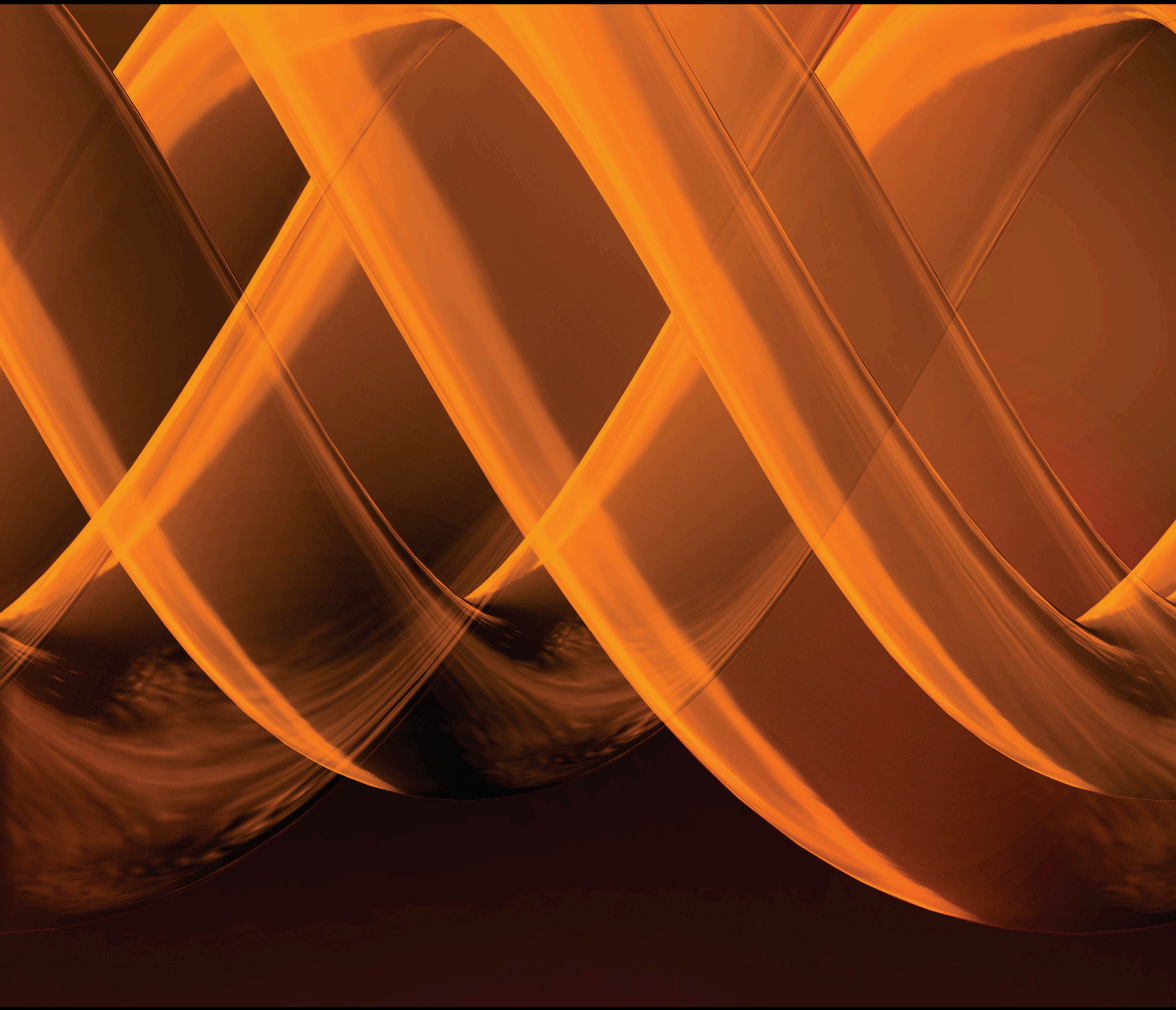
International Journal of Genomics

From Microbial Genomics to Metagenomics

Lead Guest Editor: Ravi Kant
Guest Editors: Abhishek Kumar and Tarja Sironen



From Microbial Genomics to Metagenomics

International Journal of Genomics

From Microbial Genomics to Metagenomics

Lead Guest Editor: Ravi Kant
Guest Editors: Abhishek Kumar and Tarja Sironen

Contents

Hindawi
International Journal of Genomics
Volume 2020, Article ID 9357450, 2 pages
https://doi.org/10.1155/2020/9357450

Editorial

From Microbial Genomics to Metagenomics

Ravi Kant,[1,2] Abhishek Kumar,[3,4] and Tarja Sironen[1,2]

[1]*Department of Veterinary Biosciences, Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland*
[2]*Department of Virology, Faculty of Medicine, University of Helsinki, Helsinki, Finland*
[3]*Institute of Bioinformatics, International Technology Park, Bangalore 560066, India*
[4]*Manipal Academy of Higher Education (MAHE), Manipal, 576104 Karnataka, India*

Correspondence should be addressed to Ravi Kant; ravi.kant@helsinki.fi

Received 8 May 2020; Accepted 9 May 2020; Published 1 June 2020

In the last 15 years, rapid application of next-generation sequencing (NGS) technologies has revolutionized the practice of microbial science. NGS has provided an unprecedented view on microbial diversity, and NGS technologies have allowed us to investigate complex microbial ecosystems, such as the human gastrointestinal (GI) microbiota, which consists of over three million genes from mainly Gram-positive bacteria [1–3].

Genomic comparisons of different bacterial genera and species have helped reveal the evolutionary origins of virulence and niche specification. Comparative analyses that compile the genomes of different strains from the same or different species (into what is called a "pangenome") have revealed that the gene content within an entire species or genera is much more than that of a single strain or species. Moreover, this sort of study has aided the understanding of one of the dominant genetic forces behind bacterial evolution, specifically the concept of lateral gene transfer between microorganisms [4, 5].

Steady advances in sequencing technologies have allowed us to elucidate the genetics of microbial interactions, for example, through comparative metagenomic and metatranscriptomic analyses of bacterial communities. Metagenomics is one of the most rapidly growing fields in the microbial sciences. A metagenomic approach provides an extraordinary view of the diverse microbial world in different environments, such as in human and animal body sites, marine and other water bodies, soil, and air. Metagenomic and metatranscriptomic analyses of bacterial communities can reveal the genetics of microbial interactions and have been widely used by researchers in diverse disciplines such as ecology, energy, agriculture, biotechnology, and medicine [6]. As a rapidly growing field, comparative genomics and metagenomics also present many challenges that must to be addressed.

For our special issue, we received several manuscripts and, through rigorous review, selected five for publication. In "*Streptococcus halichoeri*: Comparative Genomics of an Emerging Pathogen," K. Aaltonen et al. performed whole-genome sequencing of 20 different strains of an emerging pathogen, *S. halichoeri*, using the Illumina MiSeq platform and performed annotation using an automatic annotation pipeline RAST. The authors performed extensive all-against-all comparisons to unravel the core and pangenomes. Their findings highlight that *S. halichoeri* is a highly variable species with several virulence factors that indicate potential for significant pathogenicity. They also observed very little host species-specific markers in the genomes but instead observed a loose clustering according to species, as though adaptation is still incomplete. This suggests that the host switches into dogs, humans, and fur animals were rather recent and ongoing and possibly coincided with the beginning of the Fur Animal Epidemic Necrotic Pyoderma (FENP) epidemic. The authors also postulated that this species may have a marine origin as adhesins are the largest single category of virulence factors from the core genome.

In "Comparative Genomics of *Actinobacillus pleuropneumoniae* Serotype 8 Reveals the Importance of Prophages in the Genetic Variability of the Species," I. G. de Oliveira Pardo et al. presented the genome of *A. pleuropneumoniae* serotype 8 along with comparisons of seven genomes of seven serotype 8 clinical isolates with the other genomes of 12 serotypes. *A.*

pleuropneumoniae is the causative agent of porcine pleuropneumonia. Serotype 8 is the most widely distributed in the United States, Canada, United Kingdom, and southeastern Brazil. The proposed genomic analyses of serotype 8 genomes resulted in a set of 2352 protein-coding sequences; 76.6% of these proteins are commonly shared across all serotypes, 18.5% are shared with some serotypes, and 4.9% were differentially present, which are primarily a series of hypothetical and regulatory mobile elements. Additionally, the authors identified 30 prophage sequences, of which 16 are members of the family Pasteurellaceae. This suggests that mobile genetic elements play a role in the diversity and evolution of *A. pleuropneumoniae*.

In "Genomic Analysis of *Bacillus megaterium* NCT-2 Reveals Its Genetic Basis for the Bioremediation of Secondary Salinization Soil," B. Wang et al. reported genome sequencing of a nitrate-uptake bacterium *B. megaterium* NCT-2 using HiSeq and PacBio sequencing. The total size of this genome is 5.88 Mbp with 37.87% GC content including 10 indigenous plasmids. The *B. megaterium* NCT-2 genome contains 5606 genes, 142 tRNAs, and 53 rRNAs. The authors also described genes involved in bioremediation in secondary salinization soil.

Streptococcus parauberis is a Gram-positive, alpha-haemolytic lactic acid coccoid-shaped bacterium. This bacterium is a fish pathogen that especially affects olive flounder (*Paralichthys olivaceus*). Accordingly, *S. parauberis* is responsible for massive losses for fish farmers across different countries in Asia and Europe. *S. parauberis* is known to possess antibiotic resistance against most antibacterial drugs (such as tetracycline, oxytetracycline, and erythromycin) that would be used to treat this pathogen. Identification of alternative therapeutic agents against *S. parauberis* is therefore necessary. In "Pharmacodynamics of Ceftiofur Selected by Genomic and Proteomic Approaches of *Streptococcus parauberis* Isolated from the Flounder, *Paralichthys olivaceus*," N. Boby et al. employed an integrative multiomic strategy to present subtractive and comparative metabolic and genomic-based findings of therapeutic targets against *S. parauberis*. The authors also proposed ceftiofur as a new antimicrobial drug for treating *P. olivaceus* infected with *S. parauberis* by coupling multiomic approaches with pharmacodynamic profiles of the approved antimicrobial drugs.

In "FcircSEC: An R Package for Full Length circRNA Sequence Extraction and Classification," T. Hossain et al. presented an R package for deciphering and classifying the circRNA as FcircSEC (Full-Length circRNA Sequence Extraction and Classification). All existing tools for circRNA predictions only provide genomic coordinates of the predicted circRNA. Hence, the authors developed an R package that focused on several features, including gene annotation. This tool is capable of genomic location-based classification of circRNA, such as exonic or intronic.

This R-based tool is capable of handling datasets of several species including human data. The authors validated the resulting data using three different databases, namely, circBase, circRNADb, and PlantcircBase. This R tool FcircSEC is based on the Bioconductor package Biostring and uses the output from state-of-the-art circRNA prediction tools. The R package FcircSEC is freely available at its dedicated website (http://hpcc.siat.ac.cn/FcircSEC/Home.html), where the authors provided downloadable datasets, a reference manual, source code, and Windows binaries. The authors followed the standards of software sharing by providing this R tool at the Comprehensive R Archive Network (CRAN, https://cran.r-project.org/web/packages/FcircSEC/index.html) and also at the GitHub repository (https://github.com/tofazzal4720/FcircSEC).

Four out of five articles in this special issue are focused on comparative genomics, and one is focused on the development of a novel method. In our opinion, the articles in this special issue would allow us to explore new approaches to understand the molecular mechanisms linked to microbial function. We hope that these materials will assist and inspire readers working in the field of microbial genomics.

Conflicts of Interest

No potential conflict of interest was reported by the authors.

Acknowledgments

RK and TS were supported by the VEO-European Union's Horizon 2020 (grant no. 874735), the Academy of Finland (grant nos. 316264 and 329323), the Helsinki University Hospital Funds, and the Jane and Aatos Erkko Foundation. AK is thankful to the Department of Biotechnology, Government of India, India, for the Ramalingaswami Re-Entry Faculty Fellowship (Grant: BT/RLF/Re-entry/38/2017). We would like to extend our gratitude to all authors who contributed to this special issue and all reviewers who helped us select the highest quality manuscripts.

Ravi Kant
Abhishek Kumar
Tarja Sironen

References

[1] M. A. Malla, A. Dubey, A. Kumar, S. Yadav, A. Hashem, and E. F. Abd_Allah, "Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment," *Frontiers in Immunology*, vol. 9, 2019.

[2] D. MacLean, J. D. G. Jones, and D. J. Studholme, "Application of 'next-generation' sequencing technologies to microbial genetics," *Nature Reviews. Microbiology*, vol. 7, no. 4, pp. 96-97, 2009.

[3] J. E. Belizário and M. Napolitano, "Human microbiomes and their roles in dysbiosis, common diseases, and novel therapeutic approaches," *Frontiers in Microbiology*, vol. 6, 2015.

[4] R. Kant, J. Blom, A. Palva, R. J. Siezen, and W. M. de Vos, "Comparative genomics of Lactobacillus," *Microbial Biotechnology*, vol. 4, no. 3, pp. 323–332, 2011.

[5] R. Kant, A. Palva, and I. von Ossowski, "An in silico pan-genomic probe for the molecular traits behind Lactobacillus ruminis gut autochthony," *PLOS ONE*, vol. 12, no. 4, p. e0175541, 2017.

[6] V. Aguiar-Pulido, W. Huang, V. Suarez-Ulloa, T. Cickovski, K. Mathee, and G. Narasimhan, "Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis," *Evolutionary Bioinformatics*, vol. 12s1, no. 12s1, p. -EBO.S36436, 2016.

Hindawi
International Journal of Genomics
Volume 2020, Article ID 9084901, 11 pages
https://doi.org/10.1155/2020/9084901

Research Article

FcircSEC: An R Package for Full Length circRNA Sequence Extraction and Classification

Md. Tofazzal Hossain,[1,2] **Yin Peng**,[3] **Shengzhong Feng**,[1] **and Yanjie Wei**[1]

[1] *Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Center for High Performance Computing, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong Province, China 518055*
[2] *University of Chinese Academy of Sciences, No. 19(A) Yuquan Road, Shijingshan District, Beijing, China 100049*
[3] *Department of Pathology, The Shenzhen University School of Medicine, Shenzhen, Guangdong, China 518060*

Correspondence should be addressed to Yanjie Wei; yj.wei@siat.ac.cn

Received 18 November 2019; Revised 2 February 2020; Accepted 28 February 2020; Published 1 June 2020

Guest Editor: Ravi Kant

Circular RNAs (circRNAs) are formed by joining the 3′ and 5′ ends of RNA molecules. Identification of circRNAs is an important part of circRNA research. The circRNA prediction methods can predict the circRNAs with start and end positions in the chromosome but cannot identify the full-length circRNA sequences. We present an R package FcircSEC (Full Length circRNA Sequence Extraction and Classification) to extract the full-length circRNA sequences based on gene annotation and the output of any circRNA prediction tools whose output has a chromosome, start and end positions, and a strand for each circRNA. To validate FcircSEC, we have used three databases, circbase, circRNAdb, and plantcircbase. With information such as the chromosome and strand of each circRNA as the input, the identified sequences by FcircSEC are consistent with the databases. The novelty of FcircSEC is that it can take the output of state-of-the-art circRNA prediction tools as input and is applicable for human and other species. We also classify the circRNAs as exonic, intronic, and others. The R package FcircSEC is freely available.

1. Introduction

Circular RNAs (circRNAs) are formed by joining a downstream 3′ splice donor site and an upstream 5′ splice acceptor site in the primary transcript [1]. In most cases, circRNAs originate from exons close to the 5′ end of a protein coding gene and may consist of one or more exons. Furthermore, multiple circRNAs can be produced from a single gene. circRNAs are generated through several distinct mechanisms that rely on complementary sequences within flanking introns [2–4], exon skipping [4, 5], and exon-containing lariat precursors [6]. circRNAs were first discovered approximately 40 years ago and thought to be an RNA splicing error [7]. Until 2013, the researchers did not pay much attention in this area, but after publishing the paper [8], the circRNA research turned into a prominent field in scientific research. A significant amount of circRNAs is identified through the high-throughput RNA sequencing and bioinformatics analysis [9, 10]. In recent years, many types of circRNAs have been identified and found to be stable and abundant [2]. One of the important properties of circRNA is that they have tissue-specific expression. Several studies conclude that circRNAs are substantially enriched in brain tissues and the expression levels are dynamic during brain development of human and mice brain tissues [11–13]. circRNAs show differential expressions between primary ovarian tumors and metastatic tumors in ovarian carcinoma [14]. Some circRNAs also interact with RNA-binding proteins (RBPs) [15] although very little enrichment in binding sites of RBPs is found for circRNA sequences compared with those of its corresponding linear mRNA. The studies [8, 16, 17] reveal that circRNAs can bind to a few RNA-binding proteins (RBPs), such as Argonaute and MBL. circRNAs are conserved across different species and act as a microRNA (miRNA) sponge while miRNAs have oncogenic or tumor suppressor properties [18]. Although the function of most circRNAs is unknown, some functions of the circRNAs are known as miRNA sponges [8, 19, 20], protein translation templates [21–24], and regulation of gene expression [25–28]. Different studies suggest that circRNAs are important biomarkers for different cancers [29–31] and autoimmune diseases [32, 33], a potential noninvasive diagnosis for atherosclerosis [34], disorders of the central neural diseases [35], degenerative diseases [17], and cancers [10, 36].

Identification of circRNAs is a crucial step for circRNA research. A number of methods is available for the identification of circRNAs such as CIRI [37], circRNA_finder [38], DCC [39], find_circ [40], segemehl [41], CIRCexplorer [3], MapSplice [42], and UROBORUS [43]. Each of these methods can predict circRNAs and their position in the chromosome, but these methods cannot provide the full-length circRNA sequence. To infer/predict the function of the circRNAs, differential expression analysis and network analysis are very common, and full-length circRNA sequences are required. CIRI-full [44] can extract the full-length circRNA sequences from its own output of CIRI; however, the method fails for the unequal read lengths in a sample and does not accept the annotation file in gff format. FUCHS [45] does not provide a full-length circRNA sequence directly; only when using its output with additional software is it possible to obtain the full-length circRNA sequences. Besides, FUCHS is tested for the output of DCC only. Recently, a software tool circtools [46] has been published as a one-stop software solution for circRNA research which also uses the FUCHS module. Another method CircPrimer [47] can extract the full-length circRNA sequences although its main function is to design primers. CircPrimer cannot extract circRNA sequences other than for humans. The output of CIRI, find_circ, circRNA_finder, DCC, and segemehl gives three types of circRNAs (exonic, intronic, and intergenic). CIRCexplorer and MapSplice give two types (exonic and intronic) while UROBORUS gives only one type (exonic) of circRNAs in their output. Again, FUCHS, circtools, and CircPrimer cannot provide circRNA classification. A number of papers [48–50] classified their circRNAs as these five types: exonic, intronic, intergenic, sense overlapping, and antisense. Another paper [51] classified circRNA as exonic, intronic, intergenic, bidirectional/intragenic, and antisense. Our realization is that circRNA classification is not finished yet. The existence of exonic and intronic circRNA is supported by numerous biological experiments, but other types are rarely validated by PCR experiments. Therefore, we have classified the circRNAs as exonic, intronic, and others.

There are four available tools for extracting full-length circRNA sequences, CIRI-full, FUCHS, circtools, and CircPrimer. CIRI-full utilizes both BSJ (back-splice junction) and RO (reverse overlap) features to obtain full-length circRNA sequences. CIRI-full uses the output of CIRI, and RNA-seq data is needed to reconstruct the full-length sequence. The main limitation of CIRI-full is that it is not applicable if the sequencing read lengths are not equal for all reads in the RNA-seq data. Besides, CIRI-full does not accept the annotation file in gff format. FUCHS is developed to fully characterize candidate circRNA sequence utilizing all RNA-seq information from long reads (>150 bp). It is tested for the output of DCC only and not applicable for short reads. Besides, FUCHS cannot provide a full-length circRNA sequence directly. circtools is designed for RBP enrichment screenings and circRNA primer design, as well as circRNA sequence reconstruction. For circRNA sequence reconstruction, circtools utilizes the FUCHS module. The main function of CircPrimer is to design primers for circRNAs. Additionally, it can extract full-length circRNA sequences. It depends on the annotation information and is useful for human circRNAs only.

In this paper, we present an R package FcircSEC to extract directly the full-length circRNA sequences and to classify the circRNAs utilizing the output of circRNA prediction methods and the gene annotation information. We have followed the approach similar to CircPrimer in extracting circRNA sequences. Like CircPrimer, FcircSEC first selects the best transcript from the annotation file, then from the part of the transcript within the circRNA boundary, the introns are removed and finally combines all the exon sequences as a circRNA sequence. But our best transcript selection strategy (described in Materials and Methods) is different from CircPrimer. Even CircPrimer is applicable for human circRNAs only, but FcircSEC is useful for human and other species. FcircSEC only needs the output of the circRNA prediction tool along with the reference genome and the annotation file. The main advantage of FcircSEC is that it can use the output of many state-of-the-art circRNA prediction tools for extracting the actual sequence (with information on chromosome, circRNA start and end position, and strand). As there are no tools for full-length circRNAs for the user of the circRNA prediction tools other than CIRI and DCC, FcircSEC can be a good choice for them.

2. Materials and Methods

In our R package FcircSEC, from the gene annotation information of the reference genome, we extracted all transcripts and got the number of exons with their start and end positions for each transcript. Then, we selected the best transcript using the output of circRNA prediction methods. Finally, we extracted the full-length circRNA sequences from the selected best transcript. To check the validity of our package, we used human circRNAs from two popular databases circbase (http://circbase.org/) and circRNAdb (http://202.195.183.4:8000/circrnadb/circRNADb.php) and plant circRNAs from the plantcircbase (http://ibi.zju.edu.cn/plantcircbase/) database. The circRNA sequences obtained by FcircSEC were consistent with the databases.

The package needs three input files: (1) the four types of information (chromosome name, start position, end position, and strand of the circRNAs) from the output of circRNA prediction tools, (2) the reference genome, and (3) the annotation file corresponding to the reference genome. Inputs (2) and (3) can be downloaded from UCSC, NCBI, or any other databases, and input (1) can be obtained from circRNA prediction tools like CIRI, find_circ, circRNA_finder, DCC, CIRCexplorer, segemehl, MapSplice, and UROBORUS whose outputs have the abovementioned four types of information. The genome versions used in our

Table 1: Genome versions across species.

Species/database name	Genome/annotation version	Web links
circbase and circRNAdb (human)	hg19	http://hgdownload.soe.ucsc.edu/downloads.html#human http://202.195.183.4:8000/circrnadb/resources.php
Arabidopsis thaliana	TAIR10.38	ftp://ftp.ensemblgenomes.org/pub/plants/release-38/
Poncirus trifoliata	Citrus_clementina_v1.0	https://ftp.ncbi.nih.gov/genomes/refseq/plant/Citrus_clementina/
Glycine max	Glycine_max_v2.0.38	ftp://ftp.ensemblgenomes.org/pub/plants/release-38/
Gossypium arboreum	Gossypium_arboreum_v1.0	https://ftp.ncbi.nih.gov/genomes/refseq/plant/Gossypium_arboreum/
Gossypium hirsutum	NBI_Gossypium_hirsutum_v1.1	https://www.cottongen.org/data/download/genome_NBI_AD1/
Gossypium raimondii	Graimondii_221	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Graimondii/
Hordeum vulgare	Hv_IBSC_PGSB_v2.38	ftp://ftp.ensemblgenomes.org/pub/plants/release-38/
Oryza sativa	IRGSP-1.0.38	ftp://ftp.ensemblgenomes.org/pub/plants/release-38/
Solanum lycopersicum	SL2.50.38	ftp://ftp.ensemblgenomes.org/pub/plants/release-38/
Solanum tuberosum	SolTub_3.0.38	ftp://ftp.ensemblgenomes.org/pub/plants/release-38/
Triticum aestivum	IWGSC1.0+popseq.29	ftp://ftp.ensemblgenomes.org/pub/plants/release-29
Zea mays	AGPv4.38	ftp://ftp.ensemblgenomes.org/pub/plants/release-38/

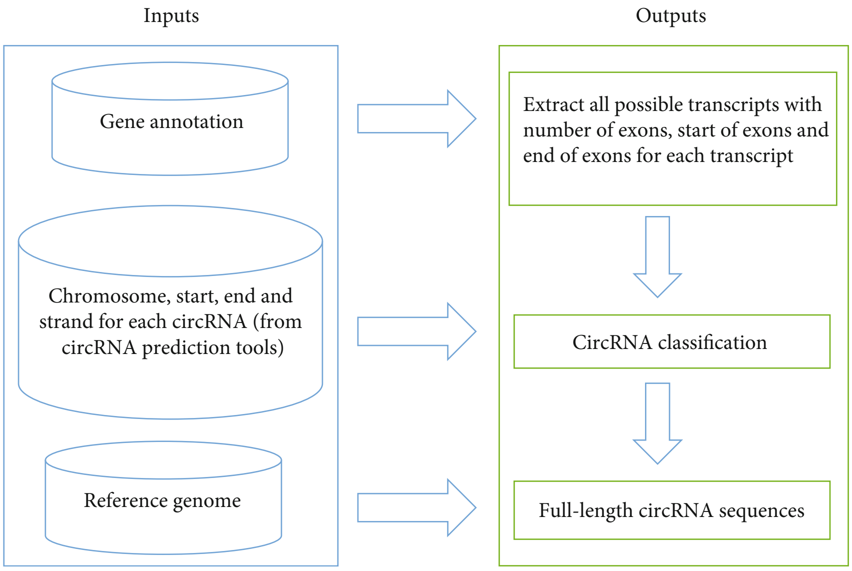

Inputs
Outputs
Gene annotation
Chromosome, start, end and strand for each circRNA (from circRNA prediction tools)
Reference genome
Extract all possible transcripts with number of exons, start of exons and end of exons for each transcript
CircRNA classification
Full-length circRNA sequences

Figure 1: Workflow of the FcircSEC package. From the gene annotation file, a nine-column transcript data file is generated for all transcripts with the number of exons and the start and end of each exon. Using the transcript data and the output of the circRNA prediction tool, the circRNA classification is done. Using the circRNA classification information and the reference genome, the full-length circRNA sequences are extracted.

analysis for different species are given in Table 1. The flowchart of FcircSEC is provided in Figure 1.

2.1. Location-Oriented Classification of Circular RNA. We classified circRNAs as three types: exonic, intronic, and others. *Exonic*: if the circRNA is originated from one or more exons of the linear transcript and the transcript strand and the circRNA strand are same, then the circRNA is exonic. *Intronic*: if the circRNA is originated from an intron of the linear transcript, then the circRNA is intronic. *Other*: if the circRNA belongs to neither exonic nor intronic, we call it as other type. The three types of the circRNAs are shown in Figure 2.

2.2. Extraction of Transcript Information from the Gene Annotation File. In this step, the input was the gene annotation file of the reference genome. The annotation file has nine columns: seqname, source, feature, start, end, score, strand, frame, and attribute. To extract the transcript information from the annotation data, the following steps were followed:

Step 1: from the attribute column of the annotation file, extract the transcript name and the gene name

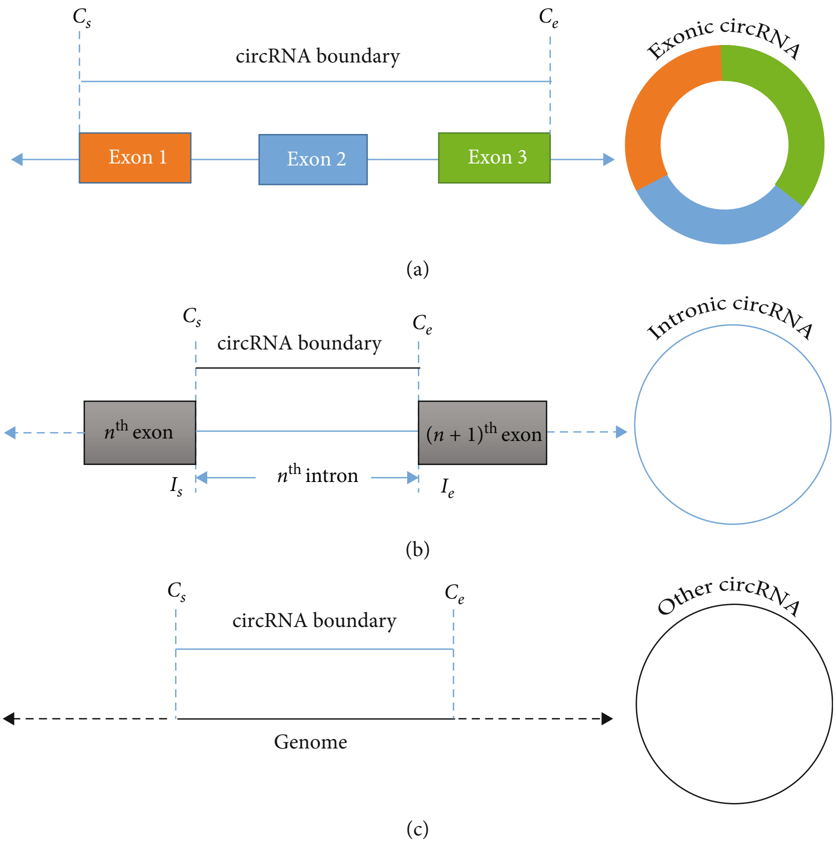


Figure 2: Location-oriented classification of circRNA. (a) The circRNA is exonic as it is formed of three exons (exons 1, 2, and 3). (b) The circRNA is intronic as it consists of one intron only. (c) The circRNA type is other as it belongs to neither exonic nor intronic.

Step 2: for each unique transcript, count the number of exons and obtain the start position and end position of each exon

Step 3: subtract 1 from the start position of the exons

Step 4: make a 9-column text file with transcript name (ID), chromosome, transcript strand, transcript start, transcript end, number of exons in each transcript, start positions of exons, end positions of exons, and gene name

2.3. Selection of the Best Transcript. The inputs of this step were the transcript data obtained from the previous Section 2.2 and the four columns (chromosome, start position, end position, and strand of circRNAs) from the output of circRNA prediction methods. In the best transcript selection, we followed two strategies. We selected the transcript whose coordinates (an interval from transcript start to end) contained the circRNA boundary. If there were several such transcripts, we selected all of them. For all possible transcripts, we checked whether the circRNA start and end positions exactly matched or not with the start of the first exon and end of the last exon, respectively. If yes, we selected that transcript as best transcript which has the longest splice sequence (sequence of all combined exons). If not, we selected the transcripts having maximum number of exons and then selected the one having the maximum length.

Let T be all transcripts extracted from the gene annotation file and O be the output from the circRNA prediction tool. For i^{th} circRNA of O, all possible transcripts T_{possible} were selected containing the circRNA boundary (e.g., transcripts 1 and 2 for circRNA 1 in Figure 3(a)). Then, the best transcript was selected using case 1 and case 2.

Case 1. For any transcript from T_{possible}, if the start position of the first exon and the end position of the last exon are exactly matched with the circRNA boundary (e.g., transcript 1 in Figure 3(a)), select that transcript. If more than one such type of transcript is selected, repeat the following steps until a single transcript is selected:

(1) select the transcript having the maximum splice length (length of all combined exons)

(2) select the transcript having the maximum transcript length

(3) select the first one

Case 2. For all transcripts from T_{possible}, if the start position of the first exon and the end position of the last exon are not exactly matched with the circRNA boundary (e.g., transcripts 1 and 2 in Figure 3(b)), select that transcript having the maximum number of exons within the boundary (e.g., transcript 1 in Figure 3(b)). If more than one such type of transcript is

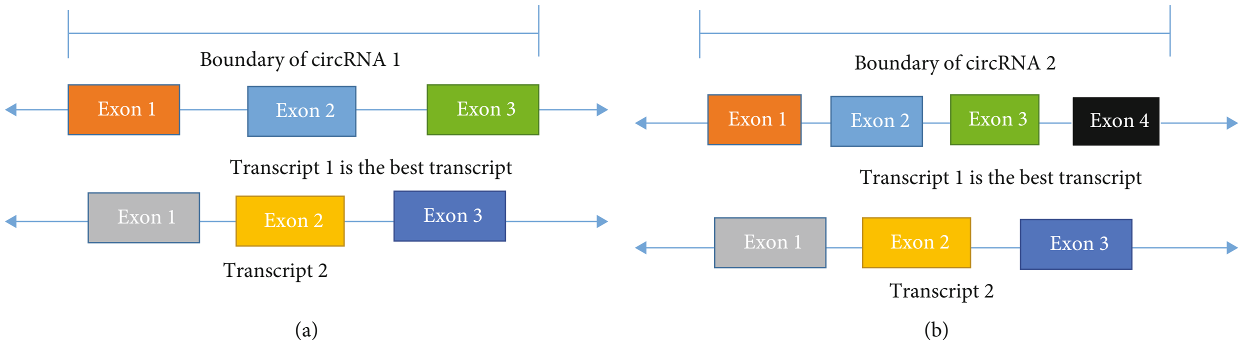


Figure 3: Best transcript selection. (a) There are two possible transcripts within the circRNA boundary. Transcript 1 is the best transcript as its start and end positions are exactly matched with those of the circRNA. (b) Within the circRNA boundary, there are two possible transcripts, and for both, the transcripts' start and end positions are not exactly matched with those of the circRNA. Transcript 1 is the best transcript as it has a larger number of exons than transcript 2.

selected, repeat the following steps until a single transcript is selected:

(1) select the transcript having the maximum transcript length

(2) select the first one

2.4. Circular RNA Classification and Sequence Extraction. The inputs of this step were the best transcript obtained from the previous Section 2.3, the four columns of the outputs of the circRNA prediction tools, and the reference genome. For any circRNA, if no best transcript is available, the corresponding circRNA was declared as "other" type. In the best transcript, if there was no exon within the circRNA boundary and an intron is contained within the circRNA boundary, we defined that circRNA as intronic. When there were some exons in the best transcript within the circRNA boundary, and the first and the last exon contained the start and end positions of the circRNA, respectively, we defined that circRNA as exonic, while the circRNA which was neither exonic nor intronic was declared as "other" type.

Let O be the output from the circRNA prediction tool and T_{best} be the best transcript for the i^{th} circRNA. Some variables were defined as

$$\text{Start} = \begin{cases} 1, & \text{circRNA start position} \geq \text{start of } 1^{\text{st}} \text{ exon of best transcript within circRNA boundary}, \\ 0, & \text{otherwise}, \end{cases}$$
$$\text{End} = \begin{cases} 1, & \text{circRNA end position} \leq \text{end of last exon of best transcript within circRNA boundary}, \\ 0, & \text{otherwise}. \end{cases} \tag{1}$$

For the i^{th} circRNA from O, the circRNA classification and sequence extraction were done using either of the following cases:

Case 1. If start = 1 and end = 1 (Figure 4(a)), the circRNA is exonic, and the sequence is composed of the exons from T_{best} within the circRNA boundary (Figure 4(a)).

Case 2. If there are no exons and only one intron in T_{best} within the circRNA boundary, the circRNA is intronic. The sequence is composed of one intron from T_{best} (Figure 4(b)).

Case 3. If case 1 and case 2 are not satisfied, the circRNA is other type. The sequence is composed of a genomic sequence from start to end of the circRNA (Figure 4(c)).

3. Results

3.1. Extraction of Transcript Data and Full-Length circRNA Sequences. We have extracted the full-length circRNA sequences for the circRNAs downloaded from three databases, circbase, circRNAdb, and plantcircbase. For circbase and circRNAdb, only the human circRNAs have been used, and for plantcircbase, plant circRNAs have been used.

We have extracted the transcript data from the gene annotation file. Using the transcript data and the output of the circRNA prediction tools, we have created the circRNA classification file which contains the circRNA classification and all the required information for getting the full-length circRNA sequences. Using the start and end positions of circRNAs obtained from the circRNA prediction tool, we have extracted the genomic sequence from the reference genome. Finally, using the circRNA classification information and

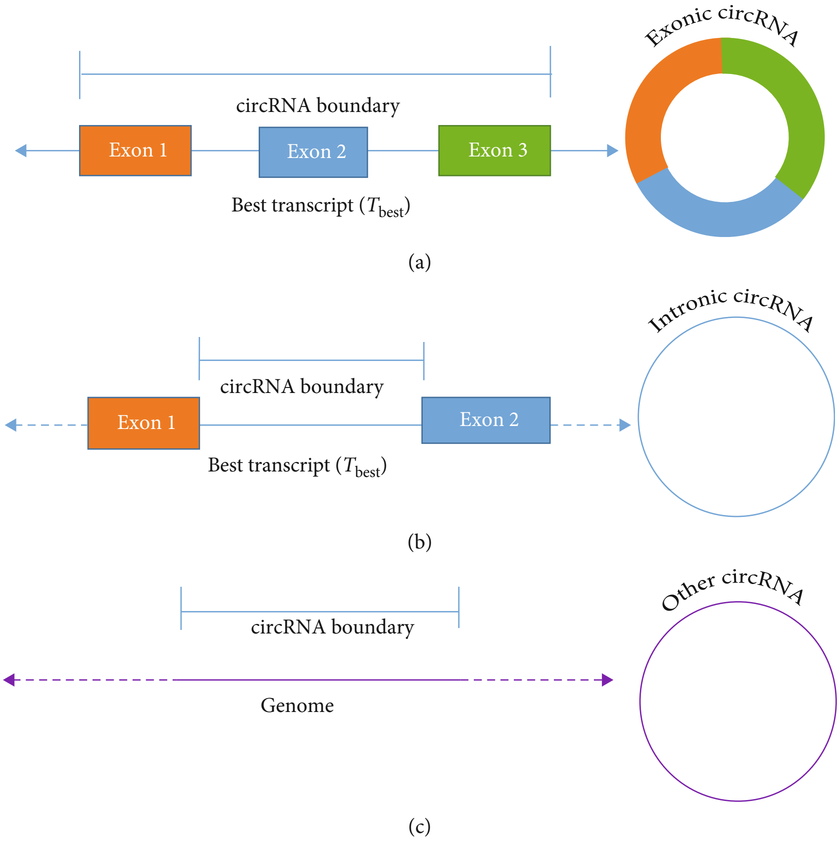


Figure 4: circRNA sequence extraction. (a) The circRNA is exonic, and its sequence is composed of combining the sequences of the three exons. (b) The circRNA is intronic, and its sequence consists of one intron sequence. (c) The circRNA is other type, and its sequence is the genomic sequence from start to end of the circRNA.

the genomic sequence, we have extracted the full-length circRNA sequences. The transcript data, circRNA classification, and full-length circRNA sequences are available at the supplementary Tables S1-S13, Tables S14-S28, and Tables S29-S43, respectively. The supplementary Tables S14-S28 (circRNA classification tables) have a total of 15 columns, and these columns represent, respectively, (1) circRNA ID, (2) chromosome, (3) circRNA start position, (4) circRNA end position, (5) circRNA strand, (6) circRNA length (7) circRNA type, (8) number of exons, (9) exon sizes, (10) exon offsets (start of each exon), (11) best transcript, (12) transcript strand, (13) transcript start, (14) transcript end, and (15) host gene.

3.2. Distribution of the circRNAs. In circbase, there are a total of 92375 human circRNAs; the extracted circRNA sequences by FcircSEC are 93.39% exonic, 0.75% are intronic, and 5.86% are others, while in circRNAdb, out of 32914 circRNAs, 99.32% are exonic, 0.02% are intronic, and 0.66% are others. Among the 67 experimentally validated circRNAs from plantcircbase, 62.69% are exonic and 37.31% are others, but no intronic is found. The classes of circRNAs for all other species are provided in Table 2. Again the distribution of the number of exons for the full-length exonic circRNAs is given in Figure 5. From Figure 5, we can observe that the median number of exons for most of the species is between 2 and 4.

3.3. Matched Sequences between Databases and FcircSEC. Since FcircSEC requires the chromosome name, start and end positions, and strand of each circRNAs as input, we have taken this information for each circRNA from the databases and extracted the full-length circRNA sequences using FcircSEC. Then, we have compared the sequences extracted by FcircSEC with those provided in the databases. During analysis, a sequence is matched if the whole sequence extracted by FcircSEC and the one provided in the database are identical (100%) and unmatched otherwise. We have calculated the proportion of matched sequences between FcircSEC and the three databases circbase, circRNAdb, and plantcircbase. Table 3 lists the proportion of matched sequences.

In circbase and circRNAdb, there are a total of 92375 and 32914 full-length human circRNA sequences, respectively. We have extracted these circRNA sequences by FcircSEC and compared them with those of the databases. From Table 3, we can see that 95.1% and 98.9% of the sequences extracted by FcircSEC are exactly matched with circbase and circRNAdb, respectively. In plantcircbase, there are 67 (out of 95143) experimentally validated full-length circRNA sequences. We have extracted these 67 circRNA sequences by FcircSEC and found that all are exactly matched with the databases. We have also extracted the full-length sequences for the rest of the 95076 circRNAs (available in Supplementary Table S31-S43).

Table 2: Different types of circRNAs classified by FcircSEC.

Species/database name	circRNA types Exonic (%)	Intronic (%)	Others (%)	Total
circbase	86267 (93.39)	695(0.75)	5413 (5.86)	92375
circRNAdb	32690 (99.32)	7 (0.02)	217 (0.66)	32914
Plantcircbase				
Arabidopsis thaliana	26643 (68.42)	1944 (4.99)	10351 (26.58)	38938
Poncirus trifoliata	242 (43.53)	5 (0.90)	309 (55.58)	556
Glycine max	2621 (49.24)	2086 (39.19)	616 (11.57)	5323
Gossypium arboreum	138 (13.41)	59 (5.73)	832 (80.86)	1029
Gossypium hirsutum	231 (46.29)	11 (2.20)	257 (51.50)	499
Gossypium raimondii	1231 (83.29)	10 (0.68)	237 (16.04)	1478
Hordeum vulgare	18 (46.15)	1 (2.56)	20 (51.28)	39
Oryza sativa	21849 (54.20)	9785 (24.27)	8677 (21.53)	40311
Oryza sativa (validated)	42 (62.69)	0 (0)	25 (37.31)	67
Solanum lycopersicum	1063 (55.83)	41 (2.15)	800 (42.02)	1904
Solanum tuberosum	584 (33.80)	3 (0.17)	1141 (66.03)	1728
Triticum aestivum	14 (15.91)	2 (2.27)	72 (81.82)	88
Zea mays	671 (20.72)	5 (0.15)	2562 (79.12)	3238

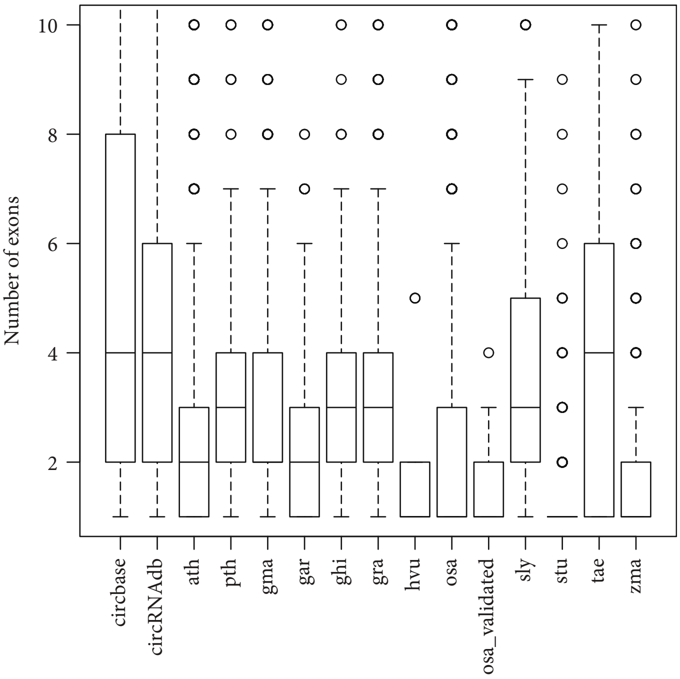


Figure 5: Boxplot of the number of exons for exonic circRNAs. Lower part of the box indicates Q_1 (1st quartile) which means 25% values of the exons are $\leq Q_1$, middle part of the box indicates Q_2(median) which means 50% values of the exons are $\leq Q_2$, and the upper part of the box represent Q_3 (3rd quartile) which means 75% values of the exons are $\leq Q_3$. The bars below and above the dashed line represent, respectively the minimum and maximum values.

Table 3: Proportion of matched sequences between databases and FcircSEC.

Database	Species	Total circRNAs	No. of matched sequences	No. of unmatched sequences	Matched percentage
circbase	Homo sapiens	92375	87840	4535	95.1%
circRNAdb	Homo sapiens	32914	32538	376	98.9%
plantcircbase	Oryza sativa (validated)	67	67	0	100%

Table 4: Comparison of FcircSEC with alternative methods.

Method	Prediction tools whose output is taken as input	Is RNA-seq data needed?	Classify circRNAs?	Limitation	Applicability
CIRI-full	CIRI	Yes	No	Not applicable for unequal read lengths in the RNA-seq data and for the annotation file in gff format	Applicable for the users of CIRI
FUCHS	DCC	Yes	No	Not applicable for short reads and cannot provide full-length circRNA sequence directly	Applicable for the users of DCC
circtools	DCC	Yes	No	Not applicable for short reads and cannot provide full-length circRNA sequence directly	Applicable for the users of DCC
CircPrimer	State-of-the-art circRNA prediction tools	No	No	Not applicable for other than human circRNAs and does not yield any information on splice sites within the circRNA sequence	Applicable for human circRNAs only
FcircSEC	State-of-the-art circRNA prediction tools	No	Yes	Does not yield any information on splice sites within the circRNA sequence	Applicable for almost all users of circRNA prediction tools

3.4. Comparison of FcircSEC with Alternative Methods. There are mainly four available tools for extracting full-length circRNA sequences: CIRI-full, FUCHS, circtools, and CircPrimer. Different methods depend on different prediction tools; for example, CIRI-full is dependent on CIRI, FUCHS is dependent on DCC, while CircPrimer and FcircSEC are not dependent on any prediction tools. Even some methods need RNA-seq data while others do not. As a result, performance of these methods is incomparable. Therefore, we have compared FcircSEC with the alternative methods in terms of application, limitation, etc. in Table 4.

From Table 4, we can see that CIRI-full takes the output of CIRI only as input, and RNA-seq data is needed to get the full-length sequence. It is not applicable if the lengths of all the reads in the RNA-seq data are not equal and if the annotation file is in gff format. Only the users of CIRI can use this tool for getting the full-length sequence. FUCHS and circtools take the output of DCC as input, and RNA-seq data is also needed to reconstruct the sequence. Both the tools are not applicable for short reads and cannot provide the full-length sequence directly. For both the tools, other software is needed to reconstruct the sequence. Both the tools are applicable for the users of DCC only. CircPrimer, although developed for designing primers, can extract the full-length sequences. But it is applicable for human circRNA only. FcircSEC can take the output of the state-of-the-art circRNA prediction tools as input. As RNA-seq data is not needed, there is no restriction in sequencing read lengths in using FcircSEC. It can take the annotation file in either gff or gtf format and is useful for human and other species. It can directly provide the full-length sequences. It can also classify circRNAs as three types (exonic, intronic, and others) while other methods cannot. The only limitation of FcircSEC is that it does not provide any information on splice sites within the circRNA sequence. In summary, we can say that FcircSEC has advantages over the existing methods.

4. Discussion

There are several circRNA prediction tools, but for only two tools CIRI and DCC, there is an existing method (CIRI-full and FUCHS) for getting the full-length sequences. For the users of other circRNA prediction tools (except CIRI and DCC), there are no existing tools for getting the full-length sequences. Although our method depends on the gene annotation information only, it will be a useful tool for the users who are interested in using the circRNA prediction tools other than CIRI and DCC.

CIRI-full and FUCHS can take the output of CIRI and DCC, respectively, as input, and hence, CIRI-full and FUCHS are applicable for the users of CIRI and DCC, respectively. circtools is also useful for DCC users as it uses the FUCHS module for circRNA sequence reconstruction. CircPrimer is applicable for human circRNAs only. Our method FcircSEC depends on the output of circRNA prediction tools, annotation information, and reference genome. FcircSEC can take the output of state-of-the-art circRNA prediction tools as input and, therefore, is applicable for almost all users of circRNA prediction tools.

Our method can extract the full-length circRNA sequence using the output of the existing circRNA prediction tools. We assume that the results of the existing circRNA prediction tools are correct, and we have not applied any filtering steps to detect the false positives. Again, within the circRNA boundary, we find a matching of the start of the first and the end of the last exons of the best transcript with the circRNA start and end positions. We assume that the circRNA contains all the intermediate exons, and we combine all the exons as a full-length circRNA. That is, we have not skipped any exons. This strategy is also used in CIRCexplorer.

FcircSEC does not take into account investigating the presence of the splice site within the circRNA sequence. For exonic circRNA, it combines all exons within the circRNA boundary to construct the full-length sequence. For the intronic and other types, it assumes that circRNAs are not spliced. By searching the databases circbase and circRNAdb, we have found that in almost all cases, the circRNA combines all exons. Besides, RNA-seq data is needed to examine the presence of a splice site within circRNAs. This is beyond the scope of the current work as FcircSEC is based on annotation information and does not take sequencing reads into account. This is the limitation of FcircSEC. We will try to overcome this limitation in the next version of the package.

Overall, the full-length sequence extraction is crucial in circRNA research. After predicting the candidate circRNAs, all the downstream analyses depend on the circRNA sequences. Therefore, FcircSEC can play an important role through extracting full-length circRNA sequences in identifying important circRNA biomarkers.

5. Conclusions

A number of methods are available in the literature for predicting the circRNA sequences. But only a limited number of methods are available for extracting full-length circRNA sequences. In this paper, we have developed an R package FcircSEC for extracting full-length circRNA sequences using the output of most of the popular circRNA prediction tools. The results of FcircSEC are consistent with the published circRNA databases and give more information that are not available in the public databases. Moreover, as for the users of the circRNA prediction tools other than CIRI and DCC, as there is no full-length circRNA sequence extraction method, FcircSEC can be a good choice for them. The R package FcircSEC is freely available at http://hpcc.siat.ac.cn/FcircSEC/Home.html.

Data Availability

For circbase and circRNAdb, the annotation file is downloaded from circRNAdb (http://202.195.183.4:8000/circrnadb/resources.php) and the reference genome hg19 is downloaded from UCSC (http://hgdownload.soe.ucsc.edu/downloads.html#human). "Arabidopsis thaliana" annotation version TAIR10.38, "Glycine max" annotation version Glycine_max_v2.0.38, "Hordeum vulgare" annotation version Hv_IBSC_PGSB_v2.38, "Oryza sativa" annotation version IRGSP-1.0.38, "Solanum lycopersicum" annotation version SL2.50.38, "Solanum tuberosum" annotation version SolTub_3.0.38, "Triticum aestivum" annotation version IWGSC1.0+popseq.29, "Zea mays" annotation version AGPv4.38 and their reference genomes are downloaded from ftp://ftp.ensemblgenomes.org/pub/plants/.

Disclosure

The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

MTH, YP, and YW designed the study. MTH developed the R package, collected data, performed computations, and tested the package. YW and SF supervised the work. All the authors' participated in writing and approved the final version of the manuscript.

Acknowledgments

We would like to thank all the members of the Computational Biology and Bioinformatics Lab, Center for High Performance Computing, SIAT, CAS for their valuable suggestions and feedback. This work was partly supported by the National Key Research and Development Program of China under grant nos. 2018YFB0204403 and 2016YFB0201305, the National Science Foundation of China under grant nos. U1813203 and 61433012, the Shenzhen Basic Research Fund under grant nos. JCYJ20180507182818013 and JCYJ20170413093358429, the China Postdoctoral Science Foundation under grant no. 2019M653132, and the CAS Key Lab under grant no. 2011DP173015. We would also like to give thanks for the funding support by the Shenzhen Discipline Construction Project for Urban Computing and Data Intelligence, Youth Innovation Promotion Association, CAS to Yanjie Wei.

Supplementary Materials

Supplementary Materials. Supplementary Tables S1–S13: transcript data for different species are given. Supplementary Tables S14–S28: circRNA classification data for different species are provided. Supplementary Tables S29–S43: full-length circRNA sequences across species are supplied. *(Supplementary Materials)*

References

[1] L. Li, Y. C. Zheng, M. R. Kayani et al., "Comprehensive analysis of circRNA expression profiles in humans by RAISE," *International Journal of Oncology*, vol. 51, no. 6, pp. 1625–1638, 2017.

[2] W. R. Jeck, J. A. Sorrentino, K. Wang et al., "Circular RNAs are abundant, conserved, and associated with ALU repeats," *RNA*, vol. 19, no. 2, pp. 141–157, 2013.

[3] X. O. Zhang, H. B. Wang, Y. Zhang, X. Lu, L. L. Chen, and L. Yang, "Complementary sequence-mediated exon circularization," *Cell*, vol. 159, no. 1, pp. 134–147, 2014.

[4] A. Ivanov, S. Memczak, E. Wyler et al., "Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals," *Cell Reports*, vol. 10, no. 2, pp. 170–177, 2015.

[5] S. Kelly, C. Greenman, P. R. Cook, and A. Papantonis, "Exon skipping is correlated with exon circularization," *Journal of Molecular Biology*, vol. 427, no. 15, pp. 2414–2417, 2015.

[6] S. P. Barrett, P. L. Wang, and J. Salzman, "Circular RNA biogenesis can proceed through an exon-containing lariat precursor," *eLife*, vol. 4, 2015.

[7] Z. Xu, P. Li, L. Fan, and M. Wu, "The potential role of circRNA in tumor immunity regulation and immunotherapy," *Frontiers in Immunology*, vol. 9, p. 9, 2018.

[8] T. B. Hansen, T. I. Jensen, B. H. Clausen et al., "Natural RNA circles function as efficient microRNA sponges," *Nature*, vol. 495, no. 7441, pp. 384–388, 2013.

[9] I. Chen, C. Y. Chen, and T. J. Chuang, "Biogenesis, identification, and function of exonic circular RNAs," *Wiley Interdisciplinary Reviews: RNA*, vol. 6, no. 5, pp. 563–579, 2015.

[10] Y. Li, Q. Zheng, C. Bao et al., "Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis," *Cell Research*, vol. 25, no. 8, pp. 981–984, 2015.

[11] A. Rybak-Wolf, C. Stottmeister, P. Glažar et al., "Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed," *Molecular Cell*, vol. 58, no. 5, pp. 870–885, 2015.

[12] X. You, I. Vlatkovic, A. Babic et al., "Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity," *Nature Neuroscience*, vol. 18, no. 4, pp. 603–610, 2015.

[13] M. T. Venø, T. B. Hansen, S. T. Venø et al., "Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development," *Genome Biology*, vol. 16, no. 1, p. 245, 2015.

[14] I. Ahmed, T. Karedath, S. S. Andrews et al., "Altered expression pattern of circular RNAs in primary and metastatic sites of epithelial ovarian carcinoma," *Oncotarget*, vol. 7, no. 24, pp. 36366–36381, 2016.

[15] M. W. Hentze and T. Preiss, "Circular RNAs: splicing's enigma variations," *The EMBO Journal*, vol. 32, no. 7, pp. 923–925, 2013.

[16] Z. Li, C. Huang, C. Bao et al., "Exon-intron circular RNAs regulate transcription in the nucleus," *Nature Structural & Molecular Biology*, vol. 22, no. 3, pp. 256–264, 2015.

[17] R. Ashwal-Fluss, M. Meyer, N. R. Pamudurti et al., "circRNA biogenesis competes with pre-mRNA splicing," *Molecular Cell*, vol. 56, no. 1, pp. 55–66, 2014.

[18] X. Zhang, Y. Peng, Z. Jin et al., "Integrated miRNA profiling and bioinformatics analyses reveal potential causative miRNAs in gastric adenocarcinoma," *Oncotarget*, vol. 6, no. 32, pp. 32878–32889, 2015.

[19] Q. Zheng, C. Bao, W. Guo et al., "Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs," *Nature Communications*, vol. 7, no. 1, article 11215, 2016.

[20] F. R. Kulcheski, A. P. Christoff, and R. Margis, "Circular RNAs are miRNA sponges and can be used as a new class of biomarker," *Journal of Biotechnology*, vol. 238, pp. 42–51, 2016.

[21] Y. Yang, X. Fan, M. Mao et al., "Extensive translation of circular RNAs driven by N^6-methyladenosine," *Cell Research*, vol. 27, no. 5, pp. 626–641, 2017.

[22] N. R. Pamudurti, O. Bartok, M. Jens et al., "Translation of circRNAs," *Molecular Cell*, vol. 66, no. 1, pp. 9–21.e7, 2017.

[23] I. Legnini, G. di Timoteo, F. Rossi et al., "Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis," *Molecular Cell*, vol. 66, no. 1, pp. 22–37.e9, 2017.

[24] H. Chen, Y. Liu, P. Li, and D. Zhu, "RE: novel role of FBXW7 circular RNA in repressing glioma tumorigenesis," *Journal of the National Cancer Institute*, vol. 111, no. 4, p. 435, 2019.

[25] W. W. Du, W. Yang, E. Liu, Z. Yang, P. Dhaliwal, and B. B. Yang, "Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2," *Nucleic Acids Research*, vol. 44, no. 6, pp. 2846–2858, 2016.

[26] K. Abdelmohsen, A. C. Panda, R. Munk et al., "Identification of HuR target circular RNAs uncovers suppression of PABPN1 translation by circPABPN1," *RNA Biology*, vol. 14, no. 3, pp. 361–369, 2017.

[27] L. M. Holdt, A. Stahringer, K. Sass et al., "Circular non-coding RNA *ANRIL* modulates ribosomal RNA maturation and atherosclerosis in humans," *Nature Communications*, vol. 7, no. 1, article 12429, 2016.

[28] W. W. Du, L. Fang, W. Yang et al., "Induction of tumor apoptosis through a circular RNA enhancing Foxo3 activity," *Cell Death and Differentiation*, vol. 24, no. 2, pp. 357–370, 2017.

[29] F. Wang, A. J. Nazarali, and S. Ji, "Circular RNAs as potential biomarkers for cancer diagnosis and therapy," *American Journal of Cancer Research*, vol. 6, no. 6, pp. 1167–1176, 2016.

[30] J. Li, J. Yang, P. Zhou et al., "Circular RNAs in cancer: novel insights into origins, properties, functions and implications," *American Journal of Cancer Research*, vol. 5, no. 2, pp. 472–480, 2015.

[31] E. Anastasiadou, L. S. Jacob, and F. J. Slack, "Non-coding RNA networks in cancer," *Nature Reviews. Cancer*, vol. 18, no. 1, pp. 5–18, 2018.

[32] L. J. Li, Q. Huang, H. F. Pan, and D. Q. Ye, "Circular RNAs and systemic lupus erythematosus," *Experimental Cell Research*, vol. 346, no. 2, pp. 248–254, 2016.

[33] G. Cardamone, E. Paraboschi, V. Rimoldi, S. Duga, G. Soldà, and R. Asselta, "The characterization of GSDMB splicing and backsplicing profiles identifies novel isoforms and a circular RNA that are dysregulated in multiple sclerosis," *International Journal of Molecular Sciences*, vol. 18, no. 3, p. 576, 2017.

[34] C. E. Burd, W. R. Jeck, Y. Liu, H. K. Sanoff, Z. Wang, and N. E. Sharpless, "Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk," *PLoS Genetics*, vol. 6, no. 12, article e1001233, 2010.

[35] W. J. Lukiw, "Circular RNA (circRNA) in Alzheimer's disease (AD)," *Frontiers in Genetics*, vol. 4, p. 307, 2013.

[36] P. Li, S. Chen, H. Chen et al., "Using circular RNA as a novel type of biomarker in the screening of gastric cancer," *Clinica Chimica Acta*, vol. 444, pp. 132–136, 2015.

[37] Y. Gao, J. Wang, and F. Zhao, "CIRI: an efficient and unbiased algorithm for *de novo* circular RNA identification," *Genome Biology*, vol. 16, no. 1, p. 4, 2015.

[38] J. O. Westholm, P. Miura, S. Olson et al., "Genome-wide analysis of Drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation," *Cell Reports*, vol. 9, no. 5, pp. 1966–1980, 2014.

[39] J. Cheng, F. Metge, and C. Dieterich, "Specific identification and quantification of circular RNAs from sequencing data," *Bioinformatics*, vol. 32, no. 7, pp. 1094–1096, 2016.

[40] S. Memczak, M. Jens, A. Elefsinioti et al., "Circular RNAs are a large class of animal RNAs with regulatory potency," *Nature*, vol. 495, no. 7441, pp. 333–338, 2013.

[41] S. Hoffmann, C. Otto, G. Doose et al., "A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection," *Genome Biology*, vol. 15, no. 2, p. R34, 2014.

[42] K. Wang, D. Singh, Z. Zeng et al., "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic Acids Research*, vol. 38, no. 18, article e178, 2010.

[43] X. Song, N. Zhang, P. Han et al., "Circular RNA profile in gliomas revealed by identification tool UROBORUS," *Nucleic Acids Research*, vol. 44, no. 9, article e87, 2016.

[44] Y. Zheng, P. Ji, S. Chen, L. Hou, and F. Zhao, "Reconstruction of full-length circular RNAs enables isoform-level quantification," *Genome Medicine*, vol. 11, no. 1, p. 2, 2019.

[45] F. Metge, L. F. Czaja-Hasse, R. Reinhardt, and C. Dieterich, "FUCHS-towards full circular RNA characterization using RNAseq," *PeerJ*, vol. 5, article e2934, 2017.

[46] T. Jakobi, A. Uvarovskii, and C. Dieterich, "circtools—a one-stop software solution for circular RNA research," *Bioinformatics*, vol. 35, no. 13, pp. 2326–2328, 2019.

[47] S. Zhong, J. Wang, Q. Zhang, H. Xu, and J. Feng, "CircPrimer: a software for annotating circRNAs and determining the specificity of circRNA primers," *BMC Bioinformatics*, vol. 19, no. 1, p. 292, 2018.

[48] J. Liu, T. Liu, X. Wang, and A. He, "Circles reshaping the RNA world: from waste to treasure," *Molecular Cancer*, vol. 16, no. 1, p. 58, 2017.

[49] Y. Wang, M. Yang, S. Wei, F. Qin, H. Zhao, and B. Suo, "Identification of circular RNAs and their targets in leaves of *Triticum aestivum* L. under dehydration stress," *Frontiers in Plant Science*, vol. 7, article 2024, 2017.

[50] Y. Shen, X. Guo, and W. Wang, "Identification and characterization of circular RNAs in zebrafish," *FEBS Letters*, vol. 591, no. 1, pp. 213–220, 2017.

[51] D. Rong, W. Tang, Z. Li et al., "Novel insights into circular RNAs in clinical application of carcinomas," *OncoTargets and Therapy*, vol. 10, pp. 2183–2188, 2017.

Hindawi
International Journal of Genomics
Volume 2020, Article ID 4850290, 17 pages
https://doi.org/10.1155/2020/4850290

Research Article

Pharmacodynamics of Ceftiofur Selected by Genomic and Proteomic Approaches of *Streptococcus parauberis* Isolated from the Flounder, *Paralichthys olivaceus*

Naila Boby, Muhammad Aleem Abbas, Eon-Bee Lee, and Seung-Chun Park

Laboratory of Veterinary Pharmacokinetics and Pharmacodynamics, College of Veterinary Medicine, Kyungpook National University, Daegu 41569, Republic of Korea

Correspondence should be addressed to Seung-Chun Park; parksch@knu.ac.kr

Received 9 December 2019; Revised 14 February 2020; Accepted 13 March 2020; Published 31 March 2020

Guest Editor: Abhishek Kumar

We employed an integrative strategy to present subtractive and comparative metabolic and genomic-based findings of therapeutic targets against *Streptococcus parauberis*. For the first time, we not only identified potential targets based on genomic and proteomic database analyses but also recommend a new antimicrobial drug for the treatment of olive flounder (*Paralichthys olivaceus*) infected with *S. parauberis*. To do that, 102 total annotated metabolic pathways of this bacterial strain were extracted from computational comparative metabolic and genomic databases. Six druggable proteins were identified from these metabolic pathways from the DrugBank database with their respective genes as mtnN, penA, pbp2, *murB*, *mur*A, coaA, and fni out of 112 essential nonhomologous proteins. Among these hits, 26 transmembrane proteins and 77 cytoplasmic proteins were extracted as potential vaccines and drug targets, respectively. From the FDA DrugBank, ceftiofur was selected to prevent antibiotic resistance as it inhibited our selected identified target. Florfenicol is used for treatment of *S. parauberis* infection in flounder and was chosen as a comparator drug. All tested strains of fish isolates with *S. parauberis* were susceptible to ceftiofur and florfenicol with minimum inhibitory concentrations (MIC) of 0.0039–1 μg/mL and 0.5–8 μg/mL, IC_{50} of 0.001–0.5 μg/mL and 0.7–2.7 μg/mL, and minimum biofilm eradication concentrations (MBEC) of 2–256 μg/mL and 4–64 μg/mL, respectively. Similar susceptibility profiles for ceftiofur and florfenicol were found, with ceftiofur observed as an effective and potent antimicrobial drug against both planktonic and biofilm-forming strains of the fish pathogen *Streptococcus parauberis*, and it can be applied in the aquaculture industry. Thus, our predictive approach not only showed novel therapeutic agents but also indicated that marketed drugs should also be tested for efficacy against newly identified targets of this important fish pathogen.

1. Introduction

Streptococcus parauberis is one of the major pathogenic bacteria which cause economic losses in the aquaculture industry in the Northeast Asia fish farming industry, including Korea. *Streptococcus parauberis*, a member of the *Streptococcaceae* family, is a nonmotile, Gram-positive, alpha-hemolytic lactic acid bacterium with a coccoid shape. It is closely related to *S. uberis* and is included in the pyogenic streptococci class. In freshwater and marine cultures, it was first found in *Scophthalmus maximus* (turbot) and is the leading cause of chronic streptococcal infection, but it was initially included in the *S. uberis* subtypes [1, 2]. Streptococcal diseases of aquaculture fish are some of the most disastrous pathogenic conditions worldwide, including different regions of Spain, the US, Korea, Japan, Israel, and Italy [3–7].

With a progressively developing aquaculture industry, bacterial pathogen infections have increased, causing exponential losses of different fish species based on their geographical regions like *Paralichthys olivaceus* (olive flounder) in South Korea and Japan, *Scophthalmus maximus* (turbot) in Spain, *Sebastes ventricosus* (sea bass species) in Japan, and *Morone saxatilis* (striped bass) in North America due to streptococcosis. It produces a considerable deficit economically to fish farmers because of its substantial impact on fish stock mortality, and its occurrence must be controlled [6, 8].

On the Korean peninsula, *Paralichthys olivaceus* (olive flounder) is one of the dominant marine and fresh water fish that has suffered consistent mortality. In every geographical region, its pathogen is unaffected and has developed resistance against many antibiotics [1, 2]. The resistance issues demand exploration of new alternative therapeutic targets for treatment of infection or finding those targets, which will enhance antimicrobial sensitivity that is already present.

With increasing knowledge regarding an infecting organism's genome, metabolism, proteomes, and molecules important for their survival, the discovery of new drug targets is easier than before. Moreover, the number of microbial species with completely sequenced genomes is increasing frequently. Currently, the number of reported prokaryotic whole genome sequences is greater than 2000. The use of computational proteomic, genomic, and bioinformatic analyses for investigation of such infectious pathogenic organisms has not only aided in the *de novo* discovery of efficacious therapeutic agents but also provided alternative uses of already available drugs [9, 10].

Various computational comparative genomic and bioinformatic analysis methodologies are established for discovery of potential therapeutic agents against many microbial agents using comparison of host and pathogen proteins [11]. Pathogenic organisms have unique essential proteins that are necessary for the survival of an organism that can be targeted for therapeutic applications for control of bacterial growth [12]. Moreover, the use of existing therapeutic agents specifically for newly identified targets will not only save drug development time but also reduce drug treatment costs.

Cephalosporins are one of the most important antibacterial classes, as four generations from this drug class have beta-lactamase resistance due to the presence of a beta-lactam ring in their substructure, similar to penicillin. Although a large number of human drugs belong to this class, their use in dairy cattle and the aquaculture industry is limited. For treatment of mastitis infections in dairy cattle, only a few cephalosporins belonging to the 1st and 2nd generation classes have been used globally under strict regulations. Later, 3rd and 4th generation drugs like ceftiofur and cefquinome, respectively, are administered for veterinary purposes [13].

Ceftiofur (sodium salt) is one of the novel 3rd generation broad-spectrum cephalosporins introduced for veterinary use but is also administered for pathogenic bacterial infections in fish and poultry as a cell wall synthesis inhibitor. Ceftiofur is marketed under the brand name Naxcel® for Pasteurella infections of bovine respiratory disease treatment [14, 15]. Numerous reports showed that ceftiofur sodium controls infections in sheep, cattle, horse, swine, balady chicken, broiler chickens, and American black duck against *Pasteurella multocida*, *P. haemolytica*, and *E. coli*. In our study, we attempted to use ceftiofur sodium pharmacodynamics as an indicator of proof of our identified target by -omic study as a potential therapeutic target [16, 17].

Florfenicol is a semisynthetic antibacterial agent with a chemical structure and spectrum of antibacterial activity like thiamphenicol. Both florfenicol and thiamphenicol are chloramphenicol analogues in which the *p*-nitro group on the aromatic ring is substituted with a sulfonyl methyl group. Florfenicol binds to the bacterial 50S ribosomal subunit and inhibits protein synthesis at the peptidyl transferase step. *In vitro* investigations with florfenicol demonstrated potent activity against several bacteria pathogenic to fish. *In vivo* efficacy against furunculosis in Atlantic salmon and classical vibriosis in cod was confirmed [18].

Here, we identified putative novel therapeutic targets against *Streptococcus parauberis* by using available metabolic and genomic pathways and also recommended a new antimicrobial drug against *S. parauberis* infections in olive flounder with aid of the pharmacodynamics profiles of the approved antimicrobial drugs florfenicol and ceftiofur. We expect that our findings will not only give novel putative agents against *S. parauberis* but also provide new foundations using comparative and subtractive genomic methodologies for general pharmacodynamics studies of existing drugs or for specific organisms.

2. Materials and Methods

2.1. Metabolic Pathway Analysis. The KEGG database (abbreviated from the Kyoto Encyclopedia of Genes and Genomes) provides chemical, genomic, and functional information of the system of an organism and is used extensively as a source of reference dataset information produced by the sequencing of genomes [19]. In the present study, the metabolic pathways, protein, and nucleotide sequences of *Streptococcus parauberis* and *Homo sapiens* (human) genomes were retrieved from the KEGG pathway database [20, 21]. Different databases and search engines used during comparative genome analyses for the purpose of identifying therapeutic drug targets are demonstrated as a flowsheet in Figure 1.

2.2. Nonhomologous Essential Protein Screening. To identify essential nonhomologous proteins of the infecting bacteria, a comparison was performed in two steps. Firstly, we compared the proteins of *S. parauberis* with the host fish *Paralichthys olivaceus* (taxid: 8255) and with *Homo sapiens* (taxid: 9606) using NCBI-Blast for proteins with a threshold of 0.005 as an expectation value (*e* value), a percentage similarity of ≤35%, and a bit score of 100 as the minimum limit [22]. In the next step, these selected no-hits were used for comparison in the database of essential genes (DEG). The nonhomologous sequences of *S. parauberis* were aligned with the experimentally verified essential genes of five streptococcal species and proteins of 38 other Gram-positive and Gram-negative bacteria in the DEG-14.5 using the DEG microbial BLASTP with a minimum possible bit score of 100 and a cutoff *e* value of 10^{-10} [23, 24].

2.3. Characterization of Pathogen Nonhomologous Essential Proteins as Drug Targets. Based on their structural and molecular nature, the selected hits from the nucleotide and gene databases were characterized as therapeutic vaccine and drug targets. Proteins present in the cytoplasm are better drug targets; on the other hand, proteins present on the surface of cellular membranes are better targeted by vaccines

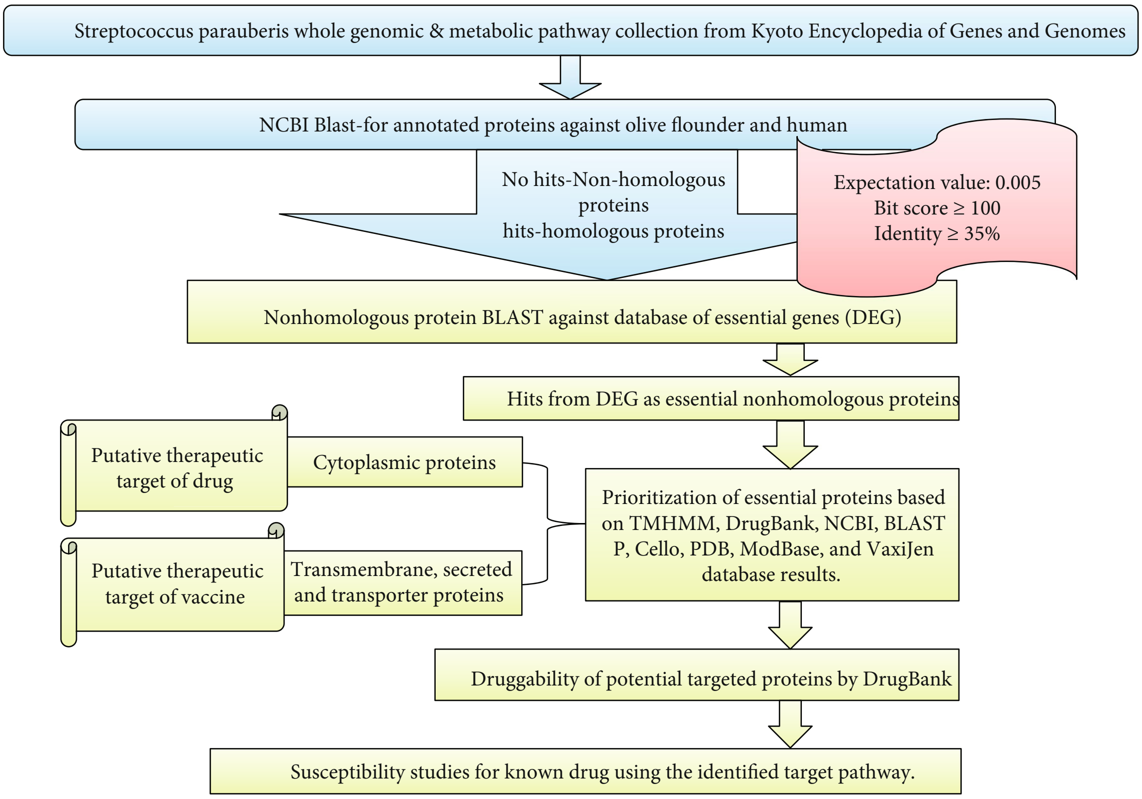

Streptococcus parauberis whole genomic & metabolic pathway collection from Kyoto Encyclopedia of Genes and Genomes
NCBI Blast-for annotated proteins against olive flounder and human
No hits-Non-homologous proteins
hits-homologous proteins
Expectation value: 0.005
Bit score ≥ 100
Identity ≥ 35%
Nonhomologous protein BLAST against database of essential genes (DEG)
Hits from DEG as essential nonhomologous proteins
Putative therapeutic target of drug
Cytoplasmic proteins
Putative therapeutic target of vaccine
Transmembrane, secreted and transporter proteins
Prioritization of essential proteins based on TMHMM, DrugBank, NCBI, BLAST P, Cello, PDB, ModBase, and VaxiJen database results.
Druggability of potential targeted proteins by DrugBank
Susceptibility studies for known drug using the identified target pathway.

Figure 1: Illustration of the comparative and subtractive genomic target identification along with susceptibility studies in *Streptococcus parauberis*. Using the different databases at each step (KEGG, NCBI-Blast, and DEG), the essential hits were selected. The other databases (CELLO, ModBase, PDB, and VaxiJen) have been used to characterize the selected nonhomologous essential proteins for their 3D structure and other physical properties. Based on this characterization, the low molecular weight (<110 kDa) cytoplasmic proteins are considered putative drug targets whereas the transmembrane proteins are putative vaccine targets. Finally, the druggability of selected essential putative targets was analyzed using the DrugBank database. Afterward, one known antimicrobial drug was selected to target the *Streptococcus parauberis* strains and to check the susceptibility of planktonic and biofilm-forming bacteria. Schematic flowchart.

[25]. Cellular localization analysis identifies proteins to their different locations on and within the pathogenic cell. The multiclass support vector machine classification database, also known as CELLO, version 2.5 (http://cello.life.nctu.edu.tw), was used for prediction of cellular and subcellular localization of the selected target proteins [26]. After cellular localization, the number of transmembrane helices in membrane proteins was predicted by TMHMM version 2.0 (http://www.cbs.dtu.dk/services/TMHMM/). This database is based on the hidden Markov model, and about 97–98% of the transmembrane helices have been determined by its experimentally predicted evidence [11, 27].

For determination of the molecular weight (MW) of the selected proteins, we used different online databases. As suggested by previous literature, smaller proteins are more soluble and easily purified, and such, they are more appropriate substances for development as drugs. Thus, we excluded proteins with MW > 110 kilodaltons [28].

Using these proteins for further analysis, the experimentally and computationally solved 3D structures were determined from the Protein Data Bank (PDB) (https://www.rcsb.org/pdb) and ModBase (https://salilab.org/modbase) databases, respectively [29, 30]. The PDB is a worldwide repository in which experimentally determined structures of proteins, nucleic acids, and complex biomolecule assemblies are deposited, and structures are explained by following their standards. ModBase is a database of protein structures developed by computational approaches and validated by statistically significant sequence alignments and model assessments [28].

In addition, for prediction of antigenic proteins, the database for protective antigen vaccine prediction, VaxiJen version 2.0, was used set with a threshold value > 0.4. The antigenic probability score, above the threshold value, represents the highest accuracy for the quantitative measure of protein sequences as protective antigens. These antigens form the basis of a subunit vaccine. A higher score of a protein refers to a higher probability for protective ability [31, 32].

The DrugBank database (version 4.3) contains unique bioinformatic and cheminformatic data of drugs and drug targets and is used for determining the druggability of essential proteins. By using default parameters, proteins were aligned for the available drug entries that included nutraceuticals, small molecule drugs, biotech (protein/peptide) drugs, and experimental drugs as approved by FDA (https://www.drugbank.ca/) [33].

2.4. In Vitro Efficacy Testing of Identified Protein Targets for Ceftiofur and Florfenicol. As ceftiofur acts as a penicillin-binding protein inhibitor, we obtained 21 strains of *S.*

Table 1: Identified metabolic pathways of *Streptococcus parauberis* on Kyoto Encyclopedia of Genes and Genomes (KEGG).

KEGG ID*	Pathways
stk00010	Glycolysis/gluconeogenesis—*Streptococcus parauberis*
stk00020	Citrate cycle (TCA cycle)—*Streptococcus parauberis*
stk00030	Pentose phosphate pathway—*Streptococcus parauberis*
stk00040	Pentose and glucuronate interconversions—*Streptococcus parauberis*
stk00051	Fructose and mannose metabolism—*Streptococcus parauberis*
stk00052	Galactose metabolism—*Streptococcus parauberis*
stk00053	Ascorbate and aldarate metabolism—*Streptococcus parauberis*
stk00061	Fatty acid biosynthesis—*Streptococcus parauberis*
stk00071	Fatty acid degradation—*Streptococcus parauberis*
stk00072	Synthesis and degradation of ketone bodies—*Streptococcus parauberis*
stk00130	Ubiquinone and other terpenoid-quinone biosynthesis—*Streptococcus parauberis*
stk00190	Oxidative phosphorylation—*Streptococcus parauberis*
stk00220	Arginine biosynthesis—*Streptococcus parauberis*
stk00230	Purine metabolism—*Streptococcus parauberis*
stk00240	Pyrimidine metabolism—*Streptococcus parauberis*
stk00250	Alanine, aspartate, and glutamate metabolism—*Streptococcus parauberis*
stk00260	Glycine, serine, and threonine metabolism—*Streptococcus parauberis*
stk00261	Monobactam biosynthesis—*Streptococcus parauberis*
stk00270	Cysteine and methionine metabolism—*Streptococcus parauberis*
stk00280	Valine, leucine, and isoleucine degradation—*Streptococcus parauberis*
stk00281	Geraniol degradation—*Streptococcus parauberis*
stk00290	Valine, leucine, and isoleucine biosynthesis—*Streptococcus parauberis*
stk00300	Lysine biosynthesis—*Streptococcus parauberis*
stk00310	Lysine degradation—*Streptococcus parauberis*
stk00330	Arginine and proline metabolism—*Streptococcus parauberis*
stk00332	Carbapenem biosynthesis—*Streptococcus parauberis*
stk00340	Histidine metabolism - *Streptococcus parauberis*
stk00350	Tyrosine metabolism - *Streptococcus parauberis*
stk00360	Phenylalanine metabolism - *Streptococcus parauberis*
stk00362	Benzoate degradation—*Streptococcus parauberis*
stk00380	Tryptophan metabolism—*Streptococcus parauberis*
stk00400	Phenylalanine, tyrosine, and tryptophan biosynthesis—*Streptococcus parauberis*
stk00430	Taurine and hypotaurine metabolism—*Streptococcus parauberis*
stk00440	Phosphonate and phosphinate metabolism—*Streptococcus parauberis*
stk00450	Selenocompound metabolism—*Streptococcus parauberis*
stk00460	Cyanoamino acid metabolism—*Streptococcus parauberis*
stk00471	D-Glutamine and D-glutamate metabolism—*Streptococcus parauberis*
stk00473	D-Alanine metabolism—*Streptococcus parauberis*
stk00480	Glutathione metabolism—*Streptococcus parauberis*
stk00500	Starch and sucrose metabolism—*Streptococcus parauberis*
stk00511	Other glycan degradation—*Streptococcus parauberis*
stk00520	Amino sugar and nucleotide sugar metabolism—*Streptococcus parauberis*
stk00521	Streptomycin biosynthesis—*Streptococcus parauberis*
stk00523	Polyketide sugar unit biosynthesis—*Streptococcus parauberis*
stk00525	Acarbose and validamycin biosynthesis—*Streptococcus parauberis*
stk00550	Peptidoglycan biosynthesis—*Streptococcus parauberis*
stk00561	Glycerolipid metabolism—*Streptococcus parauberis*
stk00562	Inositol phosphate metabolism—*Streptococcus parauberis*

Table 1: Continued.

KEGG ID*	Pathways
stk00564	Glycerophospholipid metabolism—*Streptococcus parauberis*
stk00590	Arachidonic acid metabolism—*Streptococcus parauberis*
stk00592	Alpha-linolenic acid metabolism—*Streptococcus parauberis*
stk00620	Pyruvate metabolism—*Streptococcus parauberis*
stk00622	Xylene degradation—*Streptococcus parauberis*
stk00625	Chloroalkane and chloroalkene degradation—*Streptococcus parauberis*
stk00626	Naphthalene degradation—*Streptococcus parauberis*
stk00627	Amino benzoate degradation—*Streptococcus parauberis*
stk00630	Glyoxylate and dicarboxylate metabolism—*Streptococcus parauberis*
stk00640	Propanoate metabolism—*Streptococcus parauberis*
stk00643	Styrene degradation—*Streptococcus parauberis*
stk00650	Butanoate metabolism—*Streptococcus parauberis*
stk00660	C5-Branched dibasic acid metabolism—*Streptococcus parauberis*
stk00670	One carbon pool by folate—*Streptococcus parauberis*
stk00680	Methane metabolism—*Streptococcus parauberis*
stk00730	Thiamine metabolism—*Streptococcus parauberis*
stk00740	Riboflavin metabolism—*Streptococcus parauberis*
stk00750	Vitamin B6 metabolism—*Streptococcus parauberis*
stk00760	Nicotinate and nicotinamide metabolism—*Streptococcus parauberis*
stk00770	Pantothenate and CoA biosynthesis—*Streptococcus parauberis*
stk00780	Biotin metabolism—*Streptococcus parauberis*
stk00790	Folate biosynthesis—*Streptococcus parauberis*
stk00900	Terpenoid backbone biosynthesis—*Streptococcus parauberis*
stk00910	Nitrogen metabolism—*Streptococcus parauberis*
stk00920	Sulfur metabolism—*Streptococcus parauberis*
stk00970	Aminoacyl-tRNA biosynthesis—*Streptococcus parauberis*
stk01040	Biosynthesis of unsaturated fatty acids—*Streptococcus parauberis*
stk01100	Metabolic pathways—*Streptococcus parauberis*
stk01110	Biosynthesis of secondary metabolites—*Streptococcus parauberis*
stk01120	Microbial metabolism in diverse environments—*Streptococcus parauberis*
stk01130	Biosynthesis of antibiotics—*Streptococcus parauberis*
stk01200	Carbon metabolism—*Streptococcus parauberis*
stk01210	2-Oxocarboxylic acid metabolism—*Streptococcus parauberis*
stk01212	Fatty acid metabolism—*Streptococcus parauberis*
stk01220	Degradation of aromatic compounds—*Streptococcus parauberis*
stk01230	Biosynthesis of amino acids—*Streptococcus parauberis*
stk01501	Beta-lactam resistance—*Streptococcus parauberis*
stk01502	Vancomycin resistance—*Streptococcus parauberis*
stk01503	Cationic antimicrobial peptide (CAMP) resistance—*Streptococcus parauberis*
stk02010	ABC transporters—*Streptococcus parauberis*
stk02020	Two-component system—*Streptococcus parauberis*
stk02024	Quorum sensing—*Streptococcus parauberis*
stk02060	Phosphotransferase system (PTS)—*Streptococcus parauberis*
stk03010	Ribosome—*Streptococcus parauberis*
stk03018	RNA degradation—*Streptococcus parauberis*
stk03020	RNA polymerase—*Streptococcus parauberis*
stk03030	DNA replication—*Streptococcus parauberis*
stk03060	Protein export—*Streptococcus parauberis*

Table 1: Continued.

KEGG ID*	Pathways
stk03070	Bacterial secretion system—*Streptococcus parauberis*
stk03410	Base excision repair—*Streptococcus parauberis*
stk03420	Nucleotide excision repair—*Streptococcus parauberis*
stk03430	Mismatch repair—*Streptococcus parauberis*
stk03440	Homologous recombination—*Streptococcus parauberis*
stk04122	Sulfur relay system—*Streptococcus parauberis*

*KEGG ID represents the molecular pathways within major metabolic pathways such as cellular processes, genetic information processes, metabolism, and drug and disease development. Each pathway has a three-letter organism code ("skt" for *Streptococcus parauberis* in KEGG database) followed by a five-digit number.

Table 2: Distribution of essential nonhomologous proteins in major metabolic pathways of *S. parauberis*.

Major metabolic pathway	% of essential nonhomologous proteins*
Metabolism of terpenoid and polyketides	4
Carbon metabolism	4
Cellular process	5
Metabolism of cofactors and vitamins	6
Nucleotide metabolism	6
Microbial metabolism in diverse environments	6
Energy metabolism	6
Biosynthesis of antibiotics	9
Lipid metabolism	11
Amino acid metabolism and biosynthesis	11
Carbohydrate metabolism	12
Biosynthesis of secondary metabolites	13
Glycan biosynthesis and metabolism	13
Environment information processing	19
Genetic information processing	31
Metabolic pathways	42

*Essential nonhomologous proteins of *S. parauberis* selected by NCBI-Blast and DEG database analyses.

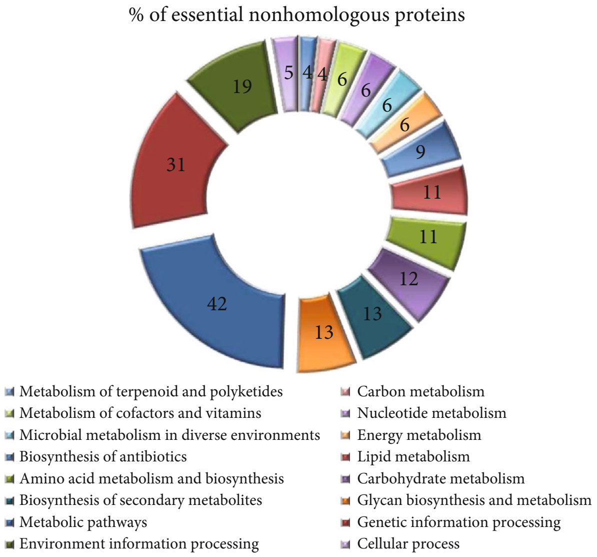


Figure 2: Distribution of essential nonhomologous proteins in major metabolic pathways of *S. parauberis*. Each color bar represents a single metabolic process. Essential nonhomologous protein distribution was analyzed using NCBI-Blast and the DEG database. The value noted in each bar represents the percentage of essential, nonhomologous proteins (out of 112 proteins as selected by DEG) involved in different metabolic pathways drawn manually from the KEGG pathway map.

parauberis isolated from fish and analyzed the functional effectiveness of identified targets in these strains by ceftiofur, and florfenicol was used for comparative effects.

2.4.1. Antimicrobial Susceptibility Profiles

(1) Minimal Inhibition Concentration (MIC) and Minimal Bactericidal Concentration (MBC). The broth microdilution method was applied to determine the susceptibility of ceftiofur sodium and florfenicol (both obtained from Sigma-Aldrich, St. Louis, MO, US) against *S. parauberis* (21 isolates and 1 KCTC 3651 strain) and *S. aureus* ATCC 29213. Generally, the MICs of the isolates were determined according to the method by the Clinical and Laboratory Standards Institute (2017), in cation-adjusted Mueller-Hinton broth (Ca-MHB; BD Bacto™) using 5×10^8 CFU/mL of bacterial concentrations [34]. The initial concentration of all antimicrobials was 1024 μg/mL in Ca-MHB, and the stock solutions were filtered using a 0.22 μm syringe filter (Merck Millipore Ltd., MA, US). The MIC_{50} values (at which 50% of the isolates were inhibited), MIC_{90} values, and MIC ranges of both antimicrobials against the isolates were determined. Bacteria were cultured overnight, adjusted with TH broth to an optical density of 0.1 at 600 nm (OD600), and aliquoted into 96-well plates. The antimicrobials were serially diluted twofold with MHB to give a concentration range of 512 μg/mL in 100 μL volumes. The growth and negative controls were prepared using Ca-MHB with and without bacteria, respectively. Microtiter plates were incubated at 30°C for 24 h. The MIC was determined as the lowest concentration at which noticeable growth was not observed. The MBC

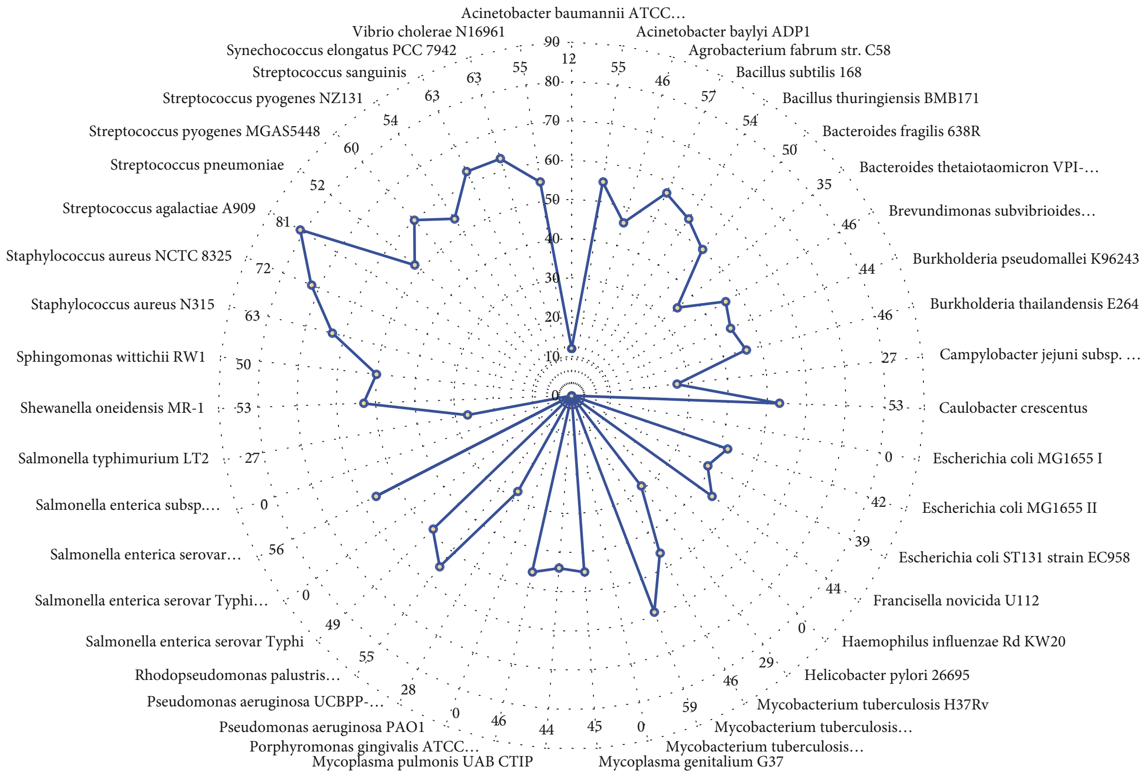


Figure 3: Frequency of hits of *S. parauberis* proteins by 43 bacteria in DEG. Distance from the center shows the extent of homologs of the 112 proteins by the essential proteins of 43 bacteria on DEG. Nonhomologous proteins of *S. parauberis* with homology were selected as essential and further characterized for selection of putative drug and vaccine targets.

(the lowest concentration that showed a 99.9% or higher killing rate) was determined by culturing 20 μL of the drug dilutions with samples from the 96-well plates on MHA plates and incubating overnight. The distribution of colonies on plates was inspected visually to assess the drug carryover effect. Studies were conducted in duplicate and were repeated at least twice on separate days.

(2) Mutation Prevention Concentration (MPC). The MPC determination for ceftiofur and florfenicol was performed as described previously [35, 36]. In summary, the tested bacteria were cultured, and after 24 h incubation in MHB, the bacterial suspensions were centrifuged at 5000g for 20 min and resuspended in 3 mL MHB to yield a concentration of 10^{10} CFU/mL. By plating, the serial dilutions of 100 mL of samples on a drug-free medium inoculum were further confirmed. After, agar plates containing seven different known concentrations of ceftiofur and florfenicol were inoculated with aliquots of 100 μL of *Streptococcus parauberis* strains (approx. 10^{10} CFU). *S. parauberis* KCTC 3651, the fully susceptible control strain, was used as a control in each experiment. The inoculated plates were incubated for 48 h at 30°C and screened visually for growth. The minimum antibiotic concentration with no bacterial colonies present was recorded as the MPC.

2.4.2. Time-Kill Curve Assay. To determine the killing dynamics of ceftiofur and florfenicol, a time-kill assay according to the CLSI (NCCLS, 1999) guidelines was performed [37]. In summary, exponentially growing overnight cultures of tested *S. parauberis* strains were diluted to ~5×10^8 CFU/mL bacterial densities and then inoculated to 5 mL brain-heart infusion (BHI) medium tubes. Later, ceftiofur and florfenicol were added to these tubes after bacterial dilutions at concentrations of 0.5, 1, 2, and 4x MIC, whereas one tube with no antibiotic was used as the control for growth. After incubating the samples at 30°C on a shaking incubator, 20 μL samples were taken from each tube at intervals of 0, 1, 2, 4, 6, 8, 12, and 24 h. Each sample was diluted 10-fold serially into aliquots and inoculated on BHI agar plates for the estimation of CFU per mL. For the two antibiotics, time-kill curves were obtained at different concentrations.

2.4.3. Biofilm Formation Quantification. For quantifying biofilm formation of *Streptococcus parauberis*, we followed the method of O'Toole (2011) with some bacterial growth modifications [38]. In summary, using the BHI broth, we prepared 100x dilutions from the 24 h inoculum of nine tested strains for determination of bacterial attachment and biofilm formation. The dilutions were transferred into 96-well microtiter polystyrene plates (Nunclmmuno™ MaxiSorp™, Nalgene, US) as eight replicates. After 48 h incubation at 30°C, we measured optical density at the wavelength of 595 nm and estimated the bacterial growth for both adherent as well as suspended cells [39]. We discarded the supernatant and rinsed the wells 3–4 times with 400 μL per rinse of distilled water. Following rinsing, wells were stained using

125 μL volume of 1% crystal violet solution. After 20 min, the dye was discarded and wells were washed using distilled water 3–4 times. The dye was dissolved using 125 μL of acetic acid (30% *v*/*v*), and we quantified the biofilm formation of tested strains by measuring OD595.

2.4.4. Minimum Biofilm Eradication Concentration (MBEC). For measurement of *S. parauberis* strain biofilm sensitivity to ceftiofur and florfenicol, the biofilms of the selected strains were grown using Calgary Biofilm Device (CBD; Innovotech, Calgary, Canada) microplates with slight modifications from the manufacturer's instructions. In summary, freshly grown overnight cultures of selected strains were diluted to 0.5 McFarland standard in broth, and 100 μL was transferred to 96-well flat-bottom microplates except for the negative control. Biofilm growth plates were immersed with modified CBD pegs and incubated to allow formation of biofilms for 48 h at 30°C. After rinsing, these biofilms containing peg lids with autoclaved distilled water were transferred to the antibiotic challenge plates with 100 μL of Ca-MHB with twofold dilutions of antibiotics in each well. After 24 h of incubation with antibiotics at 30°C, we repeated the same rinsing procedure and placed the peg lids into the biofilm recovery plate with Ca-MHB free of antibiotics. Biofilms were transferred from pegs to the wells by sonicating the recovery plates using a Branson 8200 sonicator (Emerson Electric Co., US) for 5 min at room temperature. Finally, the peg lids were replaced with standard lids, and the OD650 was recorded using a microplate colorimeter (VersaMax; Molecular Devices Corp., US). The lowest antibiotic dilution that caused inhibition of bacterial regrowth was recorded as the MBEC.

2.5. Statistical Analysis. The SAS statistical software version 9.4 (SAS Institute, Cary, NC, US) with one-way ANOVA was used for biofilm quantitation and growth of bacteria analysis, and Duncan's Multiple Range Test (DMRT) was used for evaluation of statistically significant differences among treatment groups, i.e., $p < 0.05$ was noted as statistically significant. Figure legends are used to designate statistics of various experiments.

3. Results

3.1. Comparative Genomic Analysis. For *Streptococcus parauberis*, 102 identified pathways were selected for analysis (Table 1). In comparison of the selected pathways of the pathogen with the existing 299 pathways of humans (*Homo sapiens*), 29 pathways were declared as unique to the pathogen. About 793 proteins are involved in these pathways, and out of these, only 277 were recognized as nonhomologues for both olive flounder and humans at the selected cutoff value of KEGG.

In *S. parauberis*, 112 nonhomologous proteins were identified as essential when its nonhomologous proteins were aligned with the essential protein sequences of five different streptococcal species named as *Streptococcus agalactiae A909*, *Streptococcus pneumoniae*, *Streptococcus pyogenes MGAS5448*, *Streptococcus pyogenes NZ131*, and *Streptococcus sanguinis*. These 112 essential nonhomologous proteins take part in 16 major metabolic pathways noted from the KEGG database as shown in Table 2 and Figure 2.

Table 3: *S. parauberis* nonhomologous essential proteins as subunits for vaccines with the number of transmembrane helices and antigenic characteristics predicted by TMHMM and VaxiJen databases, respectively.

Seq. KEGG ID	Genes	TMHMM*	Probable antigenicity**
STP_0829	mtlA	8	Antigen
STP_0495	atpB	5	Antigen
STP_0496	atpF	1	Antigen
STP_0274	pbpX	1	Antigen
STP_0314	dgkA	3	Antigen
STP_1416	bacA	8	Antigen
STP_1616	pbp1B	1	Antigen
STP_1749	pbp2A	1	Antigen
STP_0544	ltaS	5	Antigen
STP_1654	dppC	5	Antigen
STP_0122	lplB	4	Antigen
STP_0799	pstA	4	Antigen
STP_1210	lplB	6	Antigen
STP_1444	ecfT	4	Antigen
STP_0327	secG	2	Antigen
STP_1690	yidC, spoIIIJ, OXA1, ccfA	5	Antigen

*Number of transmembrane helices for listed membrane proteins predicted by TMHMM (version 2.0) database. **Probable antigenicity (protective antigens and vaccine subunits) predicted by VaxiJen database (version 2.0) with threshold value 0.4.

Moreover, we compared all the essential proteins of the 43 available bacteria in the DEG with no selected hits. As described in Figure 3, the least hit was recorded by *Acinetobacter baumannii* ATCC 17978 with a homology of only 12 proteins, whereas the highest homology was recorded with the essential proteins of *Streptococcus agalactiae A909* with hits of 81 proteins. On the other hand, *Escherichia coli* strain MG1655 I, strain of *Mycobacterium tuberculosis* H37Rv III, *Salmonella enterica* Serovar *Typhi Ty Paoi*, *Pseudomonas aeruginosa* PAO1, *Haemophilus influenzae* RdKW20, *Salmonella enterica* subsp. *enterica serovar*, and *Typhimurium str.* 14028S each had no homology with the nonhomologous proteins of the pathogen under investigation.

The CELLO database analysis found the presence of 77 cytoplasmic, 24 membrane, and four extracellular proteins. Moreover, five proteins were cytoplasmic as well as membrane-bound and two proteins were found in three regions of the pathogen cell (cytoplasm, membrane, and extracellular area). Similarly, according to the TMHMM database, about 26 proteins had transmembrane helices but only six were in common with the CELLO database prediction (Table S1). Furthermore, out of 26 transmembrane proteins, only 16 had an antigenic probability greater than 0.4 (Table 3).

Each essential protein's molecular weight was determined by referring to the UniProt database. As shown in

Table 4: *S. parauberis* nonhomologous essential proteins with druggability for FDA-approved drugs as inferred from the DrugBank database using BLASTP and the list of FDA-approved drugs for the targets.

KEGG	Name	Gene	DB* ID	Drug name	Drug group
STP_0292	5′-Methylthioadenosine/S-adenosylhomocysteine nucleosidase	mtnN	DB02158	(1s)-1-(9-Deazaadenin-9-Yl)-1,4,5-trideoxy-1,4-imino-5-methylthio-D-ribitol	E
			DB02281	Formycin	E
			DB00173	Adenine	A, N
			DB02933	5′-Deoxy-5′-(methylthio)-tubercidin	E
			DB08606	(3R,4S)-1-[(4-Amino-5H-pyrrolo[3,2-D]pyrimidin-7-YL) methyl]-4-[(methylsulfanyl) methyl] pyrrolidin-3-OL	E
STP_1093	Penicillin-binding protein 2B	penA	DB01066	Cefditoren	A
			DB01212	Ceftriaxone	A
			DB01140	Cefadroxil	A, VA, W
			DB00493	Cefotaxime	A
			DB00319	Piperacillin	A
			DB00607	Nafcillin	A
			DB00415	Ampicillin	A, VA
			DB00485	Dicloxacillin	A, VA
			DB01163	Amdinocillin	W
			DB01603	Meticillin	A
			DB00456	Cefalotin	A, VA
			DB00713	Oxacillin	A
			DB01331	Cefoxitin	A
			DB00567	Cephalexin	A, VA
			DB03313	Cephalosporin C	E
			DB08795	Azidocillin	A
			DB00739	Hetacillin	A, VA, W
STP_1616	Penicillin-binding protein 2	pbp2	DB04147	Lauryl dimethylamine-N-oxide	E
STP_1749	Penicillin-binding protein 2	pbp2	DB04147	Lauryl dimethylamine-N-oxide	E
	Penicillin-binding protein 1B	mrcB	DB01598	Imipenem	A
			DB01329	Cefoperazone	A
			DB01332	Ceftizoxime	A
			DB01327	Cefazolin	A
			DB01331	Cefoxitin	A
			DB01328	Cefonicid	A
			DB01415	Ceftibuten	A
			DB00430	Cefpiramide	A
			DB00438	Ceftazidime	A
			DB00274	Cefmetazole	A
			DB00303	Ertapenem	A, I
			DB01414	Cefacetrile	A
			DB04570	Latamoxef	A
			DB06211	Doripenem	A, I
			DB11367	Cefroxadine	W
	Penicillin-binding protein 1A	mrcA	DB01598	Imipenem	A
			DB01329	Cefoperazone	A
			DB01332	Ceftizoxime	A
			DB01333	Cefradine	A
			DB01327	Cefazolin	A
			DB01331	Cefoxitin	A

Table 4: Continued.

KEGG	Name	Gene	DB* ID	Drug name	Drug group
			DB01328	Cefonicid	A
			DB01415	Ceftibuten	A
			DB00430	Cefpiramide	A
			DB00438	Ceftazidime	A
			DB00274	Cefmetazole	A
			DB00303	Ertapenem	A, I
			DB01414	Cefacetrile	A
			DB04570	Latamoxef	A
			DB06211	Doripenem	A, I
STP_0791	Pantothenate kinase	coaA	DB01783	Pantothenic acid	N, VA
			DB01992	Coenzyme A	N
			DB04395	Phosphoaminophosphonic acid-adenylate ester	E
STP_0603	Isopentenyl-diphosphate delta-isomerase	fni	DB03247	Riboflavin monophosphate	E

*DB: DrugBank database; E: experimental, A: approved; VA: veterinary approved; W: withdrawn.

Table 2 and Table S1, out of 112 nonhomologous essential proteins, only one protein showed a greater weight than the selected limit of molecular weight (<110 kDa) although all of them had 3D experimental models. The druggability results of the essential proteins as determined by the DrugBank database identified six proteins, which had hits for drugs that are approved, nutraceutical, investigational, and experimental at an *e* value limit of 10^{-25} (Table 4).

The six genes for these essential proteins with a hit from the DrugBank database were mtnN ($5'$-methylthioadenosine/S-adenosylhomocysteine nucleosidase), penA (penicillin-binding protein 2B), pbp2 (penicillin-binding protein 2), *murB* (penicillin-binding protein 1B), murA (penicillin-binding protein 1A), coaA (pantothenate kinase), and fni (isopentenyl-diphosphate delta-isomerase). The number of hits increased with an increase in expectation value, as 10 and 19 essential nonhomologous proteins were hits for expectation value limits of 10^{-10} and 10^{-5}, respectively.

3.2. In Vitro Efficacy of Identified Proteins Targeted by Ceftiofur and Florfenicol

3.2.1. In Vitro Antibiotic Sensitivity Profiles including Minimal Inhibition Concentration (MIC), Minimal Bactericidal Concentration (MBC), and Mutant Prevention Concentration (MPC). The antimicrobial susceptibility profile against 22 field isolates from fish and a known strain (KCTC 3651) of *S. parauberis* was determined for the selected veterinary antimicrobial drugs. Ceftiofur sodium has an antimicrobial mechanism in common with the one of the identified therapeutic targets (PBP 2A) from the genomic analysis, and for comparative purposes, florfenicol was used.

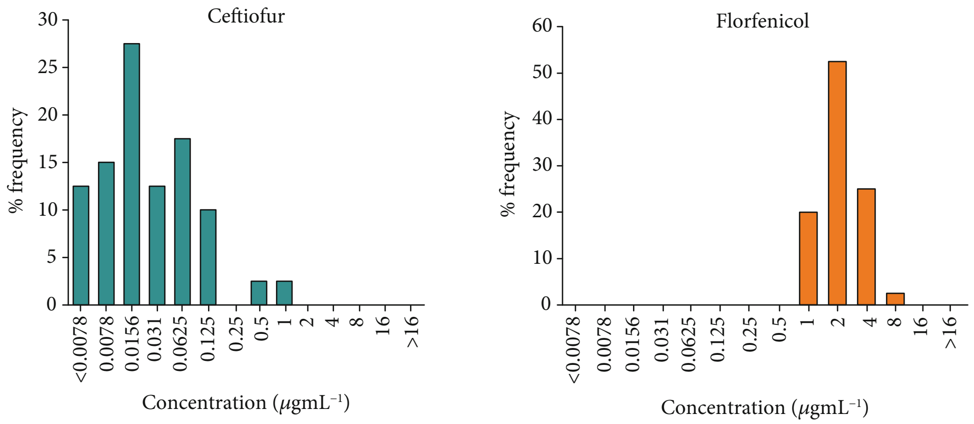
Table 5: Minimum inhibitory concentrations of ceftiofur and florfenicol against olive flounder isolated *Streptococcus parauberis*.

Parameters	Antimicrobial drugs	
	Ceftiofur	Florfenicol
MIC$_{50}$ (μg/mL)	0.0156	2
MIC$_{90}$ (μg/mL)	0.125	8
MIC$_{\text{Range}}$ (μg/mL)	0.0039-1	0.5-8
MBC$_{\text{Range}}$ (μg/mL)	0.0078-32	1-128
IC$_{50\ \text{Range}}$ (μg/mL)	0.001-0.5	0.7-2.7
R (%)	0	0
KCTC 3651 (μg/mL)	0.0078	0.5
S. aureus ATCC 29213 (μg/mL)	1	2
CLSI range for *Streptococcus viridans*		
S	≤1	≤4
R	≥4	≥16
CLSI range for *Staphylococcus sp.*		
S	≤1	≤8
R	≥4	≥32

MIC: minimum inhibitory concentration; MBC: minimum bactericidal concentration; R: rate of resistance; CLSI range: clinical breakpoints for *Streptococcus* and *Staphylococcus* sp. as defined by the Clinical and Laboratory Standards Institute.

According to the MIC and MBC testing results, all isolates were completely sensitive to ceftiofur and florfenicol with MICs ranging from 0.0039 to 1 μg/mL and 0.5 to 8 μg/mL, respectively. The MBC values against field isolates ranged from 0.0078 to 32 μg/mL and 1 to 128 μg/mL for ceftiofur and florfenicol, respectively, as shown in Table 5 and Figure 4. The MPC for both antibiotics was evaluated in triplicate, and the MPCs of ceftiofur and florfenicol were more than the twofold dilution of their MIC. Hence, the MPC/MIC ratio for ceftiofur was 8–32, but in reference to florfenicol, this ratio was noted as eight (Table 6).

3.2.2. Time-Kill Curve Assays and IC$_{50}$ Values. The inhibitory effects of ceftiofur and florfenicol were determined against *S. parauberis* using a time-kill assay by the growth inhibition



(a)

(b)

Figure 4: Minimum inhibitory concentration (MIC) frequencies. MIC frequencies observed for ceftiofur (a) and florfenicol (b) against 22 *Streptococcus parauberis* isolated strains from diseased olive flounder.

Table 6: Minimum inhibitory concentration and mutant prevention concentration comparison for ceftiofur and florfenicol against field and known strains of *S. parauberis*.

Strains		Ceftiofur			Florfenicol	
	MIC (μg/mL)	MPC (μg/mL)	MPC/MIC	MIC (μg/mL)	MPC (μg/mL)	MPC/MIC
S. parauberis 2628	1	32	32	2	16	8
S. parauberis KCCM3651	0.0078	0.0624	8	0.5	4	8

MIC: minimum inhibitory concentration; MPC: mutant prevention concentration.

method. The bacterial incubation was performed with each tested antibiotic at concentrations of 0.5, 1, 2, and 4x MIC. The killing dynamics of each antimicrobial showed that ceftiofur started to inhibit the tested strains after 4 hours, and at 12 hours, this inhibition was maximal. However, in the case of florfenicol, inhibitory activity also began after 4 hours, but maximum inhibition was noted at 24 hours. Moreover, the inhibition of bacterial growth by florfenicol was less than that of ceftiofur at 0.5x MIC (Figure 5).

By comparing the counts of bacteria after 12 hours of incubation using each antibiotic, the IC_{50} values were determined to obtain the growth inhibition concentrations of ceftiofur and florfenicol. The concentration of ceftiofur and florfenicol for 50% growth inhibition ranged from 0.001 to 0.5 μg/mL and 0.7 to 2.7 μg/mL, respectively (Table 5).

3.2.3. Biofilm Quantitation and Minimum Biofilm Eradication Concentration (MBEC). For quantitation of bacterial growth and biofilm formation activities of seven *S. parauberis* strains, BHI media and 0.5% glucose-aided BHI media were used. *S. aureus* ACTC 29213 was used as the positive control for biofilm formation (Figure 6). The growth of bacteria in BHI media ranged from OD values of 0.5 to 1.2. However, this range for growth of bacteria in 0.5% glucose-aided media was 2.1–2.4. For *S. aureus* ACTC 29213 at OD_{550}, biofilm formation in 0.5% glucose-aided media was 0.1. Although the biofilm formation in BHI for other strains was <0.05, in glucose-aided BHI media, the formation of biofilm was moderate for all strains in the range of 0.1–0.2. If the absorbance was measured at <0.10 at OD_{550}, the formation of biofilm was presumed as weak/absent, between 0.1 and 1.0, it was considered moderate, and if it was >1.0, it was determined as a strong biofilm.

The minimum biofilm eradication concentration (MBEC) was measured to check the ceftiofur and florfenicol susceptibility against biofilm-forming isolates of *S. parauberis* as shown in Table 7. Larger differences between the MBEC/MIC ratios show higher susceptibility differences of biofilm-forming bacteria versus planktonic bacteria towards ceftiofur and/or florfenicol.

4. Discussion

The increase in bacterial pathogenicity and resistance to antibiotics has provoked the interest of researchers for new studies of health and pathogenic bacterial species. Around the globe, scientists and researchers are paying more attention to novel therapeutic targets for preventing resistance to bacterial infection. The availability of vast computational parameters and -omic data has increased the identification of suitable therapeutic agents. Inhibition of essential proteins can stop bacterial growth, as they are important for survival of the bacteria. Systematic comparative analyses of *Streptococcus parauberis* found 112 potential vaccine and drug targets in the DrugBank, out of which six essential targets had druggability due to homology. We further characterized these targets using different databases to differentiate between potential drug and vaccine targets. The drug

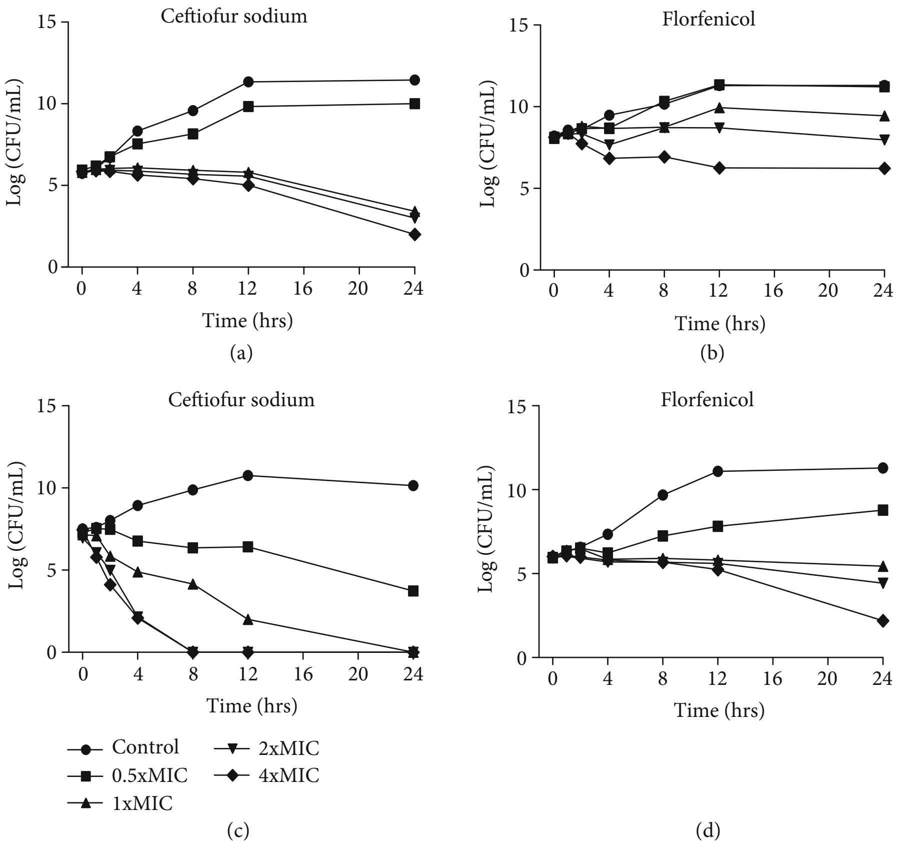


Figure 5: Time-kill curves of ceftiofur (a, c) and florfenicol (b, d) against *S. parauberis* KCTC 3651 strain (top) and *S. parauberis* S2628 strain (bottom). (a, c) represent inhibitory activities of ceftiofur at 0, 1, 2, 4, 8, 12, and 24 hours, and (b, d) represent the same for florfenicol when the drugs were exposed to exponentially growing *S. parauberis* KCTC 3651 and S2628 strains at their 0.5, 1, 2, and 4x minimum inhibitory concentration (MIC) values.

targets in question are involved in different cellular activities like metabolism of purines and pyrimidines, replication of DNA, ribosomal synthesis, ABC transporter pathway, and nucleotide metabolism.

Apart from proteins that could be targeted, we found several metabolic pathways involved in bacterial resistance against various antibiotics. Our targeted pathogen can resist the action of the antibiotic vancomycin (VCM). The vancomycin resistance pathway expression is induced by two component systems, i.e., VanS-VanR and D-Ala-D-Lac or D-Ala-D-Ser depsipeptides which can replace the D-Ala-D-Ala dipeptide, resulting in inhibition of vancomycin binding to pentadepsipeptides [D-Ser] or [D-Lac]. The variation in D-Ala-D-Lac and D-Ala-D-Ser indicates high and low levels of resistance to VCM, respectively. Peptidoglycan modifications lead to successive cell wall formation, and this may affect resistance in the present organism [40].

The pathway for cationic antimicrobial peptide (CAMP) resistance is also present. In host defense mechanisms, CAMPs play a pivotal role against microorganisms as a component of the innate immune response. In fact, CAMPs kill bacterial cells by deterioration of the integrity of bacterial inner and outer membranes. However, some bacteria have developed several resistance mechanisms like efflux pumps in membranes, external trapping mechanisms, substitution of anionic cell surface contents with cations, crosslinking and biosynthesis of cell surface components, and peptidase production of CAMPs. Similarly, *Streptococcus parauberis* can develop resistance against CAMPs [41, 42].

Strains of *Streptococcus parauberis* have previously shown resistance against different antimicrobial drugs such as tetracycline, oxytetracycline, and erythromycin [43, 44]. Because of the chance of antibiotic resistance, there is a need for new and alternative therapeutic targets against this pathogenic bacterium. In the past, the biology of microbial agents has limited the identification of new antimicrobial and vaccinating agents. However, advancement in proteomic and genomic knowledge has led to milestones in investigating new vaccines and for finding targets for effective therapeutic agents [10].

The essential nonhomologous proteins of *S. parauberis* were recognized as important potential targets, providing new perspectives on therapeutic targets in the pathogen pathways with both safety and specificity. Using different therapeutic and vaccine agents against those genomic and proteomic parts of the pathogenic bacteria that are essential for its reproduction is critical during potential drug design. These agents can interact or affect the normal replication of the targeted microorganism [45].

After a protein is established as essential, it is important to determine the protein characteristics and localization within the cell. This can assist in understanding its function

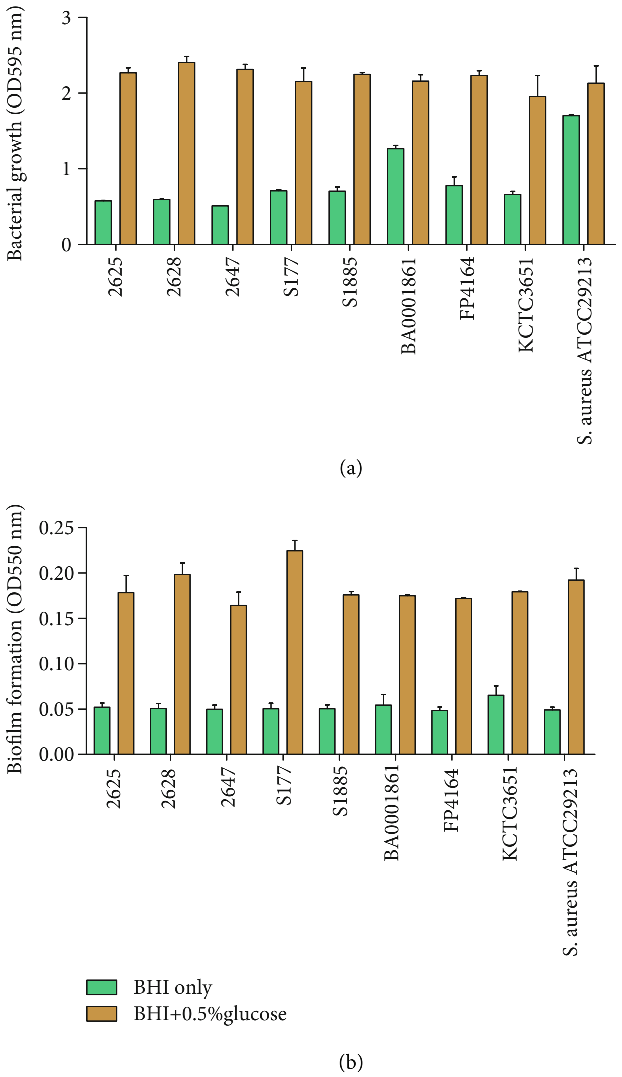


Figure 6: Quantitation of bacterial growth and biofilm formation in *Streptococcus parauberis* strains. Bacterial growth (a) vs. biofilm formation assay (b) of planktonic bacteria in brain heart infusion (BHI) and 0.5% glucose-supplemented BHI media. *S. aureus* is used as the positive control for biofilm formation whereas medium only is used as the normal control (NC). The optical density (OD) was read at wavelength of 550 nm. The absorbance was noted as <0.1 for weak biofilm formation, between 0.1 and 1 as moderate biofilm, and ≥1.0 as strong biofilm.

and nature for therapeutic targeting either as a vaccine or as a target for the antimicrobial function [46, 47]. Therefore, the identified cytoplasmic proteins can be used for targeting by antimicrobial drugs whereas the suggested membrane proteins with transmembrane helices can act as selected toxins or surface-exposed proteins and targeted by vaccine production, and these vaccines can initiate the immune response mediated by antibodies [48, 49].

The low molecular weight and druggability of these essential proteins noted according to the DrugBank database assisted in filtering possible targets. In addition to these findings, some nutraceutical, experimental, investigational, and approved therapeutic agents were assessed against the binding potential of essential nonhomologous proteins of *S. parauberis* and proved that these proteins are druggable and can be used as potential therapeutic targets. Moreover, different combinations of these drugs may be used to treat streptococcal infections of aquaculture habitats.

In this study, we identified twenty-nine FDA-approved drugs with a hit from the essential proteins of *S. parauberis*, out of which five were veterinary-approved drugs. The six identified genes targeted by these drugs are reported to have vital roles in bacterial metabolism. As noted, the mtnN gene is related to the methylthioadenosine/S-adenosylhomocysteine (MTA/SAH) nucleosidase, and its importance in bacteria has been appreciated previously. By inclusive analysis of its various roles, it is an integral component of the activated methyl cycle, which recycles adenine and methionine through S-adenosylmethionine- (SAM-) mediated methylation reactions, and also produces the universal quorum-sensing signal, autoinducer-2 (AI-2) [50]. Furthermore, murA, murB, penA, and pbp2 genes produce penicillin-binding protein types 1A, 1B, 2B, and 2, respectively, and these proteins are membrane carboxypeptidases and transpeptidases. Peptidases are required for regulation of chain length, glycan subunit polymerization, and muropeptide cross-linkages [51]. The pantothenate kinase (coaA) gene is related to the CoA biosynthesis pathway in bacteria and mammals. Pantothenate kinase is a key regulator of biosynthesis and directs the intracellular concentration of CoA through feedback regulation by CoA and its thioesters [52]. The isopentenyl-diphosphate delta-isomerase (fni) gene is related to isoprenoid metabolism. Isoprenoids play an important role in all living organisms, in mammals as steroid hormones, in plants as carotenoids, and in bacteria as ubiquinones or menaquinones. Isoprenoids are synthesized by consecutive condensations of the five-carbon precursor isopentenyl diphosphate (IPP) to its isomer dimethyl-allyl diphosphate (DMAPP). Isopentenyl diphosphate delta isomerase catalyzes an essential reaction in the biosynthesis of isoprenoids by converting IPP to DMAPP [53].

As most of the identified therapeutic targets play a pivotal role in cellular metabolism, by developing new efficacious therapeutic agents in a synchronized way, we can potentially control the infections caused by *S. parauberis*. Moreover, by using the results of this study, we can make significant innovations in testing the efficacy of available antimicrobial drugs. Here, to check the therapeutic efficacy of one of the identified targets in this study, we used an approved veterinary drug, ceftiofur, which inhibits the transpeptidation step of peptidoglycan synthesis during the formation of cell walls by binding to penicillin-binding proteins. For comparative purposes, we also studied florfenicol [15].

The *in vitro* results for antibiotic sensitivity against the tested strains revealed the greatest effects of ceftiofur sodium against fish isolates of *S. parauberis* in comparison with florfenicol. All tested strains of *S. parauberis* showed 100% susceptibility to ceftiofur and florfenicol with an MIC range of 0.0039–1 μg/mL and 0.5–8 μg/mL, respectively. These results are common with other reported susceptibility studies of

Table 7: Minimum inhibitory concentration and minimum biofilm eradication concentration comparison for ceftiofur and florfenicol against field and known strains of *S. parauberis*.

Strain	Ceftiofur			Florfenicol		
	MIC (μg/mL)	MBEC* (μg/mL)	MBEC/MIC ratio	MIC (μg/mL)	MBEC (μg/mL)	MBEC/MIC ratio
S. parauberis 2628	1	4	4	2	8	4
S. parauberis S177	0.125	256	2048	4	16	4
S. parauberis S1885	0.0039	2	512	1	64	64
S. parauberis KCTC 3651	0.0078	32	4102	0.5	64	128
S. aureus ATCC 29213	1	256	256	2	64	32

MIC: minimum inhibitory concentration; MBEC: minimum biofilm eradication concentration.

ceftiofur and florfenicol against streptococcal strains isolated from different sources [54–56].

Ceftiofur sodium and florfenicol have time-dependent inhibitory activities against *S. parauberis* strains, as bacterial growth was inhibited after 4 hours of incubation, irrespective of the antibiotic concentrations. Based on the minimum and maximum inhibitory concentrations, i.e., MIC_{50} and MIC_{90} (the concentration at which about 90 percent of the tested strains were inhibited), all tested bacterial strains were highly susceptible. We illustrated that ceftiofur has a minimum bactericidal concentration (MBC) 2–16 times greater than the MIC against *S. parauberis*, which confirmed reports by other authors [13, 16, 57, 58]. Although the IC_{50} range of ceftiofur was less than its MIC value, the IC_{50} value obtained from the time-kill curve analysis can assist in predicting the *in vivo* antimicrobial efficacy and dosage regimen according to the concentration change over time [59].

For environmental survival, one of the most important resistance mechanisms is the formation of biofilm by bacterial strains. Biofilms consist of aggregates of adherent bacteria in joint composition with proteins, polysaccharides, DNA, and lipids called an extracellular polymeric matrix. Different fish pathogenic bacteria including a few species of streptococcal bacteria like *S. mutants* have biofilm-forming potential [60–62].

Based on the present findings, the same characteristics of biofilm formation were identified within *S. parauberis* strains. Biofilm formation was enhanced with the addition of 0.5% glucose in BHI media as a source of carbohydrate. Biofilms work as the reservoir for survival of bacteria; thus, in this form, bacterial resistance to the antibiotics was increased. Due to the altered behavior of biofilm communities of bacteria, the antimicrobial susceptibility of biofilm-forming strains of *S. parauberis* was determined by the minimum biofilm eradication concentration (MBEC). The MBEC of ceftiofur was >1000 times higher than its MIC, which agrees with previous reports suggesting that the planktonic form of bacteria is 10–1000 times less resistant than bacteria of biofilm communities. In order to obtain maximum therapeutic outcomes and reduce antibiotic resistance, it is crucial to optimize the dosage of available antibiotics, and as such, agents for treatment of resistant bacterial infections are limited. Moreover, to achieve therapeutically effective antimicrobials, the application of biofilm-forming bacteria during *in vitro* susceptibility studies is more effective [63, 64].

In conclusion, our *in vitro* susceptibility studies found ceftiofur as an effective antibiotic against both planktonic as well as biofilm-forming strains of pathogenic *Streptococcus parauberis* isolated from fish. Moreover, we found that one of our identified target pathways was efficacious against all tested *S. parauberis* strains. In the future, more *in vitro* and *in vivo* findings will play a crucial role in validation of the present findings and in testing of other available antimicrobials against *S. parauberis*, for elucidation of efficacious and safe therapeutic agents. Furthermore, using available data for other identified pathways, we can develop more specific and potent novel therapeutic agents against *S. parauberis*.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

The funding agencies had no role in study design, data collection, and interpretation or the decision to submit the work for publication. The manuscript was based, in part, on the first author's doctoral dissertation at Kyungpook National University.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was financially supported by the Ministry of Small and Medium-sized Enterprises (SMEs) and Startups (MSS), Korea, under the "Regional Specialized Industry Development Program (R&D, P0002954)" supervised by the Korea Institute for Advancement of Technology (KIAT), and the National Research Foundation of Korea (NRF) as funded by the Ministry of Education (2019R1A2C2006277).

Supplementary Materials

Table S1: prioritization summary of 112 essential nonhomologous proteins of S. parauberis based on TMHMM, BLASTP, CELLO, PDB, and ModBase database results. *(Supplementary Materials)*

References

[1] S. W. Nho, J. I. Hikima, I. S. Cha et al., "Complete genome sequence and immunoproteomic analyses of the Bacterial fish pathogen *Streptococcus parauberis*," *Journal of Bacteriology*, vol. 193, no. 13, pp. 3356–3366, 2011.

[2] A. M. Williams and M. D. Collins, "Molecular taxonomic studies on *Streptococcus uberis* types I and II. Description of *Streptococcus parauberis sp. nov.*," *Journal of Applied Microbiology*, vol. 68, no. 5, pp. 485–490, 1990.

[3] J. M. Nieto, S. Devesa, I. Quiroga, and A. E. Toranzo, "Pathology of *Enterococcus* sp. infection in farmed turbot, *Scophthalmus maximus* L.," *Journal of Fish Diseases*, vol. 18, no. 1, pp. 21–30, 1995.

[4] A. Eldar and C. Ghittino, "*Lactococcus garvieae* and *Streptococcus iniae* infections in rainbow trout *Oncorhynchus mykiss*: similar, but different diseases," *Diseases of Aquatic Organisms*, vol. 36, no. 3, pp. 227–231, 1999.

[5] G. W. Shin, K. J. Palaksha, H. H. Yang et al., "Discrimination of streptococcosis agents in olive flounder (*Paralichthys olivaceus*)," *Bulletin-European Association of Fish Pathologists*, vol. 26, no. 2, p. 68, 2006.

[6] G. W. Baeck, J. H. Kim, D. K. Gomez, and S. C. Park, "Isolation and characterization of Streptococcus sp. from diseased flounder (*Paralichthys olivaceus*) in Jeju Island," *Journal of Veterinary Science*, vol. 7, no. 1, pp. 53–58, 2006.

[7] A. Eldar, P. F. Frelier, L. Assenta, P. W. Varner, S. Lawhon, and H. Bercovier, "*Streptococcus shiloi*, the name for an agent causing septicemic infection in fish, is a junior synonym of *Streptococcus iniae*," *International Journal of Systematic and Evolutionary Microbiology*, vol. 45, no. 4, pp. 840–842, 1995.

[8] A. Haines, E. Nebergall, E. Besong, K. Council, O. Lambert, and D. Gauthier, "Draft genome sequences for seven *Streptococcus parauberis* isolates from wild fish in the Chesapeake Bay," *Genome Announcements*, vol. 4, no. 4, 2016.

[9] S. Pereyre, P. Sirand-Pugnet, L. Beven et al., "Life on arginine for *Mycoplasma hominis*: clues from its minimal genome and comparison with other human urogenital mycoplasmas," *PLoS Genetics*, vol. 5, no. 10, article e1000677, 2009.

[10] D. Damte, J. W. Suh, S. J. Lee, S. B. Yohannes, M. A. Hossain, and S. C. Park, "Putative drug and vaccine target protein identification using comparative genomic analysis of KEGG annotated metabolic pathways of *Mycoplasma hyopneumoniae*," *Genomics*, vol. 102, no. 1, pp. 47–56, 2013.

[11] M. M. Parvege, M. Rahman, and M. S. Hossain, "Genome-wide analysis of *Mycoplasma hominis* for the identification of putative therapeutic targets," *Drug Target Insights*, vol. 8, 2014.

[12] M. Y. Galperin and E. V. Koonin, "Searching for drug targets in microbial genomes," *Current Opinion in Biotechnology*, vol. 10, no. 6, pp. 571–578, 1999.

[13] R. E. Hornish and S. F. Katarski, "Cephalosporins in veterinary medicine-ceftiofur use in food animals," *Current Topics in Medicinal Chemistry*, vol. 2, no. 7, pp. 717–731, 2002.

[14] B. A. Dixon and G. S. Issvoran, "The activity of ceftiofur sodium for *Aeromonas spp.* isolated from ornamental fish," *Journal of Wildlife Diseases*, vol. 28, no. 3, pp. 453–456, 1992.

[15] K. A. Al-Kheraije, "Studies on the antibacterial activity of ceftiofur sodium in vitro and birds," *Open Journal of Veterinary Medicine*, vol. 3, no. 1, pp. 16–21, 2013.

[16] R. S. Singer, S. K. Patterson, and R. L. Wallace, "Effects of therapeutic ceftiofur administration to dairy cattle on *Escherichia coli* dynamics in the intestinal tract," *Applied and Environmental Microbiology*, vol. 74, no. 22, pp. 6956–6962, 2008.

[17] A. E. Abd-Ellateif and I. M. G. El-Din, "The role of ceftiofur sodium (Excenel) in the control of *Pasteurella multocida* infection in chickens," *Proceeding of the*, pp. 632–645, 1998.

[18] B. T. Lunestad and O. Samuelsen, "4 - Veterinary drug use in aquaculture," in *Improving Farmed Fish Quality and Safety*, pp. 97–127, Woodhead Publishing, 2008.

[19] B. T. Birhanu, S. J. Lee, N. H. Park, J. B. Song, and S. C. Park, "*In silico* analysis of putative drug and vaccine targets of the metabolic pathways of *Actinobacillus pleuropneumoniae* using a subtractive/comparative genomics approach," *Journal of Veterinary Science*, vol. 19, no. 2, pp. 188–199, 2018.

[20] M. Kanehisa, S. Goto, M. Hattori et al., "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, vol. 34, no. 90001, Supplement 1, pp. D354–D357, 2006.

[21] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Research*, vol. 38, Supplement 1, pp. D355–D360, 2009.

[22] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[23] A. M. Butt, I. Nasrullah, S. Tahir, and Y. Tong, "Comparative genomics analysis of Mycobacterium ulcerans for the identification of putative essential genes and Therapeutic Candidates," *PLoS One*, vol. 7, no. 8, article e43080, 2012.

[24] R. Zhang, H. Y. Ou, and C. T. Zhang, "DEG: a database of essential genes," *Nucleic Acids Research*, vol. 32, Supplement 1, pp. D271–D272, 2004.

[25] D. Barh, S. Tiwari, N. Jain et al., "*In silico* subtractive genomics for target identification in human bacterial pathogens," *Drug Development Research*, vol. 72, no. 2, pp. 162–177, 2011.

[26] C. S. Yu, Y. C. Chen, C. H. Lu, and J. K. Hwang, "Prediction of protein subcellular localization," *Proteins: Structure, Function, and Bioinformatics*, vol. 64, no. 3, pp. 643–651, 2006.

[27] A. Krogh, B. Larsson, G. Von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001.

[28] M. Duffield, I. Cooper, E. McAlister, M. Bayliss, D. Ford, and P. Oyston, "Predicting conserved essential genes in bacteria: *in silico* identification of putative drug targets," *Molecular BioSystems*, vol. 6, no. 12, pp. 2482–2489, 2010.

[29] H. M. Berman, J. Westbrook, Z. Feng et al., "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[30] U. Pieper and M. Benjamin, "MODBASE, a database of annotated comparative protein structure models and associated resources," *Nucleic Acids Research*, vol. 42, no. D1, pp. D336–D346, 2014.

[31] I. A. Doytchinova and D. R. Flower, "VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines," *BMC Bioinformatics*, vol. 8, no. 1, p. 4, 2007.

[32] I. A. Doytchinova and D. R. Flower, "Bioinformatic approach for identifying parasite and fungal candidate subunit

vaccines," *The Open Vaccine Journal*, vol. 1, no. 1, pp. 22–26, 2008.

[33] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a comprehensive resource for *in silico* drug discovery and exploration," *Nucleic Acids Research*, vol. 34, no. 90001, pp. D668–D672, 2006.

[34] Clinical and Laboratory Standards Institute, *Methods for Determining Bactericidal Activity of Antimicrobial Agents; Approved Guideline M26-A*, Clinical and Laboratory Standards Institute, Wayne, PA, USA, 1999.

[35] J. M. Blondeau, X. Zhao, G. Hansen, and K. Drlica, "Mutant prevention concentrations of fluoroquinolones for clinical isolates of *Streptococcus pneumoniae*," *Antimicrobial Agents and Chemotherapy*, vol. 45, no. 2, pp. 433–438, 2001.

[36] Y. Dong, X. Zhao, B. N. Kreiswirth, and K. Drlica, "Mutant prevention concentration as a measure of antibiotic potency: studies with clinical isolates of Mycobacterium tuberculosis," *Antimicrobial Agents and Chemotherapy*, vol. 44, no. 9, pp. 2581–2584, 2000.

[37] C. F. Wei, J. H. Shien, S. K. Chang, and C. C. Chou, "Florfenicol as a modulator enhancing antimicrobial activity: example using combination with thiamphenicol against *Pasteurella multocida*," *Frontiers in Microbiology*, vol. 7, p. 389, 2016.

[38] G. A. O'Toole, "Microtiter dish biofilm formation assay," *Journal of Visualized Experiments*, no. 47, 2011.

[39] K. L. Frank, E. J. Reichert, K. E. Piper, and R. Patel, "*In vitro* effects of antimicrobial agents on planktonic and biofilm forms of *Staphylococcus lugdunensis* clinical isolates," *Antimicrobial Agents and Chemotherapy*, vol. 51, no. 3, pp. 888–895, 2007.

[40] P. Courvalin, "Vancomycin resistance in gram-positive cocci," *Clinical Infectious Diseases*, vol. 42, Supplement 1, pp. S25–S34, 2006.

[41] K. L. Nawrocki, E. K. Crispell, and S. M. McBride, "Antimicrobial peptide resistance mechanisms of gram-positive bacteria," *Antibiotics*, vol. 3, no. 4, pp. 461–492, 2014.

[42] J. L. Anaya-López, J. E. López-Meza, and A. Ochoa-Zarzosa, "Bacterial resistance to cationic antimicrobial peptides," *Critical Reviews in Microbiology*, vol. 39, no. 2, pp. 180–195, 2013.

[43] F. Meng, K. Kanai, and K. Yoshikoshi, "Characterization of drug resistance in *Streptococcus parauberis* isolated from Japanese flounder," *Fish Pathology*, vol. 44, no. 1, pp. 40–46, 2009.

[44] Y. K. Park, S. W. Nho, G. W. Shin et al., "Antibiotic susceptibility and resistance of *Streptococcus iniae* and *Streptococcus parauberis* isolated from olive flounder (*Paralichthys olivaceus*)," *Veterinary Microbiology*, vol. 136, no. 1-2, pp. 76–81, 2009.

[45] F. M. Mobegi, S. A. van Hijum, P. Burghout et al., "From microbial gene essentiality to novel antimicrobial drug targets," *BMC Genomics*, vol. 15, no. 1, article 958, 2014.

[46] I. Simon, M. Wright, T. Flohr, P. Hevezi, and I. W. Caras, "Determining subcellular localization of novel drug targets by transient transfection in COS cells," *Cytotechnology*, vol. 35, no. 3, pp. 189–196, 2001.

[47] E. Glory and R. F. Murphy, "Automated subcellular location determination and high-throughput microscopy," *Developmental Cell*, vol. 12, no. 1, pp. 7–16, 2007.

[48] F. Doro, S. Liberatori, M. J. Rodríguez-Ortega et al., "Surfome analysis as a fast track to vaccine discovery identification of a novel protective antigen for group B *streptococcus* hypervirulent strain COH1," *Molecular & Cellular Proteomics*, vol. 8, no. 7, pp. 1728–1737, 2009.

[49] M. C. Hung and W. Link, "Protein localization in disease and therapy," *Journal of Cell Science*, vol. 124, no. 20, pp. 3381–3392, 2011.

[50] N. Parveen and K. A. Cornell, "Methylthioadenosine/S-adenosylhomocysteine nucleosidase, a critical enzyme for bacterial metabolism," *Molecular Microbiology*, vol. 79, no. 1, pp. 7–20, 2011.

[51] K. F. KONG, L. Schneper, and K. Mathee, "Beta-lactam antibiotics: from antibiosis to resistance and bacteriology," *APMIS*, vol. 118, no. 1, pp. 1–36, 2010.

[52] R. Leonardi, S. Chohnan, Y. M. Zhang et al., "A pantothenate kinase from *Staphylococcus aureus* refractory to feedback regulation by coenzyme A," *Journal of Biological Chemistry*, vol. 280, no. 5, pp. 3314–3322, 2005.

[53] K. Kaneda, T. Kuzuyama, M. Takagi, Y. Hayakawa, and H. Seto, "An unusual isopentenyl diphosphate isomerase found in the mevalonate pathway gene cluster from *Streptomyces sp.* strain CL190," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 3, pp. 932–937, 2001.

[54] P. V. Rossitto, L. Ruiz, Y. Kikuchi et al., "Antibiotic susceptibility patterns for environmental streptococci isolated from bovine mastitis in central California dairies," *Journal of Dairy Science*, vol. 85, no. 1, pp. 132–138, 2002.

[55] T. C. S. Soares, A. C. Paes, J. Megid, P. E. M. Ribolla, K. D. S. Paduan, and M. Gottschalk, "Antimicrobial susceptibility of *Streptococcus suis* isolated from clinically healthy swine in Brazil," *Canadian Journal of Veterinary Research*, vol. 78, no. 2, pp. 145–149, 2014.

[56] J. Marie, H. Morvan, F. Berthelot-Herault et al., "Antimicrobial susceptibility of *Streptococcus suis* isolated from swine in France and from humans in different countries between 1996 and 2000," *Journal of Antimicrobial Chemotherapy*, vol. 50, no. 2, pp. 201–209, 2002.

[57] L. Dutil, R. Irwin, R. Finley et al., "Ceftiofur resistance inSalmonella entericaSerovar Heidelberg from chicken meat and humans, Canada," *Emerging Infectious Diseases*, vol. 16, no. 1, pp. 48–54, 2010.

[58] A. Franklin, "The *in vitro* bactericidal activity of danofloxacin and ceftiofur against respiratory pathogens in cattle," in *Proceedings of the 17th World Buiatrics Congress*, vol. 3, pp. 214–217, MI, USA, August 31 - September 4, 1992.

[59] N. Venisse, N. Grégoire, M. Marliat, and W. Couet, "Mechanism-based pharmacokinetic-pharmacodynamic models of *in vitro* fungistatic and fungicidal effects against *Candida albicans*," *Antimicrobial Agents and Chemotherapy*, vol. 52, no. 3, pp. 937–943, 2008.

[60] D. D. Tassew, A. F. Mechesso, N. H. Park, J. B. Song, J. W. Shur, and S. C. Park, "Biofilm formation and determination of minimum biofilm eradication concentration of antibiotics in *Mycoplasma hyopneumoniae*," *Journal of Veterinary Medical Science*, vol. 79, no. 10, pp. 1716–1720, 2017.

[61] W. Cai and C. R. Arias, "Biofilm formation on aquaculture substrates by selected bacterial fish pathogens," *Journal of Aquatic Animal Health*, vol. 29, no. 2, pp. 95–104, 2017.

[62] A. Yoshida and H. K. Kuramitsu, "Multiple *Streptococcus mutans* genes are involved in biofilm formation," *Applied and Environmental Microbiology*, vol. 68, no. 12, pp. 6283–6291, 2002.

[63] T. F. C. Mah and G. A. O'toole, “Mechanisms of biofilm resistance to antimicrobial agents,” *Trends in Microbiology*, vol. 9, no. 1, pp. 34–39, 2001.

[64] S. Mulla, A. Kumar, and S. Rajdev, “Comparison of MIC with MBEC assay for *in Vitro* antimicrobial susceptibility testing in biofilm forming clinical bacterial isolates,” *Advances in Microbiology*, vol. 6, no. 2, pp. 73–78, 2016.

Hindawi
International Journal of Genomics
Volume 2020, Article ID 4109186, 11 pages
https://doi.org/10.1155/2020/4109186

Research Article

Genomic Analysis of *Bacillus megaterium* NCT-2 Reveals Its Genetic Basis for the Bioremediation of Secondary Salinization Soil

Bin Wang,[1] **Dan Zhang**,[1] **Shaohua Chu**,[1] **Yuee Zhi**,[1] **Xiaorui Liu**,[2] **and Pei Zhou**[1]

[1]*School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China*
[2]*The International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China*

Correspondence should be addressed to Xiaorui Liu; xiaorui1211@126.com and Pei Zhou; peizhousjtu@163.com

Received 11 November 2019; Revised 1 February 2020; Accepted 8 February 2020; Published 29 February 2020

Guest Editor: Ravi Kant

Bacillus megaterium NCT-2 is a nitrate-uptake bacterial, which shows high bioremediation capacity in secondary salinization soil, including nitrate-reducing capacity, phosphate solubilization, and salinity adaptation. To gain insights into the bioremediation capacity at the genetic level, the complete genome sequence was obtained by using a multiplatform strategy involving HiSeq and PacBio sequencing. The NCT-2 genome consists of a circular chromosome of 5.19 Mbp and ten indigenous plasmids, totaling 5.88 Mbp with an average GC content of 37.87%. The chromosome encodes 5,606 genes, 142 tRNAs, and 53 rRNAs. Genes involved in the features of the bioremediation in secondary salinization soil and plant growth promotion were identified in the genome, such as nitrogen metabolism, phosphate uptake, the synthesis of organic acids and phosphatase for phosphate-solubilizing ability, and Trp-dependent IAA synthetic system. Furthermore, strain NCT-2 has great ability of adaption to environments due to the genes involved in cation transporters, osmotic stress, and oxidative stress. This study sheds light on understanding the molecular basis of using *B. megaterium* NCT-2 in bioremediation of the secondary salinization soils.

1. Introduction

Soil application of organic and inorganic fertilizers for crop and vegetable cultivation is the major source for soil nitrate-nitrogen (nitrate-N), which increases agricultural productivity. However, the vegetable yields do not increase continuously with soil nitrate-N [1]. A large accumulation of nitrate in soil results in soil secondary salinization, having various adverse effects on soil productivity, and nitrate accumulation in vegetables [2]. What is more, the reduction of nitrate to nitrite can cause various human diseases [1]. Soil secondary salinization is a severe problem in intensively managed agricultural ecosystems [3]. It is required to develop a low-cost bioremediation method to remove nitrate from soil.

In our previous study, *Bacillus megaterium* NCT-2 was isolated from the secondary nitrate-salinized soil in a greenhouse, which shows high nitrate-reducing capacity and salinity adaptation in secondary salinization soil [4]. It can remove nitrate at initial nitrate-N concentrations ranging from 100 mg/L to 1,000 mg/L and grow well in inorganic salt medium with 4.0% sodium chloride [4]. In our field trails, the concentrations of NO_3^- in both soil and plant were reduced significantly when we used the NCT-2 strain mixed with straw powder to treat secondary salinization soil (unpublished). Moreover, this strain showed significant phosphate-solubilizing ability of insoluble inorganic phosphates in the culture medium [5]. Strain NCT-2 has the potential to be utilized as a biofertilizer for bioremediation of the secondary nitrate-salinized soil and plant growth promotion [6].

The Gram-positive bacterium *Bacillus megaterium* is found in diverse habitats from soil to sediment, sea, and dried food. It was named after its big size with a volume approximately 100 times than that of *Escherichia coli* [7]. Its big size made it ideal to be used in studies of cell structure, protein localization, sporulation, and membranes [8, 9]. Due to no production of endotoxins associated with the outer membrane and no external alkaline proteases, they are used widely as desirable cloning hosts in food and pharmaceutical production processes for α- and β-amylases in the baking industry [10, 11], penicillin acylase [12–14], and vitamin B12 [15], such as *Bacillus megaterium* DSM 319, *Bacillus megaterium* QM B1551, and *Bacillus megaterium* WSH 002 [16, 17]. The genomes of them have been sequenced to gain insights into the metabolic versatility that facilitate biotechnological applications, not the bioremediation of secondary salinization soil [18, 19].

Despite the previously published work sequenced the 5.68 Mb draft genome of *B. megaterium* NCT-2 by using the Solexa platform, consisting of the 204 contigs, it focused only on the multiple alignments of nitrate assimilation-related gene sequences [20]. The functional nitrate assimilation-related genes (the nitrate reductase electron transfer subunit, the nitrate reductase catalytic subunit, the nitrite reductase [NAD(P)H] large subunit and small subunit, and the glutamine synthetase) were identified [20]. The genes that could be involved in the full potential of strain NCT-2 in the bioremediation of secondary salinization soil remain unknown. For this, we obtained its complete genome sequence by using a multiplatform strategy involving HiSeq and PacBio sequencing. Furthermore, we performed a comprehensive analysis of nitrogen metabolism and plant growth-promoting features. The comparative analysis might be helpful for use in soil bioremediation.

2. Methods

2.1. DNA Preparation and Genome Sequencing. B. megaterium NCT-2, isolated from the secondary salinized greenhouse soil in China, was cultured in a defined inorganic salt medium as previously described [4]. It was registered in China General Microbiological Culture Collection Center under CGMCC No. 4698. Genomic DNA was isolated using QIAGEN DNeasy Blood & Tissue Kit (Hilden, Germany). The concentration and quality of DNA were determined by a Qubit Fluorometer (Thermo Scientific, USA), NanoDrop Spectrophotometer (Thermo Scientific, USA), and agarose electrophoresis. The whole genome of the *B. megaterium* strain NCT-2 was sequenced by the BGI Tech Solutions Co., Ltd. (Shenzhen, China) by using Illumina Hiseq 4000 short-read sequencing platform (Illumina Inc., San Diego, CA, USA) (insert size, 500 bp; 2 × 125 bp read length) and PacBio RSII long-read sequencing platform (Pacific Biosciences of California, Inc., Menlo Park, CA, USA) (Figure S1).

2.2. Genome Assembly and Annotation. After quality control, the *de novo* assembly of the whole NCT-2 genome was performed using the RS_HGAP Assembly3 in the SMRT Analysis pipeline version 2.2.0 [21]. The HiSeq clean reads were preliminarily assembled into contigs and then were used for hybrid error correction of the subreads from PacBio. There were two rounds of error correction. One was analyzed by using SOAPsnp and SOAPIndel [22] and another was by using the Genome Analysis Toolkit (GATK) [23]. Finally, SSPACE-LongRead [24] and Celera assemble [25] were used to generate a high-quality genome. The finished NCT-2 genome was submitted to GenBank, replacing the previous version of the draft genome [20].

The protein-coding genes were predicted by using Glimmer 3.02 [26], and the tandem repeats were detected with Tandem Repeat Finder 4.04 [27]. The gene function annotation was accomplished by blasting the protein sequences against the database of Kyoto Encyclopedia of Genes and Genomes (KEGG) [28]. In addition, the RAST web server (https://rast.nmpdr.org) with the default parameters was used to catalog all the predicted genes into subsystems according to functional categories [29, 30]. CGView was used to produce the maps of the circular genomes with gene feature information [31]. Genome alignments with locally collinear blocks were performed with MAUVE [32].

2.3. Phylogenetic Analysis. The whole genome-based phylogenetic analysis was performed by using the CVTree 3.0 online server [33, 34]. Fourteen genome sequences were obtained from GenBank. A phylogenetic tree was constructed by the neighbor-joining method using MEGA analysis [35–37]. In addition, FusionDB was used to analyze the functional repertories of *B. megaterium* NCT-2 and identify the nearest "neighbors" based on the functional similarities [38, 39].

3. Results and Discussion

3.1. General Genomic Characteristics. A total of ~1,189 Mb raw data and ~1,147 Mb clean data were obtained after filtering the low-quality reads generated by the HiSeq platform. The PacBio platform yielded 48,392 polymerase reads (with the average size of 12.9 kb) and 622 Mb subreads after quality control. The complete genome was assembled by taking advantage of the higher accuracy short reads from the HiSeq platform and the long subreads from the PacBio platform. The genome consists of a circular chromosome of 5.19 Mb with an average GC content of 38.2% (accession number: CP032527.2) and ten circular plasmids designated as the plasmid pNCT2-1 to pNCT2-10 (accession numbers: CP032528.1-CP032537.1). Sequence information was visualized in CG view Server (Figure 1 and S2). The total genome size is 5.88 Mb with an average GC content of 37.87%. The whole genome contains 6,039 genes, including 5,606 coding sequences, 203 RNA genes, and 230 pseudo genes. There are 127 identified tandem repeat sequences (TRF), 83 minisatellite DNA, and 7 microsatellite DNA.

The general features of *B. megaterium* NCT-2 were compared with five genomes of *Bacillus* strains (*Bacillus megaterium* DSM 319, *Bacillus megaterium* QM B1551, *Bacillus subtilis* subsp. subtilis str. 168, *Bacillus cereus* Q1, and *Bacillus licheniformis* DSM 13) (Table 1). The genome GC contents for three *B. megaterium* strains are around 38%. Strain NCT-2 has the largest genome size and most coding sequences and RNA genes, such as 53 rRNAs and 142 tRNAs. There were 14 rRNA operons on the negative chain and one rRNA operon on the positive strand with a 16S-23S-5S organization. In addition, the positive chain had one unusual rRNA operon with a 16S-23S-5S-5S organization and a single 5S rRNA. The microbial genome size is positively correlated

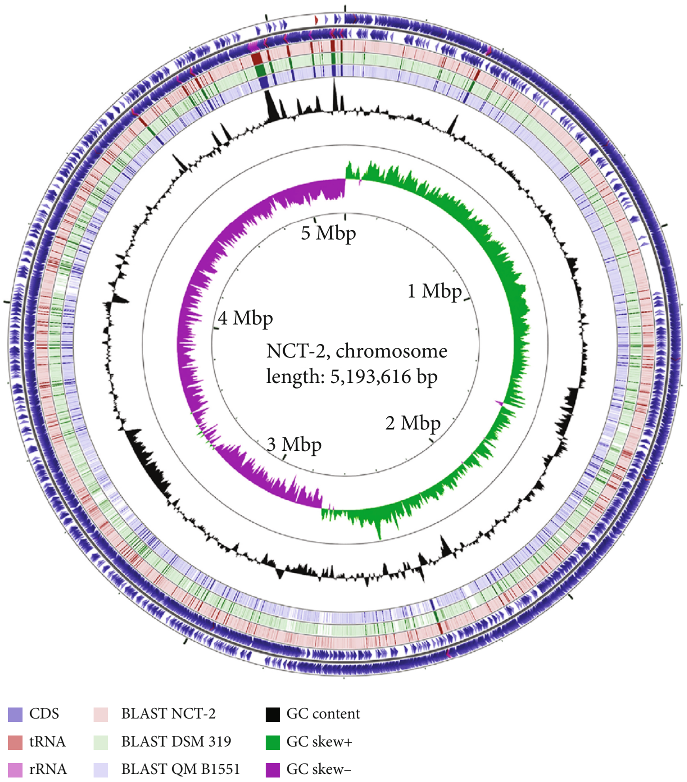


Figure 1: Genetic and physical map of the genome of *B. megaterium* NCT-2 prepared using CGView. Circles from the outside to the inside show the position of protein-coding sequences (blue), tRNA gene (red), and rRNA genes (pink) on the positive (circle 1) and negative (circle 2) strands. Circles 3-5 show the positions of BLAST hits detected through BLASTx comparisons of *B. megaterium* NCT-2 against itself (circle 3), *B. megaterium* DSM 319 (circle 4), and *B. megaterium* QM B1551 (circle 5). Circles 6 and 7 show plots of GC content and GC skew plotted as the deviation from the average for the entire sequence.

Table 1: General genome features of *B. megaterium* NCT-2 compared with other five *Bacillus* strains.

Strain	*B. megaterium* NCT-2	*B. megaterium* QM B1551	*B. megaterium* DSM 319	*B. subtilis* 168	*B. cereus* Q1	*B. licheniformis* DSM 13
Genome size (Mb)	5.88	5.52	5.10	4.22	5.51	4.22
Chromosome size (Mb)	5.19	5.10	5.10	4.22	5.21	4.22
G+C content (%)	37.8	37.97	38.1	43.5	35.5	46.2
Chromosomal G+C content (%)	38.2	38.3	38.1	43.5	35.6	46.2
Gene number	6039	5674	5245	4536	5856	4382
Coding sequence number	5606	5379	4941	4237	5513	4219
RNA gene number	203	182	153	116	137	98
rRNA genes (5S, 16S, and 23S)	53 (19, 17, 17)	37 (13, 12, 12)	33 (11, 11, 11)	30 (10, 10, 10)	39 (13, 13, 13)	21 (7, 7, 7)
tRNA gene number	142	137	114	86	93	72
Plasmid number	10	7	0	0	2	0

with their environment adaptability [40]. One typical characteristic of soil microorganisms is the high number of rRNAs, which is helpful for fast growth, successful sporulation, germination, and rapid response to changing the availability of nutrients [41–44]. These features indicate that strain NCT-2 has great ability of adaptation to various environments.

Most strains of *Bacillus megaterium* carry multiple plasmids, such as strain QM B1551 has seven resident plasmids [18], *Bacillus megaterium* strain 216 has ten plasmids [45], and *Bacillus megaterium* NBRC 15308 has six plasmids. As for the ten plasmids in strain NCT-2, the sizes range from 9,625 bp to over 132 kb making up 11.7% of the whole genome (Table S1). The plasmids have significantly lower GC contents than the chromosome (33.7-37.0% versus 38.2%). There are 761 coding sequences and 23 RNA genes. Both plasmids pNCT2-2 and pNCT2-6 had one tRNA. In addition, pNCT2-7 had 18 tRNAs, one 5S RNA, one large subunit ribosomal RNA (LSU rRNA), and one small subunit ribosomal RNA (SSU rRNA). Additional rRNA operons carried on plasmids slowed the growth rates of *E. coli* on poor carbon sources [46]. Further investigations are needed to clarify the role of plasmids in bacterial growth and adaptations to high-nitrate environments in bioremediation of the secondary salinization soils.

3.2. Phylogenetic Lineage Analysis. We used CVTree 3.0 to construct a phylogenetic tree based on the complete proteomes with *Macrococcus caseolyticus* JCSC5402 as an outgroup. The obtained tree (Figure 2(a)) indicated that *B. megaterium* NCT-2 was most homologous to *B. megaterium* DSM 319 and then *B. megaterium* QM B1551. Similarly, genome comparison using the RAST Prokaryotic Genome Annotation Server also showed that the genomic sequence of NCT-2 had a higher comparison score with *B. megaterium* QM B1551 and *B. megaterium* DSM 319 (Figure S3). Furthermore, 16S rDNA sequences from 15 *Bacillus* strains were used to construct a phylogenetic tree by MEGA7 with the neighbor-joining method. The neighbor-joining phylogenetic tree shows that strain NCT-2 is closest to *B. megaterium* QM B1551, *B. megaterium* DSM 319, and *B. megaterium* WSH 002 (Figure 2(b)). Whole-genome alignment of *B. megaterium* NCT-2 to closely related QM B1551 and DSM 319 by using MAUVE revealed that the chromosomes of the three strains showed overall collinearity (Figure 2(c)).

3.3. Functional Annotations of B. megaterium NCT-2. To investigate the function of the 5,606 coding sequences, the GO database, the KEGG database, the COG database, and RAST web server were used. The 3,159 genes annotated by GO were classified into biological processes, cellular components, and molecular functions (Figure S4). The top five categories were catalytic activity (1,822), metabolic process (1,786), cellular process (1,567), single-organism process (1,400), and binding (1,214).

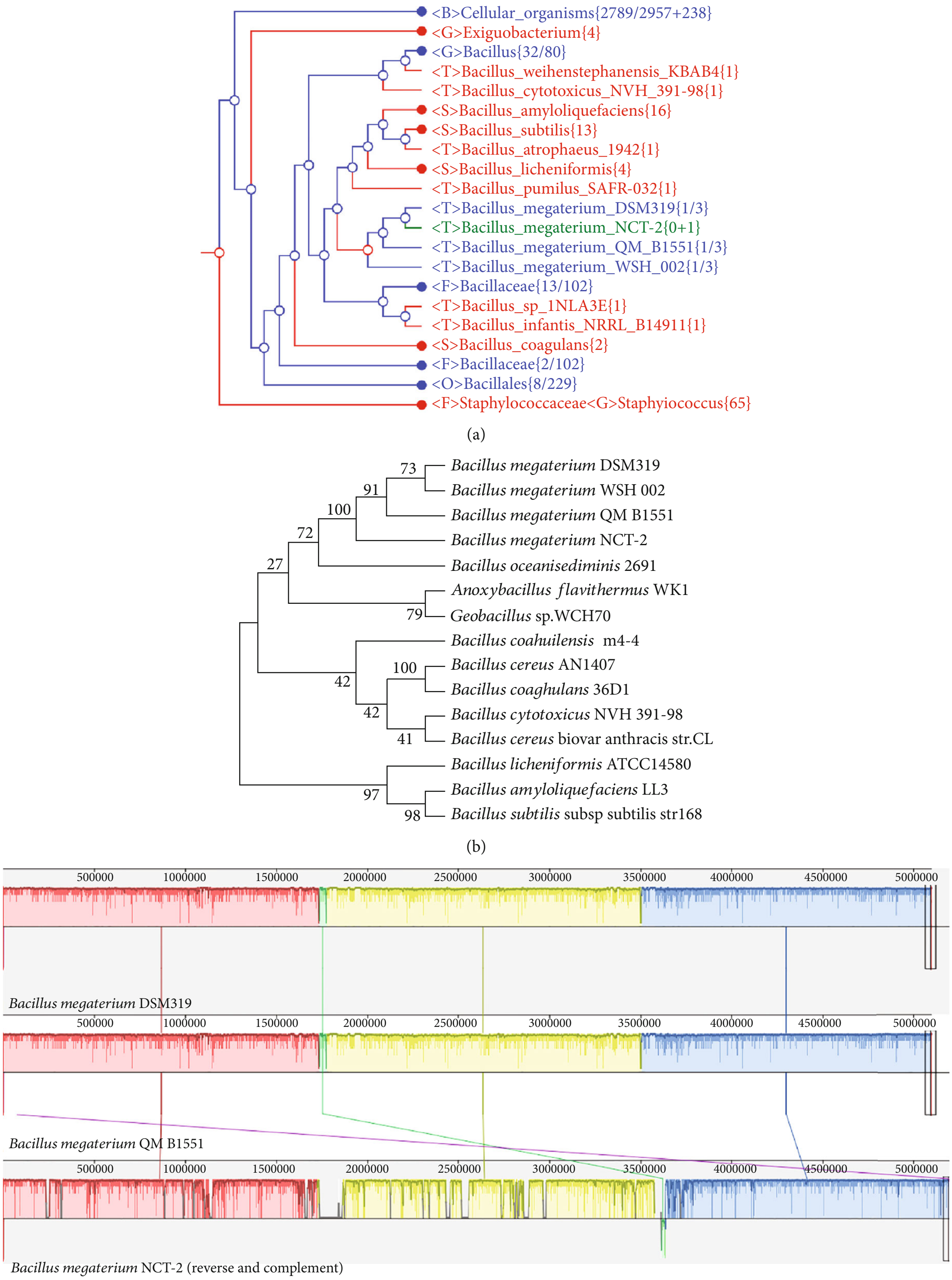
2,338 chromosomal genes (44%) were assigned into 477 subsystems by RAST (Figure S5a). Subsystem category comparisons among six related *Bacillus* strains showed that the number of genes involved in "Amino Acids and Derivatives" and "Carbohydrates" was highest in the genome of the six strains (Figure 3(a)). In addition, *Bacillus megaterium* has more genes involved in "Cofactors, Vitamins, Prosthetic Groups, Pigments." The top five categories in strain NCT-2 were the "Amino Acids and Derivatives" (538), "Carbohydrates" (500), "Cofactors, Vitamins, Prosthetic Groups, Pigments" (340), "Protein Metabolism" (283), and "Fatty Acids, Lipids, and Isoprenoids" (180).

Likewise, 2,962 genes annotated by the KEGG database were assigned to 38 pathways (Figure 3(b)). The top five enriched pathways were "Biosynthesis of other secondary metabolism" (710), "Signaling molecules and interaction in Environmental information processing" (542), "Substance dependence" (540), "Nucleotide metabolism" (475), and "Immune disease" (472).

Like most strains of *B. megaterium*, which carry more than four plasmids, strain NCT-2 harbors ten indigenous plasmids. Only 75 genes (10%) were assigned into 37 subsystems by RAST (Figure S5b), including genes for riboflavin metabolism, butanol biosynthesis, and xylose utilization, and parts of genes in benzoate degradation and metabolism of central aromatic intermediates. There are also genes for cobalt-zinc-cadmium resistance, oxidative stress, and nitrosative stress.

3.4. Microbial Functional Similarities. The translated protein sequence of *B. megaterium* NCT-2 was downloaded from RAST and submitted to the FusionDB web server (https://services.bromberglab.org/fusiondb/mapping) [38]. The submitted proteome (containing 5,364 proteins) matched to 3,662 FusionDB functions, while 228 proteins could not be mapped to any function in their database. The functional similarities of *B. megaterium* NCT-2 with 1,374 taxonomically distinct bacteria (with similarity > 40%) were shown in Table S2, most of them were soil bacterium. Strain NCT-2 is most functionally similar to *B. megaterium* DSM 319 (90%) and *B. megaterium* QM B1551 (89%). The functional relationships among nine *Bacillus* strains were demonstrated by the fusion+ networks (Figure 4(a)). There were 1,290 functions shared by all of them. The common functional annotations related to nitrogen metabolism were nitrite transporter NirC, nitrogen-fixing NifU domain protein, nitroreductase, nitrate transport protein, and 2-nitropropane dioxygenase. Notably, there are 3,047 functions shared among three strains of *B. megaterium* (strain NCT-2, strain QM B1551, and strain DSM 319) (Figure 4(b)). Strain NCT-2 has most of the core genes and pathways, including vitamin biosynthesis and nitrogen metabolism. The nitrogen metabolism-related genes, such as those encoding nitrate transport protein, nitrate/nitrite sensor protein, nitric oxide reductase activation protein, nitrite reductase [NAD(P)H] large subunit, nitrite reductase [NAD(P)H] small subunit, nitrite transporter, nitrite-sensitive transcriptional repressor, nitrogen regulatory protein P-II, nitrogen-fixing NifU domain protein, nitroreductase, and nitroreductase family protein, were located on the chromosome of the three strains. Furthermore, only strain NCT-2 carries the gene encoding for periplasmic nitrate reductase.

3.5. Genome Inventory for Nitrogen Metabolism. In our field experiment, strain NCT-2 shows high nitrate-reducing capacity in secondary salinization soil (unpublished). The functional nitrate assimilation-related genes that are involved in the process of converting nitrate to glutamine have been identified [20]. The genes encoding nitrate and nitrite reductase were cloned and overexpressed in *Escherichia coli* [47]. Here, the whole genomic analysis also revealed the genes encoding sensor, transporter, and enzymes are involved in nitrogen metabolism. The genes were scattered in the chromosome. Genes encoding nitrite-sensitive transcriptional repressor (NsrR), which is directly sensitive to nitrosative stress, were found in both the chromosome and the plasmid (Table S3 and Figure S6). *B. megaterium* NCT-2 possessed


<B>Cellular_organisms{2789/2957+238}
<G>Exiguobacterium{4}
<G>Bacillus{32/80}
<T>Bacillus_weihenstephanensis_KBAB4{1}
<T>Bacillus_cytotoxicus_NVH_391-98{1}
<S>Bacillus_amyloliquefaciens{16}
<S>Bacillus_subtilis{13}
<T>Bacillus_atrophaeus_1942{1}
<S>Bacillus_licheniformis{4}
<T>Bacillus_pumilus_SAFR-032{1}
<T>Bacillus_megaterium_DSM319{1/3}
<T>Bacillus_megaterium_NCT-2{0+1}
<T>Bacillus_megaterium_QM_B1551{1/3}
<T>Bacillus_megaterium_WSH_002{1/3}
<F>Bacillaceae{13/102}
<T>Bacillus_sp_1NLA3E{1}
<T>Bacillus_infantis_NRRL_B14911{1}
<S>Bacillus_coagulans{2}
<F>Bacillaceae{2/102}
<O>Bacillales{8/229}
<F>Staphylococcaceae<G>Staphyiococcus{65}

(a)

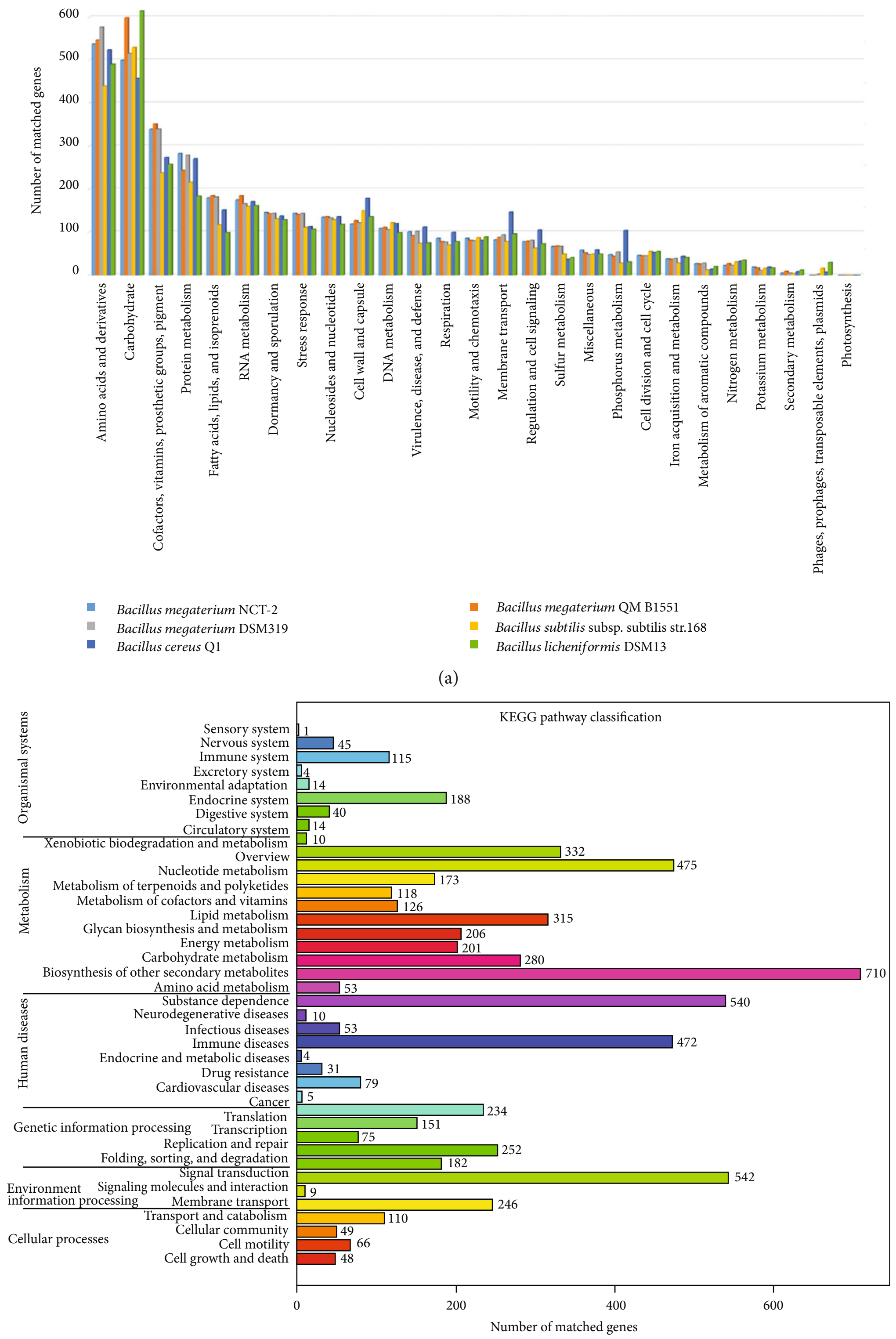
73
91
100
72
27
79
42
100
42
41
97
98
Bacillus megaterium DSM319
Bacillus megaterium WSH 002
Bacillus megaterium QM B1551
Bacillus megaterium NCT-2
Bacillus oceanisediminis 2691
Anoxybacillus flavithermus WK1
Geobacillus sp.WCH70
Bacillus coahuilensis m4-4
Bacillus cereus AN1407
Bacillus coaghulans 36D1
Bacillus cytotoxicus NVH 391-98
Bacillus cereus biovar anthracis str.CL
Bacillus licheniformis ATCC14580
Bacillus amyloliquefaciens LL3
Bacillus subtilis subsp subtilis str168

(b)

500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000
5000000
Bacillus megaterium DSM319
Bacillus megaterium QM B1551
Bacillus megaterium NCT-2 (reverse and complement)

(c)

Figure 2: Phylogenetic tree showing the position of *Bacillus megaterium* NCT-2. (a) The tree was constructed based on the frequency of >6-string predicted peptides from the whole genome sequences by using CVTree 3.0. (b) The tree is based on 16S rDNA phylogenetic analysis by using MEGA7 with the neighbor-joining method. The bootstrap consensus tree inferred from 1,000 replicates is taken to represent the evolutionary history of the taxa analyzed. (c) Chromosomal similarity among strain NCT-2, DSM 319, and QM B1551 by using Mauve alignments. Three local collinear blocks (LCBs) on the chromosomes were identified and joined by connecting lines in the three genomes.



(a)

(b)

Figure 3: Distribution of genes based on Cluster of Orthologous Groups (COG) classification and KEGG pathways in *B. megaterium* NCT-2. Only genes assigned to the COG and KEGG categories were used for analysis. (a) Functional categorization of *B. megaterium* NCT-2 compared with 5 related *Bacillus* strains based on the COG database. (b) KEGG classification of *B. megaterium* NCT-2.

nitrate/nitrite sensor protein (NaNiS) and nitrate/nitrite transporter (NaNiT) for sensing and transporting the NO_3^- and NO_2^-. In the process of nitrate and nitrite ammonification, assimilatory nitrate reductase (NaRas) and nitrite reductase (NiRas) catalyzed the reduction of nitrate to ammonia through nitrite [48]. Then, ammonia was assimilated into

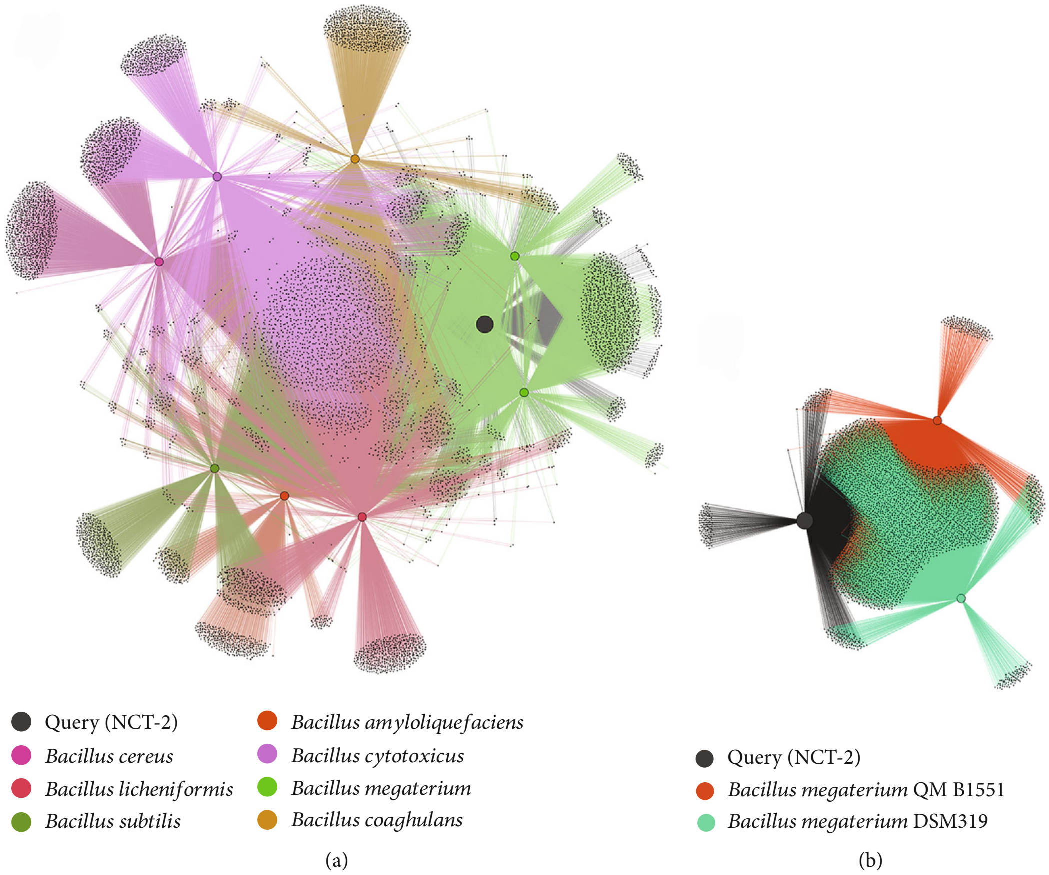


Figure 4: Functional similarity networks analysis by fusion+. (a) Functional similarity network among 9 strains of *Bacillus*. (b) Functional similarity network among 3 strains of *Bacillus megaterium*. The networks contain nodes of organisms and functions. The colored organism nodes are connected to black function nodes.

amino acids through L-Glutamine and L-Glutamate by glutamine synthetase type I (GSI), Ferredoxin-dependent glutamate synthase (GOGATF), glutamate synthase [NADPH] large chain (GOGDP1), and glutamate synthase [NADPH] small chain (GOGDP2). Ammonium transporter (Amt) was also encoded in the genome. Ammonium is an important nitrogen source for plant growth. Environmental NH_4^+/NH_3 was imported across membranes by Amt for cell growth in prokaryotes and plants [49]. Bacterial Amt proteins act as passive channels for the uncharged gas ammonia (NH_3) [50]. It means that *B. megaterium* NCT-2 might scavenge NH_4^+/NH_3 in soil instead of providing. In the face of nitrosative stress, genes encoding nitrite-sensitive transcriptional repressor (NsrR) were found in both the chromosome and the plasmid. NsrR played a pivotal role in the regulation of NirK (nitrite reductase), which was expressed aerobically in response to the increasing concentration of NO_2^- and decreasing pH [51]. However, no functional NirK could be found. Instead, two nitric oxide reductase activation proteins (NorD and NorQ) for denitrifying reductase gene clusters were found but without nitric oxide reductase, making the function of denitrification highly unlikely. Thus, the genome analysis proposed that *B. megaterium* NCT-2 could convert nitrate from secondary salinization soil into biomass through glutamate rather than reduce nitrate to nitrous oxide or dinitrogen, which are lost from the soil (Figure 5). It is an effective bioremediation approach to remove nitrate from soils.

3.6. Genes Associated with Plant Growth-Promoting Features. Our previous studies on the plant growth promotion of *B. megaterium* NCT-2 revealed that it could produce organic acids (lactic acid, acetic acid, propionic acid, and gluconic acid) and phosphatase in culture medium, showing significant phosphate-solubilizing ability [5]. Inoculation with *B. megaterium* NCT-2 significantly increased the root fresh weight of maize [6]. The genome of NCT-2 contains genes encoding for glucose 1-dehydrogenase (EC 1.1.1.47) and alkaline phosphatase (EC 3.1.3.1). Glucose dehydrogenase can oxidize glucose to gluconic acid, which is the most frequent organic acid produced by phosphate-solubilizing bacteria [52]. Additionally, the phosphate starvation system for phosphate uptake encoded by *pst*S, *pst*C, *pst*A, and *pst*B was also found in the genome. The phosphate solubilization capacity of strain NCT-2 plays a positive role in promoting plant growth by dissolving unavailable P (PO_4^{3-}) in soil to plant available forms.

Many plant growth-promoting bacteria have the ability to synthesize plant auxins (indole-3-acetic acid, IAA)

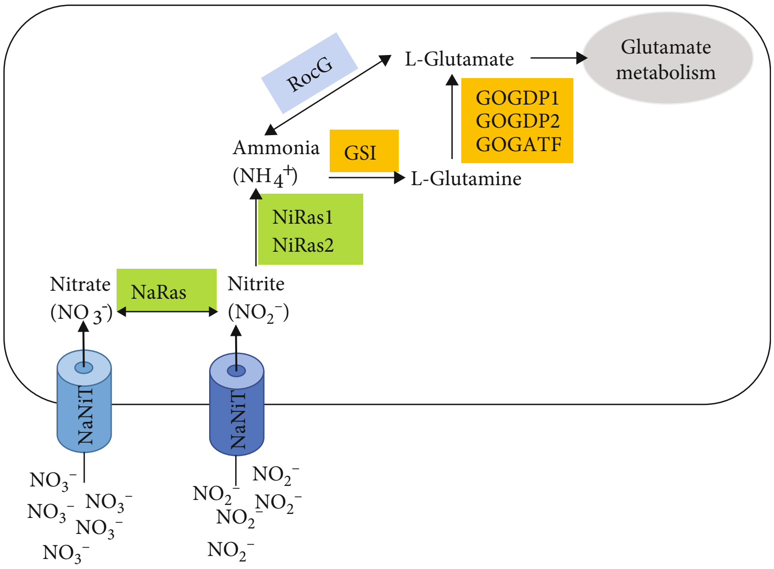


Figure 5: The proposed nitrogen metabolism in *B. megaterium* NCT-2. NaNiT: nitrate/nitrite transporter; NaRas: assimilatory nitrate reductase large subunit (EC 1.7.99.4); NiRas1: nitrite reductase [NAD(P)H] large subunit (EC 1.7.1.4); NiRas2: nitrite reductase [NAD(P)H] small subunit (EC 1.7.1.4); GSI: glutamine synthetase type I (EC 6.3.1.2); GOGDP1: glutamate synthase [NADPH] large chain (EC 1.4.1.13); GOGDP2: glutamate synthase [NADPH] small chain (EC 1.4.1.13); GOGATF: Ferredoxin-dependent glutamate synthase (EC 1.4.7.1); RocG: NAD-specific glutamate dehydrogenase (EC 1.4.1.2).

[53, 54], which is a key regulator for plant growth and development, such as cell division and elongation, lateral root production, and flowering [55]. Large-scale genomic analysis of IAA synthesis pathways suggested that plenty of bacteria could synthesize IAA via multiple incomplete pathways, and Firmicutes genomes had the simplest Trp-dependent IAA synthetic system [56]. According to the KEGG analysis, strain NCT-2 could assimilate tryptophan (Trp) (Figure S7) but had incomplete Trp-dependent IAA synthesis pathways, such as the indole-3-acetamide (IAM) pathway and indole-3-pyruvate (IPA) pathway (Figure S8). It had aldehyde dehydrogenase (NAD^+) (EC 1.2.1.3) and amidase (EC 3.5.1.4) catalyzing the final step of IAA synthesis. However, we could not find the enzymes which convert Trp into IAM and IPA. These results suggested that strain NCT-2 might synthesize IAA from intermediates.

Both the phosphate solubilization and IAA synthesis play important roles in plant growth promotion of strain NCT-2 during biocontrol and bioremediation of the secondary salinization soils.

3.7. Genes Involved in Stress Response. B. megaterium NCT-2 showed high salinity adaptation in secondary salinization soil in our previous study [4]. From the genome perspective, we can see genes involved in cation transporters (magnesium transport and copper transport system) and stress response, such as osmotic stress, oxidative stress, and detoxification. Glycine betaine, a very efficient osmoprotectant, can be synthesized or acquired from exogenous sources [57]. There are glycine betaine ABC transport systems (*opu*A, *opu*C, and *opu*D) for choline uptake and genes for the glycine betaine biosynthetic enzymes (choline dehydrogenase, *gbs*B, and betaine-aldehyde dehydrogenase, *gbs*A) in strain NCT-2 genome. Moreover, the genome contains genes encoding for superoxide dismutase (EC 1.15.1.1), catalase (EC 1.11.1.6), and ferroxidase (EC 1.16.3.1), protecting bacteria from oxidative stress. It implied that NCT-2 has great ability of adaption to environments.

4. Conclusion

A hybrid approach with multiple assembler was used to assemble the complete genome of *B. megaterium* NCT-2. The deeper investigation identified clues associated with the features of the bioremediation of secondary salinization soil and plant growth promotion at the gene level, such as nitrogen metabolism, phosphate uptake, synthesis of organic acids and phosphatase for phosphate-solubilizing ability, and Trp-dependent IAA synthetic system. Furthermore, the genes involved in cation transporters, osmotic stress, and oxidative stress implied that NCT-2 has great ability of adaption to environments. In summary, these results provide valuable genomic resources for further studies and applications of using *B. megaterium* NCT-2 in bioremediation processes of secondary salinization soil.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. The genome sequence of *B. megaterium* NCT-2 has been deposited in GenBank. The accession number for the *B. megaterium* NCT-2 chromosome is CP032527.2, and those for ten plasmids are CP032528.1, CP032529.1, CP032530.1, CP032531.1, CP032532.1, CP032533.1, CP032534.1, CP032535.1, CP032536.1, CP032537.1.

Disclosure

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Xiaorui Liu and Pei Zhou conceptualized the study. Bin Wang and Xiaorui Liu performed formal analysis. Dan Zhang and Shaohua Chu took care of funding acquisition. Bin Wang, Dan Zhang, and Shaohua Chu performed the investigation. Yuee Zhi and Xiaorui Liu performed methodology. Pei Zhou acquired resources. Yuee Zhi and Pei Zhou performed supervision of the study. Bin Wang wrote the original draft. Dan Zhang and Xiaorui Liu reviewed and edited the manuscript.

Acknowledgments

This research was funded by grants from National Key Research and Development Program (no. 2016YFD0800803), National Natural Science Foundation of China (nos. 31702003 and 31902105), Young Elite Scientists Sponsorship Program by CAST (YESS Program) (no. 2017QNRC001), Shanghai Education Development Foundation and Shanghai Municipal Education Commission "Chenguang Program" (no. 17CG07), and China Postdoctoral Science Foundation (no. 2019M651505). We are grateful to Prof. Hongyu Ou and Prof. Fei Tao (Shanghai Jiaotong University) for many helpful comments and suggestions.

Supplementary Materials

Supplementary 1. Figure S1: whole genome sequencing and assembly workflow. Supplementary 2. Figure S2: circular representation of the ten plasmids of *B. megaterium* NCT-2. Supplementary 3. Figure S3: genome similarity of strain NCT-2. The genome of strain NCT-2 was submitted to the web service RAST and was compared with genomes of other strains. A higher comparison score means higher similarity. Supplementary 4. Figure S4: histogram of GO classifications. The results are summarized in three categories: biological process (blue), cellular component (brown), and molecular function (orange). Supplementary 5. Figure S5: genes connected to subsystems according to functional categories. (a) The subsystems of genes from chromosome. (b) The subsystems of genes from plasmids. Supplementary 6. Figure S6: enzymes involved in the nitrogen metabolism of *B. megaterium* NCT-2 from KEGG. Genes of *B. megaterium* NCT-2 were shown in green boxes. Supplementary 7. Figure S7: enzymes involved in phenylalanine, tyrosine, and tryptophan biosynthesis of *B. megaterium* NCT-2 from KEGG. Genes of *B. megaterium* NCT-2 were shown in green boxes. Supplementary 8. Figure S8: enzymes involved in the tryptophan metabolism of *B. megaterium* NCT-2 from KEGG. Genes of *B. megaterium* NCT-2 were shown in green boxes. Supplementary 9. Table S1. plasmids features of *B. megaterium* NCT-2. Supplementary 10. Table S2. the functional similarities of *B. megaterium* NCT-2 with 1,374 bacterial genomes. Supplementary 11. Table S3. gene cluster involved in nitrogen metabolism of *B. megaterium* NCT-2. *(Supplementary Materials)*

References

[1] Z. H. Wang and S. X. Li, "Effects of nitrogen and phosphorus fertilization on plant growth and nitrate accumulation in vegetables," *Journal of Plant Nutrition*, vol. 27, no. 3, pp. 539–556, 2004.

[2] W. S. Shen, X. G. Lin, W. M. Shi et al., "Higher rates of nitrogen fertilization decrease soil enzyme activities, microbial functional diversity and nitrification capacity in a Chinese polytunnel greenhouse vegetable land," *Plant and Soil*, vol. 337, no. 1-2, pp. 137–150, 2010.

[3] L. Lassaletta, G. Billen, B. Grizzetti, J. Anglade, and J. Garnier, "50 year trends in nitrogen use efficiency of world cropping systems: the relationship between yield and nitrogen input to cropland," *Environmental Research Letters*, vol. 9, no. 10, article 105011, 2014.

[4] S. WeiWei, H. HongYan, W. Nan, Z. YueE, L. QunLu, and Z. Pei, "Characterization of a nitrate-uptake bacterial strain *Bacillus megaterium* NCT-2," *Fresenius Environmental Bulletin*, vol. 22, no. 2, pp. 412–417, 2013.

[5] X. Wei, X. U. Lurong, D. Zhang, Y. Zhi, and P. Zhou, "Phosphate solubilizing characteristics of a nitrate-tolerating bacterium *Bacillus megaterium*," *Acta Scientiae Circumstantiae*, vol. 35, no. 7, pp. 2052–2058, 2015.

[6] S. Chu, D. Zhang, Y. Zhi et al., "Enhanced removal of nitrate in the maize rhizosphere by plant growth-promoting *Bacillus megaterium* NCT-2, and its colonization pattern in response to nitrate," *Chemosphere*, vol. 208, pp. 316–324, 2018.

[7] P. S. Vary, R. Biedendieck, T. Fuerch et al., "*Bacillus megaterium*—from simple soil bacterium to industrial protein production host," *Applied Microbiology and Biotechnology*, vol. 76, no. 5, pp. 957–967, 2007.

[8] S. Hrafnsdottir, J. W. Nichols, and A. K. Menon, "Transbilayer movement of fluorescent phospholipids in *Bacillus megaterium* membrane vesicles," *Biochemistry*, vol. 36, no. 16, pp. 4969–4978, 1997.

[9] G. J. McCool and M. C. Cannon, "PhaC and PhaR are required for polyhydroxyalkanoic acid synthase activity in *Bacillus megaterium*," *Journal of Bacteriology*, vol. 183, no. 14, pp. 4235–4243, 2001.

[10] R. Kittsteiner-Eberle, I. Ogbomo, and H. L. Schmidt, "Biosensing devices for the semi-automated control of dehydrogenase substrates in fermentations," *Biosensors*, vol. 4, no. 2, pp. 75–85, 1989.

[11] T. Nagao, T. Mitamura, X. H. Wang et al., "Cloning, nucleotide sequences, and enzymatic properties of glucose dehydrogenase isozymes from *Bacillus megaterium* IAM1030," *Journal of Bacteriology*, vol. 174, no. 15, pp. 5013–5020, 1992.

[12] K. Suga, Y. Shiba, T. Sorai, S. Shioya, and F. Ishimura, "Reaction kinetics and mechanism of immobilized penicillin acylase from *Bacillus megaterium*," *Annals of the New York Academy of Sciences*, vol. 613, pp. 808–815, 1990.

[13] L. Martin, M. A. Prieto, E. Cortes, and J. L. Garcia, "Cloning and sequencing of the pac gene encoding the penicillin G acylase of *Bacillus megaterium* ATCC 14945," *FEMS Microbiology Letters*, vol. 125, no. 2-3, pp. 287–292, 1995.

[14] W. Panbangred, K. Weeradechapon, S. Udomvaraphant, K. Fujiyama, and V. Meevootisom, "High expression of the penicillin G acylase gene (pac) from *Bacillus megaterium* UN1 in its own pac minus mutant," *Journal of Applied Microbiology*, vol. 89, no. 1, pp. 152–157, 2000.

[15] E. Raux, A. Lanois, M. J. Warren, A. Rambach, and C. Thermes, "Cobalamin (vitamin B12) biosynthesis: identification and characterization of a Bacillus megaterium cobI operon," *Biochemical Journal*, vol. 335, no. 1, pp. 159–166, 1998.

[16] C. Korneli, F. David, R. Biedendieck, D. Jahn, and C. Wittmann, "Getting the big beast to work—systems biotechnology of *Bacillus megaterium* for novel high-value proteins," *Journal of Biotechnology*, vol. 163, no. 2, pp. 87–96, 2013.

[17] "Manipulation of *B. megaterium* growth for efficient 2-KLG production by *K. vulgare*," *Process Biochemistry*, vol. 45, no. 4, pp. 602–606, 2010.

[18] M. Eppinger, B. Bunk, M. A. Johns et al., "Genome sequences of the biotechnologically important *Bacillus megaterium* strains QM B1551 and DSM319," vol. 193, no. 16, pp. 4199–4213, 2011.

[19] L. Liu, Y. Li, J. Zhang et al., "Complete genome sequence of the industrial strain *Bacillus megaterium* WSH-002," *Journal of Bacteriology*, vol. 193, no. 22, pp. 6389-6390, 2011.

[20] W. Shi, W. Lu, Q. Liu, Y. Zhi, and P. Zhou, "The identification of the nitrate assimilation related genes in the novel *Bacillus megaterium* NCT-2 accounts for its ability to use nitrate as its only source of nitrogen," *Functional & Integrative Genomics*, vol. 14, no. 1, pp. 219–227, 2014.

[21] J. Eid, A. Fehr, J. Gray et al., "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[22] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713-714, 2008.

[23] A. McKenna, M. Hanna, E. Banks et al., "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[24] M. Boetzer and W. Pirovano, "SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information," *BMC Bioinformatics*, vol. 15, no. 1, article 211, 2014.

[25] E. W. Myers, G. G. Sutton, A. L. Delcher et al., "A whole-genome assembly of Drosophila," *Science*, vol. 287, no. 5461, pp. 2196–2204, 2000.

[26] A. L. Delcher, K. A. Bratke, E. C. Powers, and S. L. Salzberg, "Identifying bacterial genes and endosymbiont DNA with Glimmer," *Bioinformatics*, vol. 23, no. 6, pp. 673–679, 2007.

[27] G. Benson, "Tandem repeats finder: a program to analyze DNA sequences," *Nucleic Acids Research*, vol. 27, no. 2, pp. 573–580, 1999.

[28] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Research*, vol. 32, Supplement 1, pp. D277–D280, 2004.

[29] R. K. Aziz, D. Bartels, A. A. Best et al., "The RAST Server: rapid annotations using subsystems technology," *BMC Genomics*, vol. 9, no. 1, p. 75, 2008.

[30] R. Overbeek, R. Olson, G. D. Pusch et al., "The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST)," *Nucleic Acids Research*, vol. 42, no. D1, pp. D206–D214, 2014.

[31] P. Stothard and D. S. Wishart, "Circular genome visualization and exploration using CGView," *Bioinformatics*, vol. 21, no. 4, pp. 537–539, 2005.

[32] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, "Mauve: multiple alignment of conserved genomic sequence with rearrangements," *Genome Research*, vol. 14, no. 7, pp. 1394–1403, 2004.

[33] J. Qi, B. Wang, and B. I. Hao, "Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach," *Journal of Molecular Evolution*, vol. 58, no. 1, pp. 1–11, 2004.

[34] G. H. Zuo and B. L. Hao, "CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy," *Genomics, Proteomics & Bioinformatics*, vol. 13, no. 5, pp. 321–331, 2015.

[35] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[36] K. Tamura, M. Nei, and S. Kumar, "Prospects for inferring very large phylogenies by using the neighbor-joining method," *Proceedings of the National Academy of Sciences*, vol. 101, no. 30, pp. 11030–11035, 2004.

[37] S. Kumar, G. Stecher, and K. Tamura, "MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets," *Molecular Biology and Evolution*, vol. 33, no. 7, pp. 1870–1874, 2016.

[38] C. Zhu, Y. Mahlich, M. Miller, and Y. Bromberg, "fusionDB: assessing microbial diversity and environmental preferences via functional similarity networks," *Nucleic Acids Research*, vol. 46, no. D1, pp. D535–D541, 2018.

[39] C. Zhu, T. O. Delmont, T. M. Vogel, and Y. Bromberg, "Functional basis of microorganism classification," *PLoS Computational Biology*, vol. 11, no. 8, article e1004472, 2015.

[40] K. T. Konstantinidis and J. M. Tiedje, "Trends between gene content and genome size in prokaryotic species with larger genomes," *Proceedings of the National Academy of Sciences*, vol. 101, no. 9, pp. 3160–3165, 2004.

[41] J. A. Klappenbach, J. M. Dunbar, and T. M. Schmidt, "rRNA operon copy number reflects ecological strategies of bacteria," *Applied and Environmental Microbiology*, vol. 66, no. 4, pp. 1328–1333, 2000.

[42] P. M. Shrestha, M. Noll, and W. Liesack, "Phylogenetic identity, growth-response time and rRNA operon copy number of soil bacteria indicate different stages of community succession," *Environmental Microbiology*, vol. 9, no. 10, pp. 2464–2474, 2007.

[43] C. Andres-Barrao, F. F. Lafi, I. Alam et al., "Complete genome sequence analysis of *Enterobacter* sp SA187, a plant multi-stress tolerance promoting endophytic bacterium," *Frontiers in Microbiology*, vol. 8, 2017.

[44] K. Yano, T. Wada, S. Suzuki et al., "Multiple rRNA operons are essential for efficient cell growth and sporulation as well as

outgrowth in *Bacillus subtilis*," *Microbiology*, vol. 159, Part 11, pp. 2225–2236, 2013.

[45] B. C. Carlton and D. R. Helinski, "Heterogeneous circular DNA elements in vegetative cultures of Bacillus megaterium," *Proceedings of the National Academy of Sciences*, vol. 64, no. 2, pp. 592–599, 1969.

[46] B. S. Stevenson and T. M. Schmidt, "Growth rate-dependent accumulation of RNA from plasmid-borne rRNA operons in *Escherichia coli*," *Journal of Bacteriology*, vol. 180, no. 7, pp. 1970–1972, 1998.

[47] S. H. Chu, D. Zhang, D. X. Wang, Y. E. Zhi, and P. Zhou, "Heterologous expression and biochemical characterization of assimilatory nitrate and nitrite reductase reveals adaption and potential of *Bacillus megaterium* NCT-2 in secondary salinization soil," *International Journal of Biological Macromolecules*, vol. 101, pp. 1019–1028, 2017.

[48] D. J. Richardson and N. J. Watmough, "Inorganic nitrogen metabolism in bacteria," *Current Opinion in Chemical Biology*, vol. 3, no. 2, pp. 207–219, 1999.

[49] D. Kleiner, "NH4+ transport systems: CRC Press," in *Alkali cation transport systems in prokaryotes*, pp. 379–396, CRC Press, Inc., Boca Raton, Fla, 1993.

[50] E. Soupene, L. He, D. Yan, and S. Kustu, "Ammonia acquisition in enteric bacteria: physiological role of the ammonium/-methylammonium transport B (AmtB) protein," *Proceedings of the National Academy of Sciences*, vol. 95, no. 12, pp. 7030–7034, 1998.

[51] H. J. Beaumont, S. I. Lens, W. N. Reijnders, H. V. Westerhoff, and R. J. van Spanning, "Expression of nitrite reductase in Nitrosomonas europaea involves NsrR, a novel nitrite-sensitive transcription repressor," *Molecular Microbiology*, vol. 54, no. 1, pp. 148–158, 2004.

[52] N. Oteino, R. D. Lally, S. Kiwanuka et al., "Plant growth promotion induced by phosphate solubilizing endophytic *Pseudomonas* isolates," *Frontiers in Microbiology*, vol. 6, article 745, 2015.

[53] D. Duca, J. Lorv, C. L. Patten, D. Rose, and B. R. Glick, "Indole-3-acetic acid in plant-microbe interactions," *Antonie Van Leeuwenhoek International Journal of General and Molecular*, vol. 106, no. 1, pp. 85–125, 2014.

[54] O. S. Olanrewaju, B. R. Glick, and O. O. Babalola, "Mechanisms of action of plant growth promoting bacteria," *World Journal of Microbiology and Biotechnology*, vol. 33, no. 11, article 197, 2017.

[55] A. W. Woodward and B. Bartel, "Auxin: regulation, action, and interaction," *Annals of Botany*, vol. 95, no. 5, pp. 707–735, 2005.

[56] P. Zhang, T. Jin, S. Kumar Sahu et al., "The distribution of tryptophan-dependent indole-3-acetic acid synthesis pathways in bacteria unraveled by large-scale genomic analysis," *Molecules*, vol. 24, no. 7, article 1411, 2019.

[57] B. Kempf and E. Bremer, "Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments," *Archives of Microbiology*, vol. 170, no. 5, pp. 319–330, 1998.

Hindawi
International Journal of Genomics
Volume 2020, Article ID 9354204, 12 pages
https://doi.org/10.1155/2020/9354204

Research Article

Comparative Genomics of *Actinobacillus pleuropneumoniae* Serotype 8 Reveals the Importance of Prophages in the Genetic Variability of the Species

Isabelle Gonçalves de Oliveira Prado,[1] **Giarlã Cunha da Silva**,[1] **Josicelli Souza Crispim**,[1] **Pedro Marcus Pereira Vidigal**,[2] **Moysés Nascimento**,[3] **Mateus Ferreira Santana**,[1] **and Denise Mara Soares Bazzolli**[1]

[1] *Laboratório de Genética Molecular de Bactérias/Bioagro-Departamento de Microbiologia, Universidade Federal de Viçosa, Viçosa 36570-900, Brazil*
[2] *Núcleo de Análise de Biomoléculas (NuBioMol), Universidade Federal de Viçosa, Viçosa, Brazil*
[3] *Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, Brazil*

Correspondence should be addressed to Denise Mara Soares Bazzolli; dbazzolli@ufv.br

Received 9 October 2019; Accepted 3 January 2020; Published 19 February 2020

Guest Editor: Tarja Sironen

Actinobacillus pleuropneumoniae is the etiologic agent of porcine pleuropneumonia. Currently, there are 18 different serotypes; the serotype 8 is the most widely distributed in the United States, Canada, United Kingdom, and southeastern Brazil. In this study, genomes of seven *A. pleuropneumoniae* serotype 8 clinical isolates were compared to the other genomes of twelve serotypes. The analyses of serotype 8 genomes resulted in a set of 2352 protein-coding sequences. Of these sequences, 76.6% are present in all serotypes, 18.5% are shared with some serotypes, and 4.9% were differential. This differential portion was characterized as a series of hypothetical and regulatory protein sequences: mobile element sequence. Synteny analysis demonstrated possible events of gene recombination and acquisition by horizontal gene transfer (HGT) in this species. A total of 30 sequences related to prophages were identified in the genomes. These sequences represented 0.3 to 3.5% of the genome of the strains analyzed, and 16 of them contained complete prophages. Similarity analysis between complete prophage sequences evidenced a possible HGT with species belonging to the family Pasteurellaceae. Thus, mobile genetic elements, such as prophages, are important components of the differential portion of the *A. pleuropneumoniae* genome and demonstrate a central role in the evolution of the species. This study represents the first study done to understand the genome of *A. pleuropneumoniae* serotype 8.

1. Introduction

Pork is an important source of animal protein and is currently one of the most commonly consumed meat products in the world [1]. However, the use of an intensive production system has frequently given rise to the occurrence of respiratory diseases, having a major impact on production, causing significant economic losses in pig farming [2, 3]. Porcine pleuropneumonia is one of the most important respiratory diseases in pigs and is caused by the bacterium *Actinobacillus pleuropneumoniae*; this species can be divided into two biotypes according to their dependence on nicotinamide adenine dinucleotide (NAD) [4]. Currently, this species is classified into 18 serotypes based on the antigenic properties of capsule polysaccharides [5–7].

The pathogenesis of porcine pleuropneumonia is complex and involves different virulence factors produced by the bacterium [8–10]. Virulence is multifactorial and is related to a combination of factors such as toxins from the RTX family, composition and structure of capsule polysaccharides, outer membrane lipopolysaccharide (LPS), iron siderophores, biofilm formation, and adhesins [8]. In addition to the

Table 1: *A. pleuropneumoniae* genomes used in this study.

Strain/serotype	Genome size (pb)	CDS	% GC	Accession code (WGS)	Reference
4074/1	2318649	2135	41.2	CP029003.1	Xu et al., [16]
4226/2	2314083	2228	41.2	ADXN00000000.1	Zhan et al., 2010
JL03, 3	2242062	2115	41.2	CP000687.1	Xu et al., 2008
M62/4	2260565	2186	41.2	ADOF00000000.1	Xu et al., [16]
L20/5b	2274482	2168	41.3	CP000569.1	Foote et al., 2008
Femo/6	2302700	2300	41.0	ADOG00000000.1	Xu et al., [16]
AP76/7	2345435	2234	41.2	CP001091.1	Linke et al., 2008
MV460/8	2213381	2116	41.1	JSVG00000000.1	Pereira et al., [23]
MV518/8	2275540	2189	41.1	JSVZ00000000.1	
MV597/8	2219395	2115	41.1	JSVX00000000.1	
MV780/8	2274000	2179	41.1	JSVV00000000.1	
MV1022/8	2262828	2179	41.1	JSVF00000000.1	
MV5651/8	2264279	2179	41.1	JSVY00000000.1	
MIDG2331/8	2337633	2235	41.1	LN908249.1	Bossé et al., [22]
CVJ13261/9	2256417	2163	41.2	ADOI00000000.1	Xu et al., [16]
D13039/10	2266276	2155	41.2	ADOJ00000000.1	
56153/11	2257884	2154	41.2	ADOK00000000.1	
1096/12	2185499	2082	41.2	ADOL00000000.1	
N273/13	2236660	2148	41.2	ADOM00000000.1	

abovementioned virulence factors, some *A. pleuropneumoniae* serotypes present natural competence, and therefore, the occurrence of natural transformation is common in this species and dissemination of resistance genes in Pasteurellaceae family members, as *A. pleuropneumoniae*, is common [11–13]. Finally, virulence is complex, and antimicrobial resistance genes that can be encoded by both the chromosome and plasmids are essential depending on the specific niche, as in the natural hosts and in specific conditions [14, 15].

Although comparative genomic studies with different genotypes of *A. pleuropneumoniae* serotypes were carried out [16], no information on *A. pleuropneumoniae* serotype 8 was provided so far. Over the years, *A. pleuropneumoniae* serotype 8 has been neglected in identification studies due to failures in serotyping techniques, and as a result, genomic studies involving this serotype are nonexistent. Although recent studies have showed a wide distribution of this serotype in several regions, such as the United Kingdom [17, 18], North America [19], and Brazil [20], just recently *A. pleuropneumoniae* serotype 8 genome sequence was available [21, 22].

A study carried out by our research group using the alternative host, *Galleria mellonella* larvae, detected different virulence patterns in clinical isolates of serotype 8 *A. pleuropneumoniae* [23]. Based on the results obtained from that study, six isolates with different phenotypic profiles were selected for genomic sequencing. Clinical isolates from *A. pleuropneumoniae* serotype 8 have virulence complexity [9], but no specific information on genotypic variability is available so far. Then, this study is the first to describe *A. pleuropneumoniae* serotype 8 genomic from comparative analysis between *A. pleuropneumoniae* serotype 8 genomes (clinical isolates: Brazilian origin [21], one of English origin [22]) and twelve genomes of different serotypes of *A. pleuropneumoniae* deposited in databases.

2. Materials and Methods

2.1. Actinobacillus pleuropneumoniae Genome Sequences. Nineteen genomes from different serotypes of *A. pleuropneumoniae* available at the GenBank database (https://www.ncbi.nlm.nih.gov/genbank/) were used in the present study. The genomes of *A. pleuropneumoniae* serotypes 1 (4074), 13 (JL03), 5 (L20), 7 (AP76), and 8 (MIDG2331) are closed. The genomes of the other serotypes are in contigs (Table 1).

2.2. Determination of Protein-Coding Sequence Set in A. pleuropneumoniae Serotype 8. The set of clusters of the coding DNA sequences (CDS) predicted for *A. pleuropneumoniae* serotype 8 was based on the seven genomes of clinical isolates taken from pig farms, six from Brazil [21] and one from the United Kingdom [22]. In this analysis, the CD-HIT v.4.6.1 program [24, 25] was used to consider an identity threshold of 0.85 to cluster the CDS. For the functional annotation of the *A. pleuropneumoniae* serotype 8 reference genome, five databases were used: COG [26], CDD [27], PFAM [28], SMART [29], and UNIPROT [30]. The similarity searches were carried out using the BLAST algorithm [31] considering an E value ≤ 10^{-5}.

2.3. Comparative Analysis of Predicted CDS of Serotype 8 and the Other 12 Different Serotypes of A. pleuropneumoniae. In this analysis, 12 genomes of *A. pleuropneumoniae* different serotypes (1, 2, 3, 4, 5b, 6, 7, 9, 10, 11, 12, and 13) were used (Table 1). These sequences are deposited in the UNIPROT database [30]. The comparative analysis was carried out using the BLAST algorithm [31], contrasting the genomes of the serotypes analyzed against the *A. pleuropneumoniae* serotype 8 reference assembled in this study.

2.4. Analysis of A. pleuropneumoniae Orthologous Gene Groups. From the predicted CDS of the 12 different serotypes and 7 serotype 8 *A. pleuropneumoniae* genomes, a database containing 28002 CDS corresponding to all serotypes of the species was assembled. Using the CD-HIT v.4.6.1 program [24, 25], with an identity threshold of 0.70 identity to cluster the sequences, an analysis was carried out to characterize the total set of CDS of the species. The CD-HIT was used for clustering the sequence and for reducing redundancy among them, to improve the results.

The groups of CDS identified by the CD-HIT were classified as core, shared, or differential. Additionally, the predicted protein sequences of *A. pleuropneumoniae* serotype 8 were individually compared to the predicted protein sequences of the other serotypes using the BLAST algorithm [31].

2.5. Genome-Wide Analysis of Preferential Codon Usage and GC%. The analysis of the preferential use of codons and GC content was carried out using the EMBOSS program [32] for the different serotype genomes of *A. pleuropneumoniae*. The use of each synonymous codon was determined by calculating the RSCU (Relative Synonymous Codon Usage). The RSCU value calculated for each codon was the parameter used to evaluate the codon selection type, with values = 1 characteristic of codons used with equal frequency; values > 1 were positive selection and <1 negative selection.

2.6. Synteny Analysis. The analyses were derived from the closed genomes of *A. pleuropneumoniae* serotypes 03 (JL03), 5b (L20), 7 (AP76), and 8 (MIDG2331) and from the six genomes of *A. pleuropneumoniae* serotype 8 isolates of Brazilian origin (MV460, MV518, MV597, MV780, MV1022, and MV5651). Multiple alignments of the sequences of the *A. pleuropneumoniae* genomes were derived from the Progressive Mauve v.2.3.1 software program [33].

2.7. Analysis of Sequences Similar to Prophages. Sequences similar to the prophages present in all the *A. pleuropneumoniae* genomes used in this study were obtained through the PHASTER program [34]. The prophage sequences were aligned by MAFFT [35], and the alignment was edited using the GBLOCKS program [36]. A dendrogram using the Neighbor-Joining genetic distance grouping method was generated by the MEGA 6 program [37] with a bootstrap containing 2000 replicates. Prophage complete sequences were compared using BLAST [31] against the GenBank databases to identify possible horizontal gene transfers between bacteria. For this, a coverage and identity above 70% and an *E* value less than 10^{-5} were used as cutoff points. After editing, the alignment was obtained using the GBLOCKS program. Using the same complete sequences of prophages, an alignment was done with the Clustal Omega [38]. From the values of the identity matrix provided, a heat map was constructed with software R under version 3.5.1.

3. Results

3.1. Genomic Analysis of A. pleuropneumoniae Serotype 8. The total set of predicted CDS of *A. pleuropneumoniae* serotype 8 generated from the seven clinical isolates corresponded to 2352 sequences (Table 2). Of these, 1801 (76.6%) were considered core, 436 (18.5%) were shared with other serotypes, though not all, and 115 (4.9%) were predicted to be differential to *A. pleuropneumoniae* serotype 8 genomes. Among the 2352 CDS of *A. pleuropneumoniae* serotype 8, 1925 (81.8%) were categorized into the COG database. Among these were 1542 (80.1%) encode proteins with known functional categories (excluding "Unknown function" and "Prediction of general functions").

From the distinction of the core, shared and differential regions of *A. pleuropneumoniae* serotype 8 CDS and clusters of ortholog groups were analyzed. Of the 1801 sequences comprising the core portion, 1685 were affiliated to the categories of the COG database. Among these, 1358 sequences (80.6%) represent known functional categories (Table 2). The majority of sequences characterized as core are related to amino acid metabolism and transport; ribosomal translation, structure, and biogenesis processes; biogenesis of the wall, membrane, and cell envelope; and production and conservation of energy, among other activities considered essential to the survival of the pathogen (Table 2).

As regards the shared portion, of the 436 sequences, 220 were affiliated with the COG database categories, of which 166 (75.5%) were known functional categories. Most sequences are related to the metabolism and transport of inorganic ions; biogenesis of the cellular envelope; replication, recombination, and DNA repair; and metabolism in general (Table 2; Supplementary Data Table 1).

The differential portion of the *A. pleuropneumoniae* serotype 8 genomes showed 115 CDS. Only 20 sequences were affiliated to the COG database categories, of which 18 (90.0%) were known functional categories (Table 2). These differential C are related to the regulatory processes and HGT mechanisms such as plasmids and prophages (Table 2; Supplementary Data Table 1). In this portion, CDS related to resistance to antibiotics such as tetracycline and florfenicol genes, transcriptional regulators such as LysR, DNA repair protein, transposon gamma-delta resolvase, transport proteins such as sodium and glutamate symmetric acetyltransferase, and prophage-related protein-coding sequences were reported (Table 2).

3.2. The Pangenome of A. pleuropneumoniae. From the 2984 clusters obtained from the total set of CDS of the species, the general characterization of the CDS with the distinction of the core, shared, and differential portions was carried out. Of the total, 1737 clusters were characterized as core region, present in the thirteen serotypes analyzed; 756 were clusters of CDS of shared proteins, and 491 clusters corresponded to CDS of differential proteins of each serotype. As regards the total genome of each serotype, the core portion averaged 82.5% (Table 3) showing conservation among the different serotypes.

Table 2: Coding sequences of *A. pleuropneumoniae* serotype 8.

COG	Description of COG classes	Core	Shared	Differential	Total
A	RNA modification and processing	1	0	0	1
C	Conversion and production of energy	117	9	0	126
D	Cycle control and cell division, chromosome partitioning	25	3	0	28
E	Amino acid metabolism and transport	158	10	1	169
F	Nucleotide metabolism and transport	60	2	0	62
G	Carbohydrate metabolism and transport	117	11	2	130
H	Coenzyme metabolism and transport	91	10	1	102
I	Lipid metabolism and transport	39	3	0	42
J	Translation, ribosomal structure, and biogenesis	153	7	2	162
K	Transcript	77	10	7	94
L	Replication, recombination, and repair	95	20	4	119
M	Biogenesis of cell wall, membrane, and envelope	127	21	0	148
N	Cellular motility	6	1	1	8
O	Posttranslational modification, protein turnover, and chaperones	96	6	0	102
P	Metabolism and transport of inorganic ions	104	24	0	128
Q	Biosynthesis of secondary metabolites, transport, and catabolism	8	5	0	13
T	Signal transduction mechanisms	31	1	0	32
U	Intracellular traffic, secretion, and vesicular transport	33	5	0	38
V	Defense mechanisms	20	18	0	38
R	Prediction of general functions	158	23	2	183
S	Unknown function	169	31	0	200
NC	Proteins not categorized on COG	116	216	95	427
	Total of affiliated proteins	1685	220	20	1925
	Total of serotype 8 proteins	1801	436	115	2352

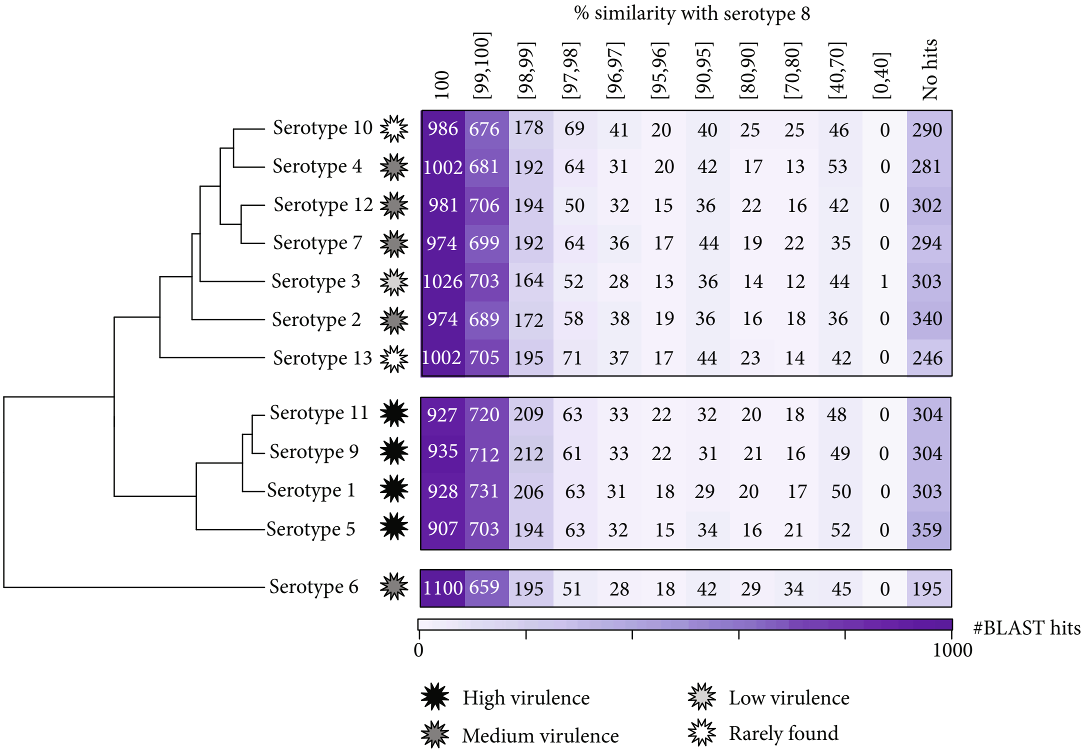
Table 3: Characterization of protein groups in different serotypes of *A. pleuropneumoniae.*

Serotype	Total proteins	Core Proteins (%)	Shared Proteins (%)	Differential Proteins (%)
Serotype 1	2176	1765 (81.1)	404 (18.6)	7 (0.3)
Serotype 2	2064	1774 (86.0)	275 (13.3)	15 (0.7)
Serotype 3	2026	1756 (86.7)	260 (12.8)	10 (0.5)
Serotype 4	2219	1790 (80.7)	325 (14.7)	104 (4.7)
Serotype 5b	2004	1765 (88.1)	208 (10.4)	31 (1.6)
Serotype 6	2211	1768 (80.0)	384 (17.4)	59 (2.7)
Serotype 7	2113	1774 (84.0)	327 (15.5)	12 (0.6)
Serotype 8*	2352	1801 (76.6)	436 (18.5)	115 (4.9)
Serotype 9	2197	1779 (81.0)	416 (18.9)	2 (0.1)
Serotype 10	2170	1774 (82.0)	321 (14.8)	75 (3.5)
Serotype 11	2184	1767 (81.0)	414 (19.0)	3 (0.1)
Serotype 12	2081	1771 (85.1)	294 (14.1)	16 (0.8)
Serotype 13	2145	1760 (82.1)	381 (17.8)	4 (0.2)

*Reference genome represents all the sequences encoding the seven clinical isolate genomes previously sequenced.

3.3. Similarity Analysis between the Amino Acid Sequences Predicted for A. pleuropneumoniae Serotype 8 and Other Serotypes. An alignment between predicted amino acid sequences of *A. pleuropneumoniae* serotype 8 was created against all other *A. pleuropneumoniae* amino acid sequences used in this study generating clusters based on the pattern of similarity (Figure 1). Of the total 2352 amino acid sequences, 2196 (93.4%) had similarity patterns higher than 95%, thus revealing high serotype 8 sequence conservation in relation to the others. Based on the analysis of the BLAST results, three main groups of similarity related to high, medium, and low virulence standards were obtained. There was a greater sharing of the predicted CDS of serotype 8 with the serotype 6 sequences, followed by serotype 3 (Figure 1).

3.4. Codon Preferential Usage. As regards the codon analysis, a high standard of conservation was observed in the use of codons among all *A. pleuropneumoniae* serotypes investigated, which includes the clinical isolates of serotype 8 analyzed in this study. In Figure 2, we have represented the use of codons by *A. pleuropneumoniae*. Codons with higher RSCU values result in higher positive selection for their respective amino acids (Figure 2). We observed no significant differences in the proportions of the use of amino acids between the different isolates nor between the serotypes. The most commonly used amino acids were leucine (L: 10.6%), alanine (A: 8.7%), isoleucine (I: 6.8%), and valine (V: 6.8%), while cysteine (C: 1.0%) and tryptophan (W: 1.2%) were the most rarely used (Figure 2).



	100	[99,100]	[98,99]	[97,98]	[96,97]	[95,96]	[90,95]	[80,90]	[70,80]	[40,70]	[0,40]	No hits
Serotype 10	986	676	178	69	41	20	40	25	25	46	0	290
Serotype 4	1002	681	192	64	31	20	42	17	13	53	0	281
Serotype 12	981	706	194	50	32	15	36	22	16	42	0	302
Serotype 7	974	699	192	64	36	17	44	19	22	35	0	294
Serotype 3	1026	703	164	52	28	13	36	14	12	44	1	303
Serotype 2	974	689	172	58	38	19	36	16	18	36	0	340
Serotype 13	1002	705	195	71	37	17	44	23	14	42	0	246
Serotype 11	927	720	209	63	33	22	32	20	18	48	0	304
Serotype 9	935	712	212	61	33	22	31	21	16	49	0	304
Serotype 1	928	731	206	63	31	18	29	20	17	50	0	303
Serotype 5	907	703	194	63	32	15	34	16	21	52	0	359
Serotype 6	1100	659	195	51	28	18	42	29	34	45	0	195

Figure 1: Analysis of similarity between predicted amino acid sequences of *A. pleuropneumoniae* serotype 8 and the other serotypes. The protein-coding sequences were clustered according to similarity in percentages to serotype 8. The serotypes were also characterized in relation to virulence.

3.5. Synteny Analysis of A. pleuropneumoniae. In the alignment between the genotype-representative contigs of the six Brazilian clinical isolates of *A. pleuropneumoniae* serotype 8 and closed genomes of serotypes 3 (JL03), 5b (L20), 7 (AP76), and 8 (MIDG2331), we verified conservation in the genome structure (Figure 3). The genomes of *A. pleuropneumoniae* share practically the same blocks, denominated LCBs: "Selecting Locally Collinear Blocks." Although genome alignment conservation was found, it was possible to observe regions of acquisition/loss of genetic material and rearrangements (Figure 3). The presence of a differential and conserved block in the genomes of serotypes 5 (L20) and 7 (AP76) as well as in four of the clinical isolates of serotype 8 (MV518, MV780, MV1022, and MV5651) was observed (Figure 3). In this block composed of approximately 42.206 pb and with GC content of 40.4%, there are sequences encoding proteins related to integrase (WP_005620278.1), pyrophosphatase (WP_005620280.1), RdgC recombination protein (WP_011848390.1), DNA methyltransferase (WP_011848391.1), antirepressor (WP_011848395.1, WP_011848414.1), DNA methylase (WP_011848397.1), endodeoxyribonuclease RuvA (WP_011848399.1), terminase (WP_043880767.1), peptidase (WP_043877971.1), and various phage proteins (WP_011848407.1, WP_011848405.1, WP_011848408.1, and WP_011848412.1). In a second differential genomic segment, present only in serotype 7, a region of inversion and rearrangement of a block of approximately 59.649 bp and GC content of 40.9% was present (Figure 3). In this segment, we found sequences corresponding to genes encoding carboxylase enzymes (WP_005617934.1, WP_005602033.1), oxidoreductases (WP_005598646.1, WP_005602054.1, and WP_012478542.1), reductase (WP_005598644.1), virulence factors involved in iron uptake (WP_005602070.1), integrase (WP_005617888.1), and transposases (WP_005599960.1).

Minor variations were also observed between the different genomes (Figure 3). Among these differences, we found a small region of approximately 13000 pb present only in the seven genomes of *A. pleuropneumoniae* serotype 8 and the CDS found were for tRNA-glutamate ligase (WP_005608501.1), tRNA-Ala (WP_005612726.1), preprotein translocase (WP_005612726.1), transcriptional regulator of the Rha family (WP_039768145.1), antirepressor (WP_058230489.1), propanediol utilization protein (WP_039768152.1), host death prevention protein family (Phd) (WP_005598318.1), YoeB toxin (WP_005605064.1), tetracyl disaccharide kinase (WP_005598320.1, WP_005608502.1), and nine sequences encoding hypothetical proteins (WP_039709488.1, WP_039768147.1, WP_039709486.1, WP_039709484.1, WP_039709483.1, WP_052250595.1, WP_014991324.1, WP_039768150.1, and WP_014991326.1). Another region of approximately 13180 pb was found in the genomes of *A. pleuropneumoniae* serotype 5 (L20), serotype 7 (AP76), and in three serotype 8 clinical isolates (MIDG2331, MV1022, and MV518). Sequences corresponding to genes present in this region were encoded for *flp*C operon (WP_039709641.1), *flp*B operon (WP_039709034.1, WP_011848427.1), fimbriae protein (WP_058230512.1, WP_039709035.1, WP_005611759.1, and WP_011848428.1), ATP-dependent ATP-RhI-RNA (WP_039709036.1), ATP-

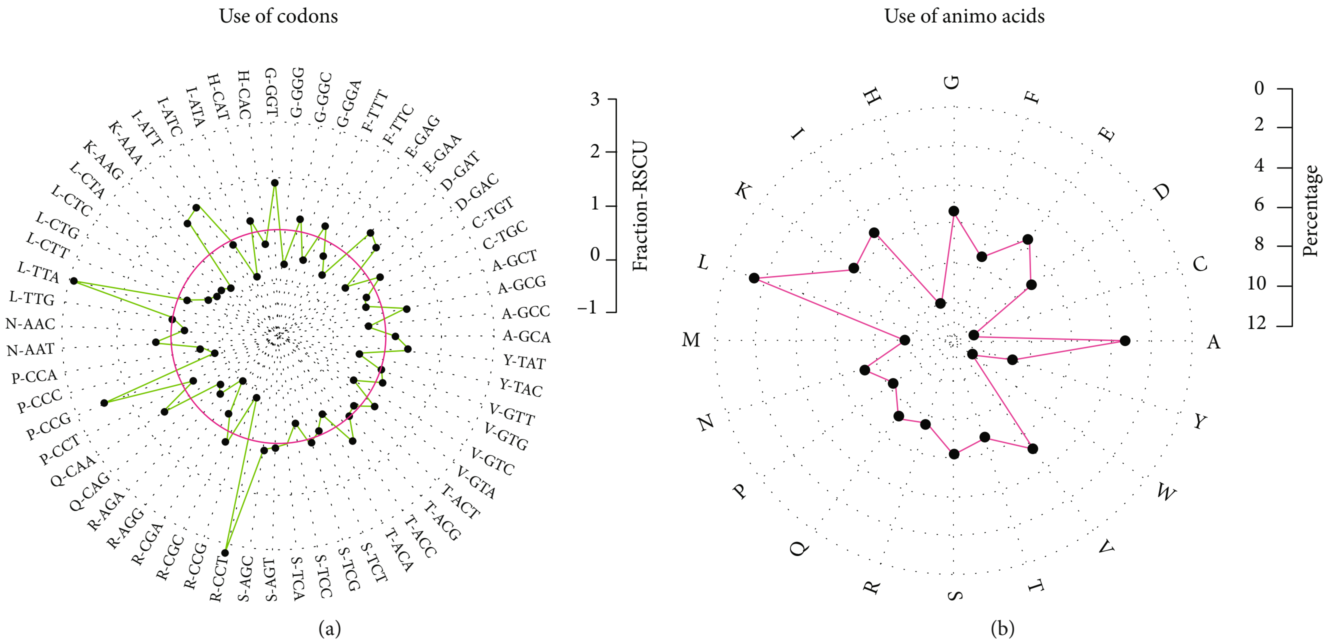


Figure 2: Use of codons and their respective amino acids of *A. pleuropneumoniae*. (a) The trend in the use of codons is represented in the circular map. Methionine, tryptophan, and stop codons were omitted. Synonymous codons for an amino acid used with equal frequency have RSCU = 1, indicated by the red circular line. (b) Percentage of the use of amino acids represented in the circular map. A: alanine; C: cysteine; D: aspartic acid; E: glutamic acid; F: phenylalanine; G: glycine; H: histidine; I: isoleucine; K: lysine; L: leucine; M: methionine; N: asparagine; P: proline; Q: glutamine; R: arginine; S: serine; T: threonine; V: valine; W: tryptophan; Y: tyrosine.

binding protein (WP_005611764.1), iron-ABC transport substrate binding (WP_005611765.1), and three hypothetical proteins (WP_005611761.1, WP_009875478.1, and WP_005611761.1).

3.6. Analysis of Prophage Sequences. 30 sequences similar to the prophages were found in the 19 strains of *A. pleuropneumoniae* analyzed. From the total of sequences similar to prophages, 16 were classified by the PHASTER program as complete, 11 as incomplete, and 3 as questionable (Supplementary Data Table 2). Incomplete and questionable sequences were considered genomic regions containing sequences derived from phages. The regions containing prophage-related genes represent 0.3 to 3.5% of the genomes analyzed (Supplementary Data Table 2). The largest prophage identified had 48.1 kb and the smallest 22.4 kb, identified in the MV1022 and M62 strains, respectively (Supplementary Data Table 2). The GC content of the identified prophages varied between 39.3 and 44.6% (Supplementary Data Table 2).

The Neighbor-Joining method to cluster analysis between sequences similar to complete prophages allowed us to identify four different clusters (P1-P4) (Figure 4). Analysis of the complete prophages using the BLAST algorithm against sequences from the GenBank database showed that the sequences contained in the P3 cluster (prophage 2 (4074 strain), prophage 1 (4226 strain), prophage 1 (CVJ13261), and prophage 2 (56153 strain)), the M62 prophage 2, and the AP76 prophage 2 share over 70% identity and were found in genomes of *Haemophilus ducreyi* (AE017143.1), *Mannheimia haemolytica* (KP137440.1), and *Actinobacillus suis* (CP009159.1), respectively (Table 4). For the other 13 complete prophages, no significant identity or sequence coverage was found in GenBank.

The heat map distribution showed a high identity among the prophage 1 from 518, 780, and 5651 (serotype 8) and femo (serotype 6) (Figure 5). This could also be observed among prophage 2 from 56153 (serotype 11) and 4074 (serotype 1) and prophage 1 from CVJ13261 (serotype 9) and 4226 (serotype 2). A considerable identity also was observed among prophage 1 from 518, 780, and 5651 (serotype 8), femo (serotype 6), and N273 (serotype 13). Similarly, it also was observed between N273 (serotype 13) and AP76 (serotype 7) and L20 (serotype 5) and AP76 (serotype 7).

4. Discussion

Analyses of GC content, codon usage, and amino acid use among the different *A. pleuropneumoniae* serotypes showed that they share a set of conserved CDS. The core portion of the genome that is well conserved among the serotypes also reinforces these results. Among the most commonly used amino acids are branched chain amino acids, such as leucine, isoleucine, and valine. These branched chain amino acids are required for the survival and virulence of *A. pleuropneumoniae* in swine, capable of synthesizing these amino acids critical for respiratory tract pathogens [39].

In the analyses of clusters from the set of CDS shared between serotypes, the pattern of clustering by similarity was compatible with the classification of serotypes into three virulence categories [16]: low, medium, and high virulence. We observed that serotype 8 shares a high number of

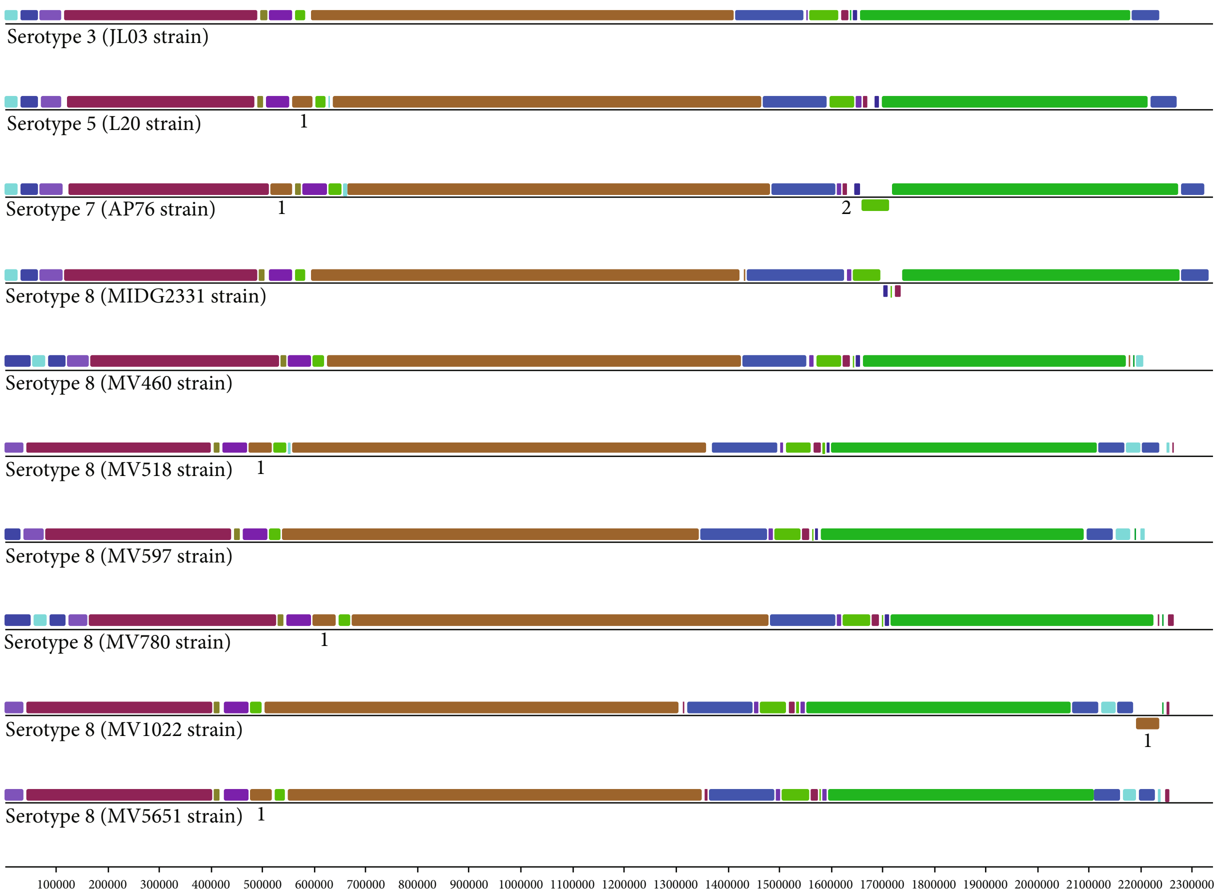


Figure 3: Synteny analysis of *A. pleuropneumoniae* genomes. The genomes represented correspond to serotype 3 (JL03 strain), serotype 5 (L20 strain), serotype 7 (AP76), and serotype 8 (MIDG2331, MV460, MV518, MV597, MV780, MV1022, and MV5651 strains). Horizontal bars represent the size of the genome (kb). The region identified in 1 represents the acquisition and loss of genomic information, and region 2 represents a recombination event.

protein-coding sequences with the serotypes characterized as having medium virulence, such as serotypes 2, 4, 6, 7, and 12. The characteristic of the serotypes considered as medium virulence category is associated with the persistence of the pathogen in the environment [8]. Additionally, a large sharing of CDS for serotype 8 proteins was observed in serotype 6, followed by serotype 3. As already reported in serotyping analyses, certain groups may cross-react and be mischaracterized. Serotypes 3, 6, and 8 of *A. pleuropneumoniae* in serotyping studies in North America constitute a single group, and discrimination of these three serotypes within this group is extremely difficult when using the antiserum technique [19].

In the COG analyses of the predicted amino acid sequences of *A. pleuropneumoniae* serotype 8, the core region is characterized by housekeeping genes. However, genes belonging to the core region may have differences in the level of DNA sequences. A number of genes classified as core like those encoding anaerobic glycerol-3-phosphate dehydrogenase subunit A (*glpA*), oxygen-independent coproporphyrinogen-III oxidase (*hemN*), heptosyltransferase family (*mutM*), tellurite resistance protein (*tehA*), sulfate transport system permease (*cysW*), thiazole biosynthesis protein (*thiH*), haloacid dehalogenase-like hydrolases (*had*) superfamily (*cof*), nucleoside diphosphate sugar epimerase, and oligopeptide transporter testify to positive selection in *A. pleuropneumoniae* [40]. In general, these genes are involved in the transporting of nutrients and cellular metabolism that show that *A. pleuropneumoniae* has responded to different environmental pressures.

In the core portion, we also found genes that, according to [41], have increased expression during the acute phase of natural infection of *A. pleuropneumoniae* in pigs. These genes were related, for example, to the assembly of curli fibers, important in the formation of biofilms [42]; to the maltose operon that may increase the competition capacity in some strains of pathogenic bacteria [43]; and to the ula operon involved with an ascorbate transport system under anaerobic conditions that can also be considered an important virulence factor for this species [41].

The accessory portion, comprising the shared portion and the differential, is characterized by genes that confer benefits to the microorganism under certain environmental conditions. The differential portion, as observed in *in silico* assays, has a strong relationship with HGT processes, containing sequence-encoding proteins common to plasmids and phages. This region can result in important adaptations, influencing the differentiated interaction of the pathogen

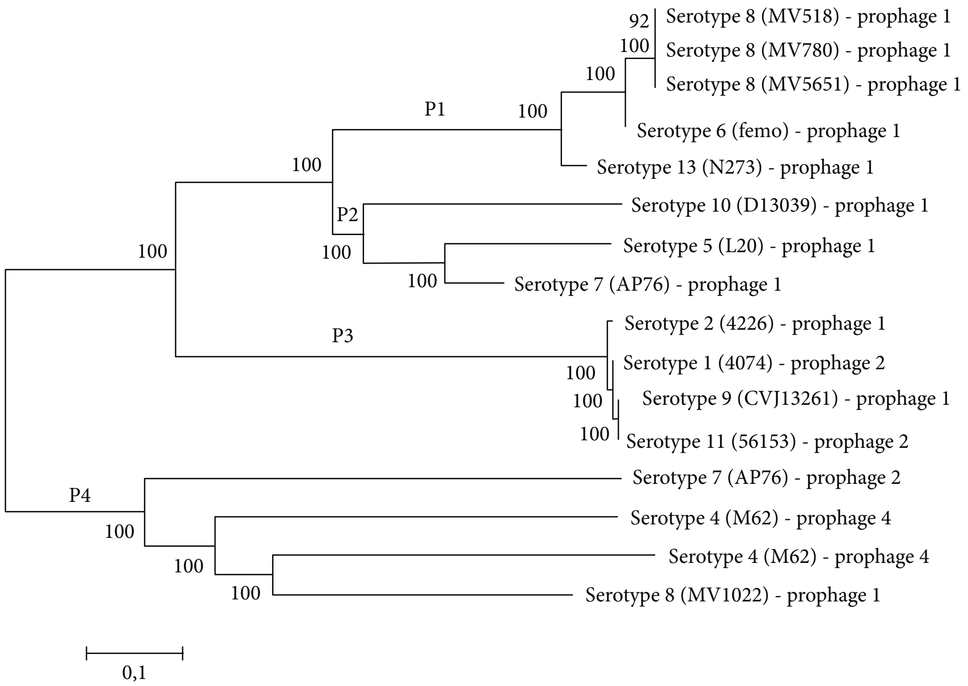

92
Serotype 8 (MV518) - prophage 1
100
Serotype 8 (MV780) - prophage 1
100
Serotype 8 (MV5651) - prophage 1
P1
100
Serotype 6 (femo) - prophage 1
100
Serotype 13 (N273) - prophage 1
P2
Serotype 10 (D13039) - prophage 1
100
100
Serotype 5 (L20) - prophage 1
100
Serotype 7 (AP76) - prophage 1
Serotype 2 (4226) - prophage 1
P3
Serotype 1 (4074) - prophage 2
100
100
Serotype 9 (CVJ13261) - prophage 1
100
Serotype 11 (56153) - prophage 2
Serotype 7 (AP76) - prophage 2
P4
100
Serotype 4 (M62) - prophage 4
100
Serotype 4 (M62) - prophage 4
100
Serotype 8 (MV1022) - prophage 1
0,1

Figure 4: Phylogenetic relationships between sequences related to complete prophages found in *A. pleuropneumoniae*. The phylogenetic tree using the Neighbor-Joining method with 2000 bootstraps was generated by the MEGA 6 program after alignment by MAFFT and GBLOCKS. The scale is represented below, with 0.1 nucleotide substitutions per site.

Table 4: Identification of complete prophage sequences of *A. pleuropneumoniae*.

Strain/serotype	Prophage sequence identified	Organism/accession	ID %	Coverage	*E* value
4074/1	2	*Haemophilus ducreyi*/AE017143.1	97	85	0.0
4226/2	1	*Haemophilus ducreyi*/AE017143.1	98	95	0.0
M62/4	2	*Actinobacillus equuli*/CP007715.1	94	7	0.0
M62/4	4	*Mannheimia haemolytica*/KP137440.1	86	76	0.0
L20/5b	1	*Mannheimia* sp./CP006942.1	87	23	0.0
Femo/6	1	*Mannheimia haemolytica*/CP004753.1	89	25	0.0
AP76/7	1	*Mannheimia* sp./CP006942.1	89	26	0.0
	2	*Actinobacillus suis*/CP009159.1	89	86	0.0
MV518/8	1	*Mannheimia haemolytica*/CP004753.1	86	12	0.0
MV780/8	1	*Mannheimia haemolytica*/CP004753.1	86	14	0.0
MV1022/8	1	*Mannheimia haemolytica*/CP004753.1	86	12	0.0
MV5651/8	1	*Mannheimia haemolytica*/CP004753.1	86	15	0.0
CVJ13261/9	1	*Haemophilus ducreyi*/AE017143.1	97	89	0.0
D13039/10	1	*Mannheimia haemolytica*/CP004753.1	83	23	0.0
56153/11	2	*Haemophilus ducreyi*/AE017143.1	97	85	0.0
N273/13	1	*Mannheimia haemolytica*/CP004753.1	86	14	0.0

with the host, as well as having an important role in the differentiation of serotypes and mechanisms of virulence. As regards the differential portion, few differential C were affiliated with the COG categories. As this part of the genome has not been studied in a judicious way, we have a great network of sequences that codify proteins characterized as hypothetical, which are not categorized in the COG analysis. Among the sequences found in the differential portion, two sequences relating to the LysR family are present in the reference genome of *A. pleuropneumoniae* serotype 8. LysR is a family of transcriptional regulators that regulate a diverse set of genes, including those involved in virulence, metabolism, quorum sensing, and motility [44]. This regulator has also been related to processes of regulation of genes that code for urease in pathogenic bacteria [45]. In the differential portion of reference genome serotype 8, sequences encoding tetracycline, florfenicol, and sulfonamide resistance proteins were also found. In previous studies, the

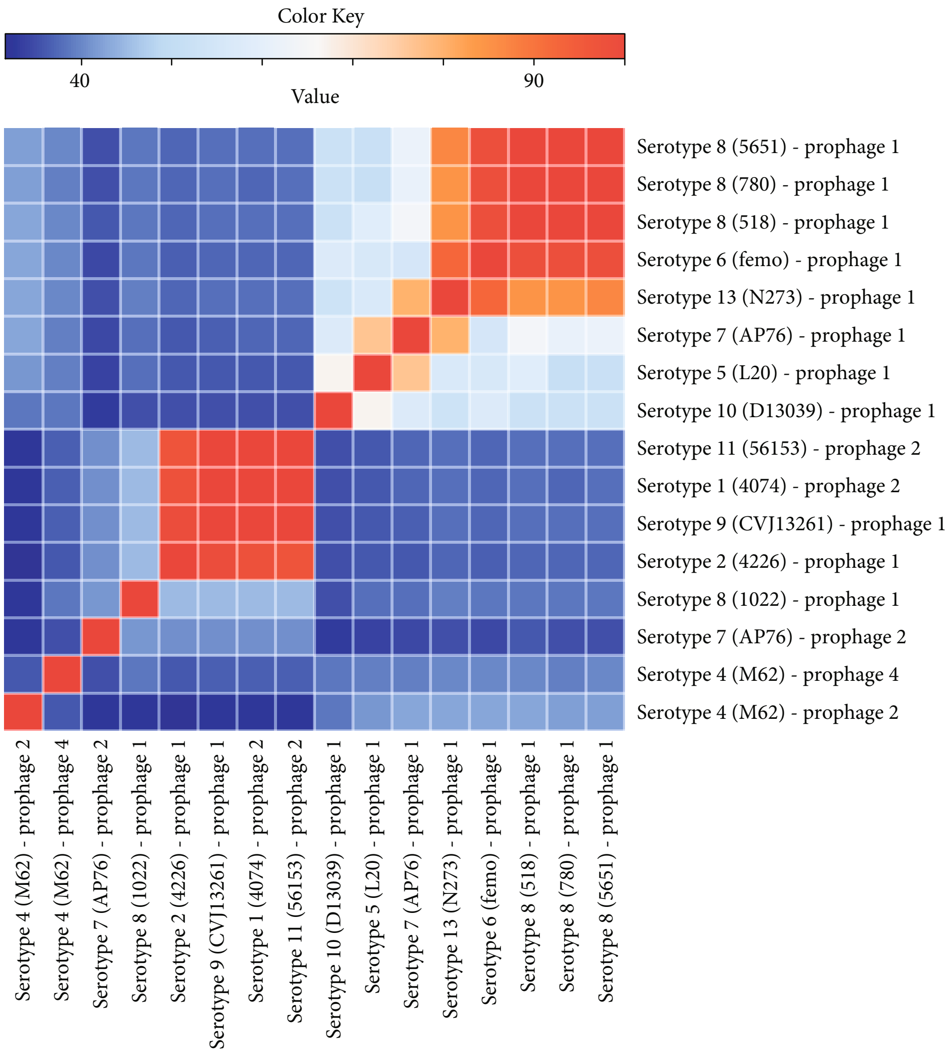


FIGURE 5: Heat map analysis from identity matrix generated by global alignment of 16 complete prophage genomes. The alignment was done using Clustal Omega, and the identity matrix generated was used to create the heat map by R software.

presence of plasmids in *A. pleuropneumoniae* and other members of the family Pasteurellaceae, conferring resistance to florfenicol, chloramphenicol, and tetracycline, has been characterized [13, 46, 47].

Alignment of the *A. pleuropneumoniae* genomes allowed for the determination of gain/loss and sequence rearrangements between serotypes. In the serotype 7 strains, there are rearrangements relating to the presence of insertion elements, indicating a process of integration of moving elements. Transposable elements have the ability to move within the genome, and their insertion close to the coding regions may alter gene expression [48]. Transposable elements if present in multiple copies can serve as sites for ectopic recombination events in the genome. Finally, these elements can incorporate additional genes and subsequently act as vectors for these genes. Any change, insertion, deletion, or rearrangement that may occur in a genome may alter the expression of adjacent genes and generate a substantial impact on gene expression and pathogenesis of the microorganism [49, 50]. The alignment of the genomes also demonstrated the existence of variations between the serotypes analyzed. The differences in alignments largely correspond to sequences relating to the HGT process such as prophages. Prophages are phages that integrate into the bacterial genome, in which they play an important role in genomic diversity and may be related to the acquisition of virulence factors for the host cell [51]. The acquisition of foreign sequences to the genome may be related to the fact that *A. pleuropneumoniae* is capable of performing natural transformation and has different levels of competence among serotypes and even among isolates of the same serotype [11, 12].

The results observed in Figure 5 showed consistent relation with the phylogenetic analysis. It was possible to see because the prophage sequences that showed considerable or high identity are present at the same or close groups in the phylogenetic tree.

In this study, 16 putative sequences related to the complete prophages were identified in the 19 genomes analyzed. Similarity analysis of the complete prophage sequences found in *A. pleuropneumoniae* against the GenBank database identified high similarity and coverage with sequences present in the genomes of *A. suis*, *M. haemolytica*, and *H. ducreyi*, which may be related to HGT among species belonging to the family Pasteurellaceae. *A. suis* is commonly found in swine as tonsil commensal, but in the presence of unknown stimuli, it may invade the bloodstream, causing septicemia and sequelae, such as meningitis and arthritis, and even lead to the death of the host [52]. On the other hand, *M. haemolytica* is frequently involved in respiratory diseases in cattle [53] while *H. ducreyi* is a bacterium that causes soft chancre, a sexually transmitted disease in humans, and which has pigs as a model for studying the disease [54, 55]. Of the 6 sequences with high similarity and coverage identified in GenBank, only prophage 4 of the M62 strain had significant alignment correspondence with the phage sequence already described in the literature. This prophage was found in *M. haemolytica* and named vB_MhM_3927AP2 by the authors, being a phage belonging to the *Myoviridae* family [56]. The remaining 13 prophages have low identity and coverage in biological databases, suggesting that they may be phages unique to this species or not reported yet.

In conclusion, the genome of *A. pleuropneumoniae* serotype 8 is conserved in relation to the other serotypes, being more related to serotypes 3 and 6, which justifies the problems of serotyping to distinguish these three serotypes. We detected strong evidence of DNA sequence acquisition and recombination in the genomes of the different isolates/serotypes, and these differences were attributed to the presence of mobile genetic material, mainly prophages. In this study, we have identified 16 complete prophages, 6 of which may have suffered HGT among species belonging to the family Pasteurellaceae. However, the other prophages seem to be exclusive of *A. pleuropneumoniae* and not yet reported in the literature. Thus, prophages seem to play a key role in the restructuring of genomes and in the emergence of new strains of this pathogen.

Data Availability

The data that support the results of this study are available in databases described in the manuscript and from the corresponding authors upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We are grateful to the Núcleo de Análise de Biomoléculas of the Universidade Federal de Viçosa for providing facilities for the implementation of the experiments and Tomas Gomes Reis Veloso for helping with additional analysis. This work was supported by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG; APQ-02732-15), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/PROEX), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Supplementary Materials

Supplementary Data Table 1: differential protein sequences found in *A. pleuropneumoniae* serotype 8. Supplementary Data Table 2: putative prophage sequences identified in the *A. pleuropneumoniae* genomes analyzed. *(Supplementary Materials)*

References

[1] OECD, *Meat consumption (indicator)*, 2019, October 2019.

[2] B. Clark, L. A. Panzone, G. B. Stewart et al., "Consumer attitudes towards production diseases in intensive production systems," *PLoS One*, vol. 14, no. 1, article e0210432, 2019.

[3] K. D. Stärk, "Epidemiological investigation of the influence of environmental risk factors on respiratory diseases in swine—a literature review," *The Veterinary Journal*, vol. 159, no. 1, pp. 37–56, 2000.

[4] S. Pohl, H. U. Bertschinger, W. Frederiksen, and W. Mannheim, "Transfer of *Haemophilus pleuropneumoniae* and the *Pasteurella haemolytica*-like organism causing porcine necrotic pleuropneumonia to the genus *Actinobacillus* (*Actinobacillus pleuropneumoniae* comb. nov.) on the basis of phenotypic and deoxyribonucleic acid relatedness," *International Journal of Systematic and Evolutionary Microbiology*, vol. 33, no. 3, pp. 510–514, 1983.

[5] R. Sárközi, L. Makrai, and L. Fodor, "Identification of a proposed new serovar of *Actinobacillus pleuropneumoniae*: serovar 16," *Acta Veterinaria Hungarica*, vol. 63, no. 4, pp. 444–450, 2015.

[6] J. T. Bossé, Y. Li, R. Fernandez Crespo et al., "Comparative sequence analysis of the capsular polysaccharide loci of *Actinobacillus pleuropneumoniae* serovars 1 -18, and development of two multiplex PCRs for comprehensive capsule typing," *Veterinary Microbiology*, vol. 220, pp. 83–89, 2018.

[7] J. T. Bossé, Y. Li, R. Sárközi et al., "Proposal of serovars 17 and 18 of *Actinobacillus pleuropneumoniae* based on serological and genotypic analysis," *Veterinary Microbiology*, vol. 217, pp. 1–6, 2018.

[8] K. Chiers, T. De Waele, F. Pasmans, R. Ducatelle, and F. Haesebrouck, "Virulence factors ofActinobacillus pleuropneumoniaeinvolved in colonization, persistence and induction of lesions in its porcine host," *Veterinary Research*, vol. 41, no. 5, p. 65, 2010.

[9] M. F. Pereira, C. C. Rossi, L. E. Seide, S. Martins Filho, C. M. Dolinski, and D. M. S. Bazzolli, "Antimicrobial resistance, biofilm formation and virulence reveal *Actinobacillus pleuropneumoniae* strains' pathogenicity complexity," *Research in Veterinary Science*, vol. 118, pp. 498–501, 2018.

[10] E. L. Sassu, J. T. Bossé, T. J. Tobias, M. Gottschalk, P. R. Langford, and I. Hennig-Pauka, "Update on *Actinobacillus pleuropneumoniae* — knowledge, gaps and challenges," *Transboundary and Emerging Diseases*, vol. 65, pp. 72–90, 2018.

[11] J. T. Bossé, S. Sinha, T. Schippers, J. S. Kroll, R. J. Redfield, and P. R. Langford, "Natural competence in strains *of Actinobacillus pleuropneumoniae*," *FEMS Microbiology Letters*, vol. 298, no. 1, pp. 124–130, 2009.

[12] J. T. Bossé, D. M. Soares-Bazzolli, Y. Li et al., "The generation of successive unmarked mutations and chromosomal insertion of heterologous genes in *Actinobacillus pleuropneumoniae*

using natural transformation," *PLoS One*, vol. 9, no. 11, article e111252, 2014.

[13] G. C. da Silva, C. C. Rossi, M. F. Santana, P. R. Langford, J. T. Bossé, and D. M. S. Bazzolli, "p518, a small *floR* plasmid from a South American isolate of *Actinobacillus pleuropneumoniae*," *Veterinary Microbiology*, vol. 204, pp. 129–132, 2017.

[14] A. Beceiro, M. Tomas, and G. Bou, "Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world?," *Clinical Microbiology Reviews*, vol. 26, no. 2, pp. 185–230, 2013.

[15] M. Schroeder, B. D. Brooks, and A. E. Brooks, "The complex relationship between virulence and antibiotic resistance," *Genes*, vol. 8, no. 1, p. 39, 2017.

[16] Z. Xu, X. Chen, L. Li et al., "Comparative genomic characterization of *Actinobacillus pleuropneumoniae*," *Journal of Bacteriology*, vol. 192, no. 21, pp. 5625–5636, 2010.

[17] C. O'Neill, S. C. P. Jones, J. T. Bosse et al., "Prevalence of *Actinobacillus pleuropneumoniae* serovars in England and Wales," *Veterinary Record*, vol. 167, no. 17, pp. 661-662, 2010.

[18] Y. Li, J. T. Bossé, S. M. Williamson et al., "*Actinobacillus pleuropneumoniae* serovar 8 predominates in England and Wales," *Veterinary Record*, vol. 179, pp. 276-277, 2016.

[19] M. Gottschalk and S. Lacouture, "*Actinobacillus* pleuropneumoniaeserotypes 3, 6, 8 and 15 isolated from diseased pigs in North America," *Veterinary Record*, vol. 174, no. 18, p. 452, 2014.

[20] C. C. Rossi, A. M. Vicente, W. V. Guimarães, E. F. Araújo, M. V. De Queiroz, and D. M. S. Bazzolli, "Face to face with *Actinobacillus pleuropneumoniae*: landscape of the distribution of clinical isolates in Southeastern Brazil," *African Journal of Microbiology Research*, vol. 7, no. 23, pp. 2916–2924, 2013.

[21] M. F. Pereira, C. C. Rossi, F. M. de Carvalho et al., "Draft genome sequences of six *Actinobacillus pleuropneumoniae* serotype 8 Brazilian clinical isolates: insight into new applications," *Genome Announcements*, vol. 3, no. 2, article e01585-14, 2015.

[22] J. T. Bossé, R. R. Chaudhuri, Y. Li et al., "Complete genome sequence of MIDG2331, a genetically tractable serovar 8 clinical isolate of *Actinobacillus pleuropneumoniae*," *Genome Announcements*, vol. 4, no. 1, article e01667-15, 2016.

[23] M. F. Pereira, C. C. Rossi, M. Vieira de Queiroz et al., "*Galleria mellonella* is an effective model to study *Actinobacillus pleuropneumoniae* infection," *Microbiology*, vol. 161, no. 2, pp. 387–400, 2015.

[24] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658-1659, 2006.

[25] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[26] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, "The COG database: a tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, no. 1, pp. 33–36, 2000.

[27] A. Marchler-Bauer, S. Lu, J. B. Anderson et al., "CDD: a conserved domain database for the functional annotation of proteins," *Nucleic Acids Research*, vol. 39, no. Database, pp. D225–D229, 2011.

[28] R. D. Finn, J. Tate, J. Mistry et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 36, pp. D281–D288, 2008.

[29] I. Letunic, R. R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork, "SMART 5: domains in the context of genomes and networks," *Nucleic Acids Research*, vol. 34, no. 90001, pp. D257–D260, 2006.

[30] C. H. Wu, R. Apweiler, A. Bairoch et al., "The universal protein resource (UniProt): an expanding universe of protein information," *Nucleic Acids Research*, vol. 34, no. 90001, pp. D187–D191, 2006.

[31] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[32] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: the European Molecular Biology Open Software Suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276-277, 2000.

[33] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, "Mauve: multiple alignment of conserved genomic sequence with rearrangements," *Genome Research*, vol. 14, no. 7, pp. 1394–1403, 2004.

[34] D. Arndt, J. R. Grant, A. Marcu et al., "PHASTER: a better, faster version of the PHAST phage search tool," *Nucleic Acids Research*, vol. 44, no. W1, pp. W16–W21, 2016.

[35] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[36] J. Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis," *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 540–552, 2000.

[37] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: molecular evolutionary genetics analysis version 6.0," *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[38] F. Sievers and D. G. Higgins, "Clustal Omega, accurate alignment of very large numbers of sequences," *Methods in Molecular Biology*, vol. 1079, pp. 105–116, 2014.

[39] S. Subashchandrabose, R. M. LeVeque, T. K. Wagner, R. N. Kirkwood, M. Kiupel, and M. H. Mulks, "Branched-chain amino acids are required for the survival and virulence of *Actinobacillus pleuropneumoniae* in swine," *Infection and Immunity*, vol. 77, no. 11, pp. 4925–4933, 2009.

[40] Z. Xu, H. Chen, and R. Zhou, "Genome-wide evidence for positive selection and recombination in *Actinobacillus pleuropneumoniae*," *BMC Evolutionary Biology*, vol. 11, no. 1, p. 203, 2011.

[41] V. Deslandes, M. Denicourt, C. Girard, J. Harel, J. H. E. Nash, and M. Jacques, "Transcriptional profiling of *Actinobacillus pleuropneumoniae* during the acute phase of a natural infection in pigs," *BMC Genomics*, vol. 11, no. 1, p. 98, 2010.

[42] M. M. Barnhart and M. R. Chapman, "Curli biogenesis and function," *Annual Review of Microbiology*, vol. 60, no. 1, pp. 131–147, 2006.

[43] S. A. Jones, M. Jorgensen, F. Z. Chowdhury et al., "Glycogen and maltose utilization by *Escherichia coli* O157: H7 in the mouse intestine," *Infection and Immunity*, vol. 76, no. 6, pp. 2531–2540, 2008.

[44] S. E. Maddocks and P. C. F. Oyston, "Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins," *Microbiology*, vol. 154, no. 12, pp. 3609–3623, 2008.

[45] J. T. Bossé, H. D. Gilmour, and J. I. MacInnes, "Novel genes affecting urease acivity in *Actinobacillus pleuropneumoniae*," *Journal of Bacteriology*, vol. 183, no. 4, pp. 1242–1247, 2001.

[46] J. T. Bossé, Y. Li, T. G. Atherton et al., "Characterisation of a mobilisable plasmid conferring florfenicol and chloramphenicol resistance in Actinobacillus pleuropneumoniae," *Veterinary Microbiology*, vol. 178, no. 3-4, pp. 279–282, 2015.

[47] Y. Li, G. C. da Silva, Y. Li et al., "Evidence of illegitimate recombination between two Pasteurellaceae plasmids resulting in a novel multi-resistance replicon, pM3362MDR, in *Actinobacillus pleuropneumoniae*," *Frontiers in Microbiology*, vol. 9, p. 2489, 2018.

[48] P. Siguier, E. Gourbeyre, and M. Chandler, "Bacterial insertion sequences: their genomic impact and diversity," *FEMS Microbiology Reviews*, vol. 38, no. 5, pp. 865–891, 2014.

[49] M. Chandler and J. Mahillon, "Insertion sequences revisited," in *Mobile DNA II*, pp. 305–366, American Society of Microbiology, 2002.

[50] T. Ooka, Y. Ogura, M. Asadulghani et al., "Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in *Escherichia coli* O157 genomes," *Genome Research*, vol. 19, no. 10, pp. 1809–1816, 2009.

[51] L. C. Fortier and O. Sekulovic, "Importance of prophages to evolution and virulence of bacterial pathogens," *Virulence*, vol. 4, no. 5, pp. 354–365, 2013.

[52] A. R. Bujold and J. I. MacInnes, "Identification of putative adhesins of *Actinobacillus suis* and their homologues in other members of the family Pasteurellaceae," *BMC Research Notes*, vol. 8, no. 1, p. 675, 2015.

[53] J. Gioia, X. Qin, H. Jiang et al., "The genome sequence of *Mannheimia haemolytica* A1: insights into virulence, natural competence, and Pasteurellaceae phylogeny," *Journal of Bacteriology*, vol. 188, no. 20, pp. 7257–7266, 2006.

[54] M. M. Hobbs, L. R. San Mateo, P. E. Orndorff, G. Almond, and T. H. Kawula, "Swine model of *Haemophilus ducreyi* infection," *Infection and Immunity*, vol. 63, no. 8, pp. 3094–3100, 1995.

[55] W. G. Fusco, N. R. Choudhary, P. A. Routh et al., "The *Haemophilus ducreyi* trimeric autotransporter adhesin DsrA protects against an experimental infection in the swine model of chancroid," *Vaccine*, vol. 32, no. 30, pp. 3752–3758, 2014.

[56] Y. D. Niu, S. R. Cook, J. Wang et al., "Comparative analysis of multiple inducible phages from *Mannheimia haemolytica*," *BMC Microbiology*, vol. 15, no. 1, p. 175, 2015.

Hindawi
International Journal of Genomics
Volume 2020, Article ID 8708305, 9 pages
https://doi.org/10.1155/2020/8708305

Research Article

Streptococcus halichoeri: Comparative Genomics of an Emerging Pathogen

Kirsi Aaltonen,[1,2] Ravi Kant,[1,2] Marjut Eklund,[3] Mirja Raunio-Saarnisto,[4] Lars Paulin,[5] Olli Vapalahti,[1,2,6] Thomas Grönthal,[3] Merja Rantala,[3] and Tarja Sironen[1,2]

[1]*Department of Veterinary Biosciences, Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland*
[2]*Department of Virology, Faculty of Medicine, University of Helsinki, Helsinki, Finland*
[3]*Department of Equine and Small Animal Medicine, Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland*
[4]*Finnish Food Authority, Seinäjoki, Finland*
[5]*Institute of Biotechnology, University of Helsinki, Helsinki, Finland*
[6]*HUSLAB, Hospital District of Helsinki and Uusimaa, Helsinki, Finland*

Correspondence should be addressed to Kirsi Aaltonen; kirsi.aaltonen@helsinki.fi

Received 22 November 2019; Accepted 14 January 2020; Published 18 February 2020

Academic Editor: João Paulo Gomes

Streptococcus halichoeri is an emerging pathogen with a variety of host species and zoonotic potential. It has been isolated from grey seals and other marine mammals as well as from human infections. Beginning in 2010, two concurrent epidemics were identified in Finland, in fur animals and domestic dogs, respectively. The fur animals suffered from a new disease fur animal epidemic necrotic pyoderma (FENP) and the dogs presented with ear infections with poor treatment response. S. halichoeri was isolated in both studies, albeit among other pathogens, indicating a possible role in the disease etiologies. The aim was to find a possible common origin of the fur animal and dog isolates and study the virulence factors to assess pathogenic potential. Isolates from seal, human, dogs, and fur animals were obtained for comparison. The whole genomes were sequenced from 20 different strains using the Illumina MiSeq platform and annotated using an automatic annotation pipeline RAST. The core and pangenomes were formed by comparing the genomes against each other in an all-against-all comparison. A phylogenetic tree was constructed using the genes of the core genome. Virulence factors were assessed using the Virulence Factor Database (VFDB) concentrating on the previously confirmed streptococcal factors. A core genome was formed which encompassed approximately half of the genes in Streptococcus halichoeri. The resulting core was nearly saturated and would not change significantly by adding more genomes. The remaining genes formed the pangenome which was highly variable and would still evolve after additional genomes. The results highlight the great adaptability of this bacterium possibly explaining the ease at which it switches hosts and environments. Virulence factors were also analyzed and were found primarily in the core genome. They represented many classes and functions, but the largest single category was adhesins which again supports the marine origin of this species.

1. Introduction

Streptococcus halichoeri was first described in 2004. It was isolated from grey seals (Halichoerus grypus) wherein it derives its name. The bacteria were found from wounds that had been inflicted by other seals, but evidence of systemic infection was also found [1]. *S. halichoeri* is one of only three Streptococcus species associated with marine mammals; the other two are *Streptococcus phocae* and *Streptococcus iniae*; the latter has also been found in farmed marine aquacultures and humans and has a significant pathogen potential [1]. *S. halichoeri* was found to be Gram-positive and belonging to the Lancefield group B. They are cocciforms that grow in pairs or short chains. In addition, they are nonhemolytic, facultatively anaerobic, and catalase-negative [1].

S. halichoeri has subsequently been found in humans in 2014 in a man with postoperative empyema [2] and in a diabetic man with infectious cellulitis [3]. In addition, several isolates have been obtained from human blood of both septicemic patients and others with unknown symptoms in the United States of America. The bacterium is hence considered an emerging pathogen that can cause serious disease in humans [4]. Subsequently, this bacterium has also been found in Steller sea lion (*Eumetopias jubatus*) [5] and the European badger (*Meles meles*) in which it was found to cause serious clinical symptoms [6].

During 2007, a novel and severe disease emerged in fur animals and was named "fur animal epidemic necrotic pyoderma" (FENP). The pathogen most strongly associated with the disease was *Arcanobacterium phocae* [7] previously only detected in marine mammals. The source of the original host shift of *A. phocae* is thought to have been infected seal meat used as feed [8]. During the investigation of the FENP outbreak, a Streptococcus species, previously undetected in fur animals, was also found in many samples, especially from mink. It was unclear whether this *Streptococcus* spp. together with a few other bacteria contributed to the disease [7]. The Streptococcus was later identified as *Streptococcus halichoeri*. While investigating the FENP epidemic, it has been found additionally in a quality control sample of herring from the Gulf of Finland used to prepare feed for the mink (unpublished observation). At the same time period as the outbreak of FENP, Finnish pet dogs were afflicted with an ear infection with poor treatment response and *S. halichoeri* was isolated from samples of the diseased dogs. Later on, this bacterium was also isolated from skin infections of dogs. These isolates were characterized by Eklund and colleagues [9] using conventional bacteriological, biochemical, and sequencing methods.

Our study focuses on further characterizing this emerging pathogen through sequencing the whole genomes of 20 isolates. This approach permits analysis of the core genome and virulence factors of *S. halichoeri*, allows more reliable phylogenetic analyses and attempts to trace the direction and frequency of previous bacterial introductions and (cross-species) transmissions.

2. Materials and Methods

2.1. Bacterial Isolates, Growth Conditions, and DNA Extraction. The bacterial strains ($n = 20$) included in this project have been characterized by Eklund et al., and the selection of strains was based on clustering in PFGE [9]. Ten isolates were from canine infections, five from mink, two from Finnraccoon, and one from a blue fox. The canine isolates are all from diagnostic samples of superficial or deep pus from cases of otitis or dermatitis. The clinical significance is unclear as all of these findings were of mixed culture most often together with *Staphylococcus pseudintermedius.* The eight fur animal isolates are all from severe dermatitis lesions. Findings from Nordgren et al. [10] suggest a possible role for *S. halichoeri* in the pathogenesis of FENP. Also, two reference strains were included, one from a seal (CCUG 48324) [1] and one from human isolate (CCUG 67100) [4]. The origin of the strains and their characteristics are listed in Table 1.

The bacteria were grown on blood agar plates with 4% defibrinated sheep blood, overnight. A single colony was then inoculated into 2 ml of Super Broth medium. They were grown at +37°C with mild shaking for 24 hours and harvested by centrifugation for 5 minutes at 4,500 g. The cells were stored in -20°C awaiting extraction. The DNA was extracted using the Epicentre by Lucigen MasterPure Gram Positive DNA Purification Kit (Lucigen Corp., Wisconsin, USA) according to the kit instructions. An overnight lysozyme treatment, stated optional in the kit, was used to ensure bacterial lysis.

2.2. Genome Sequencing and Annotation. Genomes of the 20 *S. halichoeri* isolates were sequenced at the Institute of Biotechnology (University of Helsinki, Finland) using next-generation sequencing platforms. Genomic DNA (0.5 mg) was sheared using a Bioruptor NGS Sonicator (Diagenode) to approximately 600 bp fragments. The fragments were polished, A-tailed, and ligated to a TruSeq truncated adapter. Purification of the ligation reaction was done using AMPure XP beads (Agencourt, Beckman Coulter). PCR of the libraries were done using Phusion Hot Start II DNA Polymerase (Thermo Fisher) and index P7 primers and full-length P5 adapter primers. The reactions were pooled and purified with AMPure XP beads. Size selection of the pool was done according to Lundin et al. [11]. The obtained library pool was paired-end sequenced on a MISeq Sequencer using the v3 600 cycle kit (Illumina).

Genomes of the 20 newly sequenced *S. halichoeri* strains were deposited in GenBank under the accession numbers listed in Table 1. The annotation was performed using the assembled DNA sequences of the 20 new draft genomes from these isolates. The genomes were run through an automatic annotation pipeline RAST (Rapid Annotation using Subsystem Technology) [12], followed by manual curation in few cases.

2.3. Orthologous Gene Prediction and Genome Sequence Comparison. Identification of orthologous genes for 20 *S. halichoeri* genomes was performed by an all-against-all comparison of the genes of all genomes using blastp [13] with the standard scoring matrix BLOSUM62 and an initial *E*-value cut-off of $1e^{-05}$. The bit score of every blast hit was set into proportion to the best bit score possible, the bit score of a hit of the query gene against itself. The outcome for this was a score ratio value (SRV) between 0 and 100 that reflected the quality of the hit much better than the raw blast bit score [14].

Two genes were acknowledged orthologous if a reciprocal best blast hit existed among them, and both hits had an SRV > 32. The SRV threshold is computed from distribution of blast hits between analyzed sequences as described in the supplement of Blom et al. [15]. Based on this orthology principle, the core genome was calculated as the set of genes that had orthologous genes in all other analyzed strains.

The pangenome was estimated as the set of all unique genes of a set of genomes. All genes of one reference genome were considered the basic set for the calculation. Afterwards, the genes of a second genome were matched with this set, and all genes in the second genome that had no orthologous gene in the starting gene set were added to this set. This process

Table 1: A general overview of thirteen *Streptococcus halichoeri* genomes.

Strain	Host source	Infected organ	Geographic region	Year	Genome accession no	Status	Coverage	Contigs	Size (Mbps)	G+C (%)	ORFs	Proteins	Ref.
P154	Dog	Ear	Finland, Uusimaa	2010	WLZC00000000	Draft	112x	42	2.06	41.3	2025	1978	This study
P376	Dog	Ear	Finland, Uusimaa	2012	WLZD00000000	Draft	90x	47	2.07	41.2	2036	1991	This study
P380	Dog	Ear	Finland, Uusimaa	2012	WLZE00000000	Draft	109x	70	2.00	41.4	1991	1953	This study
P399	Dog	Skin	Finland, Uusimaa	2012	WLZF00000000	Draft	204x	244	2.26	41.5	2198	2150	This study
P408	Dog	Ear	Finland, Uusimaa	2012	WLZG00000000	Draft	120x	33	1.91	41.7	1892	1847	This study
P791	Dog	Ear	Finland, Northern Savonia	2012	WLZH00000000	Draft	130x	50	1.93	41.6	1889	1846	This study
P993	Dog	Skin	Finland, Uusimaa	2015	WLZI00000000	Draft	90x	42	2.05	41.4	2019	1973	This study
P994	Dog	Skin	Finland, Uusimaa	2015	WLZJ00000000	Draft	115x	40	2.04	41.4	2006	1961	This study
P1033	Dog	Ear	Finland, Uusimaa	2015	WLZK00000000	Draft	117x	32	2.02	41.1	1992	1956	This study
P1063	Dog	Skin	Finland, Uusimaa	2015	WLZL00000000	Draft	120x	43	1.95	41.7	1919	1884	This study
S157 B-4*	Mink	Eye	Finland, Ostrobothnia	2010	WLZM00000000	Draft	90x	62	2.03	41.3	1970	1937	This study
S171 B-3*	Finnraccoon	Skin	Finland, Ostrobothnia	2010	WLZN00000000	Draft	79x	76	2.02	41.5	1971	1934	This study
S173 B-1*	Blue fox	Eye	Finland, Ostrobothnia	2010	WLZO00000000	Draft	103x	54	1.93	41.7	1882	1859	This study
S185 B-2*	Finnraccoon	Paw	Finland, Ostrobothnia	2011	WLZP00000000	Draft	89x	59	1.94	41.7	1891	1858	This study
S186 B-6* c	Mink	Eye	Finland, Ostrobothnia	2011	WLZQ00000000	Draft	75x	44	1.98	41.5	1924	1898	This study
S212 B-7*	Mink	Skin	Finland, Ostrobothnia	2012	WLZR00000000	Draft	61x	34	2.00	41.5	1976	1937	This study
S244 B-5*	Mink	Skin	Finland, Ostrobothnia	2013	WLZS00000000	Draft	130x	56	1.94	41.7	1866	1849	This study
S258 B-8*	Mink	Paw	Finland, Ostrobothnia	2015	WLZT00000000	Draft	75x	41	2.00	41.4	1988	1947	This study
CCUG48324	Seal	Lung	UK, Scotland, Inverness	2003	WLZU00000000	Draft	280x	54	1.89	41.6	1873	1829	This study
CCUG67100	Human	Blood	USA, NC, Raleigh	2015	WLZV00000000	Draft	94x	34	1.91	41.7	1877	1849	This study

*B-1-8 are the strain markers used by Eklund et al. [9].

was iteratively repeated for all genomes of the compared set, leading to the pangenome. The circular plot comparing 20 genomes was generated with BioCircos [16].

2.4. Phylogenetic Construction. The phylogenetic tree was calculated using a somewhat modified version of the pipeline proposed by Zbodnov and Bork [17]. Alignments of each core gene set are compiled using MUSCLE [18], the numerous resulting multiple alignments were concatenated, and poorly aligned positions were removed using GBLOCKS [19]. The trimmed multiple alignment was used to create a phylogenetic tree using the neighbour-joining operation of PHYLIP [20].

2.5. Identification of the Putative Virulence Factors in the Genomes. The Virulence Factor Database (VFDB) [21] as well as known virulence factors of Streptococci was used as guidelines when choosing the putative virulence factors to be sought. The Virulence Factor Database is based on experimentally validated or strongly suspected bacterial virulence factors from multiple bacterial species. There are listed known virulence factors from twenty different species of streptococci encompassing 56 different strains with 75 recognized virulence factors. The closest genetic relative to *S. halichoeri* in this database is *Streptococcus agalactiae* which also correlated with the identified virulence factors. Two different approaches were used to analyze these factors, utilizing either the core genome or the annotated pangenome. This enabled recognition of the virulence factors with most importance as well as the more dispensable ones.

3. Results and Discussion

3.1. General Features of the Genomes of 20 Streptococcus halichoeri Isolates. In this study, we have constructed sequences of 20 *S. halichoeri* isolates via high-throughput sequencing. The assembled draft sequences were initially annotated using an automated pipeline for gene identification and then afterwards improved by additional manual curation. Plasmid DNA sequences were excluded from this annotation process. A list of the annotated genes predicted for the 20 newly sequenced genomes is given as supporting information (Table S1), with each genomic sequence deposited into GenBank (Table 1). The general features of the 20 new *S. halichoeri* genomes included and analyzed in this study are presented in Table 1. Here, genomes were characterized from dog [10], mink [5], Finnraccoon [2], human [1], blue fox [1], and seal [1] hosts. Till date, there are no other *S. halichoeri* genome sequences present in the NCBI RefSeq database making this study the first genomic study of *S. halichoeri*.

Despite the fact that all 20 genomes are draft assemblies, they still represent good quality sequence data for performing genomic comparisons (Figure S1). The average coverage of the genome sequencing ranges widely from 61-fold (S212) to 280-fold (CCUG48324). Furthermore, the number of contigs in the assembled genomes was between 32 and 244 (P1033 and P399, respectively). The genome size was ranging between 1.89 (CCUG48324) and 2.26 (P399) Mbps. The total GC content varies only slightly and ranged between 41.2 and 41.8%. The numbers of predicted protein-encoding open-reading frames (ORFs) in the 20 isolates varied from 1,873 (CCUG48324) to 2,198 (P399) suggesting reasonable diversity in the species of *S. halichoeri*.

3.2. Phylogeny. Pangenomic studies are usually performed without referencing the individual ecological niches the isolates are derived from. However, the host source of the bacterial strains should be considered an essential parameter for the pangenome to be deciphered correctly. Reconstructing a core genome-based phylogenic tree from our 20 *S. halichoeri* strains offers additional understanding between the incidental phyletic associations and of any common origins by presenting possible correlations. Comparison of the 20 genomes illustrated in Figure S1 shows the wide strain diversity of *S. halichoeri*.

A phylogenic tree of 20 *S. halichoeri* strains was constructed using a multiple alignment of 1,456 core proteins as illustrated in Figure 1. Most *S. halichoeri* genomes grouped together into individual clades according to their host-derived origins. Interestingly, there were two separate clades; one was roughly dominated by dogs while the other clade comprised mostly of mink, blue foxes, and Finnraccoons. Similar results were seen by Eklund et al. when partial sequences were compared [9]. The clustering of the dog strains together seems expected, as sharing the same host would likely reflect a same origin for these strains while also niche adaptation could play a role. The grouping of mink-, Finnraccoon-, and blue fox-associated strains was also expected, indicating a common origin of *S. halichoeri* strains in these animals. The human strain (CCUG67100) clustered closely together with three of the dog strains (P399, P408, and P1033) indicating a potential zoonotic connection. It is noteworthy however that the human strain is from the United States of America and the dog strains are all from Finland. Another interesting finding was that the seal strain (CCUG48324) was somewhat different from all the other *S. halichoeri* strains. Most of the mink-associated strains were scattered in different clades except for two strains (S258 and S212), which formed a separate, distinct clade suggesting that there has been more than one introduction of this bacteria into the fur animal community. Interestingly, the dog strain P791 did not cluster with any of the clades and when investigating this further, it was found that this dog came from a different geographic area (eastern Finland) than all the other dogs which were all from the Uusimaa region (south) of Finland. There is also a single mink strain which groups together with the dog strains. Dogs and mink do have some contact within farms so direct transmission is not impossible although likely rare.

3.3. Pan, Core, and Accessory Genomes of 20 S. halichoeri Strains. We used the genome sequences of 20 *S. halichoeri* isolates to construct the pangenome. The numerous genetic loci from the pangenome essential and necessary for the survival of the bacteria is the core genome of a particular species. These genes are largely involved with different metabolic, catabolic, transport activities and degradation of nucleic

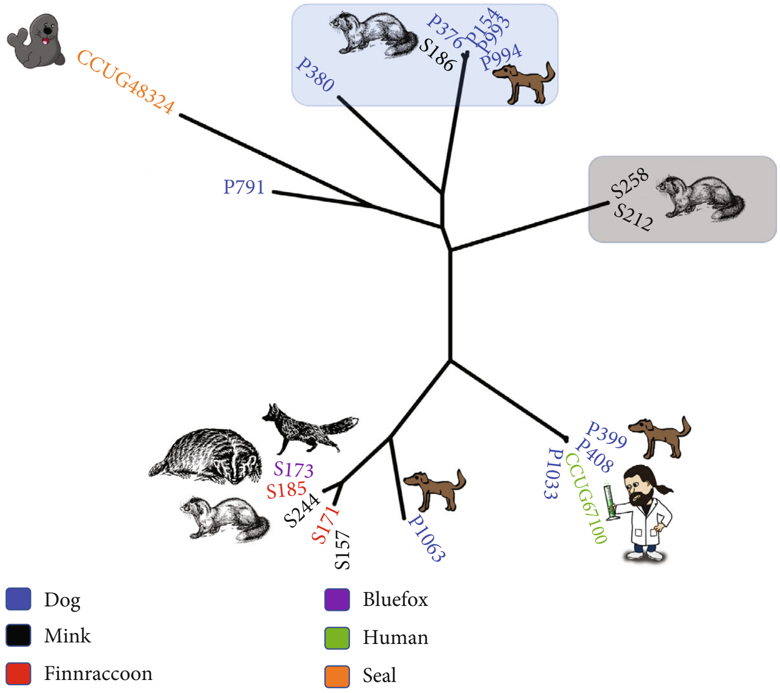


Figure 1: Phylogenetic tree based on core genome (1,456 genes).

acids, ribosomes, and proteins essential for basic housekeeping functions [22, 23]. As a group, these 20 genomes yielded a pangenome of 3,433 genes (Table S1), of which only 42% (1,456 genes) formed the core genome (Table S2), revealing a slightly high interspecies diversity (Figure 2) [24–28]. When the number of genes in pangenome was plotted against the number of *S. halichoeri* genomes using Heap's Law calculation (Tettelin et al., [22, 23]), the obtained α-value of 0.81 indicated that the pangenome is still open (Figure 3). In a detailed examination of the pangenome development data, it was noticed that the pangenome curve starts to level at approximately 3,000 genes. Genomes added after 13th genome contribute only few genes to the pangenome implying pangenome of *S. halichoeri* is eventually proceeding to a closed status. Addition of a few more strains would eventually close the pangenome representing the entire genetic repertoire of *S. halichoeri*. Similar trends were observed with the core genome development plot (Figure 3) with fewer gene reduction from the core genome after the 8th genome. The comparatively low number of core genes (1,456) in *S. halichoeri* species indicates a broad genome structure, suggesting a large accessory genome. Even with likely possibility of moderately growing pangenome, *S. halichoeri* are undoubtedly a dynamically evolving species with multiple habitats.

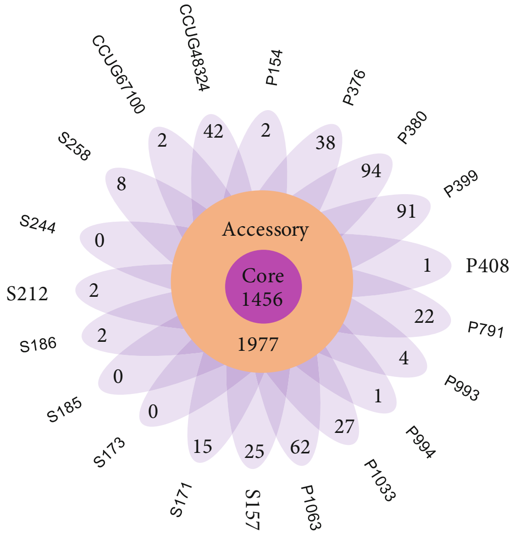


Figure 2: *Streptococcus halichoeri* pangenome (3,433 genes) representing the individual strain-specific genes.

The part of *S. halichoeri* pangenome which is not included in the core genome is generally referred to as an accessory genome. These genes are apparently not essential but can provide reasonable advantages to different strains of this species. Accessory genome basically outlines the diversity of the *S. halichoeri* species [22, 23]. The 20 *S. halichoeri* strains encompass an accessory genome of 1,977 genes with 438 genes (Figure 2) belonging to strain-specific genes that can only be found in one strain of *S. halichoeri* but absent in all other strains (also called unique genes). The numbers of unique genes per each genome are indicated in Figure 2. Two of the *S. halichoeri* strains (P380, P399) with the highest numbers of contigs (70, 244) could have many partial/split genes which can inflate the count for unique genes in these strains. Interestingly, majority of the dispensable genes of *S. halichoeri* were annotated as hypothetical proteins or proteins with an unknown function (Table S1), and as most variations exist with unknown and uncharacterized

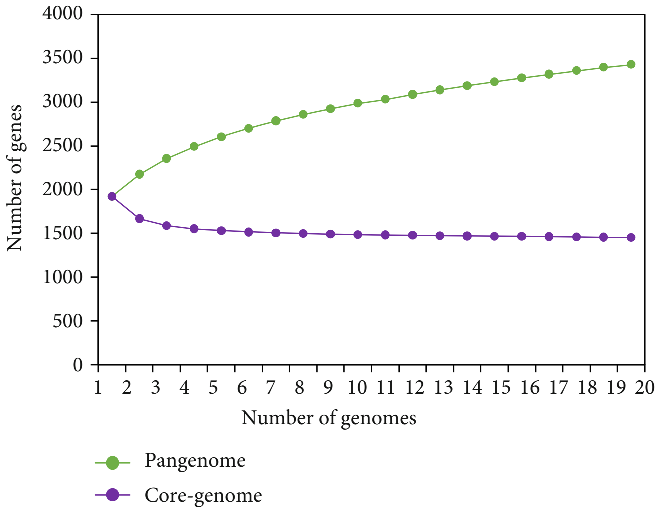


Figure 3: Pangenome development plot of *S. halichoeri*.

functionalities, it is very problematic to associate any type of adaptive role or benefits for the *S. halichoeri* strains. Nevertheless, in a small number of strains, their unique genes were annotated with a selection of predicted functions from transport and metabolism to phage-related proteins, transposases, and mobile elements.

3.4. Virulence. The bacterium is able to jump species and adapt with ease to new environments. This in addition to pathogenic potential could be explained through virulence factors. We were able to identify 19 different streptococcal virulence factors in the core genome. These belong to 5 categories based on their mechanism. These factors are listed in Table 2. The category most represented in the core virulence factors is adhesion-associated products followed by proteases and toxins. These would benefit bacteria inhabiting skin and mucosa and also contribute to necrotizing infection. We identified a non-streptococci-specific factor, multidrug resistance gene which helps bacteria fight host-derived antibacterials and hormones as well as some antibiotics [29]. We further identified a core protein capsule biosynthesis protein capA which is a suspected virulence factor and enables the bacteria to survive in high salt concentrations [30]. This protein is especially interesting as it would support the hypothesis that this bacterium is of marine origin. The core genome also carries two separate antigen A genes which are used in diagnostics and vaccines in other bacteria and may also have immunomodulatory or evasive functions [31, 32]. Interestingly, the strains also had hemolysin and catalase genes despite the original isolates testing catalase negative. When studying the dog and fur animal isolates, Eklund et al. found varying degrees of catalase activity [9] and the genomic findings support this. The expression of the gene may be dependent on environmental factors and warrants further study.

Table 2: Virulence factors found in the core genome of *Streptococcus halichoeri*.

Group	Virulence factor
Adherence	Putative choline binding protein
	Fibronectin-binding protein
	Fibronectin/fibrinogen-binding protein
	Laminin-binding surface protein
	M-like protein
	Sortase A, LPXTG specific
	Collagen-like surface protein
	Streptococcal lipoprotein rotamase A
Enzyme	Enolase
	Streptodornase D
Manganese uptake	Pneumococcal vaccine antigen A homolog
Protease	C3-degrading proteinase
	Immunoglobulin G-endopeptidase (IdeS)/Mac/secreted immunoglobulin-binding protein (Sib38)
	Serine protease, DegP/HtrA, do-like
	Streptococcal cysteine protease (streptopain)/ streptococcal pyrogenic exotoxin B (SpeB)
	Streptokinase
Toxin	C3 family ADP-ribosyltransferase
	CAMP factor
	Hemolysin III
Immune evasion	Multidrug resistance protein*
Capsid	Capsule biosynthesis protein capA*

*These proteins are additional to the known streptococcal virulence factors.

Evidence of many potential mobile genetic elements (MGE) in the genomes was also noted. Several operon-like clusters related to phages were found throughout the genomes. This suggests phages act as transporters of important genes between bacterial hosts. The presence of plasmids, phages, and integrative conjugative element (ICE) indicates the possibility of lateral gene transfer. Other streptococcal species have been found to have similar attributes, most

Table 3: Virulence factors found in the accessory genome.

Group	Virulence factor
Adherence	Agglutinin receptor Choline-binding protein A Antiphagocytic M protein
Enzyme	Phage hyaluronidase Mitogenic factor 2
Protease	C5a peptidase
Superantigen	Streptococcal pyrogenic exotoxin A (SpeA) Streptococcal pyrogenic exotoxin K (SpeK)

notably *S. canis* which is the closest genetic relative of *Streptococcus halichoeri* [33].

The accessory genome had a further eight virulence factors. These did not correlate between the different host species of the isolates except the agglutinin receptor, also adherence enabler, which was absent only in the human strain but present in all the others. These virulence factors are listed in Table 3. Streptococcal pyrogenic exotoxin A (SpeA) was only present in one strain P380. Equally interesting is the absence of pili-associated genes. Pili are common in pathogenic streptococci and assist with adherence. *S. halichoeri* had multiple adherence genes but not this very common one. Pili also enable motility and are especially found in intestinal bacteria which may suggest another niche for *S. halichoeri.* Earlier results by Eklund et al. showed antibiotic resistance to erythromycin, clindamycin, and tetracycline in selected dog strains and tetracycline resistance in two mink strains. The core genome had only one antibiotic resistance gene patB, but the pangenome had further three ermB, tetO, and inuC. The ermB is often found in streptococci and is known to code for erythromycin and clindamycin resistance. Tetracycline resistance is more commonly coded by the tetM gene, but tetO is also found. All of these are usually in transposons so would possibly be found in the missing parts of the current genomes.

Interestingly, the majority of these well-established streptococcal virulence factors were found in the core genome despite it representing less than half of the genes in a given isolate. This highlights the importance of these genes to the survival of the species. More virulence factors and putative factors can be found in the genome especially as we learn more about the hypothetical proteins within.

4. Conclusions

We find that *S. halichoeri* is a highly variable species with several virulence factors which suggest potential for significant pathogenicity. This is supported by the relatively severe human cases as well as the data on the seal and badger isolates. The many varieties of tissue, host selection, and geographic diversity suggest a diverse niche wherein the potential for lateral gene transfer gives way for a rapid adaptation to new growth environments. The core genome is saturated, but the fact that the already large dispensable genome is still somewhat incomplete suggests we have yet to see the full potential of this bacterium's adaptability and host species flexibility.

We found very little host species-specific markers in the genomes but rather loose clustering according to species as though adaptation is still incomplete. This suggests the host switches into dogs, humans, and fur animals which were rather recent and ongoing, possibly coinciding with the beginning of the FENP epidemic. Some further analysis of the virulence factors is called for as there are many more not directly associated with streptococci but which could play a critical role in the pathogenesis of this bacterium. Expression studies should also be made to verify the role and activity of these genes.

Genetic factors such as great numbers of adhesins and salt tolerance proteins as well as the fact that the first isolates were from marine mammals suggest this bacterium may have marine origins. This would also correlate well with the known history of FENP pathogen and that *A. phocae* also associates with seals. In the FENP study, *S. halichoeri* was mainly found in mink which are fed with locally caught fish much more than Finnraccoons and foxes. This together with our finding of *S. halichoeri* from a batch of herring would suggest a possible source of transmission. The Finnish dogs are also fed with raw fish occasionally, but we do not know how often or in what quantities so it is difficult to assess the level of risk and potential exposure. Other possible routes of infection may occur between the animals, both dogs and fur animals, especially in crowded farm environments, between a dog and an owner while other routes may not have been yet found. Fish handling by humans alone has been connected with infections by other marine mammal-associated pathogens. On the other hand, the recent isolation of *S. halichoeri* in a clinical sample from a badger with no contact to marine environment [6] suggests the ecological niche of this bacterium may already be much wider and possibly underdiagnosed. This is supported by at least one human finding wherein no contact to marine environment could be shown. The presence of this plausibly pathogenic bacterium in domestic dogs suggests further an opportunity for more zoonotic transfers making it important to alert diagnostic laboratories in both human and veterinary medicine.

Current data is not enough to confirm or rule out any suggested transmission or entry routes and this requires further studies. The pathogenic potential of this bacterium should also be studied more. Altogether, this study shows the great adaptability of *Streptococcus halichoeri* and we are yet to see the full potential of this emerging pathogen.

Data Availability

The genomes constructed in this study have been submitted into the NCBI genome database and can be accessed with the codes indicated in Table 1.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Kirsi Aaltonen and Ravi Kant equally contributed to this study.

Acknowledgments

Päivi Laamanen is acknowledged for library construction and sequencing. This study was supported by the Reino Rinne Foundation and the Finnish Veterinary Research Foundation.

Supplementary Materials

Supplementary 1. Figure S1: a circular plot of the different strains of Streptococcus halichoeri depicting the core genes, accessory genome, and general characteristics.

Supplementary 2. Table S1: the predicted annotations of Streptococcus halichoeri pangenome.

Supplementary 3. Table S2: the predicted annotations of Streptococcus halichoeri core genome.

References

[1] P. A. Lawson, G. Foster, E. Falsen, N. Davison, and M. D. Collins, "Streptococcus halichoeri sp. nov., isolated from grey seals (*Halichoerus grypus*)," *International Journal of Systematic and Evolutionary Microbiology*, vol. 54, no. 5, pp. 1753–1756, 2004.

[2] R. M. Foo and D. Chan, "A fishy tale: a man with empyema caused by *Streptococcus halichoeri*," *Journal of Clinical Microbiology*, vol. 52, no. 2, pp. 681-682, 2014.

[3] P. Giudice, C. Plainvert, T. Hubiche, A. Tazi, A. Fribourg, and C. Poyart, "Infectious cellulitis caused by Streptococcus halichoeri," *Acta Dermato-Venereologica*, vol. 98, no. 3, pp. 378-379, 2018.

[4] P. L. Shewmaker, A. M. Whitney, and B. W. Humrighouse, "Phenotypic, genotypic, and antimicrobial characteristics of *Streptococcus halichoeri* isolates from humans, proposal to rename *Streptococcus halichoeri* as *Streptococcus halichoeri* subsp. *halichoeri*, and description of *Streptococcus halichoeri* subsp. *hominis* subsp. nov., a bacterium associated with human clinical infections," *Journal of Clinical Microbiology*, vol. 54, no. 3, pp. 739–744, 2016.

[5] K. Lee, J. Y. Kim, S. C. Jung, H. S. Lee, M. Her, and C. Chae, "First isolation of *Streptococcus halichoeri* and *Streptococcus phocae* from a Steller sea lion (*Eumetopias jubatus*) in South Korea," *Journal of Wildlife Diseases*, vol. 52, no. 1, pp. 183–185, 2016.

[6] B. Moreno, R. Bolea, M. Morales, I. Martín-Burriel, C. González, and J. J. Badiola, "Isolation and Phylogenetic Characterization of *Streptococcus halichoeri* from a European Badger (*Meles meles*) with Pyogranulomatous Pleuropneumonia," *Journal of Comparative Pathology*, vol. 152, no. 2-3, pp. 269–273, 2015.

[7] H. Nordgren, K. Aaltonen, T. Sironen et al., "Characterization of a new epidemic necrotic pyoderma in fur animals and its association with *Arcanobacterium phocae* infection," *PLoS One*, vol. 9, no. 10, article e110210, 2014.

[8] C. Bröjer, *Pododermatitis in Farmed Mink in Canada, [M.S. thesis]*, The University of Guelph, Guelph, Canada, 2000.

[9] M. Eklund, T. Sironen, M. Raunio-Saarnisto et al., "Comparison of Streptococcus halichoeri from canine and fur animal infections: biochemical patterns, molecular characteristics and genetic relatedness," *Acta Veterinaria Scandinavica*, 2020.

[10] H. Nordgren, K. Aaltonen, M. Raunio-Saarnisto, A. Sukura, O. Vapalahti, and T. Sironen, "Experimental infection of mink enforces the role of *Arcanobacterium phocae* as causative agent of fur animal epidemic necrotic pyoderma (FENP)," *PLoS One*, vol. 11, no. 12, article e0168129, 2016.

[11] S. Lundin, H. Stranneheim, E. Pettersson, D. Klevebring, and J. Lundeberg, "Increased throughput by parallelization of library preparation for massive sequencing," *PLoS One*, vol. 5, no. 4, article e10029, 2010.

[12] R. K. Aziz, D. Bartels, A. A. Best et al., "The RAST server: rapid annotations using subsystems technology," *BMC Genomics*, vol. 9, no. 1, p. 75, 2008.

[13] S. F. Altschul, T. L. Madden, A. A. Schaffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[14] E. Lerat, V. Daubin, and N. A. Moran, "From gene trees to organismal phylogeny in prokaryotes: the case of the γ-proteobacteria," *PLOS Biology*, vol. 1, no. 1, article e19, 2003.

[15] J. Blom, S. P. Albaum, D. Doppmeier et al., "EDGAR: a software framework for the comparative analysis of prokaryotic genomes," *BMC Bioinformatics*, vol. 10, no. 1, p. 154, 2009.

[16] Y. Cui, X. Chen, H. Luo et al., "BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications," *Bioinformatics*, vol. 32, no. 11, pp. 1740–1742, 2016.

[17] E. M. Zdobnov and P. Bork, "Quantification of insect genome divergence," *Trends in Genetics*, vol. 23, no. 1, pp. 16–20, 2007.

[18] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[19] G. Talavera and J. Castresana, "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments," *Systematic Biology*, vol. 56, no. 4, pp. 564–577, 2007.

[20] J. Felsenstein, *PHYLIP—Phylogeny Inference Package*, Version 3.6 Seattle: Department of Genome Sciences, University of Washington, 2005.

[21] L. Chen, D. Zheng, B. Liu, J. Yang, and Q. Jin, "VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on," *Nucleic Acids Research*, vol. 44, no. D1, pp. D694–D697, 2016.

[22] H. Tettelin, V. Masignani, M. J. Cieslewicz et al., "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 13950–13955, 2005.

[23] H. Tettelin, D. Riley, C. Cattuto, and D. Medini, "Comparative genomics: the bacterial pan-genome," *Current Opinion in Microbiology*, vol. 11, no. 5, pp. 472–477, 2008.

[24] R. Kant, J. Blom, A. Palva, R. J. Siezen, and W. M. de Vos, "Comparative genomics of *Lactobacillus*," *Microbial Biotechnology*, vol. 4, no. 3, pp. 323–332, 2011.

[25] R. Kant, J. Rintahaka, X. Yu et al., "A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in *Lactobacillus rhamnosus*," *PLoS One*, vol. 9, no. 7, article e102762, 2014.

[26] R. Kant, A. Palva, and I. von Ossowski, "An *in silico* pan-genomic probe for the molecular traits behind *Lactobacillus ruminis* gut autochthony," *PLoS One*, vol. 12, no. 4, article e0175541, 2017.

[27] S. Åvall-Jääskeläinen, S. Taponen, R. Kant et al., "Comparative genome analysis of 24 bovine-associated *Staphylococcus* isolates with special focus on the putative virulence genes," *PeerJ*, vol. 6, article e4560, 2018.

[28] R. Kant, *Genomic insights about the Lactobacillus genus, [Ph.D. thesis]*, University of Helsinki, 2018.

[29] G.-T. Ho, F. M. Moodie, and J. Satsangi, "Multidrug resistance 1 gene (P-glycoprotein 170): an important determinant in gastrointestinal disease?," *Gut*, vol. 52, no. 5, pp. 759–766, 2003.

[30] H. Asakura, M. Yamasaki, S. Yamamoto, and S. Igimi, "Deletion of *peb4* gene impairs cell adhesion and biofilm formation in *Campylobacter jejuni*," *FEMS Microbiology Letters*, vol. 275, no. 2, pp. 278–285, 2007.

[31] H. B. Bilgic, T. Karagenc, S. Bakırcı et al., "Identification and analysis of immunodominant antigens for ELISA-based detection of *Theileria annulata*," *PLoS One*, vol. 11, no. 6, article e0156645, 2016.

[32] M. J. Crain, W. D. Waltman 2nd, J. S. Turner et al., "Pneumococcal surface protein A (PspA) is serologically highly variable and is expressed by all clinically important capsular serotypes of Streptococcus pneumoniae," *Infection and Immunity*, vol. 58, no. 10, pp. 3293–3299, 1990.

[33] V. P. Richards, R. N. Zadoks, P. D. Pavinski Bitar et al., "Genome characterization and population genetic structure of the zoonotic pathogen, *Streptococcus canis*," *BMC Microbiology*, vol. 12, no. 1, p. 293, 2012.