Food Microbiology 84 (2019) 103270

Contents lists available at ScienceDirect

Food Microbiology

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

Genetic diversity, virulence factors and farm-to-table spread pattern of *Vibrio parahaemolyticus* food-associated isolates

Chao Yang[a,b], Xianglilan Zhang[a], Hang Fan[a], Yinghui Li[b], Qinghua Hu[b], Ruifu Yang[a], Yujun Cui[a,*]

[a] State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing, 100071, China
[b] Shenzhen Centre for Disease Control and Prevention, Shenzhen, 518055, China

ARTICLE INFO

Keywords:
Vibrio parahaemolyticus
Genetic diversity
Virulence factors
Farm-to-table spread

ABSTRACT

Vibrio parahaemolyticus is the leading bacterial cause of seafood-associated gastroenteritis worldwide. Moreover, infections and outbreaks caused by *V. parahaemolyticus* has kept increasing over the last two decades. In this study, we investigated the genetic diversity, virulence factors and farm-to-table spread pattern of *V. parahaemolyticus* by analyzing 383 genomes of food-associated isolates. These strains were isolated from diverse sample types from six provinces of China in 2014, being classified into three tiers of the farm-to-table spread process: food production, circulation and consumption. The genetic diversity of *V. parahaemolyticus* in different classifications, including geographical location, sample type, source and spread tier, was similar, as the median number of pairwise SNPs within each classification was between 33,013 and 33,659. Specifically, there was no clear boundaries in genetic diversity of the isolates from inland vs. coastal provinces, as well as of those from freshwater vs. seawater products. Moreover, the virulence genes and genomic islands were only found in a small number of isolates, indicating a low disease risk of the food-associated isolates in this study. By further exploring 28 recently emerged clonal groups, we identified seven farm-to-table spread events, showing a common pattern of single-source radial spread accompanied with occasional gene gain/loss events. Generally speaking, our work highlighted the colonization of *V. parahaemolyticus* in inland provinces and freshwater environment, and provided a snapshot of the farm-to-table spread pattern of *V. parahaemolyticus* food-associated isolates. Our results showed the feasibility of tracking the farm-to-table spread of foodborne pathogen, which would help construct the whole genome sequencing-based molecular tracking network in the future.

1. Introduction

Vibrio parahaemolyticus is a natural inhabitant of estuarine, marine and coastal environments, being frequently isolated from many types of seafood. It is the leading bacterial cause of seafood-associated gastroenteritis worldwide (Baker-Austin et al., 2018; Nair et al., 2007; Su and Liu, 2007). Consumption of seafood, particularly of the shellfishes contaminated by *V. parahaemolyticus* beforehand, is the common cause of its infection in humans. Notably, compared to other foodborne pathogens, the infections and outbreaks caused by *V. parahaemolyticus* had been increasing over the last two decades (Baker-Austin et al., 2018). In the United States, there were approximately 35,000 human infections per year from 2000 to 2008 (Baker-Austin et al., 2018; Crim et al., 2015). In China, *V. parahaemolyticus* was the leading cause of foodborne disease outbreaks; more than 100 foodborne outbreaks and 2000 human infections caused by *V. parahaemolyticus* were laboratory confirmed per year from 2011 to 2016 (Liu et al., 2018).

Several molecular typing methods have been developed for *V. parahaemolyticus* typing, including serotyping, pulsed-field gel electrophoresis (PFGE), multi-locus sequence typing (MLST) and whole genome sequencing (WGS) (Y. Cui et al., 2015; DePaola et al., 2003; González-Escalona et al., 2008; Hazen et al., 2015; C. H. Lüdeke et al., 2014; C. H. M. Lüdeke et al, 2015; Lopatek et al., 2018; Nair et al., 2007; Yang et al., 2019). Compared to traditional methods, WGS provides the best resolution on reconstruction the relationships among samples and can be easily integrated with historical data. In addition, the information only provided by WGS can be used for in-depth analysis, such as detection of virulence genes and antibiotic resistance genes (Allard et al., 2016; Xavier Didelot et al., 2012; Franz et al., 2016; Ronholm et al., 2016). Given above significant advantages, WGS is becoming the most powerful method in foodborne pathogen surveillance, outbreak source tracking and risk assessments (Allard et al., 2016; Franz et al., 2016; Ronholm et al., 2016). Based on WGS, global *V. parahaemolyticus* isolates have been divided into four populations,

[*] Corresponding author.
E-mail address: cuiyujun.new@gmail.com (Y. Cui).

https://doi.org/10.1016/j.fm.2019.103270
Received 1 February 2019; Received in revised form 30 May 2019; Accepted 12 July 2019
Available online 13 July 2019

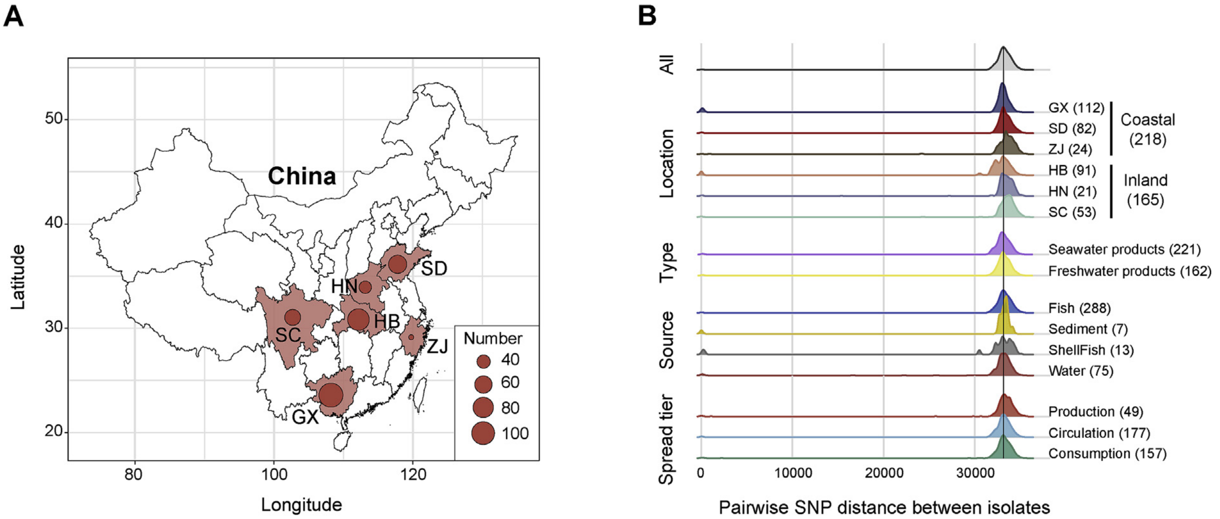

Fig. 1. Geographical distribution and genetic diversity of the 383 *V. parahaemolyticus* food-associated isolates. (A) Geographical locations of the six sampling provinces. The sampling provinces are highlighted in red background and red points, and the point size scales with the sampling number. (B) Pairwise SNP distance distribution between all food-associated isolates and between isolates of each location, type, source, and spread tier. Colors indicate different classifications. The number in brackets indicates the strain number of a classification. X-axis indicates the pairwise SNP distance and Y-axis indicates the corresponding frequency. Vertical black line indicates the median pairwise SNP distance between all isolates.

VppAsia, VppX, VppUS1 and VppUS2. The first two were found worldwide while the last two only in North America and Europe (Y. Cui et al., 2015; Yang et al., 2019).

While most *V. parahaemolyticus* strains are non-pathogenic, clinical strains are mainly attributed to several clonal groups, usually carrying the virulence factors such as thermostable direct haemolysin (*tdh*) and/or *tdh*-related haemolysin (*trh*) (Baker-Austin et al., 2018; Nair et al., 2007; Su and Liu, 2007). O3:K6 serotype and its serovariants are the most common type of strains leading to foodborne diseases worldwide (C. Han et al., 2016; D. Han et al., 2017). The disease-associated O3:K6 clonal group was firstly reported in Calcutta, India in 1996, carrying the *tdh* gene without the *trh* gene (Nair et al., 2007; Su and Liu, 2007). Soon after its emergence, it rapidly spread throughout Southeast Asian countries, then spread to the America and Europe in 1997, and subsequently to Africa in 2004, making it a pandemic group (Ansaruzzaman et al., 2005; Bag et al., 1999; Martinez-Urtaza et al., 2004, 2005). In 2012, the highly pathogenic sequence type 36 (ST36) clonal group was first found to spread from the Pacific Northwest to Northeast coast in the United States, and then to the Northwest Spain, becoming the second cross-continental spread isolates of this species. Notably, ST36 isolates carried both of the *tdh* and *trh* genes, which could be associated with the increased virulence (Martinez-Urtaza et al., 2017, 2018).

In addition to the well-known virulence genes of *tdh* and *trh*, genome sequencing revealed that two type III secretion systems (T3SS1 and T3SS2) were also related to the pathogenicity of *V. parahaemolyticus* (Makino et al., 2003). T3SS1 genes were ubiquitous in *V. parahaemolyticus* and were associated with the cytotoxic activity; while T3SS2 genes were mainly found in clinical isolates and were associated with the enterotoxicity (Ham and Orth, 2012; K.-S. Park et al., 2004a, 2004b; Zhang et al., 2012). There are two versions of T3SS2, including T3SS2α and T3SS2β. The *tdh* gene and T3SS2α encoding genes are adjacent to each other, both locating on the confirmed pathogenicity island VPaI-7 (see below); while the *trh* gene and T3SS2β genes are adjacent, both locating on the homologous genome island of VPaI-7. (Boyd et al., 2008; Hurley et al., 2006; Okada et al., 2009). Two types VI secretion systems, including T6SS1 and T6SS2, were identified in *V. parahaemolyticus*. T6SS1 was more frequently found in clinical isolates than environmental isolates, contributing to the adhesion to host cells (Y. Yu et al., 2012). However, the role of T6SS1 in pathogenicity has not been demonstrated (Ceccarelli et al., 2013). Comparative genomic analysis of pandemic and non-pandemic isolates revealed seven genomic islands (GIs), including VPaI-1–7. VPaI-1, VPaI-4, VPaI-5 and VPaI-6 are only found in pandemic isolates (Boyd et al., 2008; Hurley et al., 2006). These GIs range from 10 to 81 kb in size, and their average GC content is lower than that of the overall genomes, indicating that they were acquired by horizontal gene transfer (HGT). Except for the pathogenicity island of VPaI-7, three GIs, including VPaI-1, VPaI-2, and VPaI-6, encoded putative virulence genes, which were possibly involved in the pathogenicity (Boyd et al., 2008; Hurley et al., 2006).

Clinical strains only compose a small fraction of the overall diversity of *V. parahaemolyticus*. On the contrary, high level of genetic diversity was found among environmental strains (DePaola et al., 2003; González-Escalona et al., 2008; Hazen et al., 2015; C. H. Lüdeke et al., 2014; C. H. M. Lüdeke et al., 2015; Lopatek et al., 2018), which was considered as the result of high recombination rate and enormous population size (Y. Cui et al., 2015; Yan et al., 2011; Yang et al., 2018; Yang et al., 2019). Exploring the environmental reservoir of *V. parahaemolyticus* would provide a more comprehensive insight into this foodborne pathogen. Aiming to investigate the genetic diversity, virulence factors and farm-to-table spread pattern of *V. parahaemolyticus* food-associated isolates, we analyzed 383 genomes of *V. parahaemolyticus* food-associated strains, which were isolated from six provinces of China in 2014, covering the whole process from food production (farm), circulation to consumption (table).

2. Materials and methods

2.1. *V. parahaemolyticus* strain collections

All the 383 *V. parahaemolyticus* strains used in this study were isolated from food-associated samples during the routinely food safety surveillance by the provincial centers for disease control and prevention (CDC) of China in 2014, as a part of the TraNet (National Food Disease Molecular Tracing Network) and foodborne disease surveillance and outbreak reporting system (W. Li et al., 2018; Liu et al., 2018). These strains were isolated from six provinces of China, including three coastal provinces: Guangxi (GX, 112 isolates, taking up 29% of all the isolates), Shandong (SD, 91, 24%), Zhejiang (ZJ, 21, 5%), and three inland provinces: Hubei (HB, 53, 14%), Henan (HN, 82, 21%), Sichuan (SC, 24, 6%) (Fig. 1A). They were isolated from two types of samples: freshwater products (162, 42%) and seawater products (221, 58%), and four types of sources: fish (288, 75%), sediment (7, 2%), shellfish (13, 3%), aquaculture water (75, 20%). Based on detailed sampling location information, we classified these isolates into three tiers of farm-to-table spread process: production (isolates from aquaculture farm sites, 49 isolates, 13%), circulation (from aquatic market and supermarket sites, 177, 46%) and consumption (from restaurant sites, 157, 41%). The background information of these 383 *V. parahaemolyticus* food-associated isolates was listed in Supplementary Table 1.

2.2. Genome dataset, SNP calling and phylogeny construction

All the 383 genomes used in this study belong to a dataset of 1103 *V. parahaemolyticus* genomes, which had been employed in our previous studies (Yujun Cui et al., 2018; Yang et al., 2018; Yang et al., 2019) that separately focused on global population structure, co-adaptation evolution and bacterial recombination scaled effective population size. In this study, by complementing the meta-information of the food-associated isolates, including their geographical location, sample type, source and spread tier, we investigated the genetic diversity, virulence factors, and especially the spread within clonal frame of the 383 food-associated strains, to understand the farm-to-table spread pattern of *V. parahaemolyticus* food-associated isolates. SNPs were identified as previously described (Y. Cui et al., 2015; Yang et al., 2019). Briefly, the assemblies were aligned against a reference genome RIMD 2210633 (NC_004603.1, NC_004605.1) using MUMmer (Delcher et al., 2003) to generate the whole genome alignments and identify SNPs in the core genome (regions presented in all isolates). Raw sequencing reads were mapped to the assemblies to evaluate the SNP accuracy using SOAPaligner (R. Li et al., 2009). We filtered out the SNPs located in repetitive regions and with low sequence quality (quality score $<$ 20 or was covered by $<$ 10 reads). After filtering, 562,172 SNPs in total were identified from the 383 isolates, which were used in the neighbor-joining tree construction using TreeBeST 1.9.2 (http://treesoft.sourceforge.net/treebest.shtml). The population assignment information of these isolates was taken from our previous study (Yang et al., 2019). The phylogenetic tree was visualized using ggtree (G. Yu et al., 2017).

By calculating the pairwise SNP distance between all the 383 isolates, we defined 47 semi-clonal groups (SCG) among which the sequence differences are less than 2000 SNPs, with each SCG including 2 to 24 isolates. Because core genome size was negatively correlated with the number of strains, which could affect the size of the SNP set, we then recalled SNPs for each SCG to gain a higher resolution. 0–1833 SNPs were separately identified from each of the 47 SCGs. We used ClonalFrameML (X. Didelot and Wilson, 2015) to identify the recombination regions for each SCG. Whole genome alignment of each SCG was used to construct the maximum likelihood starting tree using RAxML (Stamatakis, 2014), and non-core regions were ignored in the calculation. SNPs located in the recombination regions were removed from each SCG. After filtering the recombined SNPs, we re-calculated the pairwise SNP distance between isolates of each SCG to identify clonal groups (CGs). Strains within the CGs are the derivatives of a common ancestor, thus they are appropriate candidates to trace the farm-to-table spread process through the phylogeny-based method. CG threshold was defined as pairwise SNP distance between isolates less than 10. The threshold can be more stringent, at the expense of fewer identification of farm-to-table spread events. However, even if the threshold was set to 5, the observed farm-to-table spread pattern will not change. Based on the threshold of 10, we identified 28 CGs from the 47 SCGs, with the number of strains of each CG ranging from 2 to 14. The other 19 SCGs do not include CGs under the threshold of pairwise SNP distance less than 10. We found that seven of the 28 CGs contained strains of different farm-to-table spread tiers. For each of the seven CGs, we selected one outgroup strain, i.e. the strain with minimum average SNP distance to the strains of each CG, based on the pairwise SNP distance matrix of all the 383 isolates. We repeated the SNP calling and recombination detection process for strains of each of the seven CGs and the corresponding outgroup strain, and the non-recombined SNPs were used in GrapeTree (Zhou et al., 2018) to construct the minimum spanning trees.

2.3. Virulence genes, genomic islands and accessory gene gain/loss detection

We detected the presence/absence of 13 previously described virulence genes/genomic islands (Boyd et al., 2008; Hurley et al., 2006; Makino et al., 2003; Y. Yu et al., 2012), including *tdh* gene (VPA1314)*, trh* gene (AB455531.1, 23,444–24,013)*,* T3SS1 (VP1658-VP1702), T3SS2 (T3SS2α: VPA1335-VPA1370, T3SS2β: AB455531.1, 39,778–73,263), T6SS1 (VP1386-VP1414), T6SS2 (VPA1025-VPA1046), VPaI-1 (VP0380-VP0403), VPaI-2 (VP0635-VP0643), VPaI-3 (VP1071-VP1094), VPaI-4 (VP2131-VP2144), VPaI-5 (VP2900-VP2910), VPaI-6 (VPA1253-VPA1270), and VPaI-7 (VPA1312-VPA1398). The sequences of the above virulence genes and genomic islands in RIMD 2210633 were used as references, except for the *trh* gene and T3SS2β genes, which were absent in the genome of RIMD 2210633, and hence the sequences of the genes in strain TH3996 was used as references (Okada et al., 2009). We mapped the raw sequencing reads of each isolate to the sequences of these genes and genomic islands using SOAPaligner (R. Li et al., 2009). For each gene/genomic island, if the overall sequence mapping coverage is larger than 70% and its depth is larger or equal to ten, it was considered as present, and otherwise, it was considered absent.

The assemblies of the seven CGs were further used in Prokka (Seemann, 2014) for gene annotation, and the annotation results (GFF3 files) were used in Roary (Page et al., 2015) to identify the pan genome and generate the matrix of the gene presence/absence of each CG. We then verified the gene presence/absence by mapping the raw sequencing reads of each isolate to each accessory gene using SOAPaligner (R. Li et al., 2009) based on the same threshold (coverage $>$ 70% and depth $\geq$ 10). The reads mapping result would be treated as the gold standard and used in further analysis if the reads mapping result was different from the Roary result. In total of 202 accessory genes were identified in the seven CGs, with four ~ 127 accessory genes for each CG. We used eggNOG-Mapper (Huerta-Cepas et al., 2017) to annotate the COG classifications of these accessory genes, 81 of 202 accessory genes were classified into 13 COG classifications.

3. Results

3.1. Genetic diversity of V. parahaemolyticus food-associated isolates

We calculated the pairwise SNP distance between the 383 food-associated isolates and between isolates of each province, type, source, and spread tier (Fig. 1). The median pairwise SNP distance between all isolates was 33,264, indicating a high genetic diversity in *V. parahaemolyticus* food-associated isolates. Specifically, we found that the genetic diversity of isolates from different provinces, sample types, sources, and farm-to-table spread tiers was similar to each other. Focusing on isolates from each province, their median pairwise SNP distances ranged from 33,013 to 33,659, with a difference of 1.9% in maximum, respectively. Surprisingly, the median pairwise SNP distances of isolates from coastal and inland provinces were 33,173 and 33,214, with a difference of 0.1%, respectively. This result indicated a similar diversity for the isolates from coastal and inland provinces. Besides, the median SNP distances of the isolates from seawater and freshwater products were also similar, with values of 33,222 and 33,298 (0.2% difference), respectively. Similar median pairwise SNP distances were also found in isolates from different sources and spread tiers. For each source, the median SNP distance ranged from 33,172 to 33,311 (0.4% difference). For each farm-to-table spread tier, the median SNP distance ranged from 33,236 to 33,420 (0.6% difference).

To explore the relationships of isolates within and between different

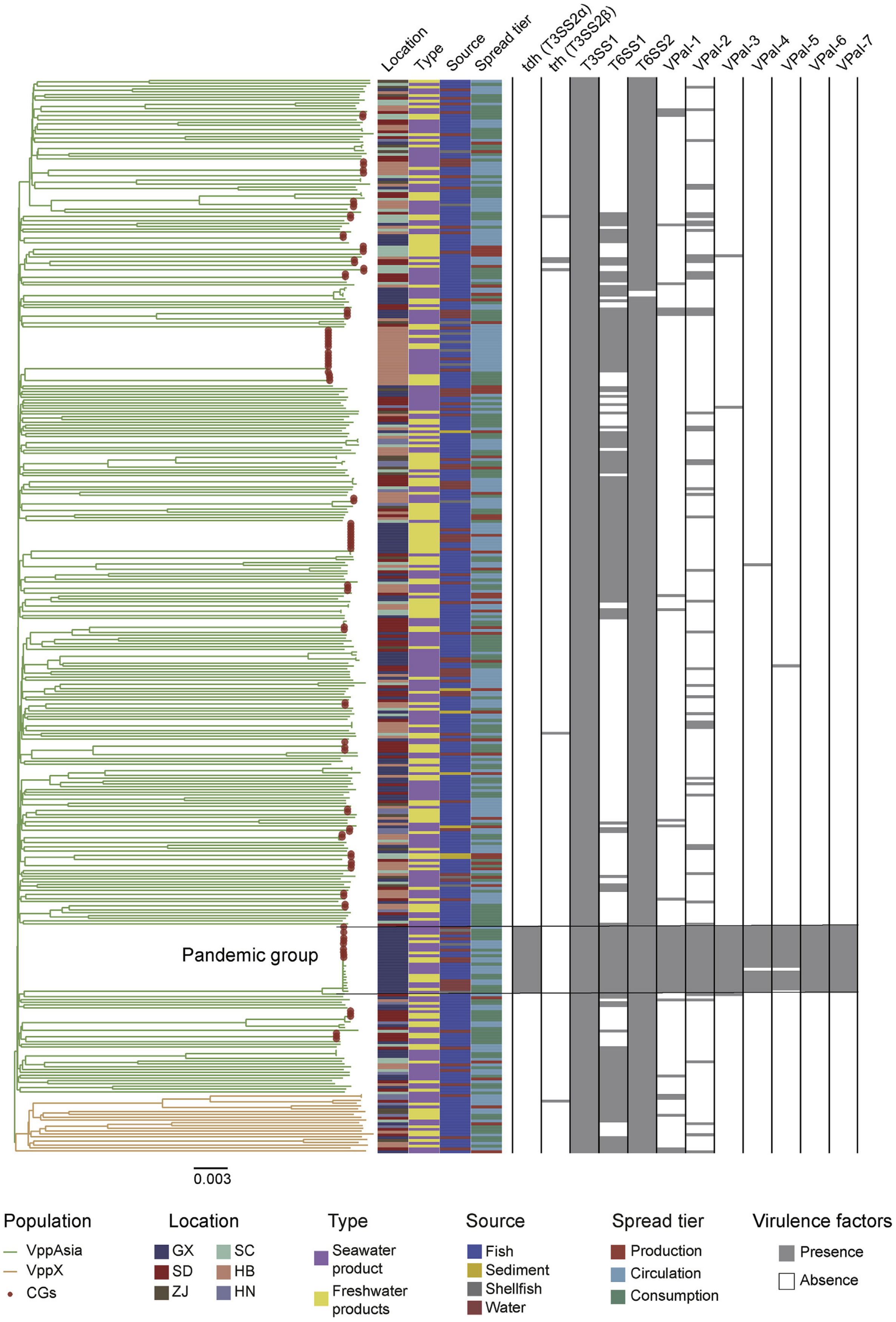

Fig. 2. Neighbor-joining tree and the distributions of virulence gene/genomic island in the 383 *V. parahaemolyticus* food-associated isolates. Branch colors of the NJ tree indicate the populations of *V. parahaemolyticus*, red points in the tree tips indicate the 28 CGs (pairwise SNP distance $<$ 10). The colors of vertical bars from left to right indicate the isolation locations, types, sources, farm-to-table spread tiers, and virulence gene/genomic island distributions. Grey color indicates the presence of gene genomic islands, white color indicates their absence.

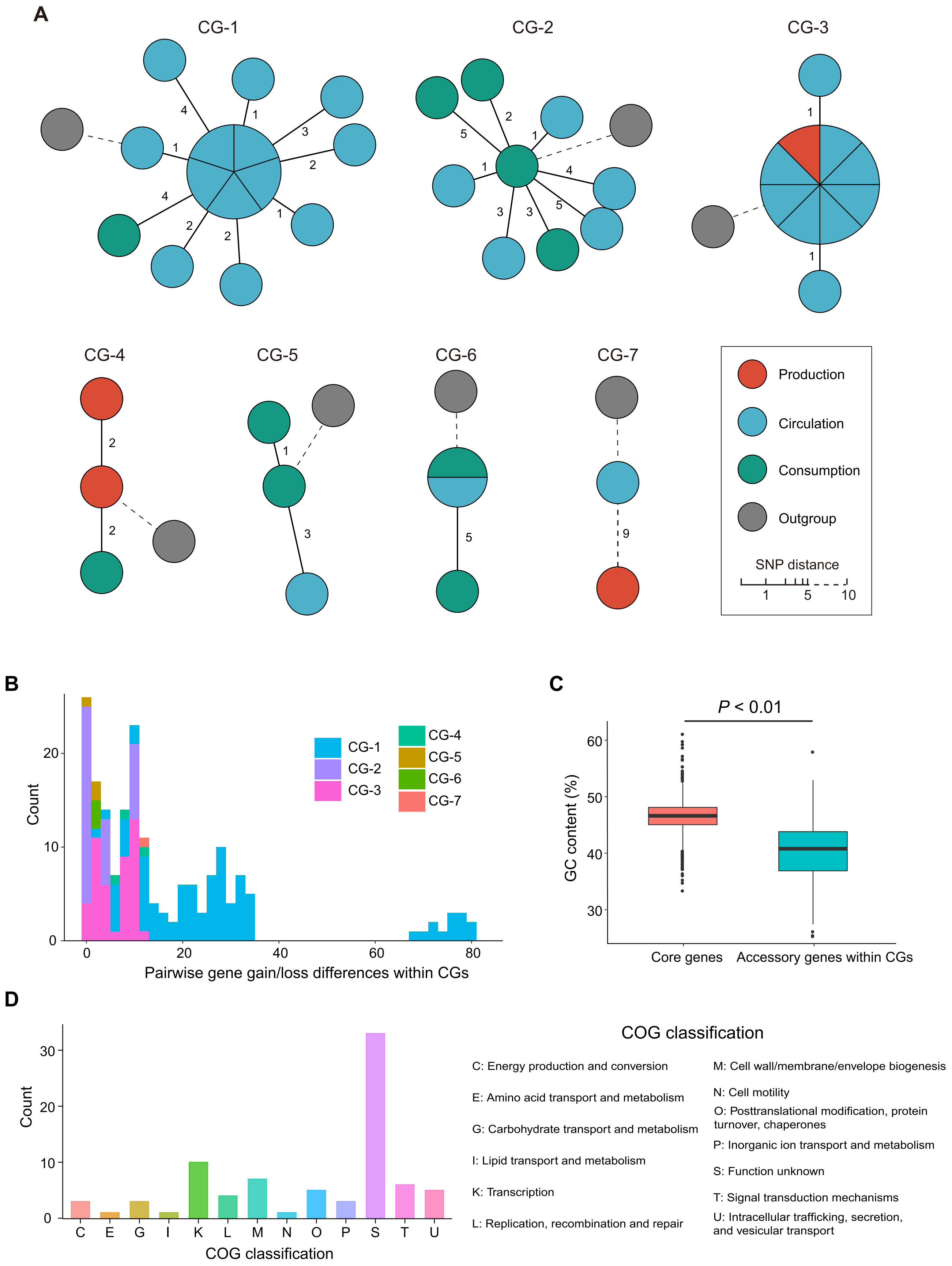

Fig. 3. Farm-to-table spread of *V. parahaemolyticus* food-associated isolates. (A) Minimum spanning trees (MSTrees) of the seven CGs based on non-recombined SNPs. Each node indicates a strain, and the colors indicate different farm-to-table spread tiers, with red for production, blue for circulation, green for consumption. Grey node indicates the outgroup strain. The number next to the branches indicates the SNP distance between two nodes. (B) Distribution of pairwise gene gain/loss differences between isolates of each CG. The colors indicate different CGs. (C) GC content of core and accessory genes of the seven CGs. (D) COG classifications of the accessory genes of the seven CGs.

classifications (locations, sample types, sources, and spread tiers), we constructed a neighbor-joining (NJ) tree based on 562,173 genome-wide SNPs (Fig. 2). Most of the isolates (363 isolates, 95%) belonged to the VppAsia population, with the others being assigned to VppX. A small number of closely related strain clusters were found in the NJ tree. Taking 2000 as the threshold of pairwise SNP distance between strains, we identified 47 "semi-clonal" groups (SCG). Each SCG contained 2–24 closely related strains, which were generally isolated from the same province, i.e., strains within the same SCG were isolated from the same province. Strains from categories of other classifications, including from those of sample types, sources and spread tiers, were largely scattered in the NJ tree. For each of the above classification, there was no clear association observed between categories and phylogenetic clusters.

3.2. Virulence genes and genomic islands

To measure the disease risk of these *V. parahaemolyticus* food-associated isolates, we detected the presence/absence of 13 previously described virulence genes and genomic islands in each isolate. The 13 virulence genes and genomic islands include two hemolysins (*tdh* and *trh*), two type III secretion systems (T3SS1 and T3SS2α/β), two type VI secretion systems (T6SS1 and T6SS2) and seven genomic islands (VPaI-1–7) (Boyd et al., 2008; Hurley et al., 2006; Makino et al., 2003; Y. Yu et al., 2012) (Fig. 2). The most well-known virulence genes *tdh* and *trh* were found in 6% and 2% isolates, separately. To be specific, all the *tdh* gene positive isolates formed a close cluster in the NJ tree, which belonged to the pandemic O3:K6 group (Fig. 2). Given that T3SS2α genes and *tdh* gene are adjacent, T3SS2α genes showed the same distribution as *tdh* gene. Similarly, the same distribution was found in T3SS2β genes and *trh* gene, as those two gene types are adjacent. T3SS1 genes were present in all the isolates, which was consistent with previous studies that they were core genes of *V. parahaemolyticus* (Ceccarelli et al., 2013; Wang et al., 2015), indicating that they could be necessary genes for the survival of *V. parahaemolyticus* in the environment. T6SS1 genes were present in nearly half the number of strains (48%) with a scatter distribution in the NJ tree, and T6SS2 genes were found in almost all strains (99%). The role of T6SSs in pathogenicity is still unproved (Ceccarelli et al., 2013; Wang et al., 2015), however, they were proposed to be involved in the environment fitness. T6SS1 was found to be active under warm marine-like conditions, while T6SS2 was active under low salt conditions (Salomon et al., 2013). Besides, T6SS1 has bacteriolytic activity against other bacteria, which is a competitive advantage in marine environment when competing for a niche (Salomon et al., 2013). The high frequency of T3SS1, T6SS1 and T6SS2 in food-associated isolates suggested that they may be related to the environmental fitness, rather than to the pathogenicity. The seven genomic islands were mainly found in pandemic group strains, and five of them (VPaI-1–5) were also found in a few non-pandemic isolates. While VPaI-1, VPaI-4, and VPaI-5 had been proposed to be unique to pandemic isolates (Hurley et al., 2006), our result showed a broader distribution of these GIs, which could be resulted from the HGT.

3.3. Farm-to-table spread pattern

High recombination rate of *V. parahaemolyticus* hampered the direct application of phylogeny-based method in source tracking and spread analysis in the whole species level, because recombination disrupted vertical genetic signals, which prevented us from inferring the real evolutionary history (Posada and Crandall, 2002; Schierup and Hein, 2000). Here, we only focused on recently emerged clonal groups, in which the recombination was limited. We identified the recombined regions for each of the 47 SCGs using ClonalFrameML (X. Didelot and Wilson, 2015), removed the SNPs located in recombined regions, and then re-calculated the pairwise SNP distance. Finally, we identified 28 clonal groups (CGs) in which pairwise SNP distance was less than 10 (Fig. 2). Strains within each CG were closely related to each other, which were most likely the derivatives of a common ancestor.

We found that all the strains of each CG were isolated from the same province, no *trans*-province spread was observed. Moreover, seven CGs contained strains from different tiers of farm-to-table spread process. We constructed minimum spanning trees (MSTrees) using non-recombined SNPs for these seven CGs to characterize the farm-to-table spread pattern of *V. parahaemolyticus* (Fig. 3A). The MSTrees revealed a radial spread pattern, in which one source generated most of the descendants while secondary spread link was rare. Besides, we found strains from each farm-to-table spread tier can be the source of other tiers. Specifically, the MSTrees revealed five types of spread routes: production to circulation (CG-3), production to consumption (CG-4), circulation to consumption (CG-1, CG-6), consumption to circulation (CG-2, CG-5), and circulation to production (CG-7).

To investigate the gene variations during the spread process, we constructed pan-genome for each CG using Roary (Page et al., 2015), and calculated the pairwise accessory gene presence/absence difference between isolates within each CG. Neither virulence genes nor genomic island gain/loss was observed. However, we found the gain/loss of a small number of other accessory genes, with the median number of ten as the gene presence/absence difference between pairwise strains of each CG (Fig. 3B). The GC content of these variant genes was significantly lower than that in the core genes (40.1% vs 46.5%, $P < 0.01$, Student's *t*-test) (Fig. 3C), indicating that the accessory gene gain/loss could be resulted from HGT. We further classified these variant genes into 13 COG classifications (Fig. 3D). There were 40% genes function unknown, 10% genes transcription associated, and less than 10% genes respectively located in other COG classifications.

4. Discussion

Recent advances in clinical clones associated with human diseases, e.g. O3:K6 serogroup and ST36 strains (Martinez-Urtaza et al., 2017, 2018; Nair et al., 2007; Su and Liu, 2007), greatly enhanced our understanding of the *V. parahaemolyticus* pathogenesis. However, most *V. parahaemolyticus* strains are not pathogenic, and clinical strains represent only a small fraction of the overall *V. parahaemolyticus* diversity. Exploring the environmental reservoir of *V. parahaemolyticus* isolates will provide a more comprehensive insight into this foodborne pathogen. Food-associated isolates are very likely to enter the human intestinal tract and may be at risk of causing disease, so they are important candidates for studying the pathogenesis of *V. parahaemolyticus*.

The high genetic diversity of *V. parahaemolyticus* food-associated isolates had been determined by different kinds of molecular typing methods, e.g. serotyping, PFGE, MLST and WGS (DePaola et al., 2003; González-Escalona et al., 2008; Hazen et al., 2015; C. H. Lüdeke et al., 2014; C. H. M. Lüdeke et al., 2015; Lopatek et al., 2018). Here using a cross-sectional genome dataset, we further gave insight into the high diversity in *V. parahaemolyticus* food-associated isolates. We showed that the genetic diversity of *V. parahaemolyticus* isolates of different provinces, types, sources and spread tiers was similar, no significant enrichment or bottleneck in a certain classification was observed.

The environment of different classifications varied greatly, such as the intestinal environment of different aquatic hosts. The similar genetic diversity among isolates of different classifications indicated that *V. parahaemolyticus* can freely spread among different niches, highlighting its strong adaptation ability. This could also explain its broad geographical dispersal and high abundance in oceans and seafood.

During the whole 383 strains, 41% of them were isolated from inland provinces, and 42% were from freshwater products. Besides, the genetic diversity of isolates from inland and coastal provinces, as well as of those from freshwater and seawater products was similar (< 0.2% SNP difference). Although it is possible that some isolates from inland provinces/freshwater could be sourced from contamination of seawater products, because the products of freshwater and seawater could be contacted with each other during transportation and storage. However, such possibility cannot fully explain the high prevalence, broad geographical distribution and high genetic diversity of isolates from inland provinces and freshwater. Therefore, we hypothesized that *V. parahaemolyticus* was very likely to have established stable environmental reservoirs in inland provinces and freshwater environment in China. This hypothesis was further supported by two evidences. Firstly, 17 isolates (isolated from sediment, water and fish) from inland provinces were isolated from the production tier of freshwater aquaculture farm sites. The median pairwise SNP distance of these isolates was 33,447, which was as high as the average genetic distance at the whole species level. Secondly, the nationwide food safety surveillance in China in 2016 found that the prevalence of *V. parahaemolyticus* in freshwater fish and shellfish was as high as 19% (613/3226), and at least 46% strains were isolated from inland provinces (central and western China) (Y. Li et al., 2019). The colonization of *V. parahaemolyticus* in inland provinces and freshwater is unique to our knowledge, because *V. parahaemolyticus* was usually considered as a type of bacteria in brackish waters. The mechanism of colonization in freshwater of *V. parahaemolyticus* is worth to be discovered in the future study.

Most environmental *V. parahaemolyticus* isolates were non-pathogenic (Baker-Austin et al., 2018; Nair et al., 2007; Su and Liu, 2007), and similar finding was acquired in this study. Specifically, we found that the positive rate of previously reported virulence genes and genomic islands were low in food-associated isolates, except for the T3SS1 and two T6SSs, indicating the *V. parahaemolyticus* food-associated isolates have a low risk of disease in general. However, we found that 6% food-associated isolates belonged to the pandemic group, which carried all the virulence genes and genomic islands and thus have a high risk of disease transmission. All of the above 6% isolates were from a coastal province named Guangxi, where the *V. parahaemolyticus* prevalence was high (22% isolation rate in southern China) (Y. Li et al., 2019). The high isolation rate of pandemic isolates in Guangxi province compared to other provinces indicated a high risk of disease transmission, which provides clues for further food safety surveillance.

Phylogeny reconstruction was the powerful and widely used method for inferring the evolutionary history of pathogens and for source tracking during infectious disease outbreak (Allard et al., 2016; Xavier Didelot et al., 2012; Franz et al., 2016; Ronholm et al., 2016). However, direct application of this method could be problematic because of the presence of homologous recombination, which lead to unreliable tree topology and branch length estimation bias (Posada and Crandall, 2002; Schierup and Hein, 2000). Notably, homologous recombination has been reported to be the driving force in the evolution of *V. parahaemolyticus* (Y. Cui et al., 2015; González-Escalona et al., 2008; Martinez-Urtaza et al., 2017; Vos and Didelot, 2009; Yan et al., 2011), the ratio of substitutions caused by recombination relative to mutation can be as high as 40 (Vos and Didelot, 2009). By focusing on the recently emerged clonal groups in which we can recognize and exclude the influence of recombination on phylogeny reconstruction, we performed the preliminary exploration of the farm-to-table spread pattern of *V. parahaemolyticus* food-associated isolates. Seven possible farm-to-table spread events were observed, and four of them supported the spread flow was from farm to table, proving the feasibility of our method. However, three of them indicated the consumption to circulation (CG-2, CG-5) spread, and circulation to production (CG-7) spread, which were inconsistent with the initial hypothesis of the spread from farm to table. This might be due to inadequate sampling that the true source of the isolates was unsampled. However, the reverse spread can exist in reality. For instance, the unsold aquatic products on the market could be put back to aquaculture farms, which lead to the observed spread flow that from tier of circulation to production. With the development of the molecular tracing network and the accumulation of genome data, the comprehensive spread net will be exhibited in the future.

With the development of WGS, the sequencing price keeps decreasing and the speed and accuracy keep improving. As the most powerful method in foodborne pathogen surveillance and outbreak source tracking, WGS has been used routinely in some institutes and authorities, such as US Food and Drug Administration (FDA), Public Health England and Statens Serum Institut of Denmark (Allard et al., 2016; Franz et al., 2016; Ronholm et al., 2016). After the promulgation of Food Safety Law of People's Republic of China in 2009, the food safety risk surveillance and analysis framework of China has been developing rapidly (W. Li et al., 2018; Wu and Chen, 2018; Wu et al., 2018). Constructing WGS-based foodborne pathogen surveillance and disease outbreak reporting system is one of the most important tasks in the future. Our work provided the initial landscape of the WGS-based genetic diversity, disease risk assessment and spread patterns of *V. parahaemolyticus* food-associated isolates. The genome dataset and analysis method will be useful in further work.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2017YFC1601503), the National Key Program for Infectious Diseases of China (No. 2018ZX10101003 and 2018ZX10714-002) and Sanming Project of Medicine in Shenzhen (No. SZSM201811071).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fm.2019.103270.

References

Allard, M.W., Strain, E., Melka, D., Bunning, K., Musser, S.M., Brown, E.W., Timme, R., 2016. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. J. Clin. Microbiol. 54 (8), 1975–1983.

Ansaruzzaman, M., Lucas, M., Deen, J.L., Bhuiyan, N., Wang, X.-Y., Safa, A., Sack, D.A., 2005. Pandemic serovars (O3: K6 and O4: K68) of Vibrio parahaemolyticus associated with diarrhea in Mozambique: spread of the pandemic into the African continent. J. Clin. Microbiol. 43 (6), 2559–2562.

Bag, P.K., Nandi, S., Bhadra, R.K., Ramamurthy, T., Bhattacharya, S., Nishibuchi, M., Nair, G.B., 1999. Clonal diversity among recently emerged strains ofVibrio parahaemolyticus O3: K6 associated with pandemic spread. J. Clin. Microbiol. 37 (7), 2354–2357.

Baker-Austin, C., Oliver, J.D., Alam, M., Ali, A., Waldor, M.K., Qadri, F., Martinez-Urtaza, J., 2018. Vibrio spp. infections. Nat. Rev. Dis. Primers 4 (1), 8. https://doi.org/10.1038/s41572-018-0005-8.

Boyd, E.F., Cohen, A.L., Naughton, L.M., Ussery, D.W., Binnewies, T.T., Stine, O.C., Parent, M.A., 2008. Molecular analysis of the emergence of pandemic Vibrio parahaemolyticus. BMC Microbiol. 8, 110. https://doi.org/10.1186/1471-2180-8-110.

Ceccarelli, D., Hasan, N.A., Huq, A., Colwell, R.R., 2013. Distribution and dynamics of epidemic and pandemic Vibrio parahaemolyticus virulence factors. Front. Cell Infect. Microbiol. 3, 97.

Crim, S.M., Griffin, P.M., Tauxe, R., Marder, E.P., Gilliss, D., Cronquist, A.B., Prevention, 2015. Preliminary incidence and trends of infection with pathogens transmitted commonly through food - foodborne Diseases Active Surveillance Network, 10 U.S. sites, 2006-2014. MMWR Morb. Mortal. Wkly. Rep. 64 (18), 495–499.

Cui, Y., Yang, C., Qiu, H., Wang, H., Yang, R., Falush, D., 2018. The landscape of coadaptation in Vibrio parahaemolyticus. bioRxiv. https://doi.org/10.1101/373936.

Cui, Y., Yang, X., Didelot, X., Guo, C., Li, D., Yan, Y., Yang, R., 2015. Epidemic clones, oceanic gene pools, and eco-LD in the free living marine pathogen Vibrio parahaemolyticus. Mol. Biol. Evol. 32 (6), 1396–1410. https://doi.org/10.1093/molbev/msv009.

Delcher, A.L., Salzberg, S.L., Phillippy, A.M., 2003. Using MUMmer to Identify Similar Regions in Large Sequence Sets. John Wiley & Sons, Inc.

DePaola, A., Ulaszek, J., Kaysner, C.A., Tenge, B.J., Nordstrom, J.L., Wells, J., Gendel, S.M., 2003. Molecular, serological, and virulence characteristics of < em > Vibrio parahaemolyticus < /em > isolated from environmental, food, and clinical sources in North America and asia. Appl. Environ. Microbiol. 69 (7), 3999–4005. https://doi.org/10.1128/aem.69.7.3999-4005.2003.

Didelot, X., Bowden, R., Wilson, D.J., Peto, T.E., Crook, D.W., 2012. Transforming clinical microbiology with bacterial genome sequencing. Nat. Rev. Genet. 13 (9), 601.

Didelot, X., Wilson, D.J., 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput. Biol. 11 (2), e1004041. https://doi.org/10.1371/journal.pcbi.1004041.

Franz, E., Gras, L.M., Dallman, T., 2016. Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. Curr. Opin. Food Sci. 8, 74–79.

González-Escalona, N., Martinez-Urtaza, J., Romero, J., Espejo, R.T., Jaykus, L.-A., DePaola, A., 2008. Determination of molecular phylogenetics of Vibrio parahaemolyticus strains by multilocus sequence typing. J. Bacteriol. 190 (8), 2831–2840.

Ham, H., Orth, K., 2012. The role of type III secretion system 2 in Vibrio parahaemolyticus pathogenicity. J. Microbiol. 50 (5), 719–725.

Han, C., Tang, H., Ren, C., Zhu, X., Han, D., 2016. Sero-prevalence and genetic diversity of pandemic V. Parahaemolyticus strains occurring at a global scale. Front. Microbiol. 7, 567. https://doi.org/10.3389/fmicb.2016.00567.

Han, D., Yu, F., Tang, H., Ren, C., Wu, C., Zhang, P., Han, C., 2017. Spreading of pandemic Vibrio parahaemolyticus O3:K6 and its serovariants: a Re-analysis of strains isolated from multiple studies. Front. Cell Infect. Microbiol. 7, 188. https://doi.org/10.3389/fcimb.2017.00188.

Hazen, T.H., Lafon, P.C., Garrett, N.M., Lowe, T.M., Silberger, D.J., Rowe, L.A., ... Sobecky, P.A., 2015. Insights into the environmental reservoir of pathogenic Vibrio parahaemolyticus using comparative genomics. Front. Microbiol. 6, 204. https://doi.org/10.3389/fmicb.2015.00204.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P., 2017. Fast genome-wide functional annotation through orthology assignment by eggnog-mapper. Mol. Biol. Evol. 34 (8), 2115–2122. https://doi.org/10.1093/molbev/msx148.

Hurley, C.C., Quirke, A., Reen, F.J., Boyd, E.F., 2006. Four genomic islands that mark post-1995 pandemic Vibrio parahaemolyticus isolates. BMC Genomics 7, 104. https://doi.org/10.1186/1471-2164-7-104.

Lüdeke, C.H., Fischer, M., LaFon, P., Cooper, K., Jones, J.L., 2014. Suitability of the molecular subtyping methods intergenic spacer region, direct genome restriction analysis, and pulsed-field gel electrophoresis for clinical and environmental Vibrio parahaemolyticus isolates. Foodb. Pathog. Dis. 11 (7), 520–528.

Lüdeke, C.H.M., Gonzalez-Escalona, N., Fischer, M., Jones, J.L., 2015. Examination of clinical and environmental Vibrio parahaemolyticus isolates by multi-locus sequence typing (MLST) and multiple-locus variable-number tandem-repeat analysis (MLVA). Front. Microbiol. 6 (564). https://doi.org/10.3389/fmicb.2015.00564.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25 (15), 1966–1967. https://doi.org/10.1093/bioinformatics/btp336.

Li, W., Wu, S., Fu, P., Liu, J., Han, H., Bai, L., Guo, Y., 2018. National molecular tracing network for foodborne disease surveillance in China. Food Control 88, 28–32.

Li, Y., Pei, X., Yan, J., Liu, D., Zhang, H., Yu, B., Yang, D., 2019. Prevalence of foodborne pathogens isolated from retail freshwater fish and shellfish in China. Food Control 99, 131–136.

Liu, J., Bai, L., Li, W., Han, H., Fu, P., Ma, X., Guo, Y., 2018. Trends of foodborne diseases in China: lessons from laboratory-based surveillance since 2011. Front. Med. 12 (1), 48–57. https://doi.org/10.1007/s11684-017-0608-6.

Lopatek, M., Wieczorek, K., Osek, J., 2018. Characterization and genetic diversity of Vibrio parahaemolyticus isolated from seafoods. Appl. Environ. Microbiol. AEM 84 (16) e00537-18.

Makino, K., Oshima, K., Kurokawa, K., Yokoyama, K., Uda, T., Tagomori, K., Iida, T., 2003. Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae. Lancet 361 (9359), 743–749. https://doi.org/10.1016/S0140-6736(03)12659-1.

Martinez-Urtaza, J., Lozano-Leon, A., DePaola, A., Ishibashi, M., Shimada, K., Nishibuchi, M., Liebana, E., 2004. Characterization of pathogenic Vibrio parahaemolyticus isolates from clinical sources in Spain and comparison with Asian and North American pandemic isolates. J. Clin. Microbiol. 42 (10), 4672–4678.

Martinez-Urtaza, J., Simental, L., Velasco, D., DePaola, A., Ishibashi, M., Nakaguchi, Y., Pousa, A., 2005. Pandemic vibrio parahaemolyticus O3: K6, Europe. Emerg. Infect. Dis. 11 (8), 1319.

Martinez-Urtaza, J., Trinanes, J., Abanto, M., Lozano-Leon, A., Llovo-Taboada, J., Garcia-Campello, M., Gonzalez-Escalona, N., 2018. Epidemic dynamics of Vibrio parahaemolyticus illness in a hotspot of disease emergence, galicia, Spain. Emerg. Infect. Dis. 24 (5), 852–859. https://doi.org/10.3201/eid2405.171700.

Martinez-Urtaza, J., van Aerle, R., Abanto, M., Haendiges, J., Myers, R.A., Trinanes, J., Gonzalez-Escalona, N., 2017. Genomic variation and evolution of Vibrio parahaemolyticus ST36 over the course of a transcontinental epidemic expansion. mBio 8 (6). https://doi.org/10.1128/mBio.01425-17.

Nair, G.B., Ramamurthy, T., Bhattacharya, S.K., Dutta, B., Takeda, Y., Sack, D.A., 2007. Global dissemination of Vibrio parahaemolyticus serotype O3:K6 and its serovariants. Clin. Microbiol. Rev. 20 (1), 39–48. https://doi.org/10.1128/CMR.00025-06.

Okada, N., Iida, T., Park, K.S., Goto, N., Yasunaga, T., Hiyoshi, H., Honda, T., 2009. Identification and characterization of a novel type III secretion system in trh-positive Vibrio parahaemolyticus strain TH3996 reveal genetic lineage and diversity of pathogenic machinery beyond the species level. Infect. Immun. 77 (2), 904–913.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31 (22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

Park, K.S., Ono, T., Rokuda, M., Jang, M.H., Okada, K., Iida, T., Honda, T., 2004a. Functional characterization of two type III secretion systems of Vibrio parahaemolyticus. Infect. Immun. 72 (11), 6659–6665.

Park, K.S., Ono, T., Rokuda, M., Jang, M.H., Iida, T., Honda, T., 2004b. Cytotoxicity and enterotoxicity of the thermostable direct hemolysin-deletion mutants of Vibrio parahaemolyticus. Microbiol. Immunol. 48 (4), 313–318.

Posada, D., Crandall, K.A., 2002. The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. 54 (3), 396–402. https://doi.org/10.1007/s00239-001-0034-9.

Ronholm, J., Nasheri, N., Petronella, N., Pagotto, F., 2016. Navigating microbiological food safety in the era of whole-genome sequencing. Clin. Microbiol. Rev. 29 (4), 837–857. https://doi.org/10.1128/CMR.00056-16.

Salomon, D., Gonzalez, H., Updegraff, B.L., Orth, K., 2013. Vibrio parahaemolyticus type VI secretion system 1 is activated in marine conditions to target bacteria, and is differentially regulated from system 2. PLoS One 8 (4), e61086.

Schierup, M.H., Hein, J., 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156 (2), 879–891.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30 (9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Su, Y.C., Liu, C., 2007. Vibrio parahaemolyticus: a concern of seafood safety. Food Microbiol. 24 (6), 549–558. https://doi.org/10.1016/j.fm.2007.01.005.

Vos, M., Didelot, X., 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3 (2), 199.

Wang, R., Zhong, Y., Gu, X., Yuan, J., Saeed, A.F., Wang, S., 2015. The pathogenesis, detection, and prevention of Vibrio parahaemolyticus. Front. Microbiol. 6, 144.

Wu, Y. n., Chen, J. s., 2018. Food safety monitoring and surveillance in China: past, present and future. Food Control 90, 429–439.

Wu, Y. n., Liu, P., Chen, J. s., 2018. Food safety risk assessment in China: past, present and future. Food Control 90, 212–221.

Yan, Y., Cui, Y., Han, H., Xiao, X., Wong, H.C., Tan, Y., Zhou, D., 2011. Extended MLST-based population genetics and phylogeny of Vibrio parahaemolyticus with high levels of recombination. Int. J. Food Microbiol. 145 (1), 106–112. https://doi.org/10.1016/j.ijfoodmicro.2010.11.038.

Yang, C., Cui, Y., Didelot, X., Yang, R., Falush, D., 2018. Why panmictic bacteria are rare. bioRxiv. https://doi.org/10.1101/385336.

Yang, C., Pei, X., Wu, Y., Yan, L., Yan, Y., Song, Y., Cui, Y., 2019. Recent mixing of Vibrio parahaemolyticus populations. ISME J. 1.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.Y., 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Method. Ecol. Evol. 8 (1), 28–36.

Yu, Y., Yang, H., Li, J., Zhang, P., Wu, B., Zhu, B., Fang, W., 2012. Putative type VI secretion systems of Vibrio parahaemolyticus contribute to adhesion to cultured cell monolayers. Arch. Microbiol. 194 (10), 827–835. https://doi.org/10.1007/s00203-012-0816-z.

Zhang, L., Krachler, A.M., Broberg, C.A., Li, Y., Mirzaei, H., Gilpin, C.J., Orth, K., 2012. Type III effector VopC mediates invasion for Vibrio species. Cell Rep. 1 (5), 453–460.

Zhou, Z., Alikhan, N.F., Sergeant, M.J., Luhmann, N., Vaz, C., Francisco, A.P., Achtman, M., 2018. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res. 28 (9), 1395–1404. https://doi.org/10.1101/gr.232397.117.