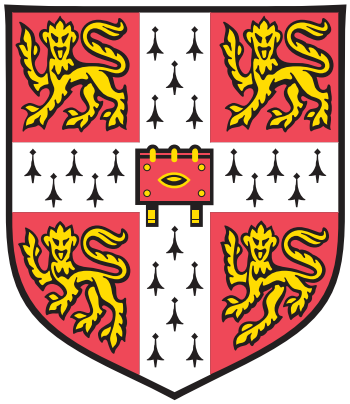
The Origins and Evolution of *Vibrio cholerae* O1 El Tor



Ankur Mutreja

Corpus Christi College

University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

August 2013

Declaration

This dissertation describes my work undertaken at the Wellcome Trust Sanger Institute between May 2010 and August 2013, under the supervision of Profs. Gordon Dougan, Nicholas Thomson and Julian Parkhill in fulfillment of the requirements for the degree of Doctor of Philosophy, at Corpus Christi College, University of Cambridge.

This thesis is the result of my own work and where the work done in collaboration is presented, it is indicated in the text. The work described here has not been submitted for a degree, diploma or any other qualification at any other university or institution. I confirm that this dissertation does not exceed the page limit specified by the Biology Degree Committee.

Ankur Mutreja

Cambridge, August 2013

Publications

Mutreja, A., Kim, D.W., Thomson, N., *et al*. (2011) Evidence for multiple waves of global transmission within the seventh cholera pandemic, *Nature*, **477(7365)**, 462-465.

Mutreja, A. (2012) Bacterial frequent flyers, *Nature Reviews Microbiology,* **10(11)**, 734.

Kiiru, J.N., Mutreja, A., *et al.* (2013) A study on the geophylogeny of clinical and environmental *Vibrio cholerae* in Kenya, *PLoS ONE,* **8(9),** e74829.

Ali, M., Mutreja, A., *et al.* (2014) Genomic Epidemiology of Vibrio cholerae O1 associated with Floods, Pakistan, 2010, *Emerging Infectious Diseases,* **20(1)**, 13-20.

Acknowledgements

I would like to sincerely thank my supervisors, Profs. Gordon Dougan, Nicholas Thomson and Julian Parkhill for allowing me to do this project, and for guiding, advising, and encouraging me throughout my PhD. I am very grateful to my thesis committee advisors Drs. Julian Rayner and Matt Berrimen, and Prof. James Wood for incredibly useful discussions and constructive criticism. My utmost gratitude goes to the Wellcome Trust for funding my research and maintenance costs.

I am very thankful to all our collaborators around the world who contributed to the *V. cholerae* collections that are at the core of this PhD. Jan Holmgren, Michael Lebens, Sam Kariuki, John Clemens, Alejandro Cravioto, Dong Wook Kim, Habib Bukhari, Cecil Czerkinsky and GB Nair played huge roles in making sure that representative samples were collected from the cholera affected sites and shipped to the WTSI .

I am extremely grateful to WTSI sequencing and pipeline teams: Michael Quail, Richard Rance, David Harris, Elizabeth Gibson, Craig and Nicola Corton, Hilary Browne, Graham Rose, Karen Brooks, Christine Burrows, Louise Clark, Vicky Murray, Scott Thurston, Andries van Tonder, and Danielle Walker have all done a tremendous job in generating the sequencing data used for this analysis. I would also like to thank Jacqui Keane and her team for providing the troubleshooting help in solving informatics related problems.

I am very grateful to Maria Fookes, who kindly helped me with the phylogenetic analyses during my rotation on this project and provided creative ideas when I first started. I am thankful to Simon Harris for sharing his expertise with me and helping me analyse results, especially during the early days of my project. Also, I would like to convey my thanks to Tom Connor for sharing his knowledge of Bayesian analysis software.

I also want to thank Sophie Palmer, Theresa Feltwell, Annabel Smith and Christina Hedberg- Delouka for looking after me at all the steps of my PhD and keeping checks that I was progressing well. I would also like to thank the travel office team, Jeanne

Cook and Anne Wombwell for their cooperation in arranging travel to important meetings and very useful conferences during my PhD. I wouldn't be here without the support of my colleagues and friends, Ravi Verma, Gaurav Godara, Deepak Singh Rana, Deepak Agrawal, Abhinav Prasad and Popoola Olalekan among several others whose words and presence kept me positive at all times.

Dedicated to

My Parents and Sister

"Always appreciate what you have and what you are getting from life, otherwise you risk losing it. Be very proud of yourself."

Contents

Figures

Chapter 1 figures

Chapter 2 figures

Chapter 3 figures

Chapter 4 figures

Chapter 5 figures

Tables

Chapter 2 tables

Chapter 3 tables

Chapter 5 tables

Appendix

Abstract

The Origins and Evolution of *Vibrio cholerae* O1 El Tor

Cholera, like plague, is an ancient disease of great historical importance, the spread of which was originally believed to be *via* bad air or 'miasma'. In 1854, a London based physician, John Snow, first provided epidemiological evidence for the connection between contaminated drinking water and cholera and approximately 50 years later *Vibrio cholerae*, the etiological agent of cholera was identified by Robert Koch. Cholera is still common in many regions of the world despite having one of the simplest known treatment regimes: oral rehydration. In fact, the increase in the incidence of cholera since 2007 has highlighted the risk our globalized community still faces.

Currently, scientists lack a detailed understanding of how *V. cholerae* transmits and evolves, although water is recognized as a critical factor. This PhD project exploits whole genome sequence data to investigate the evolution of *V. cholerae*, focusing on serogroup O1. Phylogenetic trees based on whole genome sequencing data obtained from over 1000 *V. cholerae* representative of seventh pandemic El Tor, classical and non-O1/O139 isolates collected from across the world where cholera occurs were used to determine evolutionary patterns and relationships. In cases where detailed phenotypic information or meta-data was available, phylogeny was used alongside clinical, phenotypic and geographical information to track and understand the global and regional spread of cholera.

The genotypic basis underpinning the basis of the Ogawa to Inaba serotype change was investigated using *V. cholerae* sampled during a phase III vaccine trial undertaken in Kolkata, India and these mechanisms were defined.

Based on my mining of the phylogeny and whole genome data on which it is based, informative SNPs were selected for the basis of a simple and mobile SNP genotyping scheme. Multiplex ligation-assisted probe amplification (MLPA) was selected as the most suitable laboratory based molecular technique to detect the canonical SNPs and two kits were designed for the use of scientific and public health communities in developing countries.