

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

The Origins and Evolution of <i>Vibrio cholerae</i> O1 El Tor	i
Declaration.....	ii
Publications	iii
Acknowledgements	iv
Dedicated to	vi
Contents	vii
Figures.....	x
Tables	xii
Abstract.....	xiii
1. Introduction.....	1
1.1. Cholera	1
1.1.1. Overview.....	1
1.1.2. Cholera Pandemics	2
1.2. <i>Vibrio</i> bacteria	5
1.2.1. <i>V. cholerae</i> : the species and classification.....	6
1.2.2. Ecology of <i>V. cholerae</i>	9
1.2.3. Epidemiology.....	11
1.2.4. <i>V. cholerae</i> infection and symptoms.....	13
1.2.5. Diagnosis	14
1.2.6. Treatment and prevention	16
1.2.7. Molecular basis of pathogenesis and cholera virulence factors.....	18
1.2.8. CTX and other <i>V. cholerae</i> toxins	20
1.2.9. <i>Vibrio</i> pathogenicity and seventh pandemic islands	24
1.2.10. Multiple antibiotic resistance cassettes.....	26
1.2.11. Typing schemes for <i>V. cholerae</i>	27
1.3. Whole genome sequencing.....	29
1.3.1. Next generation sequencing.....	30
1.3.1.1. New sequencing technologies.....	30
1.3.1.2. Next-Generation bioinformatics tools	31
1.3.2. Understanding bacterial evolution and transmission using genomics	33
1.4. <i>V. cholerae</i> genomics and genetic diversity.....	34
1.5. Aims and objectives of this study.....	36
2. Genomic variation in global <i>V. cholerae</i> spanning a century	38
2.1. Introduction	38
2.2. Bacterial isolates.....	40
2.3. Results and discussion.....	44
2.3.1. Global phylogeny of the <i>V. cholerae</i> species	44
2.3.2. Evolution of the seventh pandemic O1 El Tor <i>V. cholerae</i>	47
2.3.3. The three waves of seventh pandemic O1 El Tor <i>V. cholerae</i>	50
2.3.4. The origins of O139 serogroup strains	53
2.3.5. Evidence within the global phylogeny of intercontinental transmission.....	54
2.3.6. Patterns of gene acquisition and loss in the seventh pandemic	54
2.3.7. Variations in CTX and their phylogenetic distribution	55

2.3.8. Variations in SXT and its phylogenetic distribution	58
2.3.9. WASA-1 and other markers of the West Africa/South American (WASA) clade.....	63
2.3.10. Recombination	67
2.4. Conclusion and lessons from global phylogeny	68
3. Case studies on the regional evolution of <i>V. cholerae</i> O1 El Tor.....	69
3.1. Introduction.....	69
3.2. Bacterial isolates.....	72
3.2.1. Pakistan <i>V. cholerae</i> collection.....	72
3.2.2. Kenyan <i>V. cholerae</i> collection.....	73
3.2.3. Mexican <i>V. cholerae</i> collection	76
3.3. Results and discussion.....	78
3.3.1. Whole genome phylogeny of 2010 Pakistan flood <i>V. cholerae</i>	78
3.3.2. Evidence for a strict <i>V. cholerae</i> molecular clock in Pakistan	81
3.3.3. Sub-clade signature deletions within the genomes of Pakistan <i>V. cholerae</i>	83
3.3.4. Diversity within <i>V. cholerae</i> circulating in Kenya	84
3.3.5. The phylogeny of Kenyan <i>V. cholerae</i> based on whole genome sequences	86
3.3.6. Genomic features of Kenyan O1 El Tor sub-clades	90
3.3.7. A novel <i>ctxB</i> gene in some Kenyan non-O1 environmental isolates	93
3.3.8. Whole genome phylogeny of Mexican strains	94
3.3.9. Genomic islands and new markers in the Mexican <i>V. cholerae</i> genomes.....	97
3.4. Conclusion and lessons from the regional case studies.....	100
4. The genetic basis of serotype variation in <i>V. cholerae</i> samples during clinical trial in Kolkata, India	103
4.1. Introduction.....	103
4.2. Results and discussion.....	105
4.2.1. <i>wbeT</i> sequence analysis.....	105
4.2.2. Mapping <i>V. cholerae</i> from a vaccine trial performed in Kolkata to the global El Tor phylogeny.....	109
4.3. Lessons learned and questions arising from this study	111
5. Expanded analysis of the seventh pandemic <i>V. cholerae</i> lineage and design of PCR based SNP typing assays	113
5.1. Introduction.....	113
5.2. Results and discussion.....	117
5.2.1. SNPs for genotyping.....	117
5.2.1.1. Selection of canonical SNPs	117
5.2.1.2. Phylogenetic analysis on selected SNPs.....	124
5.2.2. Phylogeny expansion and MLPS kits	125
5.2.2.1. Global dissemination of wave-3 in 3 sub-waves	125
5.2.2.2. Design of the MLPA based SNP-genotyping assays.....	130
5.3. Lessons learned from the expanded phylogeny and importance of SNP genotyping.....	131
6. Conclusion and future directions	134
6.1. Conclusion.....	134
6.2. Future directions	135
6.2.1. Further expansion of the sequenced <i>V. cholerae</i> collection	135
6.2.2. Studies investigating the evolution of <i>V. cholerae</i> within cities, countries and continents	136

6.2.3. A combined transcriptomics and proteomics study of intestinal tissues taken from mice at different stages of <i>V. cholerae</i> infection	137
6.2.4. A study designed to investigate household and community level spread of <i>V. cholerae</i>	138
Methods.....	139
References.....	142
Appendix.....	154

Figures

Chapter 1 figures

1.1 Timeline showing pandemics.....	3
1.2 One of the first maps showing the spread of cholera	4
1.3 <i>Vibrio cholerae</i> bacterium.....	7
1.4 O-antigen serogroups and their properties	8
1.5 <i>V. cholerae</i> life cycle.....	10
1.6 <i>V. cholerae</i> circulation between host and environment	12
1.7 Cholera symptoms.....	14
1.8 <i>V. cholerae</i> colonies on selective media	15
1.9 Molecular mechanism of working of cholera toxin	19
1.10 Expression of virulence factors at different times.....	20
1.11 Genetic structure of CTX phage	22
1.12 Genetic differences between different CTX phages	23
1.13 Toxin co-regulated phage gene cluster	24
1.14 Genetic structure of Vibrio pandemicity island - 2	25
1.15 Genetic structure of SXT	27
1.16 Possible arrangements of genes within CTX	28
1.17 Two chromosomes of <i>V. cholerae</i> genome.....	35

Chapter 2 figures

2.1 Global <i>V. cholerae</i> phylogeny.....	45
2.2 Phylogenetic comparison of non-conventional O1 strains	46
2.3 Phylogeny of the seventh pandemic <i>V. cholerae</i> O1 El Tor lineage	48
2.4 Linear regression plot for the seventh pandemic and its waves	49
2.5 Bayesian based phylogenetic tree for the seventh pandemic	51
2.6 Spread of the seventh pandemic plotted on the world map	53
2.7 SXT variation plot.....	62
2.8 Comparison of the seventh pandemic and SXT tree	63
2.9 Gene flux within the seventh pandemic	64
2.10 Insertion site of the West African South American island -1 (WASA-1).....	65
2.11 Recombination in the WASA-1 cluster.....	67

Chapter 3 figures

3.1 Pakistan strains in the seventh pandemic phylogeny	80
3.2 Spread of cholera in Pakistan.....	81
3.3 Scatter plot of root-to-tip distance vs. date of isolation	83
3.4 Scatter plot of root-to-tip distance vs. river source in Pakistan.....	83
3.5 Sites from where Kenyan samples were collected.....	85
3.6 Kenyan strains in the global phylogeny	87
3.7 Phylogeny of the Kenyan clade and its sub-clades	88
3.8 Novel <i>ctxB</i> of the Kenyan non-O1 strain	93
3.9 Mexican strains in the global phylogeny	95
3.10 Mexican strains in the seventh pandemic phylogeny.....	96
3.11 Alignment of the novel <i>ctxB</i> in Mexican strains.....	98
3.12 Phylogeny of the Mexican local endemic-2 lineage	100

Chapter 4 figures

4.1 Classification of <i>V. cholerae</i> into biotypes and serotypes	104
4.2 Distribution of different mutation types in the <i>wbeT</i> gene	106
4.3 Phylogenetic tree of <i>wbeT</i> from seventh pandemic strains.....	108

4.4 Percentage match and mismatch between phenotypic and genotypic results	108
4.5 Distribution of Inaba and Ogawa during a linical trial study in Kolkata, India.....	110
4.6 Kolkata clinical trial strains and correlation between with their temporal and serotypic correlation	111

Chapter 5 figures

5.1 Molecular basis of working of multiplex ligation-dependent probe amplification (MLPA) technology	116
5.2 Seventh pandemic phylogeny and selection of canonical SNPs.....	118
5.3 Phylogeny based on the selected SNPs.....	125
5.4 Phylogeny of the expanded seventh pandemic lineage.....	127
5.5 Detailed structure of the wave-3 and its 3 sub-waves.....	129
5.6 Spread of strains from the expanded collection of the seventh pandemic	132

Tables

Chapter 2 tables

2.1 Global <i>V. cholerae</i> collection.....	44
2.2 CTX types in the seventh pandemic collection.....	58
2.3 Strains carrying SXT and various antibiotic resistance genes	61
2.4 Genomic islands in the seventh pandemic	66

Chapter 3 tables

3.1 Pakistan <i>V. cholerae</i> collection.....	73
3.2 Kenyan <i>V. cholerae</i> collection.....	76
3.3 Mexican <i>V. cholerae</i> collection.....	78
3.4 Genomic islands found in Kenyan strains.....	92

Chapter 5 tables

5.1 Selected canonical SNPs.....	123
5.2 Questions that can be answered with the selected SNPs.....	124
5.3 The design of MLPA kits.....	131

Appendix

A.1 Expanded seventh pandemic <i>V. cholerae</i> collection.....	154
--	-----

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.