Phylogeography of cholera; Seventh pandemic origin and spread.

Marco Salemi[1,2], Andrew Tatem[1,3], Rebecca Gray[1,2], Yuansha Chen[1,2], and Judith A. Johnson[1,2]

[1] Emerging Pathogens Institute; [2]Department of Pathology, Immunology, and Laboratory Medicine, College of Medicine; [3] Department of Geography, College of Liberal Arts and Sciences, University of Florida, Gainesville, FL
*Correspondence: Judith Johnson, Emerging Pathogens Institute (EPI), PO Box 10009, University of Florida at Gainesville, Gainesville, Florida, 32610. Phone: (352) 273-9428. Fax: (352) 273-9430 E-mail: jajohnson@pathology.ufl.edu

Introduction

Cholera, a life-threatening diarrheal disease, is endemic in Asia where it is characterized by yearly seasonal epidemics and intermittent pandemic expansions. Seven cholera pandemics have occurred since 1817, spreading rapidly from a focus in Asia to most of the known world. The 6th pandemic, beginning in the 1880's and lasting until 1923, was caused by *V. cholerae* O group 1 with the classical biotype'(Politzer 1959). In 1961, the 7th pandemic, due to *V. cholerae* O1 biotype El Tor, erupted in Indonesia in 1961 and spread across the globe (Barua 1992). The 7th pandemic first entered Africa in 1970 (Gaffga et al 2007) and in 1991, cholera reached South America for the first time in over 100 years.

In contrast to the diversity seen within *Vibrio cholerae* as a whole, each cholera pandemic appears to be caused by a single, almost clonal strain, with only slight strain differences appearing as the pandemic moves across continents and through time (Wachsmuth et al 1993, Popovic et al 1993). The reasons why each of these strains arise, becomes pandemic and take particular geographical outbreak courses are poorly understood. Furthermore, this homogeneity has made molecular epidemiology of the 7th pandemic difficult. Proposed factors affecting cholera dynamics include genetic changes in the pathogen, climate variation, landuse and human and trade movement by sea and land (Barua 1992). Understanding how and why new pandemic strains evolve and spread is critical for controlling cholera morbidity and mortality, and mitigating the spread and effects of future pandemics.

Materials and Methods

To understand factors driving pandemic spread of cholera we are developing a highly interdisciplinary framework that integrates state-of-the-art phylogenetic, population genetic and geospatial analysis techniques to investigate a *V. cholerae* strain collection spanning the 7th pandemic. We employed a hierarchical typing system that uses sequences from the housekeeping gene, *recA*, to provide the first divisions and variable number tandem repeat (VNTR) analysis to provide fine detail. Phylotypes of 138 isolates from a large strain collection spanning the 7th pandemic of *V. cholerae* including 72 El Tor isolates, were constructed by sequencing *recA* and 5 variable number tandem repeats (VNTR) described previously (Stine et. al. 2000; Ghosh et al., 2008). An initial tree was estimated by the neighbor joining method and statistical

significance was assessed by bootstrapping (1000 replicates) using the MEGA version 4.1 software package. To estimate the genealogy and the evolutionary timescale of *V. cholerae* we assemble a data set with known sampling date and used the Bayesian framework implemented in BEAST software package version 1.5.3 under an uncorrelated log-normal relaxed clock model, the HKY+G model of nucleotide substitution, and two demographic models: constant population size and the non-parametric extended Bayesian skyline plot (eBSP) model. The MCMC analysis was run until convergence with sampling every 10000th generation. The results were visualized in Tracer v1.4, and convergence of the Markov chain was assessed by calculating the effective sampling size (ESS) for each parameter. The maximum clade credibility tree, which is the tree with the largest product of posterior clade probabilities (MCC tree), was selected from the posterior tree distribution after 10% burnin using the program TreeAnnotator v1.5.3. Final trees were manipulated in FigTree v1.3.1 for display.

Results and Discussion

recA sequences separated El Tor and classical biotypes (Figure 1). El Tor- like pre-7th pandemic strains (isolated in 1905 and 1933) fell between these clusters, but other pre-7th pandemic isolates (1958-1960) clustered with 7th pandemic strains. VNTR analysis provided further discrimination within clusters (Figure 2). E-burst identified 1 clonal complex within the El Tor strains that included isolates from seven countries at the start of the pandemic (Figures 3 & 4).

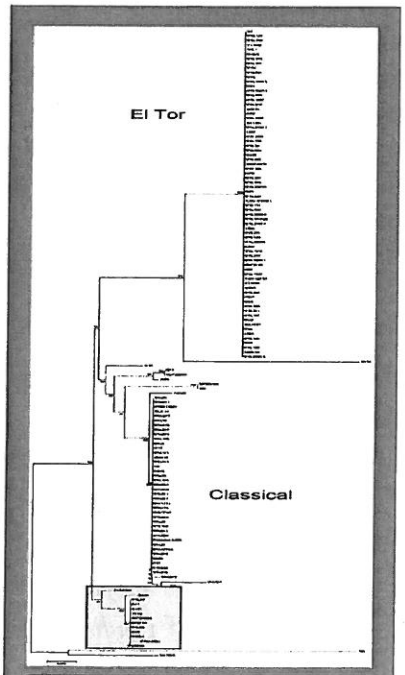
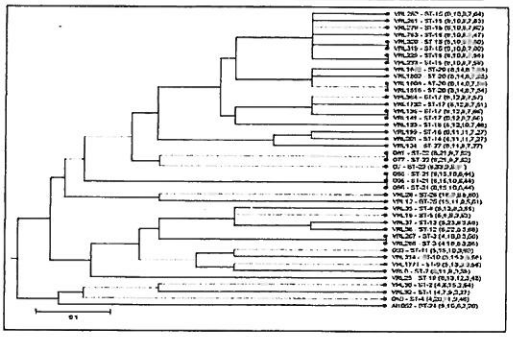


Figure 1. UPGMA tree of toxigenic *V. cholerae* based on *recA* sequences. Blue box contains early pre-pandemic and modern isolates

Figure 2 – Cluster analysis using concatenated *recA* and VNTR sequences



This combination of sequencing methods provides excellent discrimination among toxigenic O1 *V. cholerae* strains and can be used as input for phylogeographic analysis (Lemey et al. 2009). Once a robust phylogeo-graphic database of the seventh pandemic has been constructed it can be integrated with spatial datasets describing incidence, climatic variations, and human travel patterns across the time period under study.

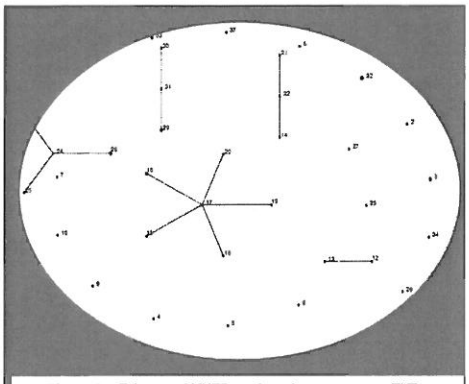


Figure 3 – E-burst of VNTR and recA sequences. El Tor phylotypes found in the 1960s are colored to correspond with Figure 9

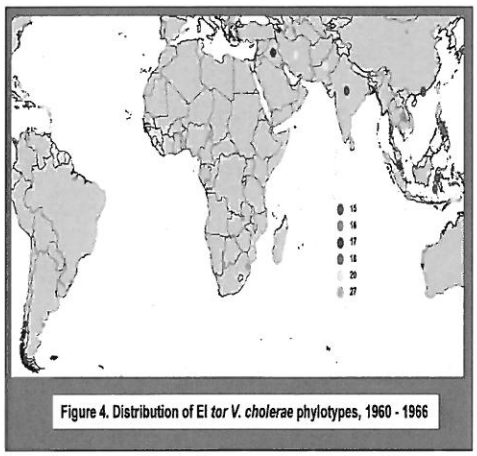


Figure 4. Distribution of El *tor V. cholerae* phylotypes, 1960 - 1966

Determining which of these factors or combination of factors best describes the evolution, dispersal and outbreak patterns seen through the proposed phylogeographic framework will provide a unique understanding of the main causes of the past pandemics, and valuable information to guide future control.

Conclusions

- Split-decomposition analysis in conjunction with high-resolution Bayesian phylogenetic methods can be used to detect and account for recombination
- Divergence between biotypes occurred more recently than previously reported
- The 7th pandemic began as a clonal outbreak
- As the pandemic spread geographically and over time, El Tor strains accumulated mutations that can be mapped to follow the fate of subpopulations in time and space

References

Barua, D. 1992. History of cholera. *In* D. Barua and W.B. Greenough III (ed.), Cholera. Plenum Medical Book Co New York, pp. 1-36

Bruen TC, Herve P, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**:2665-2681.

Feng L, Reeves PR, Lan R, Ren Y, Gao C, Zhou Z, Ren Y, Cheng J, Wang W, Wang J, Qian W, Li D, Wang L. A Recalibrated Molecular Clock and Independent Origins for the Cholera Pandemic Clones. PLoS ONE 2008;3: e4053. oi:10.1371/journal.pone.0004053.

Gaffga NH, Tauxe RV, Mintz ED. Cholera: a new homeland in Africa? Am J Trop Med Hyg. 2007;77:705-13

Ghosh R, Nair GB, Tang L, Morris JG, Sharma NC, Ballal M, Garg P, Ramamurthy T, Stine OC. Epidemiological study of *Vibrio cholerae* using variable number of tandem repeats. FEMS Microbiol Lett. 2008;**288**:196-201.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comp.* Biol. 2009. **5(9)**:e1000520. Epub 2009 Sep 25.

Pollitzer, R. *Cholera.* Geneva: World Health Organization. 1959.

Popovic T, Bopp C, Olsvik O, Wachsmuth K. Epidemiologic application of a standardized ribotype scheme for *Vibrio cholerae* O1. J Clin Microbiol 1993;31:2474-2482.11.

Salemi M, Gray RR, Goodenow MM. An exploratory algorithm to investigate intra-host recombinant viral sequences. *Molecular Phylogenet Evolution 2008;***49**:618-628.

Stine OC, Sozhamannan S, Gou Q, Zheng S, Morris JG, Johnson JA. Phylogeny of *Vibrio cholerae* based on *recA* sequence. *Infect Immun* 2000;**68**:7180-7185.

Wachsmuth IK, Evins GM, Fields PI, Olsvik O, Popovic T, Bopp CA, Wells JG, Carrillo C, Blake PA. The molecular epidemiology of cholera in Latin America. J Infect Dis 1993;167:621-626.