Multicountry genomic analysis underscores regional cholera spread in Africa

Gerald Mboowa[1#], Nathaniel Lucero Matteson[2#], Collins Kipngetich Tanui[1#], Mpanga Kasonde[3], Guyguy Kusanzangana Kamwiziku[4,5], Olusola Anuoluwapo Akanbi[6], Jucunú Johane Elias Chitio[7], Mathews Kagoli[8], René Ghislain Essomba[9,10], Alisen Ayitewala[11], Isaac Ssewanyana[11], Valentina Josiane Ngo Bitoungui[9,12], Adrienne Aziza Amuri[5,13], Andrew S Azman[14,15,16], Olajumoke Atinuke Babatunde[6], Blaise Mboringong Akenji[9], Anaïs Broban[17], Espoir Bwenge Malembaka[14,18], Francis Ongole[11], Chimaobi Emmanuel Chukwu[6], Nália Ismael[7], Otridah Kapona[3], Osvaldo Laurindo[7], Placide Kingebeni Mbala[4,5], Georges Alain Etoundi Mballa[19], Imelda Carlos Zulfa Miambo[7], Alex Ansaye Mwanyongo[8], Grace Najjuka[11], Joseph Mutale[3], Kunda Musonda[3], Allan Muruta Niyonzima[20], Mirriam Ethel Nyenje[8], Michael Popoola[6], Doreen Mainza Shempela[21], Christiane Medi Sike[10,22], Sofião Manjor Sitoe[7], Dorcas Waruguru Wanjohi[1], Placide Okitayemba Welo[23], Mtisunge Yelewa[8], Sebastian Yennan[6], Lucius Ziba[8], CholGEN Consortium, Joseph Ephram Bitilinyu-Bangoh[8], Roma Chilengi[3], Hamsatou Hadja[9,24], Jide Idris[6], José Paulo Maurício Langa[7], Daniel Mukadi-Bamuleka[25], Susan Nabadda[11], Amanda K Debes[26], David A Sack[26], Jean Kaseya[1], Yenew Kebede Tebeje[1], Shirlee Wohl[2*], Sofonias Kifle Tessema[1*]

Affiliations

[1] Africa Centres for Disease Control and Prevention, Addis Ababa, Ethiopia
[2] Division of Infectious Diseases, Brigham and Women's Hospital, Boston, United States
[3] Zambia National Public Health Institute, Lusaka, Zambia
[4] Department of Medical Biology, University of Kinshasa, Kinshasa, Democratic Republic of the Congo
[5] Institut National de Recherche Biomédicale, Kinshasa, Democratic Republic of the Congo
[6] Nigeria Centre for Disease Control and Prevention, Abuja, Nigeria
[7] Instituto Nacional de Saúde, Marracuene, Mozambique
[8] Public Health Institute of Malawi, Lilongwe, Malawi
[9] National Public Health Laboratory, Ministry of Public Health, Yaoundé, Cameroon
[10] Faculty of Medicine and Biomedical Sciences, University of Yaounde 1, Yaoundé, Cameroon
[11] National Health Laboratory & Diagnostic Services (NHLDS/CPHL), Ministry of Health, Kampala, Uganda
[12] Faculty of Medicine and Pharmaceutical Sciences, University of Dschang, Dschang, Cameroon
[13] Faculty of Medicine, University of Kinshasa, Kinshasa, Democratic Republic of the Congo
[14] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, United States
[15] Geneva Centre for Emerging Viral Diseases, Geneva University Hospitals, Geneva, Switzerland
[16] Division of Tropical and Humanitarian Medicine, Geneva University Hospitals, Geneva, Switzerland
[17] Epicentre, Paris, France
[18] Center for Tropical Diseases and Global Health, Université Catholique de Bukavu, Bukavu, Democratic Republic of the Congo
[19] Department for the Control of Diseases, Epidemics and Pandemics, Ministry of Public Health, Yaoundé, Cameroon
[20] Integrated Epidemiology, Surveillance and Public Health Emergencies, Ministry of Health, Kampala, Uganda
[21] Churches Health Association of Zambia, Lusaka, Zambia
[22] Hôpital Laquintinie de Douala, Douala, Cameroon
[23] National Program of Elimination of Cholera, Ministry of Health, Kinshasa, Democratic Republic of the Congo
[24] Yaoundé General Hospital, Yaoundé, Cameroon
[25] Rodolphe Merieux Laboratory INRB-Goma, Goma, Democratic Republic of the Congo
[26] Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, United States

[#] Contributed equally
[*] Jointly supervised this work
Correspondence: SofoniasT@africacdc.org (S.K.T.), swohl@bwh.harvard.edu (S.W.)
CholGEN Consortium authors are listed in Supplementary Information

ABSTRACT
Cholera remains a significant public health burden in many countries in sub-Saharan Africa, though the exact mechanisms of bacterial emergence and spread remain largely undefined. We generated genomic data from 728 *Vibrio cholerae* O1 isolates predominantly collected between 2019-2024 to create the largest dataset of *V. cholerae* genomes sequenced locally in Africa. This dataset enabled us to interrogate recent patterns of spread, including the rapid circulation of the AFR15 lineage associated with unusually large outbreaks in Southern Africa. We provide evidence for the movement of the AFR15 lineage into new African Member States and confirm previously observed differences in *V. cholerae* transmission dynamics in West versus East Africa, though cross-border transmission is prevalent on both sides of the continent. Despite observed differences, evolutionary processes are similar across lineages and we find no evidence for significant changes in antimicrobial resistance genotypes. Overall, our findings emphasize the importance of regionally coordinated cross-border surveillance and interventions, while also demonstrating the critical role of locally generated genomic data in understanding the spread of cholera in Africa.

INTRODUCTION
Cholera, a diarrheal disease caused by the bacterium *Vibrio cholerae*, poses a significant public health threat globally. There have been seven major global outbreaks (pandemics) of cholera, with the current outbreak primarily affecting and persisting in Africa since its introduction from Asia in 1970 [1,2]. Seventh pandemic cholera is typically caused by the *V. cholerae* O1 serogroup and El Tor biotype (referred to as 7PET *V. cholerae*), and several prior studies have documented the intermittent and often seasonally driven outbreaks observed in many African countries [3–5]. These recurring outbreaks underscore the ongoing challenges in controlling and preventing cholera, particularly in resource-limited settings, and emphasize the need for comprehensive control strategies that use all available tools to better understand the epidemiology and dynamics of cholera transmission.

Severe, protracted, and persistent cholera outbreaks in several countries necessitate improved cholera surveillance and control [6]. Over 92,000 cholera cases were reported across 16 African countries from January to May 2024 [7], with Zambia (20,113 cases), the Democratic Republic of the Congo (DRC; 16,539 cases), and Mozambique (7,762 cases) among the most affected [8]. Mozambique and Malawi also experienced severe outbreaks in 2022 [9], with the outbreak in Malawi recorded as one of the largest (over 56,000 cases) and deadliest (>3% case fatality rate) outbreaks in history [10,11]. Extreme weather events and the COVID-19 pandemic, with its disruptions to healthcare systems and changes in societal behaviors, likely contributed to the high morbidity and mortality of these outbreaks. However, simultaneous outbreaks in neighboring countries have been previously observed in this region and across Africa, suggesting that cholera transmission is regional and crosses borders, which has critical implications for its control and elimination [12,13].

Investigating transmission dynamics and unusual disease presentation (for example, high case fatality rates) requires an in-depth look at not only epidemiological and clinical data, but at the *V. cholerae* bacterium itself. Whole genome sequencing provides information about the specific strain(s) circulating, which can determine whether outbreaks in neighboring countries are connected, or if recent changes in the bacterial genome may explain observed trends. As a result, genomic analysis has recently emerged as an important tool in understanding the spread of cholera in sub-Saharan Africa [2]. Using whole genome sequences generated from *V. cholerae* isolates around the continent, previous studies first identified at least three waves of global transmission from Asia to Africa during the seventh pandemic [14], which was later divided into at least 17

independent introductions of 7PET into Africa, deemed the AFR1-AFR17 (or equivalently known as T1-T17) lineages [2,10,15–17]. These lineages have since been used to connect recent African and Middle Eastern outbreaks [16], to demonstrate regional transmission patterns in both East [17–19] and West Africa [13], and to suggest a possible transmission route underlying recent cases [10].

Although genomic analyses have been effectively used to better understand cholera transmission on the African continent, previous work has also highlighted the gaps in our current understanding and the need for additional surveillance and analyses. In particular, most published work focuses on outbreaks prior to 2020, when genomic surveillance infrastructure in Africa was limited. Therefore, an updated picture of *V. cholerae* diversity is needed, including how transmission patterns have or have not changed in recent years, the role of environmental reservoirs in transmission, and whether changes in the *V. cholerae* bacterium may have contributed to the worst cholera outbreaks in decades. There are also growing concerns over the emergence of antibiotic resistance to drugs used to manage severe cases of cholera and other infectious diseases [20], especially because bacteria can share resistance genes via horizontal gene transfer [21]. Finally, questions remain about the ability of genomic data to provide information that can meaningfully be used to inform and prioritize intervention approaches such as organizing vaccination campaigns.

In response to these unanswered questions, the Africa Centres for Disease Control and Prevention (Africa CDC), through the Africa Pathogen Genomic Initiative, formed the Cholera Genomics Consortium in Africa ("CholGEN") in collaboration with national public health institutes and national reference laboratories from seven African Union (**AU**) Member States (Cameroon, DRC, Malawi, Mozambique, Nigeria, Uganda, and Zambia) and international partners [22]. These Member States were selected based on public health needs, represent distinct cholera transmission scenarios (primarily endemic versus primarily caused by imports), and have collectively observed all known lineages in Africa. Additionally, several of these countries share borders, which will aid in understanding the role of cross-border transmission. As part of CholGen, laboratorians and bioinformaticians from each of these member states have been working to leverage local genomic sequencing capacity built during the COVID-19 pandemic to sequence bacterial isolates from recent cholera outbreaks. The resulting data provide an updated picture of circulating cholera diversity, point to specific instances of likely cross-border transmission, and highlight the role of genomic data in understanding different transmission scenarios. These results, made possible by a highly collaborative effort, underscore the importance of multi-country solutions to cholera control, which should include coordinated genomic surveillance that can reveal both cross-border transmission events and notable changes in the bacterial genome.

RESULTS

Through CholGEN, the seven AU Member States identified, processed, and sequenced 1,288 isolates in-country (**Table 1**, **Figure 1A**), resulting in a final dataset of 728 high-quality genomes included in downstream analyses (**Supp. Data 1**; see Methods for exclusion criteria). Each participating Member State selected cholera samples for sequencing based on geographical distribution and completeness of metadata, to obtain a representative picture of national and regional cholera spread. The isolates sequenced by CholGEN represent the vast majority of cholera genomes sequenced in Africa since 2019 and the largest dataset of *V. cholerae* genomes sequenced locally in Africa (**Figure 1B**). This density of recent data enabled the subsequent analyses presented below, which may help guide targeted public health interventions tailored to the transmissibility and persistence of specific outbreaks.

	Total sequences	High quality	Included in analysis
Cameroon	94	36	31
DRC	216	155	153
Malawi	136	70	69
Mozambique	108	49	37
Nigeria	293	132	131
Uganda	73	49	47
Zambia	300	237	235
Total	*1220*	*728*	*703*

Table 1. Number of *V. cholerae* genomes collected and analyzed from each of the CholGEN AU Member States. High quality represents sequences which had >90% reads mapped to the *V. cholerae O1* reference and <10% ambiguous nucleotides. "Included in analysis" refers to high-quality sequences that had a collection date available and were not excluded by clock-rate or phylogenetic placement filters (see **Supp. Data 1**).

To assess whether the most recent cholera outbreaks were derived from the known diversity of *V. cholerae* in Africa or novel introduction events, we performed a Bayesian phylogenetic reconstruction of the third wave of 7PET, which includes all lineages known to be currently circulating in sub-Saharan Africa (AFR9-AFR17 [2]). The reconstruction included 728 high-quality genomes in this dataset and 1,764 publicly available whole genome sequences from Africa and elsewhere (**Figure 2A**, **Supp. Data 2**). We assigned newly generated sequences to lineages using their phylogenetic placement and found that they all descended from previously described introductory events (**Figure 2B**). Consistent with prior analyses, our data indicated that different regions of Africa have distinct transmission patterns. In Western and Central African Member States—Cameroon, DRC, and Nigeria—our data showed that lineages AFR10 (DRC) and AFR12 (Nigeria, Cameroon) were still circulating, consistent with previous reports [13,23]. This pushes the persistence of AFR10 and AFR12 in the region to at least 29 and 15 years, respectively. Our data also supported previous studies showing that Eastern and Southern Africa continue to maintain a greater diversity of circulating lineages [17,24,25]. We observed AFR10, AFR11, AFR13, and AFR15 circulating in the region, with two or more lineages often present in a country over a five-year period.

Many of the newly generated genomes were classified as AFR15 and may therefore provide updated insight into the emergence of this lineage, which was previously linked to the unusually large outbreak in Malawi and major outbreaks in the Middle East, South Africa, and Zimbabwe [10,26]. In addition to these locations, we identified AFR15 isolates in DRC, Mozambique, and Zambia, indicating that AFR15 continues to spread rapidly across Southern Africa and has now been introduced into Central Africa (**Figure 2B**). The newly identified presence of this lineage in DRC is likely linked to the outbreak in Zambia, since all AFR15 sequences from DRC were collected in the province of Haut Katanga, a province on the Zambian border. We also observed cholera transmission between these two countries elsewhere in the phylogeny (AFR10), suggesting that movement across national boundaries plays an important role in the maintenance of cholera in this region [27].

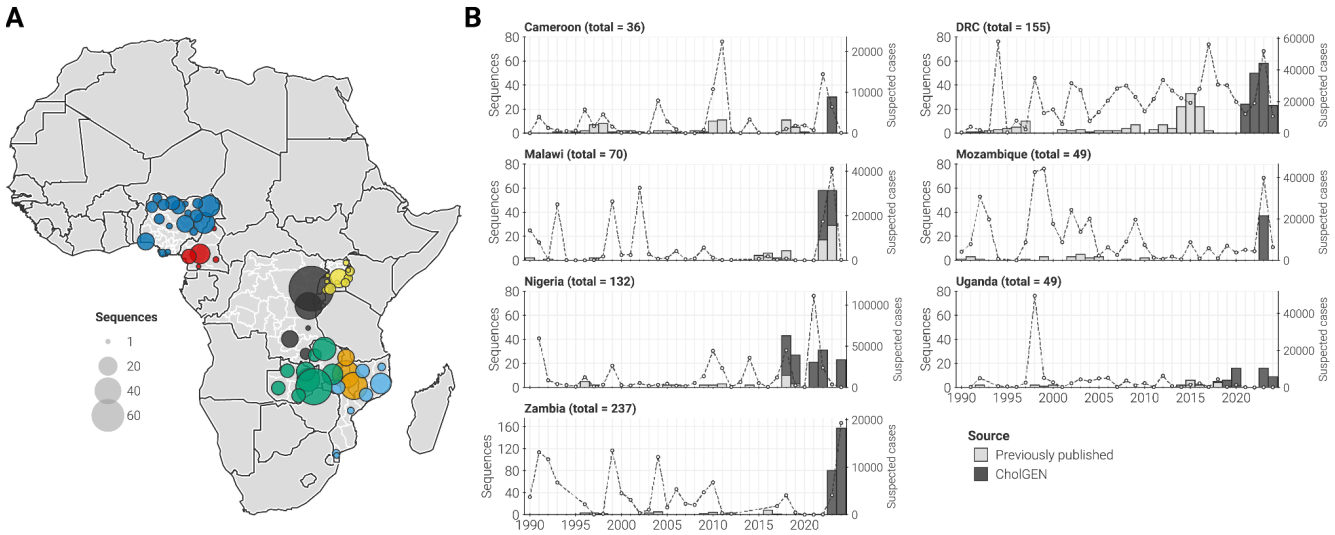


Figure 1. Genomic surveillance of *Vibrio cholerae* in seven AU Member States. (**A**) Map of Africa showing the distribution of sampling location for genomes generated. Points, located at the first-level administrative division, are colored by country of collection and sized by the number of high-quality genomes generated from that specific location. (**B**) Genomes generated per country and per year (left axis). Light gray bars: number of genomes previously published. Dark gray bars: number of genomes generated by this study. Secondary (right) axis: annual number of cholera cases reported to the World Health Organization.

Despite the likely importance of cross-border transmission, our phylogenetic analysis also suggested that there have been multiple introductions of AFR15 into Africa, rather than a single introduction identified by prior investigations [10,26,28] (**Figure 2A**). Our analysis showed that most AFR15 isolates descended from an introduction from Southeast Asia in late-2019 (95% highest posterior density **[HPD]**: February 2019 to June 2020), while a smaller subset of 11 isolates originated from an introduction from the Middle East in mid-2022 (95% HPD: December 2021 to December 2022; inset of **Figure 2A**), coinciding with outbreaks in Iraq, Lebanon, Syria, and Pakistan [26,28–30]. Isolates from all six AU Member States associated with the AFR15 lineages were linked to the earlier and larger introduction, while only Mozambique and Zimbabwe isolates descended from the later introduction. Generation and analysis of additional sequences from these AFR15 sub-clusters will reveal if major outbreaks can be attributed to one or multiple introductions.

Limited genotypic differences between lineages

We observed that the ancestral branch leading to the AFR15 lineage was long, not just in terms of time but also in terms of genetic changes (**Figure 2A** & **Supp. Figure 1**), calling into question whether large mutational changes or altered substitution rates were associated with the emergence of new strains in Africa (**Figure 2A**). To determine whether high rates of mutation accumulation accompanied the emergence of any of the lineages, we assessed whether any lineages were more divergent than expected given their sampling dates, and found that genomes from lineages observed to be currently circulating in CholGEN Member States (AFR10, AFR11, AFR12, AFR13, and AFR15) did not deviate significantly from the expected number of substitutions (**Supp. Figure 2**). We also estimated the mean substitution rate for each lineage using an uncorrelated relaxed clock rate model (see Methods), and found that all lineages share a similar substitution rate distribution (**Supp. Figure 3**). Therefore, we concluded that evolutionary processes are similar across lineages despite differences in geographic range, and that other factors must be responsible for the rapid spread of the AFR15 lineage.

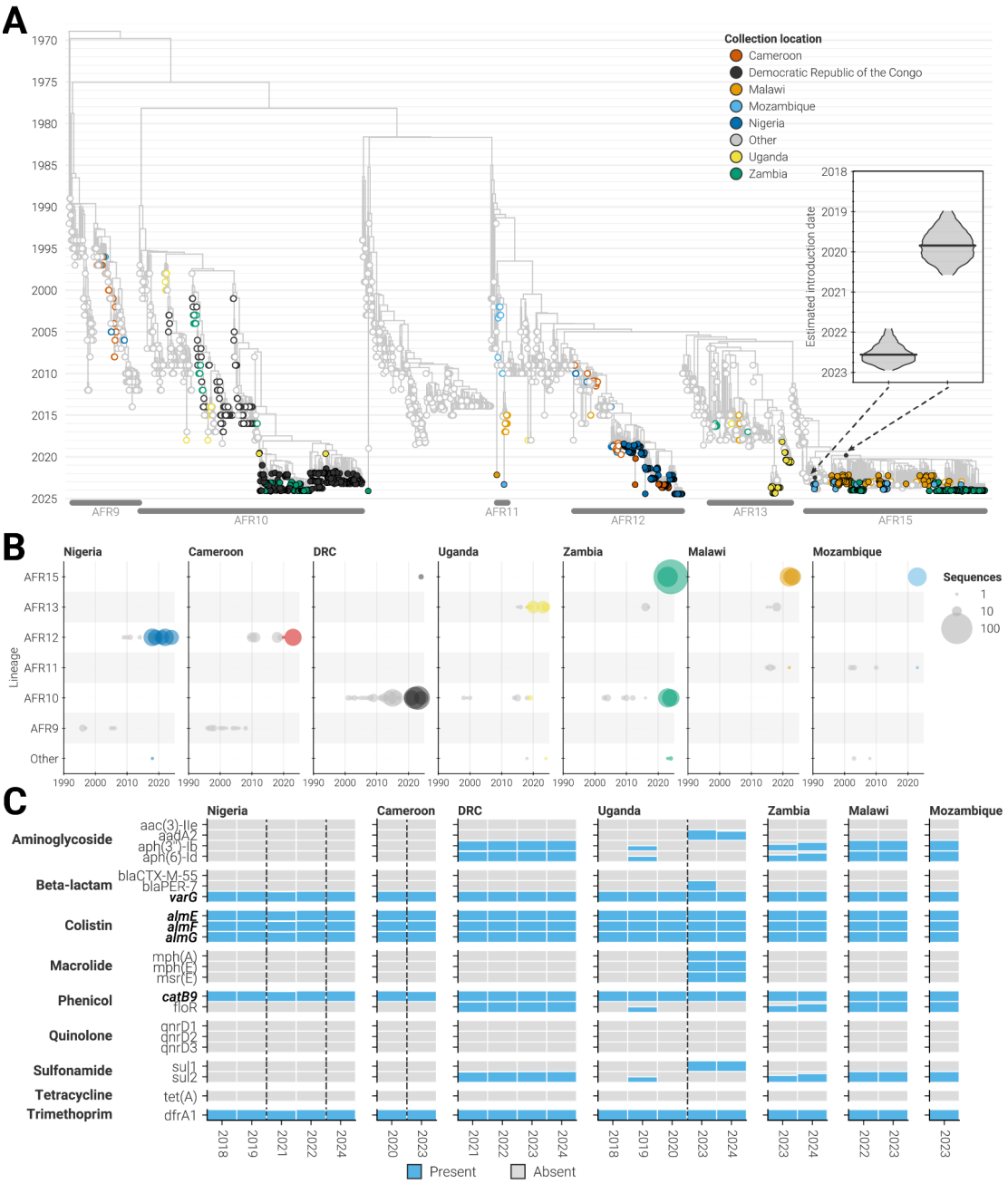


Figure 2. Phylogenetic diversity of cholera. (**A**) Maximum clade credibility tree of the third wave of 7PET *V. cholerae*. Taxa and internal nodes are colored according to their collection locations. Filled circles: genomes generated by CholGEN Member States. Open circles: previously published genomes. Black dots on nodes indicate the two putative introductions of AFR15 into Africa. Inset shows the posterior estimated date of each putative introduction. Lineages with less than 20 sequences are not labeled. (**B**) The annual number of genomes generated from each of the CholGEN Member States according to which inferred lineage they descend from. Size of circle scales is based on the number of high-quality genomes generated. Color of circle describes whether the genomes were previously published (gray) or generated by this study (other colors). Countries are ordered left-to-right based on their geographic location West-to-East. Sporadic lineages with less than 20 sequences are combined as "Other." (**C**) The proportion of isolates collected in each year and country that carry specific antimicrobial resistance genes. Blue bars: gene present. Gray bars: gene not observed. Countries are ordered as in panel B, and genes are ordered by the class of antibiotic that they provide resistance to. Bolded genes: genes located on either genomic chromosome of *V. cholerae*.

The spread of antibiotic resistance may also be associated with changes in the epidemiological patterns of bacterial spread [31] and is essential to informing best practices for cholera treatment when outbreaks occur. To provide an updated picture of the spread of resistance, we examined all isolate genomes for the presence of AMR genes and found that the AMR profiles of isolates from most CholGEN Member States did not change throughout our sampling period (**Figure 2C**), even when we looked at a fine temporal resolution (**Supp. Figure 4**). Certain resistance-associated genes, particularly those that are canonically present in the *V. cholerae* core genome (*almE*, *almF*, *almG*, *catB9*, and *varG*), were present in all isolates from all countries. Other genes were associated with the lineage circulating in a region, including *floR, sul2*, and aminoglycoside-phosphotransferase genes that are all absent in AFR12 and AFR13 strains [32]. We did not observe any genes associated with resistance to tetracyclines, which is the primary class of antibiotics used for the treatment of cholera [33]. However, several countries observed phenotypic changes in resistance during the time period associated with these samples [21,34], suggesting the need for additional studies comparing genotypic and phenotypic AMR results [35], as well as continued monitoring of resistance in affected regions.

We found a single exception to the above in our dataset. We observed that isolates from Uganda gained several AMR genes from 2020 to 2023, including *aad2*, *blaPER-7*, *mph(A)*, *mph(E)*, *mrs(E)*, and *sul1* (**Figure 2C**). This was not due to the introduction of a new lineage because only AFR13 was observed in Uganda during this period (**Figure 2B**). AMR in cholera is often associated with the acquisition of plasmids containing multidrug resistance elements, so we cross-referenced the AMR genes to determine if they were present in the *V. cholerae* genome or on mobile genetic elements. We found that the AMR genes acquired by the isolates in Uganda by 2023 were not on the core genome, and were all found on the IncA/C plasmid, which is known to be a primary source of multidrug resistance for cholera [36,37]. This plasmid has been found sporadically in historical 7PET isolates (**Supp. Figure 5**) and was not found in isolates from Uganda collected in 2018-2019, though it was observed in isolates from the 2018-2019 AFR13 outbreak in Yemen and in three Lebanese isolates that phylogenetically clustered with the 2023 isolates from Uganda [28,38]. Phylogenetic placement of 2023 sequences from Uganda containing the plasmid (which were collected from districts along the shore of Lake Victoria) suggest a recent acquisition (**Figure 2A**). These findings point to possible intercontinental spread of cholera strains carrying IncA/C and support the connection between outbreaks in the Middle East and Africa.

Cross-border transmission is frequent but rarely observed

The results above suggest that transmission patterns, rather than genotypic differences, may be crucial for understanding and preventing future cholera spread. The recency of collected isolates and close proximity of the CholGEN AU Member States, including those with ongoing AFR15 outbreaks, enabled us to take a close look at the location and frequency of transmission of AFR15 and other lineages across international borders. These results, which include identification of numerous examples of cross-border transmission across all lineages currently circulating in Africa, point to the importance of regional coordination for both surveillance efforts and outbreak interventions such as vaccination campaigns.

To closely examine cross-border transmission, we quantified the timing and number of geographic transitions (also called Markov jumps [39]), between AU Member States across the full posterior of a Bayesian phylogeographic reconstruction. While we found multiple prominent examples of cross border transmission in the phylogeny (**Figure 3A**), our initial analysis suggested that these transitions were infrequent during the third wave of 7PET (average rate of 0.04 location transitions per year; 95% HPD: 0.03-0.05 transitions per year). However, historical sampling of isolates has been inconsistent and likely biased, and we found that location

transitions were unevenly distributed over time (**Figure 3B**). Closer inspection revealed that the timing of location transitions was moderately correlated with sampling (R^2 value = 0.48; p-value < 0.001; **Figure 3C**). This result suggests that international transmissions may be more frequent and consistent than what was captured in the dataset, but only observed whenever genomic surveillance was adequate. Therefore, we expect that increased surveillance may identify locations with yet-unsampled outbreaks, allowing for more effective targeting of containment efforts to locations that are continuously reseeding transmission.

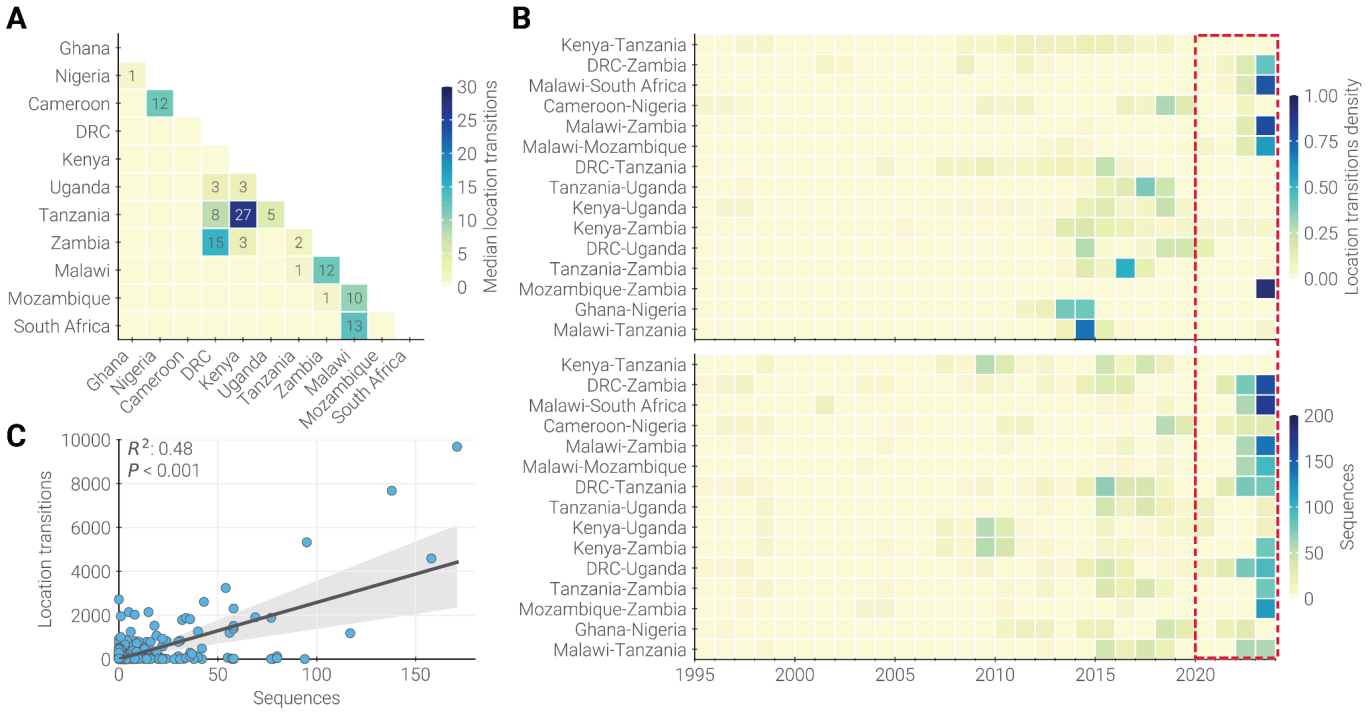


Figure 3. Dynamics of cross-border transmission. (**A**) Median number of location transitions observed between each pair of countries across the posterior distribution of trees from the discrete state analysis. Countries were included as states in the discrete state reconstruction if they had more than 20 cholera genomes in our dataset (see Methods). Only locations with a median number of location transitions greater than 0 are labeled. Countries are ordered west-to-east, north-to-south, and so that neighboring countries are next to each other. (**B**) Top panel: annual density of location transitions for each location pair with median number of location transitions greater than 0 across the posterior distribution of trees. Order of locations in each pair is alphabetical and does not reflect directionality. Bottom panel: annual number of genomes available for each location pair. Red box highlights period of increased sampling (due to CholGEN) corresponding to higher density of observed location transitions. (**C**) Correlation between the number of genomes available and the magnitude of location transitions for a given year and location pair. R^2 and P value were determined using ordinary least squares regression.

Future surveillance should be directed towards high diversity locations

Our results indicate that sampling plays an important role in observing cholera transmission, and suggest that moving towards routine genomic surveillance may support international cholera elimination goals by revealing transmission patterns that are currently missed. Therefore, we attempted to estimate un-captured transmission and determine if additional surveillance is equally needed everywhere or if there are certain lineages, countries, or regions that are most in need of additional sampling. To do so, we developed a metric to assess the value of generating a new genome that involves evaluating the sampling strategy, phylogenetic diversity, and total information captured by sequencing.

We took this holistic approach to assessing the value of generating new sequences because both sampling strategy and underlying transmission patterns influence what can be learned from sequencing. To evaluate the impact of the sampling strategy, we built on a technique commonly used in ecology [40,41] and estimated how many new mutations would be expected when adding a new genome to the analysis, referred to here as phylogenetic diversity [42,43]. While this value provides an estimate of the potential information gained from additional sequencing, it must also be considered in the context of the total number of sequences previously generated. We also recognize that sequencing can still be useful in cases of limited phylogenetic diversity if there is other information to be gained, for example if additional sequencing may reveal new introductions into a country or region. Therefore, we used a rarefaction approach [44,45] to quantify how much we would likely learn about introductions from additional sequencing. We applied this comprehensive metric to data from each CholGEN Member State and saw that the amount of diversity and information gained per new genome varied between locations and surveillance intensities (**Figure 4A-G**). Below, we focus on three examples—Uganda, DRC, and Zambia—that illustrate the range of phylogenetic diversity captured and additional information revealed given the differing surveillance strategies employed.

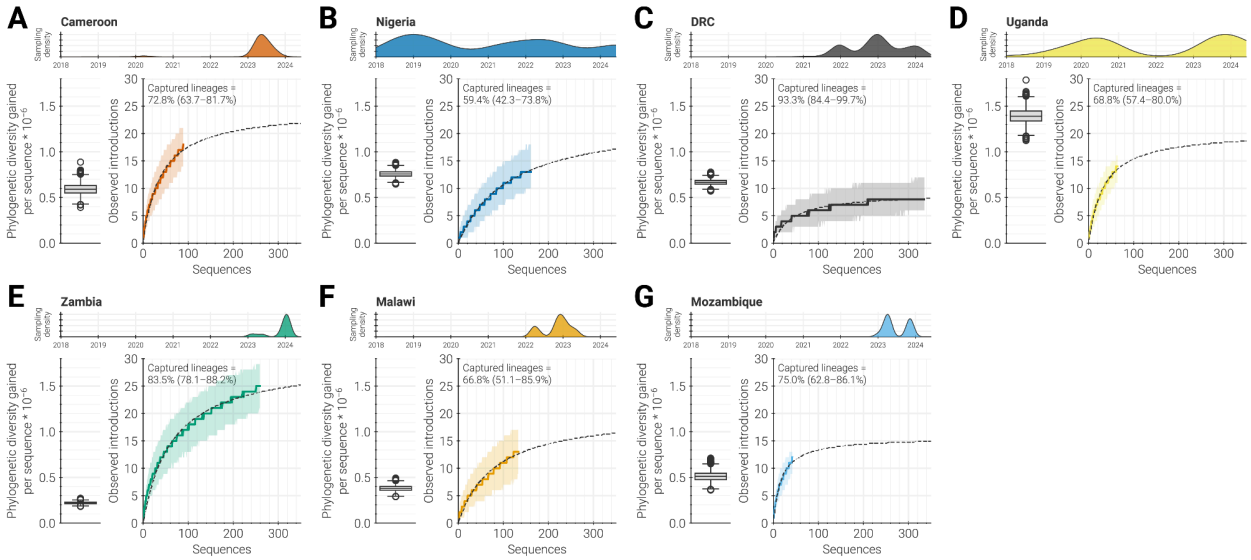


Figure 4. Assessing current surveillance. For each CholGEN Member State (**A-G**), metrics that estimate the value of sequencing new *V. cholerae* isolates. Top: temporal distribution of genomic surveillance conducted by CholGEN within each Member State. Density curves for each country are normalized so that they have the same maximum value (see **Figure 1B** for actual values). Lower left: substitutions per site gained by each genome sequenced by CholGEN (distribution across each tree in the posterior distribution of the phylogeographic analysis). Lower right: rarefaction curves describing the rate at which introductions into the country are observed given the number of genomes collected. A solid line and shaded area represents the rarefaction curve and its uncertainty, while the dashed black line represents the Michaelis-Menten model fit. The percentage of total introductions that were observed and 95% confidence intervals are recorded in the upper right corner of each country's panel. Distributions for rarefaction curves and percentages of total introductions that were observed represent 1000 random trials wherein a tree was independently sampled from the posterior and a rarefaction curve calculated.

First, new genomes from Uganda revealed marked phylogenetic diversity and numerous introductions (**Figure 4D**). This result is expected, as CholGEN samples from Uganda cover multiple years and longer sampling periods should capture more diversity assuming a constant clock rate [46]. It also aligns with transmission dynamics previously described in Uganda, where repeated introductions (rather than local spread) sustain cholera and thereby increase phylogenetic diversity [24,47]. These findings suggest that continued surveillance

using a similar strategy, where genomic data is consistently collected over time, would likely reveal further introductions and greater diversity. This could help identify additional sources of cholera or new transmission routes that could be targeted for intervention.

In contrast to Uganda, new genomes from DRC revealed relatively less phylogenetic diversity and fewer introductions, though we estimated that there are far fewer introductions into DRC than into Uganda (**Figure 4C**). This is likely due to distinct transmission patterns in DRC, where cholera is highly persistent (**Figure 2B**). These results suggest that further surveillance may uncover more phylogenetic diversity, but it is unlikely to reveal new introductions without a change in the surveillance strategy or overall transmission pattern within the country. Therefore, routine surveillance will most likely provide a representative view of the situation in DRC without requiring intensive sampling of every outbreak. However, it will be important to ensure that sampling in DRC is representative across space in addition to over time. For instance, if surveillance is not geographically representative, it may be useful to first capture geographic gaps to ensure lineages are not being missed before finalizing a surveillance strategy.

In contrast to Uganda and DRC, new genomes from Zambia showed comparatively low levels of phylogenetic diversity. However, Zambia is an example of a situation in which phylogenetic diversity was low but the potential to capture additional information (for instance, number of introductions) was very high, as new data uncovered many additional introductions despite genetic similarity (**Figure 4E**). This finding aligns with the relatively limited range of collection dates for Zambia sequences generated by CholGEN, as well as with evidence for the close transmission relationships of Zambia with Malawi, DRC, and other locations. To better capture the diversity of circulating lineages and the locations that seed cholera in Zambia, additional surveillance would need to be conducted over a longer period. Additionally, during periods of rapid spread by a single lineage, resources may be better spent capturing temporal or geographic diversity rather than sequencing large numbers of likely identical sequences from concentrated areas or clonal outbreaks.

DISCUSSION

In this manuscript, we highlight how the CholGEN initiative led to the enhancement of local cholera genomic surveillance in several AU Member States, leading to the largest dataset of high-quality *V. cholerae* genomes sequenced locally in Africa. These genomes fill critical gaps in our understanding of recent cholera circulation, revealing distinct transmission patterns across different African regions and highlighting the rapid spread of the AFR15 lineage, including its expansion into additional Member States like the DRC. Despite the association of AFR15 with unusually large outbreaks in Malawi, Zambia, and beyond, we did not identify significant genotypic differences compared to other lineages previously identified in the region. Antimicrobial resistance signatures were also consistent across lineages and outbreaks. We found evidence of cross-border transmission between neighboring countries, and our data suggests that these events may be more frequent than currently recognized, which could have significant implications for how local experts develop surveillance strategies and cholera control efforts.

The proximity of CholGEN Member States and their distribution across Africa allowed us to compare the spread of cholera within and between different regions, particularly for the most recent outbreaks on the continent. This work provides further evidence that cholera readily spreads within certain regions of Africa, indicating that cross-border transmission has a prominent role in the maintenance of cholera on the continent. That said, these analyses clearly show that sampling strategy plays a significant role in the interpretation of phylogenetic analyses, and that our ability to comment on the directionality or absolute number of international

transmissions is limited. Refining our understanding of cholera transmission will require continuous and routine genomic surveillance, both during and between cholera outbreaks. This will both fill in gaps in our current understanding and provide a more systematic dataset that can inform future prioritization through the sequencing information metrics described in **Figure 4** above. By evaluating these metrics in the context of the transmission patterns present in a country, surveillance efforts can be more effectively tailored to meet specific needs and optimally allocate resources.

It is important to recognize, however, that the real barriers to sampling are often not in the strategy, but in the realities of sample collection, transport, sequencing, and pairing with epidemiological metadata. Although all sequencing described in this manuscript was performed from cultured isolates, we were only 60% effective in generating high-quality whole genome sequences. This is likely due to a combination of isolate contamination, storage conditions not being conducive to preserving bacterial cultures, and other challenges associated with multi-step processes like whole genome sequencing. To enable better surveillance of cholera and other local pathogens, we need to improve the full pathway leading to sequencing, from sample collection and *V. cholerae* identification through sample transport and laboratory processing. Additionally, it may be important to explore sequencing methods that do not rely on bacterial culture, such as direct sequencing from stool [17,48]. This will increase the total number of samples that can be selected for sequencing and open the door to improved environmental surveillance including viable but nonculturable specimens that may play an important role in understanding transmission [49].

Even once sequencing challenges are addressed, there are important next steps to take in terms of translating findings to public health impact. For example, we found that isolates in Uganda acquired a plasmid associated with AMR, but without phenotypic data paired to these samples, it is difficult to understand the implications of such an acquisition. Further in-depth studies comparing antimicrobial resistance phenotype to genotype [35] are needed to assess which genotypes correlate with actual antibiotic resistance, and which phenotypic changes are caused by *de novo* evolution of resistance-conferring mutations versus the acquisition of mobile genetic elements from other bacterial organisms (*V. cholerae* and otherwise) through horizontal gene transfer. This information would provide a clear and actionable benefit to public health, as genotypic surveillance could therefore also serve the role of current antimicrobial susceptibility assays, enabling the use of one holistic assay (whole genome sequencing) to obtain the results of what is currently at least two separate experiments.

Additionally, while our work provides further evidence of cross-border transmission of cholera, coordination with local epidemiology and policy teams is needed to figure out how this transmission is happening and how best to stop it. Regional analyses, as demonstrated here, can identify broad transmission patterns, such as our finding that transmissions mostly occur between neighboring countries. However, more local, targeted investigations using sequencing to confirm epidemiological hypotheses are needed to determine fine-scale patterns that can more directly enhance containment and intervention strategies. Beyond transmission, future studies could also address the effectiveness of interventions directly, for example by looking more closely at mutations in *V. cholerae* genes implicated in vaccine effectiveness.

Through the creation of CholGEN, we have laid the foundations for genomics-informed decision-making for cholera containment and elimination. In this manuscript, we showcase how developing and extending local sequencing and analytical capacity resulted in the discovery of new findings related to *V. cholerae* spread. Crucially, we have also created a collaborative, multicountry group working to advance cholera containment and elimination together. Our next steps include more in-depth analyses of the collected data, including

exploration of virulence factors, vaccine targets, and within-country cholera spread. We will also use the information gathered here to inform prospective surveillance in all seven AU Member States. Coupled with effective data sharing and collaboration, we hope that these data, and the conclusions we draw from them, will bring us one step closer to the goal of global cholera elimination by 2030.

METHODS

Ethics statement. This research utilized samples collected through routine national surveillance programs and all data and samples were anonymized to ensure the privacy and confidentiality of the individuals involved. The use of these samples was approved by respective Ethical Review Boards/Institutions in each country, ensuring that the research adhered to the highest ethical standards and legal requirements. In Cameroon, ethical approval was granted by the National Committee on Ethics in Research for Human Health (2024/02/1640/CE/CNERSH/SP). In DRC, ethical approval was granted by the Board of the Ethics Committee of the School of Public Health at the University of Kinshasa (ESP/CE/148/2023 and ESP/CE/149/2023). In Malawi, this work was approved by the National Health Sciences Research Committee (Protocol #867). In Mozambique, ethical approval was granted by the National Bioethics Committee for Health (335/CNBS/23). In Nigeria, ethical approval was not required as it is based on data from Nigeria's national surveillance program, collected by the Nigeria Centre for Disease Control. In Uganda, ethical approval was granted by the Uganda Ministry of Health National Health Laboratory Services (UNHL-2024-88).

Sample collection, cholera confirmation, and bacterial culture. *V. cholerae* isolates were collected from clinically suspected cholera cases in CholGEN Member States between 2018 and 2024. In each Member State, isolates came from patients from cholera treatment centers during outbreaks or from endemic areas. In all seven countries, the Cholera Rapid Diagnostic Test positive samples were transported in Cary-Blair media to laboratories then cultured on Thiosulfate-citrate-bile salts-sucrose (**TCBS**) agar medium and incubated for 20–24 hours. Phenotypic identification of *V. cholerae* colonies was based on morphology, motility, and biochemical characteristics (positive oxidase, saccharose, indole, and gelatinase). *V. cholerae* isolates were confirmed with agglutination tests with anti-O1 or anti-O139 serum (WHO antisera).

Genomic DNA extraction and quantification. DNA extraction was performed at multiple laboratories across the CholGEN participating countries: NPHL (Yaounde, Cameroon), CPHL (Kampala, Uganda), INS (Maputo, Mozambique), ZNPHI (Lusaka, Zambia), PHIM (Lilongwe, Malawi), INRB (Kinshasa, DRC), and NCDC (Abuja, Nigeria). Confirmed V. cholerae isolates were retrieved from storage and subcultured on selective TCBS, MH, and/or HCK agar plates and incubated at 37°C overnight. In all countries, *V. cholerae* DNA was extracted using the standard protocol from the Qiagen QIAamp DNA Mini Kit, with a final elution volume of 200 µL as per the manufacturer's instructions. Genomic DNA concentrations were measured with the Qubit Fluorometer 4.0 (Thermo Fisher), following the dsDNA HS assay standard protocol and stored at 4°C.

Library preparation and sequencing. Illumina library preparation and sequencing was performed by laboratories in CholGEN participating countries. DNA samples were normalized to 0.6 ng/µL from which 30 µL was used as input for the Illumina DNA Prep Kit (Illumina). Library concentrations were assessed using the Qubit High Sensitivity DNA Kit (Invitrogen), and library size distributions were measured using a BioAnalyzer High Sensitivity DNA Kit (Agilent). Genomic libraries from individual samples that had expected size distributions and had a concentration greater than 1 ng/µL were normalized and pooled in equimolar amounts at 2 nM. The 2 nM library pool was sequenced on either an Illumina MiSeq or NextSeq 2000 using 300 cycles kits. Quality assessments of short reads were performed using FastQC [50].

Reference-based Genome Assembly. We developed a pathogen-agnostic bioinformatics pipeline to generate consensus sequences from raw sequencing reads, called *bacpage*, which is available as a local command line tool (available on Github: https://github.com/CholGen/bacpage; with an online manual: https://cholgen.github.io) on the cloud-based Terra platform [51], where a majority of analyses were performed. Briefly, as part of the pipeline, paired-end reads were aligned against the *Vibrio cholerae* N16961 isolate (accession AE003852/AE003853) using bwa-mem v0.7.17 [52]. Variants compared to the reference were identified using BCFTools v1.20 [53]. Variants were retained if they met the following criteria: (1) variant quality score of at least 20, (2) mapping score of at least 30, (3) supported by at least 15 reads, (4) present in at least 50% of reads covering a position, and (5) supported from both strands. A consensus sequence for each sample was generated by applying the supported variants to a concatenated reference genome.

Consensus sequences for each sample were kept for subsequent analyses if at least 90% of their reads mapped to the reference genome, had median coverage across all positions of the reference genome of at least 15, and had less than 10% ambiguous nucleotides.

***Vibrio cholerae* Genomic Data.** The sequences generated in this study were combined with 1772 publicly available third wave (AFR9 onwards) genomes that were found by searching literature, Sequence Read Archives (**SRA**), and VibrioWatch for *V. cholerae* isolates from Africa and Asia (**Supp. Data 2**). Where raw sequencing data was available, we generated reference-based assemblies using the same pipeline and reference as described above. In other cases, where only assembled contigs were available, variants compared to the N16961 reference were identified using snippy v4.6.0 (https://github.com/tseemann/snippy) with the `--ctgs` flag and applied to the reference genome using BCFTools.

Antimicrobial Resistance Profiling. Short reads were assembled *de novo* using the TheiaProk Illumina paired-end sequencing workflow publicly available on DockStore [54]. Briefly, the workflow trims low quality reads and portions of reads using fastp [55] and removes sequencing adaptor sequences using BBDuk [56]. Assembly was performed on high-quality reads with the Shovill pipeline (https://github.com/tseemann/shovill) which uses the SKESA assembler [57]. We then used AMRFinderPlus v3.11.20 with database version 2023-09-26.1 to identify antimicrobial resistance genes and point mutations in the assemblies [58].

Epidemiological Data. We obtained yearly case counts for each country participating in CholGEN from data curated by the cholera taxonomy project at the Johns Hopkins Bloomberg School of Public Health (https://cholera-taxonomy.middle-distance.com/). Data contributions to the team have come from many sources, though the annual case counts are heavily dependent on the WHO Annual Cholera Reports and the Global Health Observatory data repository.

Phylogenetic Analysis. We used the `phylogeny` subcommand of bacpage to generate a recombination-masked maximum likelihood phylogeny. Briefly, the pipeline performs the following steps. The complete genomic dataset was assembled by concatenating the individual reference-based assemblies along with the N16961 reference into a pseudo-alignment. Problematic sites including known recombinant regions were masked using a custom GFF file and script. Novel recombinant regions were masked using gubbins v2.3.4 [59]. To reduce the computational burden of the phylogenetic analysis, we filtered the pseudo-alignment to only include variable positions using SNP-sites v2.5.1 [60]. We constructed a maximum likelihood phylogenetic tree for the dataset using IQ-TREE2 and an GTR substitution model [61,62].

The resulting phylogeny was rooted on the reference genome. To further reduce computation burden of the analysis we extracted the subtree descending from the most recent common ancestor of lineages AFR9-15, which corresponds to the third wave of 7PET. The subtree was time-resolved using Treetime v0.11.3 [63]. Taxa deviating more than three interquartile distances from the clock-rate regression were pruned from the phylogeny. Additionally, we randomly resolved polytomies in the phylogeny by adding zero-length branches with gotree [64].

We refined the time-resolved phylogeny using BEAST v1.10.5 [65]. We specified a GTR substitution model with gamma distributed rate heterogeneity under an uncorrelated relaxed molecular clock and a constant coalescent tree prior. For the uncorrelated relaxed molecular clock, we specified an informative lognormal distribution with a mean of $8.4*10^{-7}$ substitutions/site/year and a standard deviation of $1.4*10^{-6}$ consistent with Weill et al. 2017 [2]. To account for utilizing an SNP alignment rather than a full-genome alignment, we additionally supplied the number of constant sites in the cholera reference genome.

We ran an MCMC chain of 300 million steps utilizing the BEAGLE computational library [66]. Parameters and trees were sampled every 10,000 and 100,000 steps, respectively. Convergence and mixing of the MCMC chains were assessed with Tracer v1.7.2 and Beastiary v1.8.3 [67,68]. We generated a second chain by restarting the initial chain at the 200 millionth state with a different random seed and running it for an additional 100 million states, resulting in two chains of 100 million states each after discarding burnin. After combining chains, all estimated continuous parameters were determined to have effective sample sizes of greater than 100. BEAST XML and log files for the phylogenetic analysis are available on GitHub (https://github.com/CholGen/RegionalAnalysis-2024).

Lineage Assignment. Sequences collected from Africa were assigned to the lineages initially described by Weill et al. 2017 [2]. Assignment was performed by phylogenetic placement using a custom script. Briefly, for each lineage, we identified the node corresponding to the most recent common ancestor (**MRCA**) of lineage representatives collected from prior publications [2,16,26,28,69]. All sequences descending from the MRCA were assigned to the lineage. Sequences that were collected from Africa but did not descend from a previously described lineage were assigned a value of “sporadic.”

Phylogeographic Reconstruction. We performed a discrete state ancestral reconstruction on geographic states using BEAST [39]. This analysis reconstructed location-transition history across an empirical distribution of 2000 time-calibrated trees samples from the posterior tree distribution estimated above [70,71]. To conduct the analysis, we assigned each taxon a discrete location state based on the country where it was collected. To limit the number of transition rates needing to be estimated, we binned non-African sequences into a single “Other” state, and African countries with less than 20 reported sequences into a single “Other-Africa” state. Ultimately we used 14 distinct locations states. We assumed that geographic transition rates were reversible and used a symmetric substitution model. We used Bayesian stochastic search variable selection to infer non-zero migration rates [39]. The MCMC algorithm was run for 500,000 generations and sampled every 500 generations. We used the `TreeMarkovJumpHistoryAnalyzer` from the pre-release version of BEAST v1.10.5 to obtain the Markov jump estimates and their timings from the posterior tree distribution and assumed that they are a suitable proxy for the transmission between two locations. We used TreeAnnotator v1.10 to construct a maximum clade credibility (**MCC**) tree which we visualized with baltic (https://github.com/evogytis/baltic). BEAST XML and log files for the discrete phylogeographic analysis are available on GitHub (https://github.com/CholGen/RegionalAnalysis-2024).

Introduction Identification. We identified country-specific strains from the posterior tree distribution of the discrete phylogeographic reconstruction using a custom script. Country-specific strains were defined as two or more taxa from the same country that descend from a single, shared introduction of cholera into the country from another country. Descendant taxa from outside the specific country, do not belong to the strain. This concept and the algorithm for identifying strains is derived from the definition of Du Plessis et al. 2021 [72] (see the citation for an illustration of this concept, wherein the term transmission lineage is used in place of strain).

Phylogenetic diversity. We calculated the amount of evolutionary history contributed by a set of genomes as their phylogenetic diversity [43,73,74]. For each tree in the posterior distribution of the phylogenetic analysis, we converted the units of branch lengths from years to substitutions/site by multiplying branch lengths by the estimated substitution rate for each branch. For each country, we marked taxa that were collected from the country and were generated by CholGEN participants. We then calculated the phylogenetic diversity contributed by CholGEN for a given country as the difference between the total branch length of the tree and the total branch length of the tree after pruning the marked taxa. Lastly, this value was divided by the total number of sequences marked for a given country.

Rarefaction analysis. For each country, we identified unique introductions into each country and identified genomes that descended from those introductions for each tree in the posterior. For varying sample sizes, ranging from 1 to the total number of genomes collected for a country, genomes were randomly drawn without replacement and the number of unique introductions in the sample was computed. To incorporate uncertainty from the discrete state reconstruction, we performed this rarefaction analysis 1000 times, using a randomly sampled tree each time. For each trial, we parametrically estimated the total number of introductions into each country by fitting the rarefaction curve with a Michaelis-Menten equation. The Michaelis-Menten equation is a two parameter model with the following form:

$$y = \frac{V^{*}x}{K+x}$$

where x is the number of genomes sampled, y is the number of introductions observed, V is the asymptotic estimate of the total number of introductions, and K is a shape parameter describing the rate at which new genomes reveal introductions.

Data availability. Raw sequencing reads are available on NCBI under the BioProject accession ID PRJNA1145341 and Sequence Read Archive accession IDs are provided in **Supp. Data 1**. Accession IDs for the publicly available sequences acquired from NCBI or VibrioWatch are provided in **Supp. Data 2**.

Code availability. Code for all analyses and figure generation, as well as XMLs and log file for BEAST analyses are available at: https://github.com/CholGen/RegionalAnalysis-2024.

ACKNOWLEDGEMENTS
We extend our sincere gratitude to the Africa Centres for Disease Control and Prevention (Africa CDC), Africa Public Health Foundation (APHF), and Broad Institute of MIT and Harvard, for their invaluable support and collaboration. This work was supported by the Bill & Melinda Gates Foundation (INV-047157 and INV-047156).

AUTHOR CONTRIBUTIONS

G.M., N.L.M., C.K.T., S.W., and S.K.T. conceptualized the study. G.M., N.L.M., A.S.A., and S.W. contributed the methodology. N.L.M. provided software. G.M. and N.L.M. conducted the formal analysis. M.K., G.K.K., O.A.A., J.J.C., Mathews Kagoli, R.G.E., A.A., I.S., V.J.N., A.A.A., O.A.B., B.M.A., A.B., E.B.M., F.O., C.E.C., N.I., O.K., O.L., P.K.M., G.A.E., I.C.M., A.A.M., G.N., J.M., K.M., A.M.N., M.E.N., M.P., D.M.S., C.M., S.M.S., D.W.W., P.O.W., M.Y., S.Y., L.Z., and CholGEN Consortium authors provided resources. G.M., M.K., G.K.K., O.A.A., J.J.C., Mathews Kagoli, R.G.E., A.A., I.S., V.J.N., A.A.A., O.A.B., B.M.A., A.B., E.B.M., F.O., C.E.C., N.I., O.K., O.L., P.K.M., G.A.E., I.C.M., A.A.M., G.N., J.M., K.M., A.M.N., M.E.N., M.P., D.M.S., C.M., S.M.S., D.W.W., P.O.W., M.Y., S.Y., L.Z., and CholGEN Consortium authors curated data. G.M., N.L.M., C.K.T., S.W., and S.K.T. wrote the original draft of the manuscript. All authors reviewed and edited the manuscript. N.L.M. performed visualization. J.E.B., R.C., H.H., J.I., J.P.L., D.M., S.N., A.K.D., D.A.S., J.K., Y.K.T., S.W., and S.K.T. supervised the study. G.M., C.K.T., S.W., and S.K.T. undertook project administration. A.K.D., D.A.S., and S.K.T. acquired funding.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1. Gaffga NH, Tauxe RV, Mintz ED. Cholera: a new homeland in Africa? Am J Trop Med Hyg. 2007;77: 705–713.

2. Weill F-X, Domman D, Njamkepo E, Tarr C, Rauzier J, Fawal N, et al. Genomic history of the seventh pandemic of cholera in Africa. Science. 2017;358: 785–789.

3. Rebaudet S, Sudre B, Faucher B, Piarroux R. Environmental determinants of cholera outbreaks in inland Africa: a systematic review of main transmission foci and propagation routes. J Infect Dis. 2013;208 Suppl 1: S46–54.

4. Rebaudet S, Sudre B, Faucher B, Piarroux R. Cholera in coastal Africa: a systematic review of its heterogeneous environmental determinants. J Infect Dis. 2013;208 Suppl 1: S98–106.

5. Hu D, Liu B, Feng L, Ding P, Guo X, Wang M, et al. Origins of the current seventh cholera pandemic. Proc Natl Acad Sci U S A. 2016;113: E7730–E7739.

6. Global Task Force for Cholera Control. Ending Cholera A Global Roadmap to 2030. Available: https://www.gtfcc.org/wp-content/uploads/2020/09/ending-cholera-a-global-roadmap-to-2030.pdf

7. Africa CDC weekly event based surveillance report, July 2024. In: Africa CDC [Internet]. Africa Centres for Disease Control and Prevention; 1 Jul 2024 [cited 16 Jul 2024]. Available: https://africacdc.org/download/africa-cdc-weekly-event-based-surveillance-report-july-2024/

8. World Health Organization. Multi-country outbreak of cholera: External Situation Report. 2024 Jun. Report No.: 15. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20240619_multi-country_outbreak-of-cholera_sitrep_15.pdf?sfvrsn=80f3d7af_3&download=false

9. United National Central Emergency Response Fund. CERF Allocation: Mozambique. 2024 Jan. Available: https://cerf.un.org/what-we-do/allocation/2024/summary/24-RR-MOZ-62928

10. Chabuka L, Choga WT, Mavian CN, Moir M, Tegally H, Wilkinson E, et al. Genomic epidemiology of the cholera outbreak in Malawi 2022-2023. medRxiv. 2023. p. 2023.08.22.23294324. doi:10.1101/2023.08.22.23294324

11. Miggo M, Harawa G, Kangwerema A, Knovicks S, Mfune C, Safari J, et al. Fight against cholera outbreak, efforts and challenges in Malawi. Health Sci Rep. 2023;6: e1594.

12. Unicef, Others. Cholera epidemiology and response factsheet Cameroon. WHO; 2013.

13. Ekeng E, Tchatchouang S, Akenji B, Issaka BB, Akintayo I, Chukwu C, et al. Regional sequencing collaboration reveals persistence of the T12 Vibrio cholerae O1 lineage in West Africa. Elife. 2021;10. doi:10.7554/eLife.65159

14. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 2011;477: 462–465.

15. Benamrouche N, Belkader C, Njamkepo E, Zemam SS, Sadat S, Saighi K, et al. Outbreak of Imported Seventh Pandemic Vibrio cholerae O1 El Tor, Algeria, 2018. Emerg Infect Dis. 2022;28: 1241–1245.

16. Weill F-X, Domman D, Njamkepo E, Almesbahi AA, Naji M, Nasher SS, et al. Genomic insights into the 2016-2017 cholera epidemic in Yemen. Nature. 2019;565: 230–233.

17. Xiao S, Abade A, Boru W, Kasambara W, Mwaba J, Ongole F, et al. New Vibrio cholerae sequences from Eastern and Southern Africa alter our understanding of regional cholera transmission. medRxiv. 2024. doi:10.1101/2024.03.28.24302717

18. Langa JP, Sema C, De Deus N, Colombo MM, Taviani E. Epidemic waves of cholera in the last two decades in Mozambique. The Journal of Infection in Developing Countries. 2015;9: 635–641.

19. Moore S, Worku Demlie Y, Muluneh D, Dunoyer J, Hussen M, Wossen M, et al. Spatiotemporal dynamics of cholera epidemics in Ethiopia: 2015-2021. Sci Rep. 2024;14: 7170.

20. Okeke IN, Aboderin OA, Byarugaba DK, Ojo KK, Opintan JA. Growing problem of multidrug-resistant enteric pathogens in Africa. Emerg Infect Dis. 2007;13: 1640–1646.

21. Dengo-Baloi LC, Semá-Baltazar CA, Manhique LV, Chitio JE, Inguane DL, Langa JP. Antibiotics resistance in El Tor Vibrio cholerae 01 isolated during cholera outbreaks in Mozambique from 2012 to 2015. PLoS One. 2017;12: e0181496.

22. Tanui CK, Tessema SK, Tegegne MA, Tebeje YK, Kaseya J. Unlocking the power of molecular and genomics tools to enhance cholera surveillance in Africa. Nat Med. 2023;29: 2387–2388.

23. Irenge LM, Ambroise J, Mitangala PN, Bearzatto B, Kabangwa RKS, Durant J-F, et al. Genomic analysis of pathogenic isolates of Vibrio cholerae from eastern Democratic Republic of the Congo (2014-2017). PLoS Negl Trop Dis. 2020;14: e0007642.

24. Bwire G, Sack DA, Almeida M, Li S, Voeglein JB, Debes AK, et al. Molecular characterization of Vibrio cholerae responsible for cholera epidemics in Uganda by PCR, MLVA and WGS. PLoS Negl Trop Dis. 2018;12: e0006492.

25. Mwape K, Kwenda G, Kalonda A, Mwaba J, Lukwesa-Musyani C, Ngulube J, et al. Characterisation of Vibrio cholerae isolates from the 2009, 2010 and 2016 cholera outbreaks in Lusaka province, Zambia. Pan Afr Med J. 2020;35: 32.

26. Smith AM, Sekwadi P, Erasmus LK, Lee CC, Stroika SG, Ndzabandzaba S, et al. Imported Cholera Cases, South Africa, 2023. Emerg Infect Dis. 2023;29: 1687–1690.

27. Mwaba J, Debes AK, Shea P, Mukonka V, Chewe O, Chisenga C, et al. Identification of cholera hotspots in Zambia: A spatiotemporal analysis of cholera data from 2008 to 2017. PLoS Negl Trop Dis. 2020;14: e0008227.

28. Fayad AA, Rafei R, Njamkepo E, Ezzeddine J, Hussein H, Sinno S, et al. An unusual two-strain cholera outbreak in Lebanon, 2022-2023: a genomic epidemiology study. bioRxiv. 2023. p. 2023.11.29.569232. doi:10.1101/2023.11.29.569232

29. Qamar K, Malik UU, Yousuf J, Essar MY, Muzzamil M, Hashim HT, et al. Rise of cholera in Iraq: A rising concern. Ann Med Surg (Lond). 2022;81: 104355.

30. Eneh SC, Admad S, Nazir A, Onukansi FO, Oluwatobi A, Innocent DC, et al. Cholera outbreak in Syria amid humanitarian crisis: the epidemic threat, future health implications, and response strategy - a review. Front Public Health. 2023;11: 1161936.

31. Oprea M, Njamkepo E, Cristea D, Zhukova A, Clark CG, Kravetz AN, et al. The seventh pandemic of cholera in Europe revisited by microbial genomics. Nat Commun. 2020;11: 5347.

32. Monir MM, Islam MT, Mazumder R, Mondal D, Nahar KS, Sultana M, et al. Genomic attributes of Vibrio cholerae O1 responsible for 2022 massive cholera outbreak in Bangladesh. Nat Commun. 2023;14: 1154.

33. Leibovici-Weissman Y 'ara, Neuberger A, Bitterman R, Sinclair D, Salam MA, Paul M. Antimicrobial drugs for treating cholera. Cochrane Database Syst Rev. 2014;2014: CD008625.

34. Kuma GK, Opintan JA, Sackey S, Nyarko KM, Opare D, Aryee E, et al. Antibiotic resistance patterns amongst clinical Vibrio cholerae O1 isolates from Accra, Ghana. International Journal of Infection Control. 2014;10. Available: http://www.ijic.info/article/view/13152

35. Fuesslin V, Krautwurst S, Srivastava A, Winter D, Liedigk B, Thye T, et al. Prediction of Antibiotic Susceptibility Profiles of Vibrio cholerae Isolates From Whole Genome Illumina and Nanopore Sequencing Data: CholerAegon. Front Microbiol. 2022;13: 909692.

36. Ambrose SJ, Harmer CJ, Hall RM. Compatibility and entry exclusion of IncA and IncC plasmids revisited: IncA and

IncC plasmids are compatible. Plasmid. 2018;96-97: 7–12.

37. De R. Mobile Genetic Elements of Vibrio cholerae and the Evolution of Its Antimicrobial Resistance. Frontiers in Tropical Diseases. 2021;2. doi:10.3389/fitd.2021.691604

38. Lassalle F, Al-Shalali S, Al-Hakimi M, Njamkepo E, Bashir IM, Dorman MJ, et al. Genomic epidemiology reveals multidrug resistant plasmid spread between Vibrio cholerae lineages in Yemen. Nat Microbiol. 2023;8: 1787–1798.

39. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5: e1000520.

40. Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GA, Kent J. Biodiversity hotspots for conservation priorities. Nature. 2000;403: 853–858.

41. Winter M, Devictor V, Schweiger O. Phylogenetic diversity and nature conservation: where are we? Trends Ecol Evol. 2013;28: 199–204.

42. Mace GM, Gittleman JL, Purvis A. Preserving the tree of life. Science. 2003;300: 1707–1709.

43. Dudas G, Batson J. Accumulated metagenomic studies reveal recent migration, whole genome evolution, and undiscovered diversity of orthomyxoviruses. J Virol. 2023;97: e0105623.

44. Black A, Moncla LH, Laiton-Donato K, Potter B, Pardo L, Rico A, et al. Genomic epidemiology supports multiple introductions and cryptic transmission of Zika virus in Colombia. BMC Infect Dis. 2019;19: 963.

45. Wilkinson E, Giovanetti M, Tegally H, San JE, Lessells R, Cuadros D, et al. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science. 2021;374: 423–431.

46. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale rates of evolutionary change in bacteria. Microb Genom. 2016;2: e000094.

47. Bwire G, Ali M, Sack DA, Nakinsige A, Naigaga M, Debes AK, et al. Identifying cholera “hotspots” in Uganda: An analysis of cholera surveillance data from 2011 to 2016. PLoS Negl Trop Dis. 2017;11: e0006118.

48. Bénard AHM, Guenou E, Fookes M, Ateudjieu J, Kasambara W, Siever M, et al. Whole genome sequence of Vibrio cholerae directly from dried spotted filter paper. PLoS Negl Trop Dis. 2019;13: e0007330.

49. Alam M, Sultana M, Nair GB, Siddique AK, Hasan NA, Sack RB, et al. Viable but nonculturable *Vibrio cholerae* O1 in biofilms in the aquatic environment and their role in cholera transmission. Proc Natl Acad Sci U S A. 2007;104: 17801–17806.

50. Babraham bioinformatics - FastQC A quality control tool for high throughput sequence data. [cited 15 Jul 2024]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

51. Yuen D, Cabansay L, Duncan A, Luu G, Hogue G, Overbeck C, et al. The Dockstore: enhancing a community platform for sharing reproducible and accessible computational protocols. Nucleic Acids Res. 2021;49: W624–W632.

52. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN]. 2013. Available: http://arxiv.org/abs/1303.3997

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

54. Libuit KG, Doughty EL, Otieno JR, Ambrosio F, Kapsak CJ, Smith EA, et al. Accelerating bioinformatics implementation in public health. Microb Genom. 2023;9. doi:10.1099/mgen.0.001051

55. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. Imeta. 2023;2: e107.

56. BBMap. In: SourceForge [Internet]. 15 Jul 2022 [cited 12 Jul 2024]. Available: https://sourceforge.net/projects/bbmap/

57. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. 2018;19: 153.

58. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11: 12728.

59. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43: e15.

60. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016;2: e000056.

61. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37: 1530–1534.

62. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci (Am Math Soc). 1986;17: 57–86.

63. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4: vex042.

64. Lemoine F, Gascuel O. Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. NAR Genom Bioinform. 2021;3: lqab075.

65. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4: vey016.

66. Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, Swofford DL, et al. BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. Syst Biol. 2019;68: 1052–1061.

67. Wirth W, Duchene S. Real-Time and Remote MCMC Trace Inspection with Beastiary. Mol Biol Evol. 2022;39. doi:10.1093/molbev/msac095

68. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol. 2018;67: 901–904.

69. Chaguza C, Chibwe I, Chaima D, Musicha P, Ndeketa L, Kasambara W, et al. Genomic insights into the 2022–2023*Vibrio cholerae*outbreak in Malawi. bioRxiv. 2023. doi:10.1101/2023.06.08.23291055

70. Bedford T, Riley S, Barr IG, Broor S, Chadha M, Cox NJ, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature. 2015;523: 217–220.

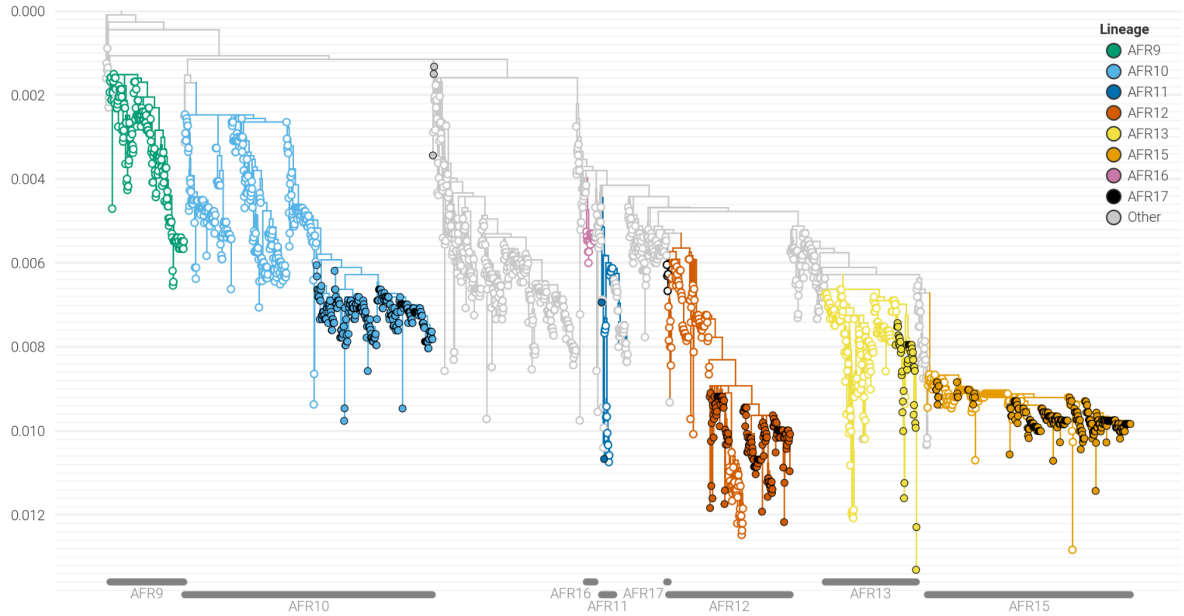
71. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014;10: e1003932.

72. du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 2021;371: 708–712.

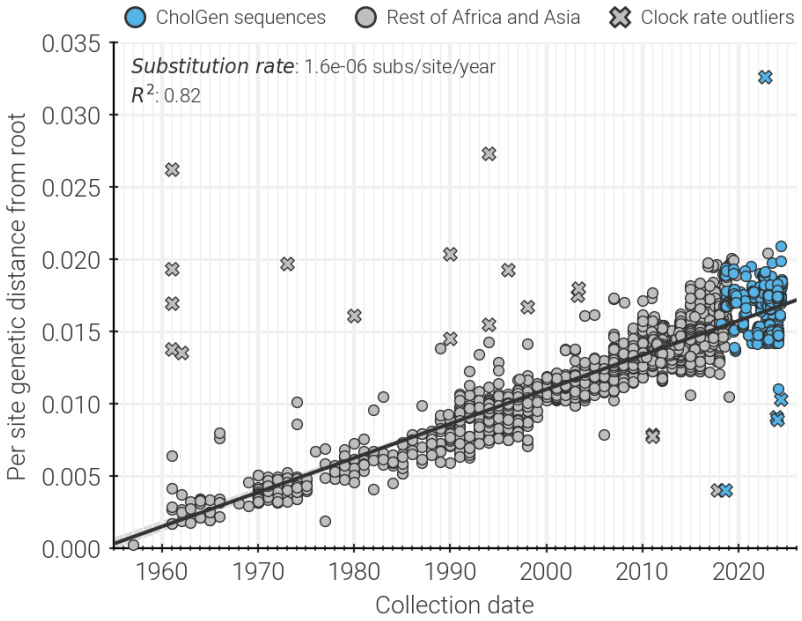
73. Felsenstein J. Phylogenies and the comparative method. Am Nat. 1985. Available: https://www.journals.uchicago.edu/doi/abs/10.1086/284325

74. Lum D, Rheindt FE, Chisholm RA. Tracking scientific discovery of avian phylogenetic diversity over 250 years. Proc Biol Sci. 2022;289: 20220088.

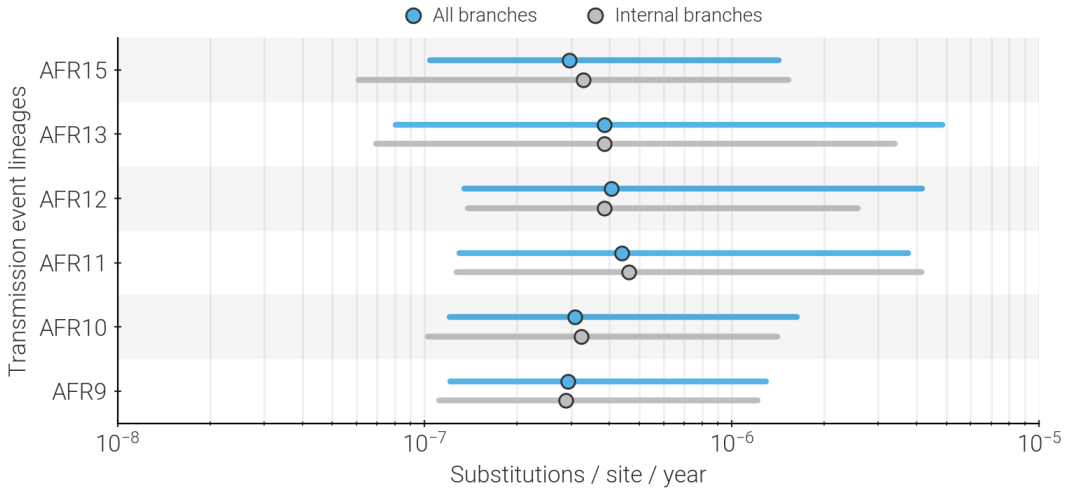
SUPPLEMENTAL FIGURES



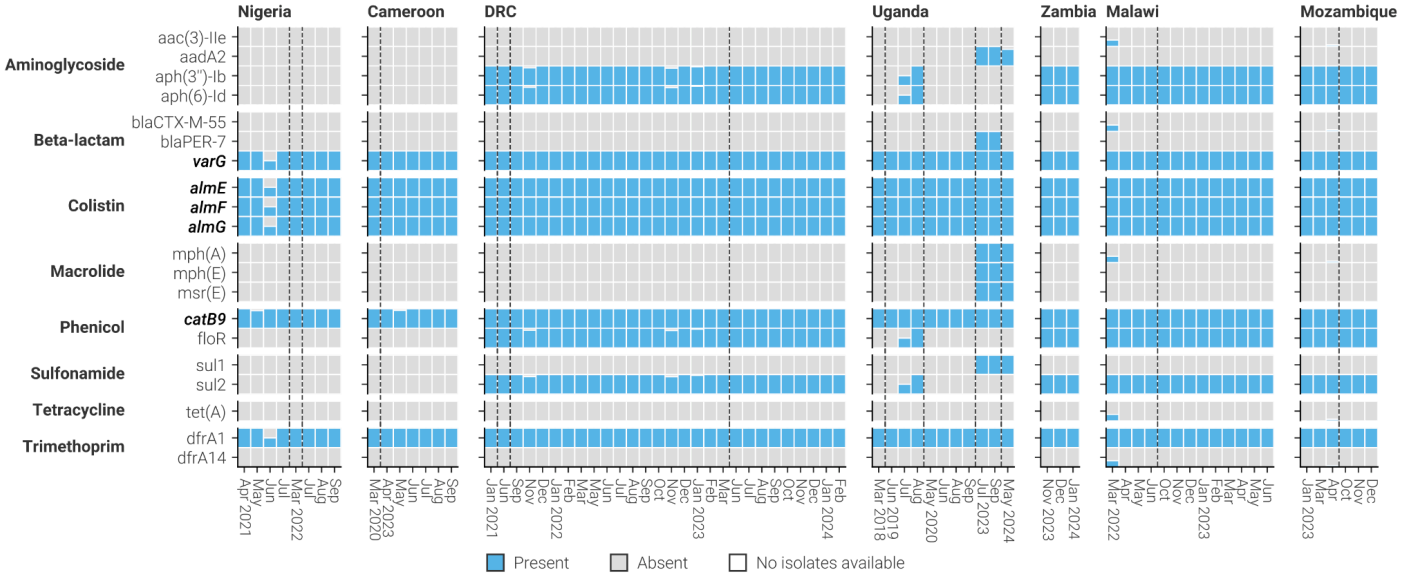
Supplemental Figure 1. Evolutionary history of the third wave of 7PET. Maximum-likelihood phylogeny with taxa labeled by the assigned lineage. Only lineages known to circulate in sub-Saharan Africa are colored. Filled circles: genomes generated by CholGEN Member States. Open circles: previously published genomes.



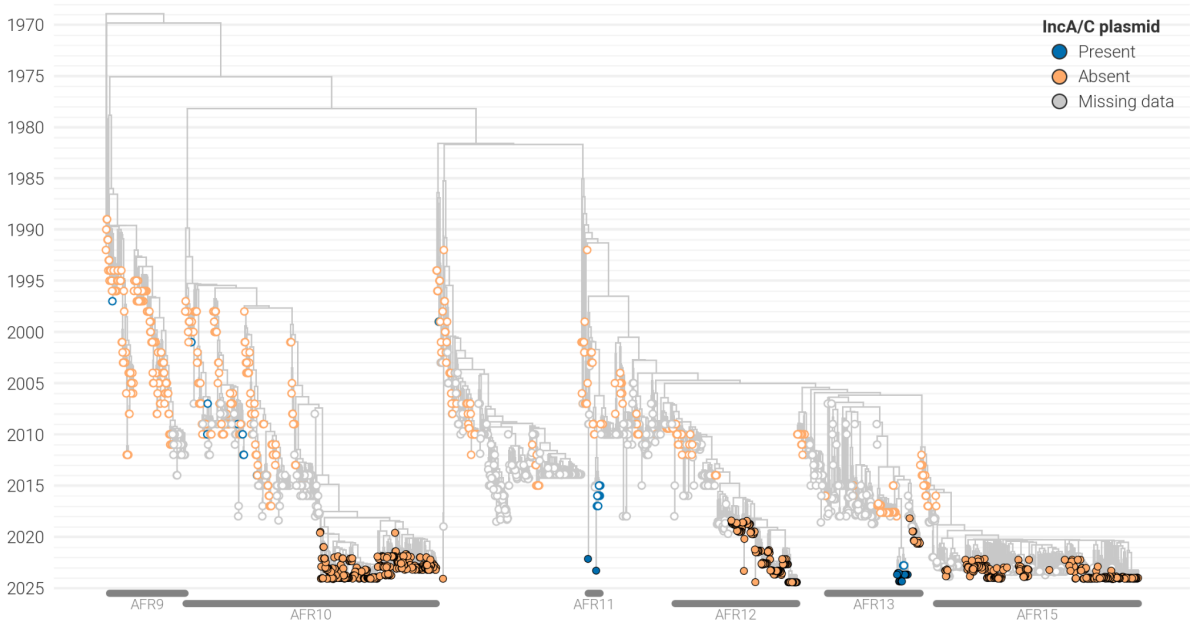
Supplemental Figure 2. Regression of root-to-tip genetic distances against sampling dates, for sequences belonging to the 7th pandemic clone of cholera. Sequences in gray represent those previously published, while sequences in blue represent those generated by CholGEN. Sequences indicated by a cross mark are more than three interquartile distances from the regression and have been removed from BEAST analyses.



Supplemental Figure 3. Distribution of substitution rates found on branches belonging to each of the currently circulating lineages. For each lineage, the range of substitution rates estimated for each branch is shown. Dot indicates median value while bar indicates 95% HPD. Blue represents the range of all branches, while gray indicates internal branches only. Note that the x-axis is log-scaled.



Supplemental Figure 4. Antimicrobial resistance profile of cholera in CholGEN Member states. The proportion of isolates collected in each month and country which carry specific antimicrobial resistance genes. Blue bars represent presence and gray represents absence of the gene. Countries are ordered generally West-to-East, and genes are organized by the class of antibiotic that they provide resistance to. Bolded genes are present on the two genomic chromosomes of cholera, whereas unbolded genes are likely present on mobile genetic elements. Dashed vertical lines indicate a break in the temporal x-axis.



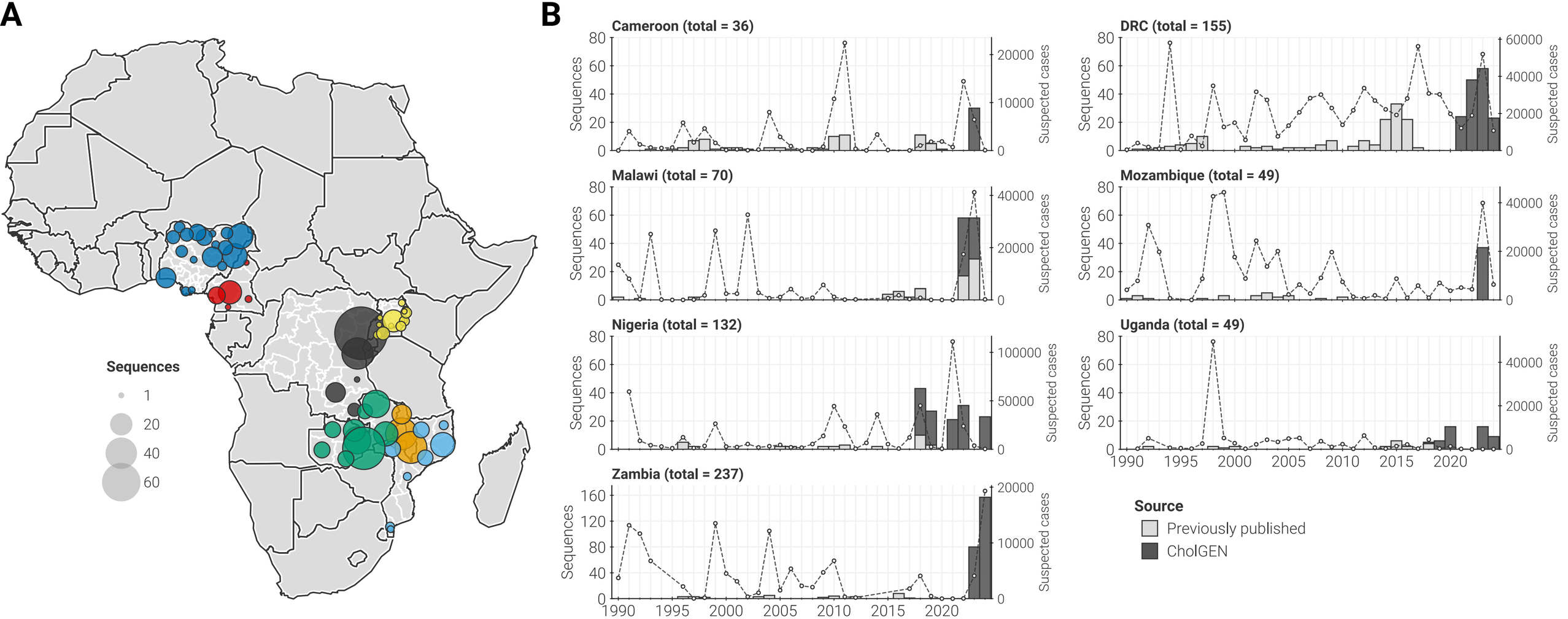
Supplemental Figure 5. Acquisition of the IncA/C plasmid amongst 7PET isolates. Maximum likelihood tree with taxa labeled by whether the IncA/C plasmid was detected or not. Filled circles: genomes generated by CholGEN Member States. Open circles: previously published genomes.

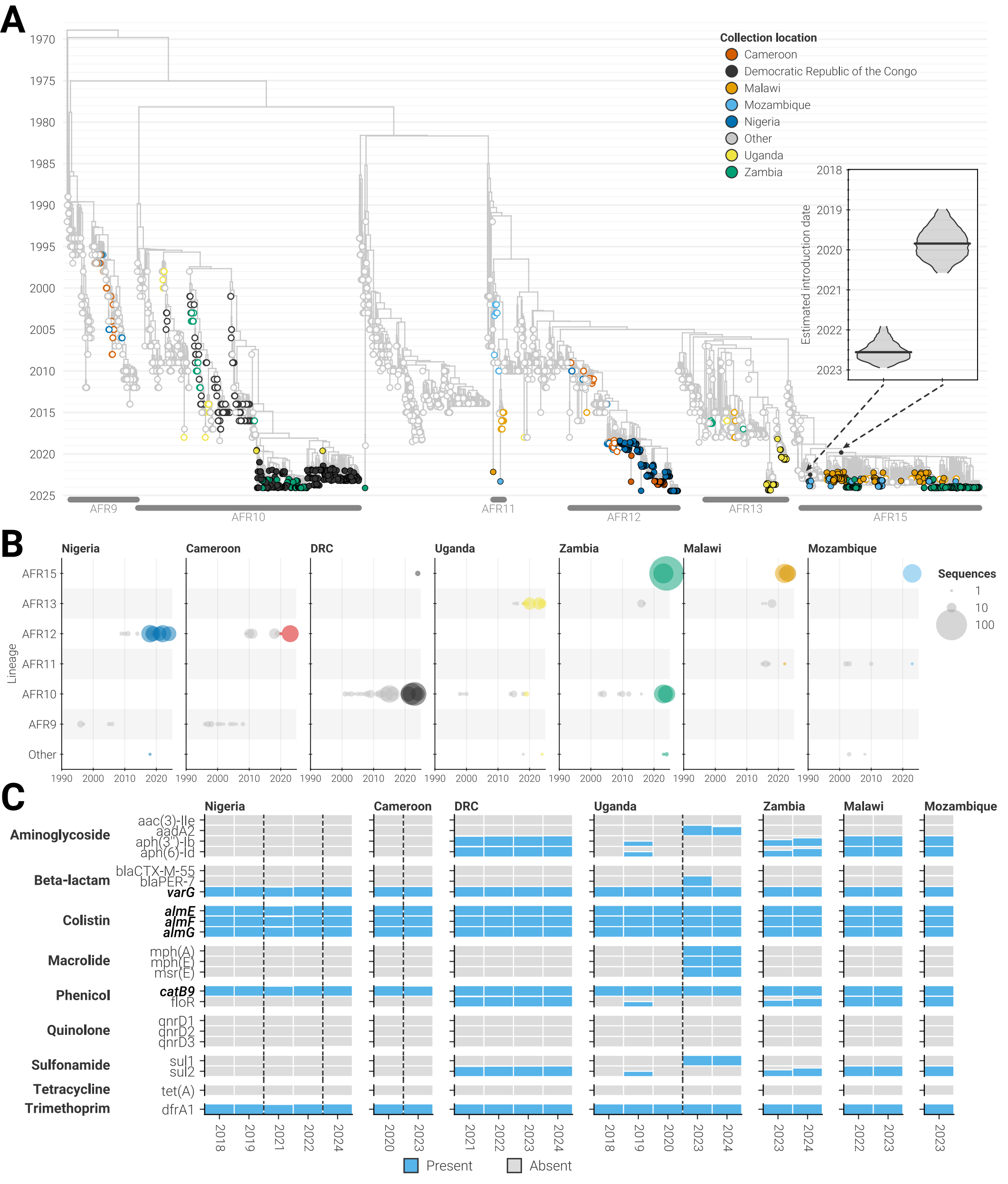
SUPPLEMENTARY INFORMATION

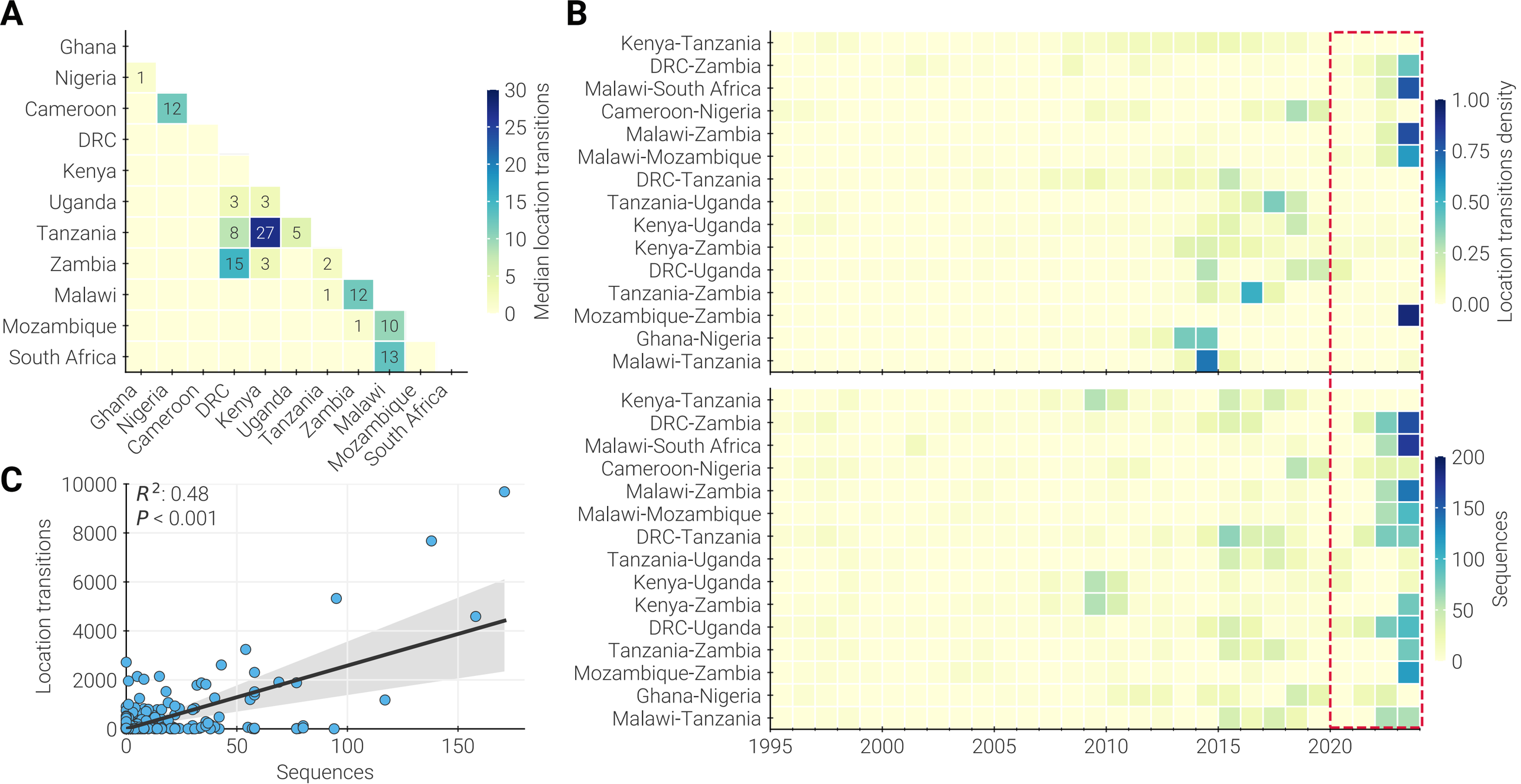
Supplementary Note 1. CholGEN Consortium Authors.

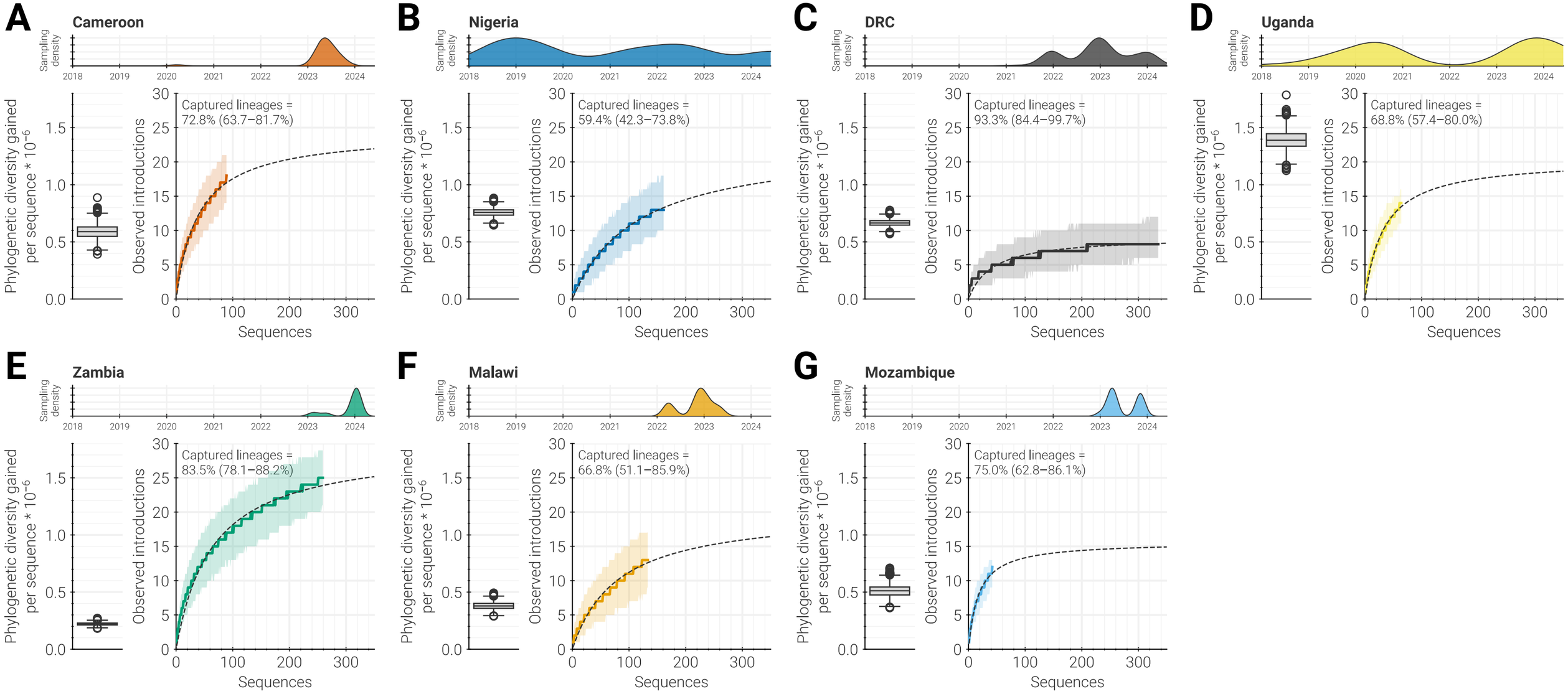
Supplementary Data 1. Details of V. cholerae O1 isolates generated by this study. Includes metadata, sequencing and assembly statistics, and antimicrobial resistance profiling data for all sequences included in analyses.

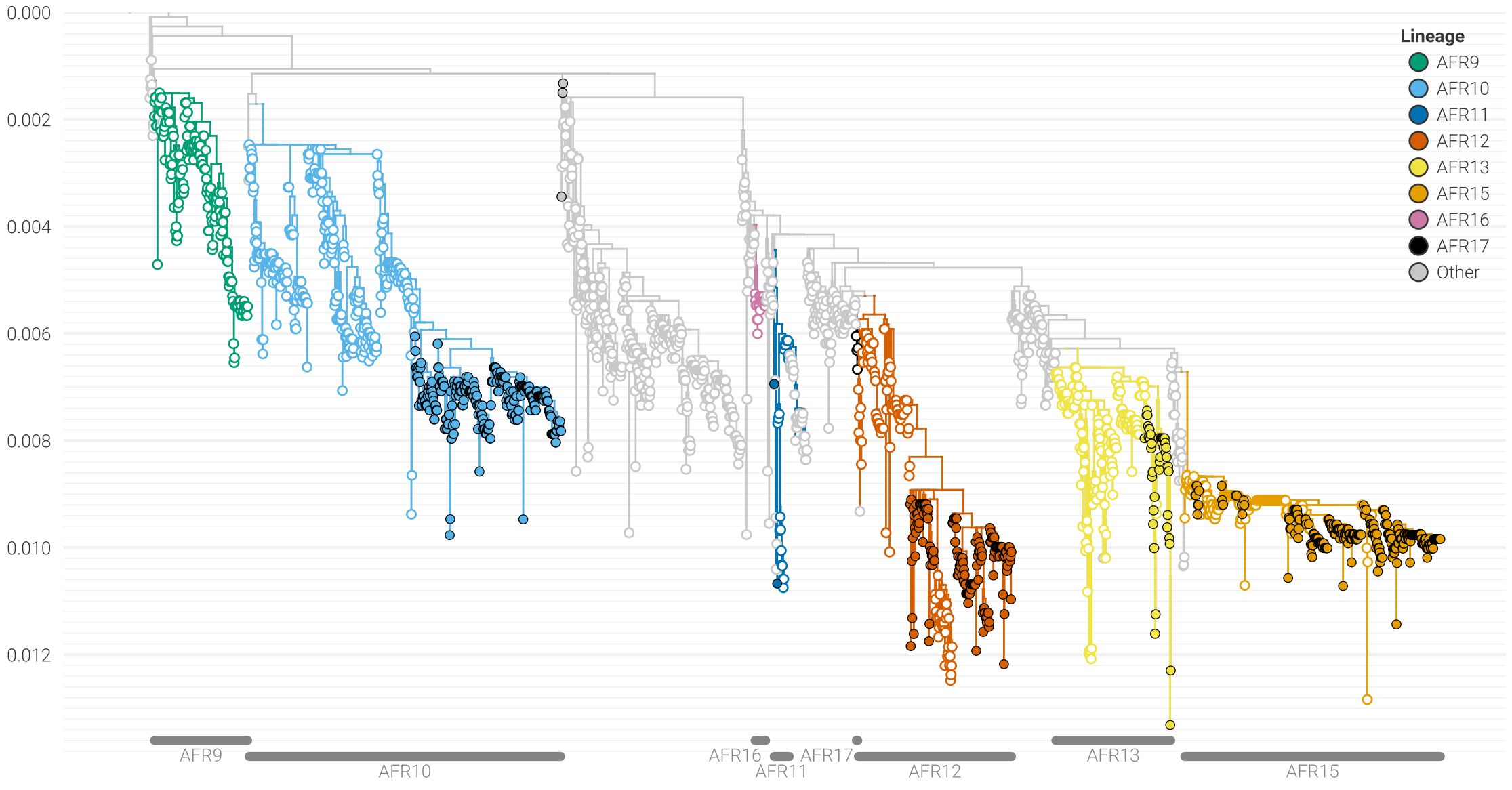
Supplementary Data 2. List of publicly available V. cholerae O1 isolates used in this study. Metadata including collection location, year, assigned lineage, and accession number for all sequences included in the background dataset for this study.

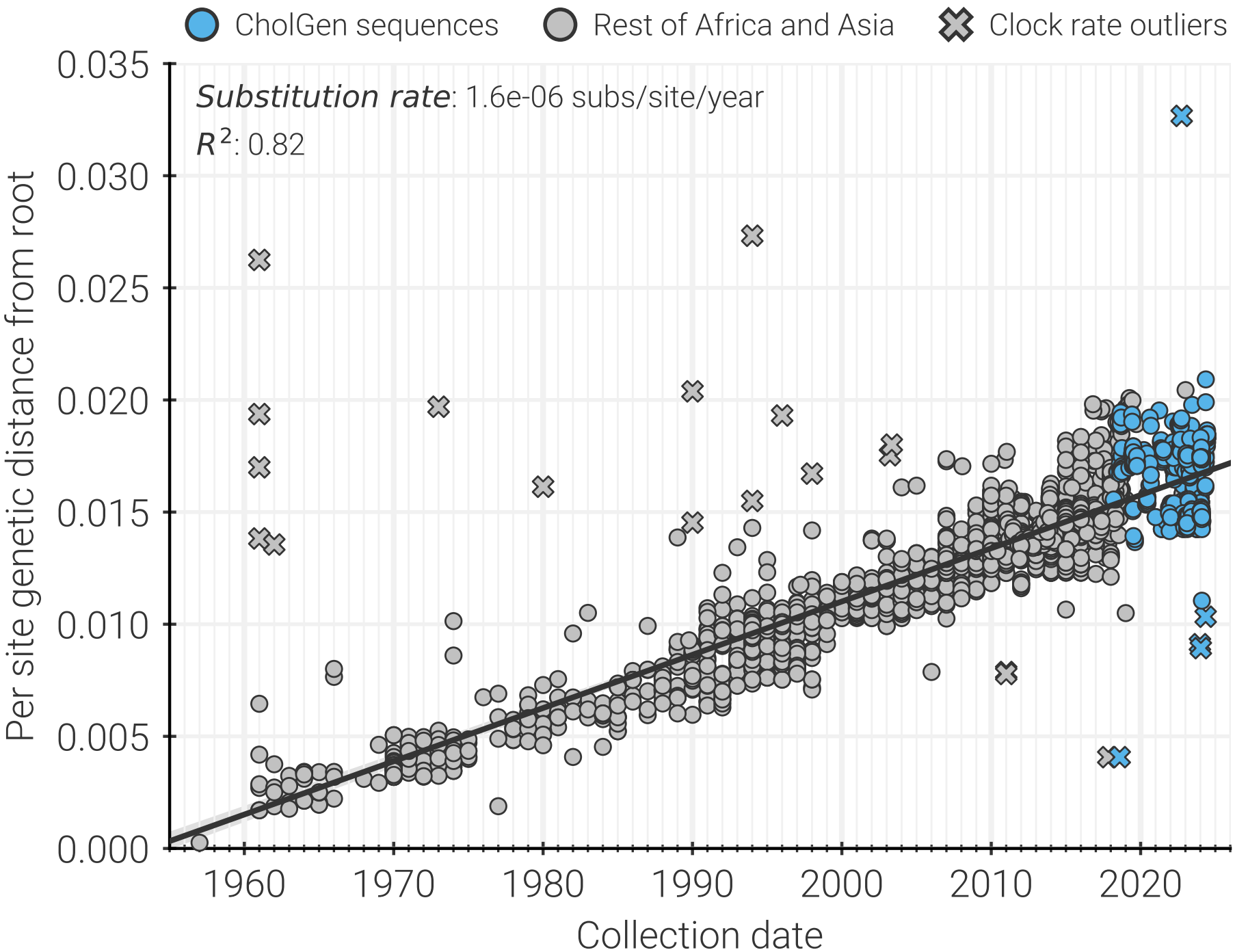

A
Sequences
1
20
40
60
B
Cameroon (total = 36)
DRC (total = 155)
Malawi (total = 70)
Mozambique (total = 49)
Nigeria (total = 132)
Uganda (total = 49)
Zambia (total = 237)
Sequences
Suspected cases
1990
1995
2000
2005
2010
2015
2020
Source
Previously published
CholGEN

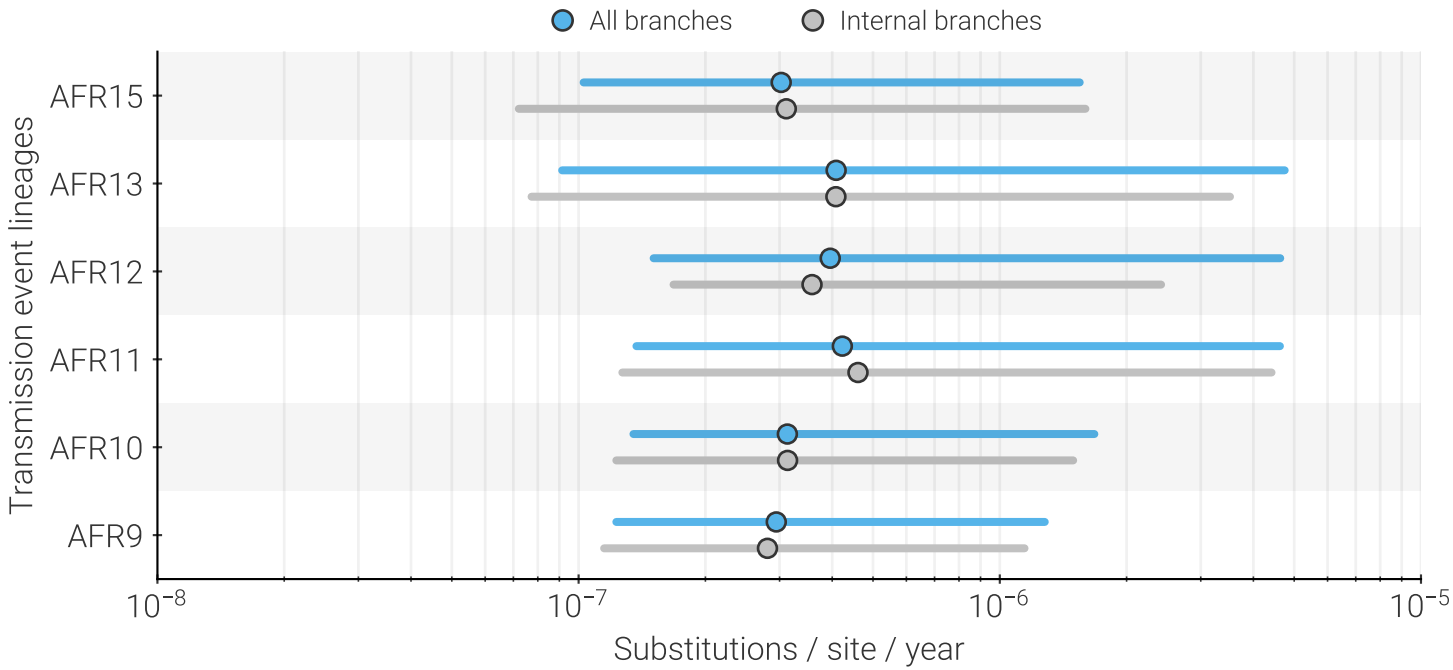

A
Collection location
Cameroon
Democratic Republic of the Congo
Malawi
Mozambique
Nigeria
Other
Uganda
Zambia
Estimated introduction date
AFR9
AFR10
AFR11
AFR12
AFR13
AFR15
B
Nigeria
Cameroon
DRC
Uganda
Zambia
Malawi
Mozambique
Lineage
Sequences
C
Aminoglycoside
Beta-lactam
Colistin
Macrolide
Phenicol
Quinolone
Sulfonamide
Tetracycline
Trimethoprim
aac(3)-IIe
aadA2
aph(3'')-Ib
aph(6)-Id
blaCTX-M-55
blaPER-7
varG
almE
almF
almG
mph(A)
mph(E)
msr(E)
catB9
floR
qnrD1
qnrD2
qnrD3
sul1
sul2
tet(A)
dfrA1
Present
Absent

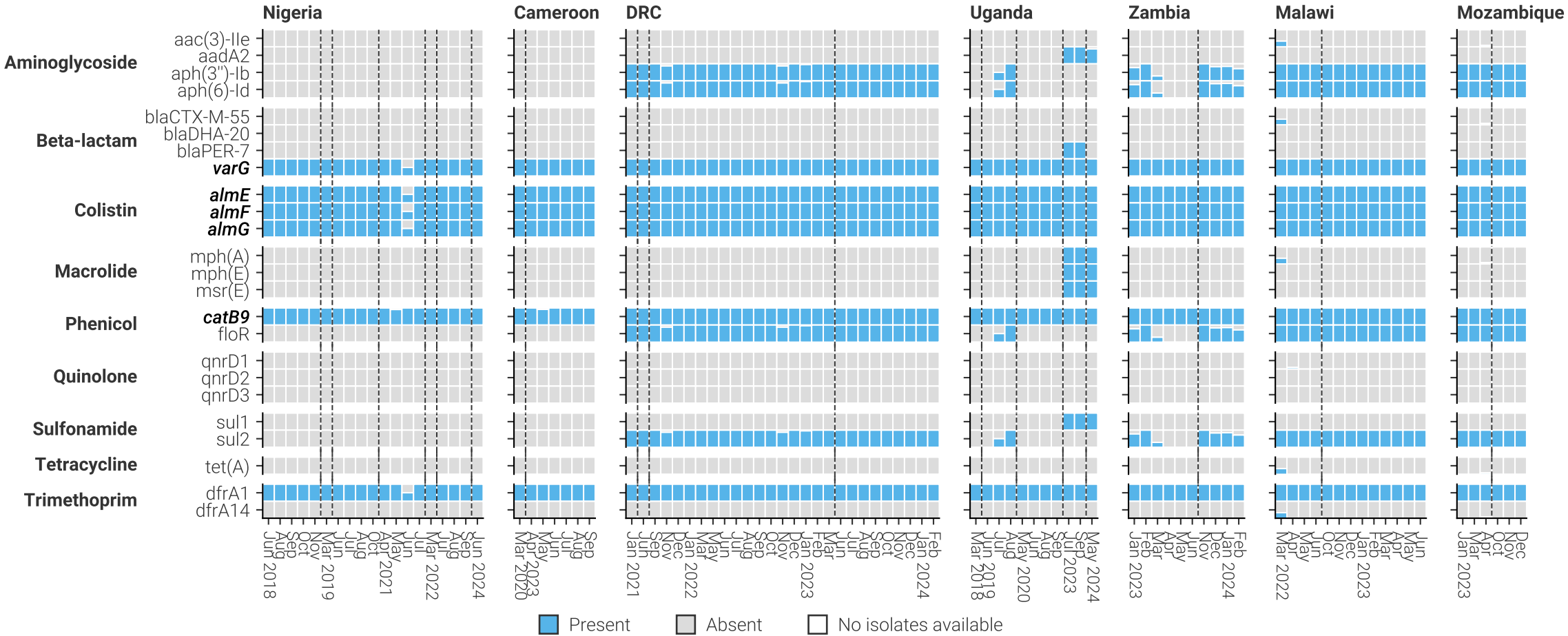

A
Ghana
Nigeria
Cameroon
DRC
Kenya
Uganda
Tanzania
Zambia
Malawi
Mozambique
South Africa
Median location transitions
B
Kenya-Tanzania
DRC-Zambia
Malawi-South Africa
Cameroon-Nigeria
Malawi-Zambia
Malawi-Mozambique
DRC-Tanzania
Tanzania-Uganda
Kenya-Uganda
Kenya-Zambia
DRC-Uganda
Tanzania-Zambia
Mozambique-Zambia
Ghana-Nigeria
Malawi-Tanzania
Location transitions density
Sequences
1995
2000
2005
2010
2015
2020
C
R^2: 0.48
P < 0.001
Location transitions
Sequences

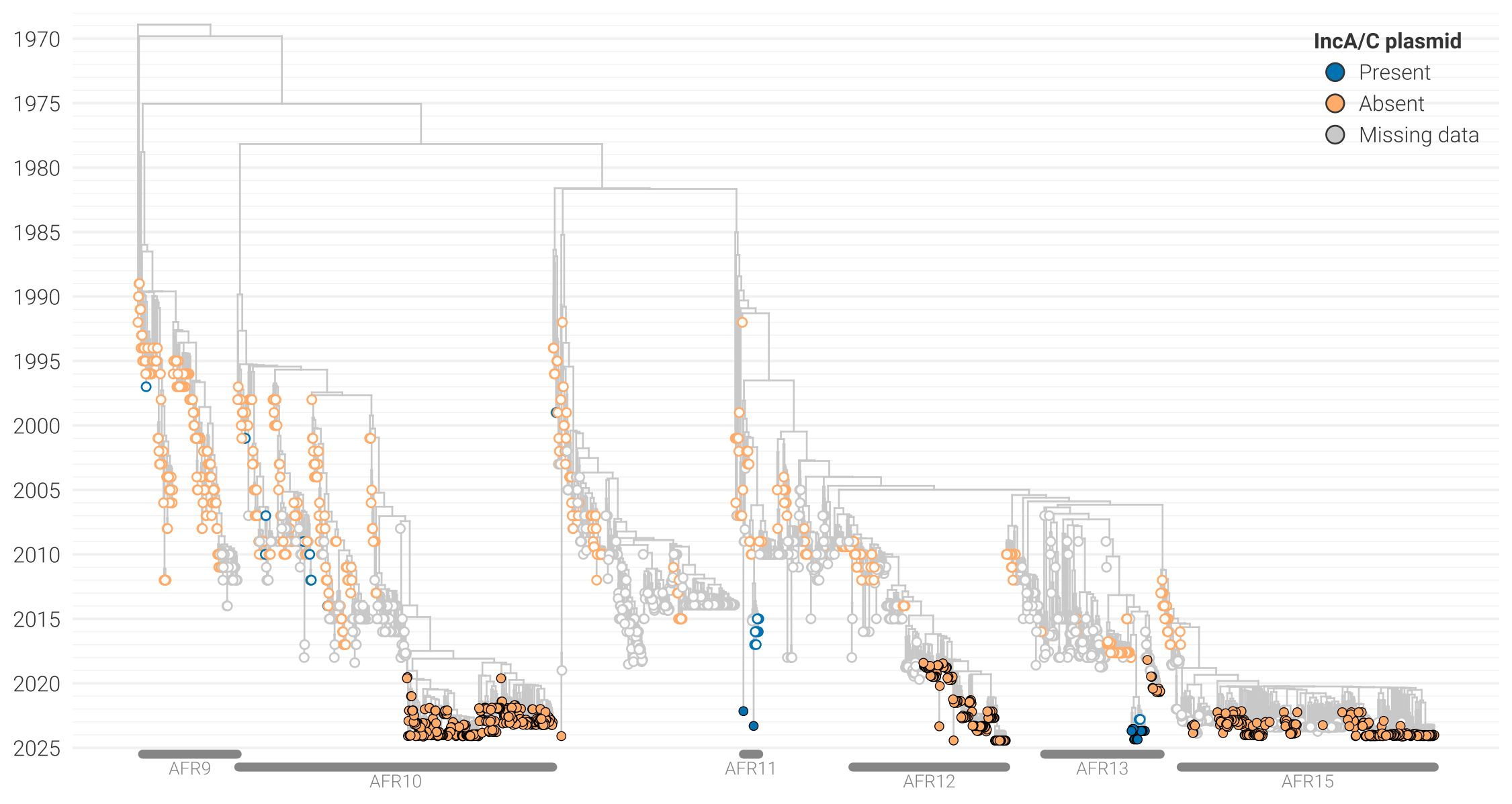

A
Cameroon
Sampling density
2018 2019 2020 2021 2022 2023 2024
Phylogenetic diversity gained per sequence * 10−6
0.0 0.5 1.0 1.5
Observed introductions
0 5 10 15 20 25 30
Captured lineages = 72.8% (63.7–81.7%)
0 100 200 300
Sequences
B
Nigeria
Sampling density
2018 2019 2020 2021 2022 2023 2024
Phylogenetic diversity gained per sequence * 10−6
0.0 0.5 1.0 1.5
Observed introductions
0 5 10 15 20 25 30
Captured lineages = 59.4% (42.3–73.8%)
0 100 200 300
Sequences
C
DRC
Sampling density
2018 2019 2020 2021 2022 2023 2024
Phylogenetic diversity gained per sequence * 10−6
0.0 0.5 1.0 1.5
Observed introductions
0 5 10 15 20 25 30
Captured lineages = 93.3% (84.4–99.7%)
0 100 200 300
Sequences
D
Uganda
Sampling density
2018 2019 2020 2021 2022 2023 2024
Phylogenetic diversity gained per sequence * 10−6
0.0 0.5 1.0 1.5
Observed introductions
0 5 10 15 20 25 30
Captured lineages = 68.8% (57.4–80.0%)
0 100 200 300
Sequences
E
Zambia
Sampling density
2018 2019 2020 2021 2022 2023 2024
Phylogenetic diversity gained per sequence * 10−6
0.0 0.5 1.0 1.5
Observed introductions
0 5 10 15 20 25 30
Captured lineages = 83.5% (78.1–88.2%)
0 100 200 300
Sequences
F
Malawi
Sampling density
2018 2019 2020 2021 2022 2023 2024
Phylogenetic diversity gained per sequence * 10−6
0.0 0.5 1.0 1.5
Observed introductions
0 5 10 15 20 25 30
Captured lineages = 66.8% (51.1–85.9%)
0 100 200 300
Sequences
G
Mozambique
Sampling density
2018 2019 2020 2021 2022 2023 2024
Phylogenetic diversity gained per sequence * 10−6
0.0 0.5 1.0 1.5
Observed introductions
0 5 10 15 20 25 30
Captured lineages = 75.0% (62.8–86.1%)
0 100 200 300
Sequences


0.000
0.002
0.004
0.006
0.008
0.010
0.012
Lineage
AFR9
AFR10
AFR11
AFR12
AFR13
AFR15
AFR16
AFR17
Other
AFR9
AFR10
AFR16
AFR11
AFR17
AFR12
AFR13
AFR15


CholGen sequences
Rest of Africa and Asia
Clock rate outliers
Substitution rate: 1.6e-06 subs/site/year
R2: 0.82
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Per site genetic distance from root
1960
1970
1980
1990
2000
2010
2020
Collection date


All branches
Internal branches
Transmission event lineages
AFR15
AFR13
AFR12
AFR11
AFR10
AFR9
10^{-8}
10^{-7}
10^{-6}
10^{-5}
Substitutions / site / year


Nigeria
Cameroon
DRC
Uganda
Zambia
Malawi
Mozambique
Aminoglycoside
aac(3)-IIe
aadA2
aph(3'')-Ib
aph(6)-Id
Beta-lactam
blaCTX-M-55
blaDHA-20
blaPER-7
varG
Colistin
almE
almF
almG
Macrolide
mph(A)
mph(E)
msr(E)
Phenicol
catB9
floR
Quinolone
qnrD1
qnrD2
qnrD3
Sulfonamide
sul1
sul2
Tetracycline
tet(A)
Trimethoprim
dfrA1
dfrA14
Present
Absent
No isolates available


IncA/C plasmid
Present
Absent
Missing data
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
AFR9
AFR10
AFR11
AFR12
AFR13
AFR15