Infection, Genetics and Evolution 82 (2020) 104316

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Research Paper

Comparative core/pan genome analysis of *Vibrio cholerae* isolates from Pakistan

Samia Zeb[a,b,*], Sardar Muhammad Gulfam[c], Habib Bokhari[a,*]

[a] Department of Biosciences, COMSATS University Islamabad (CUI), Islamabad, Pakistan
[b] Department of Microbiology, Hazara University, Mansehra, Pakistan
[c] Department of Electrical and Computer Engineering, COMSATS University Islamabad (CUI), Islamabad, Pakistan

A B S T R A C T

Cholera is an endemic disease in many regions of Asia including, Pakistan. *Vibrio cholerae*, the causative agent of cholera, is considered as one of the best adapted bacteria due to its ability to withstand severe environmental stresses. The *V. cholerae* genome is very plastic with many gene additions and deletions. In this study, we sought to understand the diversity of *V. cholerae* genes in two Pakistani subclades [e.g. Pakistani subclade I (PSC I) and Pakistani subclade II (PSC II)]. We have analyzed 44 PSC I and 56 PSC II strains, respectively. By analyzing our data, it was concluded that subclade group 2 (PSC II) has 2967 core genes repositories, while the PSC 1 group has just 1062 core genes. It was observed that the pangenome in the PSC II group is open while the pan-genome in PSC I are closed. It was also noted that the number of accessory genes (n = 2500) is higher in the PSC I group compared to the PSC II group (n = 550). Furthermore, analysis extended to the study of unique gene profiles suggested that all strains of the PSC II group have unique genes. One strain among the PSC II group had a high number of unique genes (n = 2612). However, in the PSC I group, only a few strains had unique genes with a maximum of 86 unique genes being found in a single strain. Core phylogeny of PSC I indicated that just three groups initially arose from a single common ancestor. At the same time, a complex pattern of evolution was found in the PSC II phylogenetic tree based on core gene information. This comparative genomic analysis has revealed 'waves' of *V. cholerae* evolution and information on its transmission and ability to modify its genetic content to survive in different environmental conditions. Here, we have investigated how the versatility of *V. cholerae*, a bacterium that persists across different habitats, is reflected in its genome. The data generated during the study should be extremely beneficial in defining the evolutionary relationship as well as diversity between *V. cholerae* subclades. It will also benefit epidemiological studies and the design of better treatment strategies for controlling epidemics.

1. Introduction

Vibrio cholerae is a globally dispersed pathogen that has evolved with humans for centuries and is the causal agent of the disease cholera, which is a potentially severe intestinal illness that affects ~1–5 million people and is responsible for up to 140,000 deaths annually (Ali et al., 2015). Even with modern established treatments and preventive measures, *V. cholerae* has continued to emerge as a dangerous pathogen. This is especially true in Southeast Asia, where the yearly appearance of cholera cases follow predicted patterns or "seasons" (Russell, 1925; Pascual et al., 2002). Aquatic ecosystems are often its natural habitat (Reidl and Klose, 2002). On the bases of somatic (O) antigens, 200 serogroups have been identified to date (Chatterjee and Chaudhuri, 2003; Seed et al., 2012). Only isolates in serogroup O1 (consisting of two biotypes, classical and El Tor, and the serotypes Ogawa and Inaba) and O139 have been identified as agents of cholera epidemics and pandemics (Harris et al., 2012).

Our knowledge of the epidemiology, etiology, and evolution of diarrheal diseases has now improved with the help of whole sequencing and population genomics (Wilson, 2012). By these bioinformatics approaches, strains of *V. cholerae* have been sequenced and compared worldwide over the course of century, clarifying the history of the current pandemic (Mutreja et al., 2011). This has shown that this pandemic is the result of a single clonal expansion of one *V. cholerae* O1 El Tor ancestor, accompanied by horizontal gene transfer (HGT) events involving toxin and antibiotic resistance genes (Chun et al., 2009). HGT, along with natural selection and gene duplication, drives the adaptive evolution of microbial genomes, whereas their relative importance is not well understood (Hemme et al., 2016; Zhang et al., 2017). HGT events generally occur through mobile genetic elements (GMEs) that bring genes, similar to functional genes that aid organism survival and adaptation, to the pathogen. Recent studies have indicated that more than 20% of the microbial genes are acquired via HGT (Popa et al., 2011; Polz et al., 2013). This effect brings new genes into a genome, even from taxonomically unrelated species.

The *V. cholerae* genome has the ability of transformation. The lateral or horizontal transfer of genes like virulence genes by phage (Faruque et al., 2005), pathogenicity islands, and other accessory genetic

* Corresponding authors at: Department of Biosciences, COMSATS University Islamabad (CUI), Islamabad, Pakistan.
E-mail addresses: samiazeb.84@gmail.com (S. Zeb), habib@comsats.edu.pk (H. Bokhari).

https://doi.org/10.1016/j.meegid.2020.104316
Received 28 October 2019; Received in revised form 31 March 2020; Accepted 3 April 2020
Available online 08 April 2020

elements provide insights into how bacterial pathogens emerge and evolve to become new strains (Boyd, 2008). More recently, comparative genomics has been applied to answer epidemiological questions such as the origin of a specific pathogen like *V. cholerae*. It proves that the current Haitian outbreak of cholera that started in 2010 is caused by a strain of Asian origin (Chin et al., 2011; Hendriksen et al., 2011). Azarian et al. (2014) compared 60 clinical and environmental isolates collected in Haiti from 2010 to 2012 by whole-genome sequencing and performed a single nucleotide polymorphism (SNP) analysis. They found that the strains from the years 2011 and 2012 have rapidly diverged from the 2010 ancestral strain that initiated the outbreak, suggesting evolution driven by positive selection in a new environment (Azarian et al., 2014). The microbial genome is divided into core and accessory elements, which combined constitute the pan-genome (Tettelin et al., 2008). A pan-genome is the complete repertoire of genes in a bacterial species, which includes the core genome containing genes present in all strains, a dispensable genome containing genes present in two or more strains, and unique genes specific to single strains (Medini et al., 2005).

V. cholerae has evolved to possess complex signal transduction and gene regulatory systems to survive and grow under various environmental conditions (Bhadra et al., 2008). Several studies have proposed that the origin of the serogroup O139 strain was an El Tor strain that obtained the O139 biosynthesis genes (as well as the SXT element and a capsule) via antigenic switching from a donor strain (Berche et al., 1994; Mooi and Bik, 1997; Stroeher et al., 1997; Waldor and Mekalanos, 1994). Recently, it has been proposed based on comparative sequence analysis that an O22 serogroup may be a possible donor for the O139 serogroup (Dumontier and Berche, 1998; Yamasaki et al., 1999). Unique genes play a significant role in organism evolution, especially the genes responsible for virulence, disease, and defense and niche adaptations. Therefore, in this study, we have focused on the identification of these unique genes as well. Cholera has been endemic in southern Asia since its recorded history and there is a region-specific distribution pattern of these diarrheal pathogens that needs to be taken into account for disease control and vaccine development (Lozano et al., 2012).

In this study, we characterized the genomic diversity of pre-isolated 100 isolates of *V. cholera* from Pakistan taken the 2009–2013 cholera outbreak. In our previous studies, Shah et al. (2014) characterized them into two subclades, called Pakistani subclade I (PSC I) and Pakistani subclade II (PSC II). This study has been conducted as at present, there is a need for a comprehensive core and pan-genome analysis of this pathogen to learn its evolution pattern, virulence, and ability to combat stresses. Bacterial Pan-genome Analysis (BPGA), an ultra-fast software package that provides comprehensive genome analysis details of microorganisms (Chaudhari et al., 2016) has been used in this study.

2. Material and methods

The study was approved by the Departmental Ethical Review Board, Biosciences, CUI, Islamabad. In this research, the authors extended the previous studies (Zeb et al., 2019a, 2019b) and did a comparative genomic analysis of pre-isolated species. During initial study, one of authors isolated 100 *V. cholerae* strains from clinical (serogroup O1, biotype El Tor) samples taken during cholera outbreaks and epidemics in Pakistan from 2009 to 2013. Further, serotyping and biotyping was performed, categorized them into two groups named Pakistani subclade I (PSC I) and Pakistani subclade II (PSC II). The sequencing was performed using an NGS platform. The methods for isolation, library preparation, and sequencing have already been discussed in the manuscript published (Shah et al., 2014) as a result of the initial study. The tools used for the analysis were Prokka, BPGA, and Rapid Annotations using Subsystems Technology (RAST).

The BPGA tool selected for the core and pan-genome analysis, requires the data to be inputted in a unique file format. To create standard input files that could be fed to BPGA, initially, annotation was carried out by another tool, Prokka software v1.12. Prokka is a whole-genome annotation tool used for quick annotation of bacterial, archaeal, and viral genomes to produce standards-compliant output files. In Prokka, genomes up to a size of 4 Mbp can be quickly and wholly annotated in about 10 min on a quad-core computer and scales well to 32 core SMP systems. It produces GFF3, GBK, FAA, ERR, FNN, FNA, FSA, SQN, and TBL files as output. These files are ready to be used as input files in different software. For this research, the version Prokka v1.12, was used.

From the files obtained from Prokka for all our strains, GBK and FAA output files were then used as input data files on BPGA software, to get the core and pan-genome details. BPGA has seven functional modules. We used six modules of this pipeline. Their details are given below.

2.1. Pan-genome profile analysis

Pan/core genome size of the strain can be determined by this module of the pipeline. It gives pan and core genome curves and their different gene families' distribution within the genomes under study. It can also determine the state of a genome as opened or closed.

2.2. Pan-genome sequence extraction

This module yields core, accessory, and unique protein families.

2.3. Exclusive gene family analysis

This module of the BPGA pipeline can identify the orthologous protein families that contain genes exclusively from a specific genome of the dataset (i.e., unique genes or singletons). The module also extracts the protein sequences of such exclusively present/absent families.

2.4. Atypical GC content analysis

This module does the atypical GC analysis, based on which genes are further categorized into core, accessory, and unique genes.

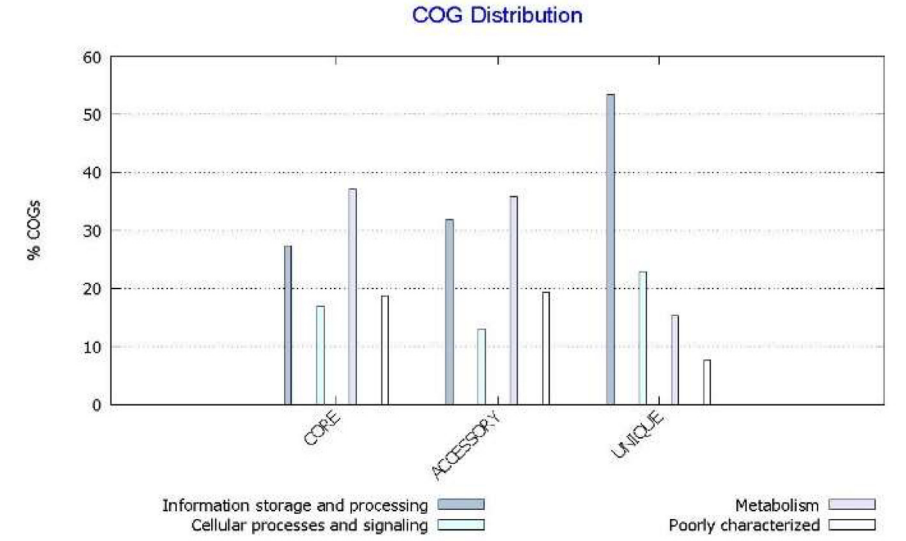
2.5. COG & KEGG determination

This is another module in the BPGA pipeline which can be used for COG and KEGG pathway determination. Each orthologous protein cluster is BLAST against reference COG and KEGG databases to assign COG and KEGG ids to all representative protein sequences in a specific cluster. Further, the percentage frequencies of COG and KEGG categories are calculated for core genes, accessory genes, and singletons (strain-specific genes).

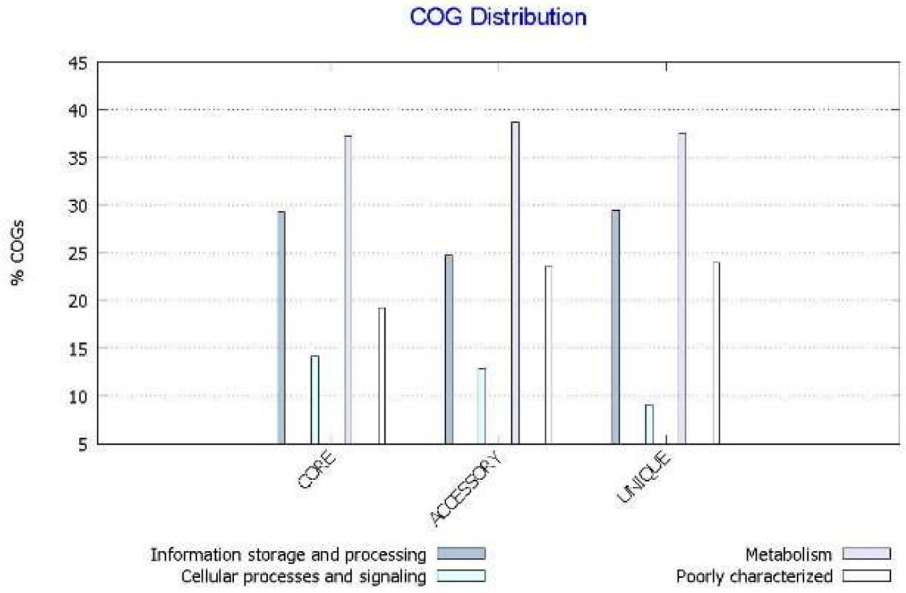
2.6. Phylogenetic tree construction

Phylogenetic analysis is carried out based on concatenated core gene sequences. BPGA can perform this evolutionary analysis by first aligning concatenated core genes and then calculating the gene matrix. The calculation of this matrix is based on similarity or dissimilarity among orthologous gene clusters. The initial step involves the extraction of protein sequences from 20 random orthologous gene clusters. BPGA is designed to have MUSCLE that performs concatenate multiple sequences alignments following which a neighbor-joining phylogenetic tree can be constructed.

To address the diversity of the pathogen, RAST, an annotation server was further used to obtain the comparative analysis of genes related to virulence, disease and defense, stress response, and pathogenicity islands. These included function-based and sequence-based comparisons of PSC I and PSC II and the particular strains from both subclades that had a maximum number of unique genes. For ease of



(a) PSC I



(b) PSC II

Fig 1. COG details of PSC I & PSC II.

understanding, the unique strains were assigned the same name given to them at the time of their isolation, that is, NP14 and D4, and the strains without any unique genes were referred to as normal PSC I and normal PSC II.

3. Results

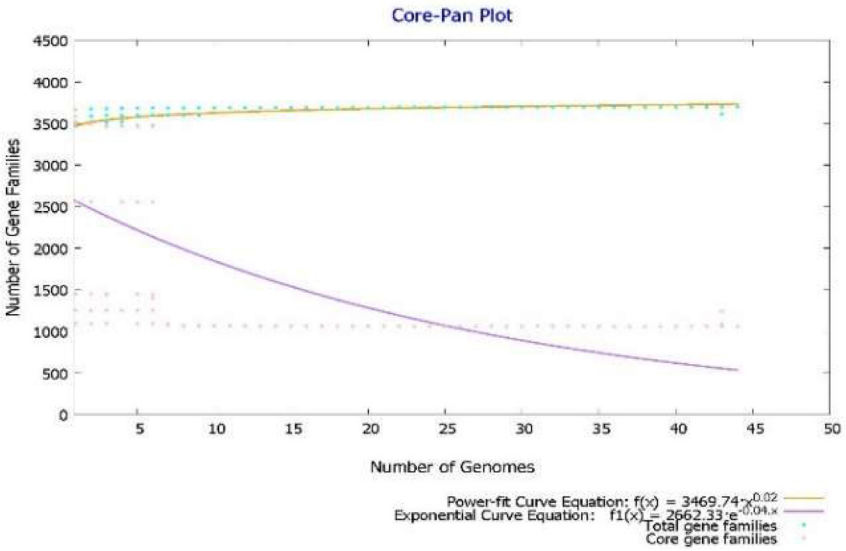
The BPGA pipeline identified functional genes present in all strains (i.e., the core genome), accessories genes, unique genes, and exclusively absent genes.

3.1. Comparative COG distribution/details of PSC I and PSC II

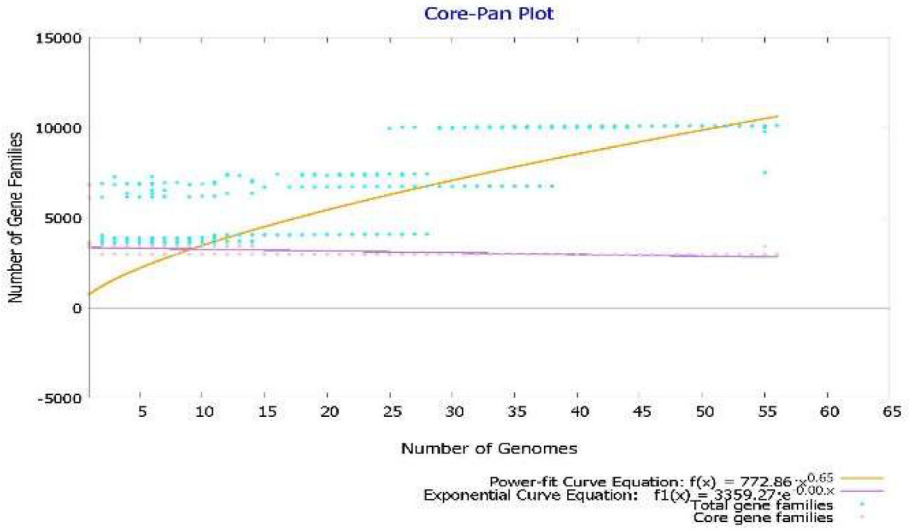
COG distributes all genomic sequences into three major categories, namely, core, accessory, and unique genes, which are shown in Fig. 1. All these categories consist of genes for information storage and processing, cellular processes and signaling, metabolism, and some poorly characterized genes. There are more unique genes (55%) for information storage and processing in PSC I as compared to PSC II. Furthermore, PSC II has double the unique genes for metabolism as compared to PSC I. This means these strains are more resistant to harsh nutrient conditions or can use different substrates to obtain energy.

3.2. Comparative core pan dot plot of PSC I and PSC II

The comparative core pan dot plot of PSC I and PSC II is shown in Fig. 2. It can be seen that PSC I appears closed as the power fit curve does not meet the exponential curve. The genome size of PSC I does not increase and the number of total gene families remains constant, while in PSC II, the size of pan-genome and the number of sequenced genomes



(a) PSC I



(b) PSC II

Fig. 2. Core pan dot plot of PSC I & PSC II.

increases.

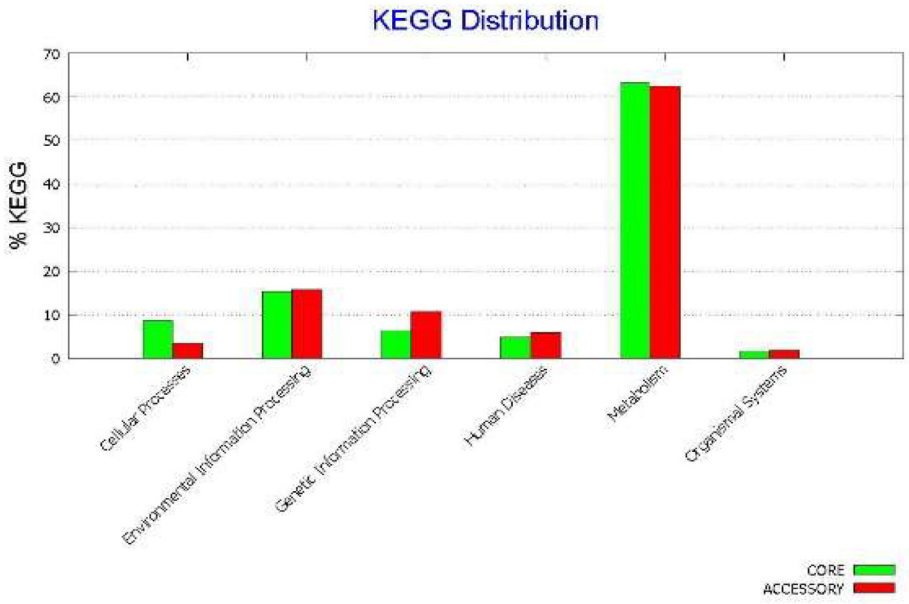
3.3. Comparative KEGG distribution/details of PSC I and PSC II

The comparative KEGG distribution/details of PSC I and PSC II are plotted in Fig. 3. The KEGG percentage shows more gene clusters for metabolism in PSC II as compared to PSC I. Interestingly, in PSC I there are no unique genes for any of the KEGG categories. In PSC II, KEGG has a greater percentage for metabolic genes that have a high number of core, accessory, and unique genes. Environmental information processing genes are the second-highest category of genes repository in PSC II.

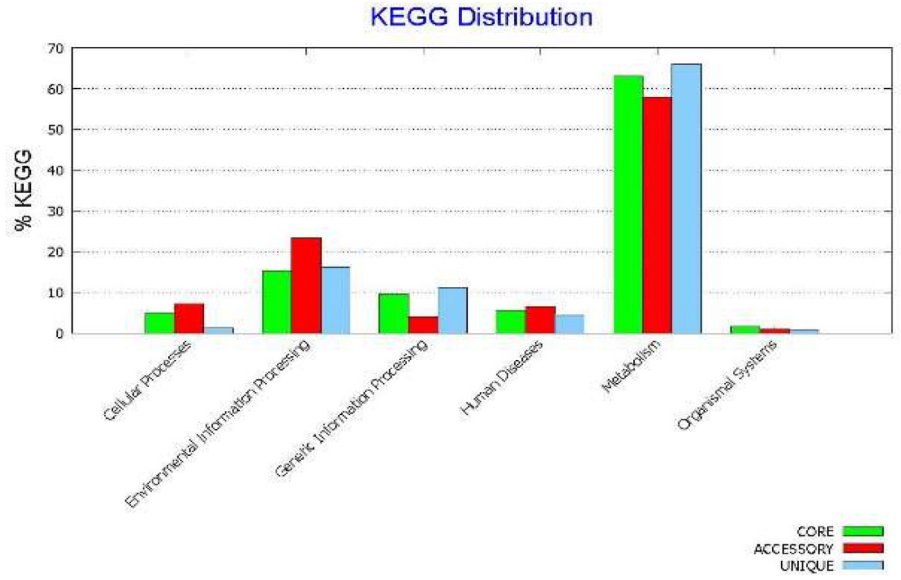
3.4. Whole genomic picture of PSC I and PSC II isolates

The whole genomic picture of PSC I and PSC II isolates is shown in Table 1 and Table 2. The detailed pan-genome analysis of the PSC I group reveals 1062 core genes. Similarly, the number of accessory genes in almost every strain of PSC I is relatively high and constant, i.e., 2440 to 2460, except for three strains that have a deficient number of accessory genes at 191, 388, and 33. Further analysis of these three strains concludes that there are exclusively absent genes as well and each exclusively absent gene lacks 24, 3, and 180 genes, respectively. Only one strain here seems to possess a reasonable number of unique genes (i.e., 86 unique genes) without any decrease in accessory genes or any other gene loss.

In the detailed pan-genome analysis of PSC II, the core gene number is two times that of PSC I core gene size and accessory gene number is about four to five times lower than PSC I accessory gene size. The exact core gene number is 2967 and accessory gene number ranges between 504 and 540, except for two strains where it increases to a large number (i.e., 3821 and 3811) and for one strain, where it is as low as 51. Like PSC I, in PSC II, the number of exclusively absent genes is also found only in three strains. One strain that lacks 433 genes has 340 unique genes as well. The second strain that lacks 8 genes has an interestingly



(a) PSC I



(b) PSC II

Fig. 3. KEGG details of PSC I & PSC II.

high number of unique genes (i.e., 2612). The third strain that lacks just 2 genes has 91 unique genes as well. Overall, 2 to 4 unique genes are found in almost every strain of PSC II.

3.5. Core phylogeny

The core phylogeny of PSC I indicates that just three groups initially arise from a single common ancestor. The topmost cluster contains a particular strain shown in the red box in which accessory genes drop to just 199 and it also lacks 24 genes exclusively. Similarly, a strain with 86 unique genes is found to be closely related to another strain in which the accessory gene number drops to 388, shown at the base of the tree indicated in the yellow box in Fig. 4(a). Further, the purple box indicates the most evolved strain in which the accessory gene number drops to just 33.

An intricate pattern of evolution is found in the PSC II phylogenetic tree shown in Fig. 4(b). The phylogeny of PSC II isolates shows that its topmost clusters are comprised of five strains. Here, one strain from 2010 isolates is found as closely related to the 2009 isolates. The cluster at base of the tree contains all the 2013 isolates except one, which is clustered with the 2010 isolates. It is found to be significantly divergent from the rest of the 2013 isolates. Its correlation can be predicted from the above picture where it is found to have low accessory genes along with the 340 unique genes in its genome. 344 genes are found to be

Table 1
PSC I pan-genome details.

Strains	1	2	3	4	...	43	44
Core Genes	1062 genes						
Accessory Genes	388	33	199	2500 genes			
Exclusively Absent Genes	3	180	24	Nil			
Unique Genes	Nil						86 genes

Table 2
PSC II pan-genome details.

Strains	1	2	3	4	...	55	56
Core Genes	2967 genes						
Accessory Genes	51	3811	3821	550 genes			
Exclusively Absent Genes	433	2	Nil			8	Nil
Unique Genes	340	91	17	Nil		2612	115

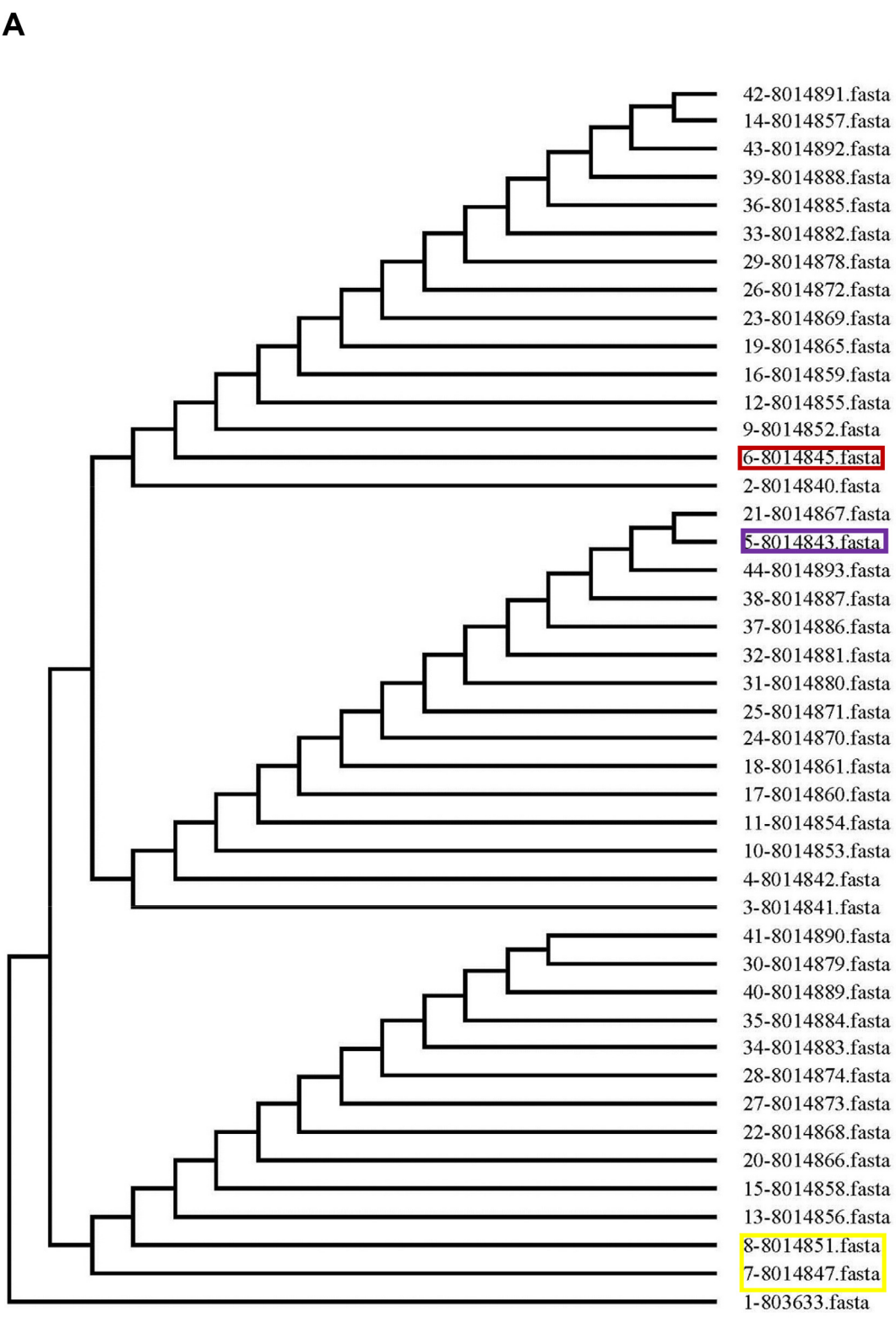


Fig. 4. (a): Core phylogenetic tree of PSC I. (b): Core phylogenetic tree of PSC II.

exclusively absent in its genomic repository. Isolates of 2011 are also found to be closely related to the 2010 isolates.

3.6. Analysis of selective strains from both subclades having the maximum number of unique genes

Strains from both subclades are selected based on their maximum unique gene repository and further comparatively analyzed with a strain without any unique genes within the same subclade. For ease of understanding, the unique strains are given the same name as that from the time of their isolation, that is, NP14 (which is the unique strain in PSC I group), and D4 (which is the unique strain in PSC II). The strains without any unique genes are referred to as normal PSC I and PSC II.

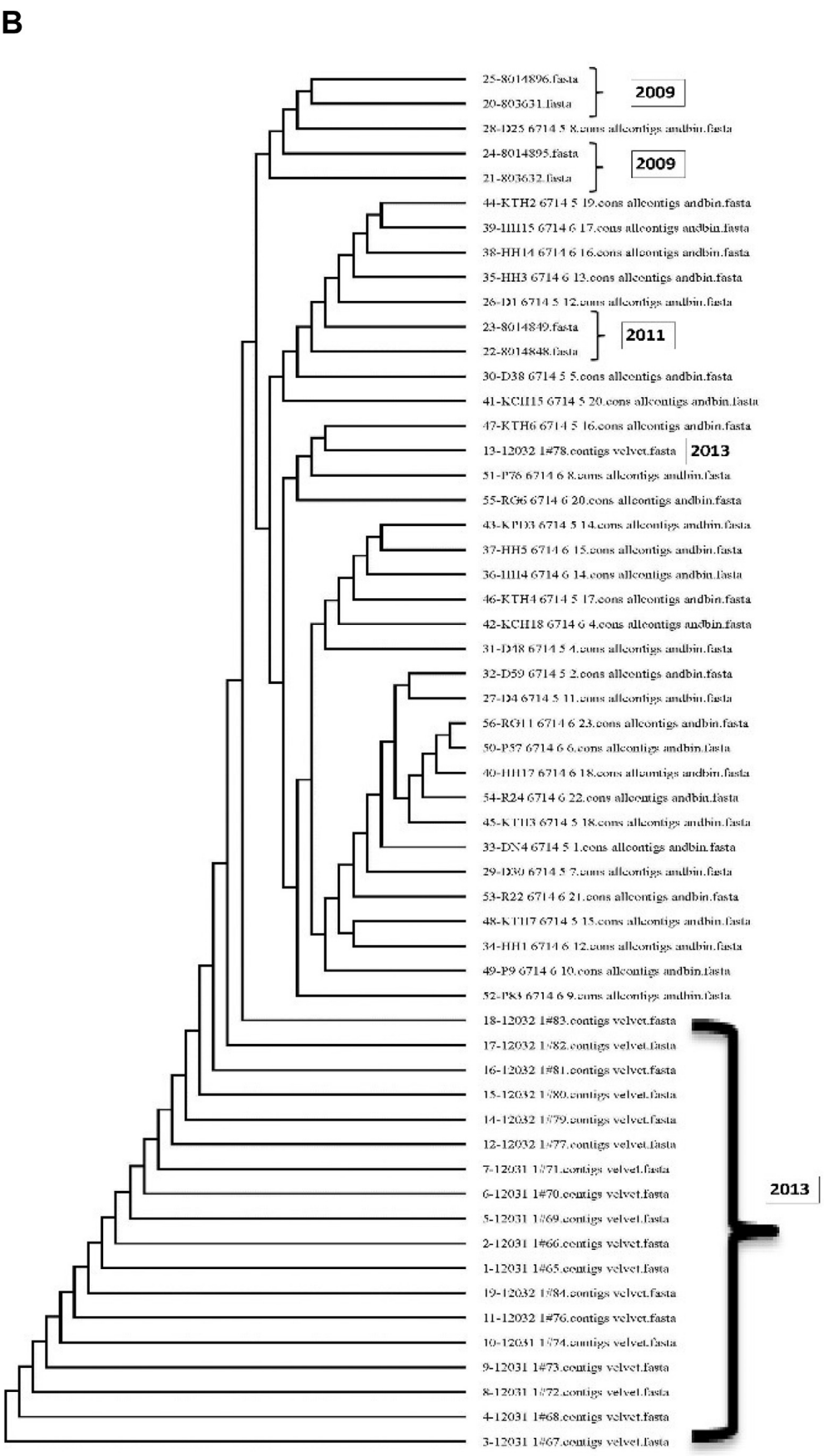


Fig. 4. (*continued*)

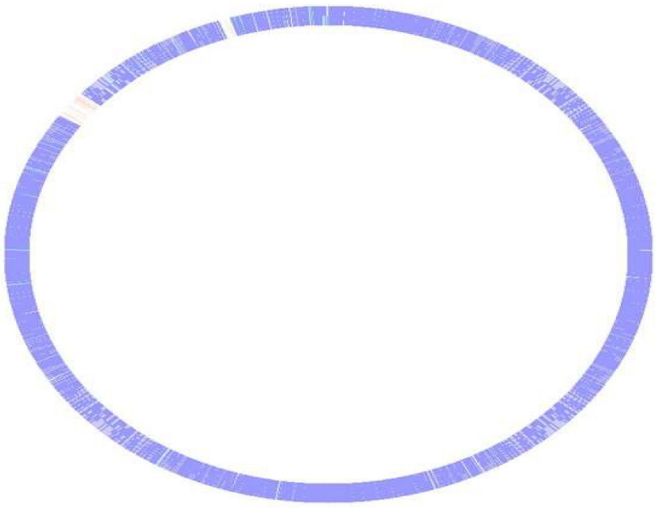


Fig. 5. Sequence based comparison of NP14 with normal PSC I without any unique gene.

3.6.1. Sequence-based comparison of NP14 with normal PSC I

Sequence-based identity of NP14 and normal PSC I without any unique gene is shown in Fig. 5. In this comparison, it is found that although NP14 has unique genes, overall, most of the genome is identical to the PSC I strain that did have any unique gene in its genetic repository. The percent protein sequence identity is shown in Table 3a. This table shows the bidirectional and unidirectional hits. The bidirectional hit indicates the best hits for proteins present in both genomes under comparison, while the unidirectional hit shows hits for sequences present in one of the compared genomes. The table also shows that as most proteins are 99 to 100% similar, we look for either bidirectional or unidirectional hits. A small part of the protein is non-identical. Further analysis of non-identical genomic sequences reveals that unique sequences of NP 14 contribute this non-identical portion.

3.6.2. Function-based comparison of NP 14 with normal PSC I

The function-based analysis of the selected genetic group/category includes the comparison of virulence, disease and defense genes, the region encoding genes for phage, prophage, and transposable elements, and stress response. The details are shown in Table 3b and 3c, and Fig. 6. In the function-based analysis of NP14 and normal PSC I, it is found that NP14 has 1 unique gene for pathogenicity islands, 4 genes for cold shock, 2 for heat shock, 5 for osmotic stress, and 11 for oxidative stress. For resistance to antibiotics and toxic compounds, there are 13 genes in NP14. This strain also contains 5 genes for toxins and superantigens. On the other hand, the normal PSC I strain that do not have any unique genes lack all these essential genes.

3.6.3. Sequence-based comparison of D4 with normal PSC II

The sequence-based identity of D4 and normal PSC II without any unique genes is shown in Fig. 7. This analysis reveals that half the genome does not possess a remarkable similarity to the strain belonging to its subclade. The percent protein sequence identity is shown in Table 4a.

This table shows the best bidirectional and unidirectional hits for the D4 genome and normal PSC-II strain. The protein sequence similarity in the genome of compared strains is indicated by the intensity of the color blue. As the intensity decreases, the percentage similarity index decreases. The color pink shows the least identity as 10%. Bidirectional hits show the similarity index for genes present in both genomes, while unidirectional hits indicate the presence of specific genes solely in D4.

Table 3b
Function based comparison of NP14 with a normal PSC I.

Characteristics	NP 14	PSC I (normal)
Size	4,160,909	4,103,053
GC content	47.3	47.4
Number of subsystems	531	378
Number of coding sequences	3919	3837
Number of RNAs	80	86

Table 3c
Special functions of NP14.

Functions	No. of Genes in NP14
Pathogenicity island	1
Cold shock	4
Heat Shock	2
Osmotic stress	5
Oxidative stress	11
Resistance to antibiotic	13

3.6.4. Function-based comparison of D4 with normal PSC II

The function-based analysis of the selected genetic group/category includes the comparison of virulence, disease and defense genes, the region encoding genes for phage, prophage, and transposable elements, and stress response. The details are shown in Table 4b and 4c, and Fig. 8. It is found that in D4, there is 1 unique gene for adhesion and 2 genes for bacteriocin. For resistance to antibiotics and toxic compounds, there are 21 genes in D4. There are 8 unique genes for phage, prophage, transposable elements, and plasmids, and 2 for pathogenicity islands. To combat stresses, there are 5 genes for cold shock, 6 genes for heat shock, 16 genes for osmotic shock, and 14 genes for oxidative stress in D4. On the other hand, the normal PSC II strain does not have all these essential genes.

Table A1 in the appendix describes all the genomes with their year and group. Moreover, it is apprised that all the sequence data are already submitted to the European Nucleotide Archive (ENA) with their accession codes. However, the IDs are included under the column titled “submission name” in Table A1 for the ease of the readers. Using these IDs, any sequence and its properties can be found on the ENA database, and all the files can be downloaded, as well.

4. Discussion

The *V. cholerae* groups named PSC I and PSC II are responsible for cholera outbreaks in Pakistan between the years 2009–2013. Through this study, we have found that the genomes of *V. cholerae* are surprisingly non-conserved between all isolates over the four-year period in which they were isolated. The 100 *V. cholerae* strains that were analyzed in the present study represent the largest collection of genome-sequenced strains for this pathogenic bacterium. The multiscale comparative approach used in this work provides insights into the diversity of *V. cholerae* strains. This is demonstrated by the BPGA analysis of

Table 3a
Percent protein sequence identity.

Bidirectional best hit	100	99.9	99.8	99.5	99	98	95	90	80	70	60	50	40	30	20	10
Unidirectional best hit	100	99.9	99.8	99.5	99	98	95	90	80	70	60	50	40	30	20	10

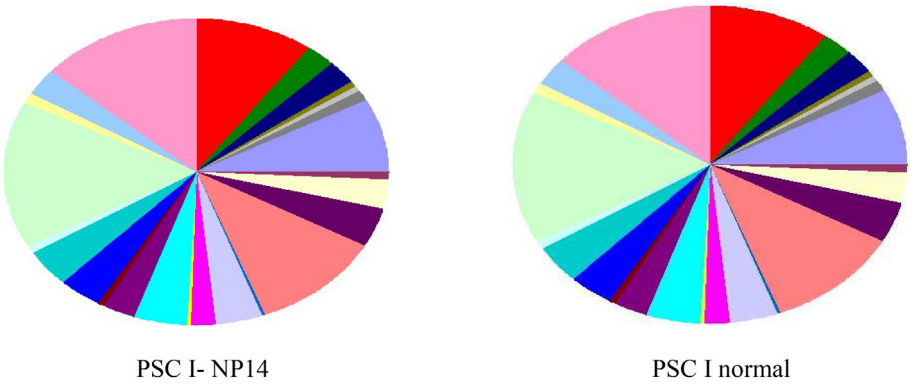


Fig. 6. Function based comparison of NP4 with PSC I without any unique gene.

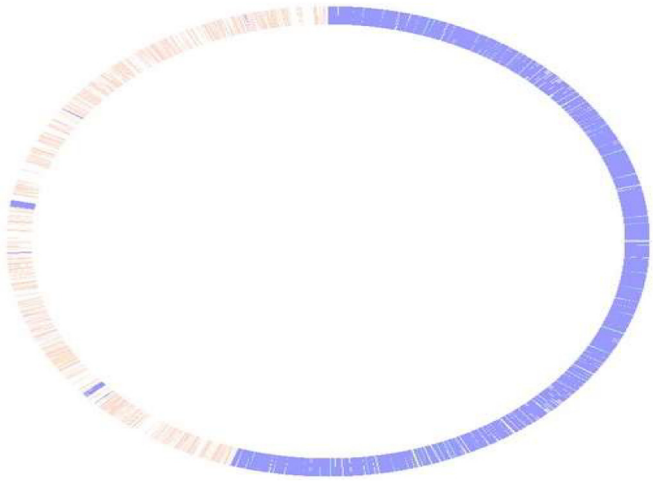


Fig. 7. Sequence based identity of D4 & PSC II without any unique gene.

Table 4b
Function based comparison of D4 with a normal PSC II.

Characteristics	D4	Normal PSC II
Size	6,937,234	4,035,656
GC content	43.2	47.5
Number of subsystems	601	377
Number of coding sequences	6619	3716
Number of RNAs	10	90

Table 4c
Special Functions of D4.

Functions	No. of Genes in D4
Pathogenicity island	2
Cold shock	5
Heat Shock	6
Osmotic stress	16
Oxidative stress	14
Resistance to antibiotic	21

whole-genome sequences which indicates that the number of accessory and unique genes varies a lot among the isolated collection.

The pan-genome analysis has revealed that the PSC I group of strains contains a core genome of 1062 (Table 1), while PSC II has 2967 (Table 2). The level of core gene content for PSC II is almost three times more than PSC I. For determining phylogenic relationships, the core genome is the optimum dataset and is most suitable for the practical determination of phylogenetic reconstruction, rather than using 16S/23S rRNA or essential housekeeping genes. In PSC I the core genome is used to determine phylogeny and the 44 V. cholerae strains are clustered in 3 groups (Fig. 4(a)). This closely related taxa-based cladogram represents the strain in which the number of accessory genes drops to just 388, arranged in a cluster along with NP14 which can be seen at the base cluster of the PSC I phylogenetic tree. NP14 has the maximum number of unique genes (i.e., 86) accompanied by a high number of accessory genes (i.e., 2522). Another strain in which the number of accessory genes drops to just 33 seems to be the most-evolved strain and is found in mid cluster end taxa branch. Interestingly, this strain exclusively lacks 180 genes as well. The third strain in which accessory gene number drops to 199 also lacks 24 genes and is clustered with the topmost taxa.

In the case of PSC II, the cladogram results in a very complicated tree (Fig. 4(b)). The D4 strain that has a high number of unique genes does not have any alteration in the number of accessory genes. It appears as the most evolved/unique strain found at the end of the cladogram taxa node. Similarly, 2 other strains isolated in 2010 that have a high number of accessory genes, that is, 3811 and 3821 genes each, accompanied by 91 and 17 unique genes, respectively, are found to be the most evolved. It is astonishing that one of the PSC II strains in which the number of core genes dropped to just 51(which is also the only strain in which the number of accessory genes have dropped), has 340 unique genes and exclusively lacks 433 genes as well. This gene belongs to the 2013 isolated collection and is clustered with 2010 isolates.

Core/pan indicate that the pan-genome size of PSC II strains increases constantly (Fig. 2). This is most probably due to an increasing number of new gene additions. This is similar to the findings of a study by Marin and Vicente (2013) in which they presented a pan-genome profile analysis where the cluster numbers of core genome are almost the same, but the pan-genomes vary greatly. Additionally, the PSC I genome is found to be closed and none of the strain of PSC I possess any unique genes for KEGG pathways (Fig. 3).

To understand this diversity, sequence-based and function-based comparisons have been carried out for selected strains that have the maximum number of unique genes, with strains from both subclades

Table 4a
Percent protein sequence identity.

Bidirectional best hit	100	99.9	99.8	99.5	99	98	95	90	80	70	60	50	40	30	20	10
Unidirectional best hit	100	99.9	99.8	99.5	99	98	95	90	80	70	60	50	40	30	20	10

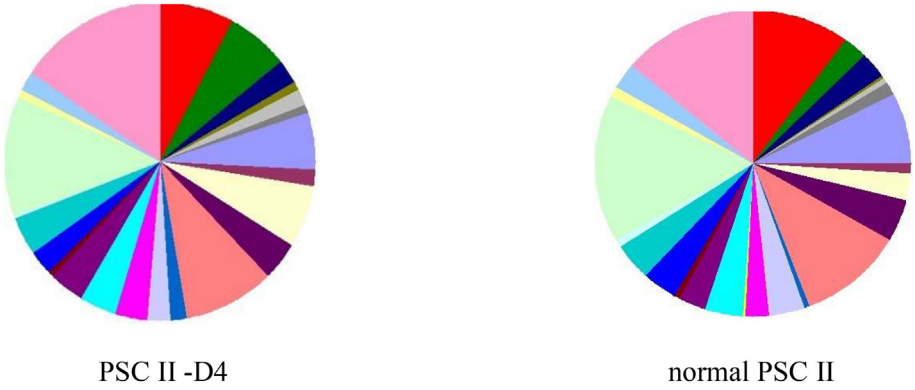


Fig. 8. Function based similarity between D4 & PSC II without any unique gene.

without any unique genes. As for this study, specific categories like genes for virulence, disease, defense, resistance to antibiotics, and stress response genes (because they occupy a different range of niches) are essential, so we have elucidated the differences between these genes.

The comparative genomic study reflects that the NP14 strain with the maximum number of unique genes has 22 genes for stress response and 13 for antibiotic resistance profile, and its normal counterpart also lacks these genes. Although their GC content is the same, NP14 has up to 531 more subsystems, while its counter strain has 378 (Table 3). Sequence-based identity shows that NP14 has a high protein-based similarity with its counter strain (Fig. 5).

The same comparative approach, when used for D4 and the normal PSC II strain without any unique genes, shows that half the genome does not possess a reasonable identity to its counter strain (Figs. 7 and 8). In D4, the GC content is 43.2 while in its counterpart strain, it is 47.5. The number of subsystems in D4 are double to that of its counterpart. The number of RNAs in D4 is 10 while its counter strain has 90 (Table 4). D4 has an additional 41 genes for stress response and 21 for antibiotic resistance that are missing in its counter strain.

Recent work indicates that the significance of measured genetic variation can be applied and interpreted as per the researcher's discretion. This data can be used to calculate a “molecular clock” like Feng et al. (2008) did in 2008, using sequence variant data of the Indonesian pre-pandemic 1937 El Tor strain and also the classical 6th and modern 7th pandemic strains to calculate it. This study noted changes in two subclades of *V. cholerae* isolated from Pakistan during 2009–2013 cholera epidemics. The genomic studies profoundly enhance our knowledge about pathogenic clones of *V. cholerae* and the existence of non-pathogenic environmental strains of this bacterial species. In fact, human activities like travel and poor sanitation seem to have been the most crucial factors that led to the spread of these pathogenic clones to Africa, Asia, Latin America, and the Caribbean. These pathogenic clades certainly have and will undergo further evolution, like the variant strains shown in this study. Analysis of the PSC I and PSC II strains that have the maximum number of unique genes reveals that the strain in PSC II with unique genes varies a lot from its normal strains without unique genes, while the strain of PSC I with unique genes does not vary as much when compared to the normal strain of PSC I.

This comparative genomics has revealed ‘waves’ of *V. cholerae* evolution. Its transmission and ability to modify its genetic content in order to survive in different environmental condition supports the concept of the molecular clock, in that, its clones are successively replaced not only over decades and centuries but over a few hours to months. Here, we investigate how the versatile *V. cholerae*, a bacterium that persists across different habitats, is reflected in its genome. The data generated during the study will be extremely beneficial in defining the evolutionary relationships as well as diversity among *V. cholerae* subclades, their epidemiological studies, and accurate treatment strategies for controlling epidemics in the future.

Author contributions

Samia Zeb designed the project, performed laboratory work and wrote the manuscript. Sardar Gulfam helped in conducting experiment and analysis of the results. Habib Bokhari has done overall supervision in all the works.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

The first author would like to acknowledge the Department of Microbiology, Hazara University Mansehra, Pakistan for providing me an opportunity to carry out this research work in COMSATS University Islamabad.

Appendix A. Appendix

Table A1
Description of all genomes with their ID and properties.

Run_Lane_Tag	Original and Submission name	ERS	Year	PSC
6714_5_1	F1DN4	ERS032845	2010	PSC-2
6714_5_2	F2D59	ERS032846	2010	PSC-2
6714_5_4	F4D48	ERS032848	2010	PSC-2
6714_5_5	F5D38	ERS032849	2010	PSC-2
6714_5_7	F7D30	ERS032851	2010	PSC-2
6714_5_8	F8D25	ERS032852	2010	PSC-2
6714_5_11	F11D4	ERS032855	2010	PSC-2
6714_5_12	F12D1	ERS032856	2010	PSC-2

(*continued on next page*)

Table A1 (*continued*)

Run_Lane_Tag	Original and Submission name	ERS	Year	PSC
6714_5_14	F14KPD3	ERS032858	2010	PSC-2
6714_5_15	F15KTH7	ERS032859	2010	PSC-2
6714_5_16	F16KTH6	ERS032860	2010	PSC-2
6714_5_17	F17KTH4	ERS032861	2010	PSC-2
6714_5_18	F18KTH3	ERS032862	2010	PSC-2
6714_5_19	F19KTH2	ERS032863	2010	PSC-2
6714_5_20	S1KCH15	ERS032864	2010	PSC-2
6714_5_21	S2KCH17	ERS032865	2010	PSC-2
6714_5_23	S4KCH16	ERS032867	2010	PSC-2
6714_6_1	S5KCH10	ERS032868	2010	PSC-2
6714_6_2	S6KCH7	ERS032869	2010	PSC-2
6714_6_3	S7KCH20	ERS032870	2010	PSC-2
6714_6_4	S8KCH18	ERS032871	2010	PSC-2
6714_6_5	S9KCH9	ERS032872	2010	PSC-2
6714_6_6	S10P57	ERS032873	2010	PSC-2
6714_6_8	S12P76	ERS032875	2010	PSC-2
6714_6_9	S13P83	ERS032876	2010	PSC-2
6714_6_10	S14P9	ERS032877	2010	PSC-2
6714_6_12	S16HH1	ERS032879	2010	PSC-2
6714_6_13	S17HH3	ERS032880	2010	PSC-2
6714_6_14	S18HH4	ERS032881	2010	PSC-2
6714_6_15	S19HH5	ERS032882	2010	PSC-2
6714_6_16	S20HH14	ERS032883	2010	PSC-2
6714_6_17	S21HH15	ERS032884	2010	PSC-2
6714_6_18	S22HH17	ERS032885	2010	PSC-2
6714_6_19	S23HH18	ERS032886	2010	PSC-2
6714_6_20	S24RG6	ERS032887	2010	PSC-2
6714_6_21	S25R22	ERS032888	2010	PSC-2
6714_6_22	S26R24	ERS032889	2010	PSC-2
6714_6_23	S27RG11	ERS032890	2010	PSC-2
8014_8_40	S1PS7	ERS135741	2011	PSC-1
8014_8_41	S2PS18	ERS136028	2011	PSC-1
8014_8_42	S3PS25	ERS135743	2011	PSC-1
8014_8_43	S4769	ERS135745	2011	PSC-1
8014_8_44	S5N5	ERS135848	2011	PSC-1
8014_8_45	S6N7	ERS135746	2011	PSC-1
8014_8_46	S7N10	ERS126526	2011	PSC-1
8014_8_47	S8NP3	ERS135849	2011	PSC-1
8014_8_48	S9NP5	ERS135749	2011	PSC-2
8014_8_49	S10NP6	ERS135750	2011	PSC-2
8014_8_50	S11NP7	ERS136029	2011	PSC-1
8014_8_51	S12NP14	ERS135752	2011	PSC-1
8014_8_52	S13CS1	ERS135753	2011	PSC-1
8014_8_53	S14CS12	ERS135851	2011	PSC-1
8014_8_54	S15CS15	ERS135755	2011	PSC-1
8014_8_55	S16CS16	ERS135756	2011	PSC-1
8014_8_56	S17CS18	ERS135852	2011	PSC-1
8014_8_57	S18770	ERS135758	2011	PSC-1
8014_8_58	S19751	ERS135759	2011	PSC-1
8014_8_59	S20759	ERS136108	2011	PSC-1
8014_8_60	S21760	ERS135761	2011	PSC-1
8014_8_61	S22754	ERS135762	2011	PSC-1
8014_8_62	S23756	ERS135854	2011	PSC-1
8014_8_63	S24758	ERS135764	2011	PSC-2
8014_8_64	S25763	ERS135765	2011	PSC-2
8014_8_65	S26753	ERS135855	2011	PSC-1
8014_8_66	S27750	ERS135767	2011	PSC-1
8014_8_67	S28703	ERS135768	2011	PSC-1
8014_8_68	S29709	ERS136031	2011	PSC-1
8014_8_69	S30719	ERS135770	2011	PSC-1
8014_8_70	S31722	ERS135771	2011	PSC-1
8014_8_71	S32729	ERS135857	2011	PSC-1
8014_8_72	S33732	ERS135773	2011	PSC-1
8014_8_73	S34736	ERS135774	2011	PSC-1
8014_8_74	S35739	ERS135858	2011	PSC-1
8014_8_75	S36742	ERS135776	2011	PSC-1
8014_8_76	S37F5	ERS135777	2011	PSC-1
8014_8_77	S38F6	ERS136032	2011	PSC-1
8014_8_78	S39764	ERS135779	2011	PSC-1
8014_8_79	S40767	ERS135780	2011	PSC-1
8014_8_82	S43A4	ERS135783	2011	PSC-1
8014_8_83	S44755	ERS135861	2011	PSC-1
8014_8_84	S45757	ERS135784	2011	PSC-1
8014_8_85	S46765	ERS135786	2011	PSC-2

(*continued on next page*)

Table A1 (*continued*)

Run_Lane_Tag	Original and Submission name	ERS	Year	PSC
8014_8_86	S47761	ERS136090	2011	PSC-2
8014_8_87	S48BW5	ERS135788	2011	PSC-2
8014_8_88	S49773	ERS135789	2011	PSC-2
8014_8_89	S50771	ERS135863	2011	PSC-2
8014_8_90	S51772	ERS135791	2011	PSC-2
8014_8_91	S52776	ERS135792	2011	PSC-1
8014_8_92	S53775	ERS135864	2011	PSC-2
8014_8_80	S41768	ERS135860	2011	PSC-2
8014_8_78	S42774	ERS135782	2011	PSC-2
8014_8_94	S55GB39	ERS135795	2011	PSC-2
8014_8_93	S54762	ERS135794	2011	PSC-2
8036_3_3	S60752	ERS135793	2011	PSC-1
8014_8_95	S56BH11	ERS136034	2011	PSC-2
8014_8_96	S57BH20	ERS135797	2011	PSC-2
8036_3_1	S58BHJ	ERS135798	2011	PSC-2
8036_3_2	S59BHA	ERS135866	2011	PSC-2

References

Ali, M., Nelson, A.R., Lopez, A.L., Sack, D.A., 2015. Updated global burden of cholera in endemic countries. PLoS Negl. Trop. Dis. 9, e0003832.

Azarian, T., Ali, A., Johnson, J.A., Mohr, D., Prosperi, M., et al., 2014. Phylodynamic analysis of clinical and environmental *Vibrio cholerae* isolates from Haiti reveals diversification driven by positive selection. MBio. 5 e01824–14.

Berche, P., Poyart, C., Abachin, E., Lelievre, H., Vandepitte, J., Dodin, A., Fournier, J.M., 1994. The novel epidemic strain O139 is closely related to the pandemic strain O1 of *Vibrio cholerae*. J. Inf. Secur. 170 (3), 701–704.

Bhadra, Rupak K., Shah, Sangita, Das, Bhabatosh, 2008. Functional analysis of the essential GTP-binding-protein-coding gene cgtA of Vibrio cholerae. Genetics and Molecular Biology 190 (13), 4764–4771.

Boyd F.(2008). Filamentous phages of *Vibrio cholerae* in: Faruque SM, Nair GB *Vibrio cholerae*: genomics and molecular biology. Horizon Scientific Press, Ltd. UK, :pp 49-66.

Chatterjee, S.N., Chaudhuri, K., 2003. (2003). Lipopolysaccharides of *Vibrio cholerae*. Biochim. Biophys. Acta (BBA) - Mol. Basis Dis. 1639, 65–79.

Chaudhari, Narendrakumar M., Gupta, Vinod Kumar, Dutta, Chitra, 2016. BPGA- an ultra-fast pan-genome analysis pipeline. Scientific Reports 6 Article number: 24373.

Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., et al., 2011. The origin of the Haitian cholera outbreak strain. N. Engl. J. Med. 364, 33–42.

Chun, J., Grim, C.J., Hasan, N.A., Lee, J.H., Choi, S.Y., et al., 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. Proc. Natl. Acad. Sci. U. S. A. 106, 15442–15447.

Dumontier, S., Berche, P., 1998. Vibrio cholerae O22 might be a putative source of exogenous DNA resulting in the emergence of the new strain of vibrio cholerae O139. FEMS Microbiol. Lett. 164, 91–98.

Faruque, S.M., Naser, I.B., Islam, M.J., Faruque, A.S.G., Ghosh, A.N., Nair, G.B., et al., 2005. Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages. Proc. Natl. Acad. Sci. U. S. A. 102, 1702–1707.

Feng, L., Reeves, P.R., Lan, R., Ren, Y., Gao, C., Zhou, Z., et al., 2008. A recalibrated molecular clock and independent origins for the cholera pandemic clones. PLoS ONE 3, e4053.

Harris, J.B., LaRocque, R.C., Qadri, F., Ryan, E.T., Calderwood, S.B., 2012. Cholera. Lancet. 379, 2466–2476.

Hemme, C.L., Green, S.J., Lavanya, R., Om, P., Angelica, P., Romy, C., Deutschbauer, A.M., Van Wu, L., He, Z., 2016. Lateral gene transfer in a heavy metal-contaminated-groundwater microbial community. Mbio 7 e02234–02215.

Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., et al., 2011. Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio. 2 e00157–11.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., ... AlMazroa, M.A., 2012. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. The lancet 380 (9859), 2095–2128.

Medini, D., Donati, C., Tettelin, H., Masignani, V., Rappuoli, R., 2005. The microbial pan-genome. Curr. Opin. Genet. Dev. 15, 589–594.

Mooi, F.R., Bik, E.M., 1997. The evolution of epidemic *Vibrio cholerae* strains. Trends Microbiol. 5, 161–165.

Mutreja, A., Kim, D.W., Thomson, N.R., Connor, T.R., Lee, J.H., et al., 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 477, 462–465.

Pascual, M., Bouma, M.J., Dobson, A.P., 2002. Cholera and climate: revisiting the quantitative evidence. Microbes Infect. 4, 237–245.

Polz, M.F., Alm, E.J., Hanage, W.P., 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 29, 170–175.

Popa, O., Hazkanicovo, E., Landan, G., Martin, W., Dagan, T., 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res. 21, 599–609.

Reidl, J., Klose, K.E., 2002. *Vibrio cholerae* and cholera: out of the water and into the host. FEMS Microbiol. Rev. 26, 125–139. https://doi.org/10.1111/j.1574-6976.2002.tb00605.

Russell, A.J., 1925. A statistical approach to the epidemiology of cholera in Madras presidency. Proc. Natl. Acad. Sci. U. S. A. 11, 653–657.

Seed, K.D., Faruque, S.M., Mekalanos, J.J., Calderwood, S.B., Qadri, F., et al., 2012. Phase variable O antigen biosynthetic genes control expression of the major protective antigen and bacteriophage receptor in *Vibrio cholerae* O1. PLoS Pathog. 8 e1002917–13.

Shah, M.A., Mutreja, A., Thomson, N., ... Wren, B.W., 2014. Genomic Epidemiology of *V. cholerae* O1 Associated with Floods, Pakistan, 2010. Emerging infectious Diseases 20 (1), 13–20.

Stroeher, U.H., Parasivam, G., Dredge, B.K., Manning, P.A., 1997. Novel *Vibrio cholerae* O139 genes involved in lipopolysaccharide biosynthesis. J. Bacteriol. 179, 2740–2747.

Tettelin, H., Riley, D., Cattuto, C., Medini, D., 2008. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11, 472–477.

Waldor, M.K., Mekalanos, J.J., 1994. *Vibrio cholerae* O139 specific gene sequences. Lancet 343, 1366.

Wilson, D.J., 2012. Insights from genomics into bacterial pathogen populations. PLoS Pathog. 8, e1002874.

Yamasaki, S., Garg, S., Nair, G.B., Takeda, Y., 1999. Distribution of *Vibrio cholerae* O1 antigen biosynthesis genes among O139 and other non-O1 serogroups of *Vibrio cholerae*. FEMS Microbiol. Lett. 179, 115–121.

Zeb, S., Ali, A., Gulfam, S.M., Bokhari, H., 2019a. Preliminary work towards finding proteins as potential vaccine candidates for vibrio cholerae pakistani isolates through reverse vaccinology. Medicina 55 (5), 195.

Zeb, S., Shah, M.A., Yasir, M., Awan, H.M., Prommeenate, P., Klanchui, A., Wren, B.W., Thomson, N., Bokhari, H., 2019b. Type iii secretion system confers enhanced virulence in clinical non-o1/non-o139 vibrio cholerae. Microbial Pathogenesis 1036.

Zhang, X., Liu, X., Liang, Y., Guo, X., Xiao, Y., Ma, L., Miao, B., Liu, H., Peng, D., Huang, W., 2017. Adaptive evolution of extreme Acidophile Sulfobacillus thermosulfidooxidans potentially 701 driven by horizontal gene transfer and gene loss. Applied & Environmental 702 Microbiology 83 AEM.03098-03016.