Original Paper

Eco-Evolutionary Drivers of Vibrio parahaemolyticus Sequence Type 3 Expansion: Retrospective Machine Learning Approach

Amy Marie Campbell[1,2], PhD; Chris Hauton[1], PhD; Ronny van Aerle[2], PhD; Jaime Martinez-Urtaza[2,3], PhD

[1]School of Ocean and Earth Science, University of Southampton, Southampton, United Kingdom

[2]Centre for Environment, Fisheries and Aquaculture Science (CEFAS), Weymouth, United Kingdom

[3]Department of Genetics and Microbiology, Autonomous University of Barcelona, Barcelona, Spain

Corresponding Author:
Jaime Martinez-Urtaza, PhD
Department of Genetics and Microbiology
Autonomous University of Barcelona
Facultat de Biociènces, oficina C3/109
Campus de la UAB, Bellaterra
Barcelona, 08193
Spain
Phone: 34 93 581 2729
Email: jaime.martinez.urtaza@uab.cat

Abstract

Background: Environmentally sensitive pathogens exhibit ecological and evolutionary responses to climate change that result in the emergence and global expansion of well-adapted variants. It is imperative to understand the mechanisms that facilitate pathogen emergence and expansion, as well as the drivers behind the mechanisms, to understand and prepare for future pandemic expansions.

Objective: The unique, rapid, global expansion of a clonal complex of *Vibrio parahaemolyticus* (a marine bacterium causing gastroenteritis infections) named *Vibrio parahaemolyticus* sequence type 3 (VpST3) provides an opportunity to explore the eco-evolutionary drivers of pathogen expansion.

Methods: The global expansion of VpST3 was reconstructed using VpST3 genomes, which were then classified into metrics characterizing the stages of this expansion process, indicative of the stages of emergence and establishment. We used machine learning, specifically a random forest classifier, to test a range of ecological and evolutionary drivers for their potential in predicting VpST3 expansion dynamics.

Results: We identified a range of evolutionary features, including mutations in the core genome and accessory gene presence, associated with expansion dynamics. A range of random forest classifier approaches were tested to predict expansion classification metrics for each genome. The highest predictive accuracies (ranging from 0.722 to 0.967) were achieved for models using a combined eco-evolutionary approach. While population structure and the difference between introduced and established isolates could be predicted to a high accuracy, our model reported multiple false positives when predicting the success of an introduced isolate, suggesting potential limiting factors not represented in our eco-evolutionary features. Regional models produced for 2 countries reporting the most VpST3 genomes had varying success, reflecting the impacts of class imbalance.

Conclusions: These novel insights into evolutionary features and ecological conditions related to the stages of VpST3 expansion showcase the potential of machine learning models using genomic data and will contribute to the future understanding of the eco-evolutionary pathways of climate-sensitive pathogens.

(JMIR Bioinform Biotech 2024;5:e62747) doi: 10.2196/62747

KEYWORDS

pathogen expansion; climate change; machine learning; ecology; evolution; vibrio parahaemolyticus; sequencing; sequence type 3; VpST3; genomics

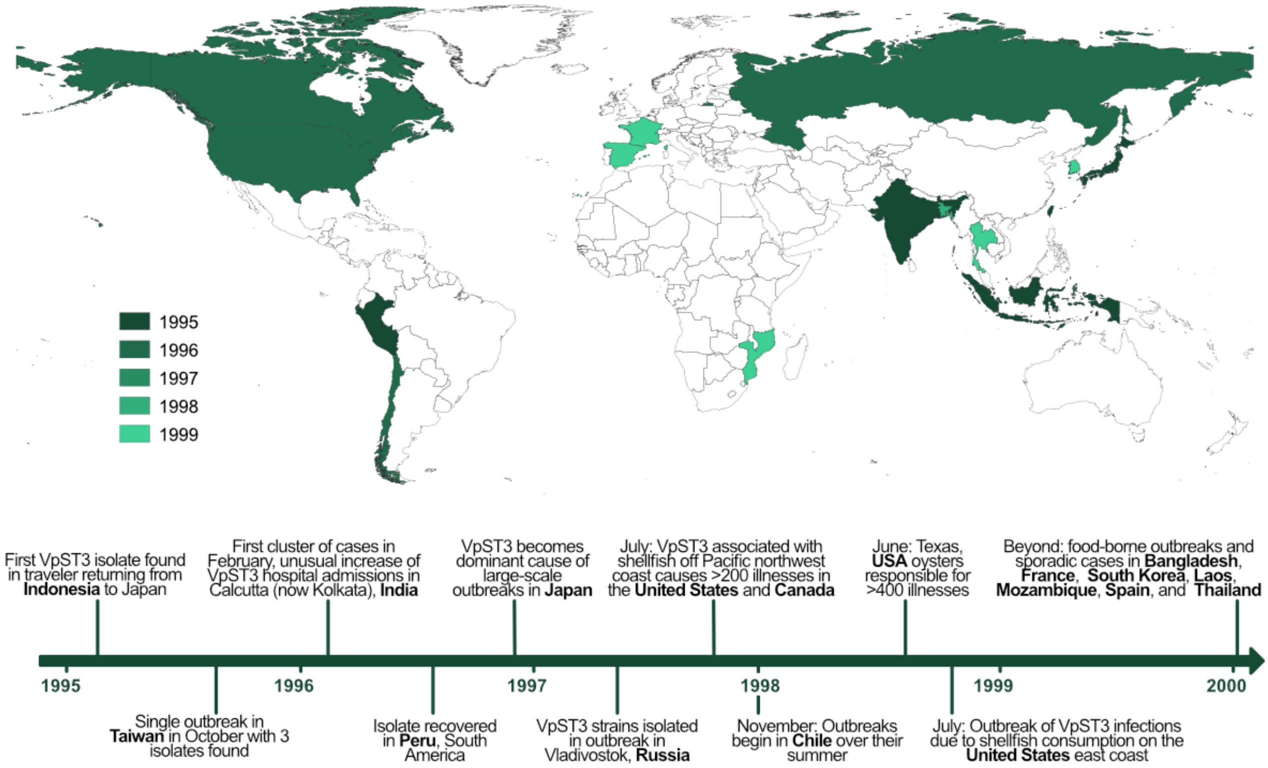
Introduction

Background

Climate change is likely to impact environmentally sensitive pathogens in terms of shifts in seasonality, expansion of suitable habitats, and the emergence and global dispersal of well-adapted variants. This has already been observed for *Vibrio parahaemolyticus* [1], a marine bacterium inhabiting coastal waters that causes acute gastroenteritis when transmitted to humans by ingestion of contaminated seafood, contributing to a large percentage of foodborne infections worldwide. Recent decades have seen this highly adaptable bacterium spread globally and increasingly cause outbreaks [1].

Before the 1990s, *Vibrio* infections were considered an exotic outcome of travel to Asia, where *Vibrio* bacteria were historically considered endemic. Up to this point, only particular strains of *Vibrio cholerae* had been designated as epidemic variants, characterized by global expansion and pandemic potential. However, transcontinental spread has now been reported for 2 *V parahaemolyticus* clonal types: sequence type 3 from Southeast Asia and, more recently, sequence type 36 from the Pacific Northwest [2,3]. The first of these instances, involving the clonal type *Vibrio parahaemolyticus* sequence type 3 (VpST3), was identified in 1996, when the unique variant, which had not been previously reported, was found to be responsible for up to 80% of the cases in a notable increase of *V parahaemolyticus* infections in Calcutta (now Kolkata), India, in 1996 [4]. This outbreak was unusual, with all recovered isolates clustered into a single homogeneous group, unlike previous outbreaks [3]. Similar isolates were then observed from outbreaks in distinct locations around the world, including Peru, Japan, Russia, Chile, and the United States [5-8] (Figure 1 [4-15]), where the variant was emerging concurrently. This included regions with conditions previously considered adverse for the presence of such pathogens. The epidemic radiations that followed in these diverse regions were the first observed for *V parahaemolyticus* and resulted in the variant supplanting local populations and rapidly becoming the most dominant *V parahaemolyticus* variant globally. As a consequence of this expansion, *V parahaemolyticus* became the second human pathogenic *Vibrio* species with an epidemic nature and, along with *V cholerae*, the only *Vibrio* species with strains capable of worldwide expansions and causing infections at a global level.

Figure 1. Timeline and map of *Vibrio parahaemolyticus* sequence type 3 (VpST3) initial expansion based on reported isolates and outbreaks in the literature (shapefile provided by Database of Global Administrative Areas).


1995
1996
1997
1998
1999
First VpST3 isolate found in traveler returning from Indonesia to Japan
First cluster of cases in February, unusual increase of VpST3 hospital admissions in Calcutta (now Kolkata), India
VpST3 becomes dominant cause of large-scale outbreaks in Japan
July: VpST3 associated with shellfish off Pacific northwest coast causes >200 illnesses in the United States and Canada
June: Texas, USA oysters responsible for >400 illnesses
Beyond: food-borne outbreaks and sporadic cases in Bangladesh, France, South Korea, Laos, Mozambique, Spain, and Thailand
1995
1996
1997
1998
1999
2000
Single outbreak in Taiwan in October with 3 isolates found
Isolate recovered in Peru, South America
VpST3 strains isolated in outbreak in Vladivostok, Russia
November: Outbreaks begin in Chile over their summer
July: Outbreak of VpST3 infections due to shellfish consumption on the United States east coast

This process of expansion results from an epidemic bacterial population structure, as described in the study by Smith et al [16]. Upon a background of numerous rapidly recombining genotypes, a limited number of very frequent genotypes are superimposed, known as clonal complexes, that have originated from highly adaptive ancestral genotypes [16]. The mechanisms behind the rise of these clonal complexes are largely unknown; yet, it is imperative to identify the conditions that allow a pathogen to emerge in such diverse locations and become dominant, as well as the drivers behind these processes, to understand and prepare for future pandemic expansions. When considering environmentally sensitive pathogens such as *Vibrio*, possible drivers can be categorized as either ecological or evolutionary. Evolutionary drivers include the processes of adaptation, mutations that increase fitness, or the uptake or horizontal transfer of beneficial accessory genes. Both ends of the spectrum of genetic diversity—generalists and specialists—are associated with pandemic expansions. Ecological drivers can range from the local environment, which affects pathogen survival and growth, to environmental corridors and transport mechanisms. Importantly, these ecological and evolutionary drivers are not exclusive and, instead, interact significantly, with this interplay known as "eco-evolutionary." A key example of this would be adaptative selection occurring after arrival to a distinct marine environment. While more studies are considering the coeffect of ecological and evolutionary factors on larger species (such as vertebrates and invertebrates), little attention has been paid to environmentally sensitive pathogens. Focusing on these pathogens would provide novel insights into how particular pathogenic strains emerge [17].

V parahaemolyticus is a uniquely placed species, with a history of pandemic expansion that facilitates the study of such eco-evolutionary drivers. First, *Vibrio* are phylogenetically diverse with highly variable genomic backgrounds shaped by recombination and horizontal gene transfer [3], from which specialized variants can emerge. Second, *V parahaemolyticus* exhibits well-characterized environmental thresholds and tolerances, rapidly responding to changes in its marine environment, such as water temperature [18-23] and salinity [18,21,24-26]. Notably, anomalously high temperatures were observed between 1996 and 1998 [27,28] around the emergence of VpST3—pertinent amid the preference of *Vibrio* for warmer waters. From a combined eco-evolutionary perspective, *Vibrio* have high genome plasticity, which facilitates rapid adaptation in response to environmental changes [29], resulting in a large diversity of causative strains and resulting infection dynamics [30,31]. It would be simplistic to assume that all these diverse *V parahaemolyticus* variants respond to environmental change homogeneously, opening up the eco-evolutionary response landscape for exploration.

Study Objectives

We reconstructed this global expansion using publicly available genome sequences of VpST3 from clinical and environmental sources, isolated from around the world over the period of expansion of this clone, to identify population structure and demographic shifts indicative of the different stages of expansion, including emergence and establishment. We investigated the possible drivers of the expansion and our ability to predict the dynamics of VpST3 by testing a range of evolutionary and ecological drivers in a combined approach using machine learning models to elucidate the complex mechanisms that, when combined, may facilitate such a rapid, global expansion. Machine learning has been credited for its ability to harness the predictive power of evolution, using pattern recognition to uncover complex associations between biological processes [32]; therefore, it is well-placed for the novel exploration of interacting eco-evolutionary mechanisms in combination. Understanding the evolutionary features and ecological conditions related to the stages of pathogen expansion is a crucial step in understanding future eco-evolutionary pathways of climate-sensitive pathogens.

Methods

The protocol for preprocessing evolutionary and ecological data into a data frame for machine learning analysis is summarized in a graphical representation in Figure S1 in Multimedia Appendix 1.

Ethical Considerations

Ethical approval was not sought for the present study because it consisted of neither human nor animal experimentation and all genomic datasets used had been previously made publicly available with reference to their ethical approval in the papers associated with these submissions.

Genomic Data

Raw sequencing datasets from a collection of 311 VpST3 isolates, representing a range of geographic areas, were acquired from public databases for genomic analyses (Table S1 in Multimedia Appendix 1). The 311 isolates covered a temporal range from 1996 to 2021, with 162 (52.1%) from Asia, 78 (25.1%) from North America, and 71 (22.8%) from South America. A series of genetic markers were used to confirm that the isolates were VpST3 using multilocus sequence typing in MLST (version 2.11) [33]. Our analysis was restricted to isolates that were submitted with accompanying isolation date and location details because such metadata were required for the downstream linkage with environmental variables. The raw sequences were processed using default parameters within Bactopia (version 2.0.2) [34], including quality filtering, assembly, and annotation. Core single nucleotide polymorphisms (SNPs) were identified across all sequences using parsnp (version 1.5.6) [35] to create a core genome alignment, mapped to the *V parahaemolyticus* reference genome RIMD2210633. Gubbins (version 3.1.6) [36] was used to remove recombining regions to provide a final nonrecombining core genome alignment.

Phylogenetic Analysis

TempEst (version 1.5.3) [37] was used to confirm a temporal signal and conformation to a molecular clock, followed by BEAST2 (version 2.7.6) [38] analysis to reconstruct the global phylogenetic dynamics of *V parahaemolyticus*, using BEAUti [39] and a structured coalescent within a MultiTypeTree template [40]. After sensitivity analyses on a range of models, the selected model used a relaxed log normal clock model and a general time reversible (GTR) substitution model, with a normal distribution substitution rate prior. The tip dates and discrete location attributes were used to situate the genomic evolution in space and time. The Markov chain Monte Carlo was run for 250 million states until all outputs converged (effective sample size >200), confirmed by Tracer (version 1.7.1) [41]. The final maximum clade credibility tree was generated using TreeAnnotator within BEAST2.

Encoding of Expansion Dynamics

The Bayesian phylogenetic analysis and subsequent tree structure informed the designation of a variety of classifications representing VpST3 dynamics. These classifications included populations within the collection, temporal divergence, the success of introductions, and the stages of expansion. We assigned each of these classifications to each of the 311 isolates, using set criteria applied to the phylogenetic tree (Textbox 1), and converted them into binary or categorical features to provide target variables for machine learning analysis.

Textbox 1. Criteria for encoding the numerical and categorical variables of expansion dynamics. The terms in parentheses refer to the column names of the expansion dynamics in the data frame input.

Populations (wave)

Populations within the *Vibrio parahaemolyticus* sequence type 3 (VpST3) collection were identified using TreeStructure (version 0.1.0) [42], which identifies genealogical patterns to infer population structure from time-scaled phylogenies by performing 100,000 tree simulations with a significance threshold set at *P*<.001

Temporal evolution (earlylate)

Very few VpST3 isolates were recovered in 2003, after the initial global population expansion; therefore, we specified this year as a split between the early colonizers found before this date and the later isolates recovered after the expansion

Success (success)

Failed introductions were monophyletic branches that did not split into further nodes in the phylogenetic tree, while successful introductions were those that saw downstream nodes in the same reported location

Stages of expansion (stages)

We split expansion into five defined stages: (1) initial introduction (the first node or nodes in a clade), (2) established population (the nodes in the clade after this introduction), (3) secondary introduction (the first node or nodes in a clade in a new location from the original introduction), (4) secondary establishment (the nodes in the clade after this introduction in the new location), and (5) bottleneck (the last node of a clade or a location within the clade)

Stages of expansion: binary (stages_binary)

A simplified version of the previous stages of expansion classification, reducing it to a binary classification of introduced isolates (the first instances in a clade or location) and established isolates (those that followed these introductory nodes within this clade)

Extraction of Evolutionary Driver Data

Genomic analysis was used to extract features representing possible evolutionary drivers for each isolate. Quantifying the gene content variation in the accessory pangenome in natural populations is important to understand the plasticity and adaptability of populations to environmental perturbations [43]. To obtain a metric of total genes present in each isolate, we used Roary (version 3.13.0) [44] to construct the pangenome and annotate each gene present in each isolate. We used Scoary (version 1.6.16) [45] to identify shell genes (present in 15%-95% of the population) whose presence was statistically associated (*P*<.01) with the previously assigned labels representing introduction, establishment, or success. We retained a selection of these that were common accessory genes (with a presence ranging from 5% to 95% across the isolates in the collection), annotated their function, and generated features representing the binary presence or absence features. We used single-likelihood ancestor counting within HyPhy (version 2.5.48) [46] to estimate the ratio of nonsynonymous to synonymous substitutions (dN/dS) and identify sites under significant diversifying or purifying selection (*P*<.05) in the genes of interest.

SNP mutations of relevance to the expansion process were selected using pcadapt (version 4.3.3) [47] for outlier detection based on population structure. The outliers were inferred based on principal component analysis, using the parameter K=2 and a desired false discovery rate threshold of 0.1 (q-threshold) to identify discriminatory SNP mutations associated with local adaptation. These SNPs were annotated to predict functional effects on genes using SnpEff (version 5.1) [48] and the *V parahaemolyticus* RIMD2210633 genome annotations as a reference. SNPs predicted to have nonsynonymous missense variants were retained for downstream analyses. We recorded the base found at this position for every isolate to assess whether these mutations would help the model define the evolutionary classification. To convert these into numerical values fit for machine learning applications, we reclassified the letters representing bases into numbers (A=1, C=2, T=3, G=4, and N-polymorphic=5).

Extraction of Ecological Driver Data

Time series data for sea surface temperature (SST) and salinity—2 of the most well-reported environmental drivers of *V parahaemolyticus* in the marine environment [49] —were acquired from the European Centre for Medium-Range Weather Forecasts Reanalysis version 5 [50] and the Met Office Hadley Centre's EN4.2.2 quality-controlled ocean dataset [51,52], respectively, covering the period from 1995 to 2021. We zonally extracted the climate time series data for the country of isolate recovery, using Database of Global Administrative Areas country zones provided as shapefiles, extending into coastal waters by 2 decimal degrees to extract the local conditions of the marine environment. Although the climate data were available at a monthly resolution, the majority of the genomic isolates only contained an annual resolution. Instead of averaging across the whole year, we created metrics for maximum, minimum, and mean values for each season across the year, alongside generated lagged variables from the previous year. Alongside environmental drivers, the seafood industry, including fisheries [53], seafood consumption and trade [54], and fish market contamination [55], has been previously hypothesized as a possible mechanism for the emergence and spread of *Vibrio* bacteria. We therefore extracted shellfish import data for each country from the FishStatJ database of the Food and Agricultural Organization of the United Nations as annual totals measured in 100 kg of net product weight [56] to explore the potential of this driver.

Machine Learning Approach

We combined the ecological and evolutionary driver metrics and the classification for each of the criteria into a single data frame for each isolate, with a total of 311 data points.

For our machine learning analysis, we chose a random forest classifier model, an ensemble learning method that uses bootstrapping across decision tree classifiers, due to its high interpretability and implemented the models using the Python module *scikit-learn* (version 1.3.0) [57]. For each expansion dynamic, we created three separate models: 1 model used only the evolutionary drivers as features, a second only ecological drivers as features, and a final model used both ecological and evolutionary drivers in a combined approach. We trained the random forest classifier (using 100 estimators, setting the maximum number of features to consider for best split to the square root of the total number of features, and using bootstrap samples to build trees) on a randomly selected subset of 70.1% (218/311) of the data, retaining the remaining 29.9% (93/311) as an unseen test dataset. In total, there were 109 features used to predict each evolutionary dynamic (Table S2 in Multimedia Appendix 1), of which 60 (55%) represented evolutionary drivers, and 49 (45%) represented ecological drivers. The classification output classes were either binary or categorical based on the expansion dynamic being predicted.

To test the accuracy of these predictions and provide insight into our ability to predict the expansion dynamics of VpST3, we reported 4 accuracy metrics, both per class and across all predictions, when the models were applied to the unseen test data. These metrics included precision (positive prediction rate, affected by false positives), recall (sensitivity rate, affected by false negatives), the F_1-score (a harmonic mean of precision and sensitivity, often used for comparative machine learning performance assessments), and overall accuracy (taking into account all components). We calculated the feature importance for all ecological and evolutionary drivers involved in each model using the Gini importance attribute within the random forest implementation in *scikit-learn* (version 1.3.0) (57), which is computed by the mean and SD of the accumulating impurity decrease within each tree due to the addition of each specific feature.

To assess the collinearity effects from cross-correlations between the ecological and evolutionary driver metrics contained in our model, we calculated the Spearman rank correlation coefficient between the driver variables. During model development of the individual ecological and evolutionary models, we selected features that did not exhibit significant (P<.05) collinearity. However, collinearity between evolutionary and ecological features in the combined model was explored, rather than omitted, to gain greater insight into potential eco-evolutionary associations. These significant relationships were visualized in a heat map using the *seaborn* (version 0.12.2) Python library [58].

Region-Specific Analysis

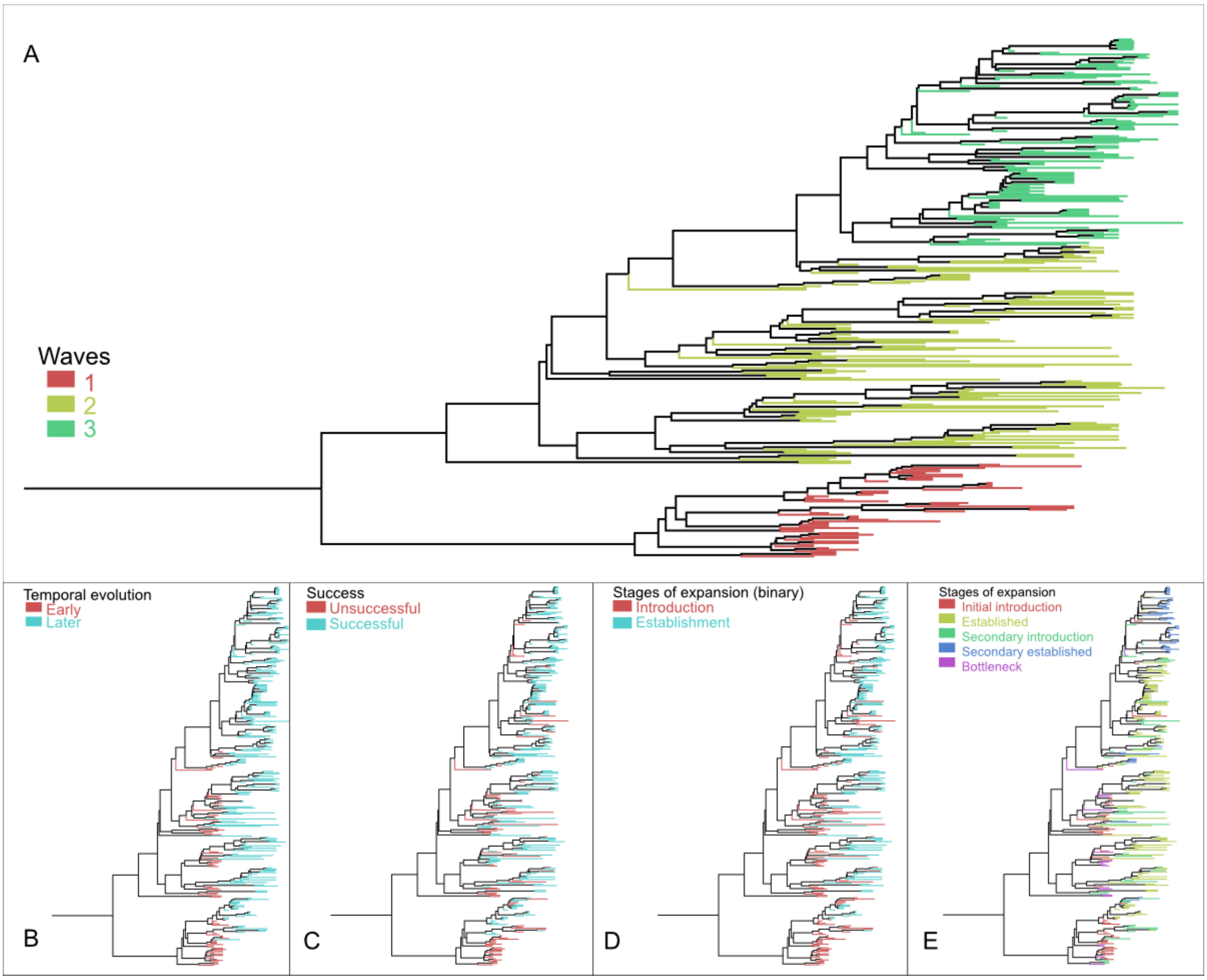
To explore the potential to understand the successful expansion of VpST3 in particular regions, we developed 2 region-specific models representing an endemic area and an area where VpST3 emerged: China and Peru, respectively. These regions were chosen because they reported the most VpST3 isolates within their respective continents and consist of distinct geographic characteristics to establish whether a regional focus on an area with specific local conditions to drive eco-evolutionary dynamics improves our ability to predict successful expansions. These models were trained and tested on regional subsets of the original data frame, with the same model parameters and features.

Results

Phylogeny Characterization

The phylogeny revealed an evolutionary population structure within the VpST3 genomes, with multiple introductions into geographically distinct locations, including secondary migrations and introductions. Our phylogenetic analysis found 3 clear population “waves” within VpST3, comprising 56 (18%), 131 (42.1%), and 124 (39.9%) of the 311 isolates (Figure 2A). In terms of temporal evolution, of the 311 isolates, 73 (23.5%) were classified as early colonizers (before 2003), and 238 (76.5%) were isolated after the initial expansion after 2003 (Figure 2B). Regarding expansion success, of the 311 isolates, 86 (27.7%) were classified as unsuccessful and 225 (72.3%) as successful (Figure 2C). With regard to the stages of expansion, of the 311 isolates, under a binary classification, 121 (38.9%) were classified as introduced, with 190 (61.1%) being classified as established (Figure 2D). When this was scaled up to the 5 stages of classification, of the 311 isolates, 49 (15.8%) were classified as initial introductions, 131 (42.1%) as established, 52 (16.7%) as secondary introductions, 58 (18.6%) as secondarily established, and finally 21 (6.8%) as representing bottlenecked populations (Figure 2E).

Figure 2. Maximum clade credibility tree with branches color coded by expansion dynamic metrics: (A) population waves, (B) temporal evolution, (C) expansion success, and (D and E) stages of expansion.



Evolutionary Features Extracted

We detected 194 potential adaptive SNP outliers within the collection of genomic isolates, of which 44 (22.7%) were predicted to be missense variants, altering an amino acid within a protein, with predicted moderate effects on particular genes (Table S3 in Multimedia Appendix 1). These SNPs were chosen as evolutionary features for the machine learning analysis. Overall, the total number of genes in each isolate ranged from 4292 to 4735, with no clear temporal signal (R^2=0.08). We identified 400 accessory genes present in 15% to 95% of the entire VpST3 collection and reduced these to 15 (3.8%) genes of interest as evolutionary features for the machine learning analysis (Table 1). This selection was based on genes that were associated with particular expansion metrics; the presence of all 15 selected genes was significantly associated (*P*<.01) with the binary classification delineating introduced and established isolates, and 5 (33%) were further associated with the successful classification metric. Annotation of these genes of interest found that most (n=5, 33%) were functionally associated with survival in the environment and tolerance to environmental conditions (Table 1). In addition, some of them (n=8, 53%) were involved in bacterial transport mechanisms, such as putrescine pathways, that promote biofilm formation. On 4 occasions, 2 versions of a gene with a similar function were identified within this group—for *pilT*, *ttcA*, CARB β-lactamase, and DeoR family transcriptional regulators. Of these 15 accessory genes, no evidence for positive diversifying selection was found (as determined by HyPhy single-likelihood ancestor counting [46]); however, 10 (67%) genes had evidence of negative, purifying selection (*P* value threshold <.10), ranging from 1 to 17 sites under purifying selection.

Table 1. Significant associations identified between accessory gene presence and key expansion dynamics.

Annotation	Function	Significance of presence association with classification labels, *P* value		
		Introduction	Establishment	Success
Lactoylglutathione lyase	Enzyme used for methylglyoxal detoxification, contributes to bacterial survival in the environment [59]	<.001	<.001	<.001
HTH[a]-type transcriptional regulator (*puuR*)	Recombinant protein, involved in putrescine pathways [60]	<.001	<.001	<.001
RNA polymerase sigma factor (*RpoS*)	Proteins that regulate transcription in bacteria, activated in response to different environmental conditions	<.001	<.001	—[b]
Type IV pilus twitching motility protein (*pilT*)	Involved in the transport (motility) of the bacteria itself, biofilms, and virulence [61]	<.001	<.001	—
Sodium:proton antiporter	Antiporters (in this case moving sodium ions in or out of a cell) play an important role in tolerance to salt stress [62]	<.001	<.001	<.001
N-carbamoylputrescine amidase (*aguB*)	Involved in biofilm production by converting N-carbamoylputrescine to putrescine [63]	<.001	<.001	<.001
Agmatine deiminase (*aguC*)	Involved in a putrescine pathway [64]	<.001	<.001	<.001
DeoR family transcriptional regulator	Primarily drives the sensing of environmental stimuli and life cycle responses [65]	<.001	<.001	—
Type IV pilus twitching motility protein (*pilT*)	Involved in the transport (motility) of the bacteria itself, biofilms, and virulence [61]	<.001	<.001	—
Carbenicillin-hydrolyzing class A beta-lactamase CARB-23	Expresses β-lactamase for resistance to antibiotic penicillins [66]	<.001	<.001	—
Ribonuclease III (*rnc*)	Modulates pathogenicity: motility, invasiveness, biofilm formation ability, and virulence [67]	<.001	<.001	—
tRNA 2-thiocytidine(32) synthetase (*ttcA*)	Involved in bacterial growth, resistance to biocides, biofilm formation, and swimming motility [68]	<.001	<.001	—
Carbenicillin-hydrolyzing class A beta-lactamase CARB-23	Expresses β-lactamase for resistance to antibiotic penicillins [66]	<.001	<.001	—
DeoR family transcriptional regulator	Primarily drives the sensing of environmental stimuli and life cycle responses [65]	<.001	<.001	—
tRNA 2-thiocytidine(32) synthetase (*ttcA*)	Involved in bacterial growth, resistance to biocides, biofilm formation, and swimming motility [68]	<.001	<.001	—

[a] HTH: helix-turn-helix.

[b] Not applicable.

Predictive Power

Overall accuracies for the different expansion metrics ranged from 0.722 to 0.967 for models using a combined eco-evolutionary approach (Figure 3). In our analysis, a combined eco-evolutionary approach almost always improved the accuracy of predicting expansion dynamics compared to using evolutionary or ecological drivers in isolation (Table 2). This was notably apparent for the predictions of the population structure within the phylogeny, in terms of the identification of 3 clear groups, in which evolutionary and ecological features individually produced accuracies of 0.733 and 0.744, respectively, but the combined approach increased the accuracy to 0.922. The only exception occurred when characterizing the success of emergence, where the ecological-only approach achieved the same accuracy as the combined approach. We could distinguish which isolates would be "successfully introduced" to an accuracy of 82% using both ecological and evolutionary data, but 13% of these were false positives, suggesting that our analysis could have overlooked a limiting factor that prevents an isolate from successfully establishing in an area.

Figure 3. Confusion matrices visualizing the predictions of random forest classifier models for each expansion dynamic when applied to unseen test datasets: (A) population waves, (B) temporal evolution, (C) expansion success, (D) binary stages of expansion, and (E) categorical stages of expansion. (B, C, and D) For binary expansion dynamics, the matrix represents (clockwise from top left) true negatives, false positives, true positives, and false negatives. (A and E) For categorical expansion dynamics, the matrix shows correct class membership and misclassified class memberships for each category.

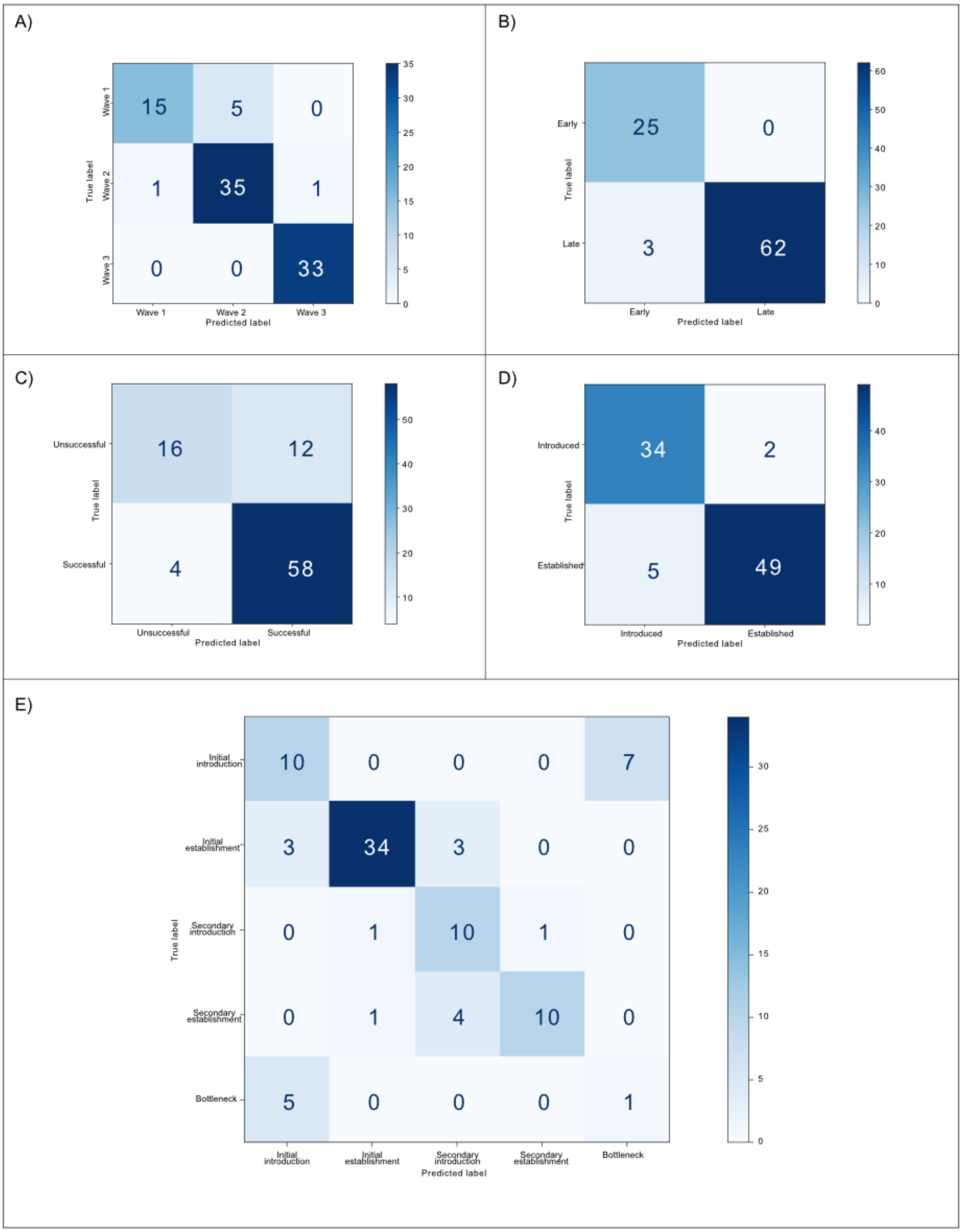


Table 2. Accuracy metrics of random forest classifier models predicting unseen test data of each expansion dynamic.

Expansion dynamics	Combined eco-evolutionary approach				Evolutionary features only				Ecological features only			
	Precision	Recall	F_1-score	Overall accuracy	Precision	Recall	F_1-score	Overall accuracy	Precision	Recall	F_1-score	Overall accuracy
Populations				0.922				0.733				0.744
Wave 1	0.938	0.750	0.833		0.529	0.450	0.486		0.900	0.450	0.600	
Wave 2	0.875	0.946	0.909		0.676	0.676	0.676		0.642	0.919	0.756	
Wave 3	0.971	1.000	0.985		0.889	0.970	0.928		0.889	0.727	0.800	
Unweighted average	0.928	0.899	0.909		0.698	0.698	0.697		0.810	0.699	0.719	
Weighted average	0.924	0.922	0.920		0.721	0.733	0.726	0.733	0.790	0.744	0.737	
Temporal evolution				0.967				0.767				0.956
Early (before 2002)	0.893	1.000	0.943		0.583	0.560	0.571		0.862	1.000	0.926	
Late (after 2003)	1.000	0.954	0.976		0.833	0.846	0.840		1.000	0.938	0.968	
Unweighted average	0.946	0.977	0.960		0.708	0.703	0.706		0.931	0.969	0.947	
Weighted average	0.970	0.967	0.967		0.764	0.767	0.765		0.962	0.956	0.956	
Stages of expansion				0.722				0.511				0.733
Initial introduction	0.556	0.588	0.571		0.412	0.412	0.412		0.526	0.588	0.556	
Established population	0.944	0.850	0.895		0.605	0.575	0.590		0.944	0.850	0.895	
Secondary introduction	0.588	0.833	0.690		0.231	0.250	0.240		0.750	0.750	0.750	
Secondary established population	0.909	0.667	0.769		0.688	0.733	0.710		0.800	0.800	0.800	
Population bottleneck	0.125	0.167	0.143		0.333	0.333	0.333		0.125	0.167	0.143	
Unweighted average	0.624	0.621	0.614		0.454	0.461	0.457		0.629	0.631	0.629	
Weighted average	0.763	0.722	0.735		0.514	0.511	0.512		0.761	0.733	0.745	
Stages of expansion (binary)				0.922				0.667				0.911
Introduction	0.872	0.944	0.907		0.583	0.583	0.583		0.868	0.917	0.892	
Establishment	0.961	0.907	0.933		0.722	0.722	0.722		0.942	0.907	0.925	
Unweighted average	0.916	0.926	0.920		0.6533	0.653	0.653		0.905	0.912	0.908	
Weighted average	0.925	0.922	0.923		0.667	0.667	0.667		0.913	0.911	0.911	
Success				0.822				0.711				0.822
Unsuccessful	0.800	0.571	0.667		0.545	0.429	0.480		0.800	0.571	0.667	
Successful	0.829	0.935	0.879		0.765	0.839	0.800		0.829	0.935	0.879	

Expansion dynamics	Combined eco-evolutionary approach				Evolutionary features only				Ecological features only			
	Precision	Recall	F_1-score	Overall accuracy	Precision	Recall	F_1-score	Overall accuracy	Precision	Recall	F_1-score	Overall accuracy
Unweighted average	0.814	0.753	0.773		0.655	0.634	0.640		0.814	0.753	0.773	
Weighted average	0.820	0.822	0.813		0.696	0.711	0.700		0.820	0.822	0.813	

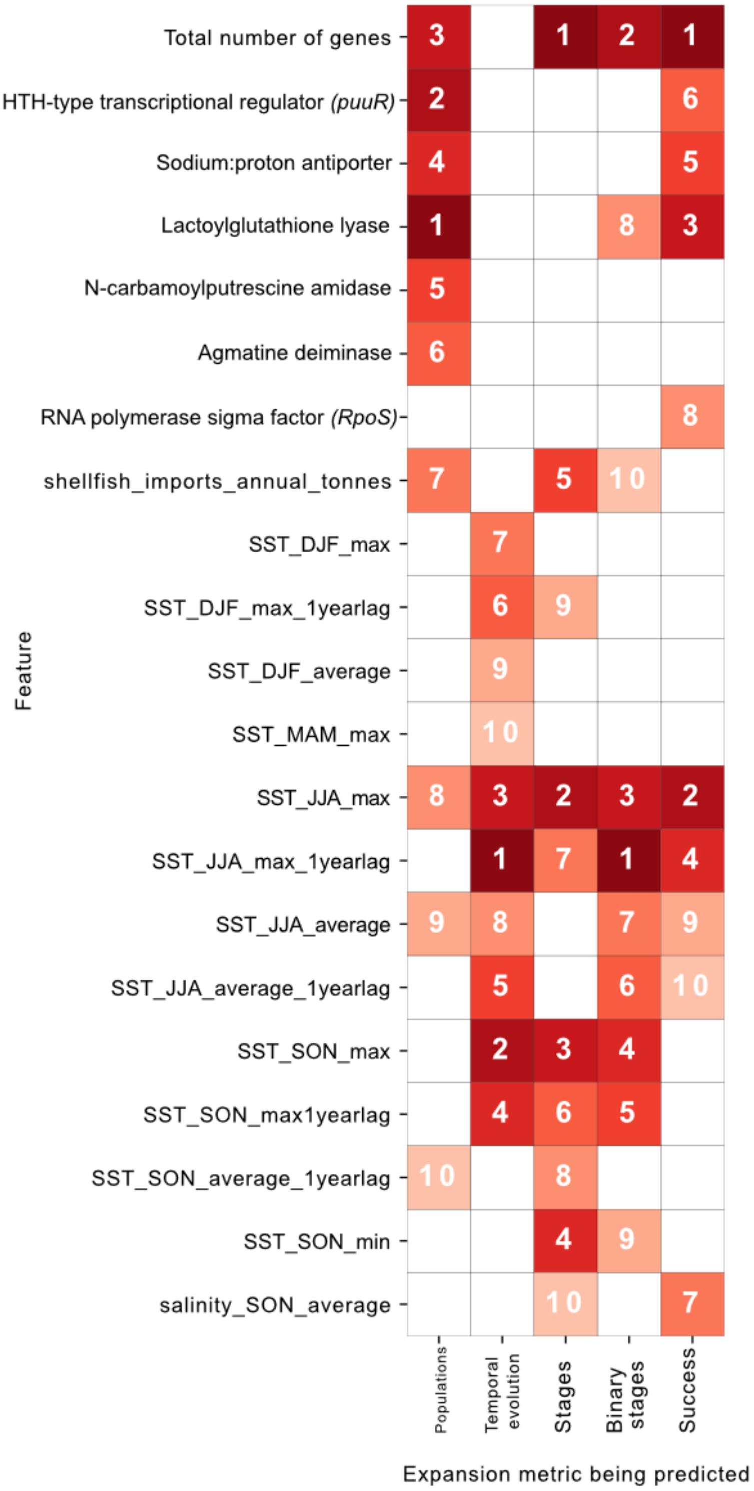
Classes that were particularly difficult to predict, with the lowest accuracies reported, were genetic bottlenecks (which were almost always misclassified as initial introductions) and the eco-evolutionary drivers that result in an isolate's failure to establish successfully. It was harder to predict initial introductions compared to predicting established populations using the categorical "stages of expansion" metric, but when this was reduced to a binary problem, accuracy increased by 0.2, suggesting that separating the stages into initial and secondary introductions (from an established population to a new area) hindered the prediction process.

Exploring the spatiotemporal presence of the errors identified when testing our eco evolutionary models on 90 unseen data points (Table S4 in Multimedia Appendix 1) revealed that the raw highest frequency of errors was found in Asia compared to other continents; however, the relative error rate considering the number of Asian isolates (n=50, 56% of the 90 data points) was the lowest across continents. Notable successes include a strong ability to predict population structure in Asia (in which all 50 samples were accurately predicted) and a low error rate (5.8%) when predicting successful expansions into the United States. Success was more difficult to predict in geographic locations with little representation in the test dataset; or example, there was 1 isolate each from Canada, Japan, Mexico, and Singapore in the test dataset, and only the success of the Canadian isolate was successfully predicted. Temporally, a greater number of errors occurred earlier in the time series, during the initial expansion period of VpST3.

Eco-Evolutionary Feature Importance

In general, ecological metrics performed stronger than evolutionary metrics individually (Figure S2 in Multimedia Appendix 1). Some of the notably important eco-evolutionary drivers included 3 accessory genes, which were almost always present in introduced isolates (and subsequently eroded) and which provided salt stress tolerance, survival advantages, and biofilm formation for motility, as well as summer maximum sea temperatures from both the year of isolate discovery and the year prior. Of the 109 total features used for training and prediction, the 10 (9.2%) strongest predictive features for each metric, based on feature importance, were collated into a data frame (Figure 4). A small range of these ecological and evolutionary metrics featured within the 10 most important features across all 5 expansion dynamics. For 1 expansion dynamic only—temporal evolution—the strongest predictive features were all environmental features, suggesting that the influence of environmental temporal trends outweighed that of the evolutionary drivers. The total number of genes was an important feature for 4 (80%) of the 5 predicted expansion dynamics and located in the top 3 most important features for each of these, suggesting that genetic diversity was a key distinguishing factor between the classes. In addition, the maximum temperatures during June, July, and August were strong predictor variables, appearing in the top 10 features of all models. Lagged sea temperature effects also offered significant information, notably the SSTs during June, July, and August from the previous year. Although salinity variables did not often appear in the top 10 features, the average salinity during September, October, and November was a useful predictor for classifying the stages of expansion and assessing the success chances of an isolate. Shellfish imports featured as important predictors in classifying population waves and the stages of expansion. Accessory gene presences were stronger predictors in classifying population waves and the chance of success than the stages of expansion themselves.

Figure 4. Feature importance for each of the expansion dynamic predictions within each random forest classifier model (where 1 indicates the most important and 10 the 10th most important). DJF: December-January-February; HTH: helix-turn-helix; JJA: June-July-August; MAM: March-April-May; SON: September-October-November; SST: sea surface temperature.



Feature	Populations	Temporal evolution	Stages	Binary stages	Success
Total number of genes	3		1	2	1
HTH-type transcriptional regulator *(puuR)*	2				6
Sodium:proton antiporter	4				5
Lactoylglutathione lyase	1			8	3
N-carbamoylputrescine amidase	5				
Agmatine deiminase	6				
RNA polymerase sigma factor *(RpoS)*					8
shellfish_imports_annual_tonnes	7		5	10	
SST_DJF_max		7			
SST_DJF_max_1yearlag		6	9		
SST_DJF_average		9			
SST_MAM_max		10			
SST_JJA_max	8	3	2	3	2
SST_JJA_max_1yearlag		1	7	1	4
SST_JJA_average	9	8		7	9
SST_JJA_average_1yearlag		5		6	10
SST_SON_max		2	3	4	
SST_SON_max1yearlag		4	6	5	
SST_SON_average_1yearlag	10		8		
SST_SON_min			4	9	
salinity_SON_average			10		7

In terms of notable relationships identified, the success of an isolate was generally associated with higher average and maximum SSTs, particularly during June, July, and August. The presence of certain accessory genes, including *puuR*, *aguB*, and *aguC*, was more important in the classification of "introduced" isolates than in the classification of "established" isolates. The isolates that were predicted to be "introduced" (as opposed to "established") almost always had these genes present compared to greater variation in the isolates that were predicted to be "established." This was even more evident in predicting the population to which an isolate belonged, where multiple accessory genes were absent in the third and most recent population wave. Shellfish imports emerged as an important driver in the distinction of the 3 separate populations, with a higher prediction range seen for isolates belonging to the third population wave.

Cross-Correlation Between Ecological and Evolutionary Drivers

We explored the relationships between the ecological and evolutionary features included in the model and found multiple significant correlations (Figure S3 in Multimedia Appendix 1). Notably, the selected accessory genes exhibited strong correlations with the SST metrics (both positive and negative) as well as with shellfish imports; some genes also correlated with the salinity metrics. In addition, some adaptive SNPs exhibited correlations; for example, the SNP at position 597 had slight negative associations with maximum SSTs and slight positive associations with minimum salinities. The total number of genes had slight positive associations with most of the SST metrics.

Region-Specific Eco-Evolutionary Models

When generating region-specific models to identify which isolates would be specifically successful in China or Peru, as representative countries, we found the model predictions for China to be largely more accurate, with accuracies ranging from 0.778 to 0.917, compared to model predictions for Peru, with accuracies ranging from 0.529 to 0.706 (Table 3). However, while the model was able to successfully classify successful isolates in China, it had difficulty in classifying the unsuccessful isolates, with poor specificity. The Peru model had more balanced predictions between these 2 classes. In both cases, the ecological features–only model was the best approach, providing the best accuracy. Total gene diversity was the top feature for the combined eco-evolutionary approach (and the evolutionary features–only model). For Peru, the remainder of the top 10 important features were ecological features; however, for China it was an even split between ecological and evolutionary drivers, including the type IV pilus twitching motility protein and the SNP at position 603 (Table S3 in Multimedia Appendix 1), which had not appeared previously among the important features. In the ecological features–only model, the top features were December to February minimum sea temperatures and June to August average temperatures a year prior for China and Peru, respectively.

Table 3. Accuracy metrics of region-specific random forest classifier models predicting unseen test data of each expansion dynamic.

Expansion dynamics	Combined eco-evolutionary approach				Evolutionary features only				Ecological features only			
	Precision	Recall	F_1-score	Accuracy	Precision	Recall	F_1-score	Accuracy	Precision	Recall	F_1-score	Accuracy
Success in China				0.889				0.778				0.917
Unsuccessful	0.500	0.250	0.330		0.167	0.250	0.200		1.000	0.250	0.400	
Successful	0.912	0.969	0.939		0.900	0.844	0.871		0.914	1.000	0.955	
Unweighted average	0.706	0.609	0.636		0.533	0.547	0.535		0.957	0.625	0.678	
Weighted average	0.866	0.889	0.872		0.819	0.778	0.796		0.924	0.917	0.894	
Success in Peru				0.529				0.647				0.706
Unsuccessful	0.500	0.375	0.429		0.750	0.375	0.500		0.714	0.625	0.667	
Successful	0.545	0.667	0.600		0.615	0.889	0.727		0.700	0.778	0.737	
Unweighted average	0.523	0.521	0.514		0.683	0.632	0.614		0.707	0.701	0.702	
Weighted average	0.524	0.529	0.519		0.679	0.647	0.620		0.707	0.706	0.704	

Discussion

Principal Findings

Our analysis suggests that VpST3, as a clonal complex, exhibited a high degree of efficacy in propagation during its expansion, evidenced by the numerous introductions in geographically distinct places at similar times. We found evolutionary features that provided mechanisms for this process, including accessory genes linked to functions that facilitate motility and biofilm formation for attachment-based transport mechanisms. The total number of genes within an isolate was an important predictor in the machine learning models for most expansion dynamics. Although we found no trend in gene numbers over time, the model associated higher gene numbers with isolates classified as within established populations, evidenced by a higher prediction range for established isolates. This suggests that isolates that became established could have acquired genes specific to survival in the local conditions, with this plasticity allowing it to colonize new geographic regions. The declining presence of certain accessory genes (*puuR*, *aguB*, and *aguC*) under purifying selection signals suggests that the genes involved in initial introduction may become less useful for population establishment, resulting in reduced selection pressure for these genes. This is corroborated by the prediction ranges of our model for "introduced" isolates, in which these genes were both important features and characterized as almost always present in introduced isolates.

Assessment of the Eco-Evolutionary Approach

Our analysis has confirmed the hypothesis that considering ecological and evolutionary features in a combined approach to explore the drivers of pathogen expansion yields higher accuracy than dealing with these drivers individually. This is a novel use of the framework described in the study by Campbell et al [17] for characterizing *V parahaemolyticus* expansion dynamics.

From the ecological perspective, SST was a strong predictor variable, as expected from well-established interactions between *V parahaemolyticus* and SST [69]; however, maximum temperatures during June, July, and August emerged as the strongest driver, alongside lagged effects from the previous year. More than two-thirds of our genomic isolates (240/311, 77.2%) were isolated in the northern hemisphere, where these months would be the warmest; this period has previously been described as the “*Vibrio* season” [69]—the characteristics of this season each year seem to drive expansion. In addition, the importance of SSTs in September, October, and November as well as in March, April, and May is pertinent to recent studies that have found expansions in seasonal suitability into cooler months, approximately a 1-month increase every 30 years [70]. Although the period from June to August is the coldest in the southern hemisphere for the South American isolates, it could still drive expansion dynamics when the maximum sea temperatures exceed the minimum for *V parahaemolyticus* survival in the environment, allowing the bacteria to persist in their environmental reservoirs until optimum conditions resume, a phenomenon known as overwintering [71]. Sea temperatures can drive both survival and community composition changes [22,72], with mostly positive associations between SST and the successful established isolates in our analyses. In laboratory studies, increases in seawater temperature have been found to upregulate the expression of virulence factors involved in adhesion processes, such as biofilm formation [73], which could facilitate transport mechanisms via attachment to marine organisms that aid expansion and settlement in new areas.

Generally, the models using evolutionary features only had a lower predictive potential; however, the inclusion of evolutionary features improved the ecological models when combined. The evolutionary features themselves potentially did not offer enough predictive information independently, but when linked to the specific local environmental conditions in which the evolutionary processes provide survival benefits, the evolutionary features were able to provide useful information within the model on pathogen expansion. The evolutionary features might lack meaning outside of ecological contexts or indeed play a different biological role in different ecological contexts. This is supported by the cross-correlations identified between several evolutionary features and the associated environmental conditions (Figure S3 in Multimedia Appendix 1), indicating that these interacting factors themselves, in the form of dynamic evolutionary responses to environmental conditions, can provide predictors of pathogen expansion. This justifies the inclusion of both ecological and evolutionary features in the same predictive model to account for the interactions between them. We observed a specific eco-evolutionary mechanism in our analysis, where SSTs were significantly associated with the presence of multiple accessory genes (Figure S3 in Multimedia Appendix 1), which could indicate an introduced selection pressure in the environment, with changes in SST representing a myriad of implications for the microbial community. However, it is important to note that these cross-correlations provide limited information and could also be purely reflecting the strength of the temporal trends of accessory gene presence, as the result of 2 concurrent or diverging trends, with sea temperature gradually increasing over the time period and accessory gene presence either increasing or decreasing steadily.

Shellfish imports were an important driver for the classification of the third population wave, which could allude to a population opportunistically taking advantage of shellfish movements as a transport mechanism. This would explain why this population has purged multiple accessory genes offering transport mechanisms, such as biofilm pathways. While the role of live aquatic animal transport in contributing to *V parahaemolyticus* expansion is currently unclear, studies have found that this method of transport introduces new populations, facilitates the exchange of genetic material, and promotes adaptation [74]. Further analysis will need to explore whether this subpopulation has undergone innovation to improve host-pathogen attachment mechanisms, particularly involving shellfish.

Few of the SNP mutations identified during outlier detection featured heavily in model decisions, despite our methodology aiming to identify mutations affecting proteins that could promote expansion dynamics. While we encoded the SNPs as categorical features in our machine learning analysis, alternative encoding techniques, such as one-hot encoding, have been explored, and it was found that including information on not only the mutation but also the position of mutation can improve accuracy [75]. Further analysis or different approaches should be explored to improve the identification of mutations critical to expansion processes.

While the models were designed generically to predict a range of expansion metrics, they could be further refined for specific purposes. There were several instances of a large discrepancy between recall and precision, particularly for smaller, underrepresented classes such as bottlenecks, which is a common issue in machine learning when dealing with imbalanced datasets. The models here were not developed individually to obtain the greatest accuracy, as the aim was to facilitate the comparison of accuracy metrics when combining ecological and evolutionary features. However, these imbalances can be remedied on a per-model basis in the future using techniques such as class weights to assign higher weights to minority classes during training or through oversampling (of the minority classes) and undersampling (of the majority classes), as demonstrated by DeLuca et al [76]. The difficulties in separating initial introductions and bottlenecks can be simplified into understanding why a particular introduction is successful or unsuccessful. We did find a potential limiting factor when predicting this success as a separate expansion metric, resulting in a high proportion of false positives where unsuccessful isolates were misclassified as successful.

We propose that a potential limiting factor here could be plankton presence, which has been found to offer nutrients for growth and host protection [77], which was not included in the analysis. This is relevant given the biofilm-related accessory genes identified, which facilitate attachment to plankton, in which these eco-evolutionary factors could combine to provide further information on isolate success. Similarly, plankton abundance was found to significantly increase the presence of 2 major virulence factors of *V parahaemolyticus*, *tdh* and *trh* [78], underlining another eco-evolutionary mechanism driving *V parahaemolyticus* dynamics. There are difficulties in quantifying marine plankton presence for such a global collection spanning decades. Earth Observation data offers a suitable source for ecological driver data in the future, providing consistent time series data at a sufficient resolution; however, satellite observations of plankton (using chlorophyll-a concentration as a proxy) are only available from late 1997; the key preceding year that represents the pivotal early introductions of the expansion of VpST3 is missing.

The spatiotemporal trends of error counts discussed (Table S4 in Multimedia Appendix 1) offer insights into model limitations and areas for future improvement, such as improving our predictive capabilities during the initial emergence of a pathogen strain and in geographic regions reporting few isolates (as is common during initial expansion).

Regional Predictive Performance for a Globally Expanding Pathogen

The difference in accuracy between the China and Peru regional models is likely due to the consequences of class imbalances. The Chinese isolates had a much higher proportion of successful isolates (108/120, 90%) than Peru (27/55, 49%), which meant that, although we were able to predict successful isolates with high precision and recall, it was very difficult to predict the minority class of unsuccessful isolates (F_1-score=0.33). Such class imbalances result in overfitting of the majority class, enabling the model to achieve a high accuracy of 90% even if it simply predicted all isolates to be successful. This can be seen in the China model using only ecological features, in which the majority class (successful isolates) was predicted perfectly due to 97% (35/36) of the data points being predicted as successful. Further evidence for overfitting is provided by a large difference between the area under the receiver operating characteristic curve values of the training and test data, which were 0.860 and 0.949, respectively. To overcome such overfitting during the future development of regional models, per-class and alternative accuracy metrics need to be considered and imbalances addressed through methods previously outlined. In the Peru model, the number of successful and unsuccessful isolates were much more balanced, resulting in lower but more balanced per-class accuracy metrics. Currently, this would suggest that we can predict the success of a pathogenic variant isolate more accurately in an endemic region than in an emerging one but at the expense of possible overfitting, providing areas for improvement. In both cases, we found that ecological drivers alone were the best approach, suggesting that the evolutionary features were introducing noise into the model. This suggests that focusing on common features in the whole group that might facilitate expansion on a global scale might not be as valuable as more region-specific evolutionary drivers, such as those representative of adaptation to local conditions of a particular region, which would need to be extracted for a more successful regional approach.

However, it is important to note that while models can be improved specifically for particular geographic regions, for example, based on the ranges of local environmental conditions, this comes at the expense of declining applicability. Such applicability could be seen as a priority for a globally expanding pathogen such as VpST3, requiring a model that is able to function in a range of distinct geographic regions. Future work could mediate this trade-off through the introduction of regional encoders as features [24] or through engineering environmental features to be more comparable, such as through normalized anomalies rather than raw values.

Future Predictive Potential

While this analysis focused on the 3 continents reporting the most VpST3 isolates (Asia, North America, and South America), in the future, the focus will need to shift to countries that lie on the periphery of the environmental tolerance ranges of *V parahaemolyticus*, representing the potential locations of future expansion. These include Europe, which, in recent years, has observed the emergence of *Vibrio* lineages and increases in vibriosis incidence as an emerging public health issue [79]. Increased genomic surveillance is required in these countries to test the ability of this framework to identify expansion potential into these new regions.

In addition, the eco-evolutionary analysis was limited by the annual resolution of the genomic isolate metadata and shellfish movement data. The majority of the isolates in our collection were submitted to public databases with limited metadata, specifying only a country and a year; however, higher-resolution metadata, such as a district and a day, week, or month, as suggested by Campbell et al [17], would greatly improve the specificity of the related ecological data that we could then append to this isolate, which is available at a very high resolution. This is particularly necessary to account for the rapid evolutionary timescales on which bacteria such as *Vibrio* function [29]. Future models would benefit from higher spatiotemporal–resolution datasets for machine learning training that facilitate the characterization of more specific eco-evolutionary drivers and increase predictive accuracy.

Conclusions

This pilot study provides a precedent for combining ecological and evolutionary driver data using machine learning to predict pathogen expansion metrics. This both aids our understanding of historic expansion and, through further refinement and development, could be operationalized into a trained database through which a new recovered isolate could be submitted and predictions made as to its introduction or establishment potential to track pathogen expansion in near real time. The current limitations preventing such operationalization include sufficient genomic surveillance, data accessibility, and interdisciplinary analysis requirements. Accuracy would need to be refined to the appropriate confidence values based on user requirements of model sensitivity. Further exploration of applicability to a range of climate-sensitive pathogens will require sufficient genomic surveillance, which is currently limited by poor spatiotemporal resolution. Combining state-of-the art analyses of both ecological and evolutionary pathogen drivers will provide new insights into future eco-evolutionary pathways of climate-sensitive pathogens.

Acknowledgments

AMC was supported by the Natural Environment Research Council (NE/S007210/1) and Centre for Environment, Fisheries and Aquaculture Science internal Seedcorn funding. JM-U was funded by grants from the Ministerio de Ciencia e Innovación (Spain; PID2021-127107NB-I00) and the Generalitat de Catalunya (Spain; 2021 SGR 00526). This work is funded within the framework of the Horizon Europe program through the project IDAlert.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables and figures.
[DOCX File , 536 KB-Multimedia Appendix 1]

References

1. Martinez-Urtaza J, Bowers JC, Trinanes J, DePaola A. Climate anomalies and the increasing risk of Vibrio parahaemolyticus and Vibrio vulnificus illnesses. Food Res Int. Aug 2010;43(7):1780-1790. [doi: 10.1016/j.foodres.2010.04.001]
2. Abanto M, Gavilan RG, Baker-Austin C, Gonzalez-Escalona N, Martinez-Urtaza J. Global expansion of Pacific Northwest Vibrio parahaemolyticus sequence type 36. Emerg Infect Dis. Feb 2020;26(2):323-326. [FREE Full text] [doi: 10.3201/eid2602.190362] [Medline: 31961301]
3. Baker-Austin C, Oliver JD, Alam M, Ali A, Waldor MK, Qadri F, et al. Vibrio spp. infections. Nat Rev Dis Primers. Jul 12, 2018;4(1):8-19. [doi: 10.1038/s41572-018-0005-8] [Medline: 30002421]
4. Nair GB, Ramamurthy T, Bhattacharya SK, Dutta B, Takeda Y, Sack DA. Global dissemination of Vibrio parahaemolyticus serotype O3:K6 and its serovariants. Clin Microbiol Rev. Jan 2007;20(1):39-48. [FREE Full text] [doi: 10.1128/CMR.00025-06] [Medline: 17223622]
5. Arakawa E, Murase T, Shimada T, Okitsu T, Yamai S, Watanabe H. Emergence and prevalence of a novel Vibrio parahaemolyticus O3:K6 clone in Japan. Jpn J Infect Dis. Dec 1999;52(6):246-247. [Medline: 10738364]
6. Gil AI, Miranda H, Lanata CF, Prada A, Hall ER, Barreno CM, et al. O3:K6 serotype of Vibrio parahaemolyticus identical to the global pandemic clone associated with diarrhea in Peru. Int J Infect Dis. Jul 2007;11(4):324-328. [FREE Full text] [doi: 10.1016/j.ijid.2006.08.003] [Medline: 17321179]
7. González-Escalona N, Cachicas V, Acevedo C, Rioseco ML, Vergara JA, Cabello F, et al. Vibrio parahaemolyticus diarrhea, Chile, 1998 and 2004. Emerg Infect Dis. Jan 2005;11(1):129-131. [FREE Full text] [doi: 10.3201/eid1101.040762] [Medline: 15705337]
8. Centers for Disease Control and Prevention (CDC). Outbreak of Vibrio parahaemolyticus infection associated with eating raw oysters and clams harvested from Long Island Sound--Connecticut, New Jersey, and New York, 1998. MMWR Morb Mortal Wkly Rep. Jan 29, 1999;48(3):48-51. [FREE Full text] [Medline: 9935142]
9. Okuda J, Ishibashi M, Hayakawa E, Nishino T, Takeda Y, Mukhopadhyay AK, et al. Emergence of a unique O3:K6 clone of Vibrio parahaemolyticus in Calcutta, India, and isolation of strains from the same clonal group from Southeast Asian travelers arriving in Japan. J Clin Microbiol. Dec 1997;35(12):3150-3155. [FREE Full text] [doi: 10.1128/jcm.35.12.3150-3155.1997] [Medline: 9399511]
10. Centers for Disease Control and Prevention (CDC). Outbreak of Vibrio parahaemolyticus infections associated with eating raw oysters--Pacific Northwest, 1997. MMWR Morb Mortal Wkly Rep. Jun 12, 1998;47(22):457-462. [FREE Full text] [Medline: 9639367]
11. Bag PK, Nandi S, Bhadra RK, Ramamurthy T, Bhattacharya SK, Nishibuchi M, et al. Clonal diversity among recently emerged strains of Vibrio parahaemolyticus O3:K6 associated with pandemic spread. J Clin Microbiol. Jul 1999;37(7):2354-2357. [FREE Full text] [doi: 10.1128/JCM.37.7.2354-2357.1999] [Medline: 10364615]
12. -. Vibrio parahaemolyticus, Japan, 1996-1998. Wkly Epidemiol Rec. Oct 29, 1999;74(43):361-363. [Medline: 10887775]
13. Chiou CS, Hsu SY, Chiu SI, Wang TK, Chao CS. Vibrio parahaemolyticus serovar O3:K6 as cause of unusually high incidence of food-borne disease outbreaks in Taiwan from 1996 to 1999. J Clin Microbiol. Dec 2000;38(12):4621-4625. [FREE Full text] [doi: 10.1128/JCM.38.12.4621-4625.2000] [Medline: 11101606]

14. Smolikova L, Lomov IM, Khomenko TV, Murnachev GP, Kudriakova TA, Fetsaĭlova OP, et al. [Studies on halophilic vibrios causing a food poisoning outbreak in the city of Vladivostok]. Zh Mikrobiol Epidemiol Immunobiol. 2001;(6):3-7. [Medline: 11881491]
15. Ansaruzzaman M, Lucas M, Deen JL, Bhuiyan NA, Wang XY, Safa A, et al. Pandemic serovars (O3:K6 and O4:K68) of Vibrio parahaemolyticus associated with diarrhea in Mozambique: spread of the pandemic into the African continent. J Clin Microbiol. Jun 2005;43(6):2559-2562. [FREE Full text] [doi: 10.1128/JCM.43.6.2559-2562.2005] [Medline: 15956363]
16. Smith JM, Feil EJ, Smith NH. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays. Dec 2000;22(12):1115-1122. [doi: 10.1002/1521-1878(200012)22:12<1115::AID-BIES9>3.0.CO;2-R] [Medline: 11084627]
17. Campbell AM, Hauton C, Baker-Austin C, van Aerle R, Martinez-Urtaza J. An integrated eco-evolutionary framework to predict population-level responses of climate-sensitive pathogens. Curr Opin Biotechnol. Apr 2023;80:102898. [FREE Full text] [doi: 10.1016/j.copbio.2023.102898] [Medline: 36739640]
18. Deeb R, Tufford D, Scott GI, Moore JG, Dow K. Impact of climate change on Vibrio vulnificus abundance and exposure risk. Estuaries Coast. Dec 7, 2018;41(8):2289-2303. [FREE Full text] [doi: 10.1007/s12237-018-0424-5] [Medline: 31263385]
19. Galanis E, Otterstatter M, Taylor M. Measuring the impact of sea surface temperature on the human incidence of Vibrio sp. infection in British Columbia, Canada, 1992-2017. Environ Health. May 27, 2020;19(1):58. [FREE Full text] [doi: 10.1186/s12940-020-00605-x] [Medline: 32460848]
20. Hartwick MA, Urquhart EA, Whistler CA, Cooper VS, Naumova EN, Jones SH. Forecasting seasonal Vibrio parahaemolyticus concentrations in New England shellfish. Int J Environ Res Public Health. Nov 07, 2019;16(22):4341. [FREE Full text] [doi: 10.3390/ijerph16224341] [Medline: 31703312]
21. Kopprio GA, Neogi SB, Rashid H, Alonso C, Yamasaki S, Koch BP, et al. Vibrio and bacterial communities across a pollution gradient in the Bay of Bengal: unraveling their biogeochemical drivers. Front Microbiol. Apr 15, 2020;11:594. [FREE Full text] [doi: 10.3389/fmicb.2020.00594] [Medline: 32351470]
22. Siboni N, Balaraju V, Carney R, Labbate M, Seymour JR. Spatiotemporal dynamics of Vibrio spp. within the Sydney Harbour Estuary. Front Microbiol. Apr 12, 2016;7:460. [FREE Full text] [doi: 10.3389/fmicb.2016.00460] [Medline: 27148171]
23. Xu M, Cao C, Wang D, Kan B. Identifying environmental risk factors of cholera in a coastal area with geospatial technologies. Int J Environ Res Public Health. Dec 29, 2014;12(1):354-370. [FREE Full text] [doi: 10.3390/ijerph120100354] [Medline: 25551518]
24. Campbell AM, Racault MF, Goult S, Laurenson A. Cholera risk: a machine learning approach applied to essential climate variables. Int J Environ Res Public Health. Dec 15, 2020;17(24):9378. [FREE Full text] [doi: 10.3390/ijerph17249378] [Medline: 33333823]
25. Froelich B, Gonzalez R, Blackwood D, Lauer K, Noble R. Decadal monitoring reveals an increase in Vibrio spp. concentrations in the Neuse River Estuary, North Carolina, USA. PLoS One. Apr 23, 2019;14(4):e0215254. [FREE Full text] [doi: 10.1371/journal.pone.0215254] [Medline: 31013284]
26. Martinez-Urtaza J, Lozano-Leon A, Varela-Pet J, Trinanes J, Pazos Y, Garcia-Martin O. Environmental determinants of the occurrence and distribution of Vibrio parahaemolyticus in the rias of Galicia, Spain. Appl Environ Microbiol. Jan 2008;74(1):265-274. [FREE Full text] [doi: 10.1128/AEM.01307-07] [Medline: 17981951]
27. Daniels NA, MacKinnon L, Bishop R, Altekruse S, Ray B, Hammond RM, et al. Vibrio parahaemolyticus infections in the United States, 1973-1998. J Infect Dis. May 2000;181(5):1661-1666. [doi: 10.1086/315459] [Medline: 10823766]
28. McLaughlin JB, DePaola A, Bopp CA, Martinek KA, Napolilli NP, Allison CG, et al. Outbreak of Vibrio parahaemolyticus gastroenteritis associated with Alaskan oysters. N Engl J Med. Oct 06, 2005;353(14):1463-1470. [doi: 10.1056/NEJMoa051594] [Medline: 16207848]
29. Oberbeckmann S, Fuchs BM, Meiners M, Wichels A, Wiltshire KH, Gerdts G. Seasonal dynamics and modeling of a Vibrio community in coastal waters of the North Sea. Microb Ecol. Apr 28, 2012;63(3):543-551. [doi: 10.1007/s00248-011-9990-9] [Medline: 22202887]
30. Banerjee SK, Rutley R, Bussey J. Diversity and dynamics of the Canadian coastal Vibrio community: an emerging trend detected in the temperate regions. J Bacteriol. Aug 01, 2018;200(15):e00787-e00717. [FREE Full text] [doi: 10.1128/JB.00787-17] [Medline: 29735763]
31. Montánchez I, Ogayar E, Plágaro AH, Esteve-Codina A, Gómez-Garrido J, Orruño M, et al. Analysis of Vibrio harveyi adaptation in sea water microcosms at elevated temperature provides insights into the putative mechanisms of its persistence and spread in the time of global warming. Sci Rep. Jan 22, 2019;9(1):289. [FREE Full text] [doi: 10.1038/s41598-018-36483-0] [Medline: 30670759]
32. Thompson CL, Alberti M, Barve S, Battistuzzi FU, Drake JL, Goncalves GC, et al. Back to the future: reintegrating biology to understand how past eco-evolutionary change can predict future outcomes. Integr Comp Biol. Feb 05, 2022;61(6):2218-2232. [doi: 10.1093/icb/icab068] [Medline: 33964141]
33. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. Dec 10, 2010;11(1):595. [FREE Full text] [doi: 10.1186/1471-2105-11-595] [Medline: 21143983]
34. Petit RA, Read TD. Bactopia: a flexible pipeline for complete analysis of bacterial genomes. mSystems. Aug 25, 2020;5(4). [doi: 10.1128/msystems.00190-20]

35. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. Nov 19, 2014;15(11):524. [FREE Full text] [doi: 10.1186/s13059-014-0524-x] [Medline: 25410596]
36. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. Feb 18, 2015;43(3):e15. [FREE Full text] [doi: 10.1093/nar/gku1196] [Medline: 25414349]
37. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. Jan 10, 2016;2(1):vew007. [FREE Full text] [doi: 10.1093/ve/vew007] [Medline: 27774300]
38. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. Apr 8, 2019;15(4):e1006650. [FREE Full text] [doi: 10.1371/journal.pcbi.1006650] [Medline: 30958812]
39. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. Aug 2012;29(8):1969-1973. [FREE Full text] [doi: 10.1093/molbev/mss075] [Medline: 22367748]
40. Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. Efficient Bayesian inference under the structured coalescent. Bioinformatics. Aug 15, 2014;30(16):2272-2279. [FREE Full text] [doi: 10.1093/bioinformatics/btu201] [Medline: 24753484]
41. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst Biol. Sep 01, 2018;67(5):901-904. [FREE Full text] [doi: 10.1093/sysbio/syy032] [Medline: 29718447]
42. Volz EM, Carsten W, Grad YH, Frost SD, Dennis AM, Didelot X. Identification of hidden population structure in time-scaled phylogenies. Syst Biol. Sep 01, 2020;69(5):884-896. [FREE Full text] [doi: 10.1093/sysbio/syaa009] [Medline: 32049340]
43. Conrad RE, Viver T, Gago JF, Hatt JK, Venter SN, Rossello-Mora R, et al. Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. ISME J. May 2022;16(5):1222-1234. [FREE Full text] [doi: 10.1038/s41396-021-01149-9] [Medline: 34887548]
44. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. Nov 15, 2015;31(22):3691-3693. [FREE Full text] [doi: 10.1093/bioinformatics/btv421] [Medline: 26198102]
45. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. Nov 25, 2016;17(1):238. [FREE Full text] [doi: 10.1186/s13059-016-1108-8] [Medline: 27887642]
46. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. May 2005;22(5):1208-1222. [doi: 10.1093/molbev/msi105] [Medline: 15703242]
47. Luu K, Bazin E, Blum MG. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour. Jan 2017;17(1):67-77. [doi: 10.1111/1755-0998.12592] [Medline: 27601374]
48. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). Oct 27, 2012;6(2):80-92. [FREE Full text] [doi: 10.4161/fly.19695] [Medline: 22728672]
49. Namadi P, Deng Z. Optimum environmental conditions controlling prevalence of Vibrio parahaemolyticus in marine environment. Mar Environ Res. Jan 2023;183:105828. [doi: 10.1016/j.marenvres.2022.105828] [Medline: 36423461]
50. Soci C, Hersbach H, Simmons A, Poli P, Bell B, Berrisford P, et al. The ERA5 global reanalysis from 1940 to 2022. Q J R Meteorol Soc. Jul 31, 2024. [doi: 10.1002/qj.4803]
51. Good SA, Martin MJ, Rayner NA. EN4: quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. JGR Oceans. Dec 11, 2013;118(12):6704-6716. [doi: 10.1002/2013jc009067]
52. Gouretski V, Cheng L. Correction for systematic errors in the global dataset of temperature profiles from mechanical bathythermographs. J Atmos Ocean Technol. 2020;37(5):841-855. [doi: 10.1175/JTECH-D-19-0205.1]
53. Sarkar BL, Nair GB, Banerjee AK, Pal SC. Seasonal distribution of Vibrio parahaemolyticus in freshwater environs and in association with freshwater fishes in Calcutta. Appl Environ Microbiol. Jan 1985;49(1):132-136. [FREE Full text] [doi: 10.1128/aem.49.1.132-136.1985] [Medline: 3977307]
54. Morris JGJ, Wilson R, Davis BR, Wachsmuth IK, Riddle CF, Wathen HG, et al. Non-O group 1 Vibrio cholerae gastroenteritis in the United States: clinical, epidemiologic, and laboratory characteristics of sporadic cases. Ann Intern Med. May 1981;94(5):656-658. [doi: 10.7326/0003-4819-94-5-656] [Medline: 7235397]
55. Pal SC, Sircar BK, Nair GB, Deb BC. Epidemiology of bacterial diarrhoeal diseases in India with special reference to Vibrio parahaemolyticus infections. In: Takeda Y, Miwatani T, editors. Bacterial Diarrheal Diseases. Dordrecht, The Netherlands. Springer; 1985:65-73.
56. FishStatJ, F. A. O. A tool for fishery statistics analysis. FAO Fisheries and Aquaculture Department, FIPS–Statistics and Information. 2017. URL: https://scholar.google.com/scholar_lookup?&title=FishStatJ%2C%20a%20tool%20for%20fishery%20statistics%20analysis&publication_year=2024 [accessed 2024-11-12]
57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12(85):2825-2830. [FREE Full text]

58. Waskom ML. seaborn: statistical data visualization. J Open Source Softw. Apr 2021;6(60):3021. [FREE Full text] [doi: 10.21105/joss.03021]
59. Chakraborty S, Gogoi M, Chakravortty D. Lactoylglutathione lyase, a critical enzyme in methylglyoxal detoxification, contributes to survival of Salmonella in the nutrient rich environment. Virulence. Dec 17, 2015;6(1):50-65. [FREE Full text] [doi: 10.4161/21505594.2014.983791] [Medline: 25517857]
60. Nemoto N, Kurihara S, Kitahara Y, Asada K, Kato K, Suzuki H. Mechanism for regulation of the putrescine utilization pathway by the transcription factor PuuR in Escherichia coli K-12. J Bacteriol. Jul 2012;194(13):3437-3447. [FREE Full text] [doi: 10.1128/JB.00097-12] [Medline: 22522900]
61. Merz AJ, So M, Sheetz MP. Pilus retraction powers bacterial twitching motility. Nature. Sep 07, 2000;407(6800):98-102. [doi: 10.1038/35024105] [Medline: 10993081]
62. Akram U, Song Y, Liang C, Abid MA, Askari M, Myat AA, et al. Genome-wide characterization and expression analysis of NHX gene family under salinity stress in Gossypium barbadense and its comparison with Gossypium hirsutum. Genes (Basel). Jul 16, 2020;11(7):803. [FREE Full text] [doi: 10.3390/genes11070803] [Medline: 32708576]
63. Liu Z, Hossain SS, Morales Moreira Z, Haney CH. Putrescine and its metabolic precursor arginine promote biofilm and c-di-GMP synthesis in pseudomonas aeruginosa. J Bacteriol. Jan 18, 2022;204(1):e0029721. [FREE Full text] [doi: 10.1128/JB.00297-21] [Medline: 34723645]
64. del Rio B, Linares DM, Ladero V, Redruello B, Fernández M, Martin MC, et al. Putrescine production via the agmatine deiminase pathway increases the growth of Lactococcus lactis and causes the alkalinization of the culture medium. Appl Microbiol Biotechnol. Jan 24, 2015;99(2):897-905. [doi: 10.1007/s00253-014-6130-8] [Medline: 25341400]
65. Pis Diez CM, Juncos MJ, Villarruel Dujovne M, Capdevila DA. Bacterial transcriptional regulators: a road map for functional, structural, and biophysical characterization. Int J Mol Sci. Feb 16, 2022;23(4):2179. [FREE Full text] [doi: 10.3390/ijms23042179] [Medline: 35216300]
66. Li P, Liu C, Li B, Ma Q. Structural analysis of the CARB β-lactamase from Vibrio parahaemolyticus facilitates application of the β-lactam/β-lactamase inhibitor therapy. Biochimie. Apr 2020;171-172:213-222. [doi: 10.1016/j.biochi.2020.03.011] [Medline: 32179166]
67. Lee J, Shin E, Park J, Lee M, Lee K. Regulator of ribonuclease activity modulates the pathogenicity of Vibrio vulnificus. J Microbiol. Dec 09, 2021;59(12):1133-1141. [doi: 10.1007/s12275-021-1518-5] [Medline: 34751908]
68. Zhou G, Wang YS, Peng H, Huang XM, Xie XB, Shi QS. Role of Ttca of Citrobacter Werkmanii in bacterial growth, biocides resistance, biofilm formation and swimming motility. Int J Mol Sci. Sep 06, 2018;19(9):2644. [FREE Full text] [doi: 10.3390/ijms19092644] [Medline: 30200616]
69. Froelich BA, Daines DA. In hot water: effects of climate change on Vibrio-human interactions. Environ Microbiol. Oct 2020;22(10):4101-4111. [doi: 10.1111/1462-2920.14967] [Medline: 32114705]
70. Trinanes J, Martinez-Urtaza J. Future scenarios of risk of Vibrio infections in a warming planet: a global mapping study. Lancet Planet Health. Jul 2021;5(7):e426-e435. [FREE Full text] [doi: 10.1016/S2542-5196(21)00169-8] [Medline: 34245713]
71. Martinez-Urtaza J, van Aerle R, Abanto M, Haendiges J, Myers RA, Trinanes J, et al. Genomic variation and evolution of Vibrio parahaemolyticus ST36 over the course of a transcontinental epidemic expansion. mBio. Nov 14, 2017;8(6):e01425-17. [FREE Full text] [doi: 10.1128/mBio.01425-17] [Medline: 29138301]
72. Logar-Henderson C, Ling R, Tuite AR, Fisman DN. Effects of large-scale oceanic phenomena on non-cholera vibriosis incidence in the United States: implications for climate change. Epidemiol Infect. Jan 2019;147:e243. [FREE Full text] [doi: 10.1017/S0950268819001316] [Medline: 31364581]
73. Billaud M, Seneca F, Tambutté E, Czerucka D. An increase of seawater temperature upregulates the expression of Vibrio parahaemolyticus virulence factors implicated in adhesion and biofilm formation. Front Microbiol. Mar 8, 2022;13:840628. [FREE Full text] [doi: 10.3389/fmicb.2022.840628] [Medline: 35350627]
74. Fu S, Wang Q, Zhang Y, Yang Q, Hao J, Liu Y, et al. Dynamics and microevolution of Vibrio parahaemolyticus populations in shellfish farms. mSystems. Jan 12, 2021;6(1):e01161-20. [FREE Full text] [doi: 10.1128/mSystems.01161-20] [Medline: 33436516]
75. Singh VK, Maurya NS, Mani A, Yadav RS. Machine learning method using position-specific mutation based classification outperforms one hot coding for disease severity prediction in haemophilia 'A'. Genomics. Nov 2020;112(6):5122-5128. [FREE Full text] [doi: 10.1016/j.ygeno.2020.09.020] [Medline: 32927010]
76. DeLuca NM, Zaitchik BF, Guikema SD, Jacobs JM, Davis BJ, Curriero FC. Evaluation of remotely sensed prediction and forecast models for Vibrio parahaemolyticus in the Chesapeake Bay. Remote Sens Environ. Dec 1, 2020;250:112016. [doi: 10.1016/j.rse.2020.112016]
77. Hartwick MA, Berenson A, Whistler CA, Naumova EN, Jones SH. The seasonal microbial ecology of plankton and plankton-associated Vibrio parahaemolyticus in the Northeast United States. Appl Environ Microbiol. Jul 13, 2021;87(15):e0297320. [FREE Full text] [doi: 10.1128/AEM.02973-20] [Medline: 33990304]
78. Turner JW, Malayil L, Guadagnoli D, Cole D, Lipp EK. Detection of Vibrio parahaemolyticus, Vibrio vulnificus and Vibrio cholerae with respect to seasonal fluctuations in temperature and plankton abundance. Environ Microbiol. Apr 2014;16(4):1019-1028. [doi: 10.1111/1462-2920.12246] [Medline: 24024909]

79. Amato E, Riess M, Thomas-Lopez D, Linkevicius M, Pitkänen T, Wołkowicz T, et al. Epidemiological and microbiological investigation of a large increase in vibriosis, northern Europe, 2018. Euro Surveill. Jul 2022;27(28):2101088. [FREE Full text] [doi: 10.2807/1560-7917.ES.2022.27.28.2101088] [Medline: 35837965]

Abbreviations

dN/dS: ratio of nonsynonymous to synonymous substitutions
GTR: general time reversible
SNP: single nucleotide polymorphism
SST: sea surface temperature
VpST3: Vibrio parahaemolyticus sequence type 3

Edited by S Hacking; submitted 30.05.24; peer-reviewed by M Jani, R Voicu; comments to author 18.08.24; revised version received 12.09.24; accepted 02.10.24; published 28.11.24

Please cite as:
Campbell AM, Hauton C, van Aerle R, Martinez-Urtaza J
Eco-Evolutionary Drivers of Vibrio parahaemolyticus Sequence Type 3 Expansion: Retrospective Machine Learning Approach
JMIR Bioinform Biotech 2024;5:e62747
URL: https://bioinform.jmir.org/2024/1/e62747
doi: 10.2196/62747
PMID: 39607996