Data in Brief 31 (2020) 105697

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Genome data of shrimp acute hepatopancreatic necrosis disease causative *Vibrio parahaemolyticus* strains isolated from South Korea aquaculture farms

Seong Don Hwang[a], Jee Youn Hwang[a], Saebom Sohn[b], Su Mi Kim[c], Seok Ryel Kim[c], Kwang Il Kim[d], Mun-Gyeong Kwon[a], Myunghee Jung[e,f], Kyoo-yeol Lee[e], Byeong-chul Kang[e,*], Jung Soo Seo[a,*]

[a] *Aquatic Disease Control Division, National Institute of Fisheries Science, Busan 46083, Republic of Korea*
[b] *Aquaculture Industry Division, East Sea Fisheries Research Institute, Gangneung, Gangwon-do 25435, Republic of Korea*
[c] *Aquaculture Industry Division, West Sea Fisheries Research Institute, Inchon 22383, Republic of Korea*
[d] *Pathology Division, National Institute of Fisheries Science, Busan 46083, Republic of Korea*
[e] *Data Science Center, Insilicogen Inc., Yongin-si 16954, Gyeonggi-do, Republic of Korea*
[f] *Department of Forest Science, College of Agriculture and Life Science, Seoul National University, Seoul 151-921, Korea*

ARTICLE INFO

Article history:
Received 3 February 2020
Revised 23 April 2020
Accepted 4 May 2020
Available online 13 May 2020

Keywords:
Litopenaeus vannamei
Vibrio parahaemolyticus
Genome
Plamid
Toxins

ABSTRACT

The *Vibrio parahaemolyticus* is a gram-negative bacterium, which is responsible for acute hepatopancreatic necrosis disease (AHPND) in shrimp and has various virulent factors. So, to intensify the knowledge on pathogenic mechanism, the heterogeneous *V.parahaemolyticus* strains genome are indeed. Here, genome of seven *V.parahaemolyticus* strains, which are virulent to shrimps were sequenced by PacBio platform and the virulence was confirmed through the presence of

* Corresponding authors.
E-mail addresses: bckang@insilicogen.com (B.-c. Kang), jsseosoo@korea.kr (J.S. Seo).

https://doi.org/10.1016/j.dib.2020.105697

plasmid (~69 Kb) with binary toxin genes (i.e., pirA and pirB) with PCR method.

Specifications Table

Subject	Parasitology
Specific subject area	Microbiology, Genomics
Type of data	Genome sequence, predicted genes and virulent plasmid of seven *Vibrio parahaemolyticus* strains.
How data were acquired	Complete DNA from each strain was sequenced with PacBio sequel system.
Data format	Assembled contigs and plasmids in fasta and the predicted genes co-ordinates are in gff3 file formats.
Parameters for data collection	Genomic DNA from pure culture.
Description of data collection	*Vibrio parahaemolyticus* strains were collected from hepatopancreas of infected shrimps (*Litopenaeus vannamei)* and water samples of the respective aquaculture ponds. The samples were inoculated on the thiosulfate citrate bile salt sucrose plates and incubated at 25°C for 24h.
Data source location	The shrimps (*Litopenaeus vannamei)* and water samples were collected from the aquaculture farms located in Incheon, Pyeongtaek, Gyeonggi-do, Shinan, and Jeollanam-do of Korea in 2016 of South Korea, at 2016, which were experienced massive death.
Data accessibility	This Whole Genome Shotgun project has been deposited at GenBank (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA482034) under the accession QPQB00000000 - QPQH00000000.

Value of the Data

- Draft genome of each strain, could be an additional data set to assess the virulence level upon the genetic variation encoded in plasmid.
- These pandemic strains genome can be an additional dataset to characterize the diversity of *Vibrio parahaemolyticus* virulence potential parameters, which aid to estimate/prevent the mortality rate in shrimp farms.
- These genomes could be valuable resource to conduct the metagenome/comparative genomic analysis among *Vibrio parahaemolyticus* strains.

1. Data

In this article, the Table 1, explains the summary of seven *Vibrio parahaemolyticus* shrimp pathogen strains (SC4, SK, SM3, SM4, WC15, WS, WY3) genome, which are sequenced by pacific biosciences (PacBio) and the summary of sequenced bases, which are assembled into chromosomes and plasmids. Further, the summary of structural annotations such as rRNA, tRNA and genes were also given in Table 1. In Table 2, along with the sequenced strains genome and the details of other twenty genomes, which were collected from the GenBank to assess the functional group of our sequenced genomes. These twenty genomes are grouped into four major groups, i.e., AHPND (Shrimp+/Human-), Pathogenic (Shrimp+/ Human+), Non-pandemic (Shrimp+/Human-), and Non-pathogenic (Shrimp-/Human-). Fig. 1, is the conformation of PCR gel image for the binary toxin genes (PirA and PirB) of AHPND, which explains the shrimp pathogenicity property of sequenced strains. Fig. 2, illustrate the comparative heap-map representation of other major virulent genes of *Vibrio parahaemolyticus* strains from the sequenced genomes along with the genomes selected from GenBank. Fig. 3, expains the phylogenetic

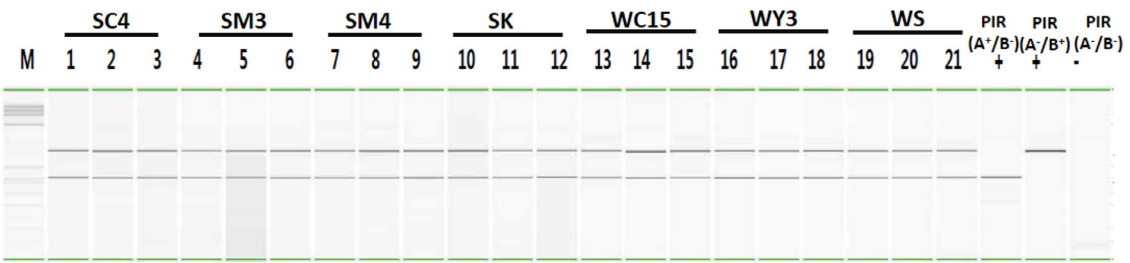

Fig. 1. PCR based virulence conformation of sequenced *Vibrio parahaemolyticus* strains with the universal markers (PirA and PirB).

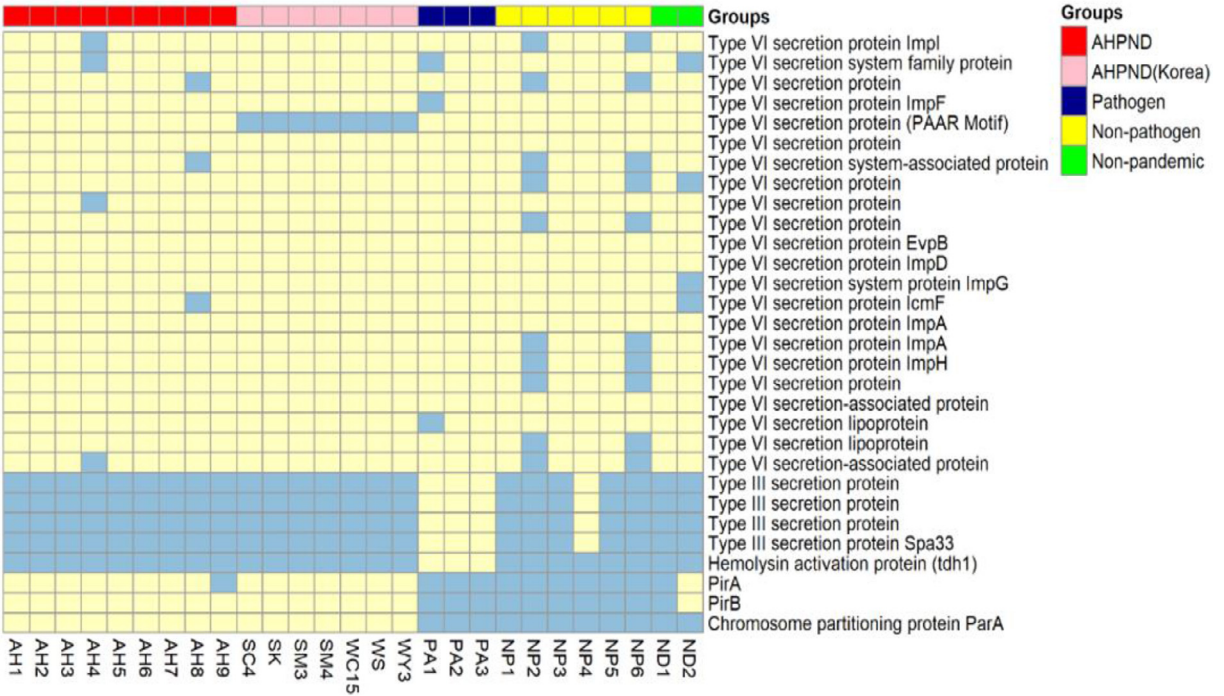

Fig. 2. Profile of virulent associated genes from the *Vibrio parahaemolyticus* sequenced genomes along with other group of genomes. AHPND (Shrimp^{+}/Human^{-}), AHPND (Korea; In this article) (Shrimp^{+}/Human^{-}), Pathogenic (Shrimp^{+}/ Human^{+}), Non-pandemic (Shrimp^{+}/Human^{-}), Non-pathogenic (Shrimp^{-}/ Human^{-}). $^{+}$/$^{-}$: pathogenic nature

Table 1
Sequenced strains genome and its deposited database details. Here, PW: Pond water; Pv: *Litopenaeus vannamei*; C: Chromosome; P: Plasmid

Strains	Genbank Accession	KCRC Accession	Contigs	Type	Total bases	GC%	Source	Genes	(r/t)RNA
SC4	QPQB00000000	KCTC 13702BP	contig.1.cir	C	3,365,393	45.33	Pv	4,883	37/131
			contig.2.cir	C	1,861,894	45.38			
			contig.4.cir	P	69,257	45.87			
SK	QPQC00000000	KCTC 13703BP	contig.1.cir	C	3,355,895	45.33	Pv	4,883	37/132
			contig.2.cir	C	1,877,787	45.35			
			contig.4.cir	P	69,261	45.87			
SM3	QPQD00000000	KCTC 13704BP	contig.1.cir	C	3,466,068	45.1	Pv	5,077	37/131
			contig.2.cir	C	1,781,672	45.62			
			contig.3.cir	P	70,608	42.09			
			contig.4.cir	P	69,262	45.87			
			contig.5.cir	P	52,744	44.77			
SM4	QPQE00000000	KCTC 13705BP	contig.1.cir	C	3,466,067	45.1	Pv	5,069	37/131
			contig.2.cir	C	1,786,479	45.6			
			contig.3.cir	P	70,152	42.09			
			contig.4.cir	P	69,260	45.87			
			contig.5.cir	P	52,746	44.77			
WC15	QPQF00000000	KCTC 13706BP	contig.1.cir	C	3,365,382	45.33	PW	4,874	37/132
			contig.2.cir	C	1,864,186	45.38			
			contig.4.cir	P	69,257	45.87			
WS	QPQG00000000	KCTC 13707BP	contig.1.cir	C	3,383,458	45.33	PW	4,764	38/133
			contig.2.cir	C	1,744,707	45.63			
			contig.3.cir	P	9,282	41.49			
			contig.4.cir	P	69,237	45.87			
WY3	QPQH00000000	KCTC 13708BP	contig.1.cir	C	3,366,116	45.33	PW	4,874	37/132
			contig.2.cir	C	1,864,201	45.38			
			contig.3.cir	P	7,588	42.07			
			contig.4.cir	P	69,261	45.87			

tree from the 46 single copy genes, which are selected from ortholog analysis conducted with OrthoMCL method. The assembled contigs and plasmids were deposited in GenBank (https://www.ncbi.nlm.nih.gov/) under the accession QPQB00000000 - QPQH00000000. Further, facilitate the easy access to the strains which used in this article, the cultures were deposited in Korean collection for type cultures (KCTC) (https://kctc.kribb.re.kr) under the accession number KCTC13702BP-KCTC13708BP (Table 1). All the figure source files were given in the supplementary folder, which named as figures source files. The README file has all the basic information of the files, which used for the figure two and three.

2. Experimental Design, Materials, and Methods

2.1. Samples

White shrimps (*Litopenaeus vannamei*) collected from the aquaculture farms located in Incheon, Pyeongtaek, Gyeonggi-do, Shinan, and Jeollanam-do of Korea at 2016, which experienced massive death. The water collected from ponds and hepatopancreas excised from shrimps were inoculated on the thiosulfate citrate bile salt sucrose plates (TCBS supplemented with 1.5% NaCl) and incubated at 25°C for 24 h. The colonies of isolated bacteria were cultured on tryptic soy agar plates (TSA supplemented with 1.5% NaCl). The AHPND property was determined by the PCR method using genomic DNA [1].

Table 2
Four group of *Vibrio parahaemolyticus* strains selected for ortholog analysis. The groups are AHPND (Shrimp$^+$/Human$^-$), Pathogenic (Shrimp$^+$/ Human$^+$), Non-pandemic (Shrimp$^+$/Human$^-$), Non-pathogenic (Shrimp$^-$/ Human$^-$). $^+/^-$: pathogenic nature

S.No	Strains	Alias	Group	Accession
1	NCKU_TV_5HP	AH1	AHPND causative strains (Shrimp$^+$ / Human$^-$)	GCF_000736315.1
2	NCKU_CV_CHN	AH2		GCF_000736325.1
3	NCKU_TV_3HP	AH3		GCF_000736335.1
4	13-028_A3	AH4		GCF_000737635.1
5	TUMSAT_DE1_S1	AH5		GCF_000591455.1
6	TUMSAT_DE2_S2	AH6		GCF_000591475.1
7	TUMSAT_D06_S3	AH7		GCF_000591495.1
8	M0605	AH8		GCF_000523375.1
9	FIM-S1708+	AH9		GCF_000732985.1
10	SNUVpS-1	ND1	Non-pandemic (Shrimp$^+$ / Human$^-$)	GCF_000315135.1
11	v110	ND2		GCF_000388025.1
12	NCKU_TN_S02	NP1	Non-pathogenic (Shrimp$^-$ / Human$^-$)	GCF_000736345.1
13	S171	NP2		GCF_000489075.1
14	VIP4-0444	NP3		GCF_000500485.1
15	VIP4-0447	NP4		GCF_000500545.1
16	VPCR-2009	NP5		GCF_000593305.2
17	VPTS-2009	NP6		GCF_000593325.2
18	RIMD2210633	PA1	Pathogenic (Shrimp$^+$ / Human$^+$)	GCF_001270945.1
19	BB22OP	PA2		GCF_000328405.1
20	Peru-466	PA3		GCF_000182345.1
21	SC4	SC4	AHPND (Korea) (Shrimp$^+$ / Human$^-$) This article	QPQB00000000
22	SK	SK		QPQC00000000
23	SM3	SM3		QPQD00000000
24	SM4	SM4		QPQE00000000
25	WC15	WC15		QPQF00000000
26	WS	WS		QPQG00000000
27	WY3	WY3		QPQH00000000

2.2. Genomic DNA isolation, sequencing, and assembly

The complete experimental procedures were conducted based on the instructions given in the respective products/kits. The below sequential steps from DNA isolation to sequencing procedures were conducted by DNALink, the authorized sequence service provider (http://www.dnalink.com/korean/index.html). Genomic DNA of these strains were extracted using the Genomic DNA Purification Kit (Qiagen, Hilden, Germany). DNA was examined by 1% agarose gel electrophoresis and quantified by a NanoDrop 2000 spectrophotometer (Thermo Scientific, MA, USA). A 10-KB library was prepared and sequenced using the PacBio (Pacific Biosciences) platform. One SMRT cell for each strain was used for sequencing, and those sequences are assembled separately by using the hierarchical genome assembly process (HGAP 2.0) [2]. Each assembly was constructed using more than 1 Gb of PacBio reads, and assembled contigs are polished with Quiver to reach final consensus accuracy.

2.3. Genome annotation

Gene prediction was carried out using Glimmer method [3], and small RNAs (i.e., rRNA and tRNA) were predicted using RNAmmer [4] and tRNAscan-SE v1.21 [5] methods respectively. Finally, the functional gene annotation was carried out based on homology searches against NR database and gene ontology (GO) database using Blast2GO method [6].

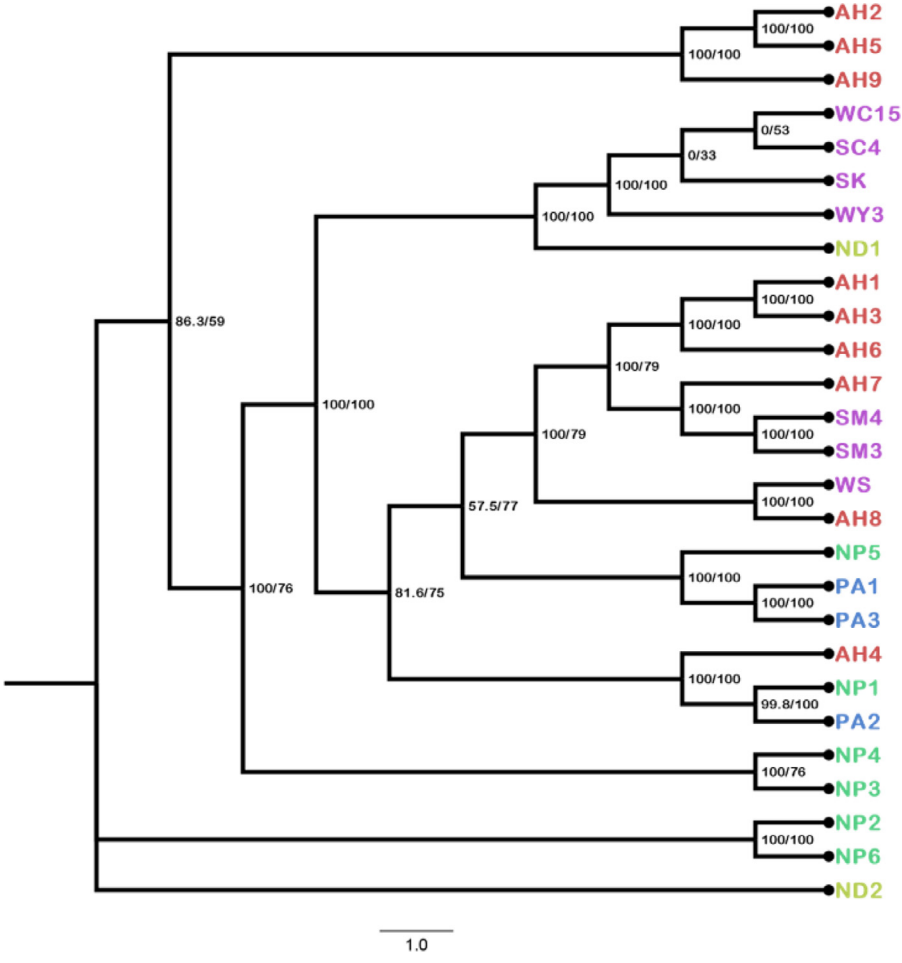

Fig. 3. Phylogenetic tree from 46 single copy genes. The maximum likelihood trees are constructed with 1,000 bootstrap iteration with other default parameters.

2.4. Polymeric chain reaction for pirA and pirB

Genomic DNA was extracted from cultured bacteria using high pure PCR template preparation kit according to manufacturer instructions (Roche Life Science). PCR reaction mixture is prepared with PuRe Taq ready-to-go PCR beads (GE Healthcare). To detect AHPND, PCR was performed with DNA templates using duplex primer sets of PirA and PirB genes (i.e., VpPirA-284F: TGACTATTCTCACGATTGGACTG, VpPirA-284R: CACGACTAGCGCCATTGTTA, VpPirB-392F: TGATGAAGTGATGGGTGCTC, VpPirB-392R: TGTAAGCGCCGTTTAACTCA) under the following PCR conditions: initial denaturation at 94°C for 3min, followed by 35 cycles of 94°C for 30 sec, 60°C for 30 sec, and 72°C for 30 sec, and a final extension at 72°C for 7 min [1]. The PirA- and PirB-cloned plasmids and Distilled water were used as the positive control and negative control, respectively. The PCR products were electrophoresed on QIAxcel (Qiagen)

2.5. Phylogenetic tree

The sequenced and selected genomes from GenBank are subjected to ortholog analysis by using OrthoMCL method [7]. The single copy genes (which have only one copy of gene in each genome) were selected from OrthoMCL cluster file. These concatenated single copy proteins are aligned using the MAFFT v7.2 [8] with default parameters. The multiple alignment initially corrected with Gblocks v0.91 [9] and subjected to phylogenetic tree reconstruction using IQ-TREE v1.5.0 [10]. The tree imported to FigTree v1.4.3. (http://tree.bio.ed.ac.uk/software/figtree/) to obtain the phylogenetic tree.

2.6. Data deposition

This Whole Genome Shotgun project has been deposited at GenBank under the accession QPQB00000000 - QPQH00000000. The microbial cultures were deposited in Korean collection for type cultures (KCTC) under the accession number KCTC13702BP-KCTC13708BP.

Acknowledgments

This work was supported by a grant from the National Institute of Fisheries Science (R2020060).

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2020.105697.

References

[1] J.E. Han, K.F.J. Tang, L.H. Tran, D.V Lightner, Photorhabdus insect-related (Pir) toxin-like genes in a plasmid of Vibrio parahaemolyticus, the causative agent of acute hepatopancreatic necrosis disease (AHPND) of shrimp, Dis. Aquat. Organ 113 (2015) 33–40 https://doi.org/10.3354/dao02830.
[2] C.-S. Chin, D.H. Alexander, P. Marks, A.A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E.E. Eichler, S.W. Turner, J. Korlach, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, Nat. Methods 10 (2013) 563 https://doi.org/10.1038/nmeth.2474.
[3] A.L. Delcher, K.A. Bratke, E.C. Powers, S.L. Salzberg, Identifying bacterial genes and endosymbiont DNA with Glimmer, Bioinformatics 23 (2007) 673–679 https://doi.org/10.1093/bioinformatics/btm009.
[4] K. Lagesen, P. Hallin, E.A. Rødland, H.-H. Stærfeldt, T. Rognes, D.W. Ussery, RNAmmer: consistent and rapid annotation of ribosomal RNA genes, Nucleic Acids Res 35 (2007) 3100–3108 https://doi.org/10.1093/nar/gkm160.
[5] T.M. Lowe, S.R. Eddy, tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence, Nucleic Acids Res 25 (1997) 955–964 https://doi.org/10.1093/nar/25.5.955.
[6] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676 https://doi.org/10.1093/bioinformatics/bti610.
[7] L. Li, C.J. Stoeckert, D.S. Roos, OrthoMCL: Identification of ortholog groups for eukaryotic genomes, Genome Res (2003) https://doi.org/10.1101/gr.1224503.

[8] K.D. Yamada, K. Tomii, K. Katoh, Application of the MAFFT sequence alignment program to large data - Reexamination of the usefulness of chained guide trees, Bioinformatics (2016) https://doi.org/10.1093/bioinformatics/btw412.
[9] G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, Syst. Biol (2007) https://doi.org/10.1080/10635150701472164.
[10] L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh, IQ-TREE, A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies, Mol. Biol. Evol 32 (2014) 268–274 https://doi.org/10.1093/molbev/msu300.