Combining epidemiologic and genomic data to better understand cholera transmission in Africa

Bethany L. DiPrete, PhD
Gillings School of Global Public Health
University of North Carolina at Chapel Hill

2024 **IDM Annual Symposium**
Global public health in a chaotic world: The role of modeling & data science
Session 1A: Genomics and Environmental Surveillance
Bill & Melinda Gates Foundation
October 1, 2024

UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

Infectious Disease
DYNAMICS
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Background

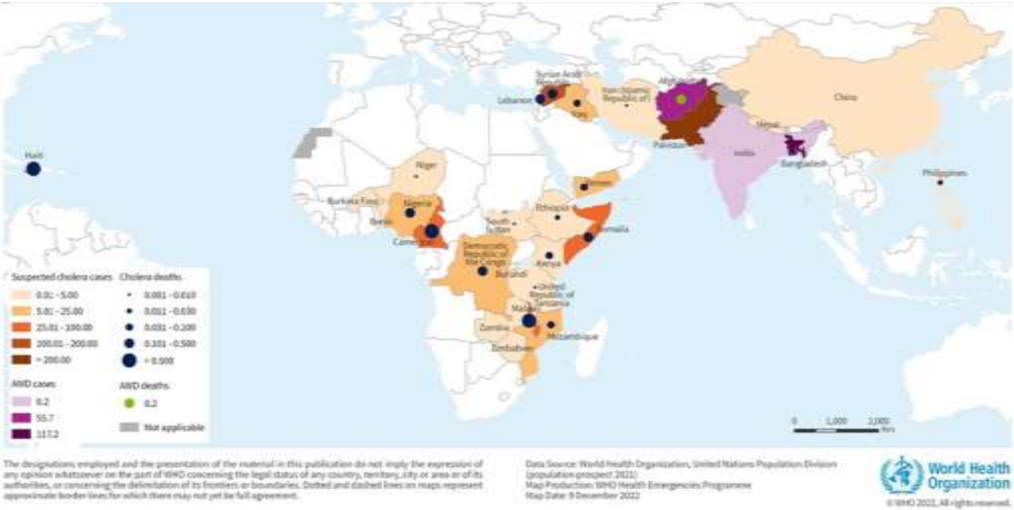
- Cholera, an acute gastrointestinal infection caused by the bacterium *Vibrio cholerae*, causes severe illness and, if untreated, death.

- From the 1800s to the early 1920s,

  there were six known cholera pandemics.

Background

- The seventh cholera pandemic began in the early 1960s and continues to cause significant morbidity and mortality globally.



The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of WHO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

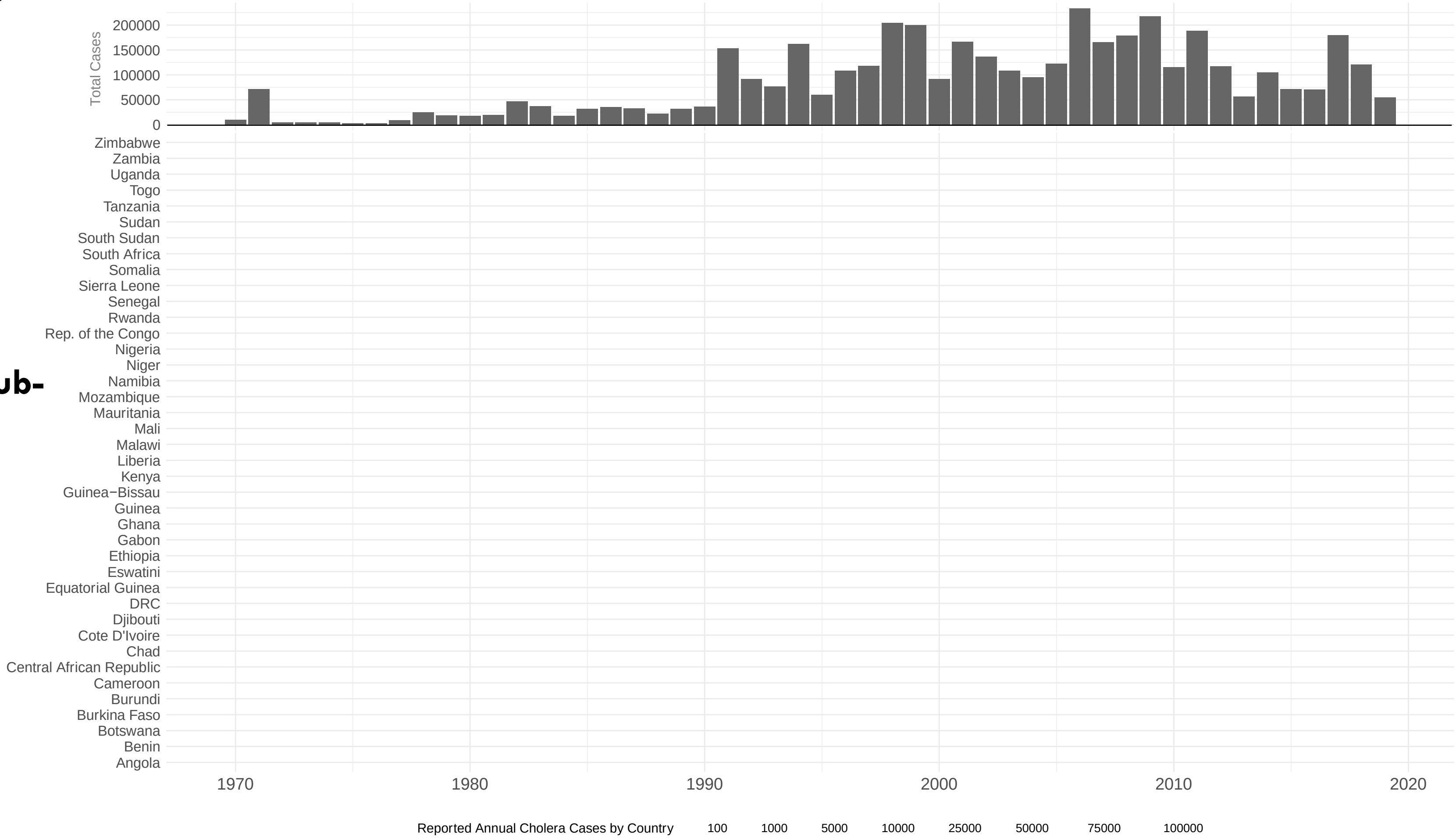
Data Source: World Health Organization, United Nations Population Division (population prospect 2021)
Map Production: WHO Health Emergencies Programme
Map Date: 8 December 2022

World Health Organization

Most of the burden of cholera is concentrated in sub-Saharan Africa, with South Asia also accounting for a significant proportion of the global cholera burden.

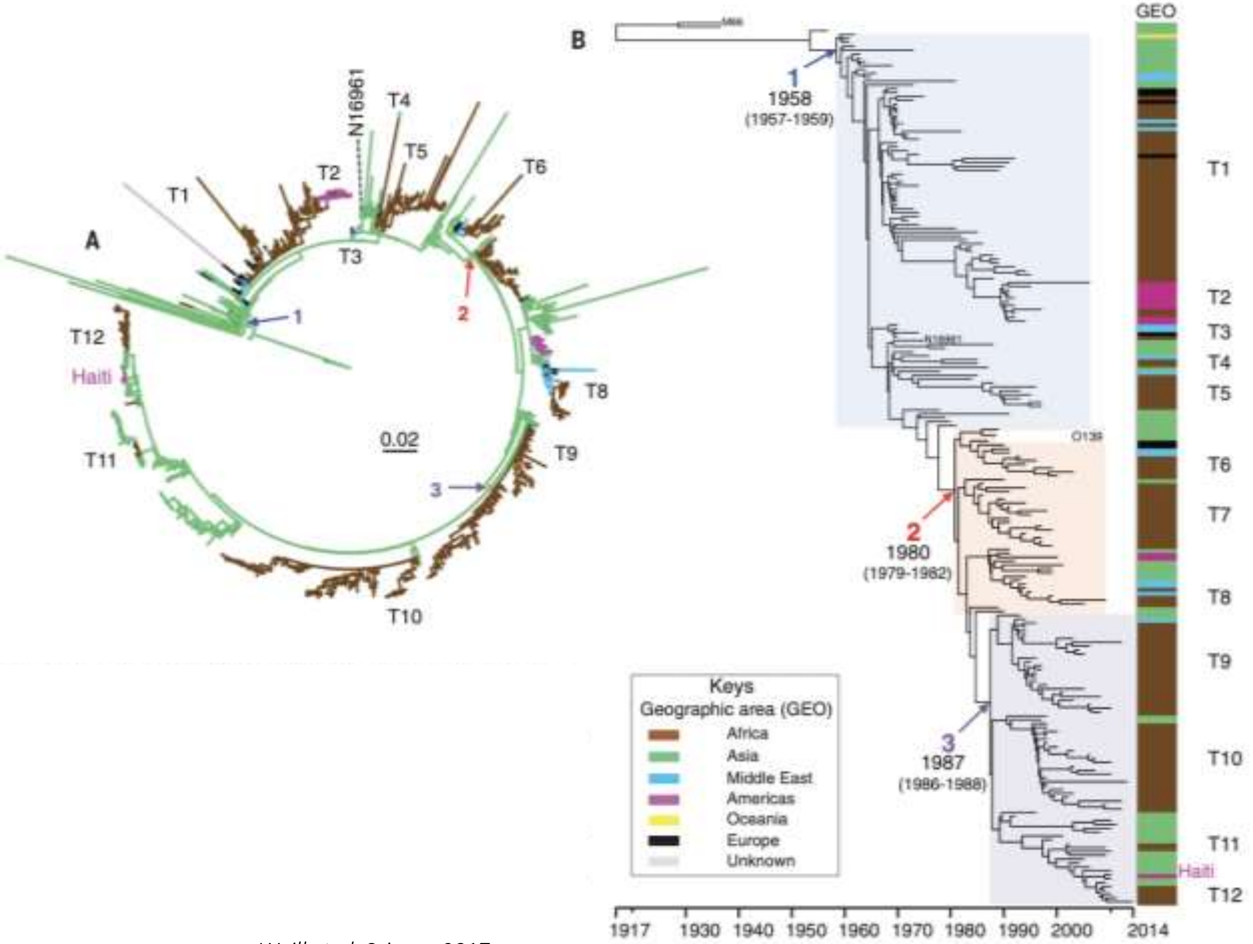
Background

Cholera burden in sub-Saharan Africa



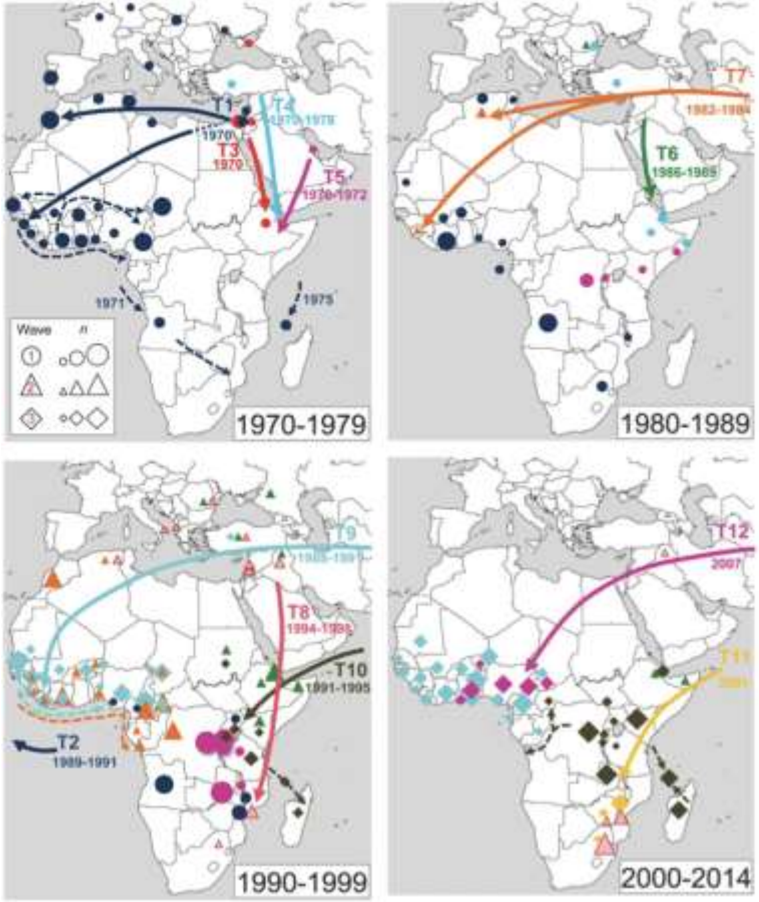
Background

Recent phylogenetic analysis found distinct introduction events into Africa



Weill et al. Science 2017

Background



Based on these findings, authors inferred propagation routes of seventh pandemic *V. cholerae* O1 El Tor in the African continent

Weill et al. Science 2017

Motivation

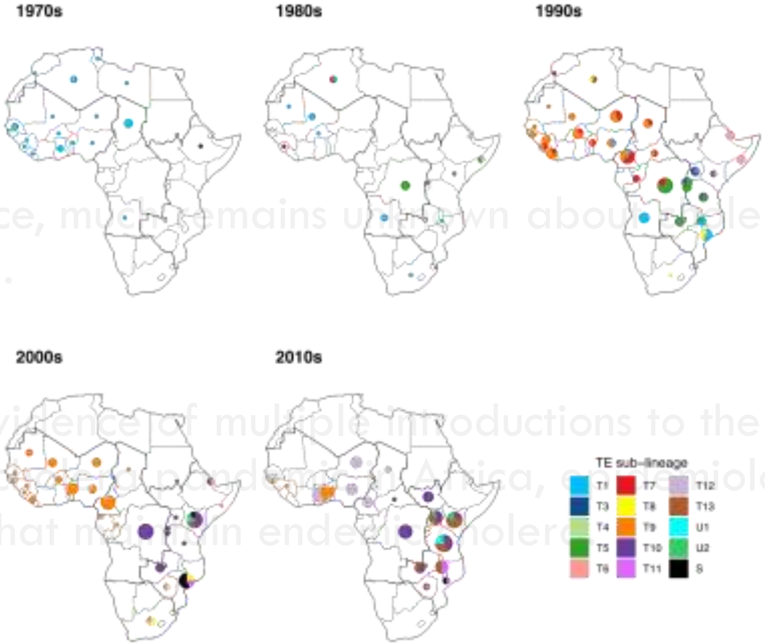
- Despite recent evidence, much remains unknown about cholera transmission dynamics within Africa.

Motivation

- Despite recent evidence, much remains unknown about cholera transmission dynamics within Africa.

- While there is clear evidence of multiple introductions to the continent that have helped sustain the seventh cholera pandemic in Africa, epidemiologic data suggests that there are also areas that maintain endemic cholera.

Motivation

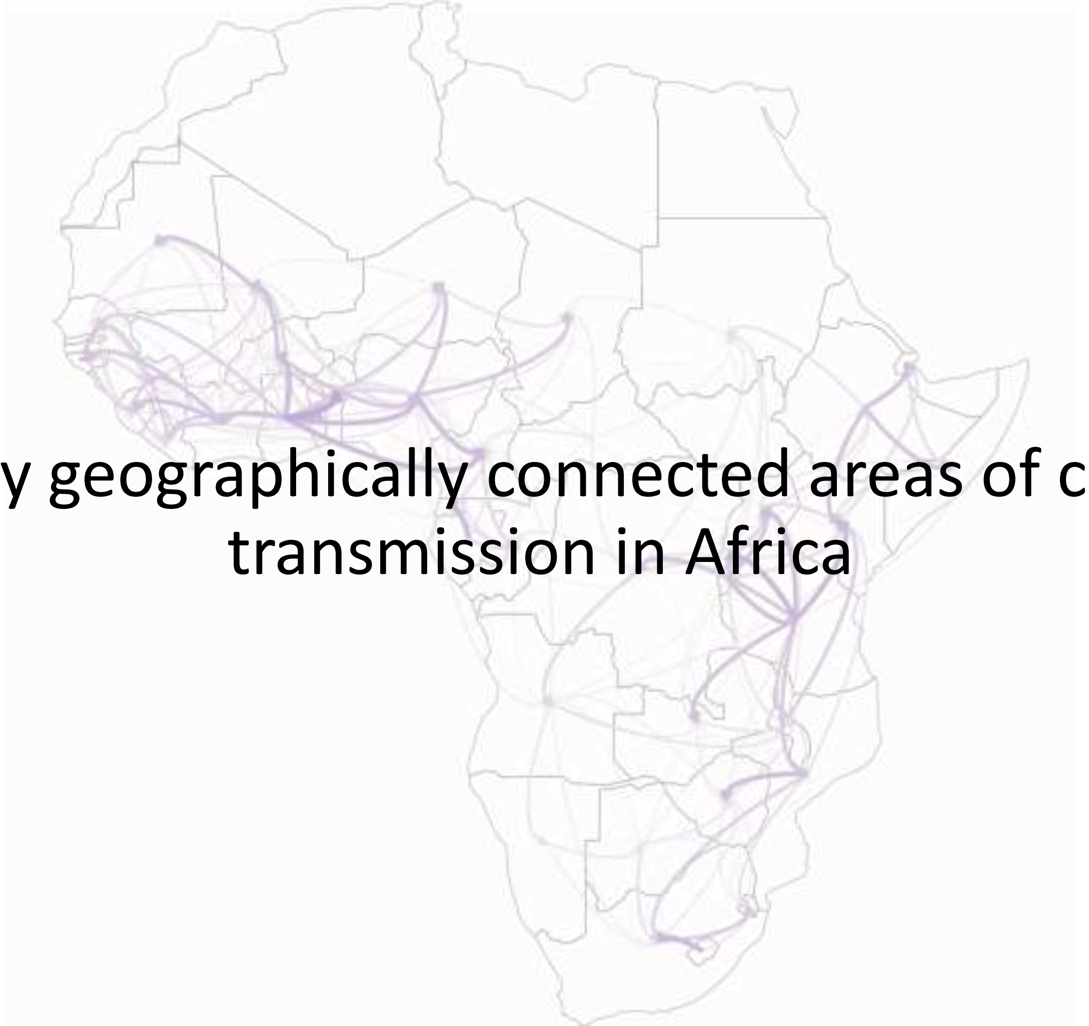


- Despite recent evidence, much remains unknown about cholera transmission dynamics within Africa.

- While there is clear evidence of multiple introductions to the continent that have sustained the seventh cholera pandemic in Africa, epidemiologic data suggests that there are also areas that maintain endemic cholera.

- Connected areas likely have correlated transmission dynamics. These basic epidemiologic units of transmission may:
  - Propagate outbreaks from intercontinental introductions
  - Maintain endemic circulation that seed outbreaks elsewhere on the continent

Objective



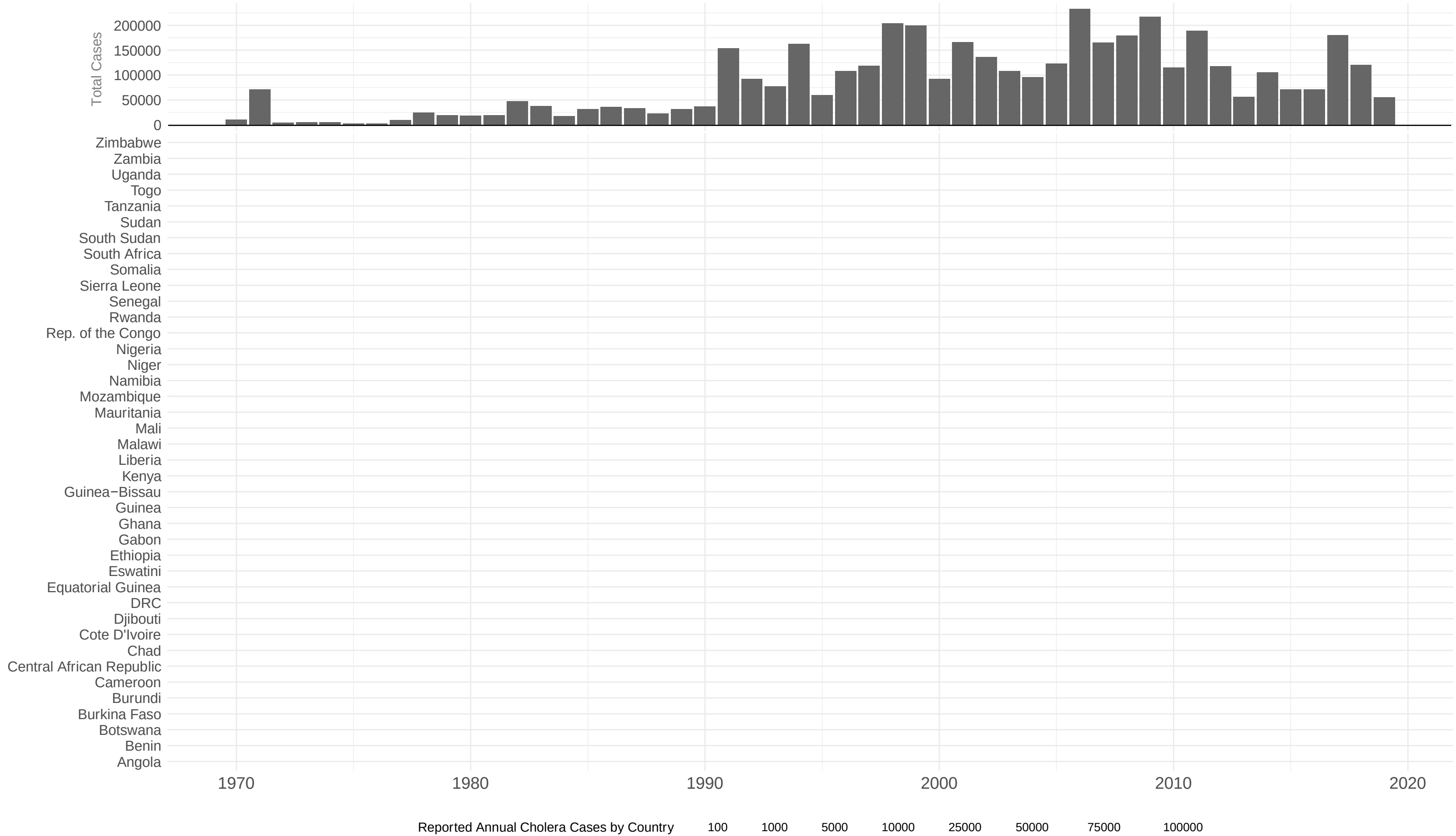
Identify geographically connected areas of cholera transmission in Africa

Data sources

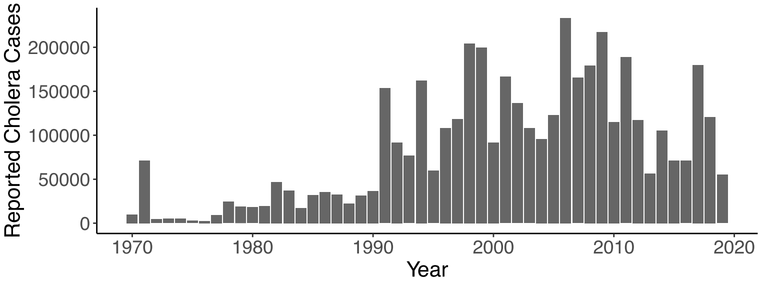
Combined molecular data with epidemiologic data of cholera incidence in sub-Saharan Africa from 1970-2020

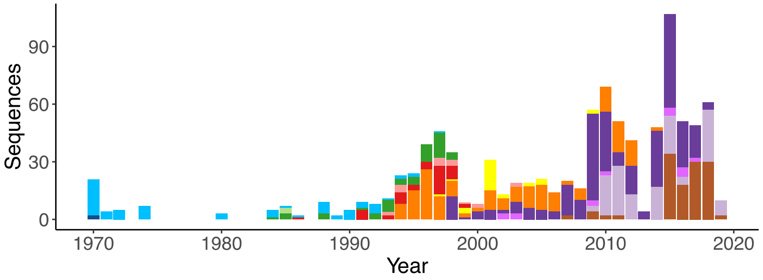
- Molecular data:
  - Publicly available cholera sequencing data from open-source repositories
    - *E.g., GenBank*
  - Metadata contains year and country for each sequenced sample
- Epidemiologic data
  - WHO reported cholera case counts aggregated by year and country

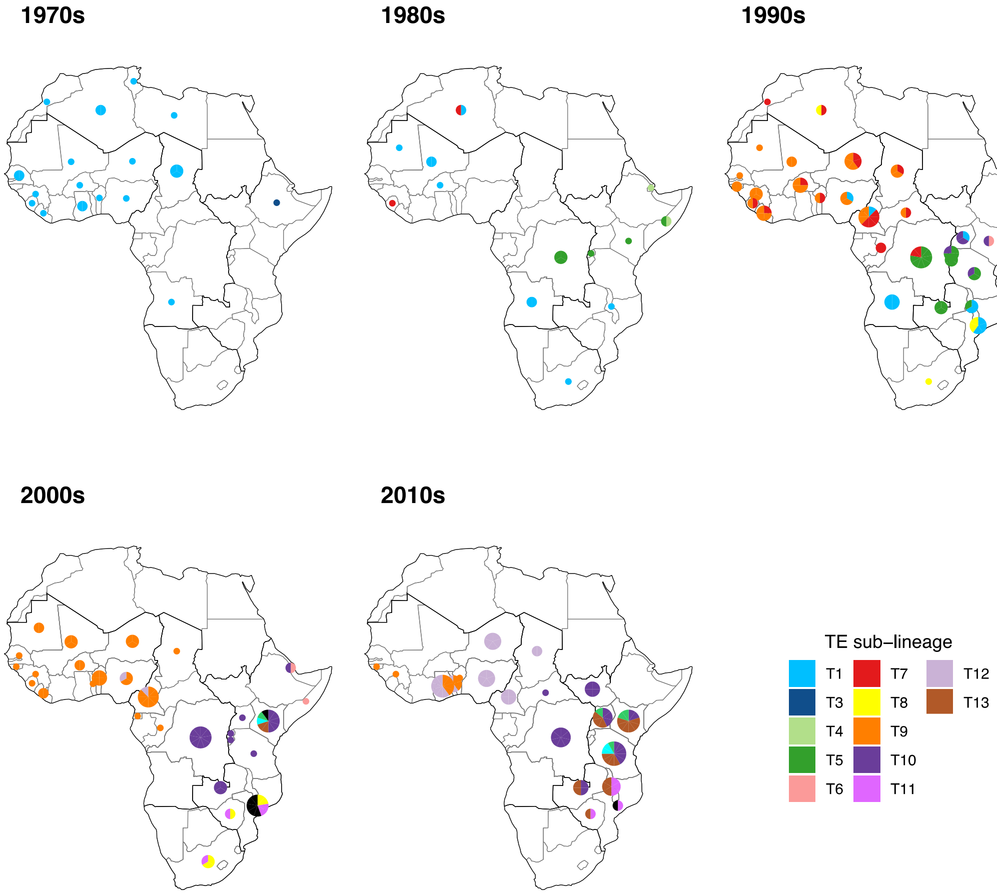
WHO Reported Cases



Publicly available sequence data

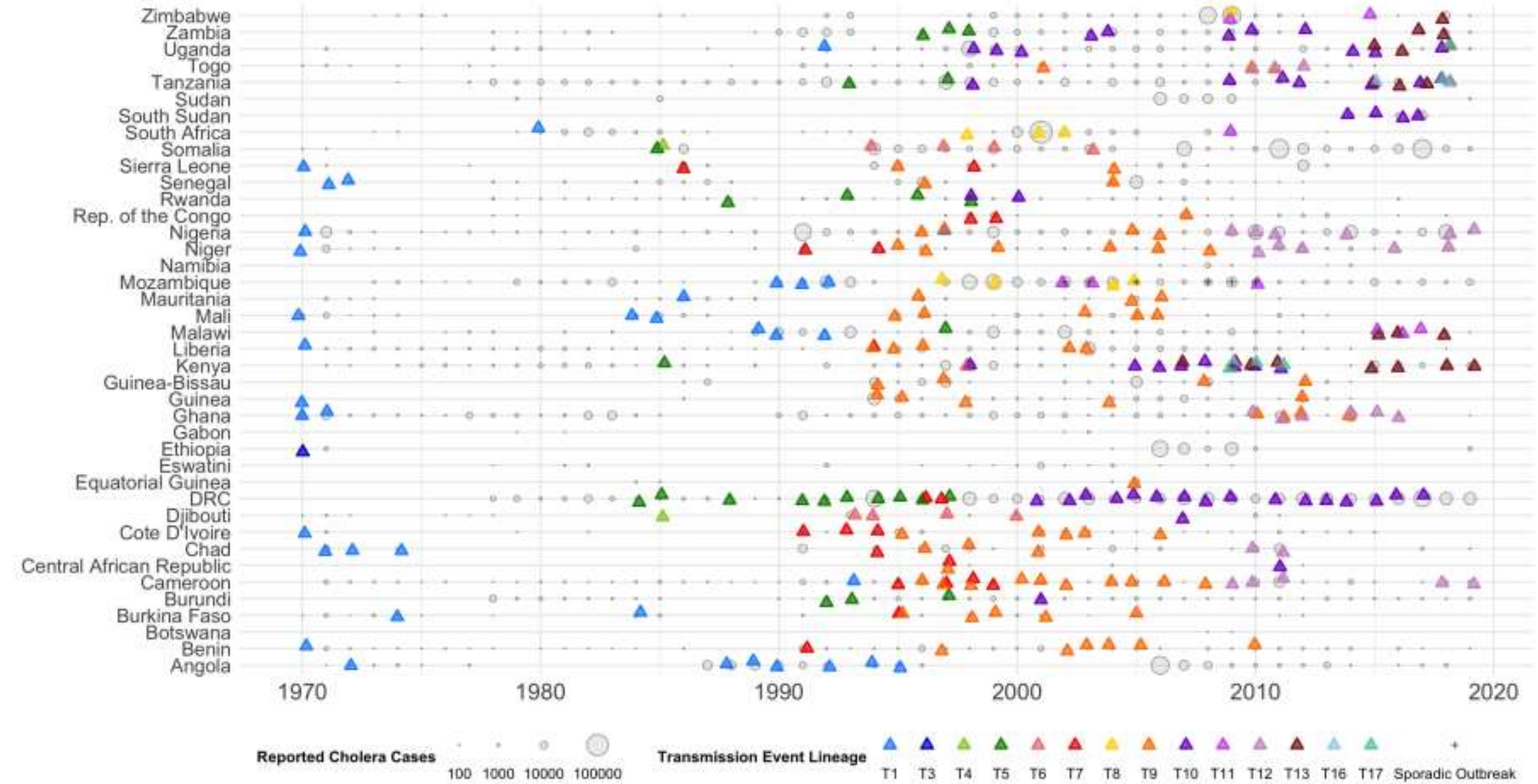




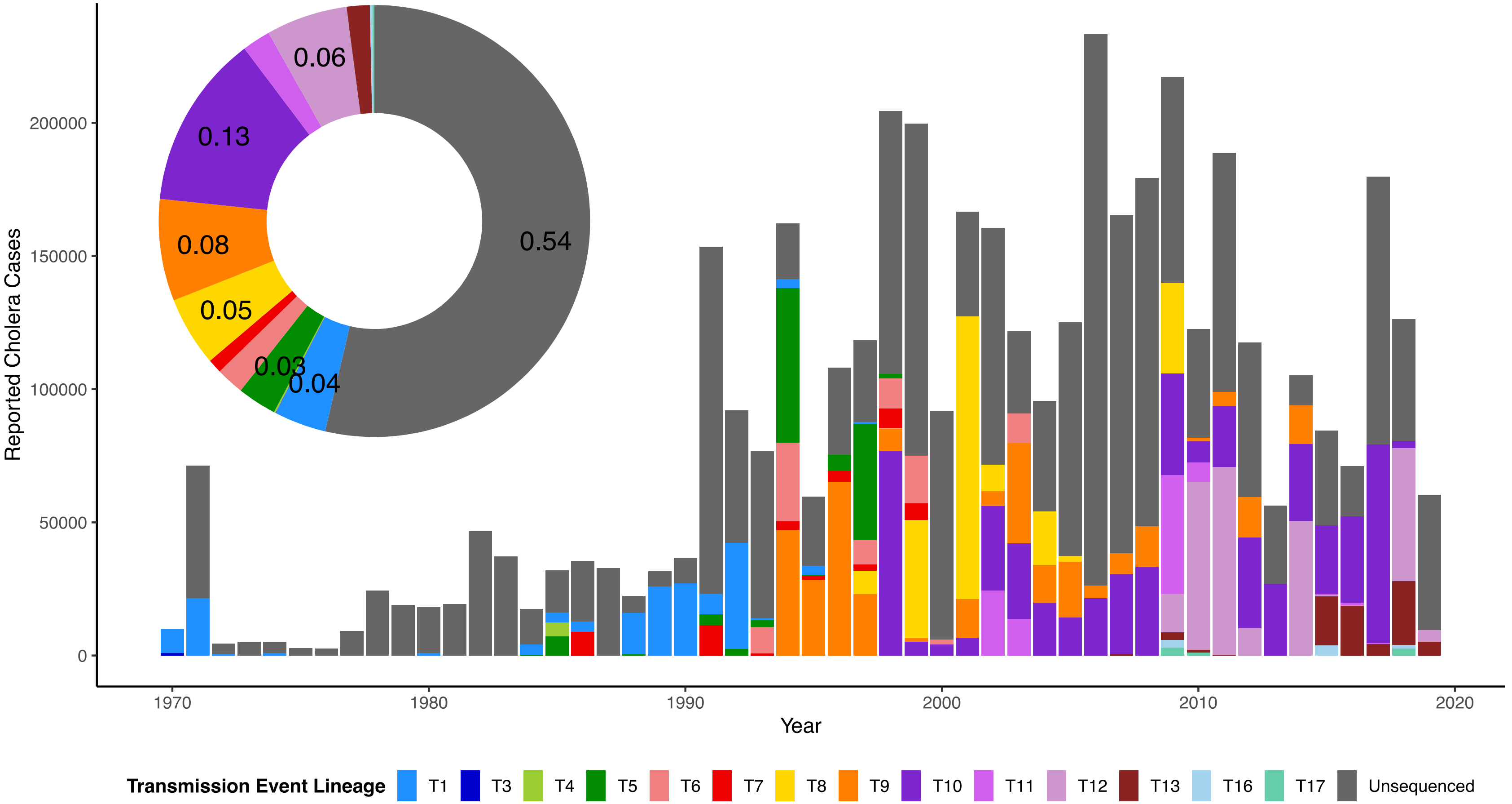


Gaps in observed data

Gaps in observed data



Gaps in observed data



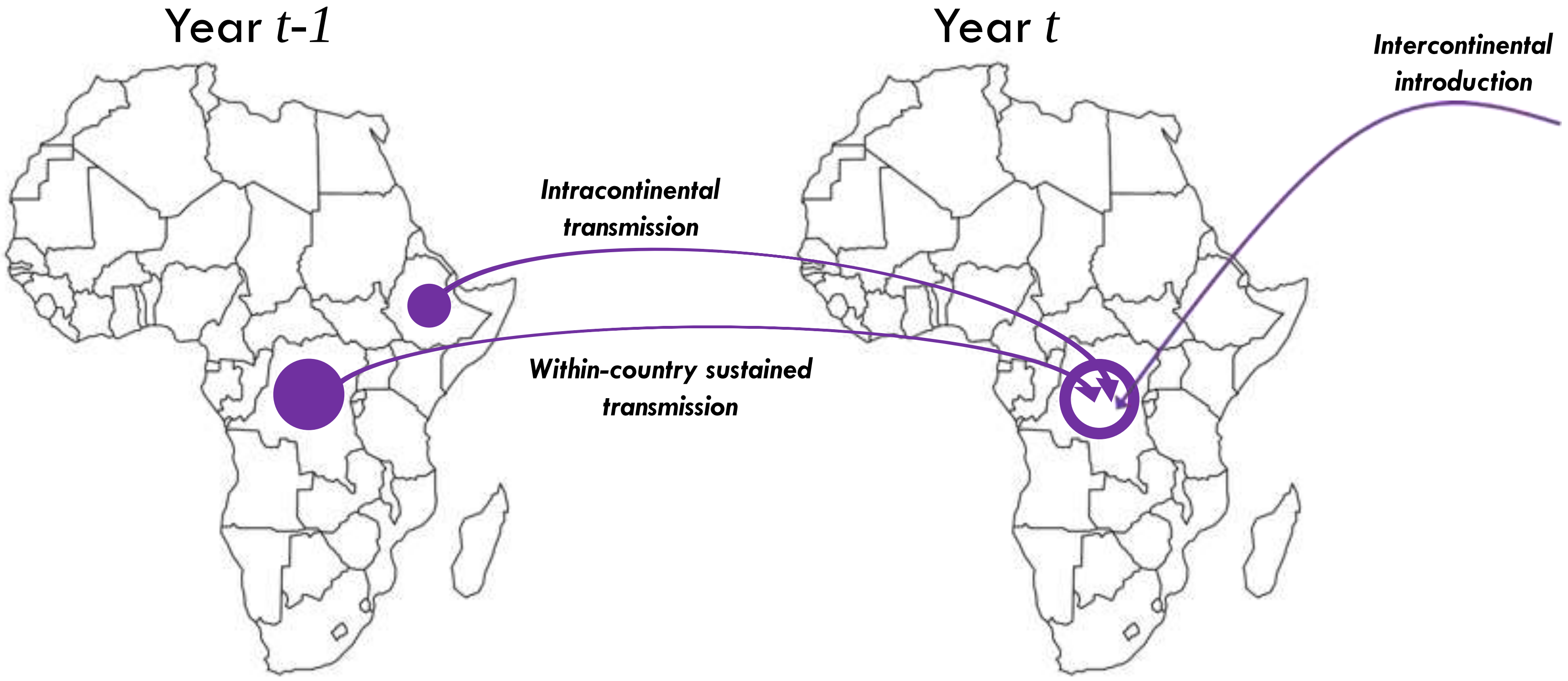
Inferring occurrence & prevalence of cholera sub-lineages to define epidemiologically relevant transmission units

Approach

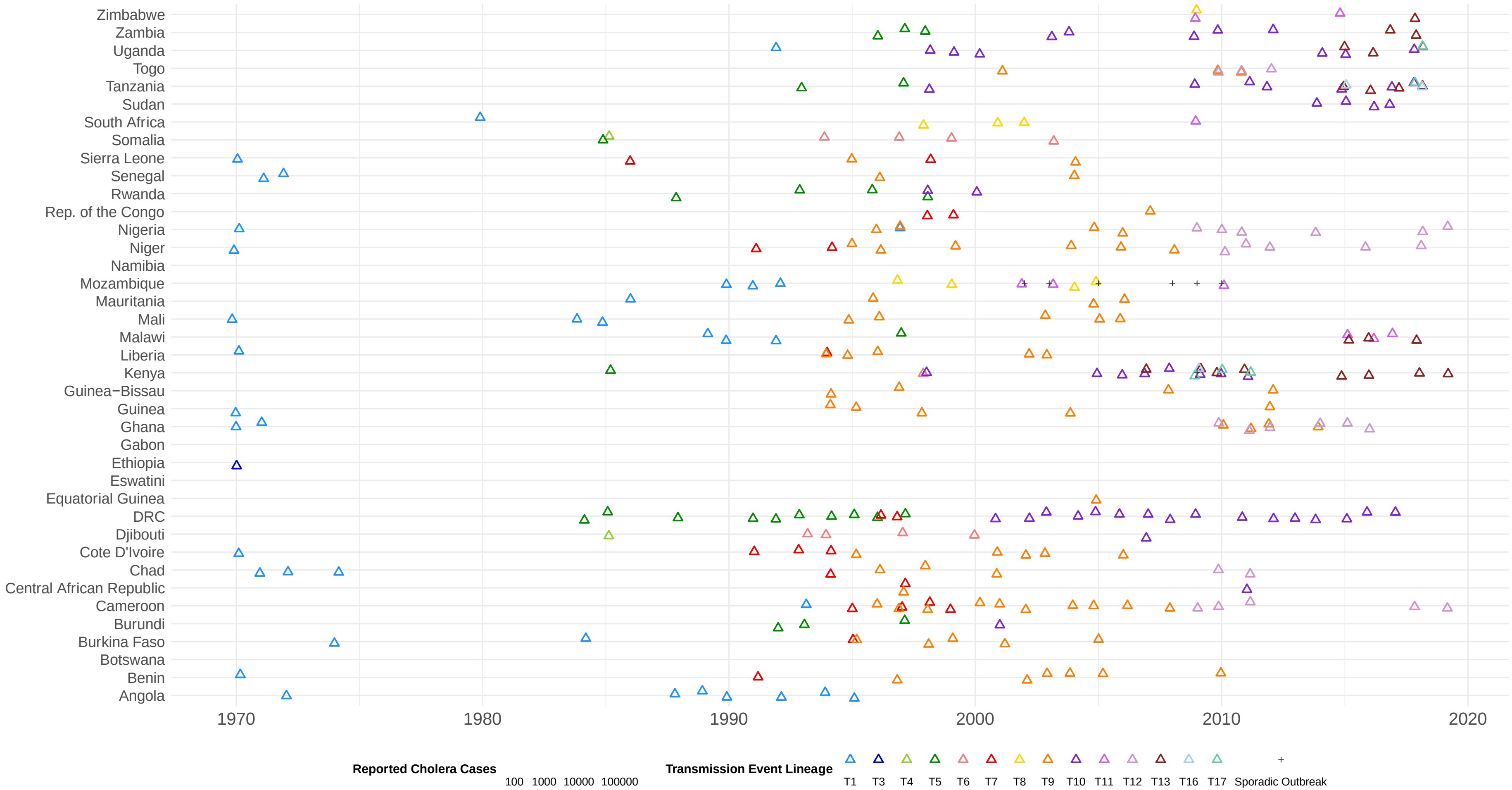
- Model occurrence and prevalence of distinct cholera sub-lineages in countries through time using a Hidden Markov Model.
  - Accounting for historical information of cholera presence

- Targets of inference:
  - strength of connectivity driving transmission between locations
  - underlying occurrence and prevalence of cholera sub-lineages in each country in all years

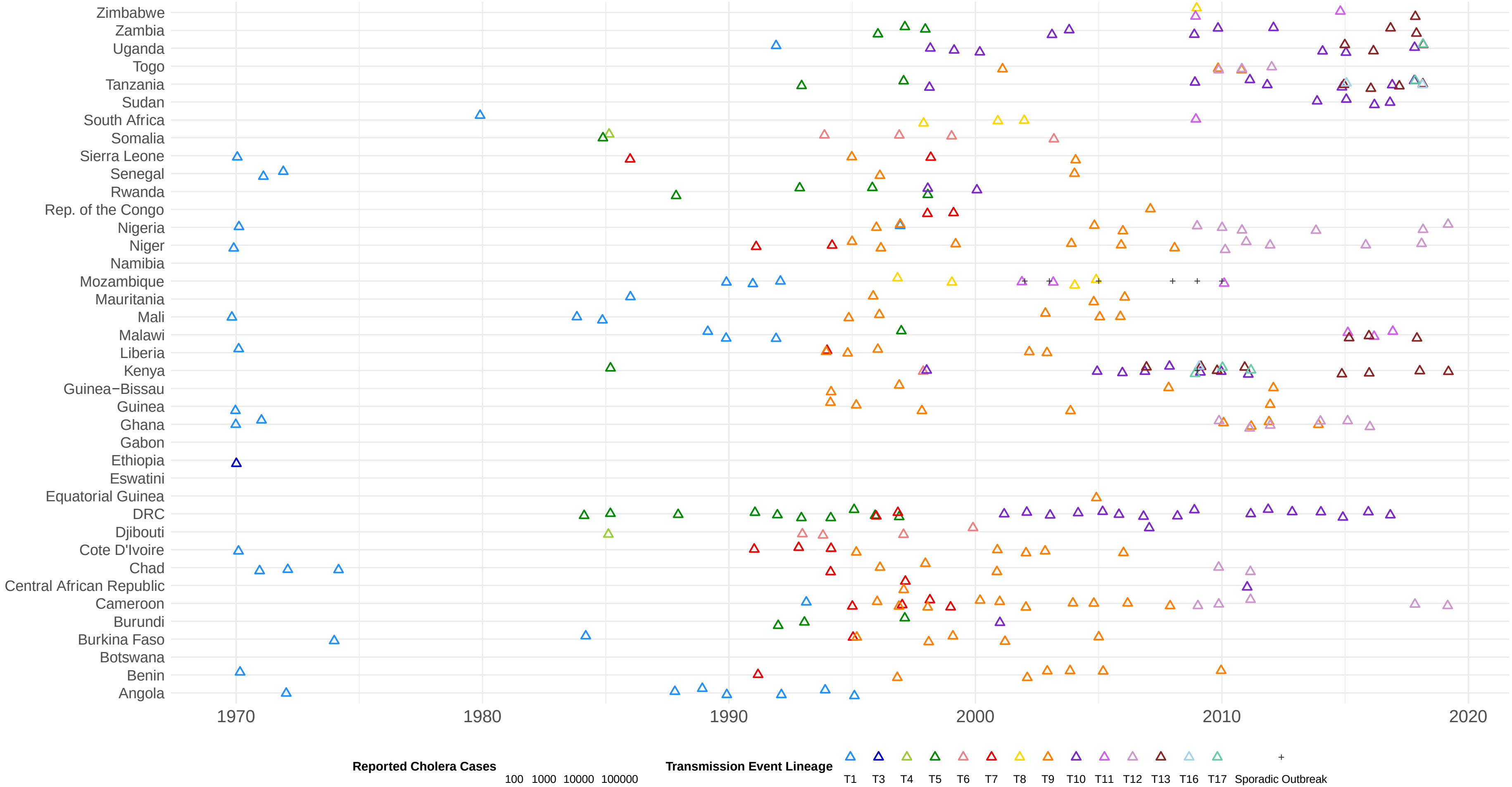
Transition Process


Year *t-1*
Year *t*
Intercontinental introduction
Intracontinental transmission
Within-country sustained transmission

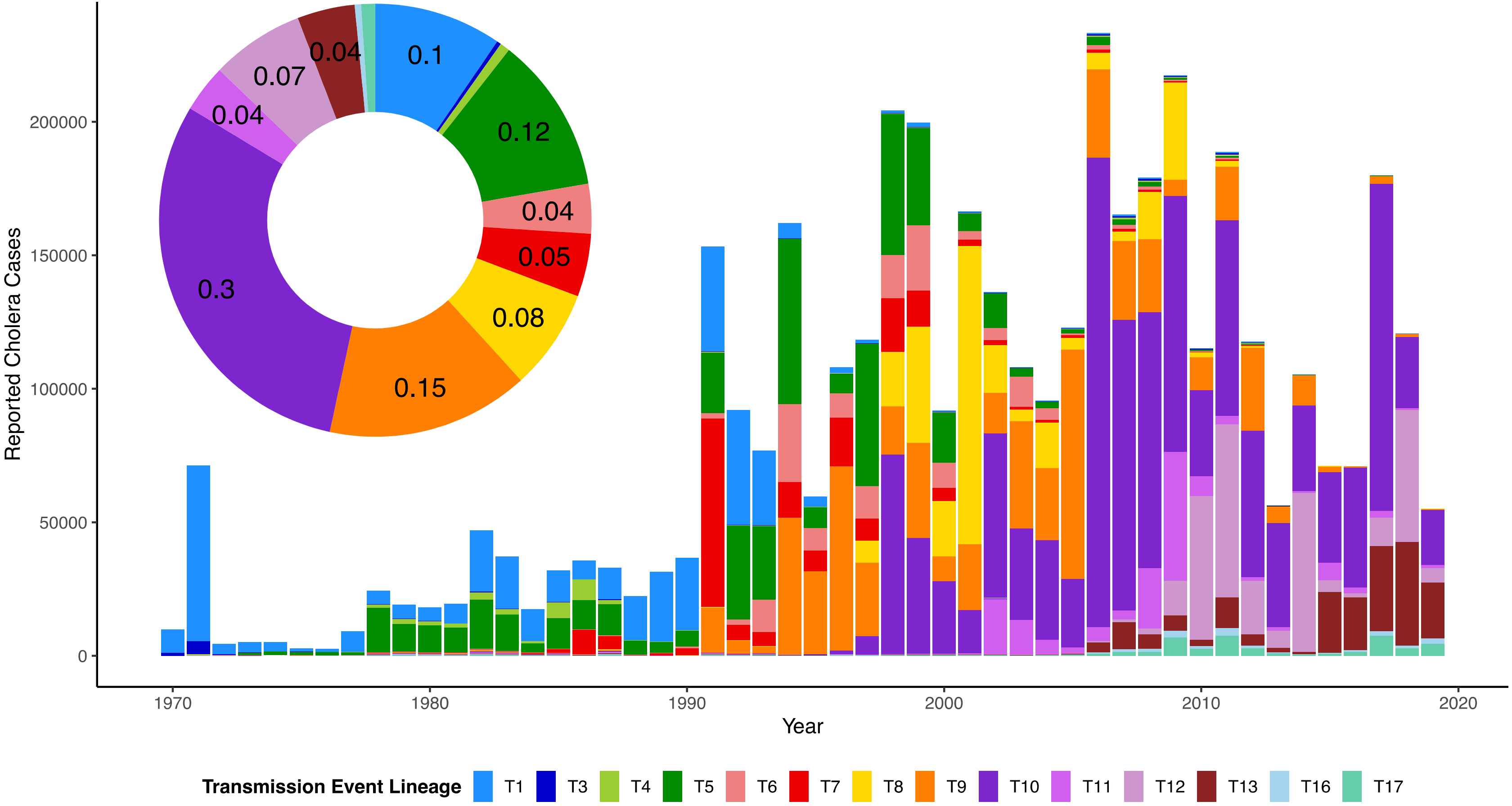
Filling in the gaps: Observed data



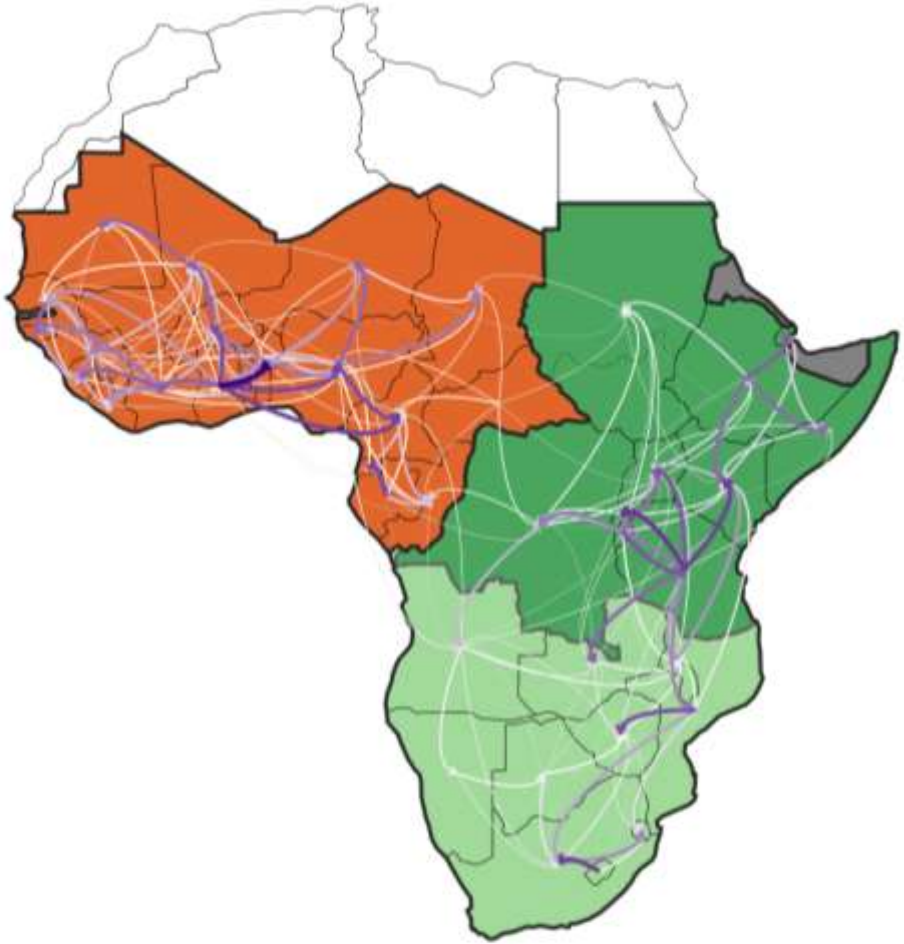
Filling in the gaps: Inferred sub-lineage presence

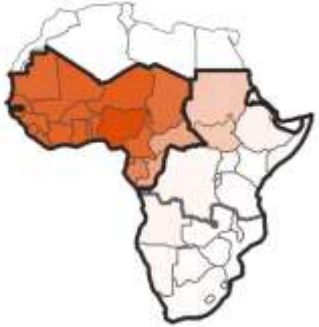


Filling in the gaps: Inferred prevalence

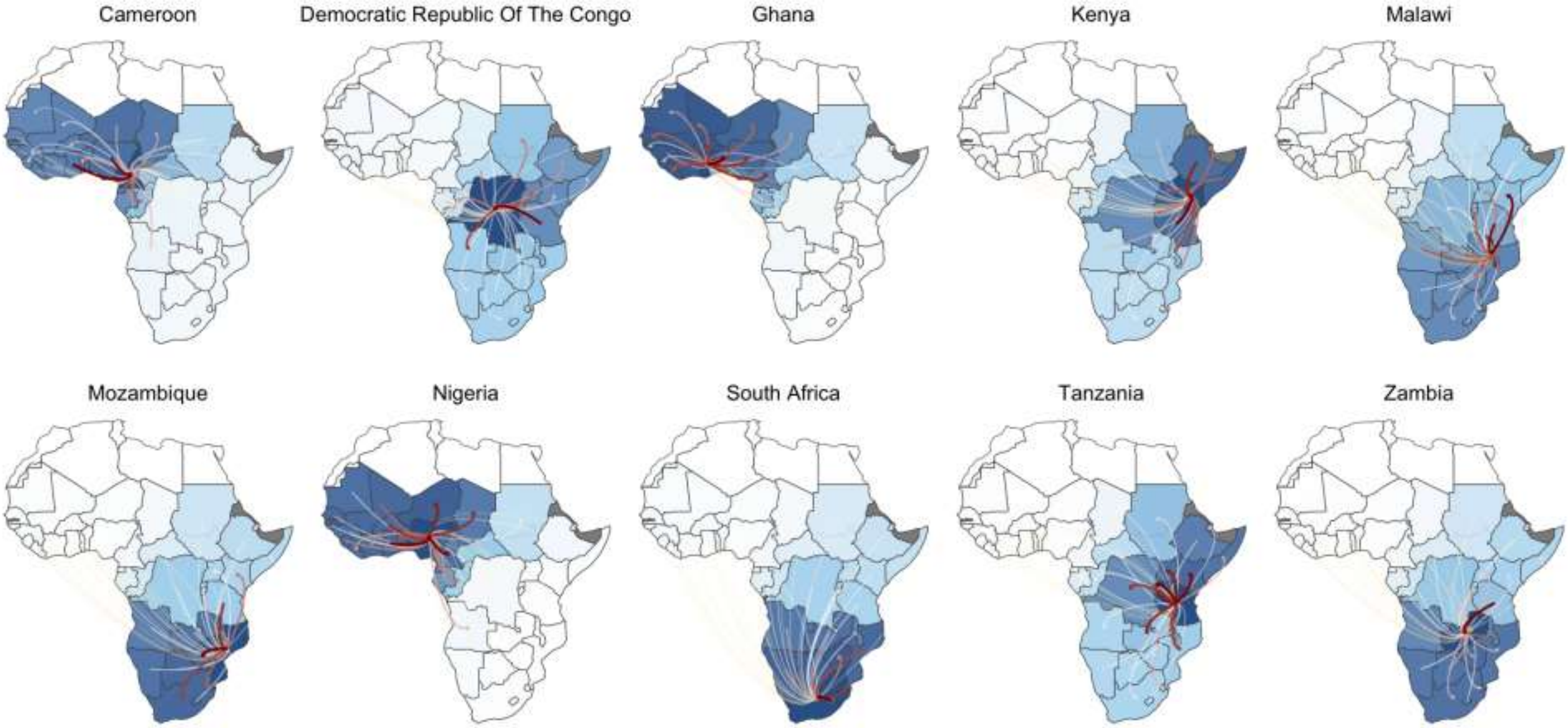


Inferred Connectivity





Inferred Connectivity



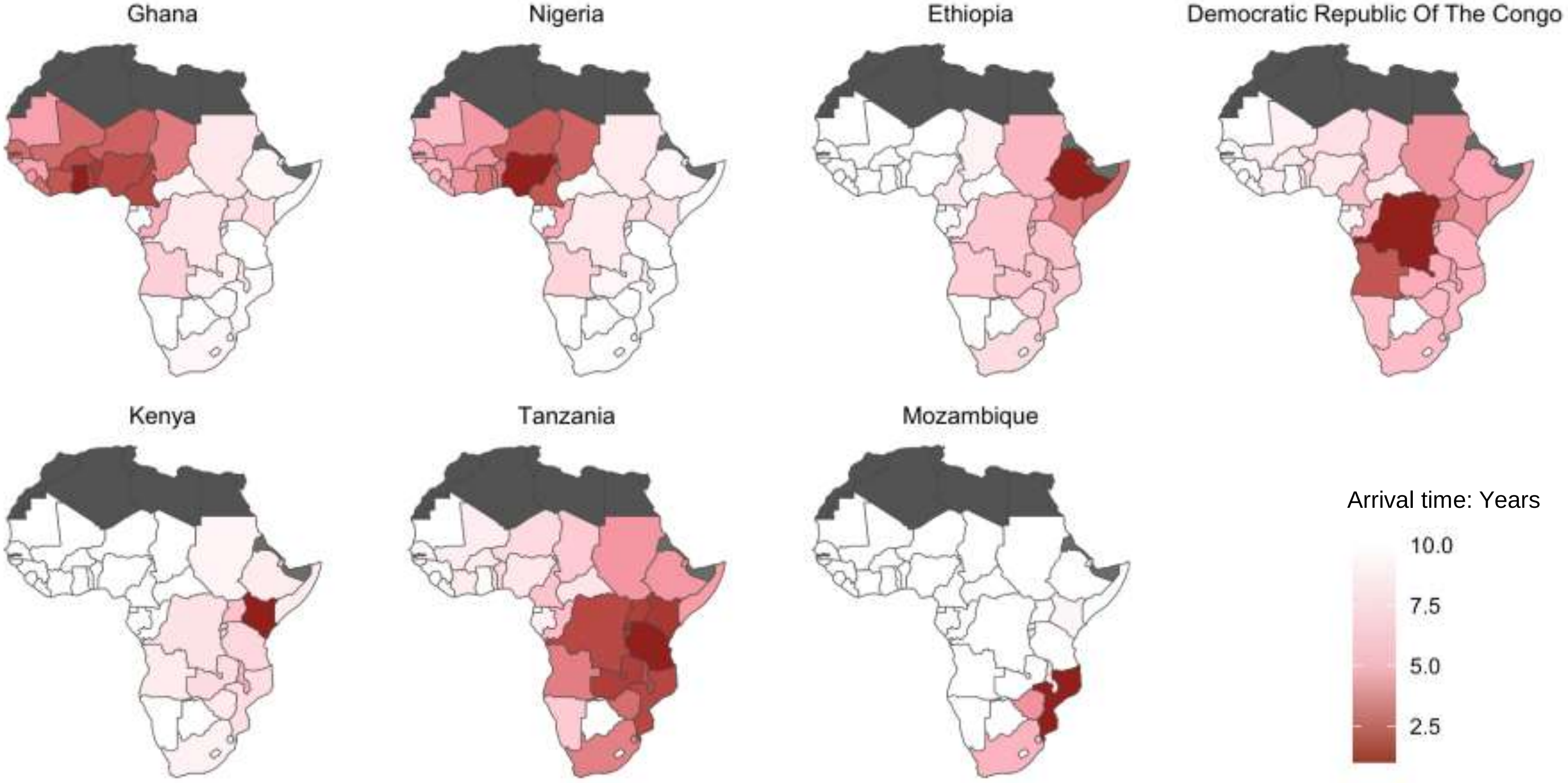
Using inferred connectivity to predict the downstream effects of a new introduction

Downstream Impacts

- We simulated the spread of a new lineage after introduction to potential seed countries.
  - Based on the same transition and observation process from the HMM and using the inferred connectivity measures from the HMM.

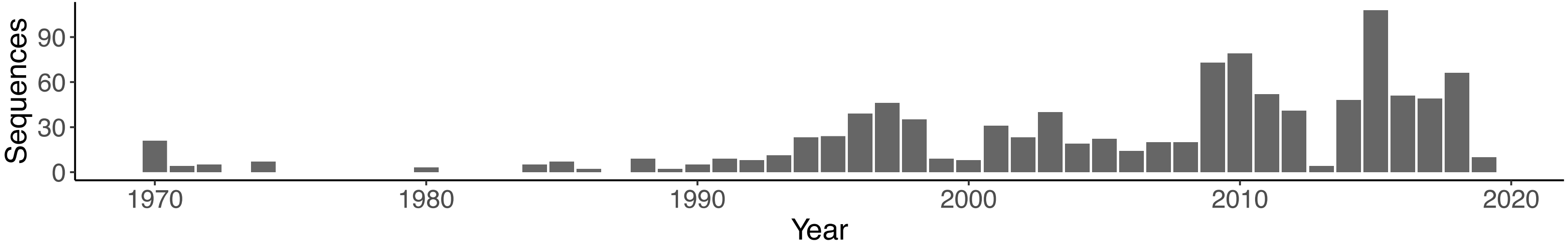
- From this simulation, we determined the mean time to arrival in each country following introduction into a single country.

Downstream impacts



Limitations

- Ultimately, sequencing remains sparse and cholera cases are often under-reported.
  - Areas with extremely sparse data can impact the ability of our model to infer underlying presence of distinct sub-lineages.



- Additional sequencing efforts can help improve our understanding of phylodynamic processes driving cholera transmission in Africa.

Implications & Future Directions

- Transmission units informing cholera control:
  - Proactive intervention:
    - identify areas where increases in cases → increase in local cholera risk in connected areas
  - Maximize indirect effects:
    - targeted vaccination and water/sanitation campaigns

- Assess drivers of cholera endemicity to determine the influence of new and re-introductions versus local undetected persistence

Image credit @WHOSouthSudan, accessed from https://www.afro.who.int/news/ministry-health-support-who-and-partners-vaccinated-over-63-000-people-oral-cholera-vaccine

ACKNOWLEDGEMENTS

Infectious Disease
DYNAMICS
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

UNC - Chapel Hill

Justin Lessler

Johns Hopkins University

Javier Perez-Saez

Andrew Azman

Brigham & Women's Hospital

Shirlee Wohl

Nathaniel Matteson

BILL & MELINDA
GATES *foundation*

The content of this presentation is solely the responsibility of the authors and does not necessarily represent the official views of the Bill & Melinda Gates Foundation.
The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the submitted abstract.

Bethany L. DiPrete, PhD · diprete@email.unc.edu
@bethanydiprete · LinkedIn: bethany-diprete

Thank You

UNC
GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Infectious Disease
DYNAMICS
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Supplement

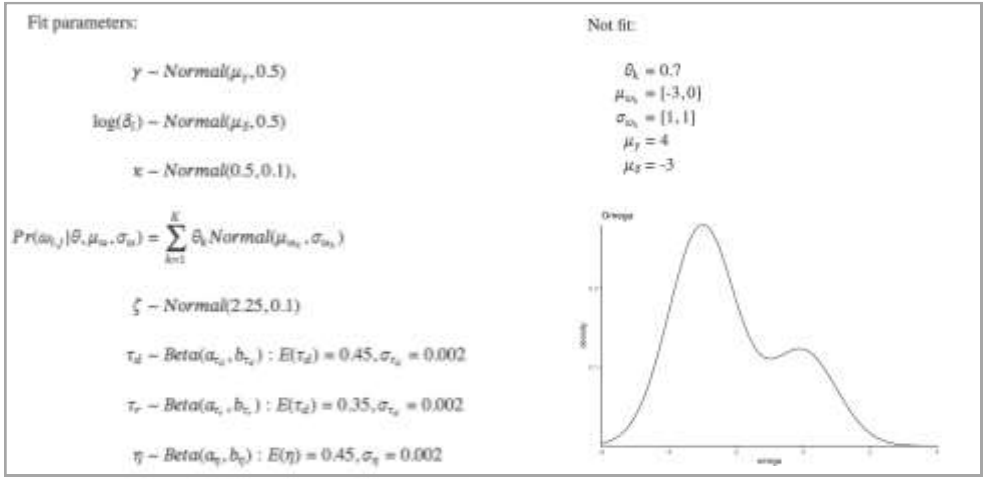
Transition process

Strain presence, $\rho_{v,i,t}$, is based on the transition process, Φ, of the probability of establishment in country i at time t, $\phi^{k,l}_{v,i,t}$.

$$\Phi^{g,k} = Pr(z_{v,i,t} = k | z_{v,i,t-1} = g)$$

The transition process, $\Phi_{g,k}$, is based on the transition matrix, Φ, and is a function of:

- $\gamma_{v,i,t-1}$, introduction rate from outside the continent
- $\xi_{i,j}$, connectivity between locations i and j, based on:
    - $\omega_{i,j}$, spatial weight (random effect), $i \neq j$
    - $d_{i,j}$, distance (km), $i \neq j$
    - pop_i, population sizes of countries i and j, $i \neq j$
    - δ, persistence of strain once it has been introduced, $i = j$
- $c_{i,t-1}$, cases in location j in the previous year
- $\lambda^*_{v,i,t-1}$, strain-specific prevalence



Fit parameters:

$$\gamma \sim Normal(\mu_\gamma, 0.5)$$

$$\log(\delta_i) \sim Normal(\mu_\delta, 0.5)$$

$$\kappa \sim Normal(0.5, 0.1),$$

$$Pr(\omega_{i,j} | \theta, \mu_\omega, \sigma_\omega) = \sum_{k=1}^{K} \theta_k Normal(\mu_{\omega_k}, \sigma_{\omega_k})$$

$$\zeta \sim Normal(2.25, 0.1)$$

$$\tau_d \sim Beta(a_{\tau_d}, b_{\tau_d}) : E(\tau_d) = 0.45, \sigma_{\tau_d} = 0.002$$

$$\tau_r \sim Beta(a_{\tau_r}, b_{\tau_r}) : E(\tau_d) = 0.35, \sigma_{\tau_d} = 0.002$$

$$\eta \sim Beta(a_\eta, b_\eta) : E(\eta) = 0.45, \sigma_\eta = 0.002$$

Not fit:

$$\theta_k = 0.7$$
$$\mu_{\omega_k} = [-3, 0]$$
$$\sigma_{\omega_k} = [1, 1]$$
$$\mu_\gamma = 4$$
$$\mu_\delta = -3$$

$$A_{v,i,t} = \{\Phi^{g,k}_{v,i,t}\}$$

$$\Phi^{g,k}_{v,i,t} = 1 - \left[(1 - (1 - e^{-\gamma_{v,i,t}})) \prod_j 1 - (1 - e^{-\phi_{v,j,t}}) \right]$$

$$\phi_{v,j,t} = (\lambda^*_{v,j,t-1} c_{j,t-1})^\eta \xi_{j,i}$$

$$\lambda^*_{v,i,t-1} = \frac{\lambda_{v,i,t-1} \alpha_{v,i,t-1}}{\sum_v \lambda_{v,i,t-1} \alpha_{v,j,t-1}}$$

where:

$$1 - e^{-\phi_{v,j,t}} = Pr(z_{v,i,t} = 1 | z_{v,j,t-1} = g)$$

$$\alpha_{v,j,t-1} = Pr(z_{v,j,t-1} = 1 | z_{v,j,t-2} = g)$$

$$\log \xi_{i,j} = \begin{cases} \log \delta_i & \text{if } i = j \\ \log\left(\kappa \frac{pop_j^{\tau_d} pop_i^{\tau_r}}{d_{i,j}^{\zeta}}\right) + \omega_{i,j} & \text{if } i \neq j \end{cases}$$

Observation process

The observed process, $\boldsymbol{\psi}_o$, is $\boldsymbol{Y}_{v,i} = \left[Y_{v,i,t}, t = 1, \dots, T\right]$, which is the observation of sequenced samples of strain v in country i and year t, and is associated with the hidden process $\boldsymbol{\Phi} = (\Phi_{v,i,t}, t = 1, \dots, T)$ of the underlying true presence of strain v in country i and year t as outlined above.

To get at prevalence, we model the probability of our observed data ($\boldsymbol{y}_t$) given the unobserved (hidden) states of presence ($z_{v,i,t} = 1$) or absence ($z_{v,i,t} = 0$) of strain v in country i at time t, where

$$\rho_{v,i,t} = Pr(z_{v,i,t} = 1),$$

$$\alpha_{v,i,t} = Pr(z_{v,i,t} = 1 | z_{v,i,t-1} = k)$$

$$Pr(y_{v,i,t} | z_{v,i,t}) = \frac{Pr(z_{v,i,t} | y_{v,i,t}) | Pr(y_{v,i,t})}{Pr(z_{v,i,t})}$$

$$= \begin{cases} 1 & \text{if } y_{v,i,t} = 0, z_{v,i,t} = 0 \\ 0 & \text{if } y_{v,i,t} \geq 0, z_{v,i,t} = 0 \\ Pr(y_{v,i,t} | \lambda_{v,i,t}) & \text{if } z_{v,i,t} = 1 \end{cases}$$

The likelihood for cases (observed cases: $c^*_{i,t}$) accounts for ≥ 1 lineage present in country i and year t and under-reporting of cases, ϵ_i :

$$Pr(c_{i,t} | z_{v,i,t}) = \begin{cases} (1 - \alpha^+_{i,t}) + \alpha^+_{i,t} Pr(c_{i,t} = 0 | \frac{1}{\epsilon_i} + N_{i,t}) & \text{if } c_{i,t} = 0 \\ \alpha^+_{i,t} Pr(c_{i,t} | \frac{c_{i,t}}{\epsilon_i}) & \text{if } c_{i,t} > 0 \end{cases}$$

where:

$$\epsilon_i \sim Beta(a_\epsilon, b_\epsilon) : E(\epsilon) = 0.8, \sigma_\epsilon = 0.1$$

$$\alpha^*_{i,t} = 1 - \prod_v 1 - \alpha_{v,i,t}$$

We can use the poisson approximation of the multinomial in the sequence observation process:

$$N_{i,t} = \sum_v y_{v,i,t},$$

$$N_{i,t} \sim \text{Poisson}(\Lambda_{i,t}),$$

$$\Lambda_{i,t} = \sum_v \lambda_{v,i,t},$$

$$Y_{v,i,t} \sim \text{Poisson}(\Lambda_{i,t} \lambda^*_{v,i,t})$$

where:

$$log(\lambda_{v,i,t}) \sim \text{Normal}\left(log\left[\frac{c_{i,t}}{\sum_{q \neq v} z_{q,i,t} + 1} \right], \sigma_\lambda \right)$$

Forward equation:

$$\Phi^{g,k}_{v,i,t} = Pr(z_{v,i,t} = k | z_{v,i,t-1} = g)$$

$$Pr(z_{v,i,t} = k | \boldsymbol{y}_{1:t-1}) = \sum_{g \in \{0,1\}} \Phi^{g,k}_{v,i,t} Pr(z_{v,i,t-1} = g | \boldsymbol{y}_{1:t-1})$$

$$\psi_{v,i,t} = Pr(y_{v,i,t} | z_{v,i,t})$$

$$\alpha^{k}_{v,i,t} = \alpha^{g}_{v,i,t-1} \Phi^{g,k}_{v,i,t} \psi_{v,i,t}$$

$$\alpha_{v,i,t} = Pr(z_{v,i,t} = 1 | z_{v,i,t-1} = g)$$

where $\alpha_{v,i,t}$ is the forward probability of presence and, as above, $\Phi_{v,i,t}$ is the transition probability, which is the probability of introduction/re-introduction (establishment) into country i at time t and $\psi_{v,i,t}$ is the observation process, as outlined below.

Backward algorithm & forward-backward algorithm:

$$\beta_{v,i,t}(k) = Pr(y_{v,i,t+1:T} | z_{v,i,t} = k)$$

$$\beta_{v,i,t-1}(g) = Pr(y_{v,i,t:T} | z_{v,i,t} = g)$$

$$= \sum_{k=1}^{K} Pr(y_{v,i,t+1:T} | z_{v,i,t} = l) Pr(\boldsymbol{y}_t | z_{v,i,t} = k)(Pr(z_{v,i,t} = k | z_{v,i,t-1} = g)$$

$$= \beta_{v,i,t} \psi_{v,i,t} \Phi^{g,k}_{v,i,t}$$

$$\rho^{k}_{v,i,t} = Pr(z_{v,i,t} = k | \boldsymbol{y}_{1:T})$$

$$= \frac{\alpha_{v,i,t}(k) \beta_{v,i,t}(k)}{Pr(\boldsymbol{y}_{1:T})}$$

$$\approx \alpha_{v,i,t}(k) \beta_{v,i,t}(k)$$