Origins of the current seventh cholera pandemic

Dalong Hu[a,b,1,2], Bin Liu[a,c,1], Lu Feng[a,b,d], Peng Ding[a], Xi Guo[a], Min Wang[a], Boyang Cao[a,b,d], Peter R. Reeves[e,3], and Lei Wang[a,b,d,f,3]

[a]TEDA Institute of Biological Sciences and Biotechnology, Nankai University, Tianjin Economic-Technological Development Area, Tianjin 300457, People's Republic of China; [b]The Key Laboratory of Molecular Microbiology and Technology, Ministry of Education, Tianjin 300457, People's Republic of China; [c]Tianjin Research Center for Functional Genomics and Biochip, Tianjin 300457, People's Republic of China; [d]Tianjin Key Laboratory of Microbial Functional Genomics, Tianjin 300457, People's Republic of China; [e]School of Life and Environmental Sciences, University of Sydney, Sydney, NSW 2006, Australia; and [f]State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin 300071, People's Republic of China

Edited by John J. Mekalanos, Harvard Medical School, Boston, MA, and approved October 11, 2016 (received for review June 2, 2016)

Vibrio cholerae **has caused seven cholera pandemics since 1817, imposing terror on much of the world, but bacterial strains are currently only available for the sixth and seventh pandemics. The El Tor biotype seventh pandemic began in 1961 in Indonesia, but did not originate directly from the classical biotype sixth-pandemic strain. Previous studies focused mainly on the spread of the seventh pandemic after 1970. Here, we analyze in unprecedented detail the origin, evolution, and transition to pandemicity of the seventh-pandemic strain. We used high-resolution comparative genomic analysis of strains collected from 1930 to 1964, covering the evolution from the first available El Tor biotype strain to the start of the seventh pandemic. We define six stages leading to the pandemic strain and reveal all key events. The seventh pandemic originated from a nonpathogenic strain in the Middle East, first observed in 1897. It subsequently underwent explosive diversification, including the spawning of the pandemic lineage. This rapid diversification suggests that, when first observed, the strain had only recently arrived in the Middle East, possibly from the Asian homeland of cholera. The lineage migrated to Makassar, Indonesia, where it gained the important virulence-associated elements** ***Vibrio*** **seventh pandemic island I (VSP-I), VSP-II, and El Tor type cholera toxin prophage by 1954, and it then became pandemic in 1961 after only 12 additional mutations. Our data indicate that specific niches in the Middle East and Makassar were important in generating the pandemic strain by providing gene sources and the driving forces for genetic events.**

Vibrio cholerae | pandemic | evolution | comparative genomics

The bacterium *Vibrio cholerae* is the causative agent of cholera, a severe and potentially life-threatening diarrheal disease that is of considerable public health concern because of its high morbidity and mortality. There have been seven cholera pandemics since 1817, and all continents except Antarctica have had significant or major incursions by one or more of them (1). As early as 1866, an intergovernmental meeting on cholera was held to develop measures for control, making *V. cholerae* one of the first infectious pathogens subject to inspection and quarantine. The current seventh pandemic began in 1961 in Makassar, Sulawesi, Indonesia, and continues to be a major health problem, with an estimated 3 million to 5 million cases of infection every year (2), including recent outbreaks in Haiti and Zimbabwe. The outbreak in Haiti after the 2010 earthquake infected nearly 700,000 people and has caused >8,500 deaths (3). In the 2008 Zimbabwe epidemic, >90,000 suspected cholera cases were reported, with >4,000 of these patients dying (4). Cholera is characterized by an extreme form of watery diarrhea, which causes dehydration that can be lethal. The major virulence factors are the TcpA pilus for attachment of bacteria to the intestinal epithelium and the cholera toxin (CTX), which is released and then enters epithelial cells, where it induces secretion of the water and salts that are the major component of the stools. The disease and virulence factors have been reviewed many times, and we suggest Harris et al. as suitable background for this paper (5).

V. cholerae strains are serogrouped based on their polysaccharide O antigens, and >200 serogroups have been identified to date (6). However, most of the serogroup diversity is in nonpathogenic environmental strains, and all pandemic strains have been from a serogroup O1 clone or a serogroup O139 variant. Serogroup O1 strains are further classified into two biotypes: "classical" (the biotype of the organisms first identified as causing cholera) and "El Tor" (named after the El Tor quarantine station in Egypt, where a strain of this biotype was first isolated in ∼1900). The two biotypes are usually distinguished according to the presence of hemolysin and the presence of the acetoin fermentation pathway, both of which are characteristic of El Tor strains. Only strains from the sixth (1899–1923) and seventh (1961 to present) pandemics have been available for modern scientific study. The sixth pandemic was caused by classical biotype strains, whereas the ongoing seventh pandemic was caused by El Tor biotype strains. Strains from the fifth pandemic were shown to be of the classical biotype, but no live strains remain. Additionally, a second pandemic genome sequence from a preserved human intestine grouped with sixth-pandemic strains in a phylogenetic tree (7), indicating that all but the seventh pandemic were caused by a single lineage of classical biotype. Although the currently available seventh- and sixth-pandemic strains share a common ancestor, we have shown that the seventh-pandemic strain did not originate directly from the

Significance

Cholera, a major disease in human history, has terrorized the world through seven pandemics. The seventh pandemic started in Indonesia in 1961 and spread globally, currently infecting 3–5 million people annually. By combining all available historical records and genomic analysis of available preseventh pandemic and some early pandemic strains, we revealed the complex six-step evolution of the pandemic strain from its probable origin in South Asia to its nonpathogenic form in the Middle East in ∼1900 to Indonesia in ∼1925, where it evolved into a pandemic strain before becoming widespread in 1961. This pathway relates to human traffic routes, including the annual Hajj pilgrimage, and involved novel niches that provided gene sources and the driving forces for stepwise evolution.

Author contributions: L.F., P.R.R., and L.W. designed research; D.H., B.L., P.D., X.G., M.W., and B.C. performed research; D.H., B.L., L.F., P.R.R., and L.W. analyzed data; and B.L., P.R.R., and L.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. CP013301–CP013320).

[1]D.H. and B.L. contributed equally to this work.

[2]Present address: School of Life and Environmental Sciences, Charles Perkins Centre, D17, University of Sydney, Sydney, NSW 2006, Australia.

[3]To whom correspondence may be addressed. Email: wanglei@nankai.edu.cn or peter.reeves@sydney.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1608732113/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1608732113

sixth-pandemic strain (8). Note that we are using the word strain for two generally accepted meanings: (*i*) the progeny of a single cell, usually first isolated as a colony (examples being “strain N16961” and “seventh pandemic strains”) and (*ii*) a lineage that occurs in nature (examples being “seventh pandemic strain” and “sixth pandemic strain”). The meaning can be deduced from the context.

Long before the seventh pandemic started in 1961, El Tor biotype strains from two small areas had been reported, commonly referred to as Middle East and Makassar strains and collectively as preseventh pandemic (or prepandemic) strains (9). The Middle East strains, which were isolated from Mecca, Saudi Arabia, and adjacent areas extending into Egypt and Iraq, were observed and studied from 1897 to 1938 (10–14) (see *SI Text* for details and sources). These strains were found in human intestines, but they were nonpathogenic because they did not cause disease or any symptoms and therefore were not subject to quarantine. It is interesting that in the period spanning from 1897 to 1938, as reported by Abdoelrachman (10), these strains were commonly found in human intestines, but no cases of cholera were reported because people found carrying these strains were not held in quarantine stations (*SI Text*). The Middle East strains that were used in the only paper published at that time (15) are no longer available (16), but some later Middle East strains remain and are included in this study. There were four cholera outbreaks caused by El Tor strains between 1937 and 1957 within ∼150 km of the town of Makassar in the island of Sulawesi, Indonesia. The disease caused by these Makassar strains resembled true cholera in severity and mortality, but lacked the capacity to spread effectively (17, 18) and was therefore distinguished from the then active sixth pandemic disease through the designation “paracholera” (17–19). This distinction was critical because individuals with cholera were subjected to quarantine, whereas those with paracholera were not (20).

A multilocus sequence typing (MLST) analysis using 26 housekeeping genes (9), including two Middle East strains [National Collection of Type Cultures 9420 (NCTC9420) and NCTC5395] and one Makassar strain (M66-2), and a genomics analysis involving one Makassar strain (M66-2) (8) both showed that those prepandemic strains branch from the lineage of El Tor biotype seventh-pandemic strains and therefore represent important precursors of the seventh-pandemic strain. In 1960, an outbreak of cholera caused by an El Tor strain occurred in Makassar (18) and spread overseas in 1961 (21), indicating that the El Tor strain had evolved to pandemic spread capacity and making 1961 the recognized start of the seventh cholera pandemic. In 1962, the World Health Organization (WHO) decided that, from then on, disease caused by El Tor strains should be considered cholera, and El Tor strains were subjected to the same controls as the sixth-pandemic strains (21). The seventh pandemic continues to this day, and, interestingly, there are also occasional outbreaks caused by prepandemic-related strains, which are more closely related to prepandemic strains than to seventh-pandemic strains in MLST analysis (9). The best-known examples of such outbreaks are the 1970s and 1980s outbreaks in the US Gulf of Carpentaria and Australia.

The phylogeny and epidemiology of the seventh pandemic have been exhaustively investigated, and a recent study that used genomic sequence analysis of >190 seventh-pandemic strains showed that the seventh pandemic is monophyletic and has spread around the world in at least three independent waves (22). The patterns of change in individual genes in the pandemic period have also been reviewed and show that change is ongoing, perhaps in relation to human immune responses (23).

Although there have been seven cholera pandemics since 1817, only seventh-pandemic strains and a few sixth-pandemic and pre-seventh-pandemic strains are currently available. The available pre-seventh-pandemic strains from the Middle East and Makassar, as well as the currently available seventh-pandemic strains, together provide an excellent opportunity for understanding how the seventh-pandemic strain originated. However, studies focused on these strains are currently rather limited. We previously used a complete genome sequence of one of them (M66-2) to identify some of the events that occurred on the pathway to the seventh-pandemic strain (8). However, for the six other available prepandemic strains, only three draft genome sequences (for A6, MAK757, and NCTC8457) have been reported. In addition, studies on the seventh pandemic itself have focused on strains isolated after 1970 (22, 24), and there were no full genome sequences of seventh-pandemic strains isolated in the first 10 y of the pandemic. Thus, the roles played by the Middle East and Makassar strains in the evolution of the seventh-pandemic strain remain unknown, as well as the many evolutionary events that occurred over the 60-y period from the first observations of El Tor strains in the Middle East through the prepandemic outbreaks in Makassar to the start of the seventh pandemic. The genetic and environmental factors that promoted the generation of the seventh-pandemic strain and the evolutionary relationships among those prepandemic strains in the Middle East and Makassar also remain unclear.

To clarify the origins of the seventh-pandemic strain and the transition from the sixth to the seventh pandemic, we obtained complete genome sequences of 10 strains, including 3 prepandemic strains and 4 very early seventh-pandemic strains, using PacBio sequencing technologies. We also included three of the prepandemic-related strains, because they appear to have diverged independently from the seventh pandemic path during the prepandemic period (9) and the details of their relationships with other prepandemic strains could be very revealing. We constructed a high-resolution phylogenic tree of the seventh pandemic lineage based mostly on full genome sequences and were able to allocate mutations, recombination events, and indels to specific branches of the tree to create a near-complete picture of the genetic changes that have occurred on each branch.

This phylogenic tree and historical records of cholera epidemiology and related social history allowed us to define six evolutionary stages leading to the seventh-pandemic strain from a nonpathogenic strain over the 60-y period. They showed that the nonpathogenic prepandemic Middle Eastern El Tor strains underwent an explosive expansion starting in or before the 1890s, and this expansion spawned the lineage that evolved to become the seventh-pandemic strain. This rapid diversification suggests that the El Tor strain had not been in the region long —it was possibly carried there from the Asian homeland of pandemic cholera by pilgrims. The main lineage gained the El Tor form of the *tcpA* gene and the classical type CTX (CTX$^{\mathrm{Cla}}$) prophage in the Middle East and became pathogenic in ∼1908. The lineage then migrated from the Middle East to Makassar. The Makassar period was a critical stage for the main lineage, leading to the development of the pandemic strain, because between 1925 and 1954, it acquired the *Vibrio* seventh pandemic I (VSP-I) and VSP-II islands, associated with the seventh pandemic, and replaced the CTX$^{\mathrm{Cla}}$ prophage with the El Tor type CTX (CTX$^{\mathrm{ET}}$) prophage, which can be distinguished from the CTX$^{\mathrm{Cla}}$ prophage primarily by the genotypes of the *ctxB* and *rstR* genes within the prophage (19). These elements were essential for the transition from the prepandemic strain to the seventh-pandemic strain, because they are absent in all prepandemic-related strains from the United States, Australia, and China, none of which have developed the ability to cause pandemic cholera. After gaining only 12 mutations within its last branch in Makassar, estimated to cover the period from 1954 to 1960, the main lineage exhibited high spread capability and erupted as the seventh-pandemic strain in 1961. We also found a high recombination frequency in the prepandemic period that declined to become negligible as pathogenicity developed, which is probably due to changes of niche while following the Asian homeland–Middle East–Makassar migration pathway of the lineage.

Results and Discussion

Choice of Strains and Genome Sequences. Genomic analysis of the prepandemic and early seventh-pandemic strains is essential for understanding the origins of the seventh pandemic and identifying key genetic events involved in the transition from prepandemic to seventh-pandemic strains in this critical period. Seven prepandemic strains are available, but there are only four genome sequences, of which only one is a complete sequence, with the others being draft genome sequences (Table S1). In this study, we obtained the complete genome sequences of an additional three prepandemic strains (two Middle East and one Makassar strains), plus four early (1961–1964) seventh-pandemic strains, which are the only complete genomes available for seventh-pandemic strains isolated during the first decade of the pandemic (1961–1970) (Table S1). We also obtained complete genome sequences of three prepandemic-related strains, isolated from the US Gulf (in 1974), Australia (in 1977), and China (in 1974) (Table S1). These prepandemic-related strains can cause cholera and have even been responsible for small outbreaks, but they did not further evolve to become pandemic. However, inclusion of these prepandemic-related strains facilitated the identification of genetic events that are critical for the pathogenicity and pandemicity of the seventh-pandemic strain, as well as the allocation of those events to specific branches on the phylogenetic tree (*Stage 3* and *Stage 5*). The final tree enabled genetic events on the pathway from the first identified El Tor strains to the outbreak of the seventh pandemic to be allocated to six stages, giving the order for the critical events that occurred in the evolution of the seventh pandemic. The additional strains also improved estimates of the time frame for these stages in the transition from the nonpathogenic form to the seventh-pandemic strain (see below).

Identification of Mutations, Recombination Events, and Indels and Allocation to Branches. Because there are so few remaining prepandemic, prepandemic-related, and early seventh-pandemic strains, we used various strategies to obtain as much information as possible from the sequence data. In particular, we optimized attribution of each single nucleotide polymorphism (SNP) to either a mutation or a recombination event (*Materials and Methods*) and used only mutational SNPs to construct an accurate phylogenetic tree; we also used these SNPs to estimate the timeframe for each evolutionary stage.

For phylogenetic analysis, in addition to the 10 complete genomes obtained in this study, we included 11 published genome sequences, comprising 5 of the 6 mostly draft genome sequences of prepandemic and prepandemic-related strains, as well as complete sequences of 6 seventh-pandemic strains isolated from 1975 to 2004 to represent the three waves of the seventh pandemic (Table S1). We did not include the sequence of strain A6, because it appears to be the same as strain C5, but under a different name. Note that details of the isolation and storage of strain C5 are given by Teppema et al. (25).

We used the 3,670,626 bp present in the 21 genomes for primary analysis and found 238 recombination events, all with donors from outside of the seventh pandemic lineage (Dataset S1). A total of 999 SNPs were attributed to mutations (Dataset S1), and these were used to generate a high-resolution phylogenetic tree (Fig. 1 and Fig. S1) that was rooted by using the sixth-pandemic strain O395. Of the 999 mutational SNPs, 958 (95.9%) are congruent with the tree, indicating the high robustness of the tree. The mutation and recombination events were then allocated to specific branches of the tree, as shown in Fig. 1 and Fig. S1. We then identified indels by examining the sequences for segments not shared by all genomes and found 102 major indels (>500 bp) (Dataset S1). To extend the data for the branches leading to the seventh-pandemic strain, we selected eight full genome sequences for separate analysis, covering the 3,794,564 bp shared by the eight strains, an increase of 123,938 bp in data for these branches. The indels present in the seven ribosomal RNA loci, the CTX region on the large chromosome, and the large integron were sometimes complex (8), and we were not able to define these as discrete events, so they were not generally included at this stage. These comprised most of the sequences still not included in the analysis, and we examined the individual genomes to complete the identification of all mutational, recombination, and indel events on key branches 2, 6, 10, 18, 19, and 22, which cover the transition of the prepandemic strain into the seventh-pandemic strain (Fig. 2 and Fig. S2). For further details, see Materials and Methods. Another major advantage of the complete genomes obtained in this study is that the mutational SNPs are reliable, even in terminal branches, in which each SNP has been observed only in that strain. When using the draft genome sequences with low coverage, the large number of SNPs because of sequencing error can seriously distort the mutation number in the terminal branch leading to that genome, and the five branches affected are shown by the dotted lines in Fig. 1.

We used the mutational SNPs and the BEAST program (26) to estimate the dates for branching at the nodes. For this estimation, we used only data that were suitable for sequence quality assessment, and on this basis we excluded strains MAK757, NCTC8457, BX330286, and 2740-80. The data from strain NCTC9420 were also excluded, because there is a frameshift mutation in the mutator gene *mutS*. The tree as first generated (Fig. S3) had E9120, the first seventh-pandemic strain isolated in 1961, diverging from the ongoing seventh pandemic lineage in 1955, which is very similar to the date obtained by Mutreja et al. (22). However, this date is not consistent with the historical record (see below). We therefore constrained the tree to reflect historical observations by imposing the condition that divergence of strain E9120 occurred in 1960 (see discussion in *Stage 6*), and the dates for nodes directly related to that event were revised accordingly (Fig. 1 and Fig. S1).

Six Stages in the Formation of the Seventh Pandemic Strain. The phylogenetic tree clearly distinguishes six stages in the evolution of the seventh-pandemic strain (Figs. 2 and 3). Most branches on the tree can be allocated to geographic regions based on the isolation locations shown in Fig. 1, and a geographic representation of the data is shown in Fig. 3. The phylogenetic analysis allocated genetic events to specific branches of the tree, based on the distribution among the strains of mutations, recombinant DNA, and genes gained or lost. Some genetic events thought to be important in the evolution of the seventh pandemic strain are shown above the bar at the top of Fig. 2.

Stage 1. Stage 1 (branch 0) covers the long period from the divergence of the sixth- and seventh-pandemic strains, which probably occurred in South and East Asia (*Stage 2*), through the first observation of the El Tor strain in the Middle East in 1897 (11) to ~1902. Because no strains are available, we have no direct knowledge of the lineage over this period.

Stage 2. Stage 2 (branch 2) covers a very short period (1902–1903) in the Middle East, during which nonpathogenic Middle East strains underwent rapid diversification and gained the El Tor form of the *tcpA* gene.

The clustering of the Middle East strains at the base of the tree (Figs. 1 and 2 and Fig. S1) is in agreement with results of previous studies (9, 22, 24), and the very short length of branch 2 makes it virtually certain that the strains along this branch were in the Middle East, in agreement with the location for the first observations of the El Tor strain. A large number of pilgrims travel to the Middle East during the annual Haj in Mecca, and there is a significant possibility that cholera carried by these pilgrims to Mecca could spread into neighboring countries (*SI Text*). Thus, government facilities were established to monitor the presence of cholera in people entering Egypt or Iraq from Saudi Arabia after the Haj in Mecca, including at El Tor on the Red Sea for travel to Egypt (27). The Middle East strains used in this study were isolated by these facilities.

During stage 2, there was explosive diversification, as shown by the complete absence of SNPs on branch 2, on which the lineage to strain NCTC9420, isolated in Cairo (Egypt), diverged from the lineage to strains isolated in El Tor (Egypt) (NCTC8457) and Baghdad (Iraq) (NCTC5395). This divergence occurred over an estimated 50-y period. If these strains had evolved from a population that had been resident in the region for many years, the chance of observing a branch with no SNPs is negligible. This finding indicates that their most recent common ancestor (MRCA) in ∼1902, and thus a representative of oldest known precursor of the seventh-pandemic strain, is not from a long-established form, but rather is a new strain that had arrived in the Middle East from elsewhere or, alternatively, had undergone a genetic change that enabled it to replace preexisting forms in a selective sweep (28). We suggest that the former is more likely, because Mecca was recognized in the 1850s as a source of cholera, leading to the establishment of quarantine and monitoring of facilities in the region. Another reason is that the El Tor strain is clearly related to the classical biotype strains of the sixth-pandemic (1899–1923) and earlier pandemic strains (29), which are thought to have arisen in South Asia (29). It seems most likely that the major divergence of the El Tor and classical biotype strains occurred in South Asia, before the El Tor strain was carried to Mecca, perhaps by pilgrims, and having become established there, was then carried to El Tor and other areas in the Middle East by pilgrims.

The date of the MRCA of the three Middle East strains (NCTC9420, NCTC8457, and NCTC5395) was estimated to be 1902, which is consistent with El Tor strains being first observed in the Middle East in 1897 (11). There is very strong evidence that the three strains used in this study were nonpathogenic (*SI Text*). Thus, we suggest that they are representatives of the nonpathogenic El Tor strain first reported by Gotschlich in 1906 (15).

The major event occurring in stage 2 was the recombination-mediated replacement of the *tcpA* gene in the *Vibrio* pathogenicity island, gaining what is known as the El Tor form of TcpA, which forms the pilus that attaches pandemic *V. cholerae* organisms to the small intestine (30) and is also the receptor for infection of the bacterium by a CTX phage (31). There is considerable variation in *tcpA* genes, which were divided into 12 clusters in a recent paper (32). Based on the phylogenetic relationships described in that paper, the *tcpA* genes in Middle East strains NCTC8457 and NCTC5395 belong to cluster 9, together with those from the serogroup O115 and O141 outbreak strains. However, the El Tor form of TcpA, present in the Middle East strain NCTC9420 and all other prepandemic, prepandemic-related, and also early seventh-pandemic strains (33), belongs to cluster 4 (the sixth pandemic *tcpA* gene is in cluster 1). The O141 outbreak strain has been reported globally, but it is not closely related to the lineage harboring the sixth- and seventh-pandemic

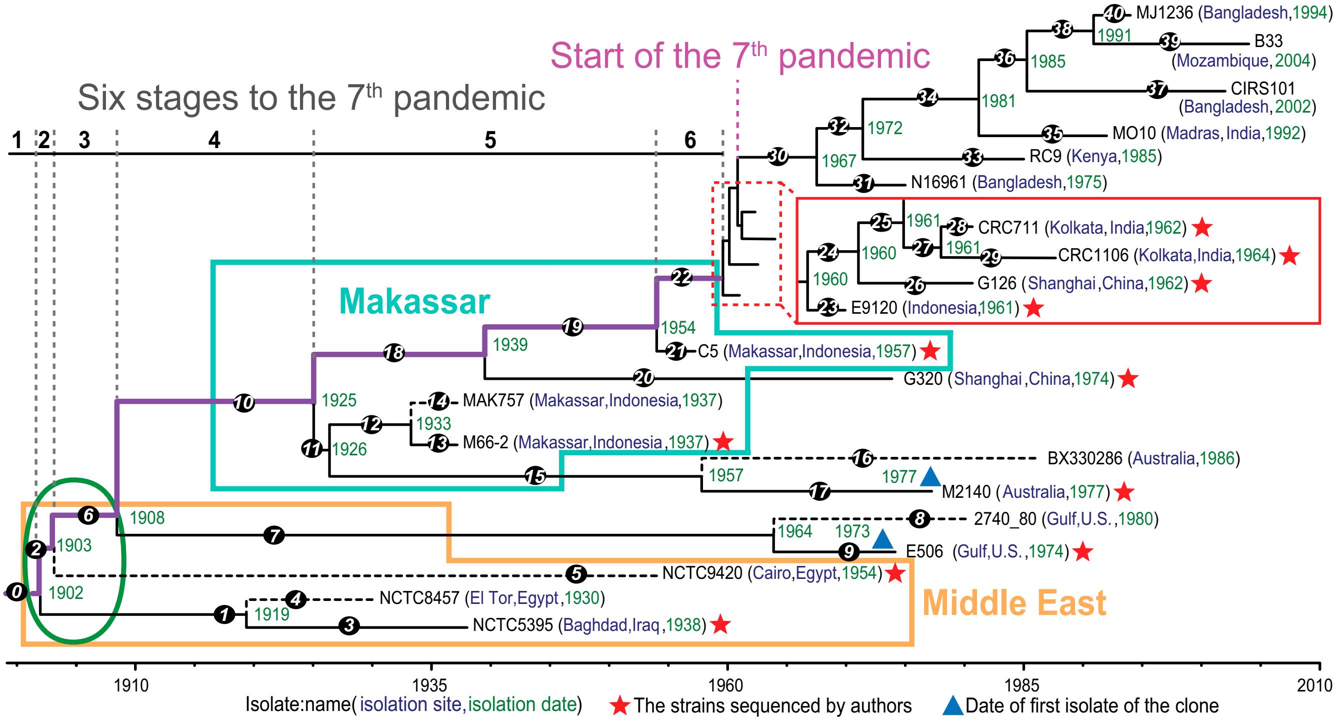

Fig. 1. The maximum-likelihood phylogenetic tree of 21 *V. cholerae* strains based on mutational differences after excluding recombination events. This is a traditional tree that was rooted using the sixth-pandemic strain O395. The branches corresponding to the six stages in the evolution of the seventh pandemic are colored in violet. The branching data were estimated by BEAST (median value) and are shown at the nodes. The lengths of the branches correspond to the times between divergence from MRCA and isolation. Genetic events (mutation, insertion, deletion, and recombination) were allocated to and shown on specific branches. The MAK757, NCTC8457, BX330286, and 2740-80 branches are shown by dotted lines and were not included in BEAST for analysis because of the lack of suitable-quality sequence data, and the dates for their divergence from MRCA were predicted based on the SNP allocations of strains M66-2, NCTC5395, M2140, and E506, respectively. The strain NCTC9420 carried a mutator gene, and there is no SNP on branch 0, which is the outermost branch of the tree; therefore, both of these branches are also shown by dotted lines. The branches representing the evolutionary stages that occurred in the Middle East and Makassar are enclosed in boxes. See Fig. S1 for a tree with genetic events allocated to branches.

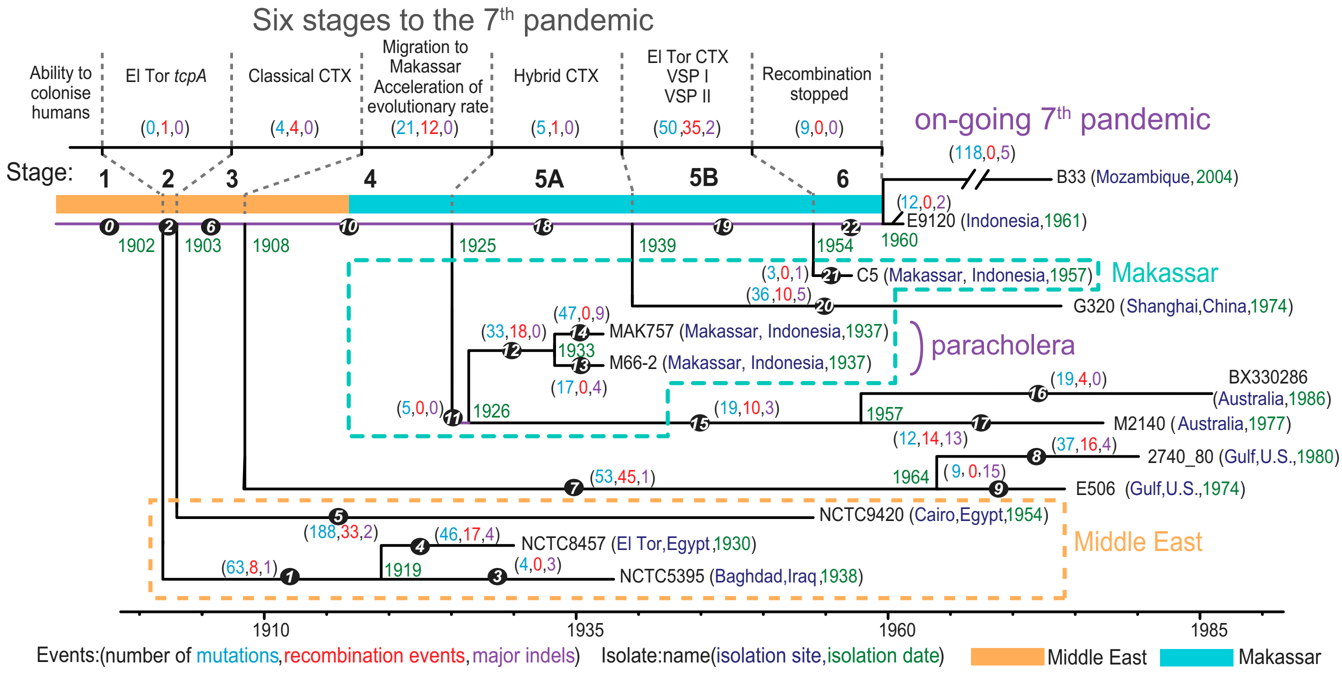
Six stages to the 7th pandemic
Ability to colonise humans
El Tor tcpA
Classical CTX
Migration to Makassar Acceleration of evolutionary rate
Hybrid CTX
El Tor CTX VSP I VSP II
Recombination stopped
(0,1,0)
(4,4,0)
(21,12,0)
(5,1,0)
(50,35,2)
(9,0,0)
on-going 7th pandemic
(118,0,5)
Stage: 1 2 3 4 5A 5B 6
0 2 6 10 18 19 22
1902 1903 1908 1925 1939 1954 1960
B33 (Mozambique,2004)
(12,0,2)
E9120 (Indonesia,1961)
(3,0,1) 21 C5 (Makassar, Indonesia,1957)
Makassar
(36,10,5) 20 G320 (Shanghai,China,1974)
(47,0,9) 14 MAK757 (Makassar, Indonesia,1937)
(33,18,0) 12 1933
13 M66-2 (Makassar, Indonesia,1937)
(17,0,4)
paracholera
(5,0,0) 11 1926
(19,4,0) 16 BX330286 (Australia,1986)
15 (19,10,3) 1957
17 M2140 (Australia,1977)
(12,14,13)
(37,16,4) 8 2740_80 (Gulf,U.S.,1980)
7 (53,45,1) 1964 (9,0,15) 9 E506 (Gulf,U.S.,1974)
5 (188,33,2) NCTC9420 (Cairo,Egypt,1954)
4 (46,17,4) NCTC8457 (El Tor,Egypt,1930)
Middle East
(63,8,1) 1 1919 3 (4,0,3) NCTC5395 (Baghdad,Iraq,1938)
1910 1935 1960 1985
Events:(number of mutations,recombination events,major indels)
Isolate:name(isolation site,isolation date)
Middle East
Makassar

Fig. 2. Details of the six stages leading to the seventh-pandemic strain. The stages on the path to the seventh-pandemic strain are shown in the horizontal bar, which comprises the branches shown in violet in Fig. 1. The branches to the prepandemic and prepandemic-related strains that demarcate the stages are shown below the bar, with details of genetic events given on each branch.

strains (34). Although there are still no data on the colonization differences related to differences in TcpA structure, we suggest that this change between the two different forms of the *tcpA* gene in stage 2 may well have been critical for the development of pathogenicity in the seventh-pandemic strain, although this gene-replacement event was clearly not needed for human colonization because all strains were isolated from humans.

We also suggest that the cluster 9 TcpA pilus plays a role in the longer period of colonization of the major Middle East strain. The incubation period for seventh-pandemic strains in humans can range from several hours to 5 d, and excretion of organisms may continue for 1–2 wk. A very small minority may continue to excrete the organism for longer periods of time (35). The rapid termination of infection is presumably due to the strong immune response observed (35). However, the Middle East strains that were studied in government facilities were detectable in human stools for ∼5 wk after the pilgrims left Mecca (11). It is probable that these strains interact with the human host very differently than pandemic strains, because they do not cause the characteristic watery diarrhea, and their colonization lasted much longer than the normal 2-wk maximum. The details of the colonization process that occurs through attachment to

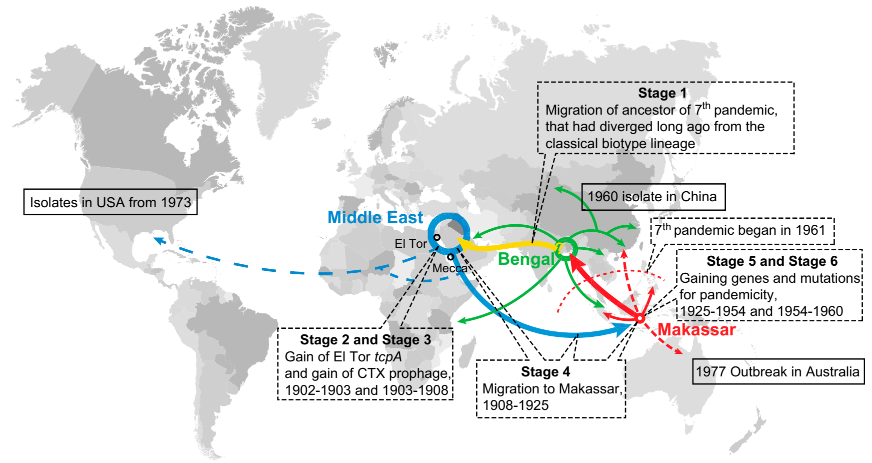
Stage 1
Migration of ancestor of 7th pandemic, that had diverged long ago from the classical biotype lineage
Isolates in USA from 1973
1960 isolate in China
Middle East
7th pandemic began in 1961
El Tor
Bengal
Mecca
Stage 5 and Stage 6
Gaining genes and mutations for pandemicity, 1925-1954 and 1954-1960
Makassar
Stage 2 and Stage 3
Gain of El Tor tcpA and gain of CTX prophage, 1902-1903 and 1903-1908
Stage 4
Migration to Makassar, 1908-1925
1977 Outbreak in Australia

Fig. 3. Migration of prepandemic strains and the locations and stages in the evolution of the seventh-pandemic strain. Three important locations in the origin and history of the seventh-pandemic strain are circled. The yellow line indicates the migration of the ancestral form; blue lines indicate migrations from the Middle East; red lines indicate migrations from Makassar; green lines indicate the early spread of the seventh pandemic from Bengal; dotted lines indicate the pathways of prepandemic-related strains to the US Gulf, Australia, and China. The six stages in the evolution of the seventh-pandemic strain are shown in the labels. The date ranges given for transmission events, which are taken from BEAST analysis, are also shown, as well as the locations of the prepandemic-related strains.

epithelial cells have been better studied in other species, in particular in Enteropathogencic *Escherichia coli* (EPEC) and the *Citrobacter rodentium* mouse model of human EPEC, in which colonization normally lasts 2–3 wk, during which the organisms are attached to the surface of epithelial cells in the colon (36). *C. rodentium* resembles *V. cholerae* (37) in its ability to elicit a strong immune response during infection, which leads to clearance of the infection (reviewed for *C. rodentium* in ref. 38). Earlier work on *C. rodentium* (39) showed that, in otherwise germ-free mice, the organisms can colonize for up to 6 wk (the longest period tested). However, colonization of the epithelial cell surface was limited to ∼2 wk, which is attributed to the development of the immune response (38). Long-term survival of the bacteria was found only in the lumen and only occurred in otherwise germ-free mice, showing that *C. rodentium* cannot compete with normally commensal organisms (38). The 5-wk-long persistence of the Middle East *V. cholerae* strain in the intestine (see above) suggests that this strain also occupies the intestinal lumen, rather than attaching to epithelial cell surfaces, as is typically the case during infection with *V. cholerae*. The implication is that the Middle East *V. cholerae* strain is competitive in the intestinal lumen, which would keep it clear of the area normally kept free of colonization by the immune system. It is clearly possible that these Middle East strains carried the cluster 9 TcpA pilus as the TcpA pili in the sixth- and seventh-pandemic strains are responsible for attachment to epithelial cells (31), and this difference may well be important for long-term survival of the Middle East strains. It may well be that the strain gained the ability to colonize the human intestine in such a commensal manner shortly before it was first detected in 1897 (*SI Text*), representing the first stage of the 60-y journey to the seventh pandemic. However, it should also be noted that long-term carriers of pandemic *V. cholerae* strains have been reported, although they are extremely rare (35).

Stage 3. Stage 3 (branch 6) is attributed to the 5-y-period spanning from 1903 to 1908 in the Middle East, during which the CTX^{Cla} prophage was obtained, and the lineage probably became pathogenic, spawning the seventh-pandemic lineage, although no pathogenic strain was recorded at that time in the Middle East. There are four mutations and four recombination events on branch 6, continuing the rapid diversification in the Middle East, by taking only 5 y to change from a nonpathogenic but human-associated form, to the common ancestor of the main lineage, which is next observed in Makassar, and the lineage to the US Gulf prepandemic-related strain. These two lineages are both pathogenic (Fig. 1) and carry the same CTX^{Cla} prophage. This CTX^{Cla} prophage is also present in the Australian prepandemic-related strain, which diverged from the Makassar 1937 outbreak strain. The CTX is responsible for the very watery diarrhea that is the hallmark of cholera, and we conclude that the prepandemic lineage gained the CTX^{Cla} prophage and also became pathogenic in stage 3. It is interesting that including sequences of US Gulf prepandemic-related strains in the analysis enabled the gain of the CTX^{Cla} prophage to be confined to a short period estimated to span from 1903 to 1908 (Fig. 1). It should be noted that, after stage 3, almost all prepandemic and prepandemic-related strains possess a CTX^{Cla} or CTX^{ET} prophage (Fig. S4), with the exceptions being strains 2740-80 and M66-2, which must have lost it (8, 40). The CTX^{Cla} prophage found in the prepandemic strains is very similar to that found in the sixth-pandemic strains, but its source within the Middle East is not known.

Stage 4. Stage 4 (branch 10) covers the migration of pathogenic prepandemic strains from the Middle East to Makassar and ends with the divergence of the lineage causing the 1937 Makassar outbreak. This branch is estimated to cover the period from 1908 to 1925.

The entry into Makassar is arbitrarily located midway along this branch (Figs. 1 and 2). There are 21 mutations and 12 recombination events allocated to the branch, each of which could have occurred in either the Middle East or Makassar. However, there are no indels recorded, and none of the mutations or recombination events involved genes known to be associated with virulence. It should be noted that the divergence of the 1937 Makassar outbreak strains at the end of branch 10 in ∼1925 is firmly located in Makassar, because the subsequent branches 11 and 18 both lead to strains isolated from Makassar outbreaks. Conversely, the 1908 node at the start of branch 10 is located in the Middle East based on the probabilities, and migration to Makassar is arbitrarily placed in the middle of branch 10 in Figs. 1 and 2.

Stage 5. Stage 5 (branches 18 and 19) covers organisms in Makassar on the direct path to the seventh-pandemic strain and runs from the 1925 divergence of the 1937 Makassar strain (and Australian prepandemic-related strain) to the 1954 divergence of the 1957 Makassar outbreak strain from the lineage to the seventh-pandemic strain (Fig. 1). During this period in Makassar (1925–1954), the prepandemic strains underwent substantial genetic changes, including 74 mutations and 36 recombination events, which replaced 114.76 kb of the genome. The lineage also gained two large insertions, the VSP-I and -II islands (41), which have been used as markers for the seventh pandemic. The nucleotide cyclase gene (*dncV*) in VSP-I was shown to be required for efficient intestinal colonization of the seventh-pandemic strain (42), and VSP-II may also contribute to pathogenicity or pandemicity, although currently there is no evidence supporting this possibility. The boundary between stages 4 and 5 is demarcated by the divergence of the Australian prepandemic-related strains, whereas the divergence of the Chinese prepandemic-related strain G320 enabled stage 5 to be divided into two segments, branches 18 and 19, spanning ∼14 and 5 y, respectively (Fig. 1). All strains after branch 18 possess *ctxB*^{ET}, and the replacement of *ctxB*^{Cla} with *ctxB*^{ET} is allocated to branch 18 (Figs. 1 and 2). Furthermore, because all strains after branch 19 possess VSP-I, -II, and *rstR*^{ET}, we allocated the gain of VSP-I and -II and the replacement of *rstR*^{Cla} with *rstR*^{ET} to branch 19. It is interesting that G320, which diverges after branch 18, has its *ctx* genes in the small chromosome only, with *ctxB*^{ET} and *rstR*^{Cla}. It could be that the common ancestor at the end of branch 18 had both the small chromosome *ctx* genes of G320 and the chromosomal *ctx* locus with its *ctxB*^{Cla} replaced with *ctxB*^{ET}. This copy could have been derived from the small chromosome through recombination and retained in G320, whereas C5 and all subsequent strains retained the copy at the chromosomal locus. Further details of variation in the CTX prophages are shown in Fig. S4. Presumably, the prepandemic lineage acquired these essential genetic elements from other *V. cholerae* strains in the Makassar region.

We suggest that the incorporation of VSP-I, -II, and CTX^{ET} played an essential role in the transition of a prepandemic strain into the seventh-pandemic strain. The reason is that these elements are absent from prepandemic-related strains isolated from the US Gulf, Australia, and China, all of which are able to cause disease, but none having the ability to spread like the seventh-pandemic strain. In the case of the US Gulf and Australian prepandemic-related strains, an alternative explanation is that the inability to spread rapidly was related to the quality of the water supply and the sewage infrastructure in those countries. However, this explanation does not apply to strain G320, because pandemic cholera persisted in China until ∼2000 (24), but prepandemic-related organisms were not involved in the pandemic (24). This finding suggests that the absence of these virulence-associated elements was an obstacle for those prepandemic-related strains in evolving pandemic capability.

Stage 6. Stage 6 (branch 22) occurred during a short period (1954–1960) in Makassar, covering the transition from the 1954 MRCA with the prepandemic strain (C5) isolated during the last Makassar outbreak, to the divergence of the short branch, to the first isolated seventh-pandemic strain (E9120). The reports from those who worked on cholera at the start of the seventh pandemic showed that, from 1961, the seventh-pandemic strains exhibited a much higher spread capability than was observed for the prepandemic strains that caused the earlier outbreaks in Makassar (17, 18). During this stage, the strain gained the high spreading capability that distinguishes paracholera from cholera, thereby becoming the seventh-pandemic strain. After El Tor disease spread overseas from Makassar, in 1962, the WHO recognized that it should be treated as cholera and be subject to quarantine (20).

There are only 12 mutational SNPs on branch 22, which must be responsible for the transition to high spread capability, because there are no other changes (Fig. S2 and Datasets S1 and S2). However, none of these mutations have an obvious connection to pathogenicity or pandemicity on the basis of current knowledge (Datasets S1 and S2), and which of these mutations is responsible for the change in epidemiology requires additional study.

The BEAST estimate for the divergence of strain E9120 from other early seventh-pandemic strains (G126, CRC711, and CRC1106) was 1955 before we constrained the date to fit the historical data (see above). The high mutation rate now inferred for the first few years of the seventh pandemic (branches 23, 24, and 27, 1960–1964; Figs. S1 and S5) may have involved further adaptation to pandemicity. An alternative explanation is that strain E9120 and the lineage that migrated worldwide actually did diverge in ∼1956, but neither came to attention before 1961. The second alternative seems highly unlikely in light of contemporary evidence that the 1960 outbreaks in Sulawesi and other areas of Indonesia (18) were the first stages of the spread of what became the seventh pandemic, reaching several countries in Asia by the end of 1961 (18, 20, 21).

Significance of the Asian Homeland–Middle East–Makassar Migration Pathway in the Generation of the Seventh-Pandemic Strain. We suggest that specific environments and human social structures in the Middle East and Makassar could have driven the evolution of the pandemic. The other *V. cholerae* strains in these regions must have served as important sources of the acquired genes that affected pathogenesis and pandemicity. The *tcpA* gene and the CTX^{Cla} prophage acquired in the Middle East (stages 2 and 3), as well as the VSP-I and -II genetic islands and the CTX^{ET} prophage acquired in Makassar (stage 5), are closely related to corresponding elements found in the sixth-pandemic strain. The sources of the Middle East *V. cholerae* strains are close to Jerusalem and Mecca, which are religious centers for several religions, including Judaism, Christianity, and Islam, and have therefore attracted pilgrims and other visitors from around the world for centuries. Additionally, Makassar was a major center for international shipping during the period discussed here, which was before airplanes replaced ships as the major vehicles for long-distance travel. The close relationship of the virulence genes gained by the pandemic lineage in the Middle East and Makassar regions to genes found in the sixth-pandemic strain suggests that donor strains may have been carried by pilgrims from South Asia, often referred to as the Asian homeland of cholera, with its major focus in Bangladesh and the adjacent Indian State of Bengal, which together cover the floodplains of the lower reaches of the Ganges and Brahmaputra rivers. Thus, we suggest that the pathway followed by the seventh pandemic lineage from Bengal to the Middle East and then Makassar is following a migration pathway through centers that had seen cholera before. We propose that this pathway was a key factor in the formation of the seventh-pandemic strain because of the continuing existence of *V. cholerae s*trains harboring genes from prior pandemics. The gain of genes from other *V. cholerae* strains as described here is probably part of an ongoing process, because similar events have been reported for recent changes in the seventh-pandemic strains (23).

Recombination Is Rare in Outbreak and Pandemic Forms, but Is Otherwise Common. We found that recombination frequency was generally high while the prepandemic lineages were present in the Middle East and Makassar. On the direct evolutionary path from the nonpathogenic El Tor strain in the Middle East to the seventh-pandemic strain (branches 2, 6, 10, 18, 19, and 22, which cover the period from 1902 to 1957), there were a total of 52 recombination events (Dataset S1 and Fig. S6), and almost all of these branches have a high proportion of SNPs due to recombination (Fig. 4 and Fig. S6). For example, strain M66-2 gained 27-fold more SNPs (1,940/70) by recombination than mutation after diverging from the main lineage at the end of branch 10 (Figs. S1, S2, and S5). This high ratio of SNPs due to recombination is comparable to the 40-fold ratio between SNPs caused by recombination and mutation that occurred during the divergence of the sixth- and seventh-pandemic stains (8). There was almost certainly continuing adaptation to the new niche of pathogens, and much of this adaptation could have been mediated by recombination, given the high number of SNPs caused by recombination and the likelihood that many of them were adaptive in their donor strains, whereas mutation is entirely random. There was also gain and loss of genes within recombinant segments, and these changes could have also been adaptive. The selective advantages of the evolutionary changes that occurred could have been related to the new environments occupied by the strains in the Middle East and Makassar.

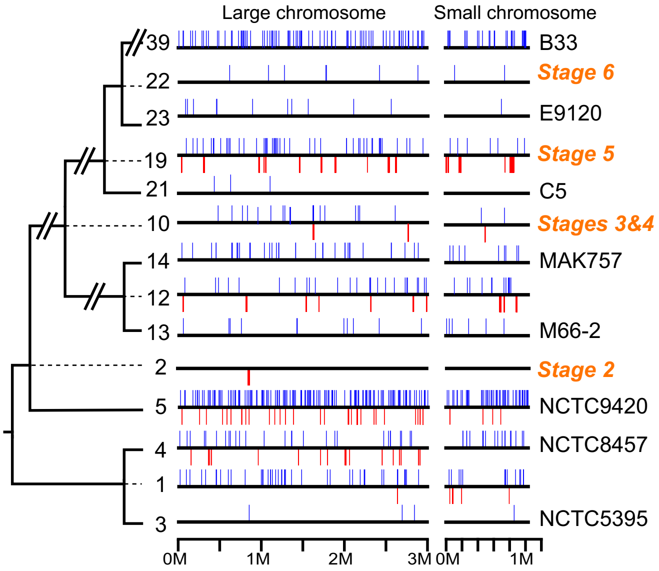

Fig. 4. The chromosomal locations of all mutation and recombination events on the major branches related to the six evolutionary stages of the seventh-pandemic strain. Mutations are marked as blue vertical bars above thick black lines, which represent the two chromosomes. The red vertical bars below show the locations of recombination events. The branch numbering is the same as in Fig. 1, but the mutations and recombination events shown include those back to the nodes. The additional events are shown on their own branches in Fig. S6, which includes all branches. Note that there are recombination events on all stage branches except for stage 6, which is short and has only 12 mutations, whereas the branch from the start of the pandemic to strain B33 has no recombination events. Additionally, for the early part of the 1937 outbreak, branch 12 has recombination events, but there are none on the terminal branches 13 and 14.

However, previous studies have shown that recombination events have been virtually absent from the seventh-pandemic strains since their origin in 1961 (22), with the only exception being the event that included the gain of the O139 O-antigen gene cluster. This low recombination frequency of seventh-pandemic strains is confirmed in this study, with the only recombination event observed for the seventh-pandemic strains from branch 6 onward being that involving the O139 O-antigen gene cluster on branch 35 to strain MO10, as shown in Fig. S6. In contrast, there were 118 mutations, but no recombination events, over 44 y from the end of branch 5 to strain B33, the most recently isolated strain in the phylogenic tree (Figs. 2 and 4). Additionally, as discussed above, recombination occurs at high frequencies in prepandemic and prepandemic-related lineages, as shown in Fig. 4 for stages 2, 3, 4, and 5 of the seventh pandemic lineage and branches 1, 4, and 5 in the Middle East. It is therefore very interesting that there are no recombination events on the terminal branches leading to strain M66-2 (branch 13) and MAK747 (branch 14) (Fig. 4), which were both isolated during the 1937 Makassar outbreak. However, there are 18 recombination events on branch 12, the shared branch that preceded branches 13 and 14, and these brought in 704 SNPs, 12-fold more than arose by mutation.

The almost total absence of recombination in the seventh-pandemic strain and in the terminal branches leading to the 1937 Makassar outbreak strains indicates a major change in the niche occupied by the outbreak and pandemic strains, with almost complete separation from environmental *V. cholerae* strains. At first, the absence of recombination in these strains may seem surprising, because the growth of *V. cholerae* strains on chitin surfaces in aquatic habitats induces competence for natural transformation (43), a condition that would be expected to occur while these pathogenic strains reside in the environment between infections. It appears that both the 1937 Makassar-outbreak strain and the seventh-pandemic strain are from a niche(s) that does not favor recombination.

The high rate of recombination in the prepandemic strains appears to be normal for *V. cholerae*. Unfortunately, there are no genome sequence-based data for recombination in environmental strains, but several MLST studies on environmental strains have shown high diversity, indicating high recombination rates, which have also been supported by several additional criteria (44–46). Each of these referenced studies used strains with a sampling bias toward human-associated strains in countries with often inadequate treatment of sewage, adding the complication that some of the “environmental” strains were from fecal contamination. However, Esteves et al. (47) reported MLST data on *V. cholerae* strains from coastal lagoons in the Mediterranean Sea, far from cholera outbreaks, and they also found high levels of recombination in what is probably a more representative collection of environmental strains.

Recombination in a seventh-pandemic strain has been observed under laboratory conditions when induced by the presence of chitin (48). Pandemic *V. cholerae* strains have also been associated with crustacean and other invertebrates, which could promote recombination, because their exoskeleton is mostly chitin. It is thus useful to consider why *V. cholerae* strains exhibited a low rate of recombination during the seventh pandemic and outbreak periods. It may simply be that *V. cholerae* spends less time in the environment during pandemic periods, with infection occurring soon after the organisms enter water, which would reduce the opportunities for recombination. This explanation is supported by the finding that organisms isolated from stool samples show increased infectivity in mice, with this increase being transient (49). However, cholera is commonly seasonal in areas where it is endemic, and rapid reinfection would not foster survival between seasons. Such rapid reinfection is also less likely in Makassar, even during periods of outbreak, because the number of outbreak cases in this region is low and often scattered.

The low levels of recombination that were observed during long-term survival of the organisms in the environment could be the result of cells of pathogenic *V. cholerae* strains existing primarily in a viable but noncultivable form (VBNC) when not in a host (48, 50). In these experiments, organisms isolated from the stools of infected persons entered the VBNC state within 24 h of being placed in water. The organisms were shown to survive in the VBNC form for more than a year and could convert to active growth when placed into rabbit ileal loops. The adoption of the VBNC form has been proposed as a survival strategy between outbreaks and would account for the seasonal occurrence of cholera in Asia, but it could also be an adaptation to avoid competition with strains already occupying the local environment.

Summary and Conclusions

In summary, our analysis provides the clearest view currently possible of the evolutionary history of the seventh-pandemic strain since the first observation of an El Tor strain. The analysis includes all seven currently available prepandemic strains, as well as a strain isolated just as the seventh pandemic started. Six stages were delineated for the genetic changes that occurred, but we can only infer the ecology of the strains at each stage by what is known of either the strains that were isolated and studied at the time or those that are still available for ongoing study.

Three clearly distinguished ecotypes were identified for the available strains within the lineage of the seventh-pandemic strain. The first ecotype is of a nonpathogenic form that colonized humans in the Middle East and was observed from 1897 to 1938. It was distinguished from the then-active sixth-pandemic strain and named El Tor, as discussed above, and strains isolated from 1930 through 1954 are still available. This form, which appears to colonize humans for periods of up to 5 wk with no adverse symptoms, was first reported publicly in 1906 (15). There is reason to think that this colonization was very different from that of modern cholera, and probably occurred in the lumen of the intestine. The second ecotype resembled the well-documented strains of the sixth and seventh pandemics in terms of disease caused, but it is much less capable of transmission. The strains encompassing this form were found in the Makassar outbreaks between 1937 and 1957. The third ecotype is that of seventh-pandemic strains, the subject of extensive experimental and epidemiological studies.

We allocated the many genetic events associated with the above to different branches in the phylogenic tree. This analysis narrowed down the genetic changes that could have contributed to the transitions between ecotypes. The genetic events important in pathogenicity or pandemicity identified in this study could be an interesting subject for future experimental studies, especially those related to the different spread capabilities of the second and the third ecotype forms. We also found that the absence of recombination in the seventh pandemic period was likewise observed in the Makassar outbreak period, indicating that *V. cholerae* strains that cause human diarrhea are not able to recombine with environmental *V. cholerae* strains. A possible reason for this finding is that these forms can exist in the VBNC state for longer periods in the environment.

Cholera is very unusual in that, not only is the pandemic form a single lineage, but ancestors of the seventh pandemic form were detected more than a century ago, and we have strains for intermediates that allow the development of that form to be followed in detail. The evolution of the seventh-pandemic strains includes several important changes mediated by gains of genes thought to come from *V. cholerae* in the environment. In general, these genes resemble those that are also present in the sixth-pandemic strains, implying that the prior presence of pathogenic *V. cholerae* was important for the evolution of pathogenicity and the development of the seventh-pandemic strain. Although only studied in detail with regard to the sixth pandemic, classical biotype strains, now thought to have been responsible for the first through sixth pandemics, seem to have played an important role in the evolution of the seventh-pandemic strain, despite the divergence of the classical and El Tor biotype strains much further back.

Materials and Methods

Genome Sequencing. We obtained complete genome sequences for 10 *V. cholerae* strains (Table S1) using PacBio technology. First, we prepared a single 10-kb library for each strain that was sequenced using C_2 chemistry in eight single-molecule real-time cells with a 90-min collection protocol on a PacBio RS. The PacBio read data for each strain were de novo assembled into complete genomes by using the PacBio hierarchical genome assembly process/Quiver software package, followed by Minimus 2, and they were polished with Quiver.

Detection of SNPs and Phylogenetic Analysis. We detected SNPs in the presently available genome sequences by aligning each to that of N16961 using BLASTn and the Mauve method (51). We then distinguished SNPs that could be attributed to mutations in the lineage from SNPs gained by recombination-mediated importation of segments derived from donor strains that carried preexisting mutations. The distribution of mutations was expected to be random, which would give an exponential distribution for the inter-SNP distances. We first classified the SNPs into different groups on the basis of their distributions in the genomes under analysis and then used a described method (8) to analyze the SNPs in each group separately to detect segments with high densities of SNPs that did not fit a general Gaussian distribution as expected for random mutations. These regions were allocated as recombination events, and SNPs in these regions were omitted when building the phylogenetic tree. Moreover, the Kolmogorov–Smirnov test was used to test the distributions of the SNPs by implementing the ks.test() function in R. A phylogenetic tree was generated by using RAxML (52) with 1,000 bootstrap samples using only mutational SNPs (999 in total). Mutations and recombination events were then allocated to specific branches. The mutations and recombination events were first grouped into different patterns according to the strains in which they occur, and then each pattern was allocated to a specific branch of the tree.

Genome Annotation and Comparative Genomics Analysis. ORFs spanning 30 amino acids in length were predicted by using Glimmer (Version 3.0) and verified manually based on the annotation of N16961. Transfer RNA and ribosomal RNA genes were predicted by using tRNAscan-SE. Artemis (53) was used to collate the data and facilitate annotation. Function predictions were based on BLASTp similarity searches in the UniProtKB, GenBank, and Swiss-Prot protein databases, as well as the clusters of orthologous groups database.

Genomes were compared by tBLASTx, BLASTn, and the Mauve method. The Artemis comparison tool was manually applied to the files generated to search for indels (such as genomic islands and phages). These were then allocated to specific branches of the tree in the same manner as for mutations and recombination events.

Because of the use of eight draft genomes of relatively low quality, we had to exclude highly variable regions when constructing the phylogenic tree, because they are very difficult to include in alignments. These regions include the RNA gene loci, the large integron and the CTX regions. For instance, only 90% (362,834 bp) of the complete genome of N16961 (4,033,460 bp) is included in the analysis for Fig. 1. To reduce this loss, we also analyzed eight selected full-genome sequences separately (Fig. S2), giving us 3,794,564 bp shared by the 8 strains. The increase of 123,938 bp of additional coverage extends the data available for the branches leading to the seventh pandemic. The better quality of the full genome sequences allowed us to identify all indels in these strains and allocate them to branches. Note that we retained the node dates obtained with the full dataset when creating the new tree. Sites with a gain or loss of 1–3 bp were treated as mutations because of strand slippage during replication. There were no sharp cut-offs, and because shorter sequences are difficult to define from Illumina data, we made another arbitrary distinction into short (4–999 bp) and major (>999 bp) indels. For the main lineage, we also investigated the regions that were excluded because of their variability, and we were able to retrieve information on additional events. The lengths of the DNA segments involved in each strain are shown in Table S1. The percentage of the genome in the shared DNA (Fig. S2) ranged from 95.1% for NCTC5395 to 99.0% for M66-2.

BEAST Analysis. We used the program BEAST (26) to infer the dates for the nodes in the tree under a relaxed molecular clock. The data were analyzed by using a coalescent constant population size and a general time-reversible model with gamma correction. The results were produced from three independent chains of 50 million stages each; these were sampled every 10,000 stages to ensure good mixing. The first 5 million stages of each chain were discarded as burn-in. The results were combined by using Log Combiner, and the phylogenetic tree built by RAxML as a target tree was generated by using Tree Annotator, both of which are components of the BEAST package.

BEAST determines node dates based on known dates of isolation and by minimizing the variation in inferred mutation rates. In this case, we also had the information that the seventh pandemic began in 1961. The tree as first generated (Fig. S3) placed the divergence of E9120 and the ongoing seventh-pandemic lineage in 1955, which is not consistent with the historical record. We therefore imposed the requirement that this divergence occur in 1960, the earliest date consistent with the historical data, and therefore changed the dates for the nodes after branches 22, 24, and 25 to 1960, 1960, and 1961, respectively, and the start of branch 22 to 1954, to make them compatible with the historical record (Fig. S1). The major effect was to reduce the length of branch 30 and increase that of branch 19, and minor changes were made to some other nodes to spread the effect in a proportional manner. The final tree is shown in Fig. 1 and Fig. S1, in which nodes are located according to the estimated branching dates. The date for the first node was set to 1902. Outgroup strain O395 was not included in the BEAST analysis because it is too divergent, but all of the SNP differences between NCTC8457/NCTC5395 and the other strains can be attributed to branch 1, because the sixth-pandemic strain O395 has the alternative allele, and branch 2 has no measurable length.

ACKNOWLEDGMENTS. We thank James Kaper for providing strain E506 (54). This work was supported by National Natural Science Foundation of China (NSFC) Key Program Grant 31530083; National 973 Program of China Grant 2013CB733904; National Special Project on Research and Development of Key Biosafety Technologies Grant 2016YFC1200104; and NSFC General Program Grants 31270133, 31470194, 31371259, 81471904, and 31270003.

1. Barua D (1992) History of cholera. *Cholera*, eds Barua D, Greenough WB, III (Plenum, New York), pp 1–36.
2. Ali M, et al. (2012) The global burden of cholera. *Bull World Health Organ* 90(3): 209–218A.
3. Pan American Health Organization and World Health Organization (2014) *Epidemiological Update - Cholera - 18 February 2014* (Pan American Health Organization/World Health Organization, Washington, DC).
4. Islam MS, et al. (2011) Phenotypic, genotypic, and antibiotic sensitivity patterns of strains isolated from the cholera epidemic in Zimbabwe. *J Clin Microbiol* 49(6): 2325–2327.
5. Harris JB, LaRocque RC, Qadri F, Ryan ET, Calderwood SB (2012) Cholera. *Lancet* 379(9835):2466–2476.
6. Chatterjee SN, Chaudhuri K (2003) Lipopolysaccharides of *Vibrio cholerae*. I. Physical and chemical characterization. *Biochim Biophys Acta* 1639(2):65–79.
7. Devault AM, et al. (2014) Second-pandemic strain of *Vibrio cholerae* from the Philadelphia cholera outbreak of 1849. *N Engl J Med* 370(4):334–340.
8. Feng L, et al. (2008) A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PLoS One* 3(12):e4053.
9. Salim A, Lan R, Reeves PR (2005) *Vibrio cholerae* pathogenic clones. *Emerg Infect Dis* 11(11):1758–1760.
10. Abdoelrachman R (1944) *Vibrio* research in the Hejaz in connection with the El Tor Problem. *Antonie van Leeuwenhoek* 10(1):93–100.
11. Ruffer MA (1907) Researches on the bacteriological diagnosis of cholera, carried out by medical officers of the sanitary, maritime and quarantine council of Egypt. *BMJ* 1(2413):735–742.
12. Flu PC (1913) Een cholera-achtlge vibrio als verwekster van een klinisch op echte asiatische cholera gelijkend ziekteproces. *Geneeskundig Tijdschrift Voor Nederlandsch-Indie* 53:771.
13. Gilmour M (1938) Rapport sur le pelerinage de 1938 [Report on the pilgrimage or 1938). *Bull Mens Soc Med Mil Fr* 30:2534–2536.
14. Geddings HD (1912) Public health weekly reports for March 15, 1912. *Public Health Rep* 27(11):371–419.
15. Gotschlich F (1906) Uber Cholera and cholerazhnliche Vibrionen unter den aus Mekka zuriickkehrenden Pilgern. *Z Hyg Infektionskr* 53:281–304.
16. Hugh R (1965) A comparison of *Vibrio cholerae* pacini and *Vibrio* El Tor pribram. *Int J Syst Evol Microbiol* 15(1):61–68.
17. Tanamal ST (1959) Notes on paracholera in Sulawesi (Celebes). *Am J Trop Med Hyg* 8(1):72–78.
18. Mukerjee S (1963) Problems of cholera (El Tor). *Am J Trop Med Hyg* 12:388–392.
19. de Moor CE (1949) Paracholera (E1 Tor): *Enteritis choleriformis* E1 Tor van Loghem. *Bull World Health Organ* 2(1):5–17.

20. Felsenfeld O (1964) Present status of the El Tor Vibrio problem. *Bacteriol Rev* 28: 72–86.
21. Felsenfeld O (1963) Some observations on the cholera (El Tor) epidemic in 1961-1962. *Bull World Health Organ*283289–296.
22. Mutreja A, et al. (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477(7365):462–465.
23. Kim EJ, Lee CH, Nair GB, Kim DW (2015) Whole-genome sequence comparisons reveal the evolution of *Vibrio cholerae* O1. *Trends Microbiol* 23(8):479–489.
24. Didelot X, et al. (2015) The role of China in the global spread of the current cholera pandemic. *PLoS Genet* 11(3):e1005072.
25. Teppema JS, Guinée PA, Ibrahim AA, Pâques M, Ruitenberg EJ (1987) In vivo adherence and colonization of *Vibrio cholerae* strains that differ in hemagglutinating activity and motility. *Infect Immun* 55(9):2093–2102.
26. Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4(5):e88.
27. Chastel C (2007) [The centenary of the discovery of the vibrio El Tor (1905) or dubious beginnings of the seventh pandemic of cholera]. *Hist Sci Med* 41(1):71–82.
28. Stine OC, Morris JG, Jr (2014) Circulation and transmission of clones of *Vibrio cholerae* during cholera outbreaks. *Curr Top Microbiol Immunol* 379:181–193.
29. Bryceson AD (1977) Cholera, the flickering flame. *Proc R Soc Med* 70(5):363–365.
30. Krebs SJ, Taylor RK (2011) Protection and attachment of *Vibrio cholerae* mediated by the toxin-coregulated pilus in the infant mouse model. *J Bacteriol* 193(19):5260–5270.
31. Rhine JA, Taylor RK (1994) TcpA pilin sequences and colonization requirements for O1 and O139 *vibrio cholerae*. *Mol Microbiol* 13(6):1013–1020.
32. Tay CY, Reeves PR, Lan R (2008) Importation of the major pilin TcpA gene and frequent recombination drive the divergence of the Vibrio pathogenicity island in *Vibrio cholerae*. *FEMS Microbiol Lett* 289(2):210–218.
33. Karaolis DKR, et al. (1998) A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. *Proc Natl Acad Sci USA* 95(6):3134–3139.
34. Haley BJ, et al. (2014) Genomic and phenotypic characterization of *Vibrio cholerae* non-O1 isolates from a US Gulf Coast cholera outbreak. *PLoS One* 9(4):e86264.
35. Kaper JB, Morris JG, Jr, Levine MM (1995) Cholera. *Clin Microbiol Rev* 8(1):48–86.
36. Collins JW, et al. (2014) *Citrobacter rodentium*: Infection, inflammation and the microbiota. *Nat Rev Microbiol* 12(9):612–623.
37. Qadri F, et al. (1997) Comparison of immune responses in patients infected with *Vibrio cholerae* O139 and O1. *Infect Immun* 65(9):3571–3576.
38. Kamada N, et al. (2015) Humoral immunity in the gut selectively targets phenotypically virulent attaching-and-effacing bacteria for intraluminal elimination. *Cell Host Microbe* 17(5):617–627.
39. Kamada N, et al. (2012) Regulated virulence controls the ability of a pathogen to compete with the gut microbiota. *Science* 336(6086):1325–1329.
40. Okada K, Roobthaisong A, Swaddiwudhipong W, Hamada S, Chantaroj S (2013) *Vibrio cholerae* O1 isolate with novel genetic background, Thailand-Myanmar. *Emerg Infect Dis* 19(6):1015–1017.
41. Dziejman M, et al. (2002) Comparative genomic analysis of *Vibrio cholerae*: Genes that correlate with cholera endemic and pandemic disease. *Proc Natl Acad Sci USA* 99(3):1556–1561.
42. Davies BW, Bogard RW, Young TS, Mekalanos JJ (2012) Coordinated regulation of accessory genetic elements produces cyclic di-nucleotides for *V. cholerae* virulence. *Cell* 149(2):358–370.
43. Meibom KL, Blokesch M, Dolganov NA, Wu CY, Schoolnik GK (2005) Chitin induces natural competence in *Vibrio cholerae*. *Science* 310(5755):1824–1827.
44. Aydanian A, et al. (2015) Genetic relatedness of selected clinical and environmental non-O1/O139 *Vibrio cholerae*. *Int J Infect Dis* 37:152–158.
45. Octavia S, et al. (2013) Population structure and evolution of non-O1/non-O139 *Vibrio cholerae* by multilocus sequence typing. *PLoS One* 8(6):e65342.
46. Li F, et al. (2014) Distribution of virulence-associated genes and genetic relationships in non-O1/O139 *Vibrio cholerae* aquatic isolates from China. *Appl Environ Microbiol* 80(16):4987–4992.
47. Esteves K, et al. (2015) Highly diverse recombining populations of *Vibrio cholerae* and *Vibrio parahaemolyticus* in French Mediterranean coastal lagoons. *Front Microbiol* 6: 708.
48. Alam M, et al. (2007) Viable but nonculturable *Vibrio cholerae* O1 in biofilms in the aquatic environment and their role in cholera transmission. *Proc Natl Acad Sci USA* 104(45):17801–17806.
49. Alam A, et al. (2005) Hyperinfectivity of human-passaged *Vibrio cholerae* can be modeled by growth in the infant mouse. *Infect Immun* 73(10):6674–6679.
50. Kamruzzaman M, et al. (2010) Quorum-regulated biofilms enhance the development of conditionally viable, environmental *Vibrio cholerae*. *Proc Natl Acad Sci USA* 107(4): 1588–1593.
51. Darling AE, Treangen TJ, Messeguer X, Perna NT (2007) Analyzing patterns of microbial evolution using the mauve genome alignment system. *Methods Mol Biol* 396: 135–152.
52. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
53. Carver TJ, et al. (2005) ACT: The Artemis Comparison Tool. *Bioinformatics* 21(16): 3422–3423.
54. Kaper JB, Nataro JP, Roberts NC, Siebeling RJ, Bradford HB (1986) Molecular epidemiology of non-O1 *Vibrio cholera*e and *Vibrio mimicus* in the US Gulf Coast region. *J Clin Microbiol* 23(3):652–654.
55. Long JG (1902) Turkey—the quaranteen camp at El Tor. *Public Health Rep* 17(20): 1156–1159.
56. Chun J, et al. (2009) Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc Natl Acad Sci USA* 106(36):15442–15447.
57. Doorenbos W, Kop J (1951) El Tor *Vibrio* in chloramphenicol estimation. *Lancet* 257: 691.
58. Sutherland H (2000) Trepang and wangkang—the China trade of eighteenth-century Makassar c. 1720s-1840s. *Authority and Enterprise Among the Peoples of South Sulawesi*, eds Tol R, van Dijk C, Acciaioli G (KITLV, Leiden, The Netherlands), pp 73–94.
59. Stacey N (2007) *Boats to Burn, Bajo Fishing Activity in the Australian Fishing Zone* (ANU E, Canberra, Australia).