Large-scale genomic epidemiology reveals the cryptic outbreaks, hidden persistent reservoirs, and spatiotemporal dynamics of Vibrio parahaemolyticus

Chao Yang
Institut Pasteur Shanghai https://orcid.org/0000-0003-0626-0586
Yinghui Li
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Min Jiang
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Lei Wang
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Yixiang Jiang
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Lulu Hu
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Xiaolu Shi
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Lianhua He
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Rui Cai
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Shuang Wu
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Yaqun Qiu
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Linying Lu
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Le Zuo
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Qiongcheng Chen
Shenzhen Center for Disease Control and Prevention, Shenzhen, China
Yarong Wu
Beijing Institute of Microbiology and Epidemiology https://orcid.org/0000-0003-4900-315X
Jaime Martinez-Urtaza
Department of Genetics and Microbiology, Facultat deBiociències, Universitat Autònoma de Barcelona, Barcelona, Spain
Chengsong Wan
Department of Microbiology, School of Public Health, Southern Medical University, Guangzhou, China;
Ruifu Yang
State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology
https://orcid.org/0000-0003-3219-7269
Yujun Cui

State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing
Qinghua Hu (huqinghua03@163.com)
Shenzhen Center for Disease Control and Prevention

Article

Keywords:

Posted Date: January 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1255817/v1

Version of Record: A version of this preprint was published at Nature Microbiology on August 2nd, 2022. See the published version at https://doi.org/10.1038/s41564-022-01182-0.

Abstract

Controlling foodborne diseases requires robust outbreak detection and a comprehensive understanding of outbreak dynamics. Here, by integrating large-scale phylogenomic analysis of 3,642 isolates and epidemiological data, we performed “data-driven” outbreak detection and described the long-term outbreak dynamics of the leading seafood-associated bacterial pathogen, *Vibrio parahaemolyticus*, in a high-prevalence city, Shenzhen, China, over a 17-year period. Different from the widely accepted notion that sporadic patients and independent point-source outbreaks dominated foodborne infections, we found that 71% of isolates from patients grouped into within-1-month clusters that differed by ≤6 SNPs, indicating putative outbreaks; 56% of these clusters contained isolates exclusively from previously defined “sporadic” patients, representing unrecognized cryptic outbreaks. Furthermore, we showed that despite the long time spans between clusters, 70% of them were genomically closely related and were inferred to arise from a small number of common sources, which provides evidence that hidden persistent reservoirs generated most of the outbreaks, rather than independent point-sources. Phylogeographical analysis further revealed the geographical heterogeneity of outbreaks and identified a coastal district as the potential hotspot of outbreaks and as the hub and major source of cross-district spread events. Our findings provide a comprehensive picture of the long-term spatiotemporal dynamics of foodborne outbreaks for the first time and present a novel perspective on the major source of foodborne infections, which will inform the design of future foodborne disease control strategies.

Introduction

Foodborne diseases (FBDs) are major and enduring threats to public health with an estimated 600 million cases each year, nearly one in ten people worldwide[1]. Because of the ubiquitous nature of foodborne pathogens in the natural environment and their ability to contaminate the farm-to-fork continuum, FBDs were usually characterized by occasional outbreaks caused by spillover of point-source contamination on the background of steady sporadic patients. Although outbreaks catch more attention, sporadic patients are generally considered to be the main burden of FBDs[2]. Effective control requires robust detection of outbreaks from baseline sporadic patients and a comprehensive understanding on the key properties of outbreak dynamics such as the spatiotemporal distribution and common sources [2,3].

Traditional FBD outbreak surveillance was passive and triggered mainly by notification of clustered cases, which made it difficult to detect and link small-scale, protracted, or cross-regional outbreaks. The establishment of molecular subtyping-based networks substantially promoted FBD surveillance and enabled outbreaks to be detected in an active “data-driven” way by identifying genetically related clusters of infections. For example, PulseNet, which is a molecular subtyping network for FBD surveillance based on pulsed-field gel electrophoresis (PFGE), has enabled nationwide and international outbreaks to be recognized through the investigation of clusters with indistinguishable PFGE patterns[4]. However, the resolutions of PFGE and other conventional typing methods were rapidly proved inadequate for outbreak confirmation, especially for highly clonal pathogens[5]. Whole-genome sequencing (WGS) provides ultimate resolution to discriminate outbreak from sporadic isolates and is becoming the reference method in FBD outbreak investigations. Moreover, WGS enables robust phylogenetic reconstruction to reveal the evolutionary relationships between isolates and identify isolates arising from common sources[2,3]. In practice, phylogenomic clusters have been demonstrated to have high concordance with epidemiological links in outbreak investigations[6,7], suggesting that these clusters can be good indicators to quantify and characterize the dynamics of FBD outbreaks, as was found for other infectious pathogens that are transmitted person-to-person[8–10]. Recent studies using phylogenomic clustering analysis to investigate FBD outbreaks showed its ability to detect cryptic outbreaks that were neglected by traditional surveillance[11,12] and to track and resolve international outbreaks[13,14]. However, to our knowledge, there is currently no study to characterize the long-term dynamics of FBD outbreaks based on large-scale WGS and epidemiological data from a population.

Vibrio parahaemolyticus (VP) is the leading cause of seafood-associated bacterial gastroenteritis worldwide[15]. Most pathogenic VP carry one or two well-known virulence genes, *tdh* and/or *trh*, that encode thermostable direct hemolysin (*tdh*) and *tdh*-related hemolysin (*trh*), respectively. Notably, since the emergence and pandemic spread of the serotype O3:K6 clonal group, VP prevalence has increased globally over the last two decades, and its geographical range has extended to high-latitude areas where it was previously unreported[15]. Given the increasing disease burden, current outbreak surveillance of VP is based mainly on passive notification, whereas WGS-based epidemiology studies have focused on characterizing a small number of outbreaks[16–20] and local epidemic dynamics in low-prevalence regions[21,22].

China has a high prevalence of VP infections, especially in southern coastal regions such as Shenzhen[23], a populous city with approximately 13 million people. Since 2002, the Shenzhen Centers for Disease Control and Prevention (CDC) began to establish an integrated laboratory-based surveillance network of FBDs, with VP as one of the target pathogens. This network consisted of three systems: Foodborne Disease Outbreak Surveillance (FDOS), Infectious Diarrheal Diseases Surveillance (IDDS), and Food Safety Surveillance (FSS), captured a dense collection of isolates from documented outbreaks (FDOS), sporadic patients (IDDS), and the environment (FSS). In this study, we sequenced all the archived VP isolates collected by the network between 2002 and 2018. By integrating large-scale phylogenomic analysis and epidemiological data, we described the genomic baseline of pathogenic VP lineages, quantified the magnitude of outbreaks in a “data-driven” way, and characterized the spatiotemporal dynamics of VP outbreaks over a 17-year period with unprecedented resolution.

Results

Dataset

A total of 3,713 archived VP isolates were successfully sequenced. After strict quality assessment (Extended Data Fig. 1), 3,642 high-quality genomes from 2,686 (74%) patients and 956 (26%) environmental isolates were obtained (Fig. 1A, Table 1, Supplementary Table 1). Among the patient isolates, 830 (31%) were from documented outbreaks captured by FDOS (FDOS-outbreaks). The remaining 1,856 (69%) were from outpatients with diarrhea at 16 sentinel hospitals (IDDS patients), which were previously considered as sporadic because there were no recognized epidemiological links. Therefore, according to the traditional surveillance, sporadic patients from IDDS compose much of the disease burden. The VP-prevalence, defined as the proportion of VP-associated outpatients to all outpatients was 4.2% (1,856/44,160). The patients had a median age of 28 years (interquartile range [IQR]: 23–35), there were more males (58%), and the clinical symptoms included diarrhea (100%), abdominal pain (78%), vomiting (39%), and fever (10%) (Table 1). Environmental isolates captured by FSS were from food (mainly fish and shellfish) and the environments of aquatic markets.

Population structure and dynamics

To characterize the circulating VP lineages, we constructed a maximum-likelihood tree of 3,642 isolates based on 1,079,993 SNPs (Fig. 2A). The tree showed there were multiple clonal groups (CGs) of two or more closely related isolates. Based on the pairwise SNP distances distribution between all the isolates (Extended Data Fig. 2), we selected ≤2,500 SNPs as the cutoff and identified a total of 176 CGs that contained 3,123 (86%) isolates. Each CG had isolates from one or more serogroup (1–5) or sequence type (1–4) and was named after the main sequence type.

A total of 2,604 (97%) patient isolates were grouped into 23 pathogenic CGs (PCGs) each with 2 to 1,946 patients (Fig. 1B,2B), whereas only 20 (2%) of the 903 environmental isolates from FSS were assigned into PCGs. Two major PCGs, pandemic PCG3 and PCG189, drove local epidemics that caused 72% (n=1,946) and 12% (n=324) of the cases, respectively. The remaining 21 PCGs contained fewer isolates, together caused 12% (n=334) of the cases and were named as PCG-others. Two well-known virulence genes, *tdh* and *trh,* were identified in 99% and 3% of the PCG isolates, respectively (Fig. 2A).

The number of PCGs fluctuated annually with a peak in 2009, then decreased and remained stable (Fig. 2C). This trend was similar to that of the number of enrolled patients (Extended Data Fig. 3). At the end of sampling, seven PCGs persisted (Fig. 2C, Extended Data Fig. 4). Pandemic PCG3 was always dominant across the sampling period, accounting for 61%–88% of the cases annually, followed by PCG189. PCG diversity varied among different districts and correlated with the VP-prevalence in outpatients (Fig. 2D). The proportion of PCGs in different districts was generally similar, and six PCGs were spread across most (≥8/10) of the districts (Fig. 2E, Extended Data Fig. 5). Additionally, there was no significant difference (p >0.05, Fisher's exact test) in the demographic and clinical characteristics between different PCGs or between PCGs and singletons (Table 1).

Clustering analysis and "data-driven" outbreak detection

The isolates from outbreaks are expected to be genetically closely related. To obtain the reference cutoff of genetic similarity between outbreak isolates, we analyzed the pairwise SNP distances between 615 isolates from 126 FDOS-outbreaks with known epidemiological links (Fig. 3A,B). As expected, most FDOS-outbreak isolates had few SNP differences with a median of 1 (IQR: 0–6). In addition, 575 (93%) isolates of 119 (94%) FDOS-outbreaks had ≤6 SNPs before and after removing SNPs caused by recombination (see methods), and few isolates were clustered with a higher SNP cutoff (Fig. 3B). The SNP distance distribution between FDOS-outbreak isolates in different PCGs were similar to the overall distribution (Extended Data Fig. 6), whereas the entire dataset had a discontinuous SNP distance distribution, with <1% pairs of isolates having ≤6 SNPs. When SNP numbers were >6, the number of paired isolates increased substantially, indicating isolates with no epidemiological links had been introduced (Fig. 3B). Therefore, we proposed ≤6 SNPs as the genetic similarity cutoff between outbreak isolates. Furthermore, we analyzed external independent datasets of 34 isolates from four outbreaks in different countries[16–20] and identified a maximum of six SNPs (Extended Data Fig. 7), showing the generalizability of this cutoff.

Under the cutoff of ≤6 SNPs, a total of 2,237 (84%) patient isolates were grouped into 221 single-linkage phylogenomic clusters (P-clusters). P-clusters represent isolates that arose from common sources, and each with 2–158 isolates from patients (Fig. 1B,3A). The durations of 215 P-clusters with detailed time records varied between 0 and 70 months (Fig. 3C); among them, 75 (35%) P-clusters with 274 patients (10% of total patients) had within-1-month durations indicating transient outbreaks. The other 140 (67%) P-clusters with 1,951 patients (73% of total patients) lasted over 1 month (Fig. 3C), indicating that persistent common sources were present and caused continuous infections. To compare these long-term P-clusters with FDOS-outbreaks (transient point-source outbreaks), we further split them into within-1-month sub-P-clusters and identified a total of 370 within-1-month P-cluster/sub-P-clusters and defined them as outbreak clusters (Ob-clusters). Taken together, Ob-clusters were genetically (≤6 SNPs), temporally (within 1-month), and spatially (citywide) closely related patient isolates, indicating putative outbreaks (Fig. 1B). The sensitivity analysis showed that different SNP cutoffs (3, 6 or 10 SNPs) and time intervals (1-week or 1-month) had limited effects on the Ob-cluster definition, accounting for a maximum of 5% reduction in number of Ob-clusters, and stricter cutoffs generally led to more Ob-clusters with smaller sizes (Extended Data Fig. 8).

There were 1,897 patients (71% of total patients) grouped into 370 Ob-clusters. Most Ob-clusters were small, with a median size of three patient isolates (IQR: 2–6; range: 2–53) (Fig. 3A). The Ob-clusters and associated patients were enriched in PCG3; 1,524 (80%) of 1,897 Ob-cluster patients and 35 (88%) of 40 large Ob-clusters with >10 patients were from PCG3, with these two proportions being significantly larger (p <0.05, Fisher's exact test) than the proportion of PCG3-associated patients (72%).

There were no significant differences (p >0.05, Fisher's exact test) in demographic and clinical symptom characteristics between Ob-cluster and non-Ob-cluster patients, except that there were significantly more males (p=0.048, Fisher's exact test) among the Ob-cluster patients (60%) than in the non-Ob-cluster patients (54%) (Table 1).

Ob-clusters *vs.* documented outbreaks

The numbers of Ob-clusters and associated patients were substantially larger than those of the FDOS-outbreaks (2.9-fold and 3.1-fold, respectively) (Fig. 3A), indicating that the number and scale of outbreaks were likely to be previously largely underestimated. Notably, a large proportion (1,259/1,856, 68%) of isolates from outpatients, which were previously considered as sporadic because no recognized epidemiological links, were assigned into Ob-clusters. Specifically, 207 (56%) of the 370 Ob-clusters contained exclusively 843 outpatient isolates, which were not captured by FDOS, that represent unrecognized cryptic outbreaks. These Ob-clusters were characterized by their small scale; 197 (95%) of them each involved ≤10 patients (Extended Data Fig. 9). However, 582 (69%) and 634 (75%) of the 843 outpatients were identified in the same sentinel hospital and district, respectively, as another patient in an Ob-cluster; i.e., they were geographically close, indicating possible epidemiological links among them. Moreover, 98 (26%) Ob-clusters contained patients from both FDOS-outbreaks (n=615) and outpatients (n=416), with the latter representing missed outbreak patients. 362 (87%) of these 416 outpatients were not identified in the district where an FDOS-outbreak was reported, which may explain the unrecognized links between them and FDOS-outbreak patients (FDOS investigations were usually limited to local districts). Taken together, if Ob-clusters are taken as the gold standard of outbreak definition, then passive notification-based FDOS had a sensitivity of 0.31 (95% confidence interval [CI]: 0.29–0.34) and specificity of 0.93 (95% CI: 0.90–0.95) for discriminating outbreaks from sporadic patients.

Genomic links between Ob-clusters over long-time spans

Despite the long-time spans (2–70 months) between Ob-clusters, 258 (70%) of 370 Ob-clusters were genomically linked, which can be assigned into 61 P-clusters. Each of these P-clusters contained 2–20 Ob-clusters, indicating that persistent common sources caused continuous outbreaks (Fig. 1B,3D). The minimum-spanning trees of some P-clusters (e.g., PC033 and PC202) were "star-like", indicating continuous outbreaks caused by a single lasting common source. The minimum-spanning trees of other P-clusters (e.g., PC052 and PC176) had “multiple stars” topologies, which may indicate related outbreaks caused by different but linked sources, or strains that colonized and evolved in a local niche and then led to multiple waves of outbreaks (Fig. 3D).

The identification of these long-term P-clusters indicated that the common sources had not been fully eliminated after the initial outbreak, but persisted in some hidden reservoirs. To quantify preventable outbreaks/patients if the hidden persistent reservoirs are eliminated, we defined continued Ob-clusters/patients (i.e., Ob-clusters/patients identified after the initial (earliest) Ob-cluster within a P-cluster) to represent preventable outbreaks/patients; the other Ob-clusters/patients were defined as initial ones (Fig. 1B,3D). A total of 197 (53%) Ob-clusters and 1,308 patients (49% of all patients) were assigned into continued Ob-clusters/patients.

Temporal dynamics of Ob-clusters

The number of Ob-clusters fluctuated annually, and initial and continued Ob-clusters became dominant alternatively, with continued Ob-clusters lagging behind the initial Ob-clusters (Fig. 4A). Since 2006, the proportion of Ob-cluster patients was stable, leading to 76%–91% patients each year, whereas the proportions of initial and continued Ob-clusters/patients were variable, with continued exceeding initial in 10 of the 13 years.

Monthly, 251 (68%) Ob-clusters and 1,727 (67%) patients were concentrated at the peak of seasonal outbreaks (July to September). The number of initial Ob-clusters showed a similar changing trend as the continued Ob-clusters, with the latter lagging behind the former. Notably, from the start of the outbreak peak in June, the number of continued exceeded initial Ob-clusters and accounted for 57% of the total patients during the 3-month peak period.

Geographical distribution and cross-district spread pattern of Ob-clusters

The geographical distribution of 370 Ob-clusters was uneven in ten districts (Fig. 4B,C). There were 311 (84%) Ob-clusters that accumulated in four northern districts with high VP-prevalence (BA, GM, LoH, LG) (Fig. 4C), and there was a significantly higher proportion (p<0.001, Fisher’s exact tests) of continued patients (975/1,808, 54%) in these four northern districts

compared to the other districts (327/886, 38%). A hotspot of Ob-clusters was identified in the western coastal district, BA, where 71% (262/370) Ob-clusters can be detected (Fig. 4B,C). Notably, this hotspot was not obvious based on the FDOS results. Only 6 (5%) of the 126 FDOS-outbreaks were reported in BA district, and 181 (69%) of the 262 Ob-clusters in this district were cryptic and were not captured by FDOS.

There were 236 (64%) cross-district Ob-clusters identified in multiple (2–6) districts, accounting for 71%–100% of total Ob-clusters in each district (Fig. 4B), suggesting extensive links among patients in different districts. In the hotspot district, 83% of the total cross-district Ob-clusters and 41%–85% of the cross-district Ob-clusters in other districts were detected, suggesting that the hotspot acted as the hub of cross-district spread. Cross-district Ob-clusters were most prevalent between the hotspot district and the three other northern districts with high VP-prevalence, where 70%–85% cross-district Ob-clusters were related to those of hotspot.

To further infer the geographical source of the 236 cross-district Ob-clusters, we firstly analyzed the locations of the earliest detected isolates from a cross-district Ob-cluster and found that 43% (101/236) cross-district Ob-clusters were detected earliest in the hotspot (Fig. 3D). Furthermore, we performed phylogeographical analysis and identified the likely source district of 162 cross-district Ob-clusters and found that 79 (49%) were in the hotspot (Fig. 3D). Subsampling analysis of cross-district Ob-clusters based on a balanced number of isolates (1–3) in each district gave similar results, with 77 (49%) of 158 inferred source districts being from the hotspot (Extended Data Fig. 10). These results provide evidence that the hotspot was also a major source of cross-district spread. Moreover, we found a correlation between geographical and pairwise SNP distance between isolates from Ob-clusters sourced from the hotspot and isolates with greater geographical distances to the hotspot had larger mean SNP distances (Extended Data Fig. 11), which may be associated with the extra diversity accumulated during cross-district spread.

Discussion

In one of the largest bacterial WGS projects conducted so far, for the first time, we used an integrated surveillance framework to characterize the long-term epidemic dynamics of the leading seafood-associated bacterial pathogen, VP, in Shenzhen, a city with a high-prevalence of VP infections over the 17-year period. We showed that pathogenic lineages were stable over a decade, with two major PCGs, PCG3 and PCG189, driving local epidemics, which may be related to their higher virulence[24] and/or adaptability. Our findings highlight the power of phylogenomic clustering analysis to quantify and describe the dynamics of outbreaks, especially small-scale, extended time-period, and cross-regional outbreaks. Through active “data-driven” analysis, we showed that 71% of isolates from patients can be grouped into Ob-clusters indicating putative outbreaks, which is different from the observation of traditional surveillance that sporadic patients compose much of the disease burden. More than half (56%) of these clusters contained isolates exclusively from previously defined “sporadic” patients, representing unrecognized cryptic outbreaks, which lead to the number difference from documented FDOS-outbreaks, and also suggest that the magnitude of outbreaks may have been substantially underestimated previously. Furthermore, we showed that despite the long time spans (2-70 months) between clusters, 70% of them were can be genomically linked and were inferred to arise from a small number of common sources, which provides evidence that continuous outbreaks caused by hidden persistent reservoirs dominated local infections, rather than independent point-source outbreaks. Finally, we identified a potential coastal outbreak hotspot district, which also acted as the hub and a major source of cross-district spread, providing a focus for future outbreak surveillance and control. Together, our findings provide a comprehensive picture of the long-term spatiotemporal dynamics of FBDs for the first time and challenge the widely accepted notion on the major source of foodborne infections, i.e., sporadic patients and independent point-source outbreaks dominated foodborne infections, which have important implications for the design of future disease control strategies.

Serotyping is still a widely used subtyping method for VP and other foodborne pathogens. However, we showed that serotypes were not highly concordant with WGS-based CGs, which is consistent with previous studies[25,26], indicating that

serotyping does not accurately describe the VP population structure. For example, more than 20 serotypes were identified in pandemic PCG3 isolates[25], and serotype O4:K12 was reported in multiple distinct lineages of VP[26]. Because WGS was costly previously, serotyping was used to confirm VP outbreaks in China, including in Shenzhen. Now that the cost has decreased, WGS is becoming the new reference method for outbreak confirmation, and genomic relatedness cutoffs to define outbreak clusters have been proposed for several foodborne pathogens[27]. However, there is currently no reference cutoff for VP, and the cutoffs for other foodborne pathogens cannot be directly applied because characteristics such as genetic diversity and mutation rate are different. The recently developed core-genome multi-locus sequence typing (cgMLST) scheme was used to investigate four VP outbreaks and a maximum of three allele differences were identified, but no genomic relatedness cutoff was proposed[20]. We analyzed >100 FDOS-outbreaks and proposed a reference cutoff of ≤6 SNPs and showed its generalizability in external datasets. Furthermore, we found that removing SNPs caused by homologous recombination had a limited impact on Ob-cluster definition, indicating that recombination analysis under the outbreak scenario is optional, even for highly recombining bacteria like VP[28,29]. With the reference cutoff, 94% of the FDOS-outbreaks can be captured; the large SNP distance between the remaining FDOS-outbreak isolates may result from strain diversity in contaminated foods[30].

Current outbreak surveillance in Shenzhen, as well as in China and many other counties, is mainly passive notification based, triggered by independent notification of clustered patients from individual hospitals. If outbreak-associated patients presented at different hospitals or even were treated by different clinical doctors, indicators of outbreak: clustered patients may be neglected. Therefore, FDOS can detect large-scale outbreaks with large numbers of patients but is not sensitive to small-scale outbreaks. The inherent limitation of FDOS makes it inevitable to underestimate the magnitude of outbreaks, especially small-scale ones. WGS-based outbreak surveillance can detect and quantify the magnitude of outbreaks by phylogenomic clustering analyses, which is active, and data driven. In this study, the numbers of Ob-clusters and associated patients we detected were approximated to be 3-fold those of the FDOS-outbreaks, indicating a substantial underestimation of outbreaks. Even using P-clusters (i.e., isolates from common sources) to quantify outbreaks, the numbers were still 1.8-fold higher than those of the FDOS-outbreaks, whereas the scale of P-cluster-associated patients was similar to that of the Ob-clusters.

Surprisingly, a high proportion (82%) of Ob-clusters contained exclusively (56%) or included (26%) outpatient isolates that were not captured by FDOS, representing cryptic outbreaks or missed outbreak patients. Most of the cryptic Ob-clusters (95%) were small scale (≤10 patients), and most of the missed outbreak patients (87%) appeared mainly in districts that were different from the district where the FDOS-outbreaks were detected, which may explain why they were unrecognized by traditional outbreak surveillance. These results highlight the advantages of centralized sequencing and analysis of data from different districts to trace outbreaks. For example, in prospective studies of other foodborne pathogens such as *Salmonella* and *Listeria monocytogenes*, centralized sequencing and analysis enabled the detection of multiple cryptic outbreaks and additional outbreak-associated patients, and follow-up investigation established the epidemiological links between patients[11,12]. However, despite the advantages of WGS-based surveillance, epidemiology and traceback evidence are still needed to confirm the links across the spread networks, and these will benefit from more precise and timely outbreak detection powered by WGS.

Transient point-source outbreaks have been considered as the common pattern of FBD outbreaks, although protracted outbreaks caused by a persistent source have also been reported[12,31]. We linked 70% of the Ob-clusters to 61 P-clusters with 2–70 months duration, suggesting that most VP outbreaks in Shenzhen were likely related to a small number of hidden persistent reservoirs. Because of its poor resolution, conventional subtyping-based surveillance cannot accurately link extended time-period outbreaks, and therefore they were treated as independent events. Notably, there were more continued Ob-clusters (n=197) and associated patients (n=1,308) than initial ones (173 Ob-clusters, 589 patients), and more than half (57%) of the Ob-clusters and patients were continued ones during the peak seasonal outbreak period. These continued Ob-clusters/patients would be preventable if the common sources were eliminated in time. These findings show the importance

of eliminating the source to resolve outbreaks, and again highlight the advantage of WGS-based surveillance in detecting extended time-period outbreaks.

Persistent reservoirs can be caused by repeated introduction from external sources and/or local colonized sources. VP infections are commonly associated with consumption of raw or undercooked contaminated seafood, and most of the seafood available in Shenzhen is imported from outside. Because the external seafood sources vary and can be affected by many determinants, the scenario of repeated external introduction may not be the general cause. Instead, pathogenic VP that have established local long-term stable environmental reservoirs in, for example, aquatic product containers and aquaculture water of seafood markets, are more likely to be the general cause of continuous outbreaks through cross-contaminated seafood. Further in-depth investigations to clarify the causes and geographical locations of the hidden persistent reservoirs will inform the design of future control and prevention strategies.

There were obvious geographical differences of VP infections in the ten districts of Shenzhen. The westernmost coastal district BA had much higher VP-prevalence (8.7%) and higher PCG diversity (21 PCGs) than the other districts (VP-prevalence: <3.6%, PCGs: ≤15). We also showed that 71% of the total Ob-clusters, and 41–85% of the cross-district Ob-clusters in other districts can be detected in BA, and nearly half of the Ob-cluster isolates were earliest detected or inferred to BA sources, indicating that BA was the potential hotspot of outbreaks and that it acted as the hub of cross-district spread and was a major source of outbreaks. The key role of BA may be related to its special geographical location on the western coast, where many ports and aquatic markets are located, and a large amount of seafood is transferred from BA and transported to other districts. Moreover, a recent metagenomics sequencing-based study showed that the abundance of harmful *Vibrio* on the western coast of Shenzhen was significantly higher than that on the eastern coast[32]. Together, these results indicated the presence of extensive links across different districts, highlighting the necessity of centralized surveillance rather than independent surveillance by ten district CDCs. Furthermore, the BA district, which was identified as the hotspot, hub, and major source of outbreaks, should be the focus of future outbreak surveillance and control.

The aim of establishing the FSS system was to prevent outbreaks and patients by actively detecting and eliminating contaminated food. However, we found that the food isolates from FSS were distinct from the patient isolates; only 2% of the FSS isolates were from PCGs and only one FSS isolate grouped into an Ob-cluster. The within-sample diversity of VP in the FSS was not considered and only one isolate was selected and sequenced for each sample in this study. Therefore, the failure to observe PCG isolates in FSS may be related to strain diversity in the sample and does not necessarily mean that none of the PCG isolates were present. Although traditional *Vibrio* selective enrichment media was usually used to facilitate the growth of VP, a recent study demonstrated that the relative abundance of pathogenic to non-pathogenic VP was substantially reduced (20-fold) after enrichment[30], thereby reducing the possibility of successfully isolating pathogenic strains. We performed the enrichment before isolating the strains, which may have led to the isolation of few pathogenic strains in FSS. This finding highlights the importance of improving culture methods.

Our study has several limitations. First and most importantly, the retrospective study design made it impossible to reinvestigate the epidemiological links among Ob-cluster patients. Although we combined genetic relatedness and spatiotemporal data to define Ob-clusters, further prospective studies are needed to better assess the concordance between Ob-clusters and epidemiological defined outbreaks. Second, only outpatients from 16 sentinel hospitals were included in this study, whereas outbreak-associated patients may go to other hospitals or not go to hospitals. Therefore, the number of Ob-clusters/P-clusters represented only the minimum estimation of the magnitude of outbreaks. Finally, only one strain from each patient/food sample was selected and sequenced, and strain diversity within a sample was not assessed. Further strain diversity studies may provide clues for the large SNP distances observed between some FDOS-outbreak isolates and the few PCG isolates from FSS.

The ultimate resolution that allows the discovery of tiny differences among isolates, and the sharp decrease in the cost of sequencing, have made WGS a viable and powerful method in routine FBD surveillance[2,3]. In addition, standardized and

digital WGS data make it possible to construct global surveillance networks, such as the SARS-CoV-2 genomic surveillance network that is playing important roles in outbreak, transmission detection, and source tracking[10,33]. However, comprehensively extracting the information from large-scale surveillance networks to facilitate the prevention and control of FBDs is still challenging. In this study, we integrated citywide “big data” of genomes and epidemiological information, to establish the model that mining previously neglected outbreaks, possible persistent common sources, the spread network and its hub and source. Our results will help to not only improve the outbreak surveillance and investigation of VP but will also provide a basis for the establishment of a “data-driven” strategy for the prevention and control of other FBDs.

Methods

Sampling framework and data collection

The Shenzhen CDC established a well-functioning FBD surveillance network consisting of three systems: Foodborne Disease Outbreak Surveillance (FDOS), Infectious Diarrheal Diseases Surveillance (IDDS), and Food Safety Surveillance (FSS). All hospitals in Shenzhen are required to report suspected outbreaks to district CDCs, followed by epidemiological investigation and laboratory confirmation. The FDOS was performed independently by ten district CDCs, then reported to the city CDC. Documented FDOS outbreaks met the following criteria: (a) two or more clustered cases of diarrhea; (b) history of dining together within 3 days of disease onset; (c) VP tested positive with the same serotype. Isolates from patients and food associated with the suspected outbreaks were collected by the FDOS, and only VP isolates with the same serotype were defined as FDOS-outbreak isolates. The IDDS routinely collected isolates from outpatients with diarrhea who presented at 16 sentinel hospitals. IDDS outpatients were previously considered as sporadic because there were no recognized epidemiological links. FSS strains were from food (mainly fish and shellfish) and the environments of aquatic markets. Only one strain from a patient/environmental sample was archived and sequenced. Because data collection is part of the infectious disease surveillance, individual informed consent was waived.

Whole genome sequencing

All the obtained VP isolates were stored in a −80°C freezer. For DNA extraction, the isolates were inoculated in thiosulfate-citrate-bile-salts-sucrose (TCBS) agar plates at 37°C. Genomic DNA was extracted using the Sodium-Tris-EDTA (STE) method. Pair-end libraries with a mean insert size of 350 bp were prepared using a NEBNext Ultra DNA Library Prep Kit (NEB). Whole genome sequencing was performed on Illumina HiSeq X-Ten platforms; the average read length was 150 bp, and an average of 1.5 Gb clean reads were generated for each strain. The genome quality assessment process was shown in Extended Data Fig. 1.

Serotyping and multi-locus sequence typing

Serotyping was performed using traditional serum slide agglutination tests and commercial antisera (Denka Seiken, Tokyo, Japan). Multi-locus sequence typing (MLST) was performed using mlst (https://github.com/tseemann/mlst) to scan the assembled genome sequences against the VP PubMLST database (https://pubmlst.org/organisms/vibrio-parahaemolyticus).

Single nucleotide polymorphisms (SNPs) calling and phylogeny

Core-genome (regions present in >99.5% strains) SNPs were identified using the Snippy pipeline v4.6.0 (https://github.com/tseemann/snippy), with RIMD 2210633 as the reference genome. Repetitive regions in the reference genome were identified using Tandem Repeats Finder (TRF) v4.07b[34] and self-aligning by BLASTN. Recombination regions of FDOS-outbreak isolates were identified using Gubbins v2.3.4[35]. Non-repetitive core-SNPs were used to construct the maximum-likelihood (ML) and minimum-spanning (MS) trees. The ML tree of 3,642 isolates was constructed using

FastTree v2.1.10,[36] and the ML trees of cross-district Ob-cluster isolates were constructed using IQ-TREE v2.0.3[37] with the auto-detected best-fitting substitution model. The MS trees of P-cluster isolates were constructed using GrapeTree v1.5.0.[38]

Phylogenomic clustering analysis

Two or more isolates were classified into CGs, P-clusters, and Ob-clusters based on pairwise SNP distances and/or time intervals (Fig. 1). First, CGs were defined to show the overall population structure as previously described[28,29]. We analyzed the pairwise SNP distances between all the isolates (Extended Data Fig. 2) and grouped the isolates into CGs using a cutoff of ≤2,500 SNPs. Second, P-clusters were defined to represent isolates that arose from common sources. Currently, there is no definition standard for such clusters. We analyzed pairwise SNP distances between isolates from FDOS-outbreaks and previously reported outbreaks[16–20], and selected ≤6 SNPs as the cutoff of P-clusters. After taking recombination into account, we detected 22 outlier FDOS-outbreak isolates that had >100 SNPs from any of the other isolates in the same FDOS-outbreak, which were unlikely to be derived from a recent common ancestor, we excluded them from the analysis. Finally, to compare with FDOS-outbreaks that taken time-intervals into account, we performed secondary clustering of P-cluster isolates based on time intervals and defined within-1-month P-cluster/sub-P-cluster isolates as outbreak clusters (Ob-clusters, Fig. 1). A 1-month window was selected because it is a common period for outbreak analysis[39] and a practical turnaround time for routine sequencing and analysis. Sensitivity analysis was performed to assess the effect of different thresholds (SNPs ≤3 and ≤10; 1-week window) on Ob-cluster definition and interpretation.

Virulence factors

ABRicate (https://github.com/tseemann/abricate) was used to detect the presence/absence of two major virulence factor genes, *tdh* and *trh*, which encode thermostable direct hemolysin and *tdh*-related hemolysin, respectively. The *tdh* or *trh* gene was defined as present if the sequence coverage was >70%.

Phylogeographical analysis

The geographical sources of cross-district Ob-clusters were inferred using SIMMAP (stochastic character mapping of discrete traits on phylogenies)[40,41]. Geographical districts were treated as discrete traits and mapped to the ML trees using the all-rates-different (ARD) model with 100 replicates. The district of the root node with posterior probability >0.7 was considered as the inferred source. Subsampling analysis was performed by selecting 1–3 isolates in each district next to the root node.

Data availability

The sequencing data have been deposited in the NCBI Sequence Read Archive under study accessions PRJNA745505. Background information of sequenced isolate is listed in Supplementary Table 1.

Declarations

Acknowledgments

We gratefully acknowledge Prof. Daniel Falush, Biao Kan, Bo Pang and Elise Tourrette for their valuable comments. This study was funded by the National Key Research and Development Program of China (No. 2018YFC1603902 and 2017YFC1601500), Sanming Project of Medicine in Shenzhen (No. SZSM201811071), China National Science and Technology Major Projects Foundation (No. 2017ZX10303406), National Natural Science Foundation of China (No. 32000008 and 31770001) and China Postdoctoral Science Foundation (No. 2020M672836).

Author contributions

Y.C., Q.H., R.Y. and C.Y. and designed, initiated, and coordinated the study. Y.L., L.L., L.Z., L.W., Y.J., Q.C., L.H., M.J., X.S., L.H., R.C., and Y.Q. contributed to data collection and management. C.Y., Y.L. and R.W. analyzed the data. All authors contributed to interpretation of the data. C.Y. wrote the first draft of the paper and Y.L., J.M.U., R.Y., Y.C. and Q.H. reviewed and revised the paper. All authors read and approved the final manuscript.

Competing interests

We declare no competing interests.

References

1. *WHO estimates of the global burden of foodborne diseases. Foodborne disease burden epidemiolgy reference group 2007-2015*. *World Health Organization* (2015) doi:10.1016/j.fm.2014.07.009.
2. Deng, X., den Bakker, H. C. & Hendriksen, R. S. Genomic Epidemiology: Whole-Genome-Sequencing–Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annual Review of Food Science and Technology* **7**, 353–374 (2016).
3. Ronholm, J., Nasheri, N., Petronella, N. & Pagotto, F. Navigating microbiological food safety in the era of whole-genome sequencing. *Clinical Microbiology Reviews* **29**, 837–857 (2016).
4. Swaminathan, B., Barrett, T. J., Hunter, S. B. & Tauxe, R. v. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging infectious diseases* **7**, 382–389 (2001).
5. Deng, X. *et al.* Genomic epidemiology of Salmonella enterica serotype Enteritidis based on population structure of prevalent lineages. *Emerging Infectious Diseases* **20**, 1481–1489 (2014).
6. Waldram, A., Dolan, G., Ashton, P. M., Jenkins, C. & Dallman, T. J. Epidemiological analysis of Salmonella clusters identified by whole genome sequencing, England and Wales 2014. *Food Microbiology* **71**, 39–45 (2018).
7. Hassoun-Kheir, N. *et al.* Concordance between epidemiological evaluation of probability of transmission and whole genome sequence relatedness among hospitalized patients acquiring Klebsiella pneumoniae carbapenemase-producing Klebsiella pneumoniae. *Clinical Microbiology and Infection* **27**, 468–e1 (2021).
8. Yang, C. *et al.* Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *The Lancet Infectious Diseases* **17**, 275–284 (2017).
9. Town, K. *et al.* Phylogenomic analysis of Neisseria gonorrhoeae transmission to assess sexual mixing and HIV transmission risk in England: a cross-sectional, observational, whole-genome sequencing study. *The Lancet Infectious Diseases* **20**, 478–486 (2020).
10. Lemieux, J. E. *et al.* Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **2507**, eabe3261 (2020).
11. Jackson, B. R. *et al.* Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. *Clinical Infectious Diseases* **63**, 380–386 (2016).
12. Moura, A. *et al.* Real-time whole-genome sequencing for surveillance of Listeria monocytogenes, France. *Emerging Infectious Diseases* **23**, 1462–1470 (2017).
13. Pijnacker, R. *et al.* An international outbreak of Salmonella enterica serotype Enteritidis linked to eggs from Poland: a microbiological and epidemiological study. *The Lancet Infectious Diseases* **19**, 778–786 (2019).
14. Pettengill, J. B. *et al.* A multinational listeriosis outbreak and the importance of sharing genomic data. *The Lancet Microbe* **1**, e233–e234 (2020).
15. Baker-Austin, C. *et al.* Vibrio spp. infections. *Nature reviews. Disease primers* **4**, 8 (2018).

16. Haendiges, J. *et al.* Pandemic Vibrio parahaemolyticus, Maryland, USA, 2012. *Emerging Infectious Diseases* **20**, 718–720 (2014).
17. Haendiges, J. *et al.* A nonautochthonous U.S. strain of Vibrio parahaemolyticus isolated from Chesapeake Bay oysters caused the outbreak in Maryland in 2010. *Applied and Environmental Microbiology* **82**, 3208–3216 (2016).
18. Haendiges, J. *et al.* Characterization of Vibrio parahaemolyticus clinical strains from Maryland (2012-2013) and comparisons to a locally and globally diverse V. parahaemolyticus strains by whole-genome sequence analysis. *Frontiers in Microbiology* **6**, 1–11 (2015).
19. Gonzalez-Escalona, N., Gavilan, R. G., Toro, M., Zamudio, M. L. & Martinez-Urtaza, J. Outbreak of Vibrio parahaemolyticus sequence type 120, Peru, 2009. *Emerging Infectious Diseases* **22**, 1235–1237 (2016).
20. Gonzalez-Escalona, N., Jolley, K. A., Reed, E. & Martinez-Urtaza, J. Defining a core genome multilocus sequence typing scheme for the global epidemiology of Vibrio parahaemolyticus. *Journal of Clinical Microbiology* **55**, 1682–1697 (2017).
21. Martinez-Urtaza, J. *et al.* Epidemic dynamics of vibrio parahaemolyticus illness in a hotspot of disease emergence, Galicia, Spain. *Emerging Infectious Diseases* **24**, 852–859 (2018).
22. Baker-Austin, C. *et al.* Genomic epidemiology of domestic and travel-associated Vibrio parahaemolyticus infections in the UK, 2008–2018. *Food Control* **115**, 107244 (2020).
23. Li, Y. *et al.* Vibrio parahaemolyticus, southern coastal region of China, 2007-2012. *Emerging Infectious Diseases* **20**, 685–688 (2014).
24. Chen, X. *et al.* A new emerging serotype of Vibrio parahaemolyticus in China is rapidly becoming the main epidemic strain. *Clinical Microbiology and Infection* (2019) doi:10.1016/j.cmi.2019.09.024.
25. Chen, Y. *et al.* Comparative genomic analysis of Vibrio parahaemolyticus: Serotype conversion and virulence. *BMC Genomics* **12**, 294 (2011).
26. Zhao, L. *et al.* Co-existence of multiple distinct lineages in Vibrio parahaemolyticus serotype O4:K12. *Microbial Genomics* (2019) doi:10.1099/mgen.0.000287.
27. Pightling, A. W. *et al.* Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. *Frontiers in Microbiology* **9**, 1–13 (2018).
28. Yang, C. *et al.* Recent mixing of Vibrio parahaemolyticus populations. *The ISME Journal* **13**, 2578–2588 (2019).
29. Yang, C. *et al.* Genetic diversity, virulence factors and farm-to-table spread pattern of Vibrio parahaemolyticus food-associated isolates. *Food Microbiology* **84**, 103270 (2019).
30. Li, Y. *et al.* Application of digital PCR and next generation sequencing in the etiology investigation of a foodborne disease outbreak caused by Vibrio parahaemolyticus. *Food Microbiology* **84**, (2019).
31. Marshall, K. E. H. *et al.* Protracted Outbreak of Salmonella Newport Infections Linked to Ground Beef: Possible Role of Dairy Cows — 21 States, 2016–2017. *MMWR. Morbidity and Mortality Weekly Report* **67**, 443–446 (2018).
32. Zhang, R. *et al.* Impact of water quality variations on the microbial metagenome across coastal waters in Shenzhen, south China. *Ocean & Coastal Management* **208**, 105612 (2021).
33. COVID, T. An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe* **1**, e99–e100 (2020).
34. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
35. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43**, e15 (2015).
36. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).

37. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).
38. Zhou, Z. *et al.* GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Research* **28**, 1395–1404 (2018).
39. Payne, M. *et al.* Enhancing genomics-based outbreak detection of endemic Salmonella enterica serovar Typhimurium using dynamic thresholds. *Microbial Genomics* **7**, (2019).
40. Bollback, J. P. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7**, 1–7 (2006).
41. Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* (2012) doi:10.1111/j.2041-210X.2011.00169.x.

Table

Table 1. Demographic and clinical symptom characteristics of *Vibrio parahaemolyticus* infections in Shenzhen, China, 2002-2018.

	Total (n=3,642)	Clonal groups (CGs)					Singleton (non-CGs, n=519)	P-clusters (n=2261)	Ob-clusters (n=1910)	Non-Ob-clusters (n=1732)
		Pathogenic CGs (PCGs)				Non-PCGs (n=483)				
		PCG3 (n=1,966)	PCG189 (n=326)	PCG-others (n=348)	Total (n=2,640)					
Type (no., frequency)										
Patient	2686 (74%)	1946 (99%)	324 (99%)	334 (96%)	2604 (99%)	21 (4%)	61 (12%)	2237 (99%)	1897 (99%)	789 (46%)
Environment	956 (26%)	20 (1%)		14 (4%)	36 (1%)	462 (96%)	458 (88%)	24 (1%)	13 (1%)	943 (54%)
Source										
IDDS	1856 (51%)	1337 (68%)	220 (67%)	246 (71%)	1803 (68%)	10 (2%)	43 (8%)	1548 (68%)	1259 (66%)	597 (34%)
FDOS	883 (24%)	622 (32%)	105 (32%)	90 (26%)	817 (31%)	30 (6%)	36 (7%)	702 (31%)	650 (34%)	233 (13%)
FSS	903 (25%)	7 (0%)		12 (3%)	20 (1%)	443 (92%)	440 (85%)	11 (0%)	1 (0%)	902 (52%)
Age										
Median (IQR)	28 (23–35), n=1910	28 (23–35), n=1397	28 (23–35), n=226	27 (23–36), n=242	28 (23–35), n=1865	29 (20–39), n=11	27 (23–37), n=34	28 (23–35), n=1630	27 (23–35), n=1412	28 (24–35), n=498
Sex (no., frequency)										
Female	827 (42%)	599 (41%)	98 (41%)	105 (43%)	802 (42%)	7 (58%)	18 (50%)	684 (40%)	593 (40%)	234 (46%)
Male	1152 (58%)	845 (59%)	142 (59%)	142 (57%)	1129 (58%)	5 (42%)	18 (50%)	1007 (60%)	872 (60%)	280 (54%)
Symptom (no., frequency)										
Diarrhea	1747 (100%)	1264 (100%)	210 (100%)	231 (100%)	1705 (100%)	9 (100%)	33 (97%)	1473 (100%)	1259 (100%)	488 (99%)
Abdominal pain	1369 (78%)	1001 (79%)	161 (77%)	173 (75%)	1335 (78%)	7 (78%)	27 (79%)	1159 (78%)	997 (79%)	372 (76%)
Vomiting	682 (39%)	508 (40%)	69 (33%)	87 (38%)	664 (39%)	5 (56%)	13 (38%)	586 (40%)	505 (40%)	177 (36%)
Fever	177 (10%)	140 (11%)	14 (7%)	19 (8%)	173 (10%)	0 (0%)	4 (12%)	148 (10%)	128 (10%)	49 (10%)

Figures

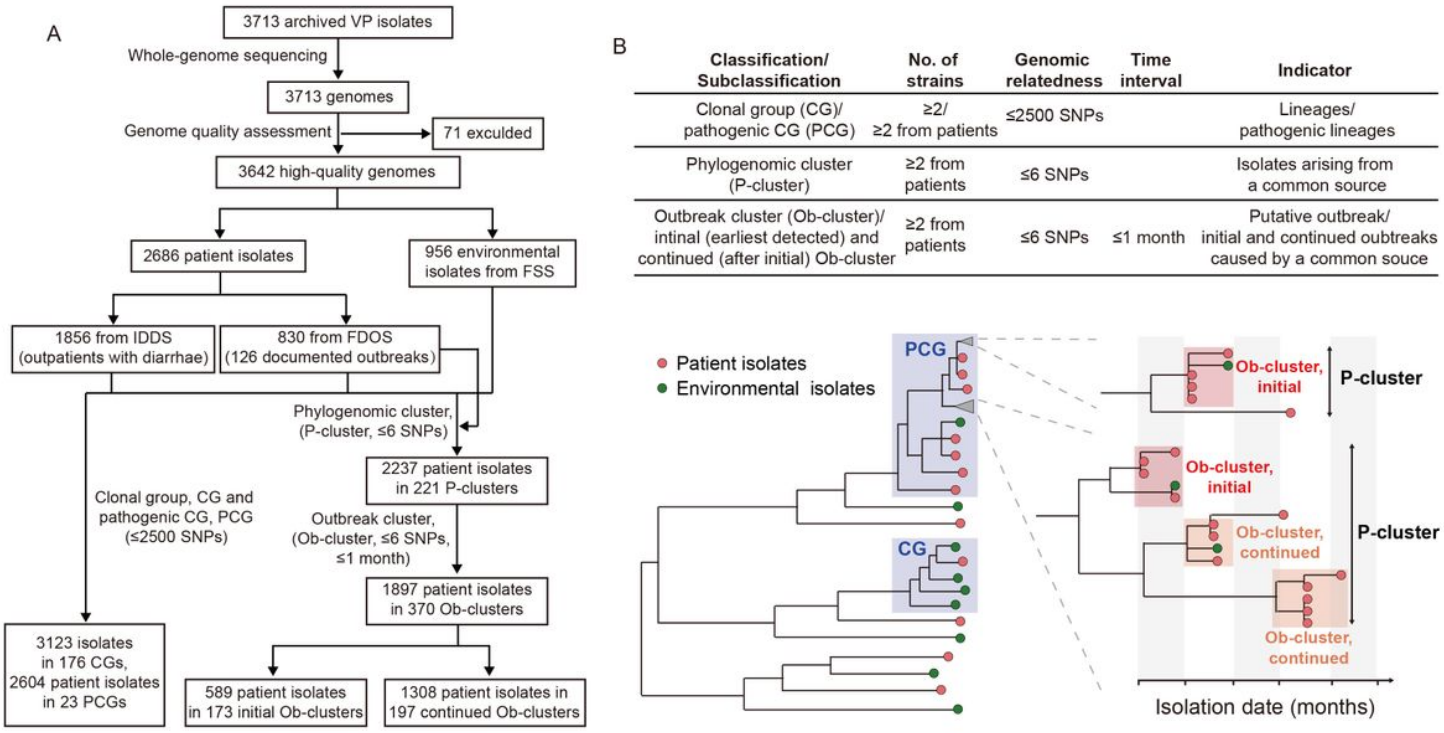

Figure 1

Study design flowchart (A) and definitions of the different classification concepts used in this study (B).

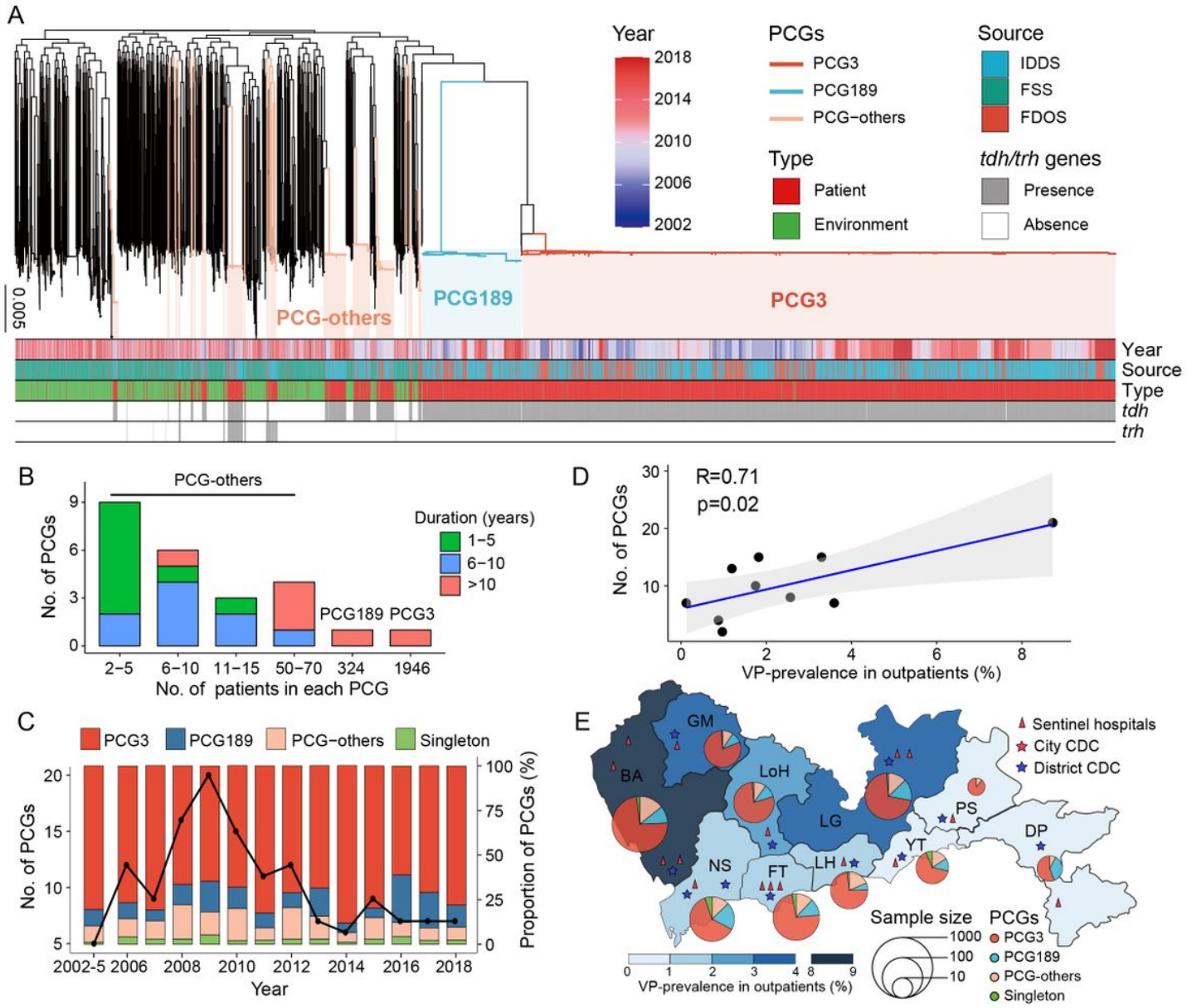

Figure 2

Population structure and spatiotemporal distribution of *Vibrio parahaemolyticus* (VP) lineages in Shenzhen, China, 2002-2018. (A) Maximum-likelihood phylogenic tree of 3,642 VP isolates. Colored branches and background indicate pathogenic clonal groups (PCGs). Bars below the tree show the isolation time, source, type, and presence/absence of two key virulence factors of the strains. (B) Size and duration distribution of PCGs. (C) Temporal dynamics of the number (line, left Y-axis) and proportion (bars, right Y-axis) of PCGs in patient isolates. Years (2002–2005) had less than 50 patient isolates were merged. (D) Correlations between VP-prevalence in outpatients with diarrhea and number of PCGs. The gray background is the 95% confidence interval. (E) Geographical distribution of PCGs in patient isolates from ten districts of Shenzhen. Pie charts show the proportion of different PCGs in each district, and their size is scaled based on the number of patient isolates.

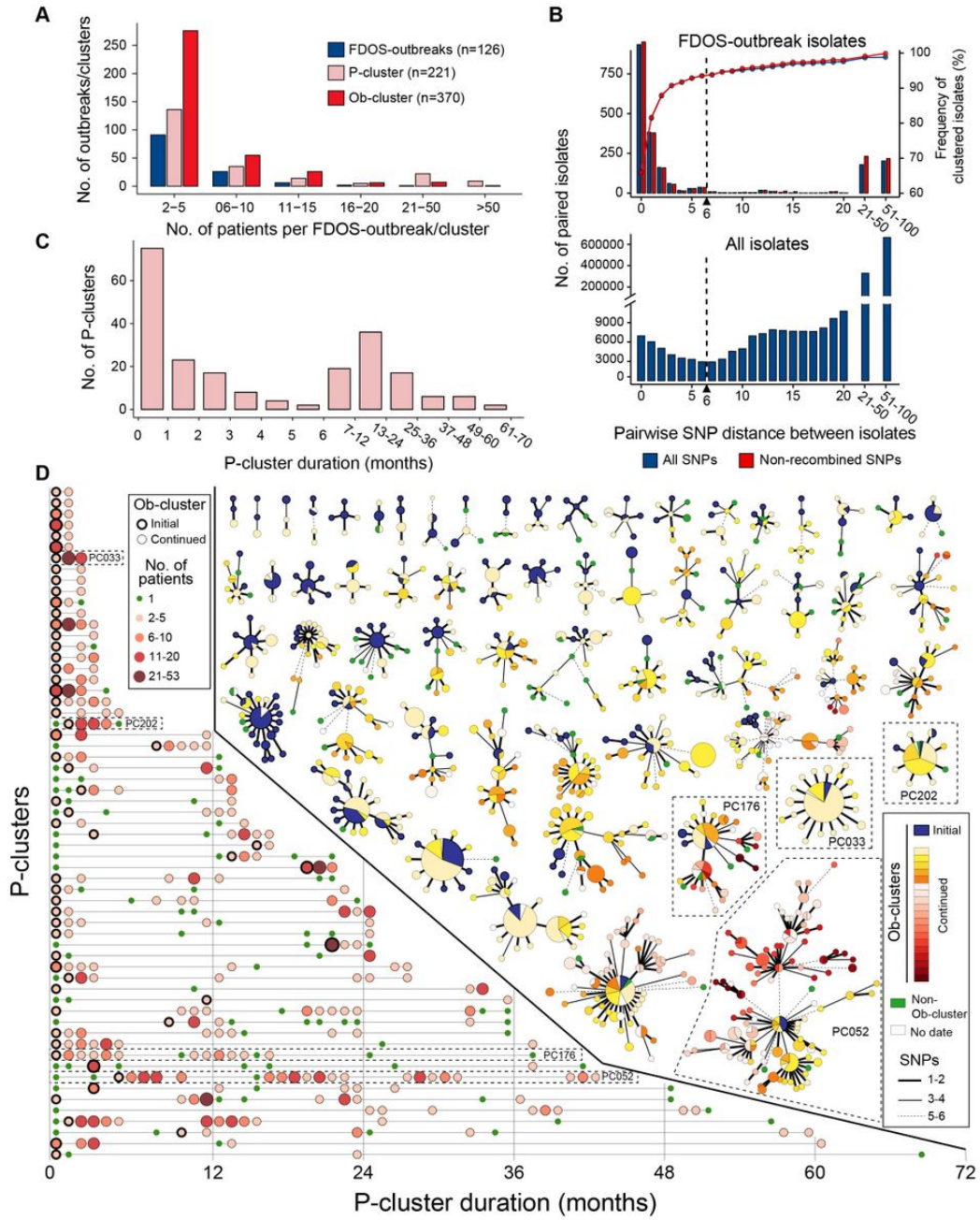

Figure 3

Phylogenomic clusters (P-clusters) and outbreak clusters (Ob-clusters). (A) Size distribution of Foodborne Disease Outbreak Surveillance (FDOS)-outbreaks, P-clusters, and Ob-clusters. (B) Pairwise SNP distances distribution (bars, left Y-axis) between FDOS-outbreaks (top) and all (bottom) isolates and the proportion of clustered FDOS-outbreak isolates under different SNP cutoffs (lines and points, right Y-axis). Blue and red indicate all the SNPs and non-recombined SNPs (after removing SNPs located in recombination regions), respectively. (C) Duration distribution of P-clusters. (D) Temporal distributions (left) and minimum spanning trees (right) of the 61 P-clusters that contained two or more Ob-clusters. Left: color and size of a point indicates the number of patients within-1-month, and bold and normal circles indicate initial and continuous Ob-clusters, respectively. Right: point colors indicate different Ob-clusters or non-Ob-clusters within a P-cluster, and point sizes are scaled based on the number of patients. The width and shape of the lines that link points indicate SNP distance.

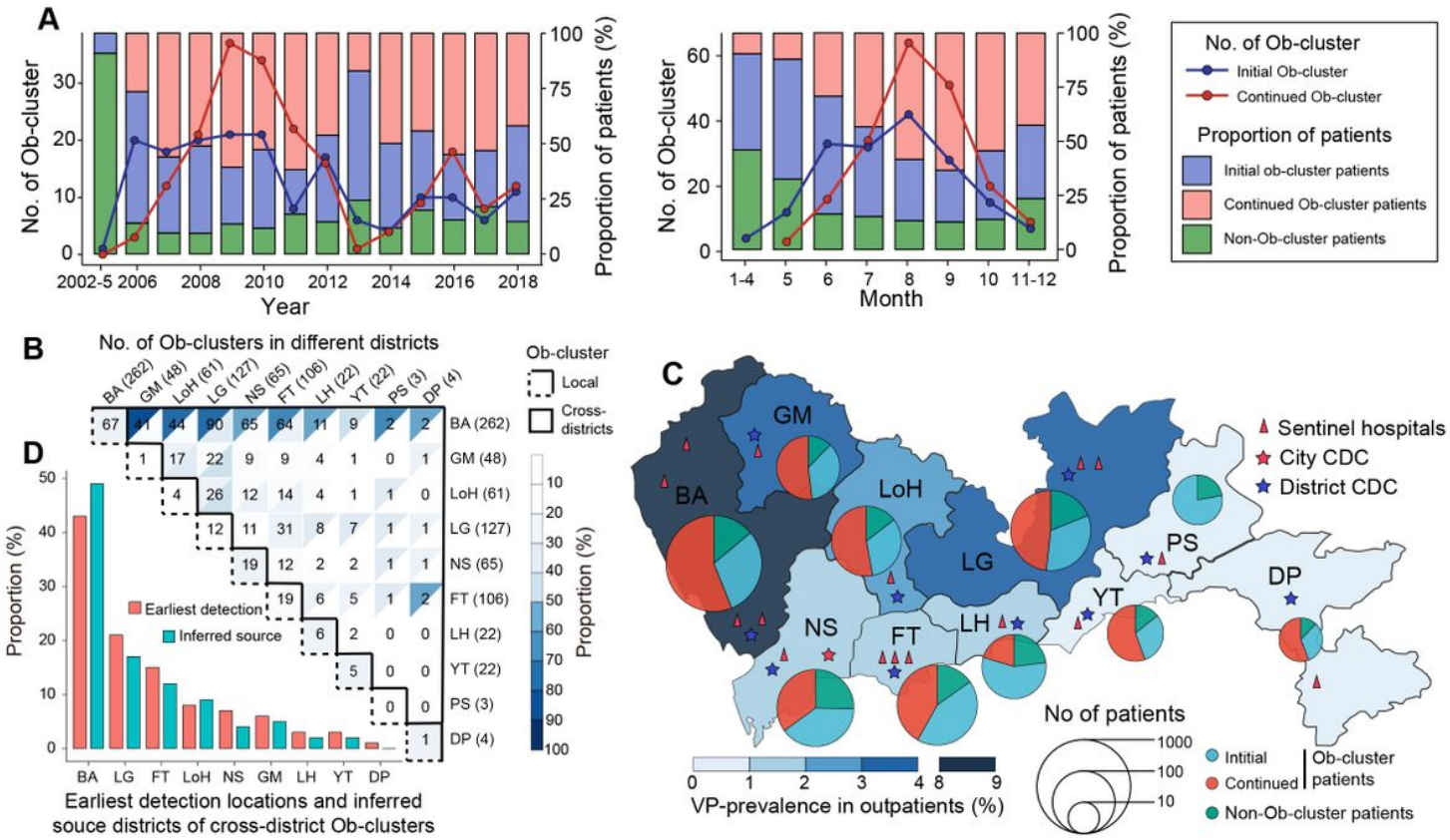

Figure 4

Spatiotemporal dynamics of outbreak clusters (Ob-clusters). (A) Temporal dynamics of the number (line, left Y-axis) and proportion (bars, right Y-axis) of initial (blue), continued (red) Ob-clusters/patients and non-Ob-cluster patients (green). Years/months with less than 50 patient isolates were merged. (B) Spatial distribution of Ob-clusters in ten districts. Numbers in brackets next to the heatmap indicate the total number of Ob-clusters in a district. Numbers in the heatmap indicate the number of local Ob-clusters or shared cross-district Ob-clusters between two districts, and colors indicate the proportions of local/cross-district Ob-clusters to the total number of the upper and right districts. (C) Geographical distribution of initial and continued Ob-cluster patients and non-Ob-cluster patients in ten districts. (D) Earliest detection locations (red) and inferred source districts (blue) of cross-district Ob-clusters.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.pdf
- ExtendedDataFig.pdf