Supplementary Information for

Toxigenic Vibrio cholerae Evolution and Establishment of Reservoirs in Aquatic Ecosystems

Corresponding author: Marco Salemi
Email: salemi@pathology.ufl.edu

This PDF file includes:

- Supplementary text
- Figures S1 to S10
- Tables S1 to S7
- Legends for Datasets S1, S2 and S3
- SI References

Other supplementary materials for this manuscript include the following:

- Datasets S1, S2 and S3 as excel file

www.pnas.org/cgi/doi/10.1073/pnas.1918763117

Supplemental Text

We consider the following mathematical model for the eco-evolutionary dynamics of Cholera:

$$
\begin{aligned}
\frac{dS(t)}{dt} &= \mu - \mu S(t) - S(t)\sum_{j=1}^{2^L} \kappa_j\left(\beta_I I_j(t) + \beta_W(t)W_j(t)\right) + 3\alpha R_3(t),\\
\frac{dI_j(t)}{dt} &= S(t)\sum_{j=1}^{2^L} \kappa_j\left(\beta_I I_j + \beta_W(t)W_j(t)\right) - \gamma I_j(t) - \mu I_j(t), \quad j = 1, \ldots, 2^L\\
\frac{dW_j(t)}{dt} &= \xi(t)\sum_{j=1}^{2^L} I_j(t) + rW_j(t) \quad 1 - \frac{\sum_{k=1}^{2^L} W_k}{K}\Bigg) - \rho_j\nu(t)W_j, \quad j = 1, \ldots, 2^L\\
\frac{dR_1}{dt} &= (1-\sigma)\gamma\sum_{j=1}^{2^L} I_j - 3\alpha R_1 - \mu R_1\\
\frac{dR_2}{dt} &= 3\alpha R_1 - 3\alpha R_2 - \mu R_2\\
\frac{dR_3}{dt} &= 3\alpha R_2 - 3\alpha R_3 - \mu R_3
\end{aligned}
\tag{1}
$$

The variables and parameters are summarized in Table S8. Our underly-ing epidemiological model is based on the "SIWR" framework, as utilized in prior models of Cholera [1, 2]. Briefly, susceptible individuals (S) can become infected with pathogen variant j (I_j) through direct host or environmental transmission routes. Infected individuals also shed pathogen into the environment (W_j) before recovery or death. Upon recovery, individuals have a gamma-distributed period of temporary immunity where R_i, $i = 1,\ 2,\ 3,$ denotes sequential stages of waning immunity. The pathogen variant j can reproduce in the environment subject to logistic growth and competition, and variant specific decay. Utiliz-ing the next-generation method as in [3], we can calculate a (variant-specific) reproduction number as

$$
\mathcal{R}_0 = \frac{1}{2}\left(\mathcal{R}_h + \mathcal{R}_w + \sqrt{(\mathcal{R}_h - \mathcal{R}_w)^2 + 4\mathcal{R}_{wh}}\right), \tag{2}
$$

where $\mathcal{R}_h = \frac{\kappa_j\beta_I}{\gamma+\mu}$, $\mathcal{R}_{wh} = \frac{\kappa_j\xi\beta_W}{\rho_j\nu(\gamma+\mu)}$, $\mathcal{R}_w = \frac{r}{\rho_j\nu}$ are the host-host, environment-host, and environment reproduction numbers.

We remark that system (1) has some similarities, but also significant differences with a previous model fit to the outbreak in Haiti [2] since our goal is to provide insight on potential evolution of Cholera based on qualitative behavior of *N_e*, rather than purely fitting the case data. In particular, we simplify the model by not explicitly including asymptomatic infection, transmission stochasticity, and the effects of temperature, in order to make room for the high-dimensional complexity of an evolutionary component described below. Additionally, we reduce transmission rates at two time-points representative of enhanced control measures and/or population behavior change in order to better match case data. In comparison, [2] utilized a time dependent environmental death rate due to phage lysis to mimic the case data. There is documentation of a national control plan being implemented beginning in 2012 and intensifying in 2013 [4, 5], which motivates our timepoints where transmission is reduced. We tested models without this second time-point of parameter change, and while we can achieve a good fit of cases under reduced or no environmental replication, we cannot recapitulate the observed dynamics of *Ne* (SI Appendix, Figure S6).

The major novelty of the model is the inclusion of evolution through multiple variants of the pathogen competing for hosts and within the environmental reservoir niche. The variants are distinguished by a “binary sequence” of length $L = n + m + k$ consisting of $n(m)$ distinct loci with alleles adapted to the aquatic reservoir (host), and k neutral loci not under selection pressure. We assume a tradeoff between these traits, which has been described in other studies [6, 7]. We denote the sequence of strain j a s $\mathbf{j} = (j_1,\ j_2, \ldots, j_L) \in \{0,\ 1\}^L$, coding the allele type at each of the L loci. When multiple mutations accumulate, we assume some epistasis in the fitness of phenotypes so that simulations can display a more reasonable slow decay of the aquatic reservoir and drop of *Ne* to zero when host trans-mission is blocked, as shown if Table S8, consistent with the hypothesis that historically in Africa outbreaks die out in a 5-10 year period. We also considered the case of independent multiplicative fitnesses. However, then the combination of multiple environmentally beneficial mutations confers a large advantage and the pathogen can persist in an aquatic reservoir even when host transmission is set to zero. Thus we conclude that assuming epistasis allows for more reasonable simulations. For each strain j, t he host transmissibility fitness factor and environmental survivability factor are

$$\kappa_j = \prod_{\ell=1}^{m+n} (j_\ell h_\ell + 1 - j_\ell)^\eta, \quad \rho_j = \prod_{\ell=1}^{m+n} (j_\ell w_\ell + 1 - j_\ell)^\eta \tag{3}$$

Here the host and environment factors at each non-neutral loci, $\ell = 1, \ldots, n + m$, a re denoted h_ℓ a nd w_ℓ, r espectively, and the epistasis is modeled through a power function of the product of mutant allele fitness factors with exponent η.

The parameters (SI Appendix, Table S8) used for the model are representative of the cholera outbreak in Haiti. However since this is a preliminary modeling effort attempting to illustrate a potential mechanism for the analyzed data, we do not perform thorough parameterization of the system and leave this for future work. Furthermore, we simplify some aspects of cholera epidemiology, for example we incorporate the influence of asymp-tomatic cases by counting the clinical (symptomatic) cases as a fraction, $q = 1/4$, of total cases. Note that while some parameter values are taken from literature, other parame-ters are calibrated to provide good qualitative fit to the data. The seasonal influence is incorporated through interpolation of monthly rainfall data, along with the assumption that precipitation increases pathogen transmission and from and to aquatic reservoir, and also increases survival in the environment. An increasing relation of environmental transmission, shedding and survival with respect to precipitation reflect the observed positive correlation between cases and rainfall, and are consistent with previous works [2]. Explicitly, $\beta_W(t) = \beta_{W,0}(1+a_1(P(t)-P))$, $\xi(t) = \xi_0(1+a_2(P(t)-P))$, $\nu(t) = \nu_0(1+a_3(P-P(t)))$, where $P(t)$ is interpolated precipitation data, P is mean monthly rainfall over study period, and a_1, a_2, a_3 are amplitudes. The number of loci, $L = n+m+k$, is chosen relatively small, along with larger mutation rate to account for less loci, increasing computation speed. Due to the small number of loci and relatively high mutation rate, the relative diversity in allele frequencies is magnified leading to larger values in *N*$_e$. We may think of the loci as a cluster of sites in the actual genome. Since we are interested in the qualitative change and relative magnitude in *N*$_e$, we multiply by a conversion fraction to account for the larger number of loci which will lower relative diversity in the actual data.

The model was coded in MATLAB, where the built-in ODE solver ODE45 was utilized for simulations. Similar to the methods in [8], we simulate mutations of loci by drawing from a binomial distribution. With a mutation rate of $\epsilon = 1.67 \times 10^{-4}$ per loci per day, we compute the number of pathogen variant mutations at fixed time steps, taken as $\Delta t = 1 \ day$, as follows. In the host setting, we consider mutations as transitions from infected individuals of one strain to another strain $(I_j \to I_k)$, similar to underlying assumptions in prior work connecting epidemic models to phlyodynamics [9]. Altering the assumption so that mutations occur upon transmission of pathogen to a new host individual did not affect qualitative results, and more detailed incorporation of mutations would require the complexity of a within-host model not in the scope of this study. Mutations in the environment occur in a subset of cell replication events (given by logistic growth term) at the fixed time steps. To improve computation speed, we assume that only one of the L loci mutates per infected case and environmental replication, i.e. the small probability of simultaneous mutations are neglected. Then for each pathogen variant $j = 1, \ldots, m$ and locus $\ell = 1, \ldots, L$, the number of mutations in time interval $(t + \Delta t)$,is given approximately by Bin$(I_j(t + \Delta t), \epsilon)$. Analogous calculations for mutations during cell replication in the environment are computed. The pathogen populations are updated accordingly, and the ODE solver is run for Δt time units and then the process repeats.

We consider the following measure of genetic diversity. The probability π_ℓ that two randomly sampled viruses differ in their allele at a particular locus, ℓ, is given by $\pi_\ell = 2p_\ell(1-p_\ell)$ where p_ℓ is the frequency of the "0 allele" in the population at locus ℓ. According to coalescent theory [10], averaging over a large number L of loci, $\pi = \frac{1}{L}\sum_{\ell=1}^{L} \pi_\ell$, gives the relationship $\pi = 2Ne_s\epsilon$, where ϵ is the mutation rate probability for each loci and *Ne*$_s$ is this effective population size measured according to the diversity at each loci in our model, many of which are under selection (hence the "s" in the subscript). Due to the small number of neutral loci and lack of stochasticity in our model, this formulation of *N e* does not sufficiently account for the contribution of genetic drift whose magnitude correlates with population size. Therefore we utilize a second measure of *Ne* based on [11], which largely reflects the population size, or more exactly, the force of infection. In a simple SIR model, this amounts to the relationship $\frac{1}{Ne_n}= \frac{\beta(t)S(t)}{\epsilon I(t)}$ [9], here the subscript "n" referring to "neutral". However in structured epidemic models the computation is more complex and by [11] can be formulated as:

$$\frac{1}{Ne_n} = \frac{2}{\epsilon}\left(S(t)\sum_{j=1}^{2^L}\kappa_j\left(\frac{p_h^2(t)\beta_I I_j(t)}{(I(t))^2} + \frac{p_h(t)p_w(t)\beta_W(t)W_j(t)}{I(t)W(t)}\right) + p_w^2(t)\frac{r(1-\frac{W(t)}{K})}{W(t)}\right),$$

where $p_h(t)$ is the probability that a lineage is inside a host at time t and $p_w(t) = 1 - p_h(t)$ is the probability that it is in the aquatic reservoir. In Volz [11], a rather complex master equation is derived for these state probabilities. Here, we take a simplified approach and just approximate $p_h(t)$ as the proportion of *total* transmissions at time t derived from the host state. In particular the total transmissions is $F(t) = F_h(t) + F_w(t)$, where

$$F_h(t) = S(t)\sum_{j=1}^{2^L}\kappa_j\beta_I I_j(t) + \xi(t)I(t), \quad F_w(t) = S(t)\sum_{j=1}^{2^L}\kappa_j\beta_W W_j(t) + r(1-\frac{W(t)}{K}).$$

Thus $p_h(t) = \frac{F_h(t)}{F(t)}$. We then formulate *Ne* as a weighted average of *Ne_s* and *Ne_n* in order to best fit the observed *Ne*. We remark that the principles behind *Ne_n* derived in [11] may break down for populations not sufficiently large as bottlenecks may reinforce the influence of selection on *Ne*. Furthermore the observed *Ne* does not resemble the case data pattern after initial outbreak. Hence, we utilize the weighted average $f_nNe_n + (1 - f_n)Ne_s$ to get the best fitting Ne under selection and bottlenecks.

An example simulation with environmental replication calibrated to both case and Ne data is presented in Fig. 4, Figure S6 and S7. For this representative simulation with environmental replication, we utilize the following control timepoints (as explained before): β_I, $\beta_{W,0}$ are reduced by factor of 0.485 at $t = 300\ days$, and additional factor of 0.58 at $t = 600\ days$ after outbreak start. In order to test envi-ronmental replication is necessary for observed *Ne*, we test model in the absence of envi-ronmental replication. Although we can mimic case data by increasing pathogen shedding rate, ξ_0, and reduced efficacy of control timepoints, removing environmental replication (by setting $r = 0$) does not allow for a good fit to observed *Ne* as shown in four example simulations (SI Appendix, Figure S6). Notice that less reduction in control efficacy (enhanced control) increases bottleneck and magnitude of drop in *Ne* during lull period, but the fall in *Ne* does not match observed Skyride plot of *Ne*. For each simulation, we assume some initial diversity and adaptations for host transmission consistent with what is expected during first rise in cases of outbreak. Observe in Figure S7 the multi-strain dynamics within hosts and aquatic reservoir.

To evaluate the effects of vaccination, we include a vaccination rate $\delta S(t)$ which transfers susceptible individuals to the R_2 recovered compartment (since vaccination induces less immunity than natural infection) (Fig. 4b, SI Appendix, Table S8). The vaccination is assumed to start around the beginning of 2015. We had data for rainfall in Oeust during 2010-2015, so in order to extend the seasonal (rainfall-dependent) environmental terms in the model in years after 2015, we utilized the average rainfall for each month over 2010-2015. Furthermore, the initial state of the model populations are assumed to be at their values for the final time (2015.9) in the example simulation with environmental replication (Fig. 4; SI Appendix, Figure S6a). In Fig. 4b and Table S8, we vary the vaccination rate δ and environmental decay rate ν_0 in order to assess the effect of control scenarios. Note that vaccinating at a rate of $\delta = .04$ per day translates into vaccination coverage of 88% (percent reduction in susceptible individuals), whereas $\delta = .01$ per day translates into vaccination coverage of 64%. Contrary to Fig. 4b, for any of the four example simulations without environmental replication, vaccination at a rate of $\delta = .01$ readily eliminates the pathogen within a year (SI Appendix, Table S8).

Supplementary Figures

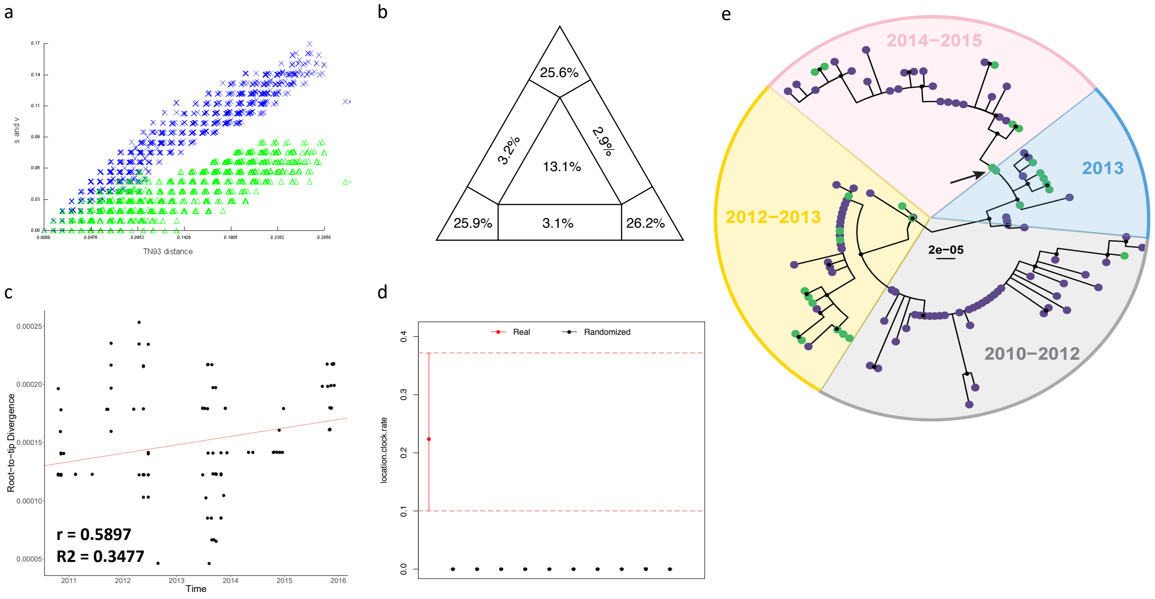

Figure S1. Estimation of phylogenetic signal and substitution saturation among all 116 environmental and clinical toxigenic *V. cholerae* O1 isolates collected in the department of Ouest, Haiti. (A) Substitution saturation was gauged by plotting pairwise nucleotide transition (s) and transversion (v) substitutions *vs.* genetic distances estimated by TN93 nucleotide substitution model for the data set. (B) Presence of phylogenetic signal was evaluated by likelihood mapping checking for alternative topologies (tips), unresolved quartets (center) and partly resolved quartets (edges) for the data set. (C) Linear regression of root-to-tip genetic distance within the ML phylogeny against sampling time for each taxa. Temporal resolution was assessed using the slope of the regression, with positive slope indicating sufficient temporal signal. Correlation coefficient "r" are reported for the data set. (D) The estimates obtained by randomizing the sampling times with 10 randomizations (black) are shown versus the estimates obtained with the correct sampling times (red), with red dashed horizontal lines showing the standard deviation on either side of the mean for the correct sampling times. (E) Maximum likelihood phylogeny inferred from the alignment including 116 environmental and clinical *V. cholerae* isolates from Haiti, 2010-2015. Each circle represents an isolate sequence colored by environmental (green) or clinical (violet) origin. Early clade comprising isolates obtained between 2010-2012 is highlighted in grey,

early middle with isolates obtained in 2012-2013 in yellow, a subsequent clade with isolates obtained in 2013, and the late clade with isolates obtained between 2014-2015 in pink.

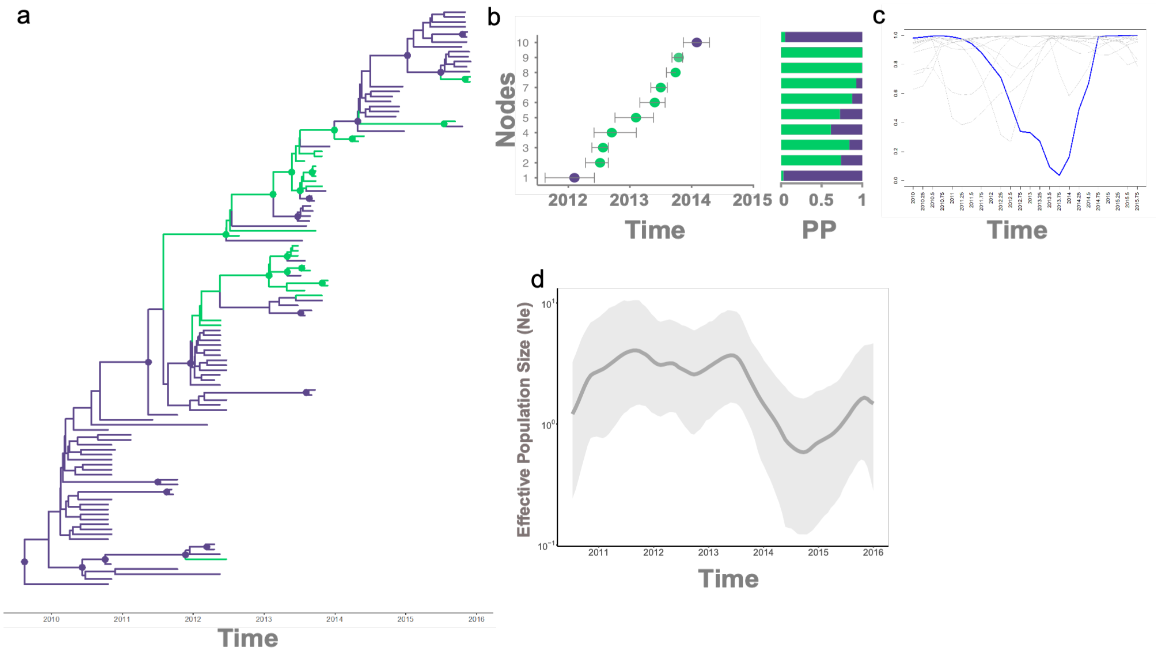

Figure S2: Contribution of toxigenic *V. cholerae* O1 environmental isolates to the evolution of the cholera epidemic in Ouest, Haiti. (A) The BASTA MCC phylogeny was obtained using the Bayesian phylogeography framework. Branch lengths are scaled in time by enforcing a strict molecular clock. Ancestral state reconstruction at each node and branch is indicated for environmental (green) clinical (violet) isolates. Circles indicated high posterior probability (PP) support on nodes (PP>0.9) for internal branches and for ancestral location at node (colored green or violet for environmental or clinical location). (B) Ancestral state reconstruction and PP at each environmental node (green) contributing to the trunk proportion of the tree, as compared to neighbors clinical nodes (violet). (C) Trunk rewards proportion (TRP) at each ancestral location state estimated over time inferred using the continuous-time Markov chain model. Blue line represents the actual TRP over time for environmental and clinical transition. The grey lines represent the null distribution of the randomized tip states (environmental and clinical) for the TRP. (D) Demographic history of *V. cholerae* in Haiti depicted by effective population size (N_e) estimates inferred from Bayesian phylogeny using the coalescent framework (Skyride demographic prior). Solid grey line corresponds to the mean N_e estimate, while gray shades indicate upper and lower bounds of

95% high posterior density interval of *Nₑ* estimates (y-axis in logarithmic scale) over time (x-axis).

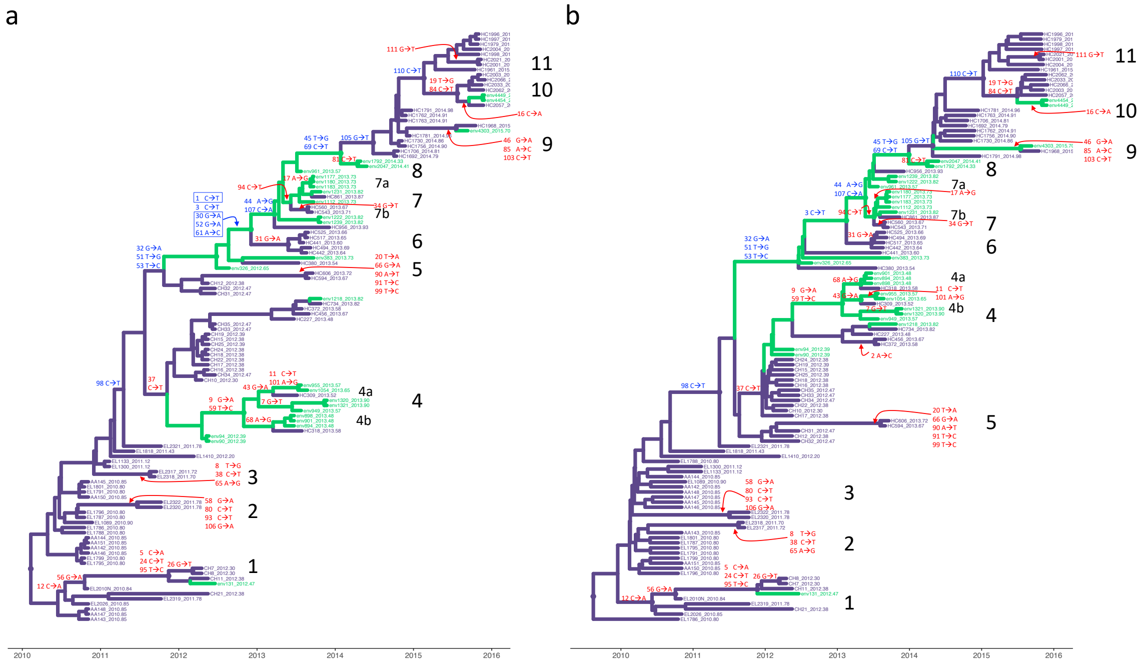

Figure S3. Mutations that differentiate surviving and waning lineages. Unambiguous and unique SNPs (uuSNPs) are traced on branches for both the MCC tree obtained applying either the (a) classical discrete phylogeography (BEAST1) and (b) BASTA (BEAST2) approaches. In blue are represented uuSNPs that define the surviving linage (located on the trunk of the tree), while in red the ones that represent the lineages waning in subsequent waves (located in internal branches that do not include terminal branches, as these are not representative of the lineage).

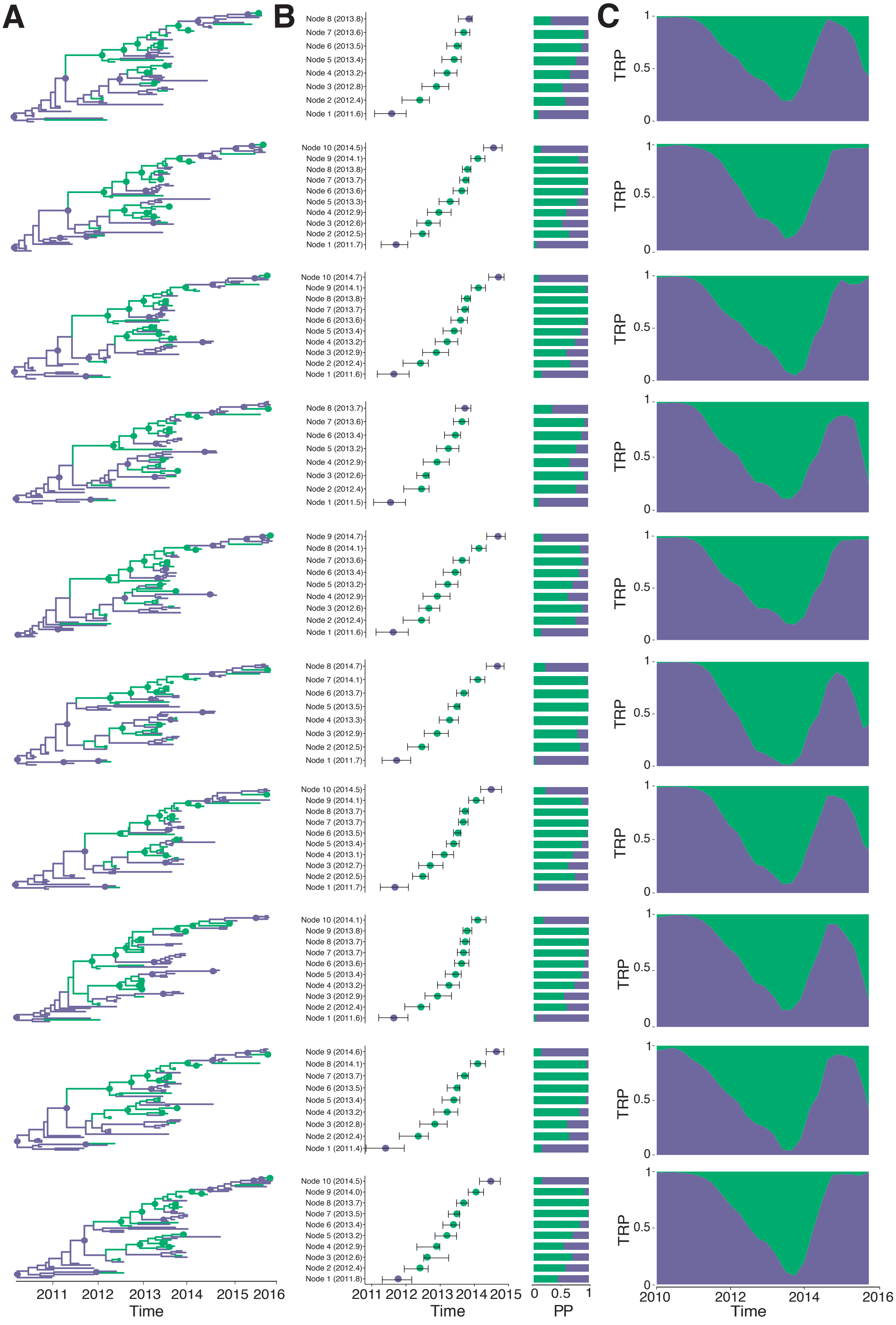
A
B
C
Node 8 (2013.8)
Node 7 (2013.6)
Node 6 (2013.5)
Node 5 (2013.4)
Node 4 (2013.2)
Node 3 (2012.8)
Node 2 (2012.4)
Node 1 (2011.6)
Node 10 (2014.5)
Node 9 (2014.1)
Node 8 (2013.8)
Node 7 (2013.7)
Node 6 (2013.6)
Node 5 (2013.3)
Node 4 (2012.9)
Node 3 (2012.6)
Node 2 (2012.5)
Node 1 (2011.7)
Node 10 (2014.7)
Node 9 (2014.1)
Node 8 (2013.8)
Node 7 (2013.7)
Node 6 (2013.6)
Node 5 (2013.4)
Node 4 (2013.2)
Node 3 (2012.9)
Node 2 (2012.4)
Node 1 (2011.6)
Node 8 (2013.7)
Node 7 (2013.6)
Node 6 (2013.4)
Node 5 (2013.2)
Node 4 (2012.9)
Node 3 (2012.6)
Node 2 (2012.4)
Node 1 (2011.5)
Node 9 (2014.7)
Node 8 (2014.1)
Node 7 (2013.6)
Node 6 (2013.4)
Node 5 (2013.2)
Node 4 (2012.9)
Node 3 (2012.6)
Node 2 (2012.4)
Node 1 (2011.6)
Node 8 (2014.7)
Node 7 (2014.1)
Node 6 (2013.7)
Node 5 (2013.5)
Node 4 (2013.3)
Node 3 (2012.9)
Node 2 (2012.5)
Node 1 (2011.7)
Node 10 (2014.5)
Node 9 (2014.1)
Node 8 (2013.7)
Node 7 (2013.7)
Node 6 (2013.5)
Node 5 (2013.4)
Node 4 (2013.1)
Node 3 (2012.7)
Node 2 (2012.5)
Node 1 (2011.7)
Node 10 (2014.1)
Node 9 (2013.8)
Node 8 (2013.7)
Node 7 (2013.7)
Node 6 (2013.6)
Node 5 (2013.4)
Node 4 (2013.2)
Node 3 (2012.9)
Node 2 (2012.4)
Node 1 (2011.6)
Node 9 (2014.6)
Node 8 (2014.1)
Node 7 (2013.7)
Node 6 (2013.5)
Node 5 (2013.4)
Node 4 (2013.2)
Node 3 (2012.8)
Node 2 (2012.4)
Node 1 (2011.4)
Node 10 (2014.5)
Node 9 (2014.0)
Node 8 (2013.7)
Node 7 (2013.5)
Node 6 (2013.4)
Node 5 (2013.2)
Node 4 (2012.9)
Node 3 (2012.6)
Node 2 (2012.4)
Node 1 (2011.8)
TRP
1
0.5
0
2011 2012 2013 2014 2015 2016
Time
2011 2012 2013 2014 2015
Time
0 0.5 1
PP
2010 2012 2014 2016
Time

Figure S4. Contribution of toxigenic *V. cholerae* O1 environmental isolates to the evolution of the cholera epidemic in Haiti between 2012 and 2014. (A) MCC phylogenies for 10 subsampled data sets of 172 environmental and clinical toxigenic *V. cholerae* O1 isolates collected between October 2010 and December 2015 (SI Appendix, Figure S3). MCC phylogenies were inferred from genome-wide hqSNP data using the Bayesian phylogeography framework implemented in BEAST package v1.8.4. Branch lengths are scaled in time by enforcing a strict molecular clock. Environmental and clinical states are indicated in green and violet, respectively. Diamonds indicated high posterior probability (PP) support on nodes (PP>0.9). (B) Ancestral state reconstruction and PP at each environmental node (green) contributing to the trunk proportion of the tree, as compared to neighbors clinical nodes (violet). (C) Trunk rewards proportion (TRP) at each ancestral location state estimated over time inferred using the continuous-time Markov chain model. Green and purple shaded areas represent the trunk proportions over time for environmental and clinical transition, respectively.

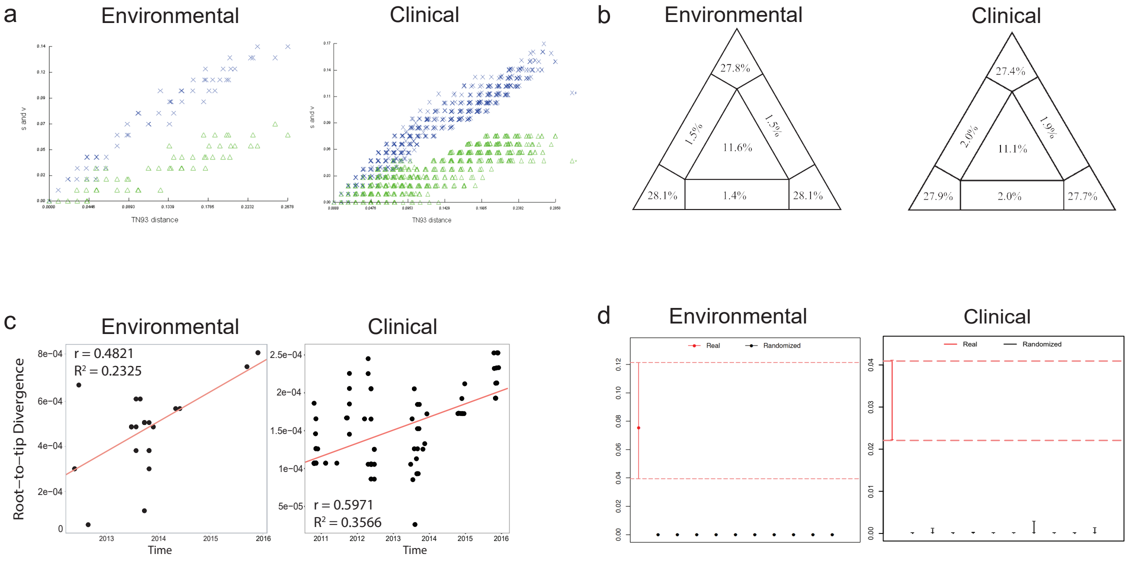

Figure S5. Estimation of phylogenetic signal and substitution saturation among the environmental only toxigenic *V. cholerae* O1 isolates and the clinical only toxigenic *V. cholerae* O1 isolates collected in the department of Ouest, Haiti. (A) Substitution saturation was gauged by plotting pairwise nucleotide transition (s) and transversion (v) substitutions *vs.* genetic distances

estimated by TN93 nucleotide substitution model for each data set. (B) Presence of phylogenetic signal was evaluated by likelihood mapping checking for alternative topologies (tips), unresolved quartets (center) and partly resolved quartets (edges) for each data set. (C) Linear regression of root-to-tip genetic distance within the ML phylogeny against sampling time for each taxa. Temporal resolution was assessed using the slope of the regression, with positive slope indicating sufficient temporal signal. Correlation coefficient "r" are reported for each data set. (D) The estimates obtained by randomizing the sampling times with 10 randomizations (black) are shown versus the estimates obtained with the correct sampling times (red), with red dashed horizontal lines showing the standard deviation on either side of the mean for the correct sampling times.

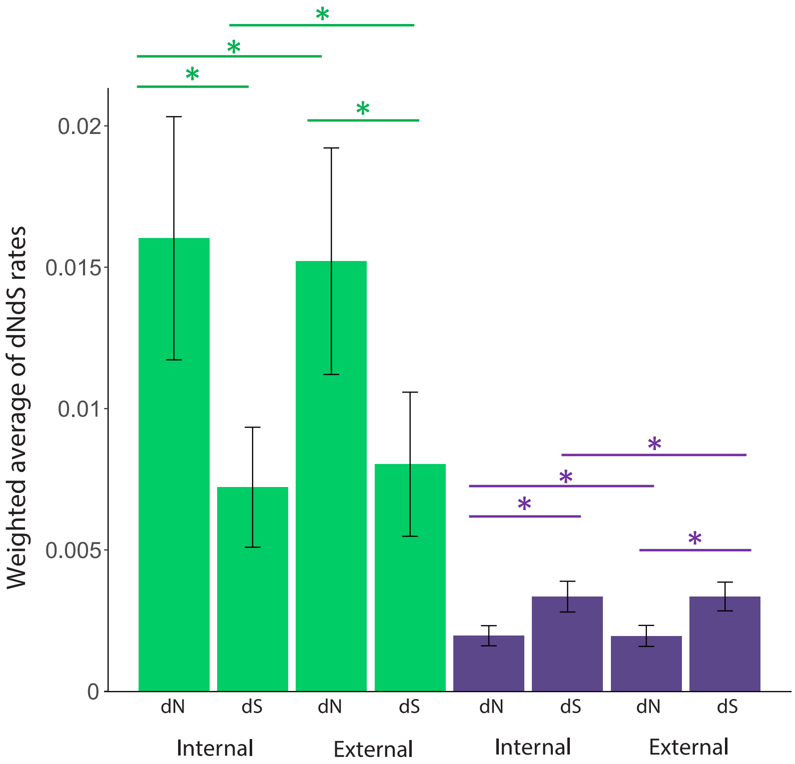

Figure S6. Weighted average of non-synonymous and synonymous substitution rates for environmental and clinical toxigenic *V. cholerae* O1 isolates. Estimates for environmental isolates (green) and clinical strains (violet) were based on 200 randomly sampled trees from the posterior distribution of molecular clock calibrated Bayesian phylogenies. Internal refers to estimate based on all internal branches of the

tree, while external refers to estimates based on terminal branches. An asterisk indicates significant ($p < 0.001$) difference between rate estimates.

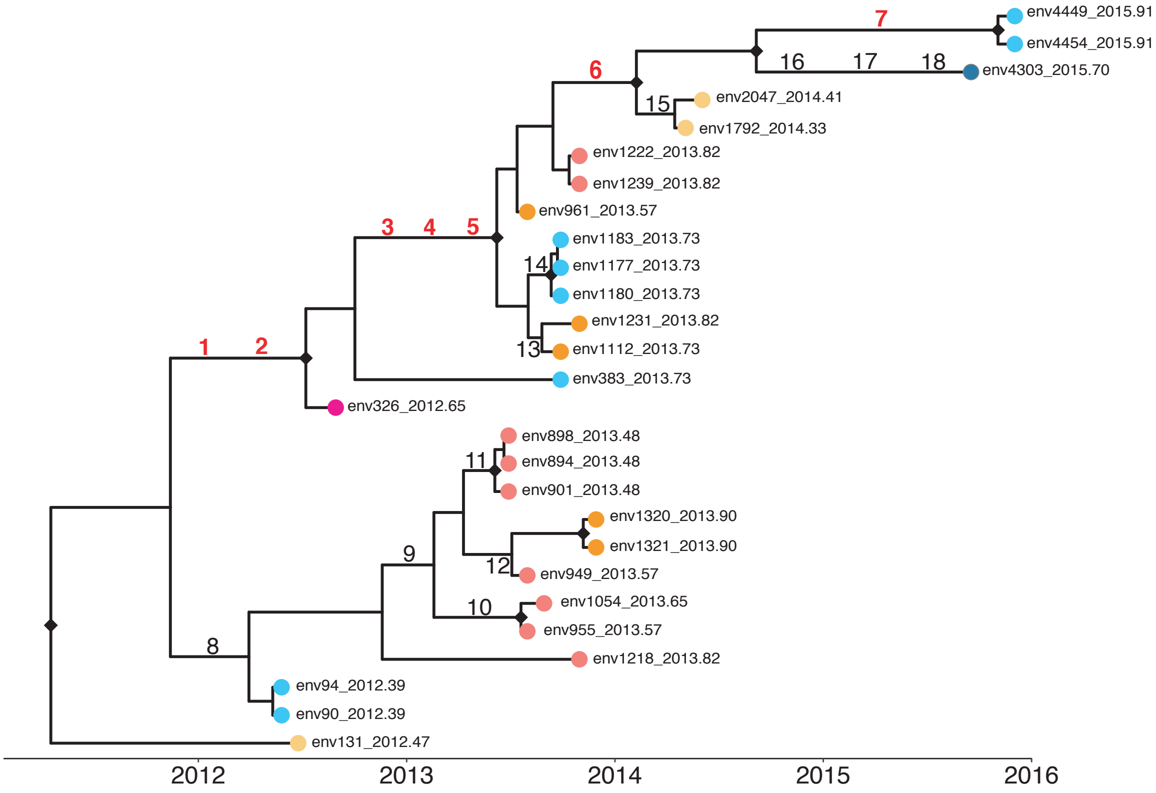

Figure S7. HqSNPs specific to environmental *V. cholerae* O1 isolates. Phylogenetic relationship of 27 environmental isolates collected between 2012 and 2015 in Haiti and non-synonymous mutations reconstructed by Bayesian inference of ancestral states, are indicated along the backbone of the tree. The hqSNPs are numbered as reported in Table S7 (SI Appendix).

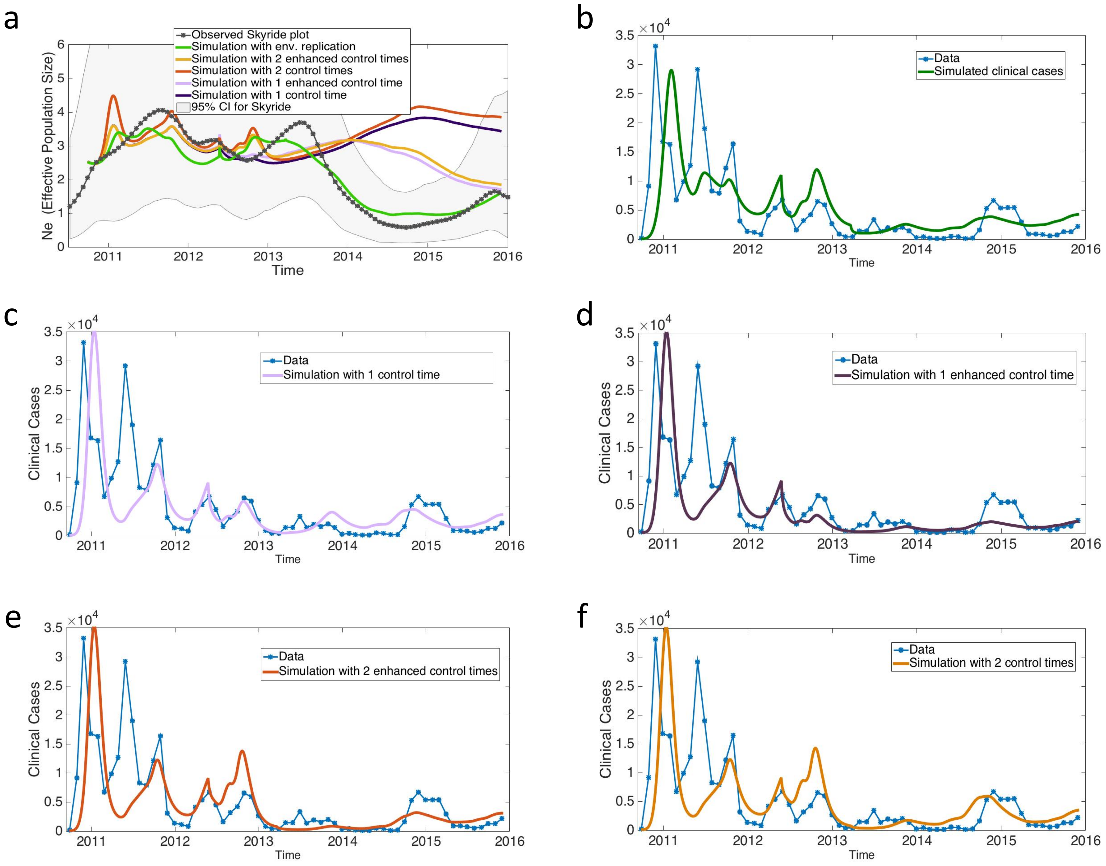

Figure S8. Calibration of the mixed transmission model of cholera. (A) *Ne* (effective population size) in simulation examples calibrated to incidence data and the observed Skyride plot of *Ne* in cases of environmental replication and without environmental replication; (B) Corresponding clinical case trajectory compared to data in the case of environmental replication, and (C), (D) (E),(F) no environmental replication with distinct control assumptions**.**

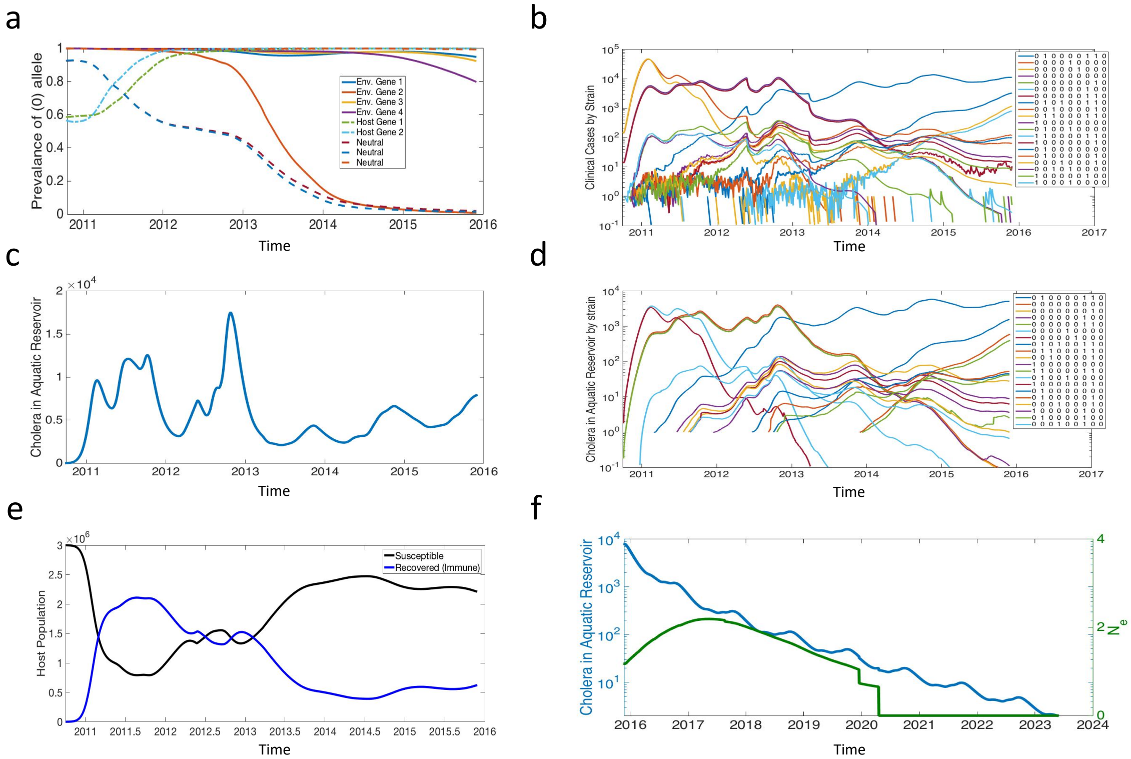

Figure S9. Calibration of the mixed transmission model of cholera: simulation with environmental replication and calibrated parameters. (A) frequencies of the (0) allele versus time for each loci, (B) the 20 strains with largest average size in host population, (C) the total aquatic reservoir pathogen concentration, (D) the 20 strains with largest average size in aquatic reservoir, (E) the total susceptible and recovered (immune) individuals, and (F) decay of Cholera in the aquatic reservoir in the absence of host transmission starting at end of 2015.

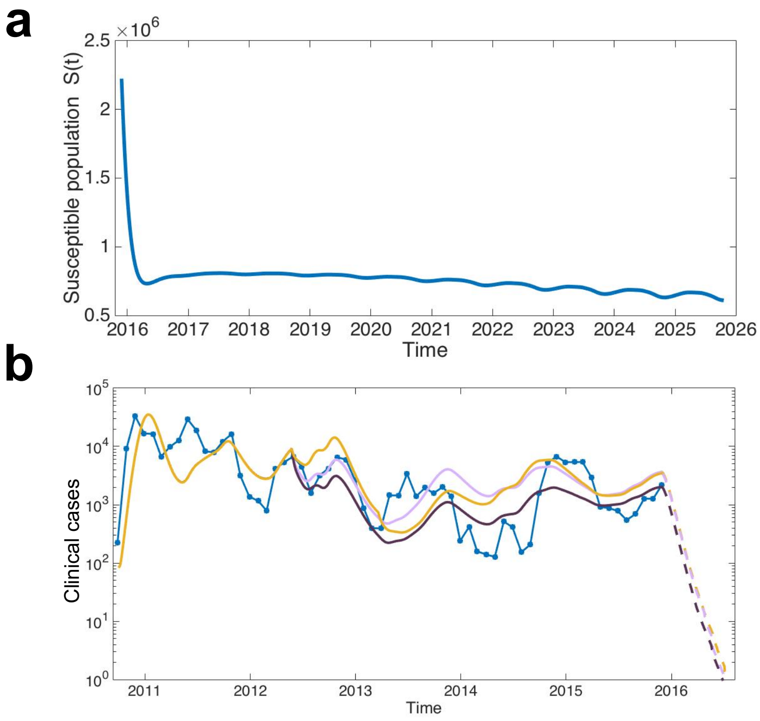

Figure S10. Vaccination prediction based on mixed transmission model of cholera. (A) Plot of susceptible versus time for simulation from Figure 4b in main text with 64% vaccination coverage in the presence of environmental replication. (B) Including vaccination at rate of δ = .01 per day for model simulations without environmental replication clears cholera rapidly within a year. Note that vaccination starts at end of 2015 and results in approximately 64% reduction in number of susceptible individuals (to about 820,000) and is roughly equivalent to the reduction modeled by Kirpich et al. [2].

Supplementary Tables

Dataset S1: List of V. cholerae O1 isolates used in this study. See supplementary excel file “Dataset S1”.

Dataset S2. List of monthly cases counts in Ouest and V. cholerae O1 isolates used in this study. See supplementary excel file “Dataset S2”.

Dataset S3. Whole genome high quality single nucleotide polymorphisms (hqSNPs). See supplementary excel file “Dataset S3”.

Table S1. Model selection of molecular clock and Bayesian demographic models to infer time-structured phylogeny for environmental and clinical *V. cholerae* O1 isolates collected in Ouest Department between October 2010 and December 2015. Log marginal likelihood (log ml) values obtained by Stepping Stone (SS) and Path Sampling (PS), are reported for models that either used as priors: strict (SC) or uncorrelated relaxed lognormal (UCLN) molecular clocks, and constant (CONST), non-parametric Bayesian skyline (BSP) or Gaussian Markov random field Bayesian skyride (GMRF) demographic models. Bayes Factor (BF) values are reported as ln(BF)ss and ln(BF)ps for PS and SS, respectively.

Model	log ml SS	Ln(BF)ss	log ml PS	Ln(BF)ps
SC CONST	-1144.57	**2.8**	-1144.70	**2.8**
UCLN CONST	-1147.41		-1147.50	
SC BSP	-1140.87	**3.4**	-1140.42	**3.3**
UCLN BSP	-1137.46		-1137.07	
SC GMRF	-1144.72	**1.8**	-1144.7	**1.2**
UCLN GMRF	-1142.89		-1143.11	
SC CONST	-1144.57	**3.7**	-1144.70	**4.3**
SC BSP	-1140.87		-1140.42	
SC CONST	-1144.57	**4.5**	-1144.70	**4.4**
SC GMRF	-1144.72		-1144.72	
SC BSP	-1140.87	**3.8**	-1140.42	**3.9**
SC GMRF	-1144.72		-1144.72	

UCLN CONST	-1147.41	**9.9**	-1147.50	**10.4**
UCLN BSP	-1137.46		-1137.07	
UCLN CONST	-1147.41	**4.5**	-1147.50	**4.4**
UCLN GMRF	-1142.89		-1143.11	
UCLN BSP	-1137.46	**5.4**	-1137.07	**6.0**
UCLN GMRF	-1142.89		-1143.11	

Table S2. Unambiguous and unique SNPs (uuSNPs) that define 11 waning lineages in the DTA and BASTA MCC phylogenies.

Waning lineages	uuSNPs
1	5 C→A (Gene Vch1786_I0077) 12 C→A (Gene Vch1786_I0381) 24 C→T (Gene X Vch1786_I0823XXX) 26 G→A (Gene lplA) 56 G→A (Gene Vch1786_I1973) 95 T→C (Gene Vch1786_II0263)
2	58 G→A (Gene Vch1786_I2046) 80 C→T (Gene Vch1786_I2813) 93 C→T (Gene Vch1786_II0121) 106 G→A (Gene speG)
3	8 T→G (Gene tgt) 38 C→T (Gene dinG) 65 A→G (Gene gyrB)
4	9 G→A (Gene Vch1786_I0310) 37 C→T (Gene ompT) 59 T→C (Gene epsK) **a:** 7 G→T (Gene Vch1786_I0159) 11 C→T (Gene recN) 43 G→A (Gene rluC) 101 A→G (Gene Vch1786_II0462) **b:** 68 A→G (Gene Vch1786_I2471)
5	20 T→A (Gene purT) 66 G→A (Gene glmU) 90 A→T (Gene Vch1786_II0065) 91 T→C (Gene Vch1786_II0065) 99 T→C (Gene Vch1786_II0384)
6	31 G→A (Gene Vch1786_I1104)
7	94 C→T (Gene Vch1786_II0117) **a:** 17 A→G (Gene Vch1786_I0588) **b:** 34 G→T (Gene Vch1786_I1144)
8	81 C→T (Gene Vch1786_I2822)
9	46 G→A (Gene Vch1786_I1618) 85 A→C (Gene Vch1786_II0051) 103 C→T (Gene Vch1786_II0532)
10	16 C→A (Gene Vch1786_I0558) 19 T→G (Gene Vch1786_I0664) 84 C→T (Gene Vch1786_II0001)

11	111 G→T (Gene Vch1786_II1009)

Table S3. Model selection of molecular clock and Bayesian demographic models to infer time-structured phylogeny for environmental *V. cholerae* O1 isolates collected between October 2010 and December 2015. Log marginal likelihood (log ml) values obtained by Stepping Stone (SS) and Path Sampling (PS), are reported for models that either used as priors: strict (SC) or uncorrelated relaxed lognormal (UCLN) molecular clocks, and constant (CONST), non-parametric Bayesian skyline (BSP) or Gaussian Markov random field Bayesian skyride (GMRF) demographic models. Bayes Factor (BF) values are reported as ln(BF)ss and ln(BF)ps for PS and SS, respectively.

Model	log ml SS	Ln(BF)ss	log ml PS	Ln(BF)ps
SC CONST	-298.08	**0.5**	-298.09	**0.5**
UCLN CONST	-298.60		-298.59	
SC BSP	-299.38	**1.3**	-299.25	**1.3**
UCLN BSP	-300.68		-300.57	
SC GMRF	-297.08	**1.7**	-296.99	**1.6**
UCLN GMRF	-298.83		-298.61	
SC CONST	-298.09	**1.0**	-298.10	**1.1**
SC BSP	-299.38		-299.25	
SC CONST	-298.09	**0.7**	-298.10	**0.5**
SC GMRF	-298.83		-298.61	
SC BSP	-299.38	**0.5**	-299.25	**0.6**
SC GMRF	-298.83		-298.61	

Table S4. Model selection of molecular clock and Bayesian demographic models to infer time-structured phylogeny for clinical *V. cholerae* O1 isolates collected between October 2010 and December 2015. Log marginal likelihood (log ml) values obtained by Stepping Stone (SS) and Path Sampling (PS), are reported for models that either used as priors: strict (SC) or uncorrelated relaxed lognormal (UCLN) molecular clocks, and constant (CONST), non-parametric Bayesian

skyline (BSP) or Gaussian Markov random field Bayesian skyride (GMRF) demographic models. Bayes Factor (BF) values are reported as ln(BF)ss and ln(BF)ps for PS and SS, respectively.

Model	log ml SS	Ln(BF)ss	log ml PS	Ln(BF)ps
SC CONST	-889.44	**1.1**	-889.54	**1.1**
UCLN CONST	-888.29		-888.39	
SC BSP	-884.66	**1.0**	-884.64	**0.9**
UCLN BSP	-883.68		-883.71	
SC GMRF	-891.13	**0.2**	-891.20	**0.3**
UCLN GMRF	-891.28		-891.53	
SC CONST	-889.44	**4.8**	-889.54	**4.9**
SC BSP	-884.66		-884.64	
SC CONST	-889.44	**1.7**	-889.54	**1.7**
SC GMRF	-891.13		-891.20	
SC BSP	-884.66	**6.5**	-884.64	**6.6**
SC GMRF	-891.13		-891.20	

Table S5. Weighted average of synonymous (K_S) and non-synonymous (K_A) substitution rates for environmental or clinical toxigenic *V. cholerae* O1 isolates.

Strains	Selection	Internal	External	*p-value**	*p-value***
Env	K_S *(SD)*	0.0072 (0.0021)	0.0080 (0.0025)	1E-04	1E-04
	K_A *(SD)*	0.016 (0.0043)	0.0152 (0.0040)	1E-04	1E-04
Cln	K_S *(SD)*	0.0033 (5.4E-04)	0.0033 (5.1E-04)	1E-04	
	K_A *(SD)*	0.0019 (3.5E-04)	0.0019 (3.7E-04)	1E-04	

Env= environmental; Cln= clinical; * intra comparison; ** inter comparison (environmental *vs* clinical).

Table S6. Ratio of non-synonymous (*dN*) and synonymous (*dS*) substitution rates for environmental or clinical toxigenic *V. cholerae* O1 isolates.

Strains	Internal *dN/ dS*	95% interval		*p-value*	External *dN/ dS*	95% interval		*p-value*
		Lower	Upper			Lower	Upper	
Env	2.22	2.20	2.23	1E-04	1.89	1.87	1.91	1E-04
Cln	0.58	0.583	0.589	1E-04	0.58	0.581	0.587	1E-04

Cln= clinical; Env= environmental

Table S7. List of genes with hqSNPs specific to environmental *V. cholerae* O1 isolates.

SNP number	Chr	Pos	Ref	Var	Codon Change	AA Change	Gene ID	N16961 Gene ID	Product	Function
1	1	1211305	G	A	aCc -> aTc	T489I	Vch1786_I1120		Conserved hypothetical protein	
2	1	2094531	T	G	Aaa -> Caa	K279Q	exeA	gspA	General secretion pathway protein A	Type II secretion system (TSII): role in virulence and survival of *V. cholerae* in different environmental niches
3	1	1675595	A	G	aAg -> aGg	K101R	Vch1786_I1539		Conserved hypothetical protein	
4	1	2372132	A	C	Ttt -> Gtt	F145V	epsG	epsG	General secretion pathway protein G	Type II secretion system (TSII): role in virulence and survival of *V. cholerae* in different environmental niches
5	2	767582	C	A	gaG -> gaT	E579D	Vch1786_II0794		Conserved hypothetical protein	
6	1	1740013	T	G	Ttc -> Gtc	F373V	Vch1786_I1601		5-methylaminomethyl-2-thiouridine-methyltransferase	involved in the biosynthesis of the modified nucleoside 5-methylaminomethyl-2-thiouridine (mnm5s2U) present in the wobble position of some tRNAs[12]
7	1	711046	T	G	Acc -> Ccc	T34P	Vch1786_I0664		Gonadoliberin III-related protein	
8	1	1464818	C	T	Gct -> Act	A140T	ompT		OmpT protein	ToxR-regulated porin OmpT [13]
9	1	2370188	T	C	Aag -> Gag	K112E	epsK		General secretion pathway protein K	Type II secretion system (TSII): role in virulence and survival of *V. cholerae* in different environmental niches
10	2	416932	A	G	gAa -> gGa	E369G	Vch1786_II0462	mcpH_1	Methyl-accepting chemotaxis protein	Chemotaxis and motility of *V. cholerae* in response to fluctuating environmental cues [14]
11	1	2618679	A	G	Acc -> Gcc	T460A	Vch1786_I2471	prlC	Oligopeptidase A	
12	1	91519	C	T	aCc -> aTc	T34I	Vch1786_I0077		Transposase Tn3	
13	1	2441365	G	A	Gtt -> Att	V370I	Vch1786_I2290		Uroporphyrin-III C-methyltransferase	

14	1	628067	A	G	gAc -> gGc	D361G	Vch1786_I0588	cpxA	Sensor histidine kinase	
15	1	2962330	C	T	Gtc -> Atc	V331I	Vch1786_I2822	rpoD	β subunit of rpoD, RNA polymerase primary sigma factor	Preferentially transcribes genes associated with exponential growth in response to nutrients [15]
16	1	1753647	G	A	cCg -> cTg	P144L	Vch1786_I1618		Flagellar hook-length control protein FliK	Controls hook-length of the flagellum and promotes biofilm formation in the extra-intestinal environment [16}
17	2	45300	A	C	gAg -> gCg	E91A	Vch1786_II0051		Deoxyguanosinetriphosphate triphosphohydrolase-related protein	
18	2	487235	C	T	Ccg -> Tcg	P295S	Vch1786_II0532		Glyceraldehyde 3-phosphate dehydrogenase	

Chr= chromosome; Pos= position in the genome; Ref= reference genome: 2010EL-1786 (GenBank: NC_016445.1 and NC_016446.1); Var= variant detected.
Table references: See references.

Table S8. Model variables and parameters

Variable/Parameter	Meaning	Value
$S(t)$	Susceptible	–
$I_j(t)$	Infected with strain j, j = 1 . . . 2^L	–
$W_j(t)$	Strain j concentration in aquatic reservoir, j = 1 . . . 2^L	–
$R_i(t)$	Recovered with temporary im-munity at stage i, i = 1, 2, 3	–
N	population size	3×10^6 [2]
μ	birth and death rate	1/(365 × 55) [2]
β_I	baseline host-host transmission rate	3.5×10^{8} [1]
$\beta_{W,0}$	baseline environment-host aver-age (seasonal) transmission rate	1.67×10^{7} [1]
γ	recovery rate	1/7 [2]
ξ_0	average (seasonal) pathogen shedding rate	0.0278 [1]
r	pathogen environmental replica-tion rate	0.02
K	pathogen environmental carry-ing capacity	1.08×10^6 [2]
v_0	baseline pathogen average (sea-sonal) environmental decay rate	0.035
σ	Case fatality rate	0.02 [2]
α	waning rate	1/365 [2]
a_1, a_2, a_3	rainfall amplitude for $\beta_W(t)$, $\xi(t)$, $v(t)$	$a_1 = 0.05, a_2 = 0.5, a_3 = 0.5$
L	# of loci	9
n, m, k	# of environmental, host-host, & neutral loci	n = 4, m = 2, k = 3
h\`[1]	host transmission fitness factor for non-neutral mutant allele (1), \` = 1, . . . n + m	0.9 ≤ h\` < 1, \` = 1, . . . , n, 1.1 ≤ h\` ≤ 1.15, \` = n + 1, . . . , m
w\`[1]	environmental decay fitness fac-tor for non-neutral mutant allele (1), \` = 1, . . . n + m	0.65 ≤ w\` < 0.81, \` = 1, . . . , n, 1.5 ≤ h\` ≤ 1.75, \` = n+1, . . . , m
η	epistasis fitness exponent	0.65
ϵ	(base) mutation rate	5×10^{5}
τ	time rescaling constant [2]	1.44
q	proportion of symptomatic (clinical) cases	1/4 [2]
f_n	neutral effective population size rescaling constant	0.15

[1] See (3) for how fitness factors combine to determine strain j host transmissibility fitness factor and environmental survivability factor, κ_j and ρ_j.

[2] A rescaling of time is chosen to better fit the Ne dynamics over time period.

References

[1] Lee EC, et al. (2017) Model distinguishability and inference robustness in mechanisms of cholera transmission and loss of immunity. Journal of theoretical biology 420:68–81.

[2] Kirpich A, et al. (2015) Cholera transmission in ouest department of haiti: dynamic modeling and the future of the epidemic. PLoS neglected tropical diseases 9(10):e0004153.

[3] Bani-Yaghoub M, Gautam R, Shuai Z, Van Den Driessche P, Ivanek R (2012) Reproduction numbers for infections with free-living pathogens growing in the environment. Journal of biological dynamics 6(2):923–940.

[4] Lucien MAB, et al. (2019) Cholera outbreak in haiti: Epidemiology, control, and prevention. Infectious Diseases in Clinical Practice 27(1):3–11.

[5] Rebaudet S, et al. (2019) The case-area targeted rapid response strategy to control cholera in haiti: a four-year implementation study. PLoS neglected tropical diseases 13(4):e0007263.

[6] Roche B, Drake JM, Rohani P (2011) The curse of the pharaoh revisited: evolutionary bi-stability in environmentally transmitted pathogens. Ecology letters 14(6):569–575.

[7] Safa A, Nair GB, Kong RY (2010) Evolution of new variants of vibrio cholerae o1. Trends in microbi-ology 18(1):46–54.

[8] van Deutekom HW, Wijnker G, de Boer RJ (2013) The rate of immune escape vanishes when multiple immune responses control an hiv infection. The Journal of Immunology 191(6):3277–3286.

[9] Volz EM, Pond SLK, Ward MJ, Brown AJL, Frost SD (2009) Phylodynamics of infectious disease epidemics. Genetics 183(4):1421–1430.

[10] Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics 10(3):195.

[11] Volz EM (2012) Complex population dynamics and the coalescent under neutrality. Genetics 190(1):187–201.

[12] Hagervall T, Edmonds C, McCloskey J, Björk G (1987) Transfer rna (5-methylaminomethyl-2-thiouridine)-methyltransferase from escherichia coli k-12 has two enzymatic activities. Journal of Bio-logical Chemistry 262(18):8488–8495.

[13] Provenzano D, Klose KE (2000) Altered expression of the toxr-regulated porins ompu and ompt dimin-ishes vibrio cholerae bile resistance, virulence factor expression, and intestinal colonization. Proceedings of the National Academy of Sciences 97(18):10220–10224.

[14] Boin MA, Austin MJ, Häse CC (2004) Chemotaxis in vibrio cholerae. FEMS microbiology letters 239(1):1–8.

[15] Carter H, Wang LF, Doi R, Moran C (1988) rpod operon promoter used by sigma h-rna polymerase in bacillus subtilis. Journal of bacteriology 170(4):1617–1621.

[16] Zhu S, Kojima S, Homma M (2013) Structure, gene regulation and environmental response of flagella in vibrio. Frontiers in microbiology 4:410.