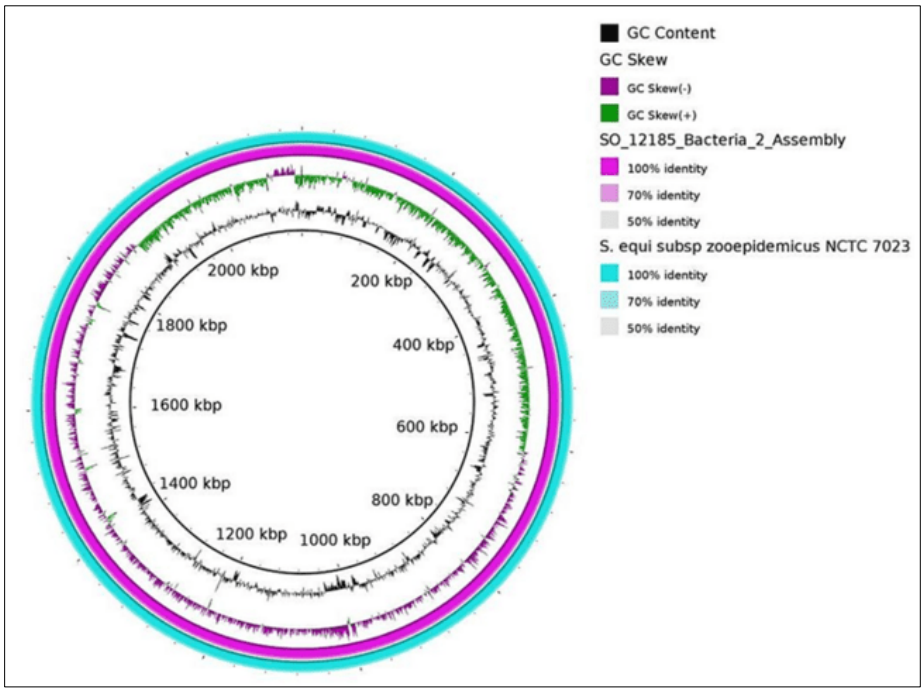

Figure 8 Circular genome comparison of *Streptococcus equi* subsp. *zooepidemicus* MTCC3523 with *S. equi* subsp. *zooepidemicus* NCTC 7023. The innermost track represents GC content (black), followed by GC skew (purple for negative skew and green for positive skew). T The outer rings display the sequence identity between the assembled genome MTCC3523 and the reference genome (*S. equi* subsp. *zooepidemicus* NCTC 7023) at different identity thresholds (100%, 70%, and 50%). The high sequence similarity suggests a close phylogenetic relationship between the two strains

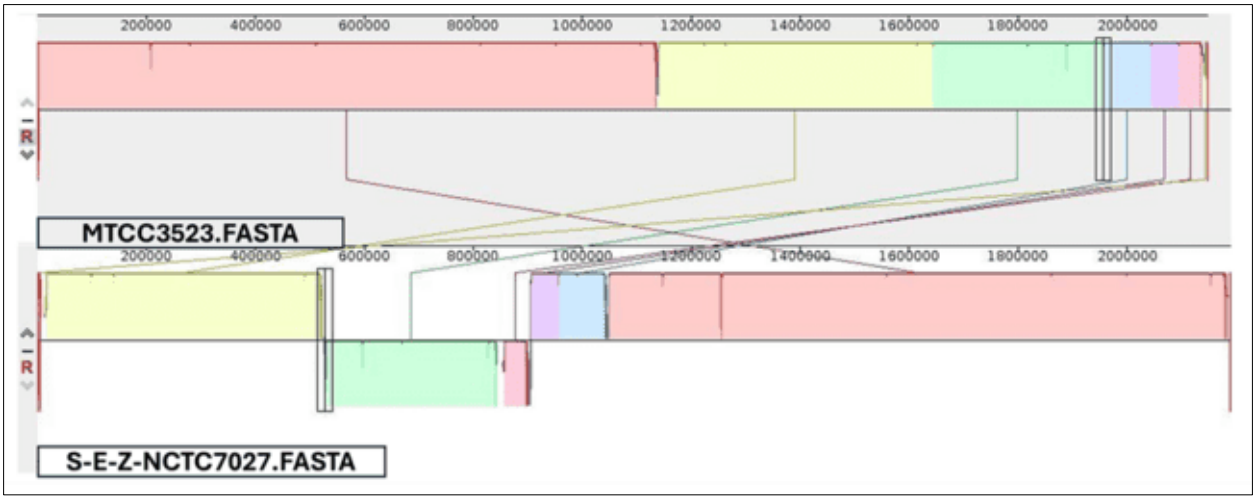

Figure 9 Whole-genome alignment of *Streptococcus equi* subsp. *zooepidemicus* MTCC3523 (MTCC3523.FASTA) and *S. equi* subsp. *zooepidemicus* NCTC 7027 (S-E-Z-NCTC7027.FASTA) using the MAUVE alignment tool. Colored blocks represent locally collinear sequence regions, with homologous segments preserved between the genomes. The presence of rearrangements, inversions, and gaps indicates structural variations between the two strains. Lines connecting the blocks highlight syntenic regions, with crossing lines showing genome rearrangements or inversions

4. Discussion

This study presents a comprehensive genomic analysis of *Streptococcus equi* subsp. *zooepidemicus* using a combination of Illumina short-read and Nanopore long-read sequencing technologies. The hybrid assembly approach provided a high-resolution draft genome with enhanced contiguity and completeness, facilitating downstream functional annotation and comparative genomics.

Oxford Nanopore sequencing technology is a fourth-generation sequencing platform that enables the real-time generation of long-read sequences. It utilizes nanopore-based electrical sensing, where single-stranded DNA passes through a protein nanopore embedded in a membrane, and the disruption of ionic current is measured to determine the nucleotide sequence. This technology provides several advantages, including the ability to generate ultra-long reads (exceeding 100 kb), real-time sequencing, and portability. These features make it particularly useful for resolving complex genomic regions, structural variations, and hybrid genome assemblies. However, Nanopore sequencing has a higher base-calling error rate compared to short-read sequencing technologies, necessitating hybrid approaches for accurate genome reconstruction(23).

Illumina sequencing, in contrast, is a second-generation sequencing technology that produces high-throughput, short-read sequences with high accuracy. It is based on sequencing-by-synthesis, where fluorescently labeled nucleotides are incorporated into DNA strands and detected in a cyclic manner. This technology is widely used for whole-genome sequencing, transcriptomics, and metagenomics due to its low error rate and cost-effectiveness. However, its short read lengths (typically 150–300 bp) limit its ability to resolve repetitive sequences and structural variations. Combining Illumina and Nanopore sequencing allows for the benefits of both technologies—high accuracy from short reads and improved genome assembly from long reads(24).

Hybrid de novo assembly is particularly valuable for bacterial genomes where the exact sequence is unknown. By integrating short, high-accuracy Illumina reads with long, more error-prone Nanopore reads, hybrid assembly methods enable the reconstruction of complete bacterial genomes with greater precision. This approach is essential for identifying novel bacterial strains, characterizing uncultured microbes, and resolving complex genomic structures such as repeat regions, plasmids, and mobile genetic elements. In cases where the bacterial genome has not been previously sequenced, hybrid assembly provides a reliable framework for constructing contiguous genomes that are more representative of the true genetic architecture(25).

The successful extraction and quality assessment of genomic DNA enabled the generation of high-quality sequencing data. The sequencing metrics indicated optimal library preparation and sequencing performance, with sufficient coverage to support hybrid assembly. The Illumina sequencing yielded approximately 1.7 million paired-end reads, while Nanopore sequencing contributed ~0.3 GB of long-read data. The integration of these datasets in Unicycler resulted in an assembled genome of approximately 2.1 Mb comprising 11 contigs, representing a significant improvement in assembly quality compared to short-read-only approaches.

Functional annotation identified 2,008 protein-coding genes, highlighting key metabolic and cellular processes. Gene ontology (GO) and KEGG pathway analyses revealed the presence of genes associated with carbohydrate metabolism, virulence factors, and antibiotic resistance, which may have implications for bacterial adaptability and pathogenicity. The KEGG pathway analysis further confirmed similarities to *S. equi* subsp. *zooepidemicus* reference strains, supporting the taxonomic classification of the isolate.

The whole-genome phylogenetic analysis and DNA-DNA hybridization confirmed that the bacterial strain is closely related to *S. equi* subsp. *zooepidemicus* NCTC 7023, with a high sequence identity of 100%. Comparative genome analysis using BRIG and Mauve revealed conserved genomic regions and synteny with the reference strain, suggesting minimal genomic rearrangements. This finding provides insight into the genetic stability of this subspecies and its evolutionary relationship within the *Streptococcus equi* lineage.

In addition to taxonomic classification, the study highlights the functional potential of the identified genes. The presence of genes involved in antibiotic resistance mechanisms is particularly significant, as it suggests a potential role in bacterial survival and adaptation under selective pressure. Further functional validation of these genes could help elucidate their exact roles and contribution to bacterial pathogenicity. Additionally, the identification of genes associated with carbohydrate metabolism indicates the ability of *S. equi* subsp. *zooepidemicus* to utilize diverse nutrient sources, which may be a crucial factor in its colonization and persistence within various host environments.

A key industrial and biomedical feature of *S. equi* subsp. *zooepidemicus* is its ability to produce hyaluronic acid (HA), a high-molecular-weight polysaccharide with applications in medicine(27), cosmetics, and biotechnology(21,22). The HA biosynthesis pathway in *S. equi* subsp. *zooepidemicus* is well characterized and involves the enzymes hyaluronan synthase (HasA), UDP-glucose dehydrogenase (HasB), and UDP-glucose pyrophosphorylase (HasC). These genes are responsible for the polymerization of HA from precursor sugars, and their presence in the assembled genome supports the strain's potential for HA production. Given the increasing demand for microbial HA synthesis as a sustainable alternative to animal-derived sources, further exploration of the regulatory mechanisms governing HA biosynthesis in *S. equi* subsp. *zooepidemicus* could provide valuable insights for industrial applications.

In contrast, *S. equi* subsp. *zooepidemicus* also produces hyaluronidase, an enzyme responsible for the degradation of hyaluronic acid. Hyaluronidase functions by cleaving HA into smaller fragments, influencing bacterial invasion, host immune evasion, and biofilm formation. The balance between HA synthesis and degradation plays a crucial role in bacterial virulence and adaptability, as excessive hyaluronidase activity can lead to reduced HA production, potentially limiting its industrial applications. In this study, we found the exact location of the hyaluronidase gene, which will help in targeted mutagenesis to reduce its activity, thereby enhancing HA yield and facilitating the production of longer HA chains. Understanding the regulatory mechanisms that control hyaluronidase expression in *S. equi* subsp. *zooepidemicus* is essential for optimizing HA production while mitigating unwanted degradation. Genetic modifications or targeted inhibition of hyaluronidase activity could provide a promising strategy to enhance microbial HA yields for commercial applications.

The findings of this study also contribute to understanding the evolutionary trajectory of *S. equi* subsp. *zooepidemicus*. Comparative genomic analyses with closely related strains indicate that horizontal gene transfer (HGT) and genomic rearrangements may have played a role in shaping the genome of this subspecies. This supports the hypothesis that bacterial adaptation to diverse ecological niches is facilitated by genomic plasticity. Future studies could further investigate the extent of HGT events and their functional significance in different bacterial populations.

Overall, the study successfully demonstrates the application of hybrid sequencing technologies for bacterial genome characterization, providing a robust framework for future investigations into the functional and comparative genomics of *S. equi* subsp. *zooepidemicus*. Future work may focus on further elucidating the functional roles of predicted genes, exploring strain-specific variations, and investigating potential virulence determinants contributing to pathogenicity and host interactions. Furthermore, whole-genome comparative analyses with pathogenic and commensal *Streptococcus* strains may provide deeper insights into the genomic features that define virulence and host specificity. Expanding the dataset with additional bacterial isolates from diverse environmental sources would enhance our understanding of genomic diversity and evolutionary dynamics within this bacterial species.

5. Conclusion

This study highlights the advantages of hybrid whole-genome sequencing using Illumina and Nanopore platforms for high-resolution genomic analysis. By combining the high accuracy of Illumina sequencing with the long-read capability of Nanopore technology, we achieved a more complete and contiguous assembly of *S. equi* subsp. *zooepidemicus* MTCC 3523. The findings provide deeper insights into bacterial functional genomics, particularly in the regulation of HA biosynthesis, paving the way for targeted genetic modifications to enhance HA yield. The results of this research hold significant implications for industrial biotechnology, offering a robust foundation for strain optimization in pharmaceutical and cosmetic applications. Future studies should focus on refining sequencing accuracy, improving genome assembly tools, and exploring advanced bioengineering strategies for microbial strain enhancement.

Compliance with ethical standards

Acknowledgments

We sincerely appreciate the invaluable support provided by Genotypic Company(28). We also extend our heartfelt gratitude to Professor Mehdi Totonchi for his scientific guidance and exceptional leadership throughout this project. Additionally, we thank Shafagh NewGene Medical Company for their funding and facilitation, which made this research possible.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

Statement of ethical approval

This project was approved by the Research Ethics Committee of Islamic Azad University, Tehran Medical Sciences University – Faculty of Pharmacy and Pharmaceutical Branches, under the ethical code IR.IAU.PS.REC.1402.621.

References

[1] Bagger FO, Borgwardt L, Jespersen AS, Hansen AR, Bertelsen B, Kodama M, Nielsen FC. Whole genome sequencing in clinical practice. BMC Med Genomics. 2024 Jan 29;17(1):39. doi: 10.1186/s12920-024-01795-w. PMID: 38287327; PMCID: PMC10823711.

[2] Mihai Pop, Genome assembly reborn: recent computational challenges, Briefings in Bioinformatics, Volume 10, Issue 4, July 2009, Pages 35366, https://doi.org/10.1093/bib/bbp026

[3] Lerminiaux N, Fakharuddin K, Mulvey MR, Mataseje L. Do we still need Illumina sequencing data? Evaluating Oxford Nanopore Technologies R10.4.1 flow cells and the Rapid v14 library prep kit for Gram negative bacteria whole genome assemblies. Can J Microbiol. 2024 May 1;70(5):178-189. doi: 10.1139/cjm-2023-0175. Epub 2024 Feb 14. PMID: 38354391.

[4] Pugh J. The Current State of Nanopore Sequencing. Methods Mol Biol. 2023;2632:3-14. doi: 10.1007/978-1-0716-2996-3_1. PMID: 36781717.

[5] Sharon BM, Hulyalkar NV, Nguyen VH, Zimmern PE, Palmer KL, De Nisco NJ. Hybrid De Novo Genome Assembly for the Generation of Complete Genomes of Urinary Bacteria using Short- and Long-read Sequencing Technologies. J Vis Exp. 2021 Aug 20;(174). doi: 10.3791/62872. PMID: 34487123.

[6] Martins VB, Moreira TD, da Silva Júnior AH, Sayer C, de Mello JM, Immich AP. Brewing Industry By-products: An Innovative Alternative to Hyaluronic Acid Biosynthesis. Current Analytical Chemistry. 2025 Jan 6.

[7] Shukla P, Srivastava P, Mishra A. On the potential activity of hyaluronic acid as an antimicrobial agent: experimental and computational validations. Bioprocess and Biosystems Engineering. 2025 Jan;48(1):27-42.

[8] Contreras Mendoza J, Arriola Guevara E, Suarez Hernández LA, Toriz G, Guatemala-Morales GM, Corona-González RI. Evaluation of mango residues to produce hyaluronic acid by Streptococcus zooepidemicus. Folia Microbiologica. 2024 Aug;69(4):847-56.

[9] Harth ML, Furlan FF, Horta AC. Microbial hyaluronic acid production in the 21 century: a roadmap toward high production, tailored molecular weight. OBSERVATÓRIO DE LA ECONOMÍA LATINOAMERICANA. 2024 Mar 27;22(3):e3913-.

[10] Bakar A, Mukhriza TM. Microbial Hyaluronic Acid Production: A Comprehensive Review of Strategies, Challenges and Sustainable Approaches. Jurnal Serambi Engineering. 2024 Jun 7;9(2):9138-50.

[11] Abdullah Thaidi NI, Mohamad R, Wasoh H, Kapri MR, Ghazali AB, Tan JS, Rios-Solis L, Halim M. Development of in situ product recovery (ISPR) system using amberlite IRA67 for enhanced biosynthesis of hyaluronic acid by streptococcus zooepidemicus. Life. 2023 Feb 16;13(2):558.

[12] Babraham Bioinformatics. (n.d.). Trim Galore! Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

[13] Wick, R. (n.d.). Porechop. Available at: https://github.com/rrwick/Porechop.

[14] Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13(6):e1005595. Published 2017 Jun doi:10.1371/journal.pcbi.1005595

[15] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. doi: 10.1093/bioinformatics/btu153. Epub 2014 Mar 18. PMID: 24642063

[16] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59-60. doi: 10.1038/nmeth.3176. Epub 2014 Nov 17. PMID: 25402007.

[17] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W182-5. doi: 10.1093/nar/gkm321. Epub 2007 May 25. PMID: 17526522; PMCID: PMC1933193.

[18] Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-ofthe-art genome-based taxonomy. Nat Commun. 2019;10(1):2182. Published 2019 May 16. doi:10.1038/s41467-019-10210-3

[19] Alikhan, NF., Petty, N.K., Ben Zakour, N.L. et al. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12, 402 (2011). https://doi.org/10.1186/1471-2164-12-402

[20] Darling, A. C., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research, 14(7), 1394–1403. https://doi.org/10.1101/gr.2289704

[21] Beheshtizadeh N, Mohammadzadeh M, Mostafavi M, Seraji AA, Ranjbar FE, Tabatabaei SZ, Ghafelehbashi R, Afzali M, Lolasi F. Improving hemocompatibility in tissue-engineered products employing heparin-loaded nanoplatforms. Pharmacological Research. 2024 Jun 19:107260.

[22] Beheshtizadeh N, Amiri Z, Tabatabaei SZ, Seraji AA, Gharibshahian M, Nadi A, Saeinasab M, Sefat F, Kolahi Azar H. Boosting antitumor efficacy using docetaxel-loaded nanoplatforms: from cancer therapy to regenerative medicine approaches. Journal of Translational Medicine. 2024 May 30;22(1):520.

[23] Heikema AP, Horst-Kreft D, Boers SA, Jansen R, Hiltemann SD, de Koning W, Kraaij R, de Ridder MA, van Houten CB, Bont LJ, Stubbs AP. Comparison of illumina versus nanopore 16S rRNA gene sequencing of the human nasal microbiota. Genes. 2020 Sep 21;11(9):1105.

[24] Stefan CP, Hall AT, Graham AS, Minogue TD. Comparison of illumina and Oxford nanopore sequencing technologies for pathogen detection from clinical matrices using molecular inversion probes. The Journal of Molecular Diagnostics. 2022 Apr 1;24(4):395-405.

[25] Cook R, Brown N, Rihtman B, Michniewski S, Redgwell T, Clokie M, Stekel DJ, Chen Y, Scanlan DJ, Hobman JL, Nelson A. The long and short of it: Benchmarking viromics using Illumina, Nanopore and PacBio sequencing technologies. Microbial genomics. 2024 Feb 29;10(2):001198.

[26] Cummings PJ, Olszewicz J, Obom KM. Nanopore DNA sequencing for metagenomic soil analysis. Journal of Visualized Experiments: JoVE. 2017 Dec 14(130):55979.

[27] Tabatabaei, S. Z., Motevaseli, E., Samiee Zafarghandi, Z., Tabatabaei, S. A. Comparison of the Overexpression of HOTAIR lncRNAs and the Downregulation of HOXD10 Between Familial and Sporadic Coronary Artery Diseases. *Iranian Heart Journal*, 2021; 22(4): 127-134.

[28] Genotypic Technology Pvt. Ltd. (2023) 'NGS Sequencing and Genome informatics services', Available at: https://www.genotypic.co.in (Accessed: 12 October 2023).