MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
KYIV NATIONAL UNIVERSITY OF TECHNOLOGIES AND DESIGN
Faculty of Chemical and Biopharmaceutical Technologies
Department of Biotechnology, Leather and Fur

QUALIFICATION THESIS

on the topic **Whole genome sequencing and genetic analysis of exopolysaccharide producing Marine bacterium P. a. Hao2018**

First (Bachelor's) level of higher education
Specialty 162 "Biotechnology and Bioengineering"
Educational and professional program "Biotechnology"

Completed: student of group BEBT-20
Xuesong XIE

Scientific supervisor
Tetiana SHCHERBATIUK, Dr. Sc., Prof.

Reviewer
Iryna VOLOSHYNA, Ph.D., As. prof.

Kyiv 2024

KYIV NATIONAL UNIVERSITY OF TECHNOLOGIES AND DESIGN

Faculty: Chemical and Biopharmaceutical Technologies
Department: Biotechnology, Leather and Fur
First (Bachelor's) level of higher education
Specialty: 162 Biotechnology and Bioengineering
Educational and professional program Biotechnology

APPROVE
Head of Department of Biotechnology, Leather and Fur, Professor,
Doctor of Technical Science
Olena MOKROUSOVA
_____
«___»_____2024

ASSIGNMENTS
FOR THE QUALIFICATION THESIS
Xuesong Xie

1. Thesis topic **Whole genome sequencing and genetic analysis of exopolysaccharide producing Marine bacterium P. a. Hao2018**
scientific supervisor Tetiana Shcherbatiuk, Dr. Sc., Prof.

approved by the order of KNUTD "__"_____2024, №____

2. Initial data for work: assignments for qualification thesis, scientific literature on the topic of qualification thesis, materials of materials of Pre-graduation practice

3. Content of the thesis (list of questions to be developed): literature review; object, purpose, and methods of the study; experimental part; conclusions

4. Date of issuance of the assignments_____

EXECUTION SCHEDULE

№	The name of the stages of the qualification thesis	Terms of performance of stage	Note on performance
1	Introduction	From 01 April 2024 to 11 April 2024	
2	Chapter 1. Literature review	From 06 April 2024 to 20 April 2024	
3	Chapter 2. Object, purpose, and methods of the study	From 21 April 2024 to 30 April 2024	
4	Chapter 3. Experimental part	From 01 May 2024 to 10 May 2024	
5	Conclusions	From 07 May 2024 to 12 May 2024	
6	Draw up a bachelor's thesis (final version)	From 20 May 2024 to 25 May 2024	
7	Submission of qualification work to the supervisor for feedback	From 25 May 2024 to 27 May 2024	
8	Submission of bachelor's thesis to the department for review (14 days before the defense)	27 May 2024	
9	Checking the bachelor's thesis for signs of plagiarism (10 days before the defense)	31 May 2024	
10	Submission of bachelor's thesis for approval by the head of the department (from 7 days before the defense)	03 June 2024	

I am familiar with the task:

Student _____Xuesong XIE

Scientific supervisor _____Tetiana SHCHERBATIUK

SUMMARY

The exopolysaccharide-producing marine bacterium P.a. Hao2018 Whole-genome sequencing and genetic analysis

– Manuscript.

Qualification thesis on the specialty 162 «Biotechnology and Bioengineering». – Kyiv National University of Technologies and Design, Kyiv, 2024.

Leckerella, also known as Salmonella (Salmonella), is a class of Gram-negative bacteria, common in soil, water, animals and plants and human intestines. Leckerella is a potentially pathogenic bacterium that can cause infections in humans and animals, leading to food poisoning or infectious diseases. In marine environments, Lecbacteria may enter marine organisms through contaminated seawater or food chains, posing a potential threat to marine ecosystems and human health. The presence of Lecella may be influenced by factors such as seawater temperature, salinity, pH, and nutrients. Studies have shown that increased seawater temperatures may lead to faster rates of growth in the ocean, increasing the risk of infection. In addition, the competition and predation relationship of other microorganisms in the ocean can also affect the number and distribution of Lecbacteria. To prevent and control the lux-bacteria pollution in the ocean, start from the source, avoid the input of pollution sources, strengthen the monitoring and detection of seawater and seafood, and ensure the safety of seafood. Scientific and rational use of Marine resources, reducing the emission of pollutants and protecting the Marine ecological environment are important measures to effectively prevent and control the lux-bacterial bacteria in the ocean.

Our research group is committed to the prevention and control of Marine microorganisms. Recently, we isolated the seedling panel of a plant with the ability to produce good biological active exopolysaccharides P.a. Hao2018 (*Leclercia peneumoniae*), from the perspective of genome, genome means help us to find that genetic information, gene function, genome structure and other aspects of

information mining, help us establish a system of biological evolution and phylogenetic research. Through whole genome sequencing and comparative genomics analysis of the bacteria, based on the functional characteristics of the nine LeKe species, we found that these species do not have the degradation of aromatic compounds in the phenol degradation pathway. Meanwhile, in the carbon source metabolism, these species do not metabolize pyruvate and alcohol, *Leclercia sp.*LTM01、*Leclercia sp*. EMC 7 is also unable to perform lactate metabolism, but they are connected through other carbon source metabolic pathways to jointly regulate carbon source utilization within organisms. The genus Lecella is a class of bacteria widely found in nature with certain pathogenicity and antibiotic resistance. The bacterium adapts to the host microenvironment in several ways, possibly acquiring resistance genes through phage horizontal gene transfer. Studies have found that this bacterium may also play an important role in biogeochemical cycles.

In general, through whole genome sequencing and analysis, basic data support can be provided for further research on the diversity, environmental impact, material degradation and infection prevention and control of the bacterium, which is helpful to deepen the understanding and prevention and control of the bacterium.

Key words: marine bacteria; Lecella; whole genome sequencing; functional analysis; drug resistance

TABLE OF CONTENTS

INTRODUCTION

In this study, through whole genome sequencing, comparative genomics analysis and functional module analysis, we gained an in-depth understanding of the genomic structure and functional characteristics, kinship and mobile gene information of this bacterium, and found that *Leclercia peneumoniae* 2018 was highly similar to other strains of Leclercia. That's about 84.5%. Comparative analysis of gene families revealed that Howe 2018 shares most gene families with other strains of Rockleia, with few specific gene families, showing a high degree of correlation. In addition, in the analysis of phage insertion elements as a whole, it was found that there are two potential phages in the Howe 2018 chromosome, containing 11 coding genes, which are closely related to their pathogenicity and antimicrobial properties. It is worth noting that as a disease-causing bacterium, Leklet also exhibits drug resistance, which will adversely affect our treatment, and these resistance genes may be acquired through phage level gene transfer. An in-depth study of this process will help us control the bacteria. Overall, as one of the important microbial communities in the ecosystem, Rockleia is resistant as a disease-causing bacterium that may adapt to the host microenvironment in a variety of ways and play an important role in geochemical cycles.

The relevance of the topic is whole genome sequencing and genetic analysis

The purpose of the study is the whole genome sequencing and analysis of Leclercia peneumoniae provided basic data support for further understanding and research on the diversity, environmental impact, material degradation, and infection prevention and control of Leclercia.

The objectives of the study study is the whole genome sequencing and analysis of *Leclercia peneumoniae* provided basic data support for further

understanding and research on the diversity, environmental impact, material degradation, and infection prevention and control of Leclercia.

The object of the study P.a. Hao2018

The subject of the study P.a. Hao2018

Research methods whole genome sequencing

The scientific novelty An in-depth study of this process will help us control the bacteria

The practical significance of the results obtained is An in-depth study of this process will help us control the bacteria

Approbation. -

CHAPTER 1 INTRODUCTION

1.1 Research status of marine bacteria

Bacteria in marine environments are important components of marine ecosystems 1. It plays an indispensable role in the Marine ecosystem and participates in the

degradation, recycling and energy conversion process of organic matter. These bacteria inhabit diverse environments, including seawater, deep sea and hydrothermal vents as well as polar environments 2. Marine bacteria can colonize the outer membrane of marine plants and animals or enter internal tissues; some bacteria can establish a highly symbiotic relationship with their specific host organisms 3. In recent years, with the thorough study on the diversity and function of marine microorganisms, the attention to marine bacteria has gradually increased.

In the past research, scholars have conducted in-depth research on Marine bacteria by means of isolation, culture and molecular biology techniques. Through the analysis of their genome structure, metabolic pathways, and growth characteristics, the important position and mechanism of action of marine bacteria in marine ecosystems are revealed 4. At the same time, the researchers also found that there are differences in the species and functions of Marine bacteria in different Marine environments, which provides important clues to further explore their functions and ecological significance in the Marine ecosystem.

Now scientists have made a series of important advances in the field of Marine bacteria research, but there are still many unknown areas and problems to be solved.

There is still a lack of systematic research on the genetic characteristics, adaptation mechanisms and interactions with other microorganisms. Therefore, vigorously

carrying out whole genome sequencing and genetic analysis of Marine bacteria will help people to more comprehensively understand the biological characteristics and functions of Marine bacteria, so as to provide a new perspective and theoretical basis for the research in the field of Marine microbial ecology.

1. 2 Maturation of bacterial whole-genome sequencing technology

Bacterial whole genome sequencing technology is a rapidly developing biotechnology field in recent years. By sequencing the complete genome of bacteria, we can have a deep understanding of the genetic characteristics and functions of bacteria 5. With the continuous progress of sequencing technology, from traditional Sanger sequencing to today's high-throughput sequencing technology, sequencing cost has been greatly reduced, sequencing speed has been greatly improved, and data quality and coverage have been significantly improved. The growth of whole genome sequencing technology provides researchers with more comprehensive and accurate bacterial genetic information, and provides strong support for the deep understanding of the genetic mechanism and biological characteristics of bacteria.

During the development of bacterial whole genome sequencing technology, new sequencing platforms and technical methods have been emerging. For example, the wide application of second-generation high-throughput sequencing technologies, such as Illumina HiSeq and MiSeq, with high throughput, high accuracy and low error rate, can be completed quickly and efficiently. At the same time, the emergence of the third generation of single-molecule sequencing technology, such as PacBio and Oxford Nanopore, breaks through the limitations of traditional sequencing technology and can realize long read long segment sequencing, providing more possibilities for the assembly of the whole bacterial genome and structural variation analysis.

With the continuous optimization and improvement of bioinformatics analysis methods, the processing and interpretation of bacterial whole genome sequencing data has become more efficient and accurate. Bioinformatics tools and software based on big data and artificial intelligence are constantly emerging, such as Bowtie, BWA,

SAMtools, etc., which can help researchers to carry out complex bioinformatics analysis such as sequence alignment, gene prediction and functional annotation, and accelerate the mining and utilization of bacterial whole genome sequencing data.

1.3 Study purpose and significance

Our research group is committed to the utilization and development of Marine microbial resources, isolated from the wrinkled plate and seedling plate, a plant with the ability to produce good biologically active exopolysaccharides P.a. Hao2018, through whole genome sequencing and comparative genomics analysis of the bacterium, we deepen our understanding of the molecular evolution, metabolic potential and ecological adaptation of marine bacteria. Moreover, the studies against the marine bacteria P.a. Hao2018 Whole-genome sequencing and genetic analysis can also provide an important reference for the development and utilization of Marine microbial resources, and promote the development of the field of Marine biotechnology. Therefore, the purpose of this study is to reveal its genetic characteristics and biological ecological significance through the genome, which will not only provide reference for the deep understanding of the life activities, genome characteristics and functions of the bacterium, but also provide scientific basis for the diversity and ecological functions of Marine microorganisms, which has important research significance and application prospects.

CHAPTER 2 Experimental materials

2.1.1 The bacteria body

A marine bacterium isolated from the microbial membrane layer on the surface of abalone seedlings, named P.a. Hao2018, served in A217 laboratory of food Building of Qilu University of Technology.

2.1.2 Culture medium

T2 medium: 10 gL of glucose-1, 10 gL of soy protein-1, Beef paste, 5 gL-1, Yeast extract of 2 gL-1, 1L of deionized water, and 18 gL of agar-1、pH 7.0

2.2 Experimental method

2.2.1 Strain culture

Ultra table table operated using a sterile inoculation ring from P.a. Hao2018 Colonies were picked from the medium and inoculated on T2 solid medium to ensure the number of colonies in each inoculation. After inoculation, the plates was placed in a thermotemperature incubator and kept at 37 C for 16 hours in favor of P.a. Hao2018 For the growth and reproduction of bacteria, repeat the previous operation steps in order to ensure that the picked colonies are single colonies and avoid confounding. Select a single P.a. Hao2018 Colonies, and transferred to tubes containing 50 ml of T2 liquid medium. The tube was placed in a 37 ° C incubator with a aker speed of 220rpm to facilitate P.a. Growth and proliferation of Hao2018, and cultured for 16 hours. Colcolonies in P.a. Hao2018 Photographs of single colonies to keep records. This process ensures that the acquisition of a single P.a. Hao2018 Strain, providing reliable basic data for subsequent experimental studies.

2.2.2 Genome extraction and sample preparation

picking P.a. Single colonies of Hao2018, which were transferred to a petri dish containing 32,000 ml of T 2 liquid medium. The plates were subsequently incubated in a constant temperature 37℃ incubator on a shaker at 220 rpm for 16 hours. Next,

at 9000 rpm, the cultured bacterial solution was placed in a 4℃ centrifuge for 5 min, and the supernatant was aspirated and discarded using a sterile pipette gun, leaving the collected bacteria. Finally, through the electronic balance weighing, we learned that the weight of the bacteria is 7 grams. The bacteria were loaded into 50-ml centrifuge tubes and sealed using a sealing membrane, while clearly marking the name of the sample and the time of collection. With dry ice, the samples were immediately transported to Shanghai Pensnor Biotech for P.a. Whole-genome sequencing work of Hao2018.

CHAPTER 3

3.1.1 Whole-genome sequencing and data quality control analysis

This project adopts the whole genome shotgun method (Whole Genome Shotgun, WGS) strategy to construct libraries of different inserts. Using the second-generation sequencing technology (Next-Generation Sequencing, NGS), based on the Illumina NovaSeq sequencing platform, while using the third-generation single molecule sequencing technology, based on the Oxford Nanopore ONT sequencing platform for these libraries to P.a. Hao2018 Performed the analysis of the whole-genome sequencing. The machine data is saved in two-end Paired-end FASTQ format. The second-generation sequencing data are analyzed by FastQC. In order to ensure the quality and accuracy of subsequent information analysis, the sequencing data is necessary to be further filtered, and the original sequencing data (rawdata) is filtered to generate high-quality sequence (high quality data). The criteria for data filtering is: joint contamination removal using AdapterRemoval (ver.2.1.7) The software removes connector contamination at the 3'end6; Then SOAPec (v2.0) software was used for quality correction of all gene sequences with a Kmer frequency of 17 7. Finally, the third generation machine data of the sample is counted.

3.1.2 Genome sequence assembly and effect evaluation

The second-generation high-throughput sequencing data obtained after adaptor decontamination and gene sequence quality correction were assembled using A5-miseq v20150522 software to obtain the contigs and scaffolds sequences8. Data in fastq format obtained from PacBio RSII third-generation sequencing were assembled using CANU software to obtain scaffold sequences9. The resulting scaffold sequence was corrected with the software was pilon software10. The resulting contigs of Illumina MiSeq and PacBio RSII sequencing data was used for collinearity analysis using mummer software to reconfirm the assembly results, determine the positional

relationship between contigs, and fill in the gap between contigs 11. Finally, the three-generation contig results were corrected using pilon software and finally were spliced to obtain the complete sequence.

3.1.3 Functional annotation of genes

(1) Protein-coding gene prediction

The Glimmer 3.0 software was used to predict the P.a. Hao2018 Genes capable of encoding proteins in the whole genome sequence 12. Set the open reading frame (ORF) length of no less than 110bp, and the remaining parameters are the default setting of Glimmer 3.0.

(2) eggNOG annotation of protein-coding genes of protein-coding genes

Using blast software for P.a. Hao2018 Genes were COG (orthologous gene cluster), and annotation alignment was completed in eggNOG (V4) database 13, The cut-off value of the sequence alignment was chosen as 1e-6, and the eggNOG number of the best hits was assigned to the corresponding protein-coding gene. The correspondence between the eggNOG number and the eggNOG classification catalogue was further used to classify each protein on the eggNOG classification catalogue.

(3) The KEGG annotation of the protein-coding genes

The KO and Pathway annotation of protein coding genes is mainly completed using the KAAS automated annotation system of KEGG, in which "For Prokaryotes" is selected for gene set and the discrimination rule for KO annotation of base coding genes is BBH14. After the KO annotation, the KO was mapped to the corresponding KEGG Pathway pathway.

GO annotation of the protein-coding genes

GO has a total of three ontology, which describe the molecular function (molecular function), the cell location (cellular component), and the biological process (biological process). GO annotation of protein coding genes was done using BLAST2GO software, GO annotation with BLAST2GO default parameters, and GOSlim annotation results were done by map2slim.

3.2 Results and analysis

3.2.1 Whole-genome sequencing and data quality control analysis

In this study, the libraries were sequenced by the Oxford Nanopore ONT sequencing platform based on the second generation sequencing technology based on the Illumina NovaSeq sequencing platform. In total, two libraries were constructed from this sequencing, and the library details are shown in Table 3-1 below.

Table 3.1 – **P.a. Hao2018 Body sequencing strategy**

Sample name	Library name	Libraries were inserted into fragment	Sequencing theplatform	Sequencing mode
TB16	--	-- kb	ONT	Standard
TB16	PE	400 bp	Illumina NovaSeq	Paired-end, 2×150bp

Second generation sequencing data collation and quality control

Table 3.2 – **Statistics of the sequencing data**

Sample name	Library name	Total Reads	Total number of base and bases (bp)	N (%)	GC (%)	Q20(%)	Q30(%)
T	PE	8,716	1,316,1	0	54.59	98.	95.

B16	,198	45,898	.0247	19	16

Note: N (%): percentage of fuzzy bases; GC (%): GC content; Q20 (%): percentage of bases with recognition accuracy above 99%; Q30 (%): percentage of bases with recognition accuracy above 99.9%.

The second-generation sequencing data based on the Illumina MiSeq sequencing platform were evaluated from the three aspects of base quality distribution, Reads average error rate distribution and Reads sequencing base content distribution to detect the quality of the data.P.a. The Hao2018 base mass distribution results show (Figure 3-1 A) that the base mass at the 5′ end and the base mass in the middle part are very average, while the average mass of the overall base is high.P.a. The average error rate distribution of Hao2018 is mainly used to evaluate the probability of sequencing error expected from the sequencing data. As shown in Figure 3-1 B, the sequence error rate at the 5′ and 3′ ends. The Base content distribution is mainly used to detect whether there are deviation from AT and GC. As shown in Figure 3-1C, the four lines A, T, G and C are basically parallel, indicating that the bases all performed well in library construction and sequencing at one time, and can conduct subsequent information analysis, but the separation occurs at the 5′ end and 3′ end. Overall, the Illumina MiSeq-generation sequencing data is of high quality, but some low quality data remained at the 5′ and 3′ ends. Therefore, the data has contaminated data and some low quality sequences for advanced filtering.

(2) Three-generation sequencing data sorting and quality control

The samples are based on the PacBio RSII sequencing platform, and the data of the third generation are compiled. The specific results are shown in Table 3-3. The sequence length distribution of the three-generation sequencing data, see Figure 3-2, shows that the data is of good quality and enables subsequent information analysis.

Table 3.1 – **The sequence length statistics of the third-generation sequencing data**

Sample	TB16
Total number of sequences	152,521
Genome Length (bp)	1,473,172,894
Longest sequence length (bp)	145,812
Minimum sequence length (bp)	482
N20 (bp)	41,266
N50 (bp)	20,114
N90 (bp)	3,673
Uncertain number of bases	0
The proportion of the number of bases to the length of the splicing sequence	0
GC content%	54.64

After the data quality control and filtering of the data of the second and third generation, the sequencing data of the second and third generation are assembled to obtain the final complete genome sequence.

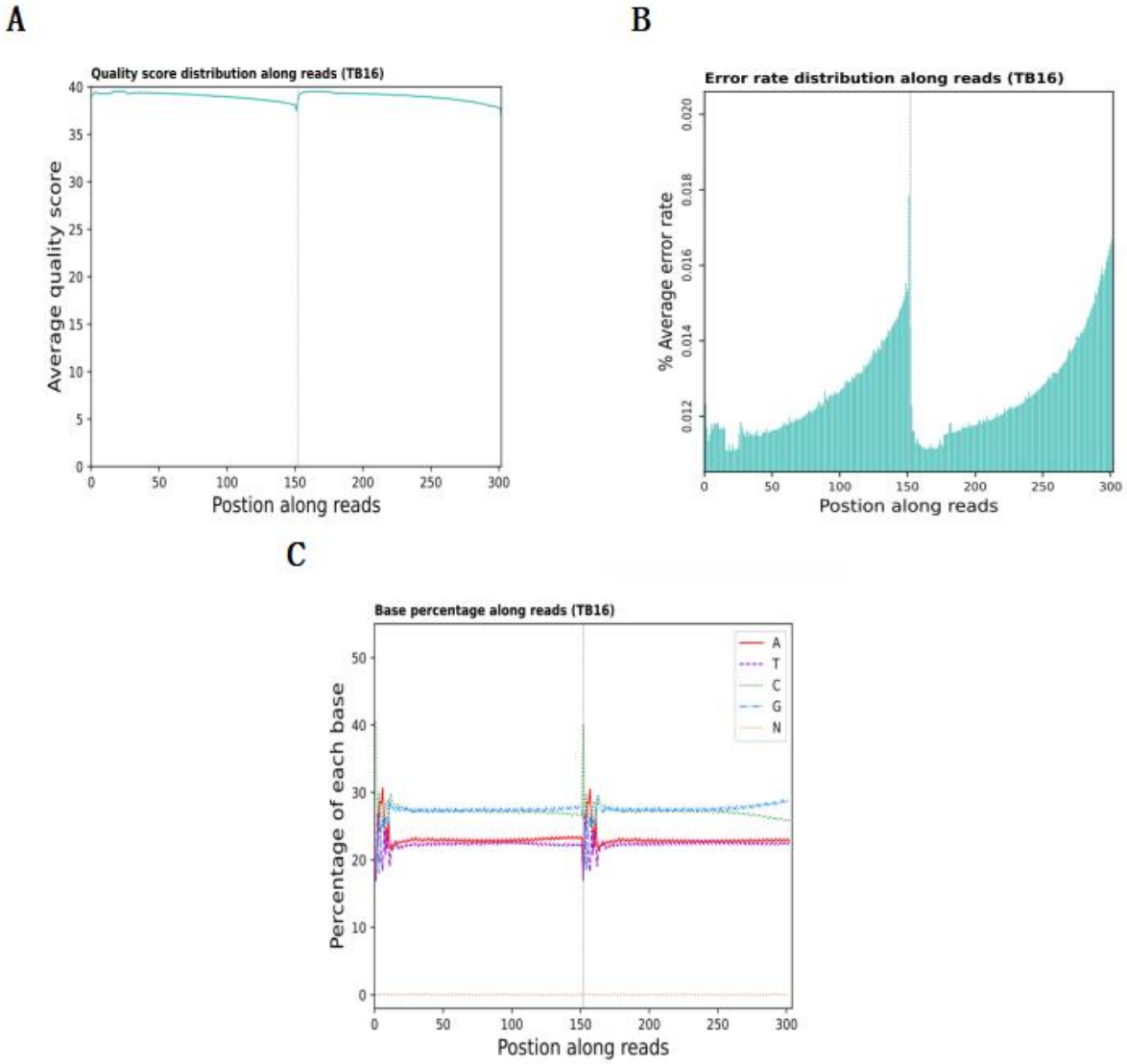


Figure 3.1 – Quality control analysis of second-generation sequencing data.

Note: A: mass distribution of bases; B: Distribution of average error rate of Reads; C: Distribution of base content of Reads sequencing.

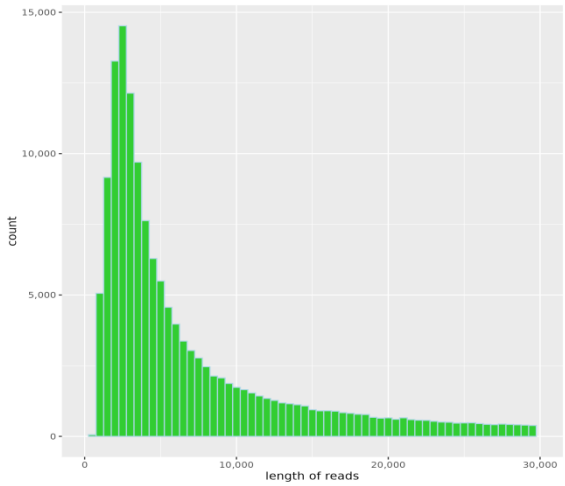


Figure 3.1 – The sequence length distribution of the third-generation sequencing data

3.2.2 Basic features of the genome

Based on the sequencing results and the analysis, P.a. The Hao2018 whole genome consists of a 4,461,691bp complete circular chromosome with a GC content of 55.12%. In the absence of a plasmid sequence, P.a. The Hao2018 genome is the chromosomal sequence.

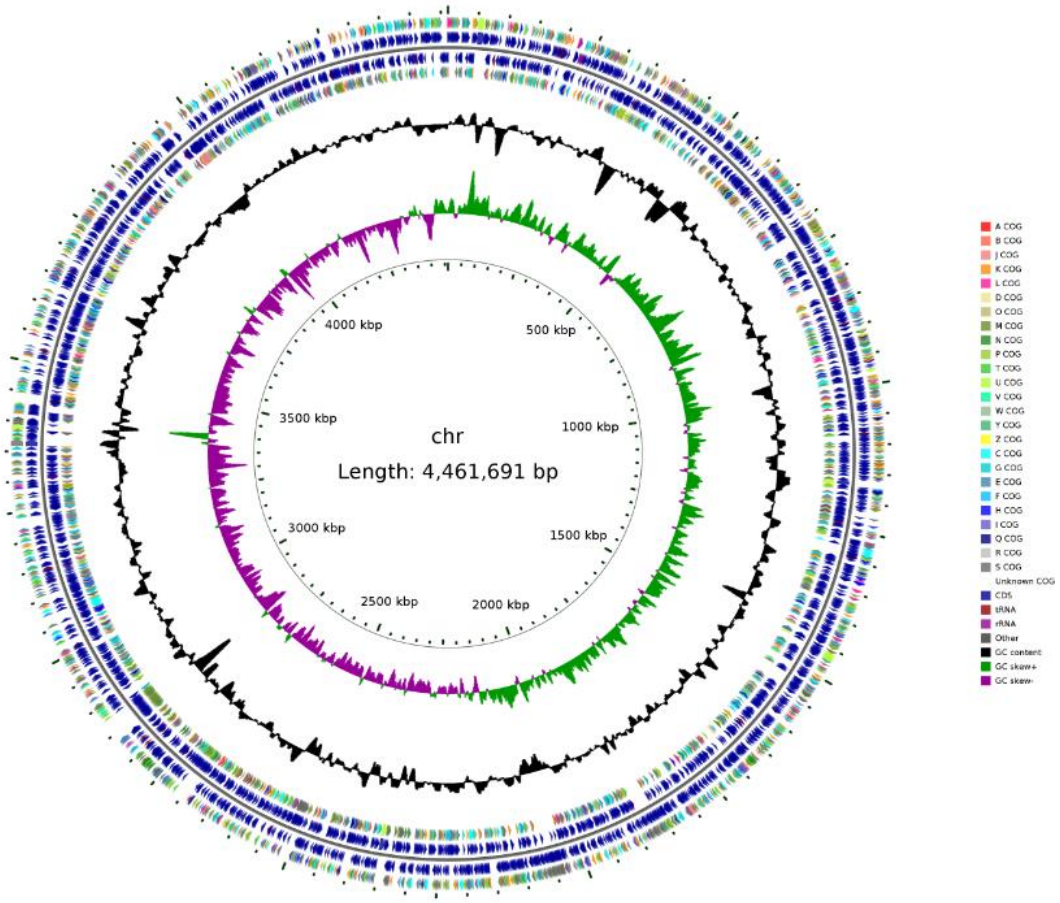


Figure 3.3 – graph 3-3P.a. Hao2018 Genome circle mapping

3.2.3 Functional annotation of the genes

P.a. The full genome sequence of Hao2018, with 4,161 protein-coding regions, represents 87.83% of the full sequence ORF. Through the functional annotation of the genes in different databases, the target genes were analyzed according to the experimental purpose. Thus, at the molecular level for the P.a. Hao2018 Potential functions are resolved.

(1) eggNOG annotation of protein-coding genes

The gene-encoding protein sequences were aligned with the protein sequences in the eggNOG (COG) database using diamond blastp to obtain the corresponding COG functional annotation results, and the proteins were functionally classified according to the COG annotation results. P.a. Hao2018 A total of 4161 protein-coding genes were annotated, For 76.64% of the whole genome sequence, As shown in Figure Figure 3-2, Transport and metabolism of carbohydrates (G, Carbohydrate transport and metabolism), amino acid transport and metabolism (E, amino acid transport and metabolism), the transport and metabolism of inorganic ions (P, Inorganic ion transport and metabolism), transcriptional correlation (K, Transcription), and cell wall / cell membrane / envelope synthesis (M, cell wall / membrane / envelope biogenesis), Of 8.31%, 7.71%, 7.28%, 7.2%, and 6.2%, respectively.

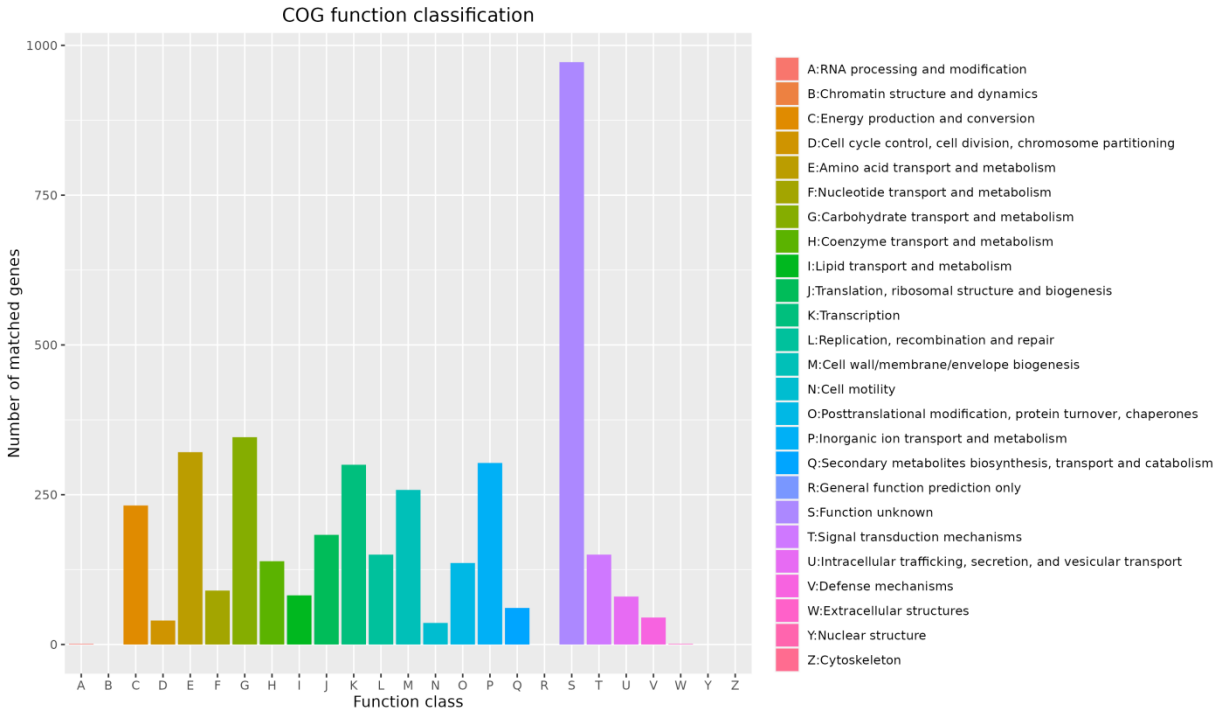


Figure 3.4 – graph 3-4P.a. Hao2018 GOG Functional annotation

(2) The KEGG annotation of the protein-coding genes

KEGG Pathway Annotation, or metabolic pathway annotation, obtaining a network of intermolecular interactions and reactions within a species. It provides information on the association between genes and metabolic pathways, helping researchers to understand the complex relationships such as signaling and metabolic pathways in biological processes. Through KEGG Pathway annotation, high-throughput data such as genomics, transcriptomics, proteomics and more data can be connected with specific biological pathways, so as to deeply explore relevant information about gene function and metabolic pathway regulation. In this study, data for P.a. The 2,767 genes of Hao2018 were KEGG annotated as shown in Figures 3 - 5. According to Figure 3-5, 279 genes were annotated for cellular processes (Cell μ lar Processes), 428 genes for Environmental Information (Environmental Information Processing), 226 genes for genetic information processing (Genetic Information Processing), and 161 genes for human diseases (Human Diseases) , New metabolism (Metabolism) was annotated with 1606 genes, and organism systems (Organismal Systems) with 67 genes. These six types, each of them, contain their own subtypes. KEGG enrichment analysis shows that signaling and cellular processes (Signaling and cellular processes), genetic information processing (Genetic information processing), and carbohydrate metabolism (Carbohydrate metabolism) are the three main metabolic pathways, and these genes control the production of exopolysaccharides.

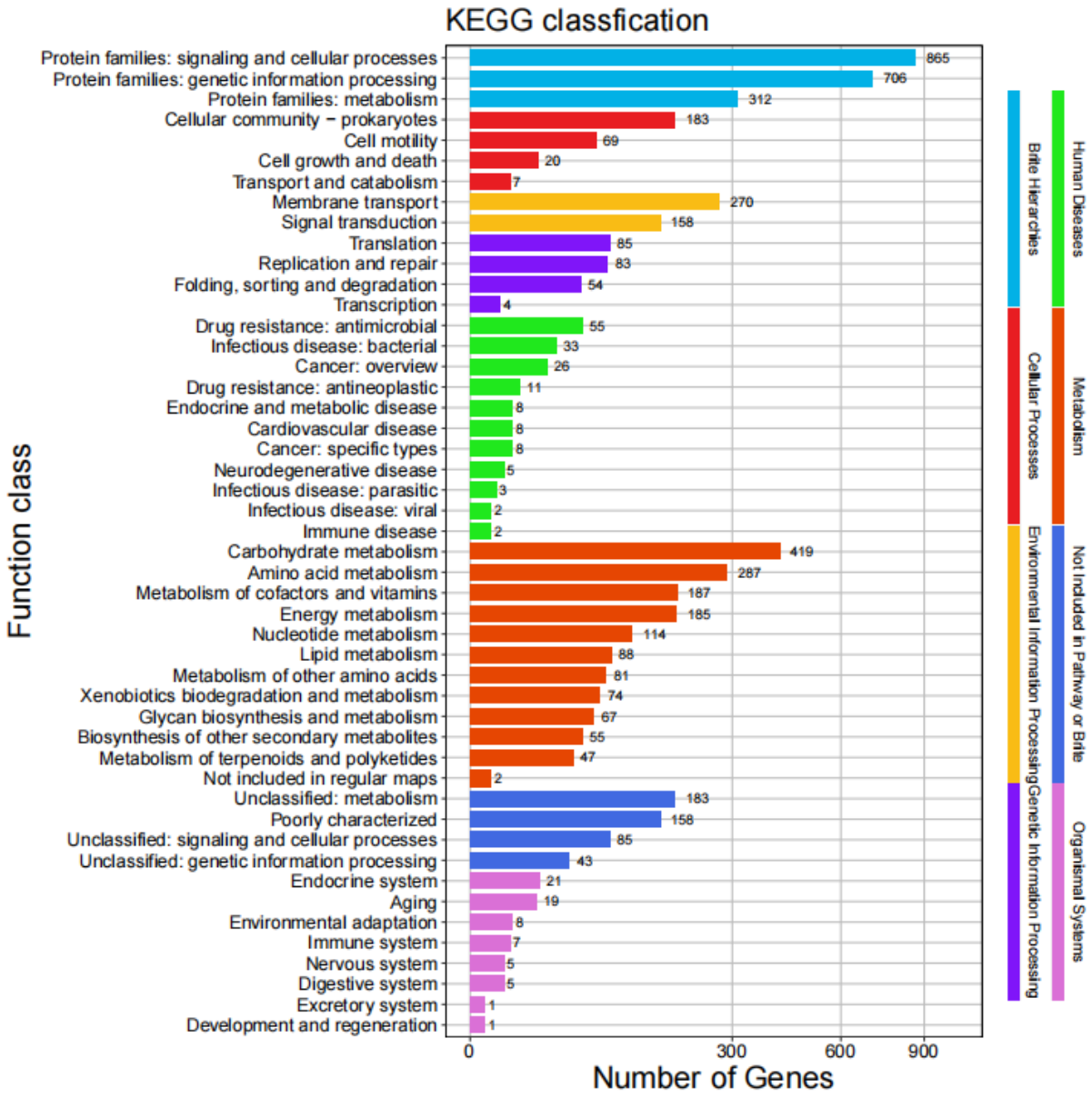


Figure 3.5 – P.a. Hao2018 KEGG, Functional annotation

(3) The GO annotation of the protein-coding genes

The strain P.a. Hao2018 Sequences were aligned with the GO (Gene Ontology) database, and the statistical analysis revealed a total of 20,870 annotated genes. As shown in Figure 3-6, the GO database annotated the proteins by cellular components (Cellular Component, CC), biological processes (Biological Process, BP), and molecular function (Molecular Function, MF). A total of 80 entries included 41,13 and 26, respectively. In the CC, The number of annotated genes was 4,799, Among them, the genes related to cell (Cell), intracellular part (Intracellular), cytoplasmic

membrane (plasma membrane), and cytoplasm (Cytoplasm) showed the highest similar trend, There are 1400,765,593,576 each; The number of genes annotated in BP was 9784, The entries involving the most number of genes include cellular nitrogen compound metabolic processes (Cellular nitrogen compound metabolic process) and biosynthetic processes (Biosynthetic process), There were 1,094 and 1085 genes, respectively; The number of genes annotated in the MF broad class was 7509, Among the entries with the most involved genes are ion binding (Ion binding), oxidoreductase activity (Oxidoreductase activity) and DNA binding (DNA binding), Their numbers were 970,465, and 433, respectively.

According to the quality control analysis results of the data, the quality of quality control from base quality distribution, average error rate distribution and base content distribution of Reads is high, and the quality value Q=40, indicating that the error of sequencing data is very low, close to 0.01% (the highest value of Q is 40). The quality control results showed the presence of little low quality data at the 5′ and 3′ ends of the sequences based on Illumina MiSeq sequencing. These low-quality sequences are short and may have an impact on the sequencing uniformity of microbial genomes. This phenomenon may be due to trace contamination of the joint at the initial departure of the sequence and the emergence of low quality data due to the gradual depletion of enzymes as sequencing proceeds. High-quality filtering of all sequencing data was performed to ensure sequence accuracy. Combining second-generation sequencing technology and third-generation sequencing technology, the P.a. Hao2018 performed deep sequencing and successfully obtained a fine map of circular chromosomes. After the loop map analysis of the chromosome, the sequences were found to be quite complete, consistent with the data quality control results. GOG, KEGG and GO annotation of protein coding genes: COG annotation showed that their protein functions involved gene transcription, transport of amino acids, carbohydrates and inorganic ions and metabolism; KEGG annotation indicated that P.a. The genes of Hao2018 are mainly involved in carbohydrate, amino acid metabolism and membrane transport; GO annotation reveals P.a. The protein function

of Hao2018 is mainly involved in major functions in biological processes. Although coding genes are annotated through different databases and from different angles, the results are basically consistent, ensuring the accuracy of predicting proteins for coding genes. Moreover, in chromosomes, some genes fail to be annotated for specific protein functions, which deserves further investigation.

3.2. Comparative analysis of gene families

In the process of genome evolution, the cluster analysis of homologous genes is the analysis of gene family (gene family). Homologous genes are genes with similar biochemical functions. The gene family of homologous proteins is also known as the "core gene family", and the other homologous protein groups are known as the "accessory" gene family. Genes different from any other two genomes are called specific gene families. Comparative analysis of gene family was performed using orthomcl (version 2.0.8) software with the alignment length set to 70%15。

3.3.An evolutionary analysis

(1) ANI analysis of the whole gene sequence

The evolutionary relationship was determined from the average nucleotide identity (ANI) using OAT 0.93.1 (Orthologous AverageNucleotide Identity Tool) tool software**Помилка! Джерело посилання не знайдено.**。

(2) Based on the single-copy gene phylogenetic tree

Sequence alignment of each single-copy gene pair was performed separately using the program Muscle (version 3.8.425)15. Unreliable sequence alignment sites were removed by Gblock (version 0.91b)16. The best amino acid substitution model was selected using ProtTest (version 3.2). The phylogenetic tree was constructed using the maximum likelihood method in IQ-TREE 3.1 software, after which the reliability of the developmental tree branches was verified (set to 1000 replicates).

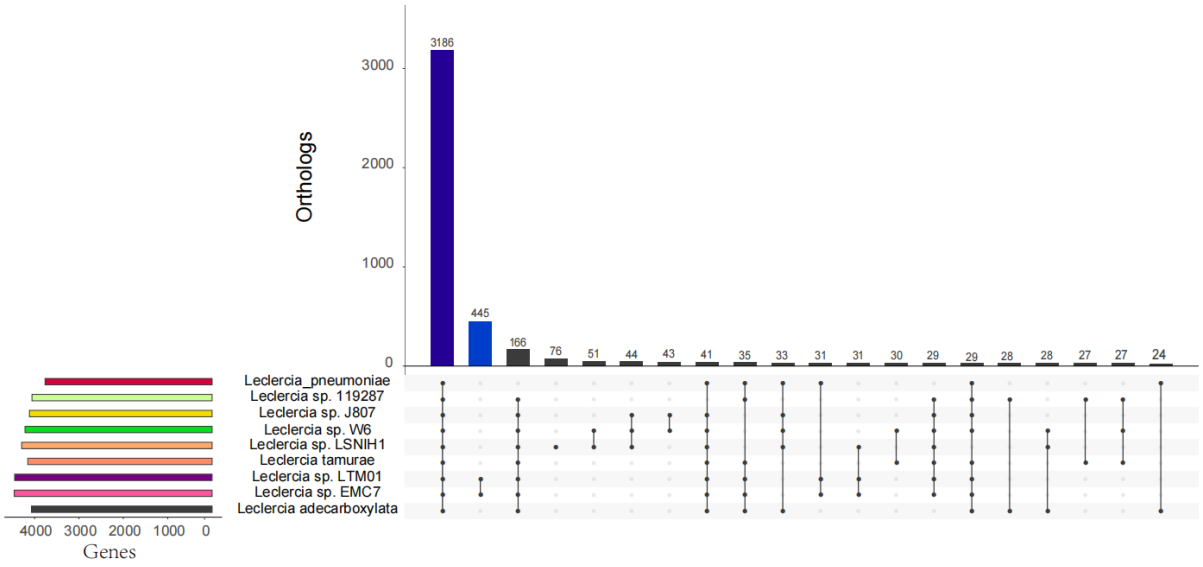
3.4. Insertion element analysis of the phage whole

The insertion element of the phage as a whole is a stretch of DNA sequence used in genetic engineering to insert a foreign gene into the phage genome. In the field of genetic engineering and biotechnology, phages are widely used as vectors to deliver foreign genes to host cells, in which insertion elements play a crucial role. The analysis of insertion elements can help researchers to better understand the structure and function of phage genome, and also provide an important reference and guidance for the practice of genetic engineering. The PHASTER software is used to predict the P.a. Hao2018 (Leclercia peneumoniae) The original phage gene 15。

3.5.Comparative analysis of gene families

(1) Comparative gene family analysis

The cluster analysis of the whole genome of nine strains (Figure 4-3) obtained 4404 gene families, and 3186 were conserved gene families, accounting for 72.34% of the total gene families, and the other 1218 were accessory gene families. Strain Leclercia sp. LSNIH1 Specific gene families exist in a number of 76. Description of the Leclercia sp. LSNIH1 Has a large specificity. In strain P.a. In Hao2018 (Leclercia peneumoniae), 41 identical gene families are found in all cases except Leclercia sp. Of the seven strains of 119287,35 identical gene families were all present except for Leclercia sp.J807、Leclercia sp.W6、Leclercia sp. In five strains of LSNIH1,33 identical gene families have Leclercia sp.J807、Leclercia sp.W6、Leclercia sp. These four strains, LSNIH1 and Leclercia adecarboxylata, have 31 identical gene families, all present in Leclercia sp.LTM01、Leclercia sp. In both EMC 7 strains, 29 identical gene families exist in both strains except Leclercia sp. Seven strains of LSNIH1. P.a. Hao2018 (Leclercia peneumoniae) This strain, almost no specific gene family, all genes are shared with other congeners.



graph 4-3P.a. Gene family analysis of the Hao2018

3.6. The evolutionary analysis

Evolutionary trees constructed based on a single or a few tandem genes have a great potential to produce different gene topologies, while only evolution constructed based on genome sequences more accurately illustrate taxonomic relationships in bacteria17. According to the results of (Figure 5-1), Leclercia _adecarboxylata, Leclercia sp.LTM01、Leclerciasp.LSNIH1、Leclercia sp.W6、Leclercia sp.J807、Leclercia sp.119287、Leclercia sp. EMC 7, Leclercia tamura e, and P.a. Hao2018 (Leclercia peneumoniae) Similar similarity. According to the phylogenetic tree, the closest related to this strain is Leclercia_adecarboxylata.

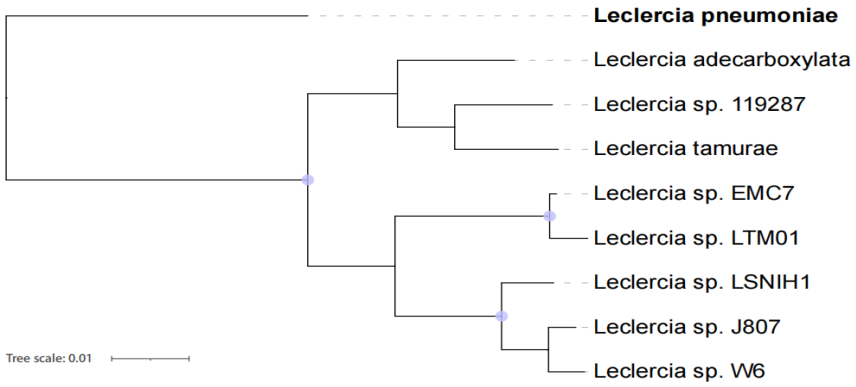


Figure 4-1 A phylogenetic tree constructed based on single-copy orthologous genes of the genus Leclercia

To further elucidate the P.a. Taxonomic relationship of Hao2018, we used the OAT software, using OrthoANI, namely the average nucleotide consistency, to evaluate the taxonomic relationship of the above nine bacterial strains21. the result shows that,P.a. The similarity coefficient between Hao2018 (Leclercia peneumoniae) and these 8 strains is about 84.5%, as shown in Figure 4-2. The conditions known as P.a. The Hao2018 strain is genus Leclercia.

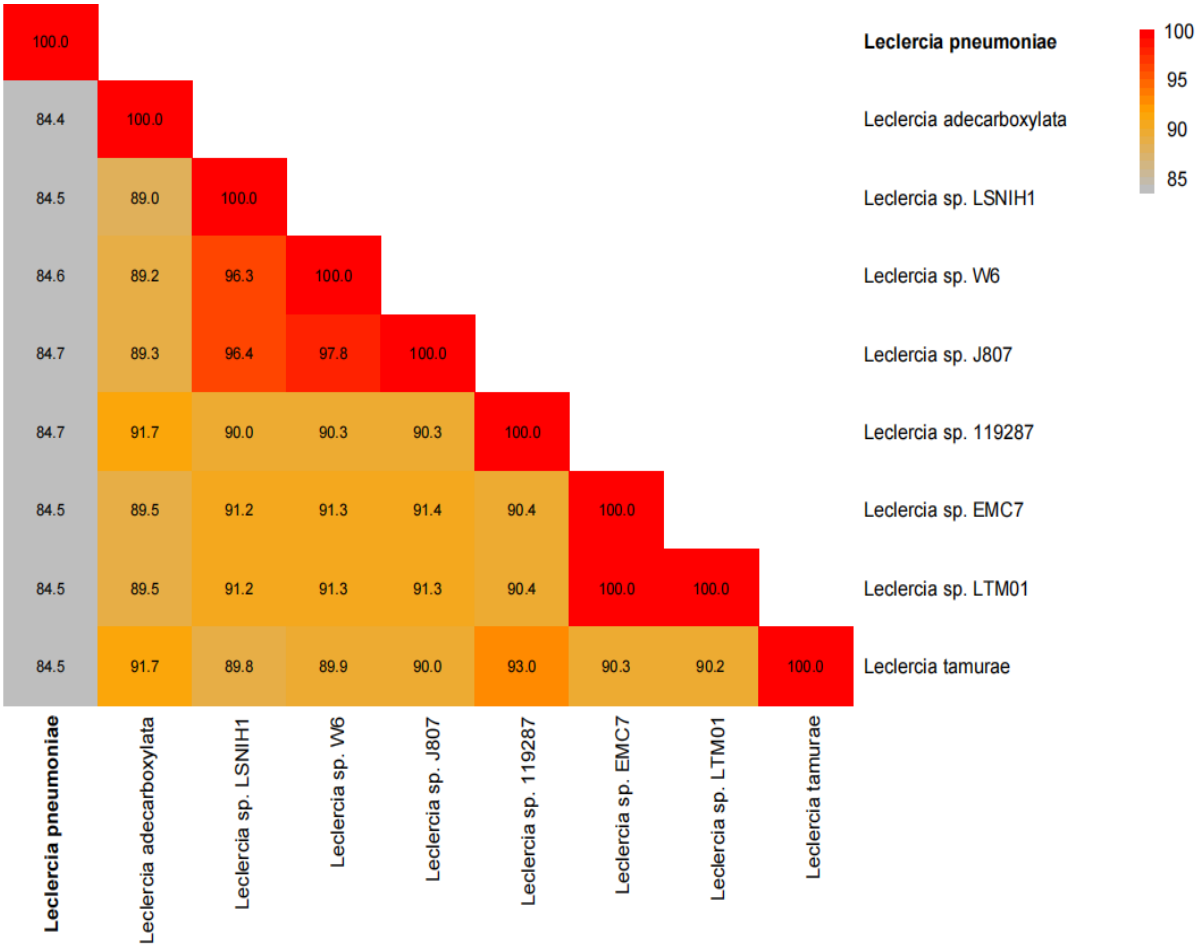


Figure 4-2 for ANI genomic similarity

3.7. Insertion element analysis of the phage whole

A prophage is a virus that integrates its own genome with the bacterium. But in terms of the horizontal gene transfer, as a carrier of horizontal transfer, bacteriophages may also promote bacterial adaptation to the host environment and be widely used24. Therefore, it is called opening the door to the study of dark matter in the biological universe. So we speculate that the bacterium may acquire new genes through horizontal gene transfer and thus adapt to the microenvironment. But these require further analysis,

Analysis of mobile genetic elements revealed that this genome does not contain plasmid sequences, but contains phage sequences, P.a. Hao2018 (Leclercia peneumoniae) has two prophages, containing 11 CDs. It has a close relationship with antibiotic resistance and antibacterial properties24. The details of the prophages are shown in Table 4-1.

Table 4-1 P.a. Hao2018 Insertion element of the phage whole

seq_name	chr|provirus_2497627_2525496	chr|provirus_2020585_2071710
length	27870	51126
topology	Provirus	Provirus
coordinates	2497627-2525496	2020585-2071710
n_genes	39	71
genetic_code	11	11
virus_score	0.9756	0.973
fdr	NA	NA
n_hallmarks	18	16
marker_enrichment	48.4928	70.6251
taxonomy	Viruses;Duplodna	Viruses;Duplodna

	viria;Heunggongvirae;Uroviricota;Caudoviricetes	viria;Heunggongvirae;Uroviricota;Caudoviricetes

3.8. Genome-wide functional annotation and visualization of functional modules

Genome-wide annotation

To P.a. Hao2018 (Leclercia peneumoniae) Eight strains performed the whole genome for functional gene annotation. In functional gene annotation, common methods include homologous alignment, functional domain prediction, etc. Homology alignment refers to comparing the sequence of the gene to be annotated with the sequence of the genes with known function to infer the function of the gene to be annotated by similarity. Functional domain prediction is done by predicting the functional domains in the protein encoded by the genes.

Visualization of the functional modules

Heatmap mapping of the functional modules of these nine bacterial strains was performed using ComplexHeatmap.

Visualization and analysis of functional modules

(1) Functional pathway of amino acid transferases

By performing the genome-wide functional annotation of the eight strains and performing the functional heat map annotation as shown in Figure 5-1, the heat map shows that in the amino acid transferase functional module, P.a.Hao2018（Leclercia peneumoniae）、Leclercia sp.LTM01、Leclercia sp .EMC7、Leclercia sp .119287 has a 4-aminobutyrate aminotransferase (4-aminobutyrate aminotransferase) synthesis gene, while the other six strains have not, and none contains a serine-pyruvate aminotransferase (Serine-pyruvate aminotransferase) synthesis gene. Aminotransferases are an important class of enzymes that play a key role in the regulation of amino acid metabolism within living organisms. Among them, 4-GABA aminotransferases and related aminotransferases play important roles in the amino

acid metabolism pathway. Aminotransferases can be divided into classes I and II, which function in different metabolic pathways. Phosphoserine aminotransferases involved in the serine metabolic pathway are important for protein synthesis. Ornithine / acetylornithine aminotransferases participate in the anabolic pathway of ornithine and have important effects on nitrogen metabolism in living organisms. Branched chain amino acid amino transferase plays a role in the metabolic pathway of branched chain amino acids, affecting protein synthesis and energy metabolism. Aspartate / tyrosine / aromatic aminotransferases and histidine phosphate / aromatic aminotransferases play important catalytic roles in amino acid metabolic pathways. Serine-pyruvate aminotransferase participates in the conversion between serine and pyruvate and is important for the balance of cellular metabolism. The study of these aminotransferases helps to deeply understand the regulatory mechanism of amino acid metabolism pathways in living organisms, which has important significance for the study of related diseases and drug development.

(2) Functional pathway of acetate conversion into ethanol

All nine strains have a functional pathway for the conversion of acetate to ethanol. Acetaldehyde is a common chemical reaction. In the continued catalytic reduction reaction, the acetaldehyde is then reduced to ethanol. Acetate is one of the precursors of acetaldehyde, which can be obtained by oxidation reaction, and acetaldehyde can be converted to ethanol through reducing reaction. This series of reactions constitutes a continuous transformation process and is one of the common reaction paths in organic synthesis.

Functional pathway of aromatic compound degradation

Of these nine strains, P.a. Hao2018 (Leclercia peneumoniae) does not have a functional pathway for the degradation of resorcinol (Catechol) and hydroquinone (Protocatechuate) in the phenylpropyl pathway, and none of these nine strains has a functional pathway for the reduction of benzoyl-coenzyme A. During biodegradation, phenol (Phenol) can be progressively degraded into phenylacetyl-Coenzyme A (Benzoyl-CoA) through a series of reactions. First, phenol will be oxidized to

resorcinol (Catechol), and after a further degradation reaction, resorcinol will be converted to proaryl succinic acid (Protocatechuate). And the proarylsuccinic acid is eventually degraded to phenylacetyl-CoA. During this degradation, phenylacetyl-CoA may also undergo a reduction reaction, which is reduced to the benzylethanol form. This step is usually catalyzed by a number of reductases, reducing phenylacetyl-CoA to benzyl ethanol, thus completing a loop in the phenol degradation pathway. According to the analysis, it indicates that these eight strains do not have the function of grading aromatic compounds.

Functional pathways of carbon source metabolism in living organisms

According to the figure, none of these nine strains able to metabolize pyruvate and alcohol, Leclercia sp.LTM01、Leclercia .sp EMC7 Unable to perform lactate metabolism. However, these bacteria are connected with each other through other carbon sources, and jointly regulate the utilization of carbon sources in living organisms, providing energy and material basis for life activities. Figure 5-2 shows the P.a. The Hao2018 (Leclercia peneumoniae) carbon cycle path, the marked red path is P.a. The carbon source metabolism of Hao2018 (Leclercia peneumoniae), Figure Figure 5-2 shows P.a. Hao2018 (Leclercia peneumoniae) Only the metabolism of acetate and organic carbon passes through it.

Functional pathways of nitrogen source metabolism in living organisms

All nine strains have functional pathways from nitrate reduction and nitrite to nitrogen, and nitrate reduction and nitrite to ammonia are two important processes in the biogeochemical cycle. In both processes, microbes play a key role in reducing nitrate and nitrite to ammonia through specific metabolic pathways. Nitrate reduction refers to the process of reducing nitrate to nitrite, and then to nitrogen or ammonia. The reduction of nitrite to ammonia is the process of reducing nitrite to ammonia. During nitrate reduction, nitrate reductase is the key enzyme capable of catalyzing the reduction of nitrate to nitrite. The nitrite produced in this process undergoes through another group of enzymes and is eventually reduced to ammonia. In the process of nitrite reduction to ammonia, nitrite reductase plays a vital role, which can directly

reduce nitrite to ammonia and complete the cycle of nitrogen. These two processes not only play an important role in nature, but also maintain the equilibrium cycle of nitrogen elements, but also play an important impact on the environment and organisms. It is through these microbial-mediated reduction reactions that nitrogen can be effectively recycled to maintain the stability of the ecosystem. Therefore, a deep investigation of the molecular mechanisms and regulatory pathways of nitrate reduction and nitrite reduction to ammonia will contribute to a better understanding of biogeochemical cycles. The nitrogen metabolism characteristics of this bacterium can also be translated through this figure. Figure Figure 5-3 shows the P.a. The nitrogen cycle path of Hao2018 (Leclercia peneumoniae), where P.a. Hao2018 (Leclercia peneumoniae) has only one route for nitrogen metabolism.

Functional pathways of sulfur source metabolism in living organisms

Through the whole genome functional heat map, it can be found that none of these nine strains has sulfur source metabolic route, so we can speculate that the lekella species do not have the function of metabolizing sulfur source.

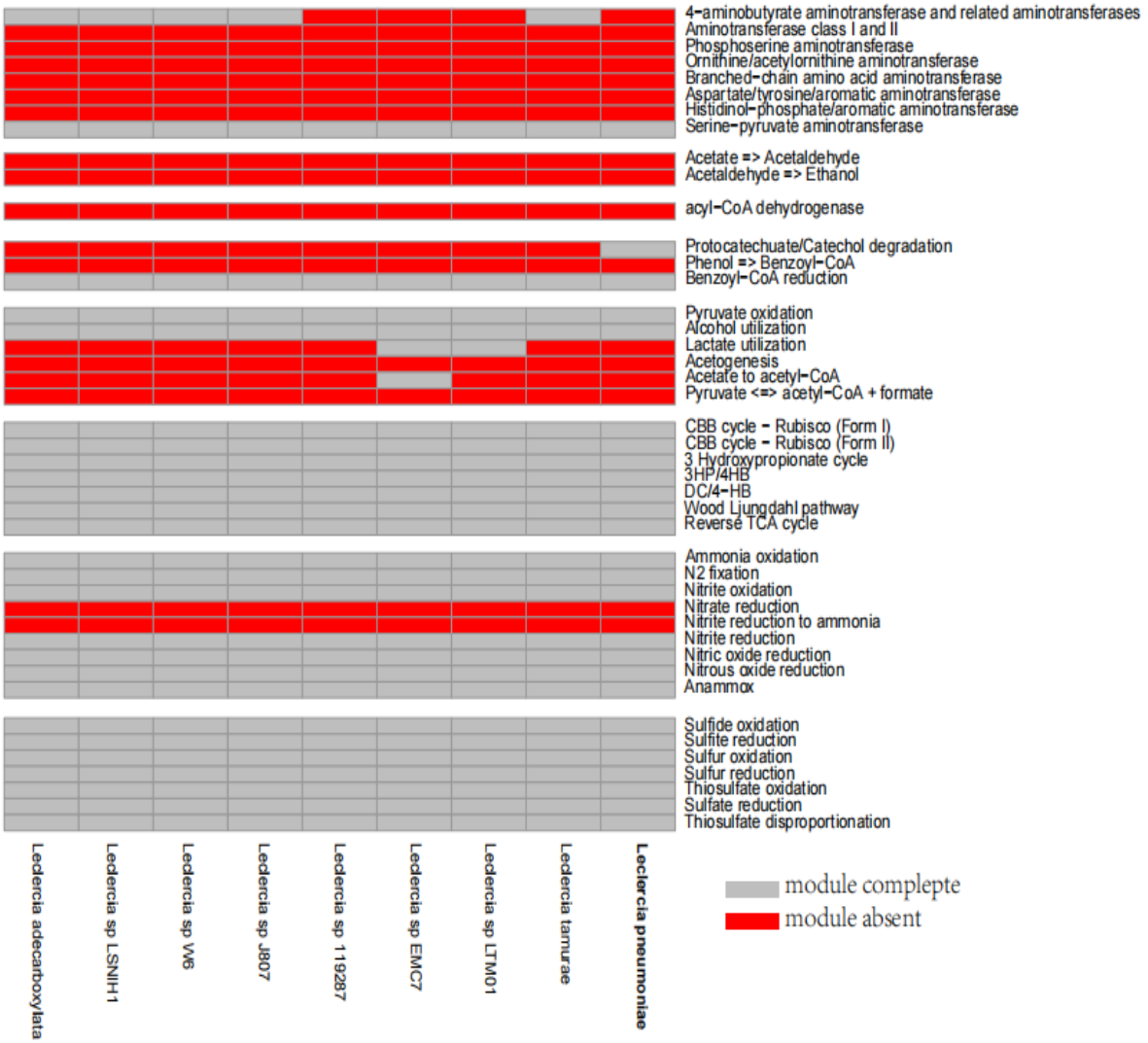


Figure 5-1 Functional heatmap

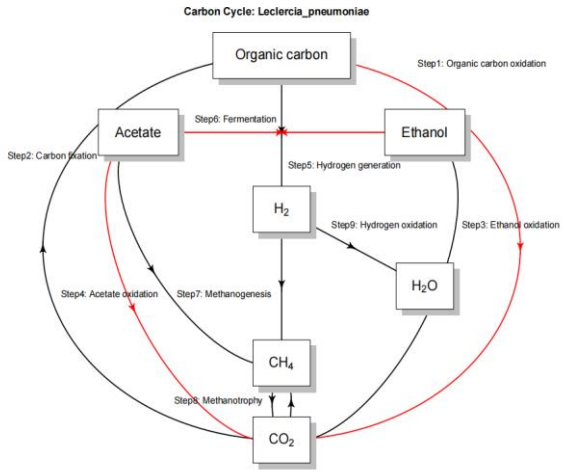


Figure 5-2 The carbon cycle path of the Leclercia pneumoniae

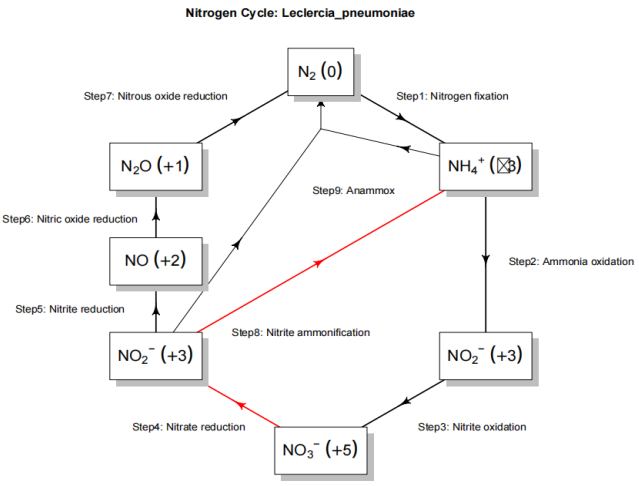


Figure 5-3 The nitrogen cycle paths of the Leclercia pneumoniae

3.9.Common functional characteristics of Lekerella

Microorganisms can resist the toxicity caused by antibiotics in many ways, among which the direct efflux of drug resistance pumps is one of the main resistance mechanisms of bacteria 25. Multidrug resistance efflux pump (Multi-Drug Resistance Efflux Pump) refers to the active membrane transport system of transmembrane proteins that can expel small organic molecules such as drugs or toxic substances such as some heavy metal ions from the intracellular to the extracellular cell membrane 25. According to the analysis of the bacterial functional module data of the nine strains, these nine strains have antibiotic transport system (Lantibiotic transport system), NisK-NisR (lantibiotic biosynthesis) two-component regulatory system and multidrug resistance efflux pump (Multi-Drug Resistance Efflux Pump) genes, which jointly regulate and resist intracellular antibiotics and toxic substances, It is mainly involved in the regulation of antibiotic transfer and excretion. They are able to recognize and transport antibiotic molecules, ensuring their smooth crossing across the cell membrane and eventually their release into the external environment. This transport process is essential for antibiotic resistance. Its presence can reduce the concentration of antibiotics and toxic substances in the bacteria body, and enhance the viability of the bacteria body 26. Among the resistance genes, the rest of these nine strains are common antibiotic resistance genes, against tetracycline and

macrocyclic lipids 27, Aminoglycosides, fluoroquinolones, micoxanamine, cephalosporin, fosfomycin, streptomycin, nitroimidazole, rifamycin and other drugs have a certain tolerance. So we can speculate that lekeria species may be resistant.

CONCLUSIONS AND OUTLOOK

The genus Leckerella is a group of bacteria widely found in nature, usually found in soil, water, animals and plants. Due to the late discovery of Lex, it has been only for 60 years since 1960. However, its discovery and understanding in China is less than 30 years, and its research nature is in a blank stage. A small number of reports mainly focus on pathogenic infections, and few reports in other fields. However, in NCBI GenBank database, there is only one case of the strain submitted in China. The submission of this sequence will enrich the gene pool of Leella.

In this study, through whole genome sequencing, comparative genomics analysis and functional module analysis, we learned about the genome structure and functional characteristics of the bacterium, relatedness and mobile genetic element information, and found high similarity between different strains of the genus Leclercia, including P.a. The similarity coefficient of Hao2018 (Leclercia peneumoniae) and other strains of genus Leclercia is about 84.5%. The comparative analysis of gene families revealed that P.a. Hao2018 Shared most of the gene families with other Leclercia strains and almost no specific gene families, showing high relatedness. Furthermore, as found in the insertion element analysis of the phage whole, P.a. The presence of two prophages, containing 11 CDs, in the Hao2018 chromosome is closely associated with its pathogenicity and antimicrobial properties. It is noteworthy that lekeria is also resistant, as a pathogenic bacterium, which would be detrimental to our treatment, and that these resistance genes, probably result from phage horizontal gene transfer. A thorough study of the process will help us to control the bacterium. Overall, the genus Leckiella, as one of the important microbial communities in the ecosystem, as a pathogenic bacterium, is drug-resistant and may adapt to the host microenvironment in various ways, and the bacterium also plays an important role in geochemical cycling.

Through the P.a. Hao2018 (Leclercia peneumoniae) whole genome sequencing and analysis also provide basic data support for the subsequent understanding and study of the diversity, environmental impact, material degradation, and infection prevention and control of Lekeria.

REFERENCE DOCUMENTATION

1. Fuhrman, J.A., Cram, J.A.& Needham, D.M.Marine microbial community dynamics and their ecological interpretation.Nat.Rev.Microbiol.13, 133‑146 (2015).
2. Gibbons, S.M.& Gilbert,J.A.Microbial diversity−exploration of natural ecosystems and microbiomes.Curr.Opin.Genet.Dev 35, 66‑72 (2015).
3. Kai, Shan, Chunlei, Wang, Wenlin, & Liu, et al.(2019).Genome sequence and transcriptomic profiles of a marine bacterium, pseudoalteromonas agarivorans hao 2018. Scientific Data.
4. Marshal Mu, Lu Desen, Zheng Weishuang, etc. Progress in new species identification and resource development of Marine bacteria in China [J]. Biological Resources, 2017,39 (06): 391-397.
5. Li Jianhong. Research on Bacillus mycoides Gnyt1 characteristics and whole genome sequencing of rhizosphere probacteria in elite plants [D]. Gansu Agricultural University, 2017.
6. Schubert M, Lindgreen S, Orlando L.AdapterRemoval v2: rapid adapter trimming, identification, and read merging[J].BMC research notes, 2016, 9(1): 88.
7. Luo R, Liu B, Xie Y, et al.SOAPdenovo2: an empirically improved memory-efficient short read de novo assembler[J].Gigascience, 2012, 1(1):18.
8. Coil D, Jospin G, Darling A E.A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data[J].Bioinformatics, 2014, 31(4): 587-589.
9. Koren S, Walenz BP, et al.Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.Genome Res,2017,27(5), 722-736.

10. Walker B J, Abeel T, Shea T, et al.Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement[J].PloS one, 2014, 9(11): e112963.
11. Kurtz S, Phillippy A, Delcher AL, et al.Versatile and open software for comparing large genomes[J].Genome Biology, 2004, 5(2): R12
12. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: Improved microbial gene identification with GLIMMER.Nucleic acids research 1999, 27(23):4636-4641.
13. Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldon T, Rattei T, Creevey C, Kuhn M et al: eggNOG v4.0: nested orthology inference across 3686 organisms.Nucleic acids research 2014, 42(Database issue):D231-239.
14. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: KAAS: an automatic genome annotation and pathway reconstruction server.Nucleic acids research 2007, 35(Web Server issue):W182-185.
15. Li L, Stoeckert C J, Roos D S.OrthoMCL: identification of ortholog groups for eukaryotic genomes[J].Genome research, 2003, 13(9): 2178-2189.
16. Lee I, Kim Y O, Park S C, et al.OrthoANI: an improved algorithm and software for calcμlating average nucleotide identity[J].International journal of systematic and evolutionary microbiology, 2016, 66(2): 1100-1103.
17. Yoon S H, Ha S M, Kwon S, et al.Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies[J].International journal of systematic and evolutionary microbiology, 2017, 67(5): 1613-1617.
18. Talavera G, Castresana J.Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments[J].Systematic biology, 2007, 56(4): 564-577.
19. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: PHAST: a fast phage search tool.Nucleic acids research 2011, 39(Web Server issue):W347-352.

20. Ding Y, Wet J R, Cavalcoli J, et al.Genome-based characterization of two prenylation steps in the assembly of the stephacidin and notoamide anticancer agents in a marine-derived Aspergillus sp[J].Journal of the American Chemical Society, 2010, 132(36): 12733-12740.
21. Lee I, Kim Y O, Park S C, et al.OrthoANI: an improved algorithm and software for calcμlating average nucleotide identity[J].International journal of systematic and evolutionary microbiology, 2016, 66(2): 1100-1103.
22. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS.PHAST: a fast phage search tool[J].Nucleic Acids Res, 2011,39(Web Server issue): W347-352.
23. Krupovic M, Prangishvili D, Hendrix RW, Bamford DH.Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere[J].Microbiol Mol Biol Rev, 2011,75(4): 610635.
24. Casjens S.Prophages and bacterial genomics: what have we learned so far?[J].Mol Microbiol,2003,49(2): 277-300.
25. Liu alcohol, Wang Dan, He Yafeng, et al. A multidrug-resistant efflux pump of bacteria [J]. Pharmaceutical Biotech, 2013,20 (02): 175-178.
26. Wu Shengliang, Ye LAN, Zhang Tongtong, et al. Whole genome analysis of a non-decarboxylated YZ 30 degrading catechol [J]. Pharmaceutical Biotech, 2022,29 (01): 8-15.
27. Gu Juepen, Gao Jie. Progress in macrolide resistance genes and new drugs [J]. Pharmaceutical Biotech, 2001 (02): 112-116.

EXPRESS ONE'S THANKS

At the end of the paper, the thank speech is a brief message to express their gratitude to the personnel who have given direct help in the research and writing process, such as instructors and other personnel. This is not only a kind of politeness, but also a respect for the work of others, is the governance scholar should think

Conclusions and Outlook

The genus Leckerella is a group of bacteria widely found in nature, usually found in soil, water, animals and plants. Due to the late discovery of Lex, it has been only for 60 years since 1960. However, its discovery and understanding in China is less than 30 years, and its research nature is in a blank stage. A small number of reports mainly focus on pathogenic infections, and few reports in other fields. However, in NCBI GenBank database, there is only one case of the strain submitted in China. The submission of this sequence will enrich the gene pool of Leella.

In this study, through whole genome sequencing, comparative genomics analysis and functional module analysis, we learned about the genome structure and functional characteristics of the bacterium, relatedness and mobile genetic element information, and found high similarity between different strains of the genus Leclercia, including P.a. The similarity coefficient of Hao2018 (Leclercia peneumoniae) and other strains of genus Leclercia is about 84.5%. The comparative analysis of gene families revealed that P.a. Hao2018 Shared most of the gene families with other Leclercia strains and almost no specific gene families, showing high relatedness. Furthermore, as found in the insertion element analysis of the phage whole, P.a. The presence of two prophages, containing 11 CDs, in the Hao2018 chromosome is closely associated with its pathogenicity and antimicrobial properties. It is noteworthy that lekeria is also resistant, as a pathogenic bacterium, which would be detrimental to our treatment, and that these resistance genes, probably result from

phage horizontal gene transfer. A thorough study of the process will help us to control the bacterium. Overall, the genus Leckiella, as one of the important microbial communities in the ecosystem, as a pathogenic bacterium, is drug-resistant and may adapt to the host microenvironment in various ways, and the bacterium also plays an important role in geochemical cycling.

Through the P.a. Hao2018 (Leclercia peneumoniae) whole genome sequencing and analysis also provide basic data support for the subsequent understanding and study of the diversity, environmental impact, material degradation, and infection prevention and control of Lekeria.