Methods

Genome sequencing

Genomic DNA for all the *V. cholerae* analysed in this study were extracted by our collaborators and shipped to the Wellcome Trust Sanger Institute (WTSI) for whole genome sequencing. Multiplex sequencing libraries of 250 bp insertion size were created for each sample using the manufacturer's protocol by the sequencing team at WTSI. The libraries were loaded on to the Illumina's GA II or HiSeq platform cell to perform 54-72-base paired-end sequencing of 12-96 separate libraries in each lane. Each library had a unique index tag and after sequencing this tag sequence information was used for assigning reads to the individual samples assisting the downstream separation of the data for each sample. All the samples achieved an average coverage of 50-200x in the regions where SNPs were called. All the data has been submitted to European Nucleotide Archive and the accession codes are listed in the strain tables provided in the chapters.

Whole genome alignment and detection of SNPs

The paired-end read data obtained was mapped to the O1 El Tor reference N16961 (for chromosome 1 and 2, the NCBI accession numbers are AE003852 and AE003853 respectively) using SMALT (http://www.sanger.ac.uk/resources/software/smalt) to obtain a whole genome alignment for all the strains in this study. For SNP calling, the default settings of nucmer program in the MUMmer package (Kurtz, *et al.*, 2004) were used. No SNPs were called from the reads that either did not map to N16961 or from the regions that were absent from the N16961 reference genome. Strict filtering of the SNPs was performed and any SNP with a quality score less than 30 was excluded. Also, a SNP was considered true only if it was present in at least 75% of the reads at any heterogeneously mapped ambiguous sites. High-density SNP clusters and the possible recombination sites were excluded using the methodology of Croucher *et al.* (Croucher, *et al.*, 2011).

Phylogenetic Analysis

Default settings of RAxML v0.7.4 (Stamatakis, 2006) were used to estimate the phylogenetic trees based on all the SNPs recorded against the reference genome as explained above. The number of SNPs on each branch were calculated by reconstructing all the polymorphic events on the tree using PAML (Yang, 2007). M66 (accession numbers CP001233 and CP001234), a pre-seventh pandemic strain and a well-known out-group for the seventh pandemic strains, was used to root the final seventh pandemic phylogenetic tree (Mutreja, *et al.*, 2011) while other trees were left un-rooted or were midpoint rooted. For visualization and ordering of the nodes, phylogenetic tree reading software Figtree (http://tree.bio.ed.ac.uk/software/figtree/) was used.

Comparative Genomics

A multi-contig draft genome was generated for each sample by assembling the paired end reads using a *de-novo* genome assembly program Velvet v0.7.03 (Zerbino and Birney, 2008). The parameters were set to give the best kmer size and at least 20x kmer coverage. Contigs were ordered using Abacas as per the reference N16961 El Tor complete genome sequence (Assefa, *et al.*, 2009; Heidelberg, *et al.*, 2000). Annotation was transferred from the reference sequence to each ordered draft assembly. Artemis Comparison Tool was used for manual comparison of the assembled genomes (Carver, *et al.*, 2008).

Linear Regression Analysis

The final phylogenetic tree was opened using Path-O-Gen v1.3 (http://tree.bio.ed.ac.uk/software/pathogen) and the root-to-tip distance data for each strain was exported to excel. This data was used to plot a linear regression curve against the year of isolation of the strain. The R-squared correlation, slope and p-values were determined using the inbuilt regression package of R-statistical environment.

Bayesian Analysis

The waves of the seventh pandemic were confirmed using BAPS (Corander, *et al.*, 2008; Corander, *et al.*, 2003). The BAPS analysis was performed on the final SNP alignment obtained after removing the recombination, which contained the unique SNP patterns from the seventh pandemic isolates. The program was run using BAPS individual mixture model and three iterations were performed independently to obtain the most optimal partitioning of the sample.

The tree was reconstructed and the ancestral or nodal dates for the strains were inferred using the Bayesian Markov Chain Monte Carlo framework (Drummond and Rambaut, 2007). The final SNP alignment without recombinant sites was used as the input dataset for BEAST in seventh pandemic each dataset (Drummond and Rambaut, 2007) and the rates of evolution on the branches of the tree were estimated using a relaxed molecular clock (Drummond, *et al.*, 2006), providing the flexibility for the rates of evolution to change amongst the branches of the tree. A coalescent constant population size and a GTR model with gamma correction were used. The results produced from three independent chains of 100 million steps each were sampled every 10,000 steps to maintain homogeneity. The first 10 million steps of each chain were binned. The results of the three chains were combined using Log Combiner, and the maximum clade credibility tree was generated using Tree Annotator software in the BEAST package (http://tree.bio.ed.ac.uk/software/beast/). ESS cut off value of 200 was used for each parameter and convergence was visually confirmed using Tracer 1.5 (http://tree.bio.ed.ac.uk/software/tracer).