New *Vibrio cholerae* sequences from Eastern and Southern Africa alter our understanding of regional cholera transmission

Shaoming Xiao[1,2]*, Ahmed Abade[3,4]*, Waqo Boru[4]*, Watipaso Kasambara[5]*, John Mwaba[6,7]*, Francis Ongole[8]*, Mariam Mmanywa[3] , Nídia Sequeira Trovão[9] , Roma Chilengi [10], Geoffrey Kwenda[10], Christopher Garimoi Orach[8,12], Innocent Chibwe[5], Godfrey Bwire[8], O. Colin Stine[13], Aaron M. Milstone[1,14], Justin Lessler[14,15,16], Andrew S. Azman[14,17,18], Wensheng Luo[2], Kelsey Murt[2], David A. Sack[2], Amanda K. Debes[2†], Shirlee Wohl[14,19 †]

[1] Division of Pediatric Infectious Disease, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[2] Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[3] Ministry of Health, Dar es Salaam, Tanzania
[4] Field Epidemiology and Laboratory Training Program, Nairobi, Kenya
[5] Ministry of Health, Lilongwe, Malawi
[6] Center for Infectious Disease Research, Zambia
[7] Department of Pathology and Microbiology, University Teaching Hospital, Lusaka, Zambia
[8] Ministry of Health, Kampala, Uganda
[9] Fogarty International Center, National Institute of Health, Bethesda, MD, USA
[10] Zambia National Public Health Institute, Lusaka, Zambia
[11] Department of Biomedical Sciences, School of Health Sciences, University of Zambia, Lusaka, Zambia
[12] Makerere University School of Public Health, Kampala, Uganda
[13] University of Maryland School of Medicine, Baltimore, USA
[14] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[15] Department of Epidemiology, Gillings School of Public Health, University of North Carolina, Chapel Hill, NC, USA
[16] Carolina Population Center, University of North Carolina, Chapel Hill, NC, USA
[17] Division of Tropical and Humanitarian Medicine, Geneva University Hospitals, Geneva, Switzerland
[18] Geneva Centre for Emerging Viral Diseases, Geneva University Hospitals, Geneva, Switzerland
[19] Division of Infectious Diseases, Brigham and Women's Hospital, Boston, MA, USA

*co-first authors
[†]co-senior authors
Correspondence: swohl@bwh.harvard.edu (S.W.)

ABSTRACT

Despite ongoing containment and vaccination efforts, cholera remains prevalent in many countries in sub-Saharan Africa. Part of the difficulty in containing cholera comes from our lack of understanding of how it circulates throughout the region. To better characterize regional transmission, we generated and analyzed 118 *Vibrio cholerae* genomes collected between 2007-2019 from five different countries in Southern and Eastern Africa. We showed that *V. cholerae* sequencing can be successful from a variety of sample types and filled in spatial and temporal gaps in our understanding of circulating lineages, including providing some of the first sequences from the 2018-2019 outbreaks in Uganda, Kenya, Tanzania, Zambia, and Malawi. Our results present a complex picture of cholera transmission in the region, with multiple lineages found to be co-circulating within several countries. We also find evidence that previously identified sporadic cases may be from larger, undersampled outbreaks, highlighting the need for careful examination of sampling biases and underscoring the need for continued and expanded cholera surveillance across the African continent.

INTRODUCTION

Cholera is an acute watery diarrheal disease that is estimated to cause approximately 95,000 deaths annually in endemic countries [1]. Most of these deaths occur during large outbreaks that are associated with specific serotypes of the *Vibrio cholerae* bacterium—O1 and O139 [2]. The ongoing seventh pandemic of cholera, caused by *V. cholerae* serogroup O1 biotype El Tor (and often referred to at the 7PET lineage), was first detected in Indonesia in 1961 and has since spread through Asia to Africa, Europe, and Latin America [3]. Africa, especially sub-Saharan Africa, has experienced a higher cholera burden during the seventh pandemic than other regions of the world [1,4,5], with numerous large outbreaks reported since 1961, including several during the last decade [6]. Although a few studies have explored the transmission patterns of *V. cholerae* in Africa, there is still a limited understanding of how cholera circulates [7].

Unlike endemic regions in Southeast Asia, cholera in Africa is often seasonal and periodic, with many countries experiencing months or years of no cholera cases between successive outbreaks. Although risk factors such as population density, seasonality and urbanization may explain some of variability in observed transmission patterns [3,8–10], within-country variation in some of these factors (e.g., seasonality) [11] highlights the limitations of these risk factors in fully capturing cholera dynamics. For example, most cholera outbreaks in Malawi have historically occurred during the rainy season, but the 2022 outbreak (which has caused approximately 60,0000 cases as of September 2023 [12]) started in the dry season [13]. It is clear that these factors alone cannot explain the repeated reemergence of cholera in regions without ongoing cases, and that fully elucidating cholera dynamics on the continent will require a better understanding of *V. cholerae* movement within and between countries. Importantly, our current lack of understanding—or perhaps even misunderstanding—of the transmission patterns of *V. cholerae* makes the prevention of outbreaks challenging, since it is difficult to prevent outbreaks without a comprehensive understanding of where they are coming from.

Genomic analysis provides some insight into the movement of pathogen lineages and has been used already to track *V. cholerae* O1 transmission in Africa. Early genomic studies showed that 7PET *V. cholerae* O1 was introduced from Asia to Africa at least 12 times since 1970 [14]. These introduction events occurred mainly in West and Southeast Africa, each defined by a specific sublineage (AFR1-AFR12) [15]. At least three additional lineages (AFR13, AFR14, and AFR15) [14–17] have been identified more recently, all following the same Asia-to-Africa introduction pattern. More recent lineages have generally appeared to replace older lineages upon introduction [14,16], though there is still limited information about how these lineages interact and circulate in sub-Saharan Africa once present.

A number of more recent studies have used genomics to characterize *V. cholerae* in Southeast Africa [18–22], a region of particularly high cholera burden [1,5,23]. However, the difficulty of sequencing cholera in low-resource settings [24] means there are still significant gaps in our understanding of which lineages were circulating at different types over the last few decades. Furthermore, there have been limited opportunities to explore potential transmission across country borders [25]. In this study, we aimed to fill some of the gaps in our understanding of 7PET *V. cholerae* O1 circulation in Southeast Africa using isolates collected from 2007–2018 in Uganda, Kenya, Tanzania, Malawi, and Zambia. In addition, we explored the potential of performing whole genome sequencing from low-cost sample preservation methods (e.g., stool or isolates on filter paper), which facilitates specimen collection in resource-constrained areas where isolates would otherwise not be captured [18,26]. We found that sequencing from these samples can produce whole genome sequences, which we used to identify multiple co-circulating *V. cholerae* lineages in Southeast Africa. Our observations both confirm broad patterns of cholera transmission across Southeast Africa and raise important questions about our ability to capture the full picture of disease emergence and spread.

RESULTS

To better understand cholera transmission in Southeast Africa, we performed whole genome sequencing (WGS) of *V. cholerae* isolates from five neighboring countries in this region: Uganda, Kenya, Tanzania, Malawi, and Zambia (**Fig 1A**). In total, we generated 118 high-quality genomes (4, 71, 22, 20, and 1 from Uganda, Kenya, Tanzania, Malawi and Zambia, respectively) from 142 sequenced isolates collected between 2007-2019 (**Supplementary Data 1**). These sequences fill a number of geographic and temporal gaps in available *V. cholerae* genomic data and demonstrate that WGS may be possible from multiple sample types and extraction methods already used in rural or remote, as well as regional, health facilities.

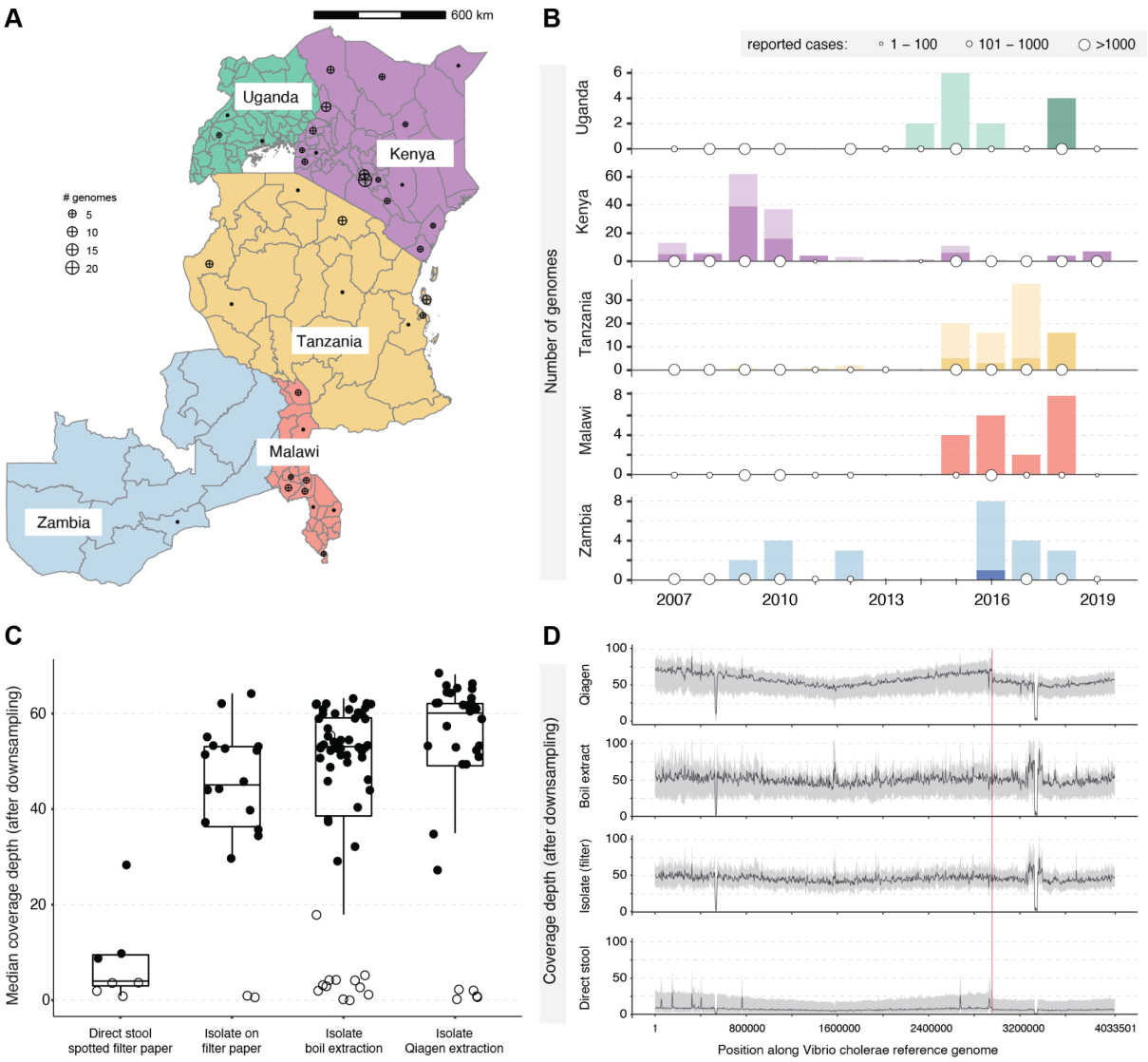


Figure 1. *Vibrio cholerae* sequences generated in this study. (**A**) Geographic distribution of genomes generated in this study. 9 samples from Kenya with no sub-country metadata are not shown. (**B**) Genomes generated in this study (dark bars) overlaid on previously published genomes (light bars), per country and per year (genomes with unknown collection year excluded) (**Supplementary Data 2**). White dots on x-axis: total number of cholera cases reported to the World Health Organization in that year [27] (**Supplementary Data 3**). (**C**) Median coverage depth, per sample type and extraction method (all samples were boil extracted except those specifically marked as Qiagen extractions), across the *V. cholerae* genome after downsampling reads to 1 million/sample. Only samples sequenced on the Illumina platform are included. 2 samples of unknown sample type were not shown. Black dots: samples classified as high-quality genomes (based on non-downsampled reads; see **Methods**); white dots: low-quality genomes excluded from further analyses. (**Supplementary Data 4**) (**D**) Distribution of coverage depth at each position along the *V. cholerae* reference genome for each sample preparation after downsampling reads to 1 million/sample. Solid black line: median coverage across all samples within the sample type group; gray bands: 20-80 quartile range, averaged across a 4000-nt sliding window [28]. (**Supplementary Data 5**) Red vertical line: boundary between chromosomes 1 and 2 of *V. cholerae* genome.

To obtain the most comprehensive picture of cholera transmission within and between Southeast African countries, we selected samples from available isolates that maximized the number of distinct geographic districts. In Kenya, for example, we sequenced isolates from 16 different regions (**Fig 1A**). We also selected samples to focus on more recent outbreaks for which no or limited sequencing data were available, providing some of the first sequences from the 2018-2019 outbreaks in Uganda, Kenya, Tanzania, and Malawi (**Fig 1B**).

Other than sequencing costs, one of the barriers to generating *V. cholerae* whole genome sequences is heterogeneity in the sample types available; sequencing typically requires relatively large amounts of purified nucleic acid, and is therefore often not attempted when only crude specimens (as opposed to cultured isolates) are available. Previous studies have shown that *V. cholerae* sequencing from Alkaline Peptone enriched (APW) stool or isolates on filter paper can be successful [26], so we attempted to take this one step further, comparing the success of various specimen types preserved on filter paper across samples from several countries and using different nucleic acid extraction techniques. Specifically, we compared sequencing from bacterial cultures, bacterial isolates preserved on filter paper, and whole stool directly spotted onto filter paper. We experimented with both column-based extraction kits as well as boil extraction methods, which can be done at reduced cost, with less laboratory equipment, and (when isolates are preserved on filter paper) without temperature-controlled supply chains.

As expected, we found that, after normalizing the total number of reads per sample (see **Methods**), column-extracted isolates produced the highest quality *V. cholerae* genomes, while direct stool spotted filter paper samples had the lowest median coverage depth across the *V. cholerae* genome (**Fig 1C**). To minimize potential platform-related effects, these results include only samples sequenced on the Illumina sequencing platform (all samples except those collected in Malawi, see **Supplementary Data 1**). Overall, the sequences from all specimen preservation types produced high quality genomes that met our inclusion criteria (see **Methods**). Although the direct stool on filter paper had the lowest median coverage, sample size for this sample type was quite small. Three of seven samples of this type still produced high quality genomes according to our thresholds, without any modification to the sequencing method used.

To ensure that our conclusions about sequencing quality were not driven by highly uneven coverage, we also looked at read depth across the whole *V. cholerae* genome (**Fig 1D**), again focusing only on samples sequenced on the Illumina platform for ease of comparison. We observed relatively even coverage across the genome, with similar trends across all sample types. These analyses show that commonly used low-cost sample preservation methods (e.g., isolates on filter paper) can be useful for genomic studies.

Using high quality genomes from all sample types, we then built a maximum likelihood tree using the 114 high quality genomes generated in this study (4 genomes with unknown collection year were excluded from downstream analyses) and 1380 previously published genomes (see **Supplementary Data 2** for details and acknowledgement). We selected these genomes to capture as much global diversity as possible, with an emphasis on capturing published genomes from sub-Saharan Africa (see **Methods** for details). Using this phylogeny, we determined that sequences generated in this study fell within the AFR10, AFR11 and AFR13 lineages (**Fig 2**). Most of these sequences belong to lineages previously identified in the country of sample origin, though we also observed some unexpected results, suggestive of a more complex picture of cholera transmission than previously thought (**Fig 3A**).

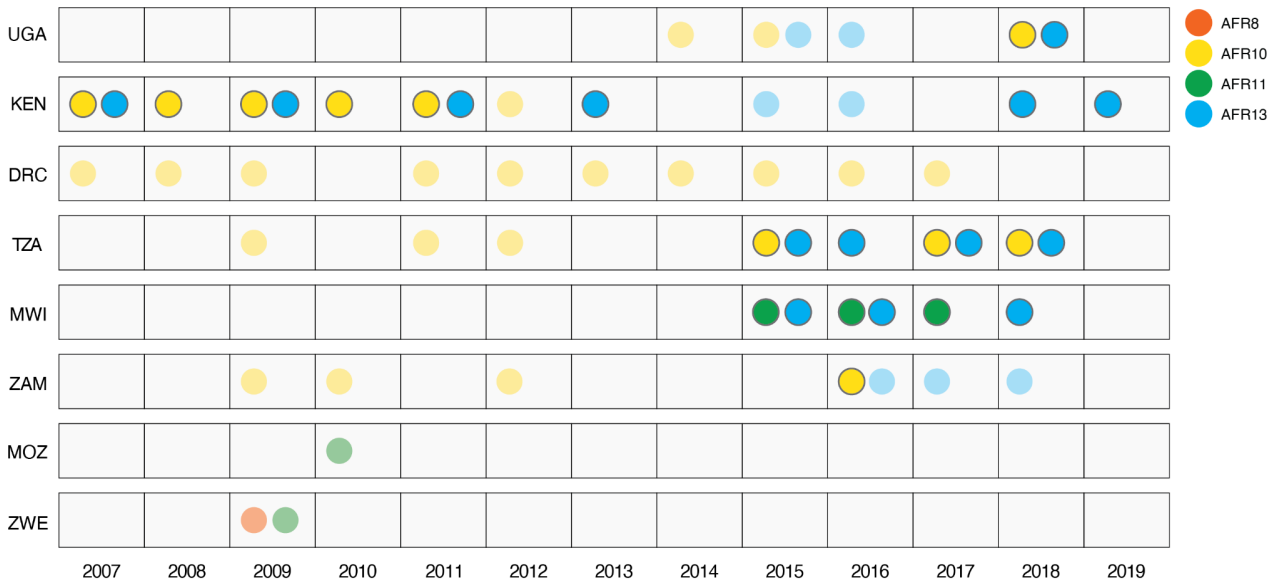


Figure 2. Distribution of lineages identified in samples from this study. Countries in Eastern and Southern Africa, ordered from North to South. Outlined circles: lineages found in genomes from this study. Faded circles (no outline): lineages found in previously published genomes. Abbreviations: UGA (Uganda), KEN (Kenya), DRC (Democratic Republic of the Congo), TZA (Tanzania), MWI (Malawi), ZAM (Zambia), MOZ (Mozambique), ZWE (Zimbabwe).

First, our data highlight that two strains previously considered to be "sporadic outbreaks" because they did not fall into one of the known AFR lineages may reflect undersampled, larger outbreaks. In both cases, these strains were previously isolated on the phylogenetic tree, suggesting they may have not spread widely within Africa once introduced [14]. However, several newly generated genomes cluster closely with these two "sporadic" cases in our phylogeny (**Fig 3B**): 21 genomes from Kenya, Tanzania and Uganda collected between 2009–2018 cluster closely with a former singleton from Kenya in 2009, and 3 sequences from Kenya and Tanzania collected between 2009–2018 cluster closely with a separate former singleton. These outbreaks, seemingly caused by independent transmission events (AFR16 and AFR17), suggest our current understanding of the transmission dynamics of *V. cholerae* O1 in Africa may be incomplete.

Additionally, although the AFR13 sub-lineage has been found in East Africa in recent years (specifically Kenya, Tanzania, Uganda and Zimbabwe, as recently as 2019) [16,18,19,21,29], we found evidence that the introduction of this lineage into Africa may have been earlier than previously thought. The previously reported AFR13 sequences were from samples collected in Kenya, Tanzania and Uganda in 2015 [16,19,21,30], and these studies estimated that this lineage emerged between 2013-2014 [16]. However, we generated 10 sequences from isolates collected in Kenya that fell within the AFR13 lineage but were collected between 2007-2011. These sequences are highly similar to others previously collected in the region and appear ancestral to sequences collected from the major outbreak in Yemen during 2016 to 2017 [16] (**Fig 3C**), thus suggesting the AFR13 lineage may have been circulating in Southeast Africa since at least 2007. Using a Bayesian phylogenetic analysis, we confirmed that the estimated emergence of the AFR13 lineage is significantly earlier than previously estimates following the addition of the sequences generated in this study (2004.17, 95% high posterior density (HPD) interval: 2003.08-2005.13) (**Fig 3D**).

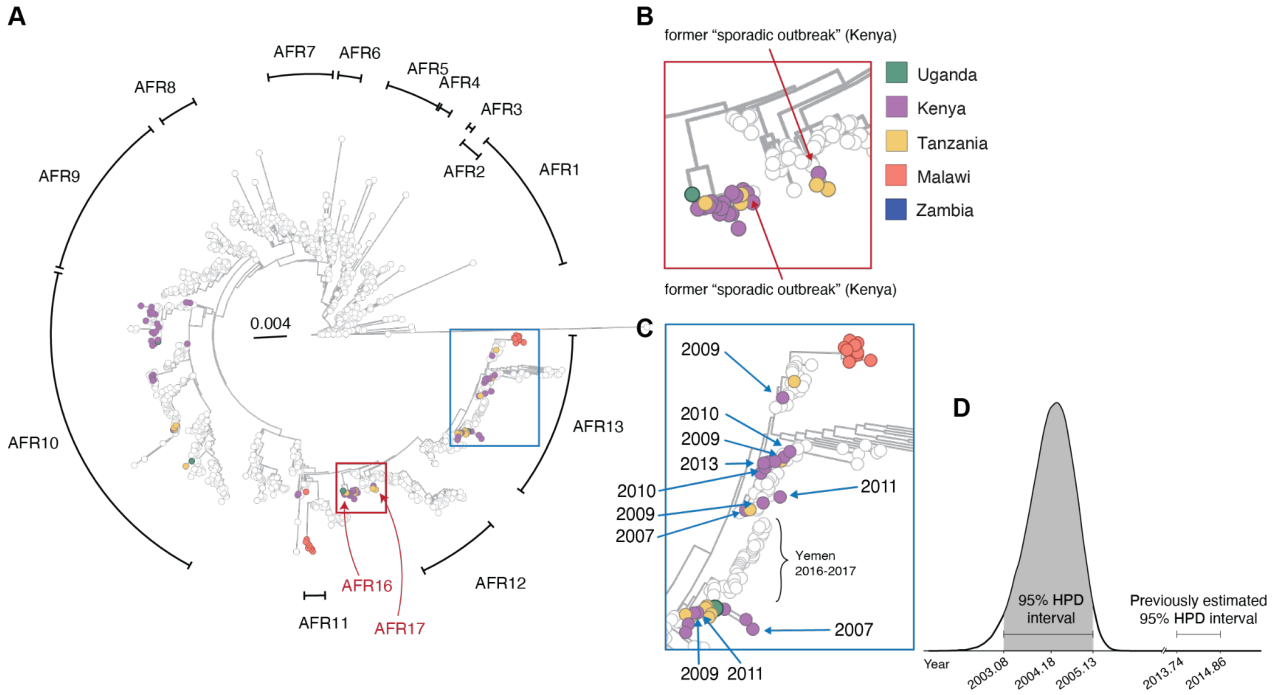


Figure 3. Phylogeny of *V. cholerae* O1 sequences. (**A**) Maximum likelihood tree of 1494 *V. cholerae* genomes. Colored tips: 114 genomes generated in this study, colored by country of sample collection. Unfilled tips: 1380 previously published genomes (see **Supplementary Data 6**). Proposed AFR16 and AFR17 nomenclature shown in red (AFR14 and AFR15 were previously identified but in samples more recent than the study period used here, and therefore are not shown). (**B**) Zoom view of the red box area shown in (A) showing two clades (AFR16 and AFR17) previously considered to be "sporadic outbreaks''. Colors as in (A). (**C**) Zoom view of the blue box area shown in (A) showing part of the AFR13 lineage. two clades previously considered to be "sporadic outbreaks''. All tips without a year label are from samples collected in 2015-2018. Colors are as in (B). (**D**) Posterior distribution of time of the most recent common ancestor (tMRCA) of the AFR13 lineage given the dataset presented in this manuscript, as compared to previous estimates [14] (see **Supplementary Data 7**).

Given the surprising nature of this finding, we attempted to confirm these results. All ten AFR13 sequences in question had high coverage depth and no evidence of contamination. We also confirmed that these isolates were brought to the United States (where the nucleic acid was extracted and they were ultimately sequenced) prior to 2012, so there was no possibility of a mix-up with other samples collected at the same site. Although it will be important to see if future genomic studies support this finding, we are including these sequences in our results because the release of these data could potentially be useful to the understanding of ongoing outbreaks.

Other sequences generated in this study (from multiple different countries) show the presence of co-circulating *V. cholerae* lineages in multiple Southeast African countries, suggesting complex relationships between the spread of different *V. cholerae* strains. Specifically, we observed multiple *V. cholerae* lineages in the same country within a single year in all five countries (**Supplementary Data 2**). For example, we observed multiple lineages circulating within Malawi in 2015 (**Fig 4A**) and 2016 (**Fig 4B**). While this is a clear example of co-circulating lineages, there were also several regions in which cholera cases were reported but no genomic data were available, or regions from which only a single sequence has been published. As above, these observations highlight the need for more dense sequencing and testing of *V. cholerae* to better understand patterns of introduction, lineage circulation, and co-circulation throughout the Southeast Africa region.

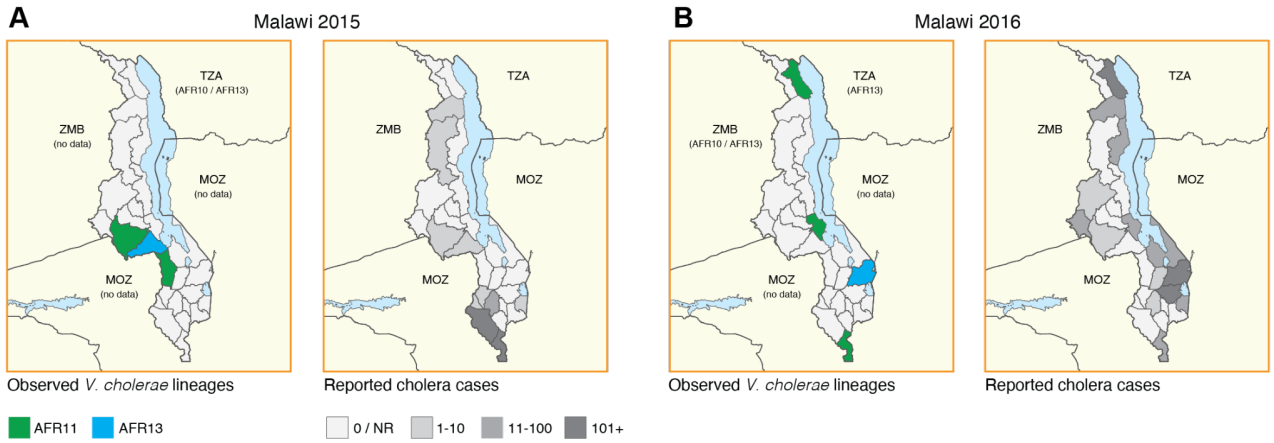


Figure 4. Co-circulation of multiple lineages at country or sub-country level. Examples of co-circulating lineages and movement of *V. cholerae* lineages (left) and reported cholera cases (right) in (**A**) 2015 and (**B**) 2016. Cholera cases from Kenya are as reported in Mutonga *et al.*[31] and reported cholera cases in Malawi are as reported to WHO[27] (**Supplementary Data 3**).

DISCUSSION

The new *V. cholerae* samples sequenced as part of this study present an updated picture of cholera transmission. We found evidence for undersampled lineages, extensive co-circulation of multiple *V. cholerae* lineages within countries, with findings suggestive of an earlier introduction of the AFR13 lineage. Taken together, these findings suggest that improved genomic surveillance of cholera may yet uncover new aspects of cholera transmission in Southeast Africa. They also highlight that insufficient or biased sampling can lead to an inaccurate or incomplete picture of transmission—a potential issue for genomic studies of any pathogen that should be addressed with careful assessment of the potential effects of biased or under-sampling.

Given the frequent movement of *V. cholerae* from Southeast Asia to Southeast Africa, addressing these sampling biases will require that genomic surveillance of cholera improves simultaneously on both continents. For example, the addition of 10 *V. cholerae* sequences from African countries in this manuscript moved the estimated time to the most recent common ancestor of the AFR13 lineage to 2004, approximately 10 years earlier than previous estimates. However, the emergence of the lineage—an understanding of which could provide clues into *V. cholerae* transmission and evolution patterns within and between outbreaks—could have happened in either Africa or Asia, and additional sampling may further refine these date estimates. Improved sampling of both regions would allow for the application of more complex phylodynamic methods for ascertaining the location of lineage emergence, as well as confirmation as to when the lineage may have been introduced into Africa.

We also need to address some of the challenges associated with generating these sequences. Decreasing reliance on culture prior to sequencing will lower the burden to real-time sequencing, and seems increasingly possible even with current sequencing techniques, as shown by our successful sequencing from multiple sample types. Our results strengthen prior evidence that WGS can be effective from isolates stored on filter paper and serve as a proof of concept that WGS is effective from stool samples spotted directly on filter paper, though further optimization will be required to ensure these methods are as cost effective and accessible as possible. As whole genome sequencing becomes more widespread, it will also be important to enable rigorous documentation and high sample quality. We have done our best here to confirm our results with limited access to the original cultures, and we have documented these challenges as a caution to future researchers.

Taken all together, our findings suggest a complex picture of *V. cholerae* transmission in Southeast Africa, and frequent co-circulation of multiple combinations of lineages. These findings also emphasize the importance of a regional approach to cholera surveillance and [25], as outbreaks in neighboring countries are connected both temporally (e.g., spikes in cases occur around the same time; **Fig 1D**) and molecularly (e.g., sequences from *V. cholerae* in multiple countries are highly related). Our hope is that this information can be useful in aiding cholera containment and mitigation, which may require cooperation across country borders. Ensuring these data continue to have public health impact will require both increased sampling (ideally as part of routine, in-country surveillance efforts) and timely coordination with epidemiologists who can interpret both routine and surprising results and ensure they are effectively shared and communicated with stakeholders across the continent.

METHODS

Data availability

Raw data for all sequenced isolates and specimens is available under NCBI BioProject accession: PRJNA616030. R code used to make figures is available here: https://github.com/HopkinsIDD/CholeraGenomics-AfricaSE. Illumina bioinformatics pipelines are available here: https://github.com/HopkinsIDD/illumina-vc. Oxford Nanopore bioinformatics pipelines are available here: https://github.com/HopkinsIDD/minion-vc.

Ethics statement

Samples from Kenya were collected as part of studies approved by the relevant Institutional Review Boards or Ethics Committees at Johns Hopkins Bloomberg School of Public Health (JHSPH; IRB numbers: 00009067, 00008982) and locally (approval number: ESRC P552/2018). Samples in Zambia, Tanzania, and Uganda were collected under JHSPH IRB 00008193, and IRB00008221, respectively, with local approval. Samples analyzed from Malawi were collected under routine public health surveillance. Sequencing and secondary analysis of these samples were determined by JHSPH to be non-human subjects research and exempt from further review.

Sample collections and bacterial culture

Samples in all countries were collected as part of outbreak investigations by the relevant National Public Health Laboratories. In Malawi, sample collection was restricted to patients with a known antibiotic consumption history. In order to confirm cholera from suspected cases in each country, specimens were first tested by cholera rapid diagnostic tests (RDTs) (Crystal VC, Arkray, Healthcare Pvt Ltd., Surat, India) or by Cholkit Ag O1 RDT (Incepta Pharmaceuticals Ltd, Bangladesh). Specimens that produced RDT positive tests were preserved in Cary Blair transport media sent from peripheral health facilities to local and/or regional laboratories and, with the exception of Malawi, spotted directly onto Whatman 903 Protein Saver Card (GE Healthcare Ltd., Forest Farm, Cardiff, UK).

For microbiological confirmation, specimens were streaked directly onto Thiosulfate Citrate Bile Salt sucrose (TCBS) agar and incubated overnight (18-24 hours) at 37°C. Immediately after inoculating the first TCBS plate, a pre-labeled APW vial was inoculated with the specimen and incubated for 4-6 hours at room temperature. After incubation, a second enriched specimen was inoculated on a TCBS plate and incubated overnight (18-24 hours) at 37°C. After overnight incubation, any cholera-like colonies were either tested via classical biochemical testing and polyvalent antisera serotyping [32] or selected with a sterile loop, resuspended in one to two drops of phosphate-buffered saline (PBS), and tested via dipstick or via polyvalent sera agglutination. All agglutination positive and dipstick-positive isolates, as well as any isolates

considered cholera suspect (demonstrating the morphology of a cholera colony), were spotted (~50uL) on filter paper and/or inoculated in AFR1N1 agar (1% tryptone and 1% NaCl) for preservation.

DNA extraction and quantification

Stool spotted filter paper and isolates on filter paper from Kenya, Zambia, Uganda, and Tanzania were sent to Baltimore, MD for extraction and molecular analysis. Each dried filter paper specimen was excised using sterile scissors and placed into a pre-labeled tube. 1 mL sterile 1X PBS was added to each sample tube and incubated for 10 minutes at room temperature. An additional 1 mL of sterile 1X PBS was then added to each sample and samples were immediately centrifuged (14,000 x g for 2 minutes) and the supernatant discarded. Subsequently, 150 µL of a 2% Chelex-100 solution (Bio-Rad) followed by 50 µL of sterile water was added to each sample. The samples were placed in a heating block at 100°C for 8 minutes and then centrifuged (14,000 x g for 2 minutes). The supernatant was transferred to a new microcentrifuge tube and either stored at -20°C or used in a PCR amplification reaction [33,34].

Glycerol stock preserved bacterial isolates received at JHSPH were revived by taking a loop from the received glycerol stock and inoculating 3 mL Luria broth and incubating overnight (18-24 hours) at 37°C. The culture was then streaked onto a TCBS plate using a four-quadrant method and incubated overnight at 37°C. A single colony was selected and inoculated into 3 mL Luria broth and incubated shaking for 8 hours at 37°C. 1 mL of the turbid culture was used as input to Qiagen QIAamp DNA Mini Kit and centrifuged for 5 minutes at 5000 x g (7500 rpm). Volume of the pellet was calculated and Buffer ATL added to a total volume of 180 µL. The remaining steps of extraction were performed according to package directions for DNA purification from tissues.

All extracted nucleic acid was confirmed using conventional PCR to be toxigenic *V. cholerae* [35] and to be serogroup O1 [36]. All DNA, regardless of type or extraction method, were quantified on the Qubit Fluorometer using the dsDNA High Sensitivity Kit according to instrument instructions.

Illumina library construction and sequencing

Illumina library preparation and sequencing was performed at JHSPH. All samples were normalized to 0.6 ng/µL and Illumina sequencing libraries were prepared according to the Nextera DNA Flex Library Prep kit. Library concentrations were measured using the Qubit High Sensitivity DNA Kit, then normalized to a final concentration of 1ng/µL. Sample quality was then assessed on Caliper LabChip and/or Agilent BioAnalyzer, and samples were excluded if they were poor quality or had a concentration below 1ng/µL. All remaining samples were then combined into a single pool. Samples were sequenced on the Illumina NovaSeq platform with 2x150 bp paired-end reads. The entire pool was run in duplicate on an Illumina NovaSeq.

Oxford Nanopore library construction and sequencing

Oxford Nanopore library preparation and sequencing was performed at JHSPH using a starting input amount of 350-1300 ng in 48 µL volume. Libraries were prepared following the SQK-LSK109 library preparation kit from Oxford Nanopore Technologies, with minor modifications as described in Ekeng et al. [25] The final pooled library was eluted in 15 µL Elution Buffer and 190 ng was diluted to 12 µL for loading onto the MinION flow cell. The MinION was run for a total of 48 hours per run and the resulting data was basecalled using Guppy version 3.0.3 with model dna_r9.4.1_450bps_fast.cfg. Adapter removal and demultiplexing was performed with Porechop [37] and reads were filtered using Filtlong [38] with the following options: '--keep_percent 90 --target_bases 800000000.'

Reference-based genome assembly

Illumina paired-end reads were aligned against *V. cholerae* O1 El Tor N16961 (accession: AE003852/AE003853). For each sample, paired-end reads from two lanes were mapped against the

reference genome using BWA version 0.7.17 [39] to produce a SAM file. The SAM files from two lanes were converted to BAM files, sorted and merged using samtools version 1.13-17 [40]. Picard version 2.26.0 [41] was used to mark duplicates in the merged BAM file and variants were called using bcftools version 1.13.35 [42]. Variants were then filtered to obtain VCF files with minimum variant quality score of 20 and a minimum mapping score of 30. Variants were also required to be present in at least 75% of reads mapped and present on both strands (≥2 read depth per stand) [16]. The criteria for inclusion of the genome for subsequent analysis were: (1) median coverage number across all positions on the genome greater than 20; and (2) at least 97% coverage of the reference genome (<%3 ambiguous bases). The complete Illumina genome assembly pipeline used in this study is publicly available at https://github.com/HopkinsIDD/illumina-vc.

For Oxford Nanopore data, reference-based genome assembly was performed as described in Ekeng et al. [25]. The complete Oxford Nanopore genome assembly pipeline used in this study is publicly available at https://github.com/HopkinsIDD/minion-vc.

Visualization of coverage depth across genomes

We compared the reference genome coverage across sample types for all Illumina-sequenced samples. First, we used seqtk version 1.2-r94 [43] to downsample the raw sequencing reads to 1 million reads per sample. We then used samtools version 1.13-17 [40] to calculate the read depth at each position across the genome. To determine the distribution of overall coverage depth by sample type, we calculated the median coverage depth across the genome for each sample. To visualize the distribution of coverage depth for each position in the genome by sample type, we calculated the median, 20th and 80th percentile coverage depth at each nucleotide position across all samples within each sample type group [28]. We then plotted the mean of each of these metrics within a 4000-nt sliding window.

Maximum likelihood estimation

We combined the sequences generated in this study with 1391 publicly available genomes (**Supplementary Data 2**). These genomes represent the majority of *V. cholerae* whole genome sequences publicly available as of July 2021. In most cases, we used reference-based assemblies generated using the same N16961 reference as used in this study [14,16,25]. However, in some cases we had to assemble raw Illumina FASTQ files or contigs. Sequences published as raw Illumina paired-end reads (accession: PRJEB30604) were downloaded and run through the assembly pipeline described above. Sequences published as contigs (accession: PRJNA729102) were processed using snippy version 4.6.0 [44]: snippy was used to produce a BAM file aligned to the N16961 reference, which was then subjected to the variant calling process described above.

After assembling our complete dataset, recombinant sites were masked in genomes generated via both Illumina and Oxford Nanopore platforms, as well as in previously published genomes. Masking was a two-part process as described in Weill et al. [14]: (1) known recombinant regions were masked using a custom GFF file (see https://figshare.com/s/d6c1c6f02eac0c9c871e); (2) genomes were concatenated into a pseudo-alignment and additional sites were masked using gubbins version 2.3.4 [45]. A maximum likelihood tree was then constructed on the masked SNP alignment using IQ-TREE version 1.6.12 with a GTR substitution model and 1000 bootstrap iterations [25]. Figtree version 1.4.4 [46] and R package ggtree version 3.4.0 [47] and treeio version 1.20.2 [48]were used for tree visualization.

Root-to-tip regression identifying outliers

The maximum likelihood trees were examined for the degree of temporal signal using TempEst [49]. 15 outlier sequences whose genetic divergence and sampling date were incongruent under a linear regression of root-to-tip divergence were identified and 11 (previously published) outliers were removed from our

dataset (**Supplementary Fig 1**). Maximum likelihood estimation was rerun as described above, resulting in a final tree with 1498 sequences and a wave 3-specific tree (see below) with 966 sequences.

BEAST analysis for wave 3 samples

Evolutionary and temporal dynamics of wave 3 sequences (see **Supplementary Fig 2**; **Supplementary Data 2**) were reconstructed with a Bayesian phylogenetic approach using Markov chain Monte Carlo (MCMC) available via the BEAST version 1.10.5 package [50] and the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD, USA [51]. A general time-reversible substitution model with gamma-distributed rate heterogeneity (GTR + gamma) was used under a strict molecular clock with a Skyline coalescent prior [14,52]. For sequences with only collection year available, the lack of tip date precision was accommodated by sampling uniformly across a one-year window from 1st January to 30th December.

Ten independent Markov chain Monte Carlo chains with three checkpoints [53] were run for 200 million steps and sampled every 20,000th generation, with at least 10% of the generations discarded as chain burn-in. All analyses were performed using the BEAGLE library to enhance computation speed [54,55]. Convergence and mixing of the chains were inspected using Tracer version 1.7.3 [56]; all continuous parameters yielded effective sample sizes greater than 200. A maximum clade credibility tree was summarized using TreeAnnotator version 1.10.5 [50] and visualization of the tree with annotations was performed with FigTree version 1.4.4 [46]. Distribution of the time to the most recent common ancestor (tMRCA) for lineage AFR13 was performed using TreeStat version 1.10.5 [57].

Geospatial data

Country and sub-country shapefiles were obtained in R from *gadm* version 4.1 [58] . The water shapefile was from World Wild Life global lakes and wetlands database [59] (retrieved from March 3, 2023). R packages *sf* version 1.0-9, *ggspatial* version 1.1.6, *ggthemes* version 4.2.4, *geodata* version 0.4-11, *raster* version 3.6-3, and *geos* version 0.2.2 were used for making maps [60–65] .

ACKNOWLEDGEMENTS

We thank David Mohr for laboratory assistance and the original authors of the sequences used in our phylogenetic analysis (references provided in **Supplementary Data 2**). We also thank Elizabeth C. Lee for help collating cholera case data. Funding for this project was provided by the National Institutes of Health under award number R01HS028634 (A.M.M), K24AI141580 (A.M.M.), and R0AI123422 (A.K.D); Bill and Melinda Gates Foundation INV-047156 (D.S., A.K.D., S.W.); and OPP1195157 (J.L. and S.W.). The opinions expressed in this article are those of the authors and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

REFERENCES

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

2. Morris, J. G., Jr. Cholera and other types of vibriosis: a story of human pandemics and oysters on the half shell. *Clin. Infect. Dis.* **37**, 272–280 (2003).

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

4. Mengel, M. A., Delrieu, I., Heyerdahl, L. & Gessner, B. D. Cholera Outbreaks in Africa. in *Cholera Outbreaks* (eds. Nair, G. B. & Takeda, Y.) 117–144 (Springer Berlin Heidelberg, 2014).

5. Lessler, J. *et al.* Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. *Lancet* **391**, 1908–1915 (2018).

6. Cholera. *Africa CDC* https://africacdc.org/disease/cholera/ (2019).

7. Kanungo, S., Azman, A. S., Ramamurthy, T., Deen, J. & Dutta, S. Cholera. *Lancet* **399**, 1429–1440 (2022).

8. Deen, J., Mengel, M. A. & Clemens, J. D. Epidemiology of cholera. *Vaccine* **38 Suppl 1**, A31–A40 (2020).

9. Zerbo, A., Castro Delgado, R. & González, P. A. A review of the risk of cholera outbreaks and urbanization in sub-Saharan Africa. *Journal of Biosafety and Biosecurity* **2**, 71–76 (2020).

10. Perez-Saez, J. *et al.* The seasonality of cholera in sub-Saharan Africa: a statistical modelling study. *Lancet Glob Health* **10**, e831–e839 (2022).

11. Sack, D. A. *et al.* Contrasting Epidemiology of Cholera in Bangladesh and Africa. *J. Infect. Dis.* **224**, S701–S709 (2021).

12. Multi-country outbreak of cholera, External situation report #7 - 5 October 2023. https://www.who.int/publications/m/item/multi-country-outbreak-of-cholera--external-situation-report--7---5-october-2023.

13. Nakkazi, E. Cholera outbreak in Africa. *Lancet Infect. Dis.* **23**, 411 (2023).

14. Weill, F.-X. *et al.* Genomic history of the seventh pandemic of cholera in Africa. *Science* **358**, 785–789 (2017).

15. Benamrouche, N. *et al.* Outbreak of Imported Seventh Pandemic Vibrio cholerae O1 El Tor, Algeria, 2018. *Emerg. Infect. Dis.* **28**, 1241–1245 (2022).

16. Weill, F.-X. *et al.* Genomic insights into the 2016-2017 cholera epidemic in Yemen. *Nature* **565**, 230–233 (2019).

17. Smith, A. M. *et al.* Imported Cholera Cases, South Africa, 2023. *Emerg. Infect. Dis.* **29**, (2023).

18. Mwaba, J. *et al.* Three transmission events of Vibrio cholerae O1 into Lusaka, Zambia. *BMC Infect. Dis.* **21**, 570 (2021).

19. Hounmanou, Y. M. G. *et al.* Genomic insights into Vibrio cholerae O1 responsible for cholera epidemics in Tanzania between 1993 and 2017. *PLoS Negl. Trop. Dis.* **13**, e0007934 (2019).

20. Hounmanou, Y. M. G. *et al.* Surveillance and Genomics of Toxigenic Vibrio cholerae O1 From Fish, Phytoplankton and Water in Lake Victoria, Tanzania. *Front. Microbiol.* **10**, 901 (2019).

21. Bwire, G. *et al.* Molecular characterization of Vibrio cholerae responsible for cholera epidemics in Uganda by PCR, MLVA and WGS. *PLoS Negl. Trop. Dis.* **12**, e0006492 (2018).

22. Chaguza, C. *et al.* Genomic insights into the 2022–2023*Vibrio cholerae*outbreak in Malawi. *bioRxiv* (2023) doi:10.1101/2023.06.08.23291055.

23. Feikin, D. R., Tabu, C. W. & Gichuki, J. Does water hyacinth on East African lakes promote cholera outbreaks? *Am. J. Trop. Med. Hyg.* **83**, 370–373 (2010).

24. Preparedness, P. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022–2032. https://www.who.int/publications/i/item/9789240046979 (2022).

25. Ekeng, E. *et al.* Regional sequencing collaboration reveals persistence of the T12 Vibrio cholerae O1 lineage in West Africa. *Elife* **10**, (2021).

26. Bénard, A. H. M. *et al.* Whole genome sequence of Vibrio cholerae directly from dried spotted filter paper. *PLoS Negl. Trop. Dis.* **13**, e0007330 (2019).

27. WHO. Cholera data 2000-2022. *Cholera cases officially reported to WHO by Member States from 2000 to 2022* https://worldhealthorg.shinyapps.io/page10cholera_data/.

28. Metsky, H. C. *et al.* Zika virus evolution and spread in the Americas. *Nature* **546**, 411–415 (2017).

29. Mashe, T. *et al.* Highly Resistant Cholera Outbreak Strain in Zimbabwe. *N. Engl. J. Med.* **383**, 687–689 (2020).

30. Kachwamba, Y. *et al.* Genetic Characterization of Vibrio cholerae O1 isolates from outbreaks between 2011 and 2015 in Tanzania. *BMC Infect. Dis.* **17**, 157 (2017).

31. Mutonga, D. *et al.* National Surveillance Data on the Epidemiology of Cholera in Kenya, 1997–2010. *J. Infect. Dis.* **208**, S55–S61 (2013).

32. Chibwe, I. *et al.* Field Evaluation of Cholkit Rapid Diagnostic Test for Vibrio Cholerae O1 During a Cholera Outbreak in Malawi, 2018. *Open Forum Infect Dis* **7**, ofaa493 (2020).

33. Kain, K. C. & Lanar, D. E. Determination of genetic variation within Plasmodium falciparum by using enzymatically amplified DNA from filter paper disks impregnated with whole blood. *J. Clin. Microbiol.* **29**, 1171–1174 (1991).

34. Debes, A. K. *et al.* Clinical and Environmental Surveillance for Vibrio cholerae in Resource Constrained Areas: Application During a 1-Year Surveillance in the Far North Region of Cameroon. *Am. J. Trop. Med. Hyg.* **94**, 537–543 (2016).

35. Nandi Bisweswar *et al.* Rapid Method for Species-Specific Identification ofVibrio cholerae Using Primers Targeted to the Gene of Outer Membrane Protein OmpW. *J. Clin. Microbiol.* **38**, 4145–4151 (2000).

36. Hoshino, K. *et al.* Development and evaluation of a multiplex PCR assay for rapid detection of toxigenic Vibrio cholerae O1 and O139. *FEMS Immunol. Med. Microbiol.* **20**, 201–207 (1998).

37. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* **3**, e000132 (2017).

38. Wick, R. Filtlong. https://github.com/rrwick/Filtlong.

39. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

41. Picard. https://broadinstitute.github.io/picard/.

42. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

43. Li, H. *seqtk: Toolkit for processing sequences in FASTA/Q formats*. (Github).

44. Seemann, T. *snippy: Rapid haploid variant calling and core genome alignment*. (Github).

45. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).

46. FigTree. http://tree.bio.ed.ac.uk/software/figtree/.

47. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

48. Wang, L.-G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and

Associated Data. *Mol. Biol. Evol.* **37**, 599–603 (2020).

49. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007 (2016).

50. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016 (2018).

51. This work utilized the computational resources of the NIH HPC Biowulf cluster. https://hpc.nih.gov/.

52. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).

53. Gill, M. S., Lemey, P., Suchard, M. A., Rambaut, A. & Baele, G. Online Bayesian Phylodynamic Inference in BEAST with Application to Epidemic Reconstruction. *Mol. Biol. Evol.* **37**, 1832–1842 (2020).

54. Suchard, M. A. & Rambaut, A. Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**, 1370–1376 (2009).

55. Ayres, D. L. *et al.* BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. *Syst. Biol.* **68**, 1052–1061 (2019).

56. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

57. Cummings, M. P. TreeStat. *Dictionary of Bioinformatics and Computational Biology* (2004) doi:10.1002/9780471650126.dob1120.

58. GADM. https://gadm.org/.

59. Global lakes and wetlands database. *World Wildlife Fund* https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database.

60. Pebesma, E. Simple features for R: Standardized support for spatial vector data. *R J.* **10**, 439 (2018).

61. Ggspatial. https://paleolimbot.github.io/ggspatial/.

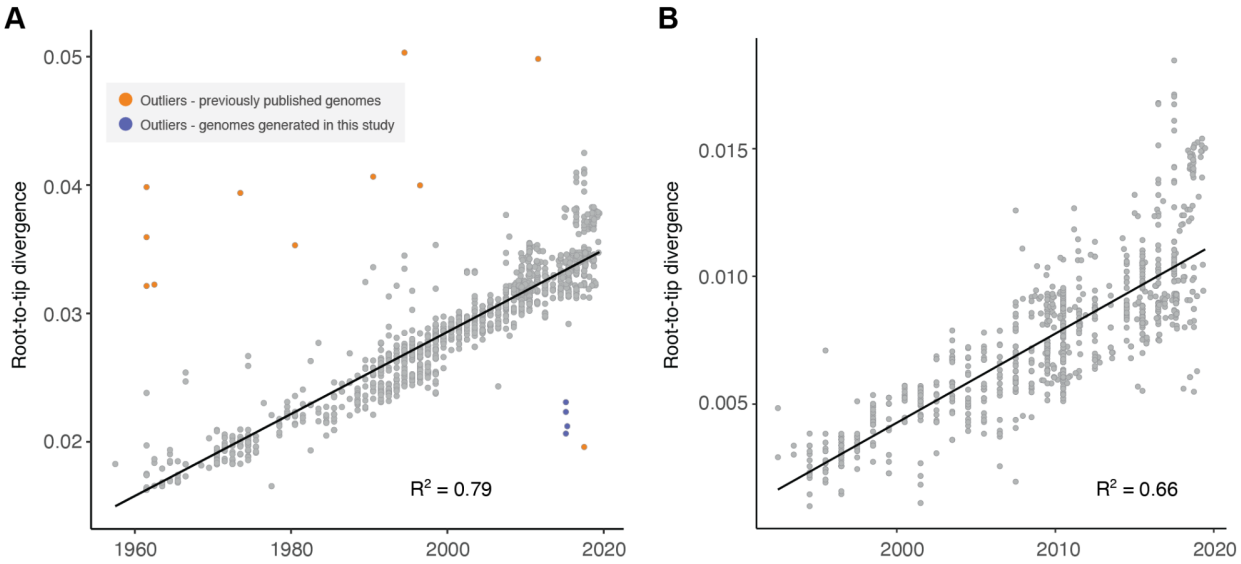
62. Arnold, J. *ggthemes: Additional themes, scales, and geoms for ggplot2*. (Github).

63. *geodata: download geographic data*. (Github).

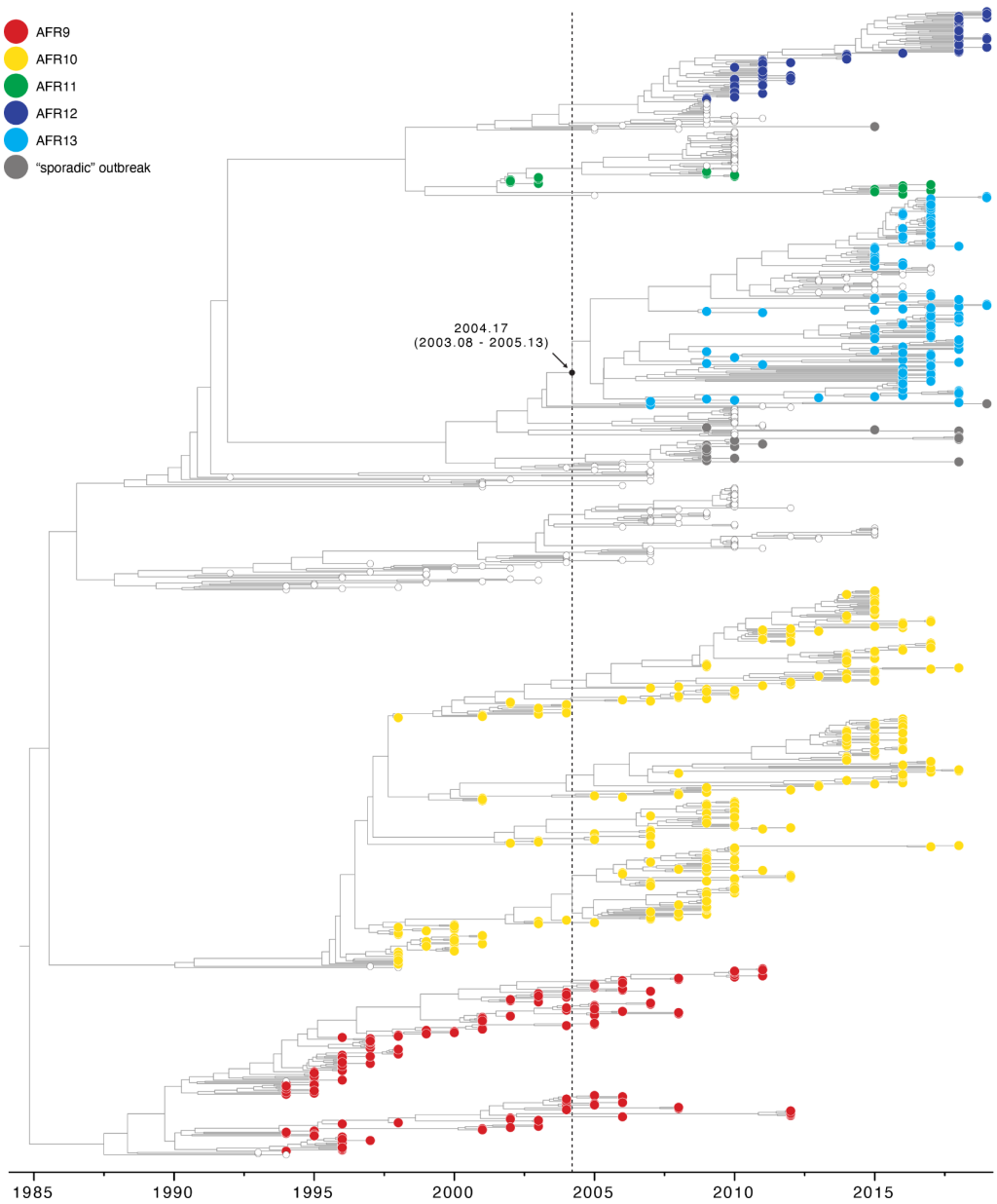
64. Spatial data science with R — R spatial. https://rspatial.org/raster/.

65. Geos. https://paleolimbot.github.io/geos/.

SUPPLEMENTARY FIGURES AND TABLES



Supplementary Figure 1. Molecular clock validation. (**A**) Root-to-tip regression of background genomes plus all 118 genomes generated in this study (n = 1509). Orange dots: outliers identified in previously published genomes; blue dots: outliers identified among the genomes generated in this study, including four genomes generated in this study that were removed due to suspected contamination. Samples represented by gray dots were included in the maximum likelihood tree analysis (N = 1494). (**B**) Root-to-tip regression of 966 wave 3 genomes included in the BEAST analysis (see **Supplementary Data 8**).



Supplementary Figure 2. Wave 3 maximum clade credibility tree. Maximum clade credibility tree depicting time-scaled phylogeny of *V. cholerae* wave 3 (see **Supplementary Data 9**). Tips are colored according to lineages, with non-AFR lineages shown as smaller white tips. Median tMRCA of the AFR13 lineage is indicated on tree (see also **Figure 3C**).

SUPPLEMENTARY INFORMATION

Supplementary Data 1. Sample metadata and sequencing metrics

Supplementary Data 2. Background sequence accessions and metadata

Supplementary Data 3. Cholera cases reported by country and year

Supplementary Data 4. Median depth distribution by sample type from sub-sampling analysis

Supplementary Data 5. Median depth distribution by nucleotide position (4000-nt aggregated) from sub-sampling analysis

Supplementary Data 6. Raw data file of maximum likelihood tree (n = 1494)

Supplementary Data 7. Raw data file for posterior distribution of tMRCA of the AFR13 sub-lineage

Supplementary Data 8. Raw data file for root-to-tip divergence of all genomes (n = 1509) and wave 3 genomes (n = 966)

Supplementary Data 9. Raw data file BEAST MCC phylogeny of wave 3 genomes (n = 966)