Received: 13 April 2022 | Revised: 29 May 2022 | Accepted: 1 June 2022

DOI: 10.1111/jfd.13668

Journal of Fish Diseases WILEY

RESEARCH ARTICLE

A genome-wide association study of the occurrence of genetic variations in *Edwardsiella piscicida*, *Vibrio harveyi*, and *Streptococcus parauberis* under stressed environments

HyeongJin Roh

Pathogens and Disease Transfer, Institute of Marine Research, Bergen, Norway

Correspondence
HyeongJin Roh, Pathogens and Disease Transfer, Institute of Marine Research, Bergen, Norway.
Email: hyeongjin.roh@hi.no; hjroh@pukyong.ac.kr

Funding information
Institute of Marine Research

Abstract

Bacterial mutation and genetic diversity in aquaculture have led to increasing phenotypic variances, which can weaken or invalidate strategies for controlling diseases. However, few studies have monitored the degree of mutation in fish bacterial pathogens caused by environmental pressure within a short period. In this study, transcriptomic sequences from *Edwardsiella piscicida*, *Vibrio harveyi* and *Streptococcus parauberis* under stressed environments were used for investigating the emergence of variants. In detail, a sub-inhibitory concentration of formalin and phenol for *E. piscicida*, sea water at 30°C for *V. harveyi* and flounder serum for *S. parauberis* were used as stressed environments, and significant single-nucleotide polymorphisms (SNPs) and/or mutation sites were investigated after culture in the ordinary liquid media (control) and the stressed environment through a genome-wide association study. As results, several SNPs or mutations during incubation were observed under different environments in *E. piscicida* and/or *V. harveyi* in the genes relevant to flagella, fimbria type 3 secretion systems, and outer and inner membranes that have been directly exposed to external environments. In particular, given that flagella and fimbriae are considered important factors in differentiating the serotypes in some bacterial pathogens, it can be speculated that different environmental pressures are the source of phenotypic or serotypic differentiation from the same origin. On the other hands, *S. parauberis* did not exhibit notable changes for 4 h when inoculated in the serum from olive flounder. The results presented in this study provide examples of possible molecular evolution in pathogens relevant to the aquaculture industry as a response to different environmental pressure.

KEYWORDS
bacteria mutation, environmental pressure, genetic variance, GWAS, SNP

1 | INTRODUCTION

Bacterial diseases are the major threats to farmed fish in aquaculture, and multiple studies have been conducted to develop methods for controlling them (e.g., Roh et al., 2016; Seo et al., 2021; Zhang et al., 2021). However, the damage caused by bacterial diseases in aquaculture has continued, and the ability to alter their genetic makeup much faster than eukaryotes is regarded as a major obstacle (Figueroa et al., 2019; Kim et al., 2021; Roh et al., 2019). For example, the emergence and spread of multi-drug-resistant bacteria in aquaculture where bacteria are required to survive against antibiotics can be originated from the pressure to evolve by modifying genes that help them adapt to the newly encountered situation (Algammal et al., 2022; Nadella et al., 2022). In addition, because most bacterial pathogens in aquaculture can be horizontally transmitted through water where many hydrophilic pollutants and chemicals are easily dissolved, different water environments can accelerate the occurrence of bacterial mutations. It is therefore very important to understand the process or patterns of genetic changes among genomes differentiated from the same strain. In recent years, the importance of interpretation and analysis of massive data has received an enormous amount of attention as technological advances are made in the field of sequencing technology, and studies investigating the genotypic characteristics between strains at the genome level have been conducted (e.g., Le et al., 2021; Roh et al., 2019, 2020; Roh & Kim, 2021). Because big data from next-generation sequencing (NGS) may contain valuable multidimensional results, these can be interpreted from multiple perspectives. However, in numerous studies, massive sequencing results have mainly been interpreted in a manner consistent with their purposes, which means some results might be important but are overlooked or underestimated (Lee et al., 2021; Montánchez et al., 2019; Yoon et al., 2020). Hence, making the most of released sequences in an open database such as a sequence read archive (SRA) and Genbank is becoming more important, given the release of a tremendous amount of sequencing results every day.

Among the methods of bacterial genomic study, the pan-genome method that profiles genomic differences by the existence of certain genes has been widely used and contributes to understanding the relationship between phenotypic and genotypic characteristics of bacterial pathogens to a certain extent (Kim et al., 2020; Roh & Kim, 2021; Rouli et al., 2015). However, because pan-genome analysis is based on the existence of genes classified as core, accessory, and singleton genes, rather than the sequence differences which are annotated as the same gene (Blom et al., 2016; Zekic et al., 2018), it is not suitable for analysing the occurrence of different single-nucleotide polymorphisms (SNPs) or mutations by strain. Even though emerging mutations and/or SNPs in certain genes or locations might be an earlier response than the differences in accessory genes, studies on SNPs and deletion and/or mutations caused by different environmental niches have been underestimated. To overcome these shortcomings, genome-wide association studies (GWASs) offer a valuable alternative. GWAS is a method for excavating significant SNPs associated with phenotypic differences by profiling all genetic variants between phenotypic groups, not at the level of genes but at each nucleotide (Lees & Bentley, 2016). The advent of GWAS in bacteria has provided insight and understanding into SNPs associated with phenotypic characteristics (Chen & Shapiro, 2015). To estimate the occurrence of SNPs or mutated sequences when the same strain is cultured under different environments through GWAS, simultaneous deep sequencing of raw files from a large number of bacterial genomes is necessary (Wang et al., 2015). Thus far, few studies have examined genomic DNA sequences from the same strain cultured in different environments, although some studies have investigated transcriptomic responses for major bacterial pathogens in aquaculture (*Vibrio harveyi*, *Edwardsiella piscicida* and *Streptococcus parauberis*) (Lee et al., 2021; Montánchez et al., 2019; Yoon et al., 2020). However, the information on nucleic acid sequences in transcriptomics that may contain promising pieces of evidence for the emergence of mutated sequences can be valuable themselves. *V. harveyi*, *E. piscicida and S. parauberis* are well-known bacterial pathogens in aquaculture, and both raw sequencing data (FastQ file) and experimental designs have been published (Lee et al., 2021; Montánchez et al., 2019; Yoon et al., 2020). In addition, their transcriptome responses have been known to change remarkably under sea water when exposed to high incubation temperature, pollutants or host serum environments, compared with cultures in liquid media at an optimal growth temperature (Lee et al., 2021; Montánchez et al., 2019; Yoon et al., 2020). These transcriptomic changes are evidence that bacterial pathogens are strongly influenced by environmental factors. Based on these studies, the purposes of this study were to (1) analyse RNA-seq raw files from several fish bacterial pathogens in various environments with a focus on mutations and SNPs, (2) identify genes that exhibit significant SNPs and mutations by different groups and (3) speculate on the microbiological meaning of these changes.

2 | MATERIALS AND METHODS

2.1 | Samples and FastQ files information

In total, 41 FastQ files for transcriptomic studies on *E. piscicida*, *V. harveyi* and *S. parauberis* have been downloaded from a SRA using the SRA-tool kit (Ver. 2.11.3-ubuntu64). The SRR accessions used in this study are summarized in Table 1 (Lee et al., 2021; Montánchez et al., 2019; Yoon et al., 2020). To observe the emergence of genetic variants stimulated by sub-lethal chemicals in *E. piscicida*, a study on *E. piscicida* CK108 transcriptomics that compared a control cultured in ordinary liquid media (brain and heart infusion) and treatments cultured in a media supplemented with either 0.0039% formalin or 0.146% phenol were used in this study (Yoon et al., 2020). In the case of *V. harveyi*, this study focussed on what kind of genetic variances occurred over time in the sea water with high water temperature compared with *V. harveyi* grown in sufficient nutrients and at the optimal temperature. In detail, the transcriptomic results of ATCC®14126™ grown in the marine broth at 26°C overnight were used as the control, while groups of *V. harveyi* transferred in sea water (SW) at 30°C for 12h, 3 and 6days were regarded as treatment groups for GWAS (Montánchez et al., 2019). Likewise, the possibilities of genetic variances in *S. parauberis* under liquid media and serum environment with the elapsed time were investigated in this study. In more detail, transcriptomic results of *S. parauberis* SPOF3K cultured in liquid media (brain and heart infusion supplemented

TABLE 1 Bacteria species, comparison groups for genome-wide association studies, experimental conditions, SRR accession numbers, and references used in this study

Species	GWAS comparison	Experimental conditions	SRR accession	Reference
Edwardsiella piscicida	• Control vs. formalin group • Control vs. phenol group	*E. piscicida* CK108 was incubated in liquid media (brain heart infusion) for 24 h at 27°C with 200 rpm shaking as the control (*n* = 3). The formalin (*n* = 4) and phenol treated groups (*n* = 4) were cultured under the same conditions but supplemented by formalin (final concentration = 0.0039%; 0.38 mM) and phenol (final concentration = 0.146%; 1. 3 mM)	SRR11652662-SRR11652676	Yoon et al. (2020)
Vibrio harveyi	• Control vs. 12 h seawater incubation at 30°C (12 h) • Control vs. 3 days seawater incubation at 30°C (3 days) • Control vs. 6 days seawater incubation at 30°C (6 days)	*V. harveyi* ATCC® 14126™ were aerobically cultured overnight in liquid media (marine broth) at 26°C as the control group (*n* = 3). The same *V. harveyi* strain which had been cultured for 12–16 h incubation in the same culturable environment as the control was diluted by sterile seawater (1:20; 20 folds dilution) and incubated at 30°C for 12 h, 3, and 6 days	SRR7058340-SRR7058351	Montánchez et al. (2019)
Streptococcus parauberis	• 1 h broth vs. 1 h serum • 2 h broth vs. 2 h serum • 4 h broth vs. 4 h serum	*S. parauberis* SPOF3K grown in liquid media (brain and heart infusion with 1% NaCl) at 26°C for 18 h with 180 rpm shaking was incubated in fresh liquid media or serum achieved from Olive flounder (~200 g). The tri-replicated cultures in both broth (control) and serum (treatment) were sampled at 1, 2, and 4 h post-incubation	SRR13349544-SRR13349561	Lee et al. (2021)

with 1% NaCl) and serum from olive flounder (*Paralichthys olivaceus*; ~200 g) for 1, 2 and 4 h were utilized (Lee et al., 2021). The summaries of the experiment designs and comparison groups for GWAS are presented in Table 1.

2.2 | Sequence analysis

The adapter and raw quality of reads in all FastQ files used in this study were trimmed using Trim Galore! (Ver. 0.6.7+ galaxy0) under the conditions of a Phred quality score threshold = 20, maximum allowed error rate = 0.1, and discarding reads that became shorter than 20 bp (Krueger, 2021). The genome sequences of *E. piscicida* CK41 (Genbank accession: CP047671.1), *V. harveyi* ATCC 14126 (Genbank accession: NZ_BCUF01000001.1) and *S. parauberis* SPOF3K (CP025420.1) downloaded from GenBank were indexed, and the trimmed FastQ files were mapped onto each indexed genome using Bowtie2 (Ver. 2.4.1) (Langmead & Salzberg, 2012). The SAM or BAM file was converted to a sorted BAM file using the sort option of samtools (Ver. 1.9) (Li et al., 2009). Based on the comparison groups of GWAS in *E. piscicida*, *V. harveyi* and *S. parauberis* described in Table 1, multiple sorted BAM files were combined to generate a bcf file in the conditions of -Ob, -m and--ploidy 1 using the mpileup option of bcftools (Danecek et al., 2016). Likewise, all sorted BAM files for each species were merged, and its bcf file was used to observe correlation missing genotypes based on the identity-by-missingness (IBM) between samples. All bcf files were converted to vcf files for further analyses using bcftools, and variant sequence frequency (VSF) was calculated for minor variant sequences from each comparison of vcf file.

2.3 | SNP calling, genome-wide association analysis and bioinformatics

All genome-wide association analyses used in this study were conducted according to the guidelines of second-generation PLINK (Ver. 1.9) (Chang et al., 2015). In brief, SNPs and/or mutations were identified with a threshold of 0.01 minor allele frequency (MAF), and ped and map files were generated using vcftools (Danecek et al., 2011). The label information and ped and/or map files were then merged, and bed, fam and bim files were used to locate missing SNPs and SNPs that exhibited significant differences by treatment with--allow-no-sex mode. To analyse the correlation of missing genotypes in each bacterial species, the bcf file containing all sorted BAM files of the same species was used. Conversely, for GWAS, the merging file containing only the sorted BAM files relevant to a GWAS comparison was used. A *p*-value less than .05 for each GWAS comparison was considered a significant SNP.

2.4 | Data analysis, visualization and statistics

The results of the correlation of missing genotypes between samples based on IBM were employed to identify the differences among all aligned sequences in each bacterial species, and hierarchical clustering with the complete method was performed using the ape package in R (Paradis & Schliep, 2019). The location and annotation of coding DNA sequences in each reference genome were predicted using RastTk pipeline (Brettin et al., 2015). All nucleic acid sequences that exceeded 0.01 MAF and SNPs belonging to genes that possessed the first to third highest numbers of significant SNPs were visualized in a Manhattan plot using qqman and ggplot2 library in R (Ver. 4.0.5) (Turner, 2014; Wickham, 2016). The gene with at least one significant SNP was mapped onto KEGG pathways using the KEGG automatic annotation server to trace the biological pathways that harboured genes with SNPs identified in each GWAS comparison (Moriya et al., 2007). For this, a single-directional best hit mode was used, and several data sets of reference bacterial species (eic and etr; *E. piscicida*, the same or closest species to *E. piscicida* (eic and etr), *V. harveyi* (vch, vcj, vco, vcm, vvu, vvy, vvm, vpa, vha, vex, vsp and van), and *S. parauberis* (spy, spn, sag, smu, stc, ssa, ssb, sgo, sez, sub, sds, sga, smb and std) were selected as the reference databases for the KEGG annotation. The ratio for the number of genes in featured pathways where more than three genes have been harboured was visualized using the Enrichment map in Cytoscape (Ver. 3.7.2) (Merico et al., 2010; Shannon et al., 2003). The significant difference in the number of genes that have at least one SNP in each KEGG pathway has been identified by the chi-square analysis using a gmodels package in R (Warnes et al., 2015).

3 | RESULTS

3.1 | SNPs and GWAS for *E. piscicida*

Although one control sample of *E. piscicida* (*E. piscicida* WT_3) was located relatively far from the other two samples, most samples were well-clustered in accordance with the phenol and formalin treatments (Figure S1a). The number of SNPs that exceeded 0.01 MAF was 1721 and 2731 in the formalin and phenol groups, respectively, and 514 and 512 were significant SNPs ($p < .05$) (Figure S1b). The number of significant SNPs belonging to each gene was then counted based on the location of significant SNPs and genes in *E. piscicida*, and 15 genes (malS, TcfC, MrdA, ClC, nrfE, SDM, L-Asc family, dppA, fucP, tolC, TREM2, LcrD, DgcZ, MetC and FimD; all abbreviations are shown in Table 2) were found to contain more than 3 significant SNPs in either the formalin or the phenol comparison (Table 2). Among these genes, MrdA, TcfC and fucP had 6, 4 and 4 significant SNPs, which were the first to third highest genes between the control and formalin groups (Figure 1A). For the control and phenol comparison, both genes (MrdA and LcrD) contained highest number of significant SNPs ($n = 6$), and fucP was the third highest ($n = 5$) (Figure 1A). Similar or identical numbers of SNPs were observed in 8 of 15 genes in both formalin and phenol groups; however, seven genes (L-Asc family, dppA, fucP, tolC, DgcZ, nrfE and FimD) were primarily observed in the formalin comparison group (Figure 1B). On the other hand, a smaller number of SNPs from these 15 genes was identified in the comparison between formalin and phenol groups (Table S1). To estimate the biological meaning of the genes where at least one SNPs or mutation was observed, all genes with the significant SNP(s) were annotated using the KEGG database; the numbers of genes mapped onto each pathway are depicted in Table S2. The proportion of the number of genes in each pathway out of all genes confirmed in both formalin and phenol comparison was calculated, and the pathways where more than three genes were assigned to at least one group are visualized in Figure 2. Five of 28 KEGG pathways (metabolic, galactose metabolism, bacterial chemotaxis, flagellar assembly and two-component systems) contained a significantly higher number of genes in phenol groups than in formalin groups (Figure 2).

3.2 | SNPs and GWAS for *V. harveyi*

In general, the clustering result of missing genotype correlations based on the IBM did not show any specific patterns among the samples, although the samples from *V. harveyi* incubated in the SW at 30°C for 6 d tended to be closer than other groups (Figure S2a). The numbers of SNPs higher than 0.01 MAF from the three comparisons (between Con and SW at 30°C for 12 h, 3 and 6 days) were 2955, 6678 and 2015, respectively, and 143, 574 and 143 sites in 12 h, 3 and 6 days groups, respectively, were significant sites ($p < .05$) (Figure S2b). The upper genes with the first to the third-largest number of significant SNPs were identified for each sampling time point, and six genes (fliN, FlaD, RcpC, DECR, PrkC and Eis) were selected in the 12-h group. Likewise, seven genes (COX1, RMD-MFP, RND-IMT, mcpB, FliN, lapD and PrkC) and six genes (FliN, RcpC, OMP, PrkC, DUF4150 and TKT) were regarded as those that had the largest number of significant SNPs in the 3 and 6 days groups, respectively (Figure 3A). Among these genes, seven (RND-MFP, RND-IMT, mcpB, COX1, lapD, Eis and TKT) exhibited 3 or more significant SNPs in a gene from at least one group (Table 3). In particular, five genes (RND-MFP, RND-IMT, mcpB, COX1 and lapD) had SNPs in only the 3 days group, and GNAT family N-acetyltransferase and Transketolase contained SNPs in the 12 h or 6 days groups (Figure 3B). Because numerous genes were dominantly enriched into KEGG pathways in the 3 days group, compared with other groups, 20 of 28 KEGG pathways had a statistically different number of genes that harboured significant SNPs between the groups (Figure 4; Table S2). In particular, KEGG pathways such as flagella assembly, ABC transporters, two-component system, biosynthesis of cofactors, microbial metabolism in diverse environments, metabolic pathways, biosynthesis of secondary metabolites, and biofilm formation—*Vibrio cholera* had a higher number of genes in the 3 days group ($p < .001$).

TABLE 2 Location, a reference code, A2 (major allele sequence)/A1 (minor allele sequence), chi-square and *p*-value of significant SNPs from GWAS for *Edwardsiella piscicida* in the formalin and phenol group, and mapping information (MI) including depth (DP) and variant sequence frequency (VSF) of each SNP

			WT vs. formalin						WT vs. phenol					
			GWAS				MI		GWAS				MI	
Gene (abbreviation)	SNP site	Ref	A2/A1	CHISQ	P	FDR	DP	VSF (%)	A2/A1	CHISQ	P	FDR	DP	VSF (%)
Periplasmic alpha-amylase (malS)	370,815	T	T/C	8	0.005	0.040	13	10	T/C	8	0.005	0.086	11	11
	370,888	T	T/C	6	0.014	0.100	9	17	T/C	10	0.002	0.086	14	8
	371,452	G	G/A	8	0.005	0.040	13	8	G/A	8	0.005	0.086	10	11
TcfC E-set like domain-containing protein[a] (TcfC)	1,238,362	A	A/G	8	0.005	0.040	6	17	G/A	4	0.046	0.185	2	50
	1,239,622	G	**G/A**	**4.8**	**0.028**	**0.103**	**47**	**6**	–				14	18
	1,240,084	A	A/G	6	0.014	0.100	4	25	G/A	4	0.046	0.185	2	50
	1,240,241	A	A/G	10	0.002	0.017	22	5	G/A	4	0.046	0.185	2	50
	1,238,381	G	–				8	0	G/A	4	0.046	0.185	2	50
	1,239,178	T	–				44	3	T/C	6	0.014	0.152	8	14
Peptidoglycan D,D-transpeptidase (MrdA)	1,394,009	A	**A/C**	**4.8**	**0.028**	**0.103**	**142**	**2**	A/C	4.8	0.028	0.185	29	8
	1,394,010	A	**A/T**	**4.8**	**0.028**	**0.103**	**144**	**2**	A/T	4.8	0.028	0.185	29	8
	1,394,012	A	**A/G**	**4.8**	**0.028**	**0.103**	**145**	**2**	A/G	4.8	0.028	0.185	29	8
	1,394,013	C	**C/G**	**4.8**	**0.028**	**0.103**	**146**	**2**	C/G	4.8	0.028	0.185	29	8
	1,394,014	C	**C/A**	**4.8**	**0.028**	**0.103**	**147**	**2**	C/A	4.8	0.028	0.185	29	9
	1,394,016	T	**T/A**	**4.8**	**0.028**	**0.103**	**147**	**2**	**T/A**	**4.8**	**0.028**	**0.185**	**30**	**8**
Chloride channel protein (ClC)	1,437,777	G	G/A	4.8	0.028	0.103	11	11	G/A	4.8	0.028	0.185	14	9
	1,437,628	C	–						C/T	4.8	0.028	0.185	27	5
	1,437,717	G	–						G/A	4.8	0.028	0.185	28	5
Cytochrome c-type heme lyase subunit (nrfE)	1,539,560	A	**A/G**	**10**	**0.002**	**0.017**	**43**	**4**	G/A	4	0.046	0.185	5	33
	1,539,936	G	G/A	6	0.014	0.100	6	50	–				4	100
	1,540,797	T	T/C	10	0.002	0.017	20	6	T/C	6	0.014	0.152	4	25
SAM-dependent methyltransferase (SDM)	1,601,113	A	A/G	10	0.002	0.017	24	14	A/G	8	0.005	0.086	12	30
	1,601,399	T	T/C	10	0.002	0.017	33	4	T/C	10	0.002	0.086	12	14
	1,601,414	A	A/G	10	0.002	0.017	26	5	A/G	10	0.002	0.086	12	13
PTS system, ascorbate-specific IIC component (L-Asc family)	2,073,595	T	**T/C**	**4.8**	**0.028**	**0.103**	**90**	**6**	–				22	33
	2,074,422	T	T/C	4.8	0.028	0.103	24	5	–				4	33
	2,074,672	C	C/T	4.8	0.028	0.103	21	10	–				7	29
	2,074,142	T	–				41	0	T/C	6	0.014	0.152	5	40
	2,074,671	T	–				21	0	T/C	6	0.014	0.152	7	14

(Continues)

TABLE 2 (Continued)

			WT vs. formalin						WT vs. phenol					
			GWAS				MI		GWAS				MI	
Gene (abbreviation)	SNP site	Ref	A2/A1	CHISQ	P	FDR	DP	VSF (%)	A2/A1	CHISQ	P	FDR	DP	VSF (%)
Dipeptide ABC transporter (dppA)	2,194,891	T	T/C	4.8	0.028	0.103	30	4	T/C	4.8	0.028	0.185	20	6
	2,194,894	T	T/C	4.8	0.028	0.103	30	4	T/C	4.8	0.028	0.185	21	6
	2,195,101	C	C/T	4.8	0.028	0.103	17	13	–				8	33
Fucose permease (fucP)	2,690,807	A	A/T	10	0.002	0.017	19	7	T/A	4	0.046	0.185	4	50
	2,690,834	G	G/A	10	0.002	0.017	20	7	–				2	100
	2,691,021	T	T/C	4	0.046	0.126	3	33	–				2	0
	2,691,140	T	A/T	4	0.046	0.126	4	25	–				1	100
TolC family protein (tolC)[a]	2,726,025	T	T/C	10	0.002	0.017	15	13	–				2	100
	2,726,173	T	T/C	4.8	0.028	0.103	11	20	–				5	50
	2,726,409	A	A/G	4.8	0.028	0.103	28	6	–				6	25
Ig-like domain-containing protein (TREM2)[a]	2,735,523	G	**G/A**	**8**	**0.005**	**0.040**	**39**	**3**	A/G	4	0.046	0.185	2	50
	2,738,408	A	A/G	6	0.014	0.100	4	33	–				2	100
	2,738,491	T	T/C	6	0.014	0.100	5	25	–				1	100
	2,734,265	A	–				24	9	A/G	4.8	0.028	0.185	9	29
	2,736,314	A	–				15	0	A/G	6	0.014	0.152	7	17
	2,736,408	T	–				10	0	T/C	6	0.014	0.152	8	25
Type III secretion inner membrane channel protein (LcrD)	3,326,526	T	**T/C**	**10**	**0.002**	**0.017**	**230**	**1**	C/T	4	0.046	0.185	3	33
	3,327,336	A	**A/G**	**4.8**	**0.028**	**0.103**	**196**	**2**	A/G	4.8	0.028	0.185	15	17
	3,327,402	A	**A/G**	**4.8**	**0.028**	**0.103**	**107**	**2**	A/G	4.8	0.028	0.185	21	25
	3,326,852	T	–				52	0	T/C	4	0.046	0.185	2	50
	3,327,486	C	–				125	0	C/A	4	0.046	0.185	2	50
	3,328,029	T	–				82	0	T/C	4	0.046	0.185	2	50
Diguanylate cyclase (DgcZ)	3,401,083	T	**T/C**	**4.8**	**0.028**	**0.103**	**106**	**2**	T/C	4.8	0.028	0.185	23	5
	3,402,611	T	**T/C**	**4.8**	**0.028**	**0.103**	**175**	**1**	-				29	4
	3,402,641	T	**T/C**	**4.8**	**0.028**	**0.103**	**197**	**1**	–				33	4
Cystathionine beta-lyase (MetC)	3,518,551	T	T/G	10	0.002	0.017	7	14	T/G	8	0.005	0.086	5	25
	3,518,851	C	T/C	4	0.046	0.126	3	67	T/C	4	0.046	0.185	4	67
	3,518,926	T	T/C	4	0.046	0.126	6	33	–				4	0
	3,519,395	G	–				17	0	G/A	4	0.046	0.185	3	50

TABLE 2 (Continued)

Gene (abbreviation)	SNP site	Ref	WT vs. formalin						WT vs. phenol					
			GWAS				MI		GWAS				MI	
			A2/A1	CHISQ	P	FDR	DP	VSF (%)	A2/A1	CHISQ	P	FDR	DP	VSF (%)
Type 1 fimbriae anchoring protein (FimD)	3,733,406	T	T/C	4.8	0.028	0.103	24	5	-				12	10
	3,734,033	T	T/C	6	0.014	0.100	5	25	-				1	100
	3,734,698	T	T/C	4	0.046	0.126	2	50	-				1	0

Note: Bold letters indicated SNPs with FDR <0.25 and DP ≥30.

[a] Gene was annotated by non-redundant (NR) protein sequence database (blastp).

3.3 | SNPs and GWAS for *S. parauberis*

In the case of *S. parauberis*, the dendrogram analysis based on missing genotype correlation revealed that the culture conditions and incubation time, especially the difference between broth and serum environments, were greatly influenced by their clustering. Although one sample (broth 4 h rep3) was abnormally closer than serum samples, most could be primarily divided by the culturable environments (broth and serum). The samples from broth environments tended to cluster well by incubation time, but not in the serum environment (Figure S3a). The number of SNPs exceeding 0.01 MAF in the comparison with each time point of broth and serum groups were 54, 43 and 65 at 1, 2 and 4 h, respectively, which was much smaller for *E. piscicida* and *V. harveyi* (Figures S3b; Figure 5). Similarly, no gene with more than two significant SNPs in all sampling time points was observed (Table 4).

4 | DISCUSSION

Although GWAS for fish bacterial pathogens is relatively underrated compared with fish GWAS for the purpose of quantitative trait locus (QTL) marker identification, a few studies have applied GWAS for fish bacterial pathogens, which has contributed to identifying genotypic characteristics in the same species (Correa et al., 2015; Holborn et al., 2018; Le et al., 2021; Rasmussen et al., 2016). Different phenotypic characteristics between strains in the same species can be created by a genetic pressure to change nucleic sequences and genes to adapt to specific environments, which has been widely observed in both prokaryotes and eukaryotes (e.g., Kim et al., 2021; Kjærner-Semb et al., 2021). This is one of the main reasons why the first step in most GWAS is to classify the phenotypic characteristics or measurements (which can be discrete units [e.g., high, intermediate and low virulence] and/or continuous values) (Chen & Shapiro, 2015). Given that bacterial genomes tend to have a much stronger linkage disequilibrium than eukaryotes, bacteria are more flexible in relation to altering their genetic information under different environments, and this can accompany mutation and sequence alternation within short time periods when they are in different culturable environments (Chen & Shapiro, 2015). It is not surprising that bacteria mutation rate can be influenced by death, population dynamics and stressful environments, and generally speaking 1–100 nucleotides in 10 million to 1 billion bases can be substituted per generation depending on environments and conditions (Chevallereau et al., 2019; Frenoy & Bonhoeffer, 2018; Westra et al., 2017). Given that bacterial doubling time is approximately 20 min in favourable environments, although the time span for bacterial replication differs owing to multiple factors (e.g., species, temperature and nutrients) (Jaruszewicz-Błońska & Lipniacki, 2017), more than one-day incubation in different conditions can be enough time to make a mutation in different sites in the genome and change the major nucleic acid sequences in their population group. In addition, the fact that relatively more frequent mutated, missed or substituted genes for a certain gene compared with

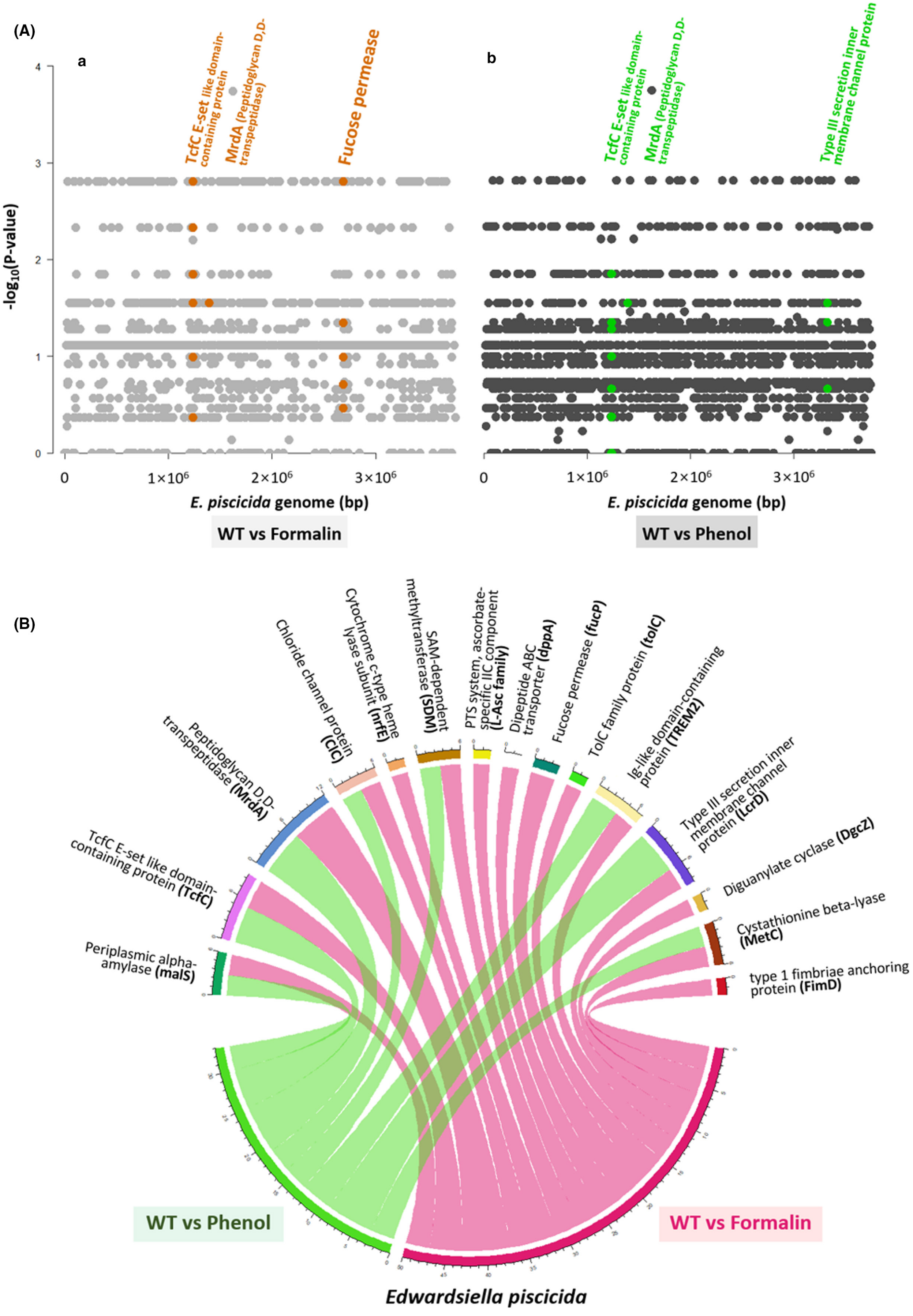
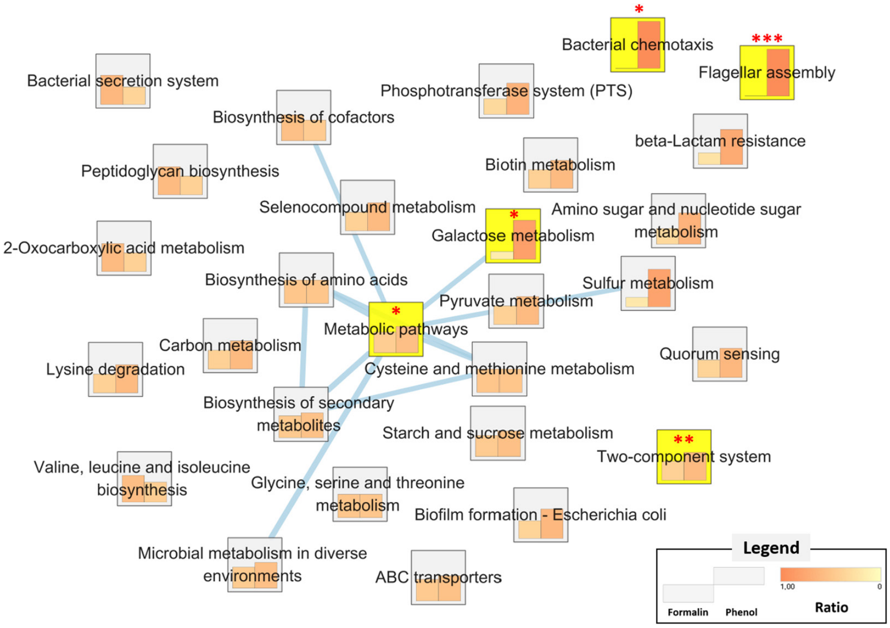

(A)
a
TcfC E-set like domain-containing protein
MrdA (Peptidoglycan D,D-transpeptidase)
Fucose permease
-log10(P-value)
0
1
2
3
4
0
1×10^6
2×10^6
3×10^6
E. piscicida genome (bp)
WT vs Formalin
b
TcfC E-set like domain-containing protein
MrdA (Peptidoglycan D,D-transpeptidase)
Type III secretion inner membrane channel protein
E. piscicida genome (bp)
WT vs Phenol
(B)
Periplasmic alpha-amylase (malS)
TcfC E-set like domain-containing protein (TcfC)
Peptidoglycan D,D-transpeptidase (MrdA)
Chloride channel protein (ClC)
Cytochrome c-type heme lyase subunit (nrfE)
SAM-dependent methyltransferase (SDM)
PTS system, ascorbate-specific IIC component (L-Asc family)
Dipeptide ABC transporter (dppA)
Fucose permease (fucP)
TolC family protein (tolC)
Ig-like domain-containing protein (TREM2)
Type III secretion inner membrane channel protein (LcrD)
Diguanylate cyclase (DgcZ)
Cystathionine beta-lyase (MetC)
type 1 fimbriae anchoring protein (FimD)
WT vs Phenol
WT vs Formalin
Edwardsiella piscicida

FIGURE 1 Manhattan plot of genome-wide association analysis (GWAS) between WT and formalin ((a)-a) or phenol group ((a)-b). Circos plot for the genes with more than 3 significant SNPs in one group. The thickness of green and pink coloured lines denotes the number of *Edwardsiella piscicida* SNPs in phenol and formalin groups, the abbreviation of each gene was written in brackets (b). The brown and green dots (a) indicate the genes that contained the first to the third highest number of significant SNPs ($p < .05$)

FIGURE 2 The ratio for the number of genes with at least one significant SNP between formalin and phenol groups belonging to KEGG pathways in *Edwardsiella piscicida*. The thickness of edges between nodes indicates similarity based on the number of shared genes. The yellow background with the asterisk in the node indicates significantly different counts based on chi-square analysis (*$p < .05$; **$p < .01$; ***$p < .001$)


*
Bacterial chemotaxis
***
Flagellar assembly
Bacterial secretion system
Phosphotransferase system (PTS)
Biosynthesis of cofactors
beta-Lactam resistance
Biotin metabolism
Peptidoglycan biosynthesis
Selenocompound metabolism
*
Amino sugar and nucleotide sugar metabolism
Galactose metabolism
2-Oxocarboxylic acid metabolism
Biosynthesis of amino acids
Sulfur metabolism
*
Pyruvate metabolism
Metabolic pathways
Carbon metabolism
Quorum sensing
Lysine degradation
Cysteine and methionine metabolism
Biosynthesis of secondary metabolites
Starch and sucrose metabolism
**
Two-component system
Valine, leucine and isoleucine biosynthesis
Glycine, serine and threonine metabolism
Biofilm formation - Escherichia coli
Microbial metabolism in diverse environments
ABC transporters
Legend
Formalin
Phenol
1.00
0
Ratio

others indirectly suggests the importance of the gene for adapting to the surrounding environment. The advance in transcriptomic sequencing technologies can provide valuable information for profiling the sequence dynamics from the survived bacteria even in the same cultures; however, most studies have focussed on mRNA expression in different environments using transcriptomic results (Le et al., 2021; Lee et al., 2021; Montánchez et al., 2019; Yoon et al., 2020). Hence, this study attempted to observe SNPs and/or mutations in the same strain with different environments and exposure times through raw sequencing files that had been used for transcriptomic studies (Lee et al., 2021; Montánchez et al., 2019; Yoon et al., 2020).

In general, GWAS have been performed using genomic information rather than the sequencing for transcriptomes. However, sequences from RNA-seq can provide an alternative source of host genetic information. For example, Berthouly-Salazar et al. (2016) have used transcriptomes easier to obtain from a non-model plant for investigating SNPs and genotyping as an alternative strategy, and Chandhini and Rejish Kumar (2019) also found that transcriptomes obtained from NGSs, such as genomic results, can serve as resources to excavate SNPs and other molecular makers. Moreover, as transcripts have been produced from only living bacteria, the result of transcriptomes has the advantage that it mainly comes from survived bacteria adapted to the changed environment regardless of the elapsed incubation time. On the other hand, one drawback of utilizing raw RNA-seq files for GWAS is that the sequencing depth can fluctuate because of the differences in bacterial mRNA expression between groups, which can cause different depth levels and negatively impact the results. Nevertheless, strategies for deep sequencing of bacterial cultures can identify the occurrence of mutations even in a minority of the major population, and this provides potential insight into bacterial evolution and adaptation with the respect to genomes. For example, the VSF and DP scores in Tables 2–4 indicate the diversity of sequence and depth rate at each position. The higher VSF with deeper DP can be interpreted as the more reliable SNPs or mutated position because it meant the more different nucleotides and sequences were mapped onto a reference genome (Danecek et al., 2011). Also, as a small number of samples used in this study can lead to the risk of false-negative (type 2 error) if Bonferroni correction would be applied, the *p*-value less than .05 without Bonferroni correction, which thresholds also used in some studies, has been used in this study (Asif et al., 2021; Kap et al., 2016; Lu et al., 2015; Zeng et al., 2015). Instead, the FDR value has been described in Tables 2–4 for evaluating the reliability of each SNP. Hence, if we keep this in mind for the interpretation of results, RNA-seq can demonstrate the overall genomic changes under different environments in pathogenic bacteria in aquaculture.

When *E. piscicida* cultured in the liquid media (broth) and a minimal sub-inhibitory concentration of formalin and phenol were compared (Yoon et al., 2020), TcfC and MrdA were the top three genes that harboured multiple SNPs and/or mutated sequences (Figure S4). In general, Tcf (typhi colonization factor) has been known to exhibit chaperone-usher fimbriae and is classified in α fimbrial clade, and its regions are composed of six operons including chaperone (TcfA), major fimbriae subunit (TcfB), usher (TcfC) and adhesion (TcfD) (Azriel et al., 2017; Leclerc et al., 2016; Yue et al., 2012). Given that Tcf has been reported as being part of the type six secretion systems and atypical fimbriae in some species, particularly *Salmonella*, it can be an

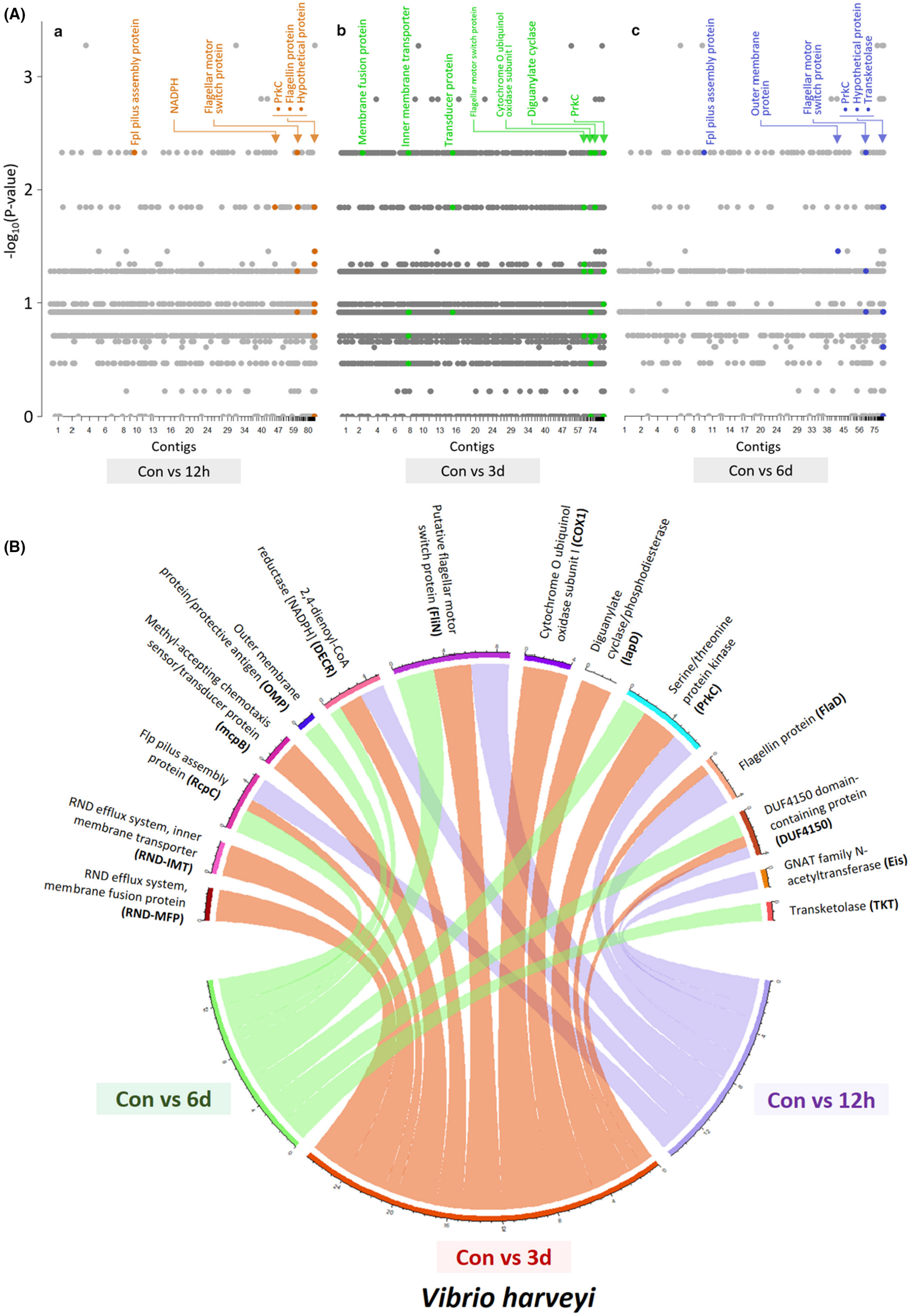

(A)
a
b
c
-log10(P-value)
0
1
2
3
Fpl pilus assembly protein
NADPH
Flagellar motor switch protein
• PrkC
• Flagellin protein
• Hypothetical protein
Membrane fusion protein
Inner membrane transporter
Transducer protein
Flagellar motor switch protein
Cytochrome O ubiquinol oxidase subunit I
Diguanylate cyclase
PrkC
Fpl pilus assembly protein
Outer membrane protein
Flagellar motor switch protein
• PrkC
• Hypothetical protein
• Transketolase
1 2 4 6 8 10 13 16 20 24 29 34 40 47 59 80
1 2 4 6 8 10 13 16 20 24 29 34 40 47 57 74
1 2 4 6 8 11 14 17 20 24 29 33 38 45 56 75
Contigs
Con vs 12h
Con vs 3d
Con vs 6d
(B)
2,4-dienoyl-CoA reductase [NADPH] (DECR)
Putative flagellar motor switch protein (FliN)
Cytochrome O ubiquinol oxidase subunit I (COX1)
Diguanylate cyclase/phosphodiesterase (lapD)
Serine/threonine protein kinase (PrkC)
Flagellin protein (FlaD)
DUF4150 domain-containing protein (DUF4150)
GNAT family N-acetyltransferase (Eis)
Transketolase (TKT)
Outer membrane protein/protective antigen (OMP)
Methyl-accepting chemotaxis sensor/transducer protein (mcpB)
Flp pilus assembly protein (RcpC)
RND efflux system, inner membrane transporter (RND-IMT)
RND efflux system, membrane fusion protein (RND-MFP)
Con vs 6d
Con vs 12h
Con vs 3d
Vibrio harveyi

FIGURE 3 Manhattan plot of genome-wide association analysis (GWAS) in *Vibrio harveyi* with 12 h ((a)-a), 3 days ((a)-b) and 6 days ((a)-c) seawater at 30°C incubation. Circos plot for the genes that have the first to third highest SNPs in one group. The thickness of the green and pink coloured lines denotes the number of *V. harveyi* SNPs at each time point (12 h, 3 and 6 days) (b). The coloured dots (brown, green and blue; a) indicate the genes that contained the first to the third highest number of significant SNPs in each time point ($p < .05$)

important virulence factor and supporting bacterial colonization in *E. piscicida* (Azriel et al., 2017; Folkesson et al., 1999). Former studies (e.g., Azriel et al., 2017; Leclerc et al., 2016) have reported that Tcf is the factor involved in host specificity, such as serotypes in *Salmonella enterica*. Switching the gene sequences involved in serotypes is noteworthy because the mutation in a Tcf region of *E. piscicida* suggested one strain can be differentiated into several serotypes under different culturable environments. Similarly, type 3 secretion system (T3SS) is one of the well-known virulence factors that contribute to invading and surviving in host cells through flagellar-like needles with injecting effector proteins, and T3SS in *E. piscicida* has been found to be significantly up-regulated under the sub-inhibitory concentration of formalin (0.38 mM) (Hou et al., 2017; Yoon et al., 2020). Notably, the group that had mutated or variant nucleic acid sequences in LcrD was broth cultured *E. piscicida* rather than formalin and phenol groups. Specifically, there was little change in the nucleotide sequence in an environment that stimulate the expression of T3SS, but a higher percentage of sequence changes was observed in broth media that does not require the high expression of T3SS. Given that numerous studies (e.g., Almaguer-Chávez et al., 2011; Barrick et al., 2009) have observed the decreasing virulence of bacteria when pathogenic bacteria are sub-cultured continuously, the sequence change of T3SS might affect the virulence of *E. piscicida*. Barrick et al. (2009) noted that *Escherichia coli* has continuously evolved through long-term sub-cultures with deletions, mutations, inversions and duplications in its genome that resulted in the loss of 1.2% of the original chromosome. This implies that the observation of mutations and/or SNPs in T3SS that has been known to the important virulence factor and factors exposed to the external environment in *E. piscicida* from the broth group can be the process of evolution to be more adaptable in the liquid media. Further study about the influence of T3SS mutation in a nutrient-enriched environment is necessary from the perspective of bacterial pathogenicity and environmental adaptation. In this context, the plethora of mutations in MrdA can be interpreted similarly to T3SS. Peptidoglycan is one of the major components of bacterial cell walls, and D-,D-transpeptidase is well-known for its strong involvement in crossing-linking peptidoglycan strands (Hugonnet et al., 2016; Triboulet et al., 2013). The peptidoglycan layer can help to maintain bacteria structures in the face of internal and outer pressures, and physiological impacts can exert on peptidoglycan cross-linking (Pidgeon et al., 2019). In particular, because D-,D-transpeptidase as penicillin-binding proteins can be easily inhibited by β-lactam antibiotics, several studies have reported that L-,D-transpeptidase, which is highly resistant to β-lactams, can be replaced when bacteria need to adapt β-lactam antibacterial agents (e.g., Pidgeon et al., 2019; Triboulet et al., 2013). Based on all these studies and the results of this study, it is speculated that D-,D-transpeptidase in *E. piscicida* is an important gene that exhibits multiple alternations and/or changes depending on environmental changes and differences. In addition, the number of genes with significant SNPs or mutations belonging to certain KEGG pathways (e.g., bacterial chemotaxis, flagellar assembly and galactose metabolisms) differed between the sub-inhibitory concentration of formalin and phenol groups. All these results support the claim that the *E. piscicida* genome can be flexibly changed by the external environment and mutations or SNPs do not occur with the same probability among all genes but instead are concentrated on specific genes.

The comparison of mutation or SNP sites in *V. harveyi*, although under the same environment (sea water at 30°C), indicated the different incubation times significantly affected the patterns of mutation. The number of SNPs that exceeded 0.01 MAF identified in the 3 days group was 6678, whereas 12 h and 6 days groups had less than half that number. This difference would have resulted in the 3 days group in most KEGG pathways having genes with at least one significant SNP or mutation site compared to others. The genes involved in flagella and flagellin components exhibited several mutations during the samplings, and these patterns were more notable as time elapsed (Figure S5). Namely, the sequence of the control group did not necessarily mean that the sequence of mutations or SNPs in the regions where mutations frequently occurred was not observed. For instance, although putative flagellar motor switch protein and flagellin protein in the control group of *V. harveyi* also have contained a piece of minor variants, the number of mutated sequences is much higher enough to switch the major sequence after incubation in high-temperature sea water (Figure S5). In addition, the variances of sequences were not identical in each group, taking into account that several 'T' were identified in the 325 position (contig 63) from only one sample belonging to the 6 days group. The mutation and differences of flagellin and flagella relevant genes may be the reason for the alternation of phenotypic characteristics. The flagellar motor switch proteins are known to promote motility and biosynthesis in flagella, and their inactivation can cause decreasing velocity (swimming ability) in highly viscous media (1% methylcellulose) and infectivity in *Borrelia burgdorferi* (Li et al., 2010). In addition, Kim et al. (2014) reported that *Vibrio vulnificus* which has a few nucleotide deletions of flagellin(s) affected the significant reduction in cytotoxicity, motility and adhesion. All these results suggest that although there may be genes and regions where mutations in bacteria occur frequently and randomly, environmental differences can create greater diversity and phenotypic characteristics. In the case of a resistance nodulation division (RND) where the system is responsible for transporting dyes, antibiotics, host molecules and detergents located between the inner and outer membrane in gram-negative bacteria, some mutations were identified in the 3 days group (Blair and Piddock, 2009). However, given that these mutations were not consistent until the 6 days group, some sequences in certain genes exhibited greater variation depending on sampling time points, although not all mutated sequences are necessarily meant to become major populations continuously.

TABLE 3 Location, a reference sequence, A2 (major allele sequence)/A1 (minor allele sequence), chi-square, and *p*-value of significant SNPs from GWAS for *Vibrio harveyi* incubated in seawater at 30°C for 12 h, 3 and 6 days, and mapping information (MI) including depth (DP) and variant sequence frequency (VSF) of each SNP

			Con vs. 12 h						Con vs. 3 days						Con vs. 6 days					
			GWAS				MI		GWAS				MI		GWAS				MI	
Gene (abbreviation)	SNP site contig)	Ref	A2/A1	CHISQ	P	FDR	DP	VSF (%)	A2/A1	CHISQ	P	FDR	DP	VSF (%)	A2/A1	CHISQ	P	FDR	DP	VSF (%)
RND efflux system, membrane fusion protein (RND-MFP)	218,971 (2)	T	–				66	0	**T/C**	**8**	**0.005**	**0.092**	**51**	**2**	–				64	0
	219,094 (2)	A	–				35	3	**A/T**	**8**	**0.005**	**0.092**	**32**	**7**	–				37	3
	219,098 (2)	T	–				35	0	**T/C**	**8**	**0.005**	**0.092**	**32**	**3**	–				36	0
RND efflux system, inner membrane transporter (RND-IMT)	51,394 (8)	C	–				25	0	C/T	8	0.005	0.092	5	20	–				15	7
	51,397 (8)	A	–				25	4	A/T	8	0.005	0.092	6	20	-				16	0
	51,789 (8)	T	–				21	0	T/G	8	0.005	0.092	10	11	–				37	0
Methyl-accepting chemotaxis sensor/ transducer protein (mcpB)	83,161 (15)	T/A	–				31	0	**A/G**	**8**	**0.005**	**0.092**	**87**	**20**	–				135	0
	83,178 (15)	T	–				24	0	**T/A**	**6**	**0.014**	**0.158**	**68**	**22**	–				110	0
	83,798 (15)	G/A	–				37	0	**A/G**	**8**	**0.005**	**0.092**	**44**	**6**	–				91	0
putative flagellar motor switch protein (fliN)	7823 (63)	T	**T/C**	**8**	**0.005**	**0.162**	**30**	**10**	T/C	6	0.014	0.158	9	22	**T/C**	**8**	**0.005**	**0.111**	**86**	**21& INDEL; TG (6)**
	7824 (63)	G	**G/T**	**8**	**0.005**	**0.162**	**30**	**3**	G/T	6	0.014	0.158	9	14	**G/T**	**8**	**0.005**	**0.111**	**85**	**7**
	8309 (63)	A	A/G	6	0.014	0.162	12	17	G/A	4	0.046	0.158	5	50	A/G	8	0.005	0.111	18	12
Cytochrome O ubiquinol oxidase subunit I (COX1)	5731 (71)	A	–				70	2	**A/G**	**8**	**0.005**	**0.092**	**42**	**5**	–				82	1
	5967 (71)	T	–				47	2	**T/C**	**8**	**0.005**	**0.092**	**31**	**7**	–				62	13
	6420 (71)	A	–				80	5	**A/C**	**8**	**0.005**	**0.092**	**47**	**11**	–				81	8
	6731 (71)	A	–				62	0	**A/G**	**8**	**0.005**	**0.092**	**32**	**3**	–				60	2
diguanylate cyclase/ phosphodiesterase (GGDEF & EAL domains) with PAS/PAC sensor(s) (lapD)	11,269 (77)	A	–				29	0	A/C	6	0.014	0.158	11	14	–				36	0
	11,718 (77)	G	–				21	0	G/A	6	0.014	0.158	5	25	–				26	0
	12,403 (77)	G	–				22	0	G/A	8	0.005	0.092	9	14	–				42	0
Serine/threonine protein kinase, regulator of stationary phase (PrkC)	21 (128)	T	A/T	4	0.046	0.162	10	20	–				2	100	T/A	6	0.014	0.144	5	60
	930 (128)	A	A/G	6	0.014	0.162	14	17	–				5	50	**A/G**	**6**	**0.014**	**0.144**	**42**	**11**
	53 (128)	C	–				15	0	C/T	8	0.005	0.092	6	17	–				20	10
	54 (128)	T	–				15	0	T/C	8	0.005	0.092	6	17	–				20	0
	57 (128)	T	–				15	0	T/G	8	0.005	0.092	6	17	–				20	0
Flagellin protein (FlaD)	325 (132)	A	**G/A**	**6**	**0.014**	**0.162**	**81**	**59**	–				52	47	–				127	56
	328 (132)	G	**A/G**	**6**	**0.014**	**0.162**	**80**	**58**	–				51	46	–				127	54
	688 (132)	G	**A/G**	**6**	**0.014**	**0.162**	**59**	**63**	**A/G**	**4.444**	**0.035**	**0.158**	**35**	**50**	–				101	51

Note: Bold letters indicated SNPs with FDR <0.25 and DP ≥30.

FIGURE 4 The ratio for the number of genes with at least one significant SNP in each time point of sea water at 30°C incubation (12 h, 3 and 6 days) belonging to KEGG pathways in *Vibrio harveyi*. The thickness of edges between nodes indicates similarity based on the number of shared genes. The yellow background with the asterisk in the node indicates significantly different counts based on chi-square analysis (*$p < .05$; **$p < .01$; ***$p < .001$)

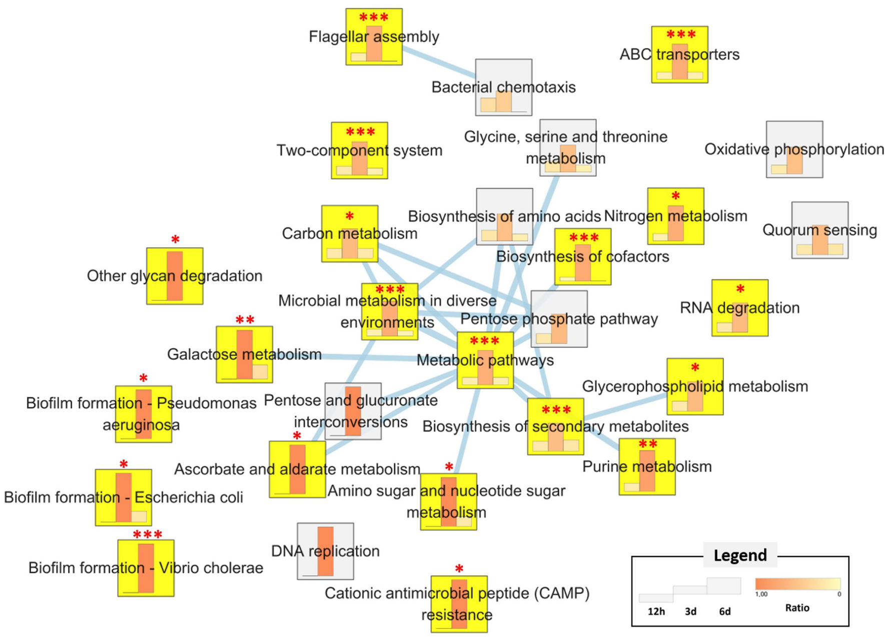
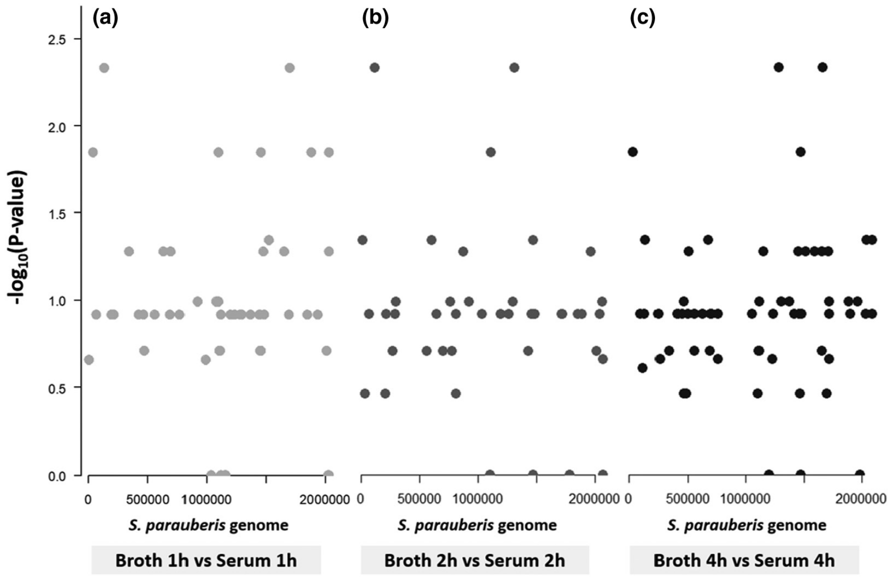

***
Flagellar assembly
***
ABC transporters
Bacterial chemotaxis
***
Two-component system
Glycine, serine and threonine metabolism
Oxidative phosphorylation
*
Carbon metabolism
Biosynthesis of amino acids
*
Nitrogen metabolism
Quorum sensing
*
Other glycan degradation
***
Biosynthesis of cofactors
***
Microbial metabolism in diverse environments
Pentose phosphate pathway
*
RNA degradation
**
Galactose metabolism
***
Metabolic pathways
*
Glycerophospholipid metabolism
*
Biofilm formation - Pseudomonas aeruginosa
Pentose and glucuronate interconversions
***
Biosynthesis of secondary metabolites
*
Ascorbate and aldarate metabolism
**
Purine metabolism
*
Biofilm formation - Escherichia coli
*
Amino sugar and nucleotide sugar metabolism
***
Biofilm formation - Vibrio cholerae
DNA replication
*
Cationic antimicrobial peptide (CAMP) resistance
Legend
12h
3d
6d
1.00
0
Ratio

FIGURE 5 Manhattan plot of genome-wide association analysis (GWAS) in *Streptococcus parauberis* for 1 h (a), 2 h (b) and 4 h (c) incubation



Regarding the comparison of *S. parauberis*, the correlation of missing genotypes between samples was well-clustered in broth and serum groups. However, no gene contained several significant mutations or SNPs, as observed in *E. piscicida* and *V. harveyi*. Because *S. parauberis* cultured in broth and serum for 4 h did not exhibit less than two times the high number of bacteria compared to 0 h (Lee et al., 2021), it is suggested that the incubation times were not sufficient to initiate significant SNPs or mutation. However, missing genotypes might be affected by gene expression because RNA-seq, files not genome sequencing results, were used in this study. According to the results of principal investigation analysis based on transcriptomic results, despite the big difference in transcriptomic responses as incubation time elapsed, closer clustering distance by incubation time rather than a different environment suggested that environmental pressure can be a much stronger factor in initiating the missing sequence or deletions.

In conclusion, this study first investigated SNPs or occurring mutations in major fish bacterial pathogens from the same strain, and successfully identified genes (especially, flagella, fimbriae and flagellin relevant genes) that have a much higher number of mutated sequences than other genes under different environments. Given that multiple studies have already demonstrated fimbriae that are directly exposed to the external environment are the important gene for differentiating bacterial serotypes, these genes can be more flexible in changing their structures which can be linked to alternation in their phenotypic characteristics. In addition, other genes that contained several significant mutations and SNPs based on GWAS can be potential biomarkers for tracking the origin and sources of the same bacterial species. Although further studies are necessary to investigate the meaning of mutations in each gene and species, the data presented in this study provide valuable information and insight into the evolution of bacterial pathogens in aquaculture.

TABLE 4 Location, a reference sequence, A2 (major allele sequence)/A1 (minor allele sequence), chi-square, and *p*-value of significant SNPs from GWAS for *Streptococcus parauberis* incubated in broth and serum for 1, 2, and 4 h, and mapping information (MI) including depth (DP) and variant sequence frequence (VSF) of each SNP

			1 h						2 h						4 h					
			GWAS				MI		GWAS				MI		GWAS				MI	
Gene (abbreviation)	SNP site	Ref	A2/A1	CHISQ	P	FDR	DP	VSF (%)	A2/A1	CHISQ	P	FDR	DP	VSF (%)	A2/A1	CHISQ	P	FDR	DP	VSF (%)
–	11,965	T	–				4	0	A/T	4.00	0.046	0.186	2	50	–				3	0
–	34,946	T	–				13	0	–				7	0	T/C	6.00	0.014	0.172	4	25
–	37,322	T	T/A	6.00	0.014	0.11	6	17	–				1	0	–				4	0
tRNA-Cys-GCA (NSUN6)	116,816	A	–				58	48	G/A	8.00	0.005	0.101	23	70	–				36	50
–	135,377	T	T/C	8.00	0.005	0.11	8	13	–				1	0	–				3	0
–	138,394	A	–				3	0	–				4	0	T/A	4.00	0.046	0.172	2	50
–	603,666	A	–				15	0	A/T	4.00	0.046	0.186	2	50	–				9	0
–	677,993	T	–				13	0	–				5	0	C/T	4.00	0.046	0.172	3	33
Replication initiator protein A (DnaA)	1,099,212	TAA	TAA/TAAA	6.00	0.014	0.11	1	20	–				6	0	–				3	0
Hypothetical protein	1,111,230	C	–				5	0	C/G	6.00	0.014	0.186	4	25	–				5	0
6-phospho-beta glucosidase (lacG)	1,288,372	A	–				4	0	–				6	0	A/G	8.00	0.005	0.152	10	10
Hypothetical protein	1,309,621	C	–				43	0	C/T	8.00	0.005	0.101	10	10	–				15	0
Phage tail length tape-measure protein T (yomI)	1,460,848	T	T/C	6.00	0.014	0.11	4	25	–				3	0	–				2	0
Phage terminase arge subunit (A)	1,471,143	T	–				3	0	T/C	4.00	0.046	0.186	2	50	T/C	6.00	0.014	0.172	3	33
–	1,526,293	T	A/T	4.00	0.046	0.168	2	50	–				1	0	–				2	0
–	1,526,442	A	G/A	4.00	0.046	0.168	2	50	–				2	0	–					
Hypothetical protein	1,662,165	G	–				6	0	–				4	0	G/A	8.00	0.005	0.152	4	25
Hypothetical protein	1,706,945	CTTTTT	CTTTTT/ CTTT	8.00	0.005	0.11	1	4	–				9	0	–				5	0
Tn5252, relaxase (Tn5252)	1,885,284	T	T/C	6.00	0.014	0.11	6	17	–				4	0	–				2	0
Hypothetical protein	2,037,162	T	T/C	6.00	0.014	0.11	3	33	–				1	0	–				1	0
PTS system, cellobiose-specific IIC component (celB)	2,040,252	T	–				7	0	–				3	0	T/C	4.00	0.046	0.172	2	50
Phosphoribosylformylglycinamidine synthase (purL)	2,089,390	ATTT	–				4	0	–				3	0	ATT/ATTT	4.00	0.046	0.172	2	50

Note: Bold letters indicated SNPs with FDR <0.25 and DP ≥30.

ACKNOWLEDGEMENTS

The author thanks Institute of Marine Research (Havforskningsinstituttet) for funding the open access of this article.

CONFLICT OF INTEREST

The author declares no competing interest exists.

DATA AVAILABILITY STATEMENT

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

REFERENCES

Algammal, A. M., Mabrok, M., Ezzat, M., Alfifi, K. J., Esawy, A. M., Elmasry, N., & El-Tarabili, R. M. (2022). Prevalence, antimicrobial resistance (AMR) pattern, virulence determinant and AMR genes of emerging multi-drug resistant *Edwardsiella tarda* in Nile tilapia and African catfish. *Aquaculture*, *548*(737), 643.

Almaguer-Chávez, J. A., Welsh, O., Lozano-Garza, H. G., Said-Fernández, S., Romero-Díaz, V. J., Ocampo-Candiani, J., & Vera-Cabrera, L. (2011). Decrease of virulence for BALB/c mice produced by continuous subculturing of *Nocardia brasiliensis*. *BMC Infectious Diseases*, *11*, 1–9.

Asif, H., Alliey-Rodriguez, N., Keedy, S., Tamminga, C. A., Sweeney, J. A., Pearlson, G., Clementz, B. A., Keshavan, M. S., Buckley, P., & Liu, C. (2021). GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Molecular Psychiatry*, *26*, 2048–2055.

Azriel, S., Goren, A., Shomer, I., Aviv, G., Rahav, G., & Gal-Mor, O. (2017). The typhi colonization factor (Tcf) is encoded by multiple non-typhoidal *salmonella* serovars but exhibits a varying expression profile and interchanging contribution to intestinal colonization. *Virulence*, *8*, 1791–1807.

Barrick, J. E., Yu, D. S., Yoon, S. H., Haeyoung, J., Tae Kwang, O., Schneider, D., Lenski, R. E., & Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia Coli*. *Nature*, *461*, 1243–1247.

Berthouly-Salazar, C., Mariac, C., Couderc, M., Pouzadoux, J., Floc'h, J., & Vigouroux, Y. (2016). Genotyping-by-sequencing SNP identification for crops without a reference genome: Using transcriptome based mapping as an alternative strategy. *Frontiers in Plant Science*, *7*, 777.

Blair, J. M., & Piddock, L. J. (2009). Structure, function and inhibition of RND efflux pumps in Gram-negative bacteria: an update. *Current opinion in microbiology*, *12*, 512–519.

Blom, J., Kreis, J., Spänig, S., Juhre, T., Bertelli, C., Ernst, C., & Goesmann, A. (2016). EDGAR 2.0: An enhanced software platform for comparative gene content analyses. *Nucleic Acids Research*, *44*, W22–W28.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., & Pusch, G. D. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, *5*, 1–6.

Chandhini, S., & Rejish Kumar, V. J. (2019). Transcriptomics in aquaculture: Current status and applications. *Reviews in Aquaculture*, *11*, 1379–1397.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*, s13742-015-0047-8.

Chen, P. E., & Shapiro, B. J. (2015). The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology*, *25*, 17–24.

Chevallereau, A., Meaden, S., van Houte, S., Westra, E. R., & Rollie, C. (2019). The effect of bacterial mutation rate on the evolution of CRISPR-Cas adaptive immunity. *Philosophical Transactions of the Royal Society B*, *374*, 20180094.

Correa, K., Lhorente, J. P., López, M. E., Bassini, L., Naswa, S., Deeb, N., Di Genova, A., Maass, A., Davidson, W. S., & Yáñez, J. M. (2015). Genome-wide association analysis reveals loci associated with resistance against *Piscirickettsia salmonis* in two Atlantic salmon (*Salmo salar* L.) chromosomes. *BMC Genomics*, *16*, 1–9.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., & Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158.

Danecek, P., Schiffels, S., & Durbin, R. (2016). *Multiallelic calling model in bcftools (−m)*. Program and documentation distributed by the author. https://samtools.github.io/bcftools/call-m.pdf

Figueroa, J., Cárcamo, J., Yañez, A., Olavarria, V., Ruiz, P., Manríquez, R., Muñoz, C., Romero, A., & Avendaño-Herrera, R. (2019). Addressing viral and bacterial threats to salmon farming in Chile: Historical contexts and perspectives for management and control. *Reviews in Aquaculture*, *11*, 299–324.

Folkesson, A., Advani, A., Sukupolvi, S., Pfeifer, J. D., Normark, S., & Löfdahl, S. (1999). Multiple insertions of fimbrial operons correlate with the evolution of *salmonella* serovars responsible for human disease. *Molecular Microbiology*, *33*, 612–622.

Frenoy, A., & Bonhoeffer, S. (2018). Death and population dynamics affect mutation rate estimates and evolvability under stress in bacteria. *PLoS Biology*, *16*, e2005056.

Holborn, M. K., Ang, K. P., Elliott, J., Powell, F., & Boulding, E. G. (2018). Genome wide association analysis for bacterial kidney disease resistance in a commercial north American Atlantic salmon (*Salmo salar*) population using a 50 K SNP panel. *Aquaculture*, *495*, 465–471.

Hou, M., Chen, R., Yang, D., Núñez, G., Wang, Z., Wang, Q., Zhang, Y., & Liu, Q. (2017). Identification and functional characterization of EseH, a new effector of the type III secretion system of *Edwardsiella piscicida*. *Cellular Microbiology*, *19*, e12638.

Hugonnet, J., Mengin-Lecreulx, D., Monton, A., den Blaauwen, T., Carbonnelle, E., Veckerle, C., Yves, V. B., van Nieuwenhze, M., Bouchier, C., & Tu, K. (2016). Factors essential for L, D-transpeptidase-mediated peptidoglycan cross-linking and β-lactam resistance in *Escherichia coli*. *Elife*, *5*, e19469.

Jaruszewicz-Błońska, J., & Lipniacki, T. (2017). Genetic toggle switch controlled by bacterial growth rate. *BMC Systems Biology*, *11*, 1–11.

Kap, E. J., Seibold, P., Scherer, D., Habermann, N., Balavarca, Y., Jansen, L., Zucknick, M., Becker, N., Hoffmeister, M., & Ulrich, A. (2016). SNPs in transporter and metabolizing genes as predictive markers for oxaliplatin treatment in colorectal cancer patients. *International Journal of Cancer*, *138*, 2993–3001.

Kim, B. S., Huh, M. D., & Roh, H. (2021). The complete genome of *Nocardia seriolae* MH196537 and intra-species level as analyzed by comparative genomics based on random Forest algorithm. *Current Microbiology*, *78*, 2391–2399.

Kim, S. Y., Thanh, X. T. T., Jeong, K., Kim, S. B., Pan, S. O., Jung, C. H., Hong, S. H., Lee, S. E., & Rhee, J. H. (2014). Contribution of six flagellin genes to the flagellum biogenesis of *Vibrio vulnificus* and in vivo invasion. *Infection and Immunity*, *82*, 29–42.

Kim, Y., Gu, C., Kim, H. U., & Lee, S. Y. (2020). Current status of pan-genome analysis for pathogenic bacteria. *Current Opinion in Biotechnology*, *63*, 54–62.

Kjærner-Semb, E., Edvardsen, R. B., Ayllon, F., Vogelsang, P., Furmanek, T., Rubin, C. J., Veselov, A. E., Nilsen, T. O., McCormick, S. D., & Primmer, C. R. (2021). Comparison of anadromous and landlocked Atlantic salmon genomes reveals signatures of parallel and relaxed selection across the northern hemisphere. *Evolutionary Applications*, *14*, 446–461.

Krueger, F. (2021). *Trim Galore. GitHub repository*. GitHub https://github.com/FelixKrueger/TrimGalore.com/fenderglass/Flye

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, *9*, 357–359.

Le, C. T., Price, E. P., Sarovich, D. S., Nguyen, T. T., Vu-Khac, H., Kurtböke, I., Knibb, W., Chen, S., & Katouli, M. (2021). Simple and cost-effective SNP genotyping method for discriminating subpopulations of the fish pathogen, *Nocardia seriolae*. *bioRxiv*. https://doi.org/10.1101/2021.12.28.474260

Leclerc, J., Quevillon, E., Houde, Y., Paranjape, K., Dozois, C. M., & Daigle, F. (2016). Regulation and production of Tcf, a cable-like fimbriae from *salmonella enterica* serovar *typhi*. *Microbiology*, *162*, 777–788.

Lee, Y., Kim, N., Roh, H., Kim, A., Han, H., Cho, M., & Kim, D. (2021). Transcriptome analysis unveils survival strategies of *streptococcus parauberis* against fish serum. *PLoS One*, *16*, e0252200.

Lees, J. A., & Bentley, S. D. (2016). Bacterial GWAS: Not just gilding the lily. *Nature Reviews Microbiology*, *14*, 406.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Li, C., Xu, H., Zhang, K., & Liang, F. T. (2010). Inactivation of a putative flagellar motor switch protein FliG1 prevents Borrelia burgdorferi from swimming in highly viscous media and blocks its infectivity. *Molecular microbiology*, *75*, 1563–1576.

Lu, J., Chu, P., Wang, H., Jin, Y., Han, S., Han, W., Tai, J., Guo, Y., & Ni, X. (2015). Candidate gene association analysis of neuroblastoma in Chinese children strengthens the role of LMO1. *PLoS One*, *10*, e0127856.

Merico, D., Isserlin, R., Stueker, O., Emili, A., & Bader, G. D. (2010). Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. *PLoS One*, *5*, e13984.

Montánchez, I., Ogayar, E., Plágaro, A. H., Esteve-Codina, A., Gómez-Garrido, J., Orruño, M., Arana, I., & Kaberdin, V. R. (2019). Analysis of *Vibrio harveyi* adaptation in sea water microcosms at elevated temperature provides insights into the putative mechanisms of its persistence and spread in the time of global warming. *Scientific Reports*, *9*, 1–12.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, *35*, W182–W185.

Nadella, R. K., Panda, S. K., Badireddy, M. R., Kurcheti, P. P., Raman, R. P., & Mothadaka, M. P. (2022). Multi-drug resistance, integron and transposon-mediated gene transfer in heterotrophic bacteria from *Penaeus vannamei* and its culture environment. *Environmental Science and Pollution Research*, *29*, 37527–37542.

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528.

Pidgeon, S. E., Apostolos, A. J., Nelson, J. M., Shaku, M., Rimal, B., Islam, M. N., Crick, D. C., Kim, S. J., Pavelka, M. S., & Kana, B. D. (2019). L, d-transpeptidase specific probe reveals spatial activity of peptidoglycan cross-linking. *ACS Chemical Biology*, *14*, 2185–2196.

Rasmussen, B. B., Grotkjær, T., D'Alvise, P. W., Yin, G., Zhang, F., Bunk, B., Spröer, C., Bentzon-Tilia, M., & Gram, L. (2016). Vibrio anguillarum is genetically and phenotypically unaffected by long-term continuous exposure to the antibacterial compound tropodithietic acid. *Applied and Environmental Microbiology*, *82*, 4802–4810.

Roh, H., Kim, B. S., Lee, M. K., Park, C., & Kim, D. (2020). Genome-wide comparison of *Carnobacterium maltaromaticum* derived from diseased fish harbouring important virulence-related genes. *Journal of Fish Diseases*, *43*, 1029–1037.

Roh, H., & Kim, D. (2021). Genotypic and phenotypic characterization of highly alkaline-resistant *Carnobacterium maltaromaticum* V-type ATPase from the dairy product based on comparative genomics. *Microorganisms*, *9*, 1233.

Roh, H. J., Kim, A., Kang, G. S., & Kim, D. (2016). Photoinactivation of major bacterial pathogens in aquaculture. *Fisheries and Aquatic Sciences*, *19*, 1–7.

Roh, H. J., Kim, B., Kim, A., Kim, N. E., Lee, Y., Chun, W., Ho, T. D., & Kim, D. (2019). Whole-genome analysis of multi-drug-resistant *Aeromonas veronii* isolated from diseased discus (*Symphysodon discus*) imported to Korea. *Journal of Fish Diseases*, *42*, 147–153.

Rouli, L., Merhej, V., Fournier, P., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, *7*, 72–85.

Seo, J. S., Kwon, M., Youn Hwang, J., Don Hwang, S., Kim, D., Bae, J., Ha Park, K., & Lee, J. (2021). Estimation of pharmacological properties of ceftiofur, an injectable cephalosporin antibiotic, for treatment of streptococcosis in cultured olive flounder *Paralichthys olivaceus*. *Aquaculture Research*, *52*, 831–841.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*, 2498–2504.

Triboulet, S., Dubée, V., Lecoq, L., Bougault, C., Mainardi, J., Rice, L. B., Ethève-Quelquejeu, M., Gutmann, L., Marie, A., & Dubost, L. (2013). Kinetic features of L, D-transpeptidase inactivation critical for β-lactam antibacterial activity. *PLoS One*, *8*, e67831.

Turner, S. D. (2014). Qqman: An R package for visualizing GWAS results using QQ and Manhattan plots. *bioRxiv*, *2014*, 005165. https://doi.org/10.1101/005165

Wang, Q., Lu, Q., & Zhao, H. (2015). A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Frontiers in Genetics*, *6*, 149.

Warnes, G. R., Bolker, B., Lumley, T., & Johnson, R. C. (2015). *Gmodels: Various R programming tools for model fitting*. R Package Version **2**.

Westra, E. R., Sünderhauf, D., Landsberger, M., & Buckling, A. (2017). Mechanisms and consequences of diversity-generating immune strategies. *Nature Reviews Immunology*, *17*, 719–728.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Yoon, J. B., Hwang, S., Baek, S., Lee, S., Bang, W. Y., & Moon, K. H. (2020). In vitro *Edwardsiella piscicida* ck108 transcriptome profiles with subinhibitory concentrations of phenol and formalin reveal new insights into bacterial pathogenesis mechanisms. *Microorganisms*, *8*, 1068.

Yue, M., Rankin, S. C., Blanchet, R. T., Nulton, J. D., Edwards, R. A., & Schifferli, D. M. (2012). Diversification of the *salmonella* fimbriae: A model of macro-and microevolution. *PLoS One*, *7*, e38596.

Zekic, T., Holley, G., & Stoye, J. (2018). Pan-genome storage and analysis techniques. *Methods in Molecular Biology*, *1704*, 29–53.

Zeng, P., Zhao, Y., Qian, C., Zhang, L., Zhang, R., Gou, J., Liu, J., Liu, L., & Chen, F. (2015). Statistical analysis for genome-wide association study. *Journal of Biomedical Research*, *29*, 285–297.

Zhang, J., Hu, Y., Sun, Q., Li, X., & Sun, L. (2021). An inactivated bivalent vaccine effectively protects turbot (*Scophthalmus maximus*) against vibrio anguillarum and Vibrio harveyi infection. *Aquaculture*, *544*(737), 158.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Roh, H. (2022). A genome-wide association study of the occurrence of genetic variations in *Edwardsiella piscicida*, *Vibrio harveyi*, and *Streptococcus parauberis* under stressed environments. *Journal of Fish Diseases*, *45*, 1373–1388. https://doi.org/10.1111/jfd.13668