Hybrid de novo Genome Assembly of Erwinia sp. E602 and Bioinformatic Analysis Characterized a New Plasmid-Borne lac Operon Under Positive Selection

Yu Xia1,2,3*, Zhi-Yuan Wei4†, Rui He1,2,3, Jia-Huan Li4, Zhi-Xin Wang4, Jun-Da Huo4 and Jian-Huan Chen4*

1 State Key Laboratory of Food Science and Technology, Jiangnan University, Wuxi, China, 2 School of Food Science and Technology, Jiangnan University, Wuxi, China, 3 Collaborative Innovation Center of Food Safety and Quality Control in Jiangsu Province, Jiangnan University, Wuxi, China, 4 Laboratory of Genomic and Precision Medicine, Wuxi School of Medicine, Jiangnan University, Wuxi, China

Our previous study identified a new β-galactosidase in Erwinia sp. E602. To further understand the lactose metabolism in this strain, de novo genome assembly was conducted by using a strategy combining Illumina and PacBio sequencing technology. The whole genome of Erwinia sp. E602 includes a 4.8 Mb chromosome and a 326 kb large plasmid. A total of 4,739 genes, including 4,543 protein-coding genes, 25 rRNAs, 82 tRNAs and 7 other ncRNAs genes were annotated. The plasmid was the largest one characterized in genus Erwinia by far, and it contained a number of genes and pathways responsible for lactose metabolism and regulation. Moreover, a new plasmid-borne lac operon that lacked a typical β-galactoside transacetylase (lacA) gene was identified in the strain. Phylogenetic analysis showed that the genes lacY and lacZ in the operon were under positive selection, indicating the adaptation of lactose metabolism to the environment in Erwinia sp. E602. Our current study demonstrated that the hybrid de novo genome assembly using Illumina and PacBio sequencing technologies, as well as the metabolic pathway analysis, provided a useful strategy for better understanding of the evolution of undiscovered microbial species or strains.

Keywords: Erwinia, hybrid sequencing, genome assembly, lac operon, bioinformatic analysis

INTRODUCTION

Erwinia is a group of the straight rod-shaped, facultative anaerobic, gram-negative bacterium of the Erwiniaeeae family of Enterobacteriaceae. Most of the Erwinia species identified by far are pathogens, saprophytes, or epiphytes of plants. It has been reported that some Erwinia species ferment lactose as a carbon source. Our previous study characterized a β-galactosidase with relatively high activity at low temperature in the Erwinia sp. strain E602 (Xia et al., 2018).

The lac operons typically involve three genes encoding the enzymes that enable bacteria to utilize lactose (Diaz-Hernandez and Santillan, 2010). The gene lacZ encodes the β-galactosidase, an enzyme that degrades lactose into monosaccharides glucose and galactose. Similarly, lacY encodes
a membrane-embedded transporter that helps bring lactose into cells. The gene lacA encodes galactoside O-acetyltransferase that catalyzes the transfer of an acetyl group from acetyl-CoA to the 6-hydroxyl of galactopyranosides, with its exact physiological function remaining unclear. In addition, the gene lacI encodes the lac repressor, which is a protein that represses the transcription of lac operon genes by binding to the promoter and preventing the operon from transcription when lactose is not available. In the presence of lactose, the lac repressor is released from the operon to allow RNA transcription (Marbach and Bettenbrock, 2012). Notably, the lac operon confers a competitive advantage for bacterial survival in the environments (Pinto et al., 2021). Therefore, the study of lac operon in the newly identified strain Erwinia sp. E602 may improve our understanding of the role of lactose metabolism in this genus.

High-throughput sequencing and de novo assembly allow obtaining the total genetic information of microbes. It is widely used to dissect the genome sequence, gene composition, and evolutionary features of novel or unknown species. Notably, long-read sequencing technologies, such as PacBio (Rhoads and Au, 2015; Zhang et al., 2021), have dramatically speeded up the understanding of microorganisms. Moreover, the whole genome sequence of microorganisms provides an important basis for post-genomics research of microorganisms, such as subsequent gene mining and functional validation. In order to further investigate the regulation mechanisms of the β-galactosidase expression in the strain Erwinia sp. E602, we sequenced its genome and conducted de novo assembly, using the hybrid Illumina and PacBio sequencing (De Maio et al., 2019). From the assembly of the Erwinia sp. E602 genome, we identified a large plasmid and a new plasmid-borne lac operon containing a lacZ gene encoding the β-galactosidase with low-temperature activity, as well as evidence supporting positive selection pressure of the lac operon.

RESULTS

De novo Assembly and Annotation of Erwinia sp. E602 Genome

As summarized in Supplementary Tables 1–3, the Illumina and PacBio reads were used to conduct de novo genome assembly. After hybrid de novo assembly using SPAdes, two circular scaffolds were obtained, including a chromosome sequence of 4,842,717 bp with a GC content of 58.91% (GenBank: CP046582.1) and a large plasmid sequence of 325,969 bp (GenBank: CP046581.1). A total of 4,739 genes, including 4,543 protein-coding, 25 rRNAs, 82 tRNAs, and 7 other ncRNAs genes were annotated as shown in Table 1 and Figures 1A,B. Among these genes, 4,472 genes including 4,296 protein-coding, 25 rRNAs, 82 tRNAs, and 7 other ncRNAs genes were located on the chromosome, while 277 protein-coding genes were located on the large plasmid.

EggNOGv5.0 was then used to predict the functions of these genes. The KEGG pathway enrichment analysis was also performed (Supplementary Material 1). Pathways with a number of genes greater than 5 were shown, and finally, 122 enriched pathways were obtained. The most-enriched pathways (gene counts) included biosynthesis of secondary metabolites (357), microbial metabolism in diverse environments (260), ABC transporters (204), biosynthesis of cofactors (156), biosynthesis of amino acids (132), two-component systems (115), carbon metabolism (98), purine metabolism (71), quorum sensing (63), pyrimidine metabolism (51), ribosome (51), cysteine and methionine metabolism (50), bacterial secretion system (48), and amino sugar and nucleotide sugar metabolism (47). By dbCAN2 database annotation (Zhang et al., 2018), 593 carbohydrates were found on the chromosome and 51 on the plasmid (Supplementary Material 1). Through Resfams database annotation (Gibson et al., 2015), we found 211 resistance genes on the chromosome and 21 resistance genes on the plasmid, respectively (Supplementary Material 1).

New Plasmid-Borne lac Operon in Erwinia sp. E602

From the gene annotation results of de novo assembly, Erwinia sp. E602 was found to contain a number of genes and pathways for lactose metabolism and regulation (Figure 2A). A list of lactose/galactose-related functional genes were shown in Table 2.

Our previous study characterized a new β-galactosidase that exhibited activity at low temperatures, indicating its potential unique biological function (Xia et al., 2018). Intriguingly, the de novo assembly results showed that the exact lacZ gene encoding the new β-galactosidase was not located on the chromosome of Erwinia sp. E602. Instead, further analysis of the sequence context surrounding the lacZ gene found that the gene was a component of a new lac operon on the large plasmid. The lac operons are known to be composed of regulatory genes, manipulated regions, and structural genes (Diaz-Hernandez and Santillan, 2010). The newly identified lac operon in the current study contained lacI, lacZ, lacY, as well as a regulatory element-CAP and a promoter, between the location of lacI and lacZ (Figure 2B). BLAST results of the plasmid-borne lac operon showed that it shares low sequence similarity with other sequences in other Erwinia strains. These results are included in Supplementary Materials. Moreover, this newly discovered lac operon, compared with the classical ones, lacked a lacA gene but had a reversed lacI-like gene encoding protein WP_212408226. Further comparison of the related genes revealed that although other lacZ, lacI and

| Type       | Chromosome | Plasmid |
|------------|------------|---------|
| RefSeq     | NZ_CP046582.1 | NZ_CP046581.1 |
| INSDC      | CP046582.1 | CP046581.1 |
| Size (Mb)  | 4.84 | 0.33 |
| GC%        | 58.9 | 59.0 |
| Protein    | 4,296 | 277 |
| rRNA       | 25 | – |
| tRNA       | 82 | – |
| Other RNA  | 7 | – |
| Gene       | 4,472 | 82 |
| Pseudogene | 62 | 5 |

TABLE 1 | Summary of Erwinia sp. E602 genome annotation.
lacY paralogs were also observed in the genome of *Erwinia* sp. E602, they did not form a canonical lac operon due to the lack of essential structure. Moreover, the *lacI* and *lacY* genes of the plasmid-borne lac operon share low similarity with their homologs in the genomes of *Erwinia* species (Table 3), suggesting that this lac operon was distinct from canonical ones. In addition, the NCBI BLAST search using the nucleic acid database was performed to identify potentially similar operon-related genes, and the results showed that the genes in this operon were more similar to those in genus *Citrobacter* (Figure 3). Likewise, the
lacZ gene of the plasmid-borne lac operon shared a relatively low identity with its paralogs in the other Erwinia species (Table 4). Instead, its lacZ and lacY had high similarities with their homologs in Citrobacter species. The genes of lacI and lacI-like protein (WP_212408226) were compared using the NCBI nucleic acid database, and no sequences with similarities greater than 80% were found. In addition, no lacA was found either on the chromosome of Erwinia sp. E602, or on its large plasmid.

**Selective Pressure in Genes in the Newly Identified lac Operon**

To better understand whether genes in the plasmid-borne lac operon are subject to natural selection during adaptation to the environment, positive selection sites were identified using branch-site models (model A vs. model A null) and site models (M1a vs. M2a; M7 vs. M8) implemented in Phylogenetic Analysis by Maximum Likelihood (PAML) (20). Due to fewer sequences similar to genes lacI and lacI-like protein (WP_212408226) in other bacteria strains, we chose the top 100 sequences ranked by identity with lacY or lacZ to be included in the two datasets separately for selection pressure analysis. The results of the branch-site model comparison showed that the gene lacY had a site at codon 187 under positive selection (Table 5). Surprisingly, there are multiple sites under strong positive selection in the gene lacZ, indicating the strong positive selection pressure in it (Table 6).

**DISCUSSION**

By using a strategy combining Illumina and PacBio sequencing technology, the de novo assembly of the whole genome of Erwinia sp. E602 was performed. Moreover, our study characterized the largest plasmid in the genus Erwinia by far, and reported a new plasmid-borne lac operon.
From the assembly, a large plasmid with a length of 325,969 bp and a total of 297 genes was characterized. Before our current study, 6 plasmids had been reported in the genus Erwinia, among which the largest plasmid was CP037949 derived from Erwinia sp. QL-Z3. That plasmid was 149,889 bp in length and encoded a total of 124 genes. Therefore, the plasmid characterized in our current de novo genome assembly of Erwinia sp. E602 is the largest plasmid found in the genus Erwinia by far. In general, the genes in plasmids might provide microorganisms with potential genetic advantages. Yet the specific role of the large plasmid in Erwinia sp. E602 still remains unclear.

The lac operon is a hallmark gene of the regulatory circuit for bacteria to regulate metabolism according to nutrient conditions in the environment (Leonard et al., 2015; Malakar, 2015; Karkare et al., 2021; Pinto et al., 2021). Through genome function annotation, a large number of genes related to lactose metabolism were found in Erwinia sp. E602, a cold-adapted strain. In particular, a new plasmid-borne lac operon was characterized in this study. The lac operon shared low similarity with homologous genes in Erwinia species. It lacked a typical lacA. Instead, the position of β-galactoside transacetylase was replaced by a lacI-like gene encoding uncharacterized function. The gene lacZ in the newly identified lac operon encodes a β-galactosidase with the characteristics of low-temperature adaptation (Xia et al., 2018), which might help the organism survive in extreme environments. Similarly, identification and isolation of a 127-kb large plasmid in a rat E. coli isolate (EC93), containing the cd1l gene, would show higher toxic potency, thus helping it to exert a competitive advantage (Waneskog et al., 2021). Plasmids are mobile parts of bacterial genomes that carry genes that benefit the survival of the organism and confer selective advantages such as antibiotic resistance, thus helping bacterial communities to quickly adapt to the environment (Bogomazova et al., 2020). Similarly, the large plasmid we found in Erwinia sp. E602 also contains a group of I-toxin-antitoxin modules, mainly including wagC, vapC, relE, stbD, yefM, and other related genes.

The plasmid-borne lac operon might enable Erwinia sp. E602 to utilize substrates from environmental resources for lactose metabolism, thereby gaining an adaptive advantage over competitors. Similar functions have been found in E. coli (Harwani, 2014). Nevertheless, our analysis showed that the lac operon in Erwinia sp. E602 had low sequence similarity to that of E. coli and was evolutionarily distant from its E. coli counterparts. Likewise, other studies reported the lac operons in specific strains (Vaughan et al., 1998; Bidart et al., 2018), which also suggested that the lac operon could vary in different habitats. Furthermore, the lacZ and lacY genes on the lac operon in Erwinia sp. E602 were under positive selection, which might benefit the bacterium in terms of survival or competition (Bundalovic-Torma et al., 2020). Meanwhile, the strain Erwinia sp. E602 was isolated from the high latitudes of northeast China and the enzyme encoded by the plasmid-borne lacZ gene exhibited good activity at low temperatures. These findings might be in line with the possible adaptation of this Erwinia strain to the cold environment.

The plasmid-borne lac operon lacked the lacA gene in Erwinia sp. E602. It has been reported that the transacetylase encoded by the lacA gene is not an essential element for lactose catabolism (Lagesen et al., 2007). In contrast, a lacI-like gene was found in the opposite strand downstream of the lac operon in Erwinia sp. E602. It remains to be elucidated whether the lacI-like gene contributes to the regulation of the lac operon.

In the current study, by combining Illumina and PacBio sequencing technologies, a hybrid de novo assembly was performed to obtain a more accurate and complete map of the genome of Erwinia sp. E602, providing a reference basis for subsequent transcriptome studies. Furthermore, a new plasmid-borne lac operon was characterized with evidence supporting evolution and adaptation to the environment of the genus Erwinia. The specific role of the large plasmid on Erwinia sp. E602 needs further studies.

### MATERIALS AND METHODS

#### Bacterial Culture and DNA Extraction

Bacterial culture and DNA extraction of Erwinia sp. E602 were performed as described in our previous report (Xia et al., 2018). The genomic DNA was used for subsequent Illumina and PacBio sequencing.

#### Illumina Sequencing and Data Processing

The library for short read sequencing was prepared from 1 μg Erwinia sp. E602 genomic DNA using NEBNext Ultra DNA Library Prep Kit (NEB, Ipswich, Massachusetts) according to the manufacturer’s instructions, assessed using a Qubit® 2.0 Fluorometer (Invitrogen, Waltham, Massachusetts) and an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, California) for quality control (QC), and sequenced on the Hi-Seq 2500 platform (Illumina, San Diego, California) using a 100 bp paired-end configure. All adapter sequences and low-quality bases contained in the reads were filtered and removed using Trimomatic (version 0.31), to keep only reads with sequencing read length more than 90 nucleotides (nt), and ensure the proportion of base quality greater than 20 (Q20) was not less than 90%.

---

**TABLE 3** | Sequence similarity of lac operon related genes.

| Title | Query coverage | E-value | Per. ident |
|-------|----------------|---------|------------|
| plasmid_238-lacI vs. genome_2601-lacI | 41% | 2.00 E-38 | 72.04% |
| plasmid_238-lacI vs. plasmid_241_wp | 4% | 7.00 E-06 | 89.29% |
| genome_2601-lacI vs. plasmid_241_wp | 3% | 0.004 | 100.00% |
| plasmid_240-lacY vs. genome_2604-bglY | No sig |
| plasmid_240-lacY vs. genome_3788-lacY | 94% | 5.00 E-161 | 70.88% |
| genome_2604-bglY vs. genome_3788-lacY | No sig |
| plasmid_239-lacZ vs. plasmid_250-lacZ3 | No sig |
| plasmid_239-lacZ vs. plasmid_240-lacY | 0 | 0.039 | 100.00% |

1. https://www.ncbi.nlm.nih.gov/genome/browse/#/plasmids/13524/
PacBio Sequencing and Data Processing

The library for single-molecule long read sequencing was also prepared from 5 µg genomic DNA using SMRTbell DNA Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, California) according to the manufacturer's instructions, assessed using a Qubit® 2.0 Fluorometer and an Agilent 2100 Bioanalyzer for quality control (QC), and sequenced on the PacBio RS System. Raw data was filtered to obtain clean data with a cutoff of minimum polymerase read quality 0.75 and minimum polymerase read length 3,500 bp.

Hybrid Assembly of Illumina and PacBio Sequencing Data

The filtered Illumina reads and PacBio subreads were then used to conduct the hybrid assembly using SPAdes software (v3.5.0) (Bankevich et al., 2012). PacBio subreads were provided...
**Gene Prediction**

The software Prokka (Seemann, 2014) (v 1.14.6) was used to predict genes in the assembly results. Carbohydrate enzymes were annotated using dbCAN2 (Zhang et al., 2018). Resistance genes were annotated using the Resfams database (Gibson et al., 2015). The prediction of rRNA was performed using rnammer1.2 (Lagesen et al., 2007). The prediction of tRNA was performed using tRNAscan-SE-1.23 (Chan and Lowe, 2019). The sRNA annotation was obtained by the software infernal-1.1rc4 (Nawrocik and Eddy, 2013). The CGView server (Stothard and Wishart, 2005) was used as an interactive comparison genomics tool to draw the circos plots.

**Functional Annotation and Pathway Enrichment Analysis**

EggNOG v5.0 (Huerta-Cepas et al., 2019) were used for the functionally annotation of genes. Pathway enrichment analysis was performed using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and KEGG Orthology (KO) databases.

Table 6 shows the positively selected sites in LacZ of Erwinia sp. E602 identified with site model M2a using Bayes Empirical Bayes analysis.

**Phylogenetic Analysis**

Phylogenetic analysis was performed using software MEGA7 (Kumar et al., 2016) using the Neighbor-Joining method. The 1,000 replicates of bootstrap test were used to evaluate the clustering of taxa at branch nodes.

**Select Pressure Assessment**

In order to analyze the selection pressure of the related genes, PAML (v4.1) was used to identify sites under positive selection in the genes with site and branch-site models (Yang, 2007). For the branch-site models, Erwinia sp. E602 was selected as the foreground branch of the branch-site model, and other bacteria were used as the background branch.

**DATA AVAILABILITY STATEMENT**

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

**AUTHOR CONTRIBUTIONS**

YX and J-HC conceived the project, planned the experiments, and participated in the results discussions. Z-YW performed the bioinformatic analysis and improved the data analysis pipelines. RH participated in the DNA extraction from bacteria. J-HL,

---

**TABLE 4** | Identity of plasmid-borne LacZ in Erwinia sp. E602 with its paralogs in other Erwinia species.

| Erwinia strains | Chromosome/Plasmid | Counts | GI (Identity%) |
|-----------------|---------------------|--------|---------------|
| Erwinia amylovora CFBP1430 | Chromosome | 3 | GI:490258078 (13%); GI:490258079 (13%); GI:490258942 (61%) |
| Erwinia billingiae Eb661 | Chromosome | 2 | GI:502966002 (65%); GI:5029665951 (16%) |
| Erwinia gerundensis E_g_EM595 | Chromosome | 2 | GI:1055842378 (64%); GI:1055871078 (16%) |
| Erwinia gerundensis E_g_EM596 | Plasmid pEM01 | 1 | GI:1055874842 (59%); GI:1055874917 (11%) |
| Erwinia sp. J780 | Chromosome | 2 | GI:1783198844 (15%); GI:1783199144 (6%) |
| Erwinia sp. QL-Z3 | Chromosome | 2 | GI:1608688023 (16%); GI:1828892330 (65%) |
| Erwinia tasmaniensis E1/99 | Chromosome | 1 | GI:501410713 (62%) |

**TABLE 5** | Bayes Empirical Bayes analysis of positively selected sites identified in lacY of Erwinia sp. E602 with the branch-site model A.

| Model | Codon | Amino acid | Probability | Post mean |
|-------|-------|------------|-------------|-----------|
| Branch-site model | 75 | T | 0.663 | |
| | 147 | G | 0.564 | |
| | 187 | G | 0.956* | |
| | 191 | T | 0.748 | |
| | 207 | S | 0.801 | |

*Posterior probability from BEB analysis > 95%; ratio for foreground branch (ω > 1).
Z-XW, and J-DH participated in the data analysis. All authors contributed to the final manuscript.

**FUNDING**

This work was supported by the Jiangsu Agriculture Science and Technology Innovation Fund (JASTIF) with the No. CX(19)3109, the National First-Class Discipline Program of Light Industry Technology and Engineering (LITE2018-14), Taihu Lake Talent Plan, and the Postgraduate Training Innovation Project of Jiangnan University: Project No. JNKY19072.

**REFERENCES**

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bidart, G. N., Rodriguez-Diaz, J., Perez-Martinez, G., and Yebra, M. J. (2018). Lactobacillus casei is involved in the transport and metabolism of the human milk oligosaccharide core-2 N-acetyllactosamine. *Sci. Rep.* 8:7152. doi: 10.1038/s41598-018-25660-w

Bogomazova, A. N., Gordeeva, V. D., Krylova, E. V., Soltynskaya, I. V., Davydova, E. E., Ivanova, O. E., et al. (2020). Mega-plasmid found worldwide confers multiple antimicrobial resistance in *Salmonella* Infantis of broiler origin in Russia. *Int. J. Food. Microbiol.* 319:108497.

Bundalovic-Torma, C., Whitfield, G. B., Marmont, L. S., Howell, P. L., and Parkinson, J. (2020). A systematic pipeline for classifying bacterial operons reveals the evolutionary landscape of biofilm machineries. *PLoS Comput. Biol.* 16:e1007721. doi: 10.1371/journal.pcbi.1007721

Chan, P. P., and Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* 1962, 1–14.

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., et al. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genom.* 5:294. doi: 10.1099/mgen.0.000294

Diaz-Hernandez, O., and Santillan, M. (2010). Bistable behavior of the lac operon in *E. coli* when induced with a mixture of lactose and TMG. *Front. Physiol.* 1:22.

Gibson, M. K., Forsberg, K. J., and Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* 9, 207–216. doi: 10.1038/ismej.2014.106

Harwani, D. (2014). Regulation of gene expression: Cryptic-glycose (bgII) operon of *Escherichia coli* as a paradigm. *Braz. J. Microbiol.* 45, 1139–1144.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Karkare, K., Lai, H. Y., Azevedo, R. B. R., and Cooper, T. F. (2021). Historical potential conflict of interest. *The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.*

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Xia, Wei, He, Li, Wang, Hua and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.