Received: 23 May 2024 | Accepted: 27 August 2024

DOI: 10.1111/1462-2920.16700

BRIEF REPORT

ENVIRONMENTAL MICROBIOLOGY

Identification of putative coral pathogens in endangered Caribbean staghorn coral using machine learning

Jason D. Selwyn[1,2] | **Brecia A. Despard**[1,2] | **Miles V. Vollmer**[1,2] | **Emily C. Trytten**[1,2] | **Steven V. Vollmer**[1,2]

[1]Marine Science Center, Northeastern University, Nahant, Massachusetts, USA

[2]Department of Marine and Environmental Sciences, Northeastern University, Boston, Massachusetts, USA

Correspondence
Jason D. Selwyn and Steven V. Vollmer, Marine Science Center, Northeastern University, Nahant, MA, USA.
Email: j.selwyn@northeastern.edu and s.vollmer@northeastern.edu

Funding information
Division of Ocean Sciences, Grant/Award Numbers: OCE-1458158, OCE-1924145

Abstract

Coral diseases contribute to the rapid decline in coral reefs worldwide, and yet coral bacterial pathogens have proved difficult to identify because 16S rRNA gene surveys typically identify tens to hundreds of disease-associate bacteria as putative pathogens. An example is white band disease (WBD), which has killed up to 95% of the now-endangered Caribbean *Acropora* corals since 1979, yet the pathogen is still unknown. The 16S rRNA gene surveys have identified hundreds of WBD-associated bacterial amplicon sequencing variants (ASVs) from at least nine bacterial families with little consensus across studies. We conducted a multi-year, multi-site 16S rRNA gene sequencing comparison of 269 healthy and 143 WBD-infected *Acropora cervicornis* and used machine learning modelling to accurately predict disease outcomes and identify the top ASVs contributing to disease. Our ensemble ML models accurately predicted disease with greater than 97% accuracy and identified 19 disease-associated ASVs and five healthy-associated ASVs that were consistently differentially abundant across sampling periods. Using a tank-based transmission experiment, we tested whether the 19 disease-associated ASVs met the assumption of a pathogen and identified two pathogenic candidate ASVs—ASV25 *Cysteiniphilum litorale* and ASV8 *Vibrio* sp. to target for future isolation, cultivation, and confirmation of Henle-Koch's postulate via transmission assays.

INTRODUCTION

The global rise in coral disease epizootics associated with human-induced climate change has caused unprecedented coral declines (Bruno et al., 2007; Burge et al., 2014; Harvell et al., 1999), especially in the greater Caribbean where white band disease (WBD) has killed up to 95% of the now-endangered *Acropora* corals since 1979 (Aronson & Precht, 2001; Gladfelter, 1982), and stony coral tissue loss disease (SCTLD) is currently causing die-offs in more than 21 common coral species (Alvarez-Filip et al., 2022; Precht et al., 2016). Despite the devastating impacts of coral diseases, specific coral pathogens have been identified in only five of the 20 or more described coral diseases (Sutherland et al., 2004). Coral bacterial pathogens have proved difficult to identify because culture-independent, genetic analyses typically identify hundreds of disease-associated amplicon sequencing variants (ASVs)/operational taxonomic units (OTUs) as candidate pathogens (e.g., Gignoux-Wolfsohn & Vollmer, 2015) coupled with difficulties culturing these putative coral pathogens to fulfil Henle-Kochs postulate in controlled transmission experiments. Large numbers of disease-associated bacteria have led to the emerging view that many coral diseases are caused by a dysbiosis between the coral host, its symbiotic algae and its associated microbiome in stressed or compromised corals (Voolstra et al., 2024), even for host-specific coral diseases with clear transmission dynamics like

the Caribbean *Acropora* WBD host-disease system (Gignoux-Wolfsohn et al., 2012).

WBD is a highly transmissible, host-specific bacterial pathogen (Gignoux-Wolfsohn et al., 2017;Kline & Vollmer, 2011; Sweet et al., 2014) that infects the Caribbean staghorn coral *A. cervicornis* (Kline & Vollmer, 2011; Ritchie & Smith, 1998), its congener the elkhorn coral *A. palmata* (Gladfelter, 1982), and their hybrids *A. prolifera* (van Oppen et al., 2000; Vollmer & Palumbi, 2002). Transmission occurs via direct contact, snail vectors and through the water column through tissue lesions (Gignoux-Wolfsohn et al., 2012). Its transmission and progression can be arrested with broad-spectrum antibiotics (Kline & Vollmer, 2011; Sweet et al., 2014) and bacterial quorum-sensing inhibitors (Certner & Vollmer, 2015, 2018). Historically two forms of WBD have been described based on the gross appearance of the disease lesions with WBD type 1 originally described as having a sharp distal disease lesion containing algal symbionts (Gladfelter, 1982) and WBD type 2 differentiated by containing a section of bleached tissue at the margin of the tissue lesion (Ritchie & Smith, 1998). In reality, WBD lesions—as a disease sign—advance at different rates even within individual coral colonies (Miller et al., 2014; Vollmer, pers. obs.), which has caused some to use the term WBD to describe these advancing disease lesions (Aronson & Precht, 2001; Gignoux-Wolfsohn et al., 2017; Kline & Vollmer, 2011; Vollmer & Kline, 2008) and others to favour the more general term of rapid tissue loss (RTL; Miller et al., 2014; Williams & Miller, 2005) further complicating the matter (sensu Miller et al., 2014).

While Henle-Kochs postulate has not been fulfilled for WBD, early bacterial culturing identified a strong association of *Vibrio charcharia* (now synonymized with *V. harveyi*) on WBD-infected *Acropora cervicornis* (Ritchie & Smith, 1998) and in situ grafting of uncharacterized *Vibrio* cultures elicited WBD disease signs (Gil-Agudelo et al., 2006). Multiple genetic surveys have since identified hundreds of disease-associated ASVs/OTUs as potential WBD pathogens from at least nine bacterial families with little consensus across studies. Our prior field surveys and tank-based transmission assays in Panama using *A. cervicornis* identified ASVs/OTUs belonging to Vibrionaceae, Flavobacteriaceae (Certner & Vollmer, 2018; Gignoux-Wolfsohn & Vollmer, 2015), Campylobacteraceae, Francisellaceae and Pasteurellaceae (Gignoux-Wolfsohn et al., 2017) as likely WBD pathogens. In situ, transmission assays to *A. cervicornis* and *A. palmata* in Florida by Rosales et al. (2019) identified four ASVs from the families Vibrionaceae, Sphingomonadaceae, Rhodobacteraceae and Cryomorphaceae that were significantly associated with disease outcomes, including one ASV, *Sphingobium yanoikuyae* (family Sphingomonadacea), that was identified as the most likely WBD pathogen based on its high frequency on diseased corals. Parasitic infection by the alpha-proteobacterium 'Candidatus Aquarickettsia rohweri' has also been associated with increased WBD susceptibility in *A. cervicornis* (Casas et al., 2004; Klinges et al., 2020).

Most 16S rRNA amplicon sequencing-based analyses of coral disease associations incorporate tens of samples comparing diseased versus healthy corals from a single location, single time point and/or single transmission experiment (e.g., Rosales et al., 2019), which, coupled with relatively high coral microbial diversity, has resulted in the apparent disagreement in bacterial disease associations across coral disease studies. In this study, we obtained 16S rRNA amplicon sequencing data from 412 corals from five *A. cervicornis*-dominated reefs in Bocas del Toro, Panama every 6 months for 2 years. We then used bacterial ASV abundances and machine learning (ML) classifiers to produce highly accurate disease prediction models and identified the top bacterial ASV features as candidate pathogens. We compared our ML approach to more traditional differential abundance analyses using mixed models to identify ASVs that were associated with disease over our 2 years of field sampling. Finally, we used a tank-based transmission experiment to identify which of our top bacterial ASVs met our expectations of being WBD pathogens.

EXPERIMENTAL PROCEDURES

Field sampling

Surveys and collections occurred on five distinct *A. cervicornis* thicket-dominated reef sites (>1500 m^2) in Coral Cay, Bocas del Toro, Panama (Table S1). At each site, permanent 50-m belt transects were established in July 2015 and surveyed every 6 months in January and July of each year until July 2017. Video surveys of the 50-m belt transects (1 m on both sides) were used to calculate *A. cervicornis* abundance (presence per m^2 quadrat) and record WBD prevalence as the percentage of *A. cervicornis* quadrats with WBD. Between January 2016 and July 2017, 20 healthy (asymptomatic) and 10 disease samples were haphazardly collected at each site (5 or 1 m apart). Diseased and healthy branches were collected underwater and transported immediately to the surface where 1 cm of the disease interface was sampled using sterilized bone cutters, placed directly into sterile 5 mL cryovials containing 2 mL of CHAOS DNA buffer (4 M guanidine thiocyanate, 0.5% N-lauroyl-sacosine, 25 mM Tris (pH 8) 0.1 M beta mercaptoethanol, Fukami et al., 2004) and stored at −20°C for genetic analysis (Table 1). Twice as many healthy versus diseased colonies were sampled to determine if the microbial genetics could identify asymptomatic (i.e., disease infected) corals among apparently healthy individuals.

TABLE 1 Sampling and sequencing summary showing the number of coral fragments that passed quality filtering sampled from each time in the field and each experimental condition in the tanks. It also shows mean sequencing depth for healthy and diseased corals along with *t*-test-derived *p*-values indicating if there are significant differences in the number of reads from healthy and diseased fragments.

				Fragments		Reads		
Source	Time	Exposure	Genotypes	Healthy	Diseased	Healthy	Diseased	*p*-Value
Field	January 2016	—	100	63	37	8864 ± 1200	9164 ± 1107	0.867
	July 2016	—	104	72	32	10,725 ± 1238	14,593 ± 1564	0.054
	January 2017	—	94	60	34	7194 ± 985	6850 ± 727	0.769
	July 2017	—	114	74	40	5941 ± 689	5548 ± 591	0.651
	Total	—	412	269	143	8212 ± 532	8817 ± 580	0.443
Tank	Pre	—	6	18	—	4892 ± 671	—	—
	Post	Control		30	—	6813 ± 908	—	—
		Disease		25	6	5795 ± 836	3076 ± 541	0.003

White band disease prevalence

Acropora cervicornis abundance and WBD prevalence were modelled using a generalized linear mixed model with a logit link function using sampling time as a fixed effect and a random effect of the site with significance assessed using a likelihood ratio test (Bates et al., 2015). Pairwise contrasts were used to identify significant differences between sampling times using the Westfall *p*-value adjustment to control the family-wise error rate (Hothorn et al., 2008; Westfall, 1997). Bocas del Toro experiences two annual temperature peaks, with one crest between April and June and the other between September and November (Figure 1B; Kaufmann & Thompson, 2005). To investigate the relationship between seawater temperature, WBD prevalence and *A. cervicornis* abundance (Figure 1A,C,D), we used linear regressions to relate the logit transformed marginal mean of WBD prevalence and *A. cervicornis* abundance at each sampling time-point to the preceding hot season's mean temperature, maximum temperature, and the number of weeks with a mean temperature of above 30°C. We also used linear regression to relate *A. cervicornis* abundance to WBD prevalence within the same time point. Sea surface temperature data was acquired from the Smithsonian Tropical Research Institute's Bocas del Toro weather station (9°21′02.96″N, 82°15′28.27″W; Paton, 2019).

Microbiome characterization

Genomic DNA was extracted from each coral sample using GenElute DNA extraction kits. 16S rRNA amplicon sequencing of the V3-V4 region was produced using Klindworth et al.'s (2013) protocol, V3-V4 (341F/785R) primer sets, and four lanes of Illumina MiSeq 2 × 300 bp sequencing. Sequenced reads can be found in the NCBI Bioproject: PRJNA1106053. 16S rRNA gene reads were quality trimmed, overlapped and assembled into ASVs using the DADA2 denoising algorithm and pipeline in R (Callahan et al., 2016; R Core Team, 2022). Chimeras were removed and taxonomy was assigned to each ASV first using a Bayesian taxonomic classifier based on the NCBI 16S microbial database and classified to the lowest taxonomic level possible with greater than 80% classification confidence (Gao et al., 2017). Any unclassified sequences were further attempted to be classified via the naïve Bayesian classifier implemented in DADA2 using DECIPHER and the suggested threshold of 0.5 (Wright, 2016) and the Silva database (Quast et al., 2013). ASV sequences were aligned using DECIPHER and a neighbour-joining tree of the aligned ASV data was constructed using PHANGORN (Schliep, 2011). The resulting ASV table, taxa table, and 16S rRNA gene tree were imported into PHYLOSEQ (McMurdie & Holmes, 2013) and merged with the sample metadata for downstream analyses.

Samples were pruned to keep only samples with more than 1000 16S rRNA gene reads and ASVs identified as cyanobacteria, mitochondria, and chloroplast sequences were removed as host or algal contaminants. After comparing rarified alpha diversity metrics, ASVs were further filtered to retain ASVs found in at least 10% of samples and across all four sampling time points to remove low abundance ASVs. Read counts of the remaining ASVs were normalized for variable sequencing depth using the trimmed mean of the M-values method with singleton pairing including the normalization of the effective library size implemented in EDGER (Robinson & Oshlack, 2010; TMMwsp; Robinson et al., 2010). Normalized read counts plus a pseudo-count of 0.5 were converted to $\log_2$ counts per million reads with all subsequent analyses being performed on these normalized and log-transformed counts per million. This normalization method, termed Elib-TMM has similar performance to the popular ANCOM-BC normalization method but also allows for full contrasts and post hocs within modelling frameworks including linear mixed-models with interactions and post hocs (Lin & Peddada, 2020).

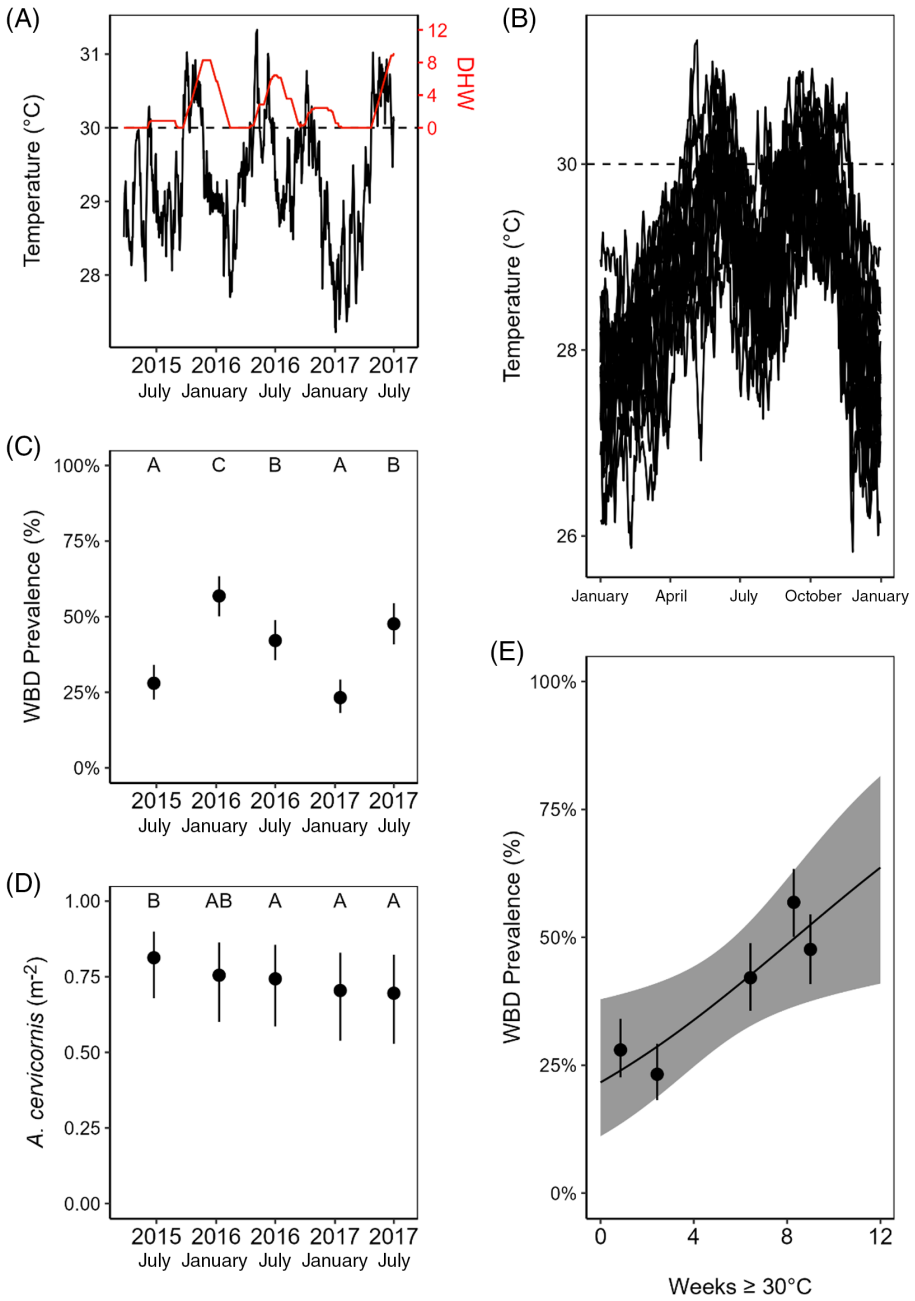

FIGURE 1 (A) Mean daily sea surface temperature (°C, black) and degree heating weeks (DHW;°C-weeks, red) during the study period. (B) Mean daily sea surface temperature visualized over one year, generated from data from 2000 to 2023, showing two annual peak times: April–June and September–November. (C) Mean White Band Disease prevalence in *Acropora cervicornis* colonies across sampling times. (D) *Acropora cervicornis* abundance (m^{-2}). (E) Relationship between mean WBD prevalence observed at each timepoint and the number of weeks ≥30°C during the preceding peak temperature season. In all panels, letters indicate significant groupings and error bars mark the 95% confidence intervals.

Community composition analysis

Differences in the alpha diversity metrics of ASV richness as well as both Shannon and inverse Simpson diversity indices were analysed using linear mixed models to test for differences in microbiome alpha diversity based on coral health and sampling time with a random effect of sampling location. To account for heteroscedasticity, the variance was modelled separately across sampling sites, times and coral health states. The richness and inverse Simpson diversity were transformed, using square-root and log transformations respectively, to meet the normality assumption of linear models, as is standard in the field. To visualize differences between healthy and diseased microbiomes and between sampling times and locations, we

estimated the microbial distance between coral samples using the Bray-Curtis distance as a measure of beta diversity (Bray & Curtis, 1957). These distances were visualized using a non-metric multidimensional scaling (NMDS) plot (Legendre & Legendre, 2012; Oksanen et al., 2013). We used a permutational ANOVA on the Bray-Curtis distances between coral samples with 10,000 permutations to test for differences in the microbial community composition based on coral health, sampling time and sampling site.

Differential abundance analysis

To act as a point of comparison with our ML approach, differential abundance analyses were performed on the field 16S rRNA amplicon sequencing data for each ASV using linear mixed-effects models with fixed effects of coral health state (diseased or healthy) and sampling time and a random effect of sample location (Bates et al., 2015) with high-quality Elib-TMM normalization while allowing for full linear mixed-model contrasts and post hocs (Lin & Peddada, 2020, 2024). The significance of fixed effects was assessed using F-tests and Kenward-Roger's method of calculating denominator degrees of freedom (Kenward & Roger, 1997). These *p*-values were then adjusted to control the false discovery rate and used to identify ASVs significantly associated with coral health state, sampling time point and interaction (Benjamini & Hochberg, 1995). To identify the subset of disease/healthy associated ASVs that were consistently differentially abundant across all four field time points, we performed a post hoc analysis to compare ASV abundance between healthy and diseased corals within each time point. We used Fisher's exact test to identify the genera overrepresented among those associated with diseased or healthy corals.

Machine learning model training and ASV feature identification

Bacterial ASVs associated with diseased corals were identified using an ensemble set of six distinct ML subcomponent models (lasso logistic regression; Friedman et al., 2010; Tibshirani, 1996), random forest (RF) (Ho, 1995; Wright & Ziegler, 2017), multilayer perceptron (Collobert et al., 2011; Falbel & Luraschi, 2023; Kuhn & Falbel, 2022), linear support vector machine (SVM) (Cortes & Vapnik, 1995; Karatzoglou et al., 2004, 2022), partial least squares (Rohart et al., 2017) and K-nearest neighbours (Cover & Hart, 1967; Fix & Hodges, 1989; Schliep & Hechenbichler, 2016). Prior to model fitting, the dataset was split into training (75%) and testing (25%) sets to reduce overfitting with all model tuning being done using the training set and the test set being used to evaluate the final model metrics (e.g., accuracy). The ASV counts were normalized using the Yeo Johnson transformation and then centred and scaled (Yeo & Johnson, 2000). The models were chosen to represent a diversity of ML classification models with the thought that they may independently select similar or different features (i.e., ASVs) for predicting coral disease state (Bolón-Canedo & Alonso-Betanzos, 2019).

Model hyperparameters were individually tuned to identify the parameter combination that minimizes the Brier score, a metric designed to penalize misclassifications and reward confident, correct classifications (Brier, 1950; Kruppa et al., 2014). All model hyperparameters were tuned by fitting the models on the training dataset using 10-fold cross-validation repeated 10 times. Each model was fit to the training dataset using an initial random grid with 50 random combinations of parameters. This random grid of fitted hyperparameters was used to initialize up to 200 iterations of Bayesian hyperparameter optimization to identify the parameter combination minimizing the Brier score (Wu et al., 2019).

Practically equivalent high-quality models of coral disease classification were defined as models that were 80% likely to be within 1% overall quality as the best-fit model. The overall quality metric used to compare models was a composite metric combining model accuracy, area under the receiver operator curve (ROC AUC), and Brier score (Derringer & Suich, 1980). Model accuracy is simply the percentage of coral fragments correctly classified (identified) as diseased or healthy. ROC AUC is an aggregate measure of model performance (true vs. false classifications) across all possible classification thresholds (Fawcett, 2006). Brier score rewards models that correctly classify coral fragments with more confidence (Brier, 1950). We combined these classification metrics into a composite metric to ensure that the models accurately and confidently predicted coral disease state. The overall quality metric for the repeated cross-validation fitted results was fitted using a hierarchical Bayesian model to identify differences in model quality (Kuhn & Silge, 2022). Random effects for repeats and folds within repeats were included to account for the repeated measurements of these data subsets (Kuhn & Silge, 2022). To ensure the models were not simply 'learning the data', we evaluated all models for accuracy, Brier score, and ROC AUC on the testing dataset after identifying the set of practically equivalent models. This ensures that the model quality metrics are similar when assessing data the models were not trained on.

Within each model, we identified ASVs important to classifying coral disease state for all of the top-quality models by calculating Monte Carlo-based Shapley Additive explanation (SHAP) values (Greenwell, 2023; Shapley, 1953; Štrumbelj & Kononenko, 2014). SHAP values were calculated independently for each model and ASV using 500 simulations with ASV importance

being calculated as the mean of the absolute values of the SHAP values (Molnar, 2022). Using an ensemble feature selection approach to take advantage of the strengths of different ML model types (Bolón-Canedo & Alonso-Betanzos, 2019; Pes, 2020; Pudjihartono et al., 2022), ASVs that were consistently highly ranked were identified by modelling ASV rankings across models using a generalized linear mixed model with a gamma distribution, a fixed effect of ASV identity, and a random effect of ML model with dispersion modelled separately for each ASV (Brooks et al., 2017). We then used post hoc tests to identify ASVs with significantly above-average rankings after adjusting for false discovery rate (Benjamini & Hochberg, 1995). We used rank-based overlap (RBO) to determine the correlation among ASV rankings for each ML model, the ensemble ranking list, and the null model list (Webber et al., 2010). To determine if the ensemble ranking appropriately represents the ML model rankings, we used a beta regression to compare the mean pairwise RBO among the ML models to the RBO between the ensemble and each ML model and to the RBO between the null model and ML models.

To identify which of the top ASVs in the ML model results are consistently associated with diseased or healthy corals in the field, we performed a post hoc analysis on the differential abundance models to compare ASV abundance between healthy and diseased corals within each timepoint.

Comparing differential abundance and machine learning

In our differential abundance and ML analyses, we identified ASVs that we defined as important and/or consistent. In the ML analysis, the important ASVs were identified as those with significantly better than average rankings and the consistent ASVs were a subset of these ASVs that were consistently differentially abundant between disease and healthy samples across all four field timepoints. Similarly, in our differential abundance analyses, important ASVs differed significantly in their abundance between diseased and healthy corals in the main effect of disease state and consistent ASVs were the subset of these ASVs which differed across all four time points. We used logistic regression to compare the percentage of important ASVs identified as consistent based on the method of determination (DA or ML) and if the ASV was healthy or disease-associated, as well as the interaction of the two. Then ASVs were grouped into categories based on whether the ASV was identified as important (or consistent) by ML, DA, both or neither.

We compared the ASV ML subcomponent rank across categories with more than five ASVs using linear mixed-effects models using the log of the subcomponent rank as the dependent variable explained by subcomponent model interacting with the consistency category (ML, DA, both or neither) nested within the importance category (ML, DA, both or neither) of each ASV and a random effect of ASV identity to account for the repeated measurements.

We calculated the Spearman correlation coefficient in diseased corals between every ASV and the ASVs identified as important and consistent by both ML and DA to understand how the ML subcomponents arrived at their rankings. We modelled the correlation with a mixed model with the consistency category nested within the importance category with random effects of both the focal ASV and correlated important/consistent ASV.

For both rank and correlation models, we assessed the significance of the independent terms using F-tests and Kenward-Roger's method of calculating denominator degrees of freedom (Kenward & Roger, 1997). For both metrics, we performed a post hoc analysis to identify differences between the group of ASVs identified as important and consistent by both ML and DA with all other categories. In the model of subcomponent ranks, we performed an additional post hoc comparing rankings of all pairs of ML models (lasso logistic regression, multilayer perceptron, RF and linear SVM) within each importance/consistency category to identify differences in subcomponent model behaviour.

Tank transmission validation of top ML associations

We used 16S rRNA amplicon sequencing data from a tank-based disease transmission experiment conducted in July 2017 to identify which of the consistently differentially expressed top ASVs predicted by the ML ensemble models displayed expected signatures of a pathogen rather than an opportunist. Six replicate fragments from six healthy coral genotypes were collected from Sebastian's reef, Bocas del Toro for the experiment and experimentally lesioned with a Waterpik to facilitate transmission (Gignoux-Wolfsohn et al., 2012). They were then distributed into three disease and three healthy-exposure 18-L recirculating tanks at ambient seawater temperatures. The three disease exposure tanks were dosed with 50 mL of disease slurry produced from 10 WBD-infected coral fragments while three healthy exposed tanks were dosed with 50 mL of healthy slurry created from 10 healthy fragments. Slurries were produced by liberating diseased or healthy coral tissue from the skeleton of sampled corals using a Waterpik containing filtered seawater (FSW) and normalizing the slurry doses to a standard ocular density of 0.6 at 600 nm. Two polyps from each coral fragment were sampled at three time points: when they were placed in the tanks (day 0), 2 days after exposure (day 2) and 8 days post-exposure or when WBD symptoms

developed, whichever occurred first (day 8). Diseased corals were removed from the tank to prevent disease amplification. The 16S rRNA gene data from the tank samples were sequenced and assembled with the field sample collections.

Differential abundance of the top ASVs from the ML models was analysed in the tank exposure experiment using a before-after control-impact design to identify ASVs associated with disease exposure and/or disease outcome. To account for the repeated measurements, we included random effects for coral fragments nested within the genotype and tank. The fixed effect treatments analysed were disease exposure and disease state (i.e., outcome) of the coral at the sampling timepoint. We used a set of a priori contrasts to categorize the top ASVs into likely pathogens compared to likely opportunists (Vega Thurber et al., 2020). The specific a priori contrasts tested for changes in abundances (1) across time following exposure to the diseased slurry as well as (2) between fragments that were exposed to diseased versus healthy slurries and (3) between disease exposed fragments that became infected versus remained healthy. These a priori contrasts were considered jointly to distinguish likely pathogens from opportunists. Specifically, both likely pathogens and opportunists were expected to have significantly higher abundances after exposure to the disease slurry. Likely pathogens were further expected to exhibit higher abundances in fragments that become infected compared to healthy fragments. Conversely, likely opportunists were predicted to exhibit higher abundances in disease-exposed compared to healthy slurry-exposed fragments, regardless of whether the coral fragment becomes infected.

We then examined the prevalence of the top putative pathogen ASVs in the field 16S rRNA gene data using a logistic regression model with fixed effects of ASV identity, sampling time, and health state and a random effect of the sampling site. To assess changes in prevalence through time, we used polynomial contrasts along with pairwise planned contrasts between healthy and diseased corals across ASVs.

RESULTS

White band disease prevalence

Between July 2015 and July 2017, annual seawater temperatures showed strong intra- and inter-annual variation (Figure 1A) with two typical peak heating events per year occurring on or around May and October (Figure 1B). Prolonged seawater temperatures above 30°C are associated with coral thermal stress events in the region and can lead to significant coral bleaching (Brown, 1997), which was observed across the Caribbean and at our study sites in Bocas del Toro, Panama in January 2016, exacerbated by the 2015/2016 El Niño (SVV unpublished data; Muñiz-Castillo et al., 2019). Degree heating weeks (DHW) above 30°C exceeded 5 weeks in January 2016, July 2016 and July 2017 (Figure 1A).

WBD prevalence varied significantly through time ($\chi^2_{(4)} = 99.98$, $p < 0.0001$, Figure 1C) ranging from a high of 56.9% (±3.4% SE) in January 2016 to lows of 28.0% (±2.9% SE) and 23.2% (±2.8% SE) in July 2015 and January 2017, respectively. WBD prevalence was significantly associated with the number of weeks with mean temperatures greater than 30°C during the preceding seasonal temperature peak ($F_{(1,3)} = 15.1$, $p = 0.03$, Figure 1E), but not the maximum or mean temperature of this period ($F_{(1,3)} = 1.9$, $p = 0.26$, $F_{(1,3)} = 3.7$, $p = 0.15$, respectively).

Acropora cervicornis declined significantly in abundance over the 2 years between July 2015 and July 2017 ($\chi^2_{(4)} = 19.58$, $p = 0.0006$, Figure 1D) dropping significantly from 0.81 m^{-2} (±0.06 SE) in July 2015 to 0.74 m^{-2} (±0.07 SE) in July 2016 and then gradually to 0.69 m^{-2} (±0.08 SE) in July 2017. This decline was not significantly correlated with the concurrent or preceding WBD prevalence at each site ($F_{(1,3)} = 0.07$, $p = 0.81$, $F_{(1,3)} = 0.19$, $p = 0.71$, respectively), or the preceding peak mean temperature ($F_{(1,3)} = 3.5$, $p = 0.16$), maximum temperature ($F_{(1,3)} = 1.8$, $p = 0.27$), or weeks above 30°C ($F_{(1,3)} = 1.2$, $p = 0.36$).

16S rRNA amplicon sequencing analyses

16S V3-V4 rRNA amplicon sequencing was obtained for 412 *A. cervicornis* (269 healthy and 143 diseased) fragments across the five sites and four time points (Table 1). We identified 9355 bacterial ASVs, including 604 ASVs that were present in more than 10% of all individuals. Of these, 342 were present across all four sampling times and present on corals in our July 2017 tank transmission validation experiment. These 9355 bacterial ASVs spanned 60 microbial classes, 137 orders, and 305 families (Figure 2). All three post-rarified alpha diversity metrics significantly differ depending on sampling time (richness: $\chi^2_{(3)} = 113.9$, $p < 0.0001$, Shannon: $\chi^2_{(3)} = 51.7$, $p < 0.0001$, and inverse Simpson diversity: $\chi^2_{(3)} = 30.7$, $p < 0.0001$) with the maximum diversity across all metrics being observed during July 2017 (richness = 118 ± 15, Shannon = 2.95 ± 0.29, inverse Simpson = 7.7 ± 1.7). There were no significant interactions between time and coral health (richness: $\chi^2_{(3)} = 1.2$, $p = 0.29$, Shannon: $\chi^2_{(3)} = 2.6$, $p = 0.45$, and inverse Simpson diversity: $\chi^2_{(3)} = 1.3$, $p = 0.71$). Healthy corals were found to have 11.5 (± 4.9, $\chi^2_{(1)} = 6.2$, $p = 0.013$) more microbial ASVs than diseased corals with no significant differences in either Shannon or inverse Simpson's

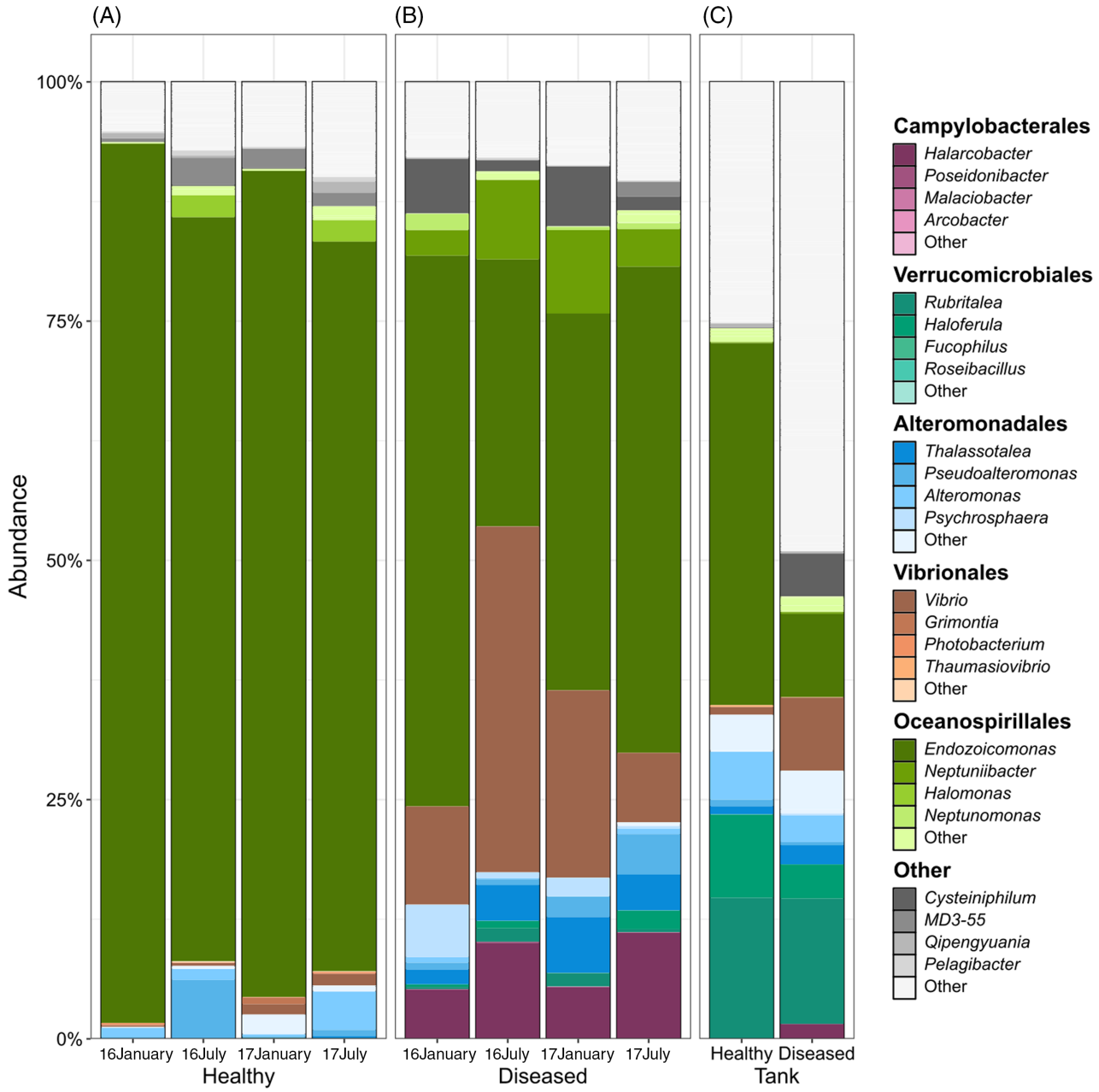

FIGURE 2 Microbial community compositions of (A) healthy, and (B) diseased coral fragments collected from the field and (C) coral fragments in the tank experiment. Colours indicate the major microbial families present in the coral microbiomes with different shades showing the dominant genera in each family.

diversity metrics ($\chi^2_{(1)} = 0.75$, $p = 0.39$, $\chi^2_{(1)} = 0.61$, $p = 0.43$, respectively).

NMDS and PERMANOVA analyses of the 342 shared ASVs show that the composition of the bacterial communities differed significantly across all levels (Table S2, Figure S1) with disease state accounting for 12% of the variation ($F_{(1,372)} = 104.4$, $p < 0.0001$), site explaining 15% of the variation ($F_{(4,372)} = 31.6$, $p < 0.0001$) and time explaining 12% of the variation ($F_{(3,372)} = 32.8$, $p < 0.0001$). The significant interactions between disease state, site and time (Table S2) indicate that the healthy and diseased microbiomes differ at each sampling time and location.

Differential abundance analyses initially showed that 244 out of the 342 ASVs (71.3%) differed significantly due to disease state (Figure 3); 82 ASVs from 19 genera were significantly more abundant on diseased corals and 162 ASVs from 45 genera were significantly more abundant on healthy corals. Further identification of ASVs which were consistently significantly differentially abundant in each sampling time reduced the number of disease-associated ASVs to

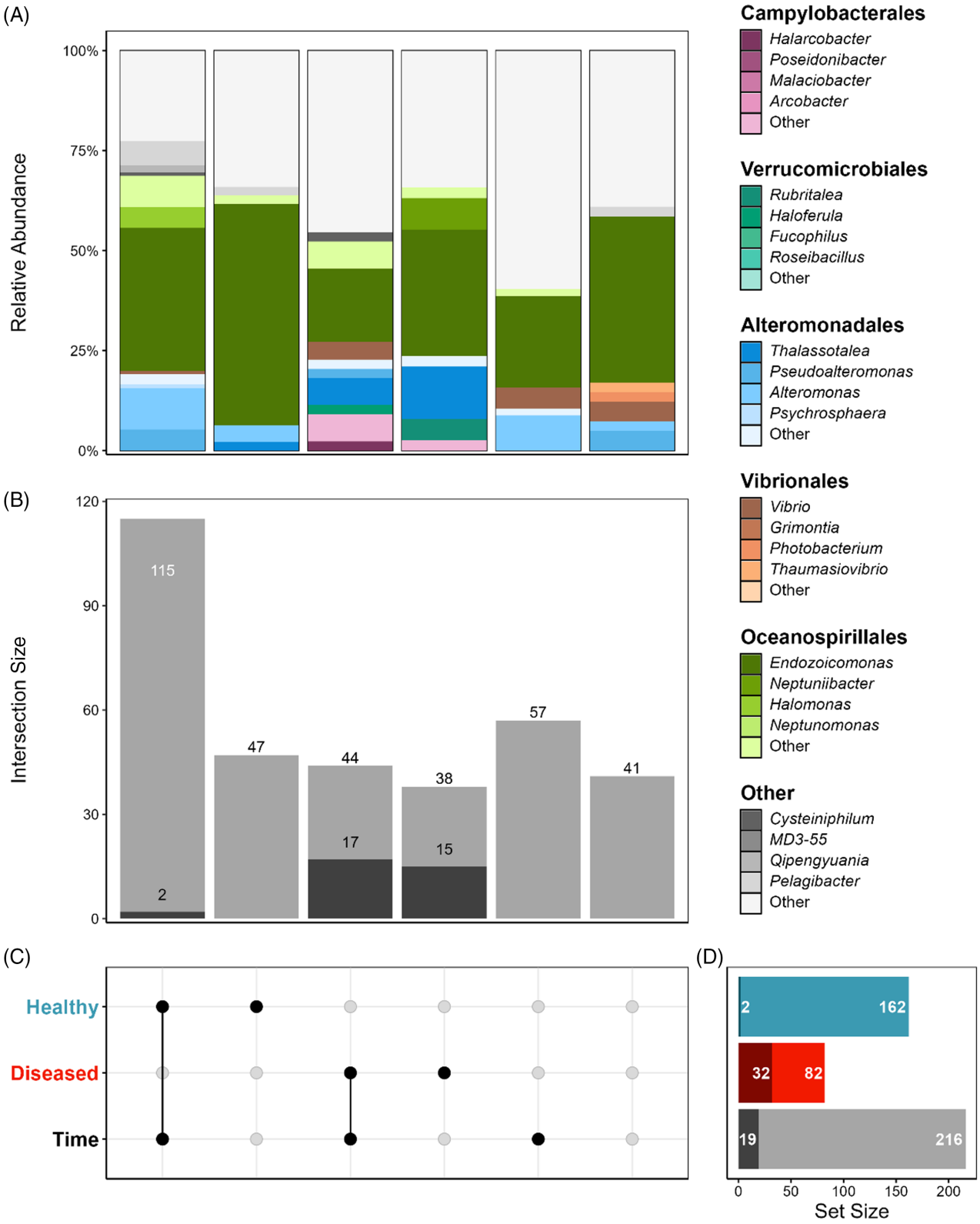

FIGURE 3 (A) Plot of the microbial community composition and (B) number of ASVs for each (C) combination of significant main effects (health state and/or time) of the linear mixed effects model. Points in panel C indicate that the ASVs are significant for the shown term. The above columns in panel B show the number of ASVs that are significant for that combination of terms and, in panel A, the microbial family composition of those ASVs (colours). (D) The total number of healthy-associated, disease-associated, and time-associated ASV panel. The darker shaded bars in panels B and D show the number of ASVs consistently differentially abundant across sampling time points.

TABLE 2 Model quality metrics showing the overall metric used to identify equivalent models and the components of the overall metric: accuracy, the area under the receiver operator curve (ROC/AUC) and Brier score. Lastly, the probability that each model is practically equivalent (within 1%) to the best overall model.

Algorithm	Overall (%)	Accuracy (%)	ROC/AUC	Brier score	Equivalence (%)
MLP	97.9 ± 0.3 (98.9)	98.0 ± 0.3 (99.0)	0.995 ± 0.001 (1)	0.017 ± 0.002 (0.01)	100.0
SVM	97.6 ± 0.2 (98.8)	96.7 ± 0.3 (98.1)	0.996 ± 0.001 (1)	0.015 ± 0.002 (0.007)	99.2
LASSO	97.4 ± 0.3 (97.4)	97.4 ± 0.3 (97.1)	0.997 ± 0.001 (0.998)	0.022 ± 0.002 (0.021)	97.7
RF	97.2 ± 0.2 (97.8)	97.7 ± 0.3 (98.1)	0.994 ± 0.001 (0.999)	0.026 ± 0.002 (0.022)	82.5
KNN	94.3 ± 0.4 (95.3)	95.4 ± 0.4 (96.2)	0.972 ± 0.004 (0.98)	0.044 ± 0.003 (0.038)	0.0
PLS	92.7 ± 0.2 (93.0)	98.0 ± 0.2 (99.0)	0.997 ± 0.001 (1)	0.094 ± 0.001 (0.096)	0.0
Null	53.0 ± 0.1 (53.0)	65.4 ± 0.1 (65.4)	0.5 ± 0 (0.5)	0.226 ± 0 (0.226)	0.0

Note: Numbers indicate the mean ± SE evaluated on the training data. Numbers in parentheses indicate the value when assessed on the test set. The solid line separates ML models that possess equivalently high-quality prediction metrics in the training dataset.

32 in 11 genera and dramatically reduced the number of healthy-associated ASVs to two ASVs in two genera. The two consistently healthy associated ASVs were *Qipengyuania* sp. (ASV40) and *Candidatus Pelagibacter ubique* (ASV207) the closest relative to the *Rickettsiales* (Le et al., 2014). Over-representation analyses of the consistently disease-associated genera identified *Thalassotalea* (four out of nine ASVs, $p = 0.002$), *Shimia* (two out of five ASVs, $p = 0.044$) and *Neptuniibacter* (two out of three ASVs, $p = 0.015$) were all significantly overrepresented among the disease associated genera relative to their abundance overall. The other consistently disease-associated ASVs include four Arcobacteraceae strains, one identified as *Halarcobacter bivalviorum* (ASV10); a Cellvibrionaceae, *Pseudoteredinibacter isoporae* (ASV31); a Crocinitomicaceae, *Crocinitomix* sp. (ASV594); 10 strains of Endozoicomonadaceae, six of which are identified as *Endozoicomonas atrinae* with the remaining four being unidentified; the Fastidiosibacteraceae *Cysteiniphilum litorale* (ASV25); five strains of Oceanospirillaceae, including the two overrepresented, unidentified *Neptuniibacter* sp.; a Pseudoalteromonadaceae *Pseudoalteromonas* sp. (ASV39); one additional unidentified Roseobacteraceae; a Rubritaleaceae *Rubritalea* sp. (ASV965); and one Vibrionaceae, *Vibrio* sp. (ASV8).

Model training and ASV identification

All six of the predictive ML models had high classification accuracy metrics (i.e., correctly predicting disease state using the ASV abundance data) of at least 92.7% with four of the six models possessing equivalent high-quality prediction metrics in the training dataset (Table 2). The multilayer perceptron (MLP) had the highest train-set quality (97.9% ± 0.3%) with the SVM (97.6% ± 0.2%, equivalence probability = 99.2%), lasso regression (97.4% ± 0.3%, equivalence probability = 97.7%) and RF (97.2% ± 0.2%, equivalence probability = 82.5%) all being equivalently effective models at classifying healthy and diseased corals based solely on their microbiomes.

Only 12 out of 412 coral samples were misclassified by any ML model. None of the 12 misclassifications were misclassified by all four equivalently effective models. Nine samples were misclassified by only one model, two samples were misclassified by both the MLP and SVM models, and one sample was misclassified by the MLP, RF and lasso regression models. Relatively even ratios of healthy and diseased samples were misclassified (5/269 healthy vs. 7/143 disease, $\chi^2_{(1)} = 2.07$, $p = 0.15$). High classification accuracies (97%+) and equal ratios of misclassifications between disease states (4.9% disease, 1.9% healthy) indicate that asymptomatic or presymptomatic corals were not common. We found no difference in the relative frequency of misclassifications between sampling sites ($\chi^2_{(4)} = 2.68$, $p = 0.61$), although there was an overabundance of misclassifications in January 2016 (7 out of 12, $\chi^2_{(3)} = 9.30$, $p = 0.03$).

ASV feature importance

To identify the most important and consistent bacterial ASVs contributing to the high predictive accuracies across the four equivalent ML models, we ranked ASVs based on SHAP values for each ML model and developed an ensemble ASV ranking. The ASV ranks across the top four ML models were highly correlated at 0.73 (±0.03), indicating that the top ASVs were consistent across models. Ensemble rankings identified a set of 24 ASVs that had significantly above-average feature importance ranks across the four models (Figure 4A). Nineteen ASVs were predictive of diseased corals and five ASVs were predictive of healthy corals (Figure 4B). Twelve of the 19 disease-associated ASVs were consistently significantly more abundant in diseased corals in the field across sampling times (Figure 4C), whereas one of the five healthy-associated ASVs (ASV40—*Qipengyuania* sp.) was consistently significantly more

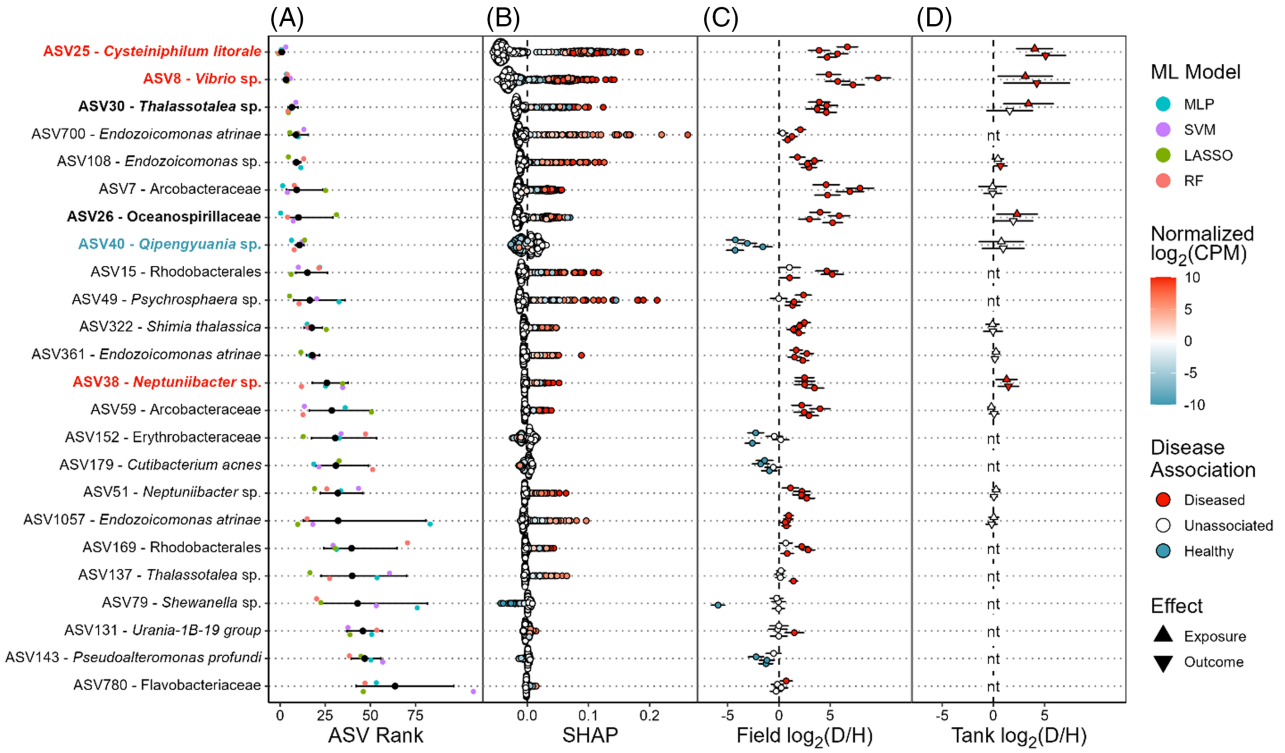

FIGURE 4 (A) Estimated model importance of all ASVs significantly above average showing modelled importance with 95% confidence intervals along with individual model rankings (colours). (B) SHAP values show the direction and magnitude by which each ASV alters the probability of coral fragments being diseased. The true disease state of the fragment is shown by the colour and the normalized amount of the ASV present in the fragment is indicated by the hue. (C) Field modelled estimates and 95% confidence intervals of the $\log_2$ fold-change in each sampling time with points coloured by significant associations with diseased (red) or healthy (blue) corals. White points are non-significant associations. (D) Tank modelled estimates and 95% confidence intervals of $\log_2$ fold-change differences between disease and healthy exposures (square) or outcomes (diamond). ASVs that were not consistently differentially expressed in the field are marked as 'nt', as they were not tested in the tank experiment. The stylization of ASV names indicates top candidates for potential pathogens (red, bold), opportunists (bold), and beneficial (blue, bold) bacteria.

abundant in healthy corals across time. The 12 disease-associated ASVs by rank importance included one *Cysteiniphilum* sp. (ASV25), one *Vibrio* sp. (ASV8), one *Thalassotalea* sp. (ASV30), three *Endozoicomonas* sp. (ASV108, ASV 361, and ASV1057), two Acrobacteraceae (ASV7 and ASV59), one *Shimia* sp. (ASV322) and three Oceanspirillaceae—two *Neptuniibacter* spp. and one uncharacterized (ASV26).

Comparing differential abundance and machine learning

While both the differential abundance and ML approaches identified different subsets of important ASVs, 12 ASVs were identified as consistently differentially abundant in both the ML and DA analyses (Figure 5). We found that ASVs identified as important by ML were 7.3× (±5.2) more likely to also be consistently compared with those identified by DA ($\chi^2_{(1)} = 5.47$, $p = 0.019$) and that healthy associated ASVs were 18.7× (±13.4) less likely to be consistently differentially abundant compared with disease-associated ASVs ($\chi^2_{(1)} = 67.2$, $p < 0.0001$). Regardless of the method used to identify ASVs as important healthy-associated ASVs were less likely to be consistent than disease-associated ASVs ($\chi^2_{(1)} = 1.66$, $p = 0.20$).

We identified seven unique groupings of ASVs depending on their categorization as important and consistent by DA and ML analyses, five of which contained more than five unique ASVs to be analysed (Figure 5D). ML subcomponent ranks differed significantly depending on both importance ($F_{(2,334)} = 184.0$, $p < 0.0001$) and consistency ($F_{(2,334)} = 37.0$, $p < 0.0001$), ML versus DA grouping categories, and their interaction with the ML subcomponent model ($F_{(6,1002)} = 2.5$, $p = 0.023$; $F_{(6,1002)} = 16.8$, $p < 0.0001$; respectively). There were no significant differences in subcomponent ranking depending on the subcomponent model alone ($F_{(3,1002)} = 2.4$, $p < 0.064$). The 12 ASVs identified as important and consistent by both approaches had significantly lower (i.e., better) average ranks (9.15 ± 1.38 SE, all $p < 0.0041$, Figure 5B) overall including against ASVs identified as important but not consistent by both approaches (23.4 ± 4.08 SE, $t_{(334)} = 3.3$, $p = 0.0041$) or ASVs identified as important

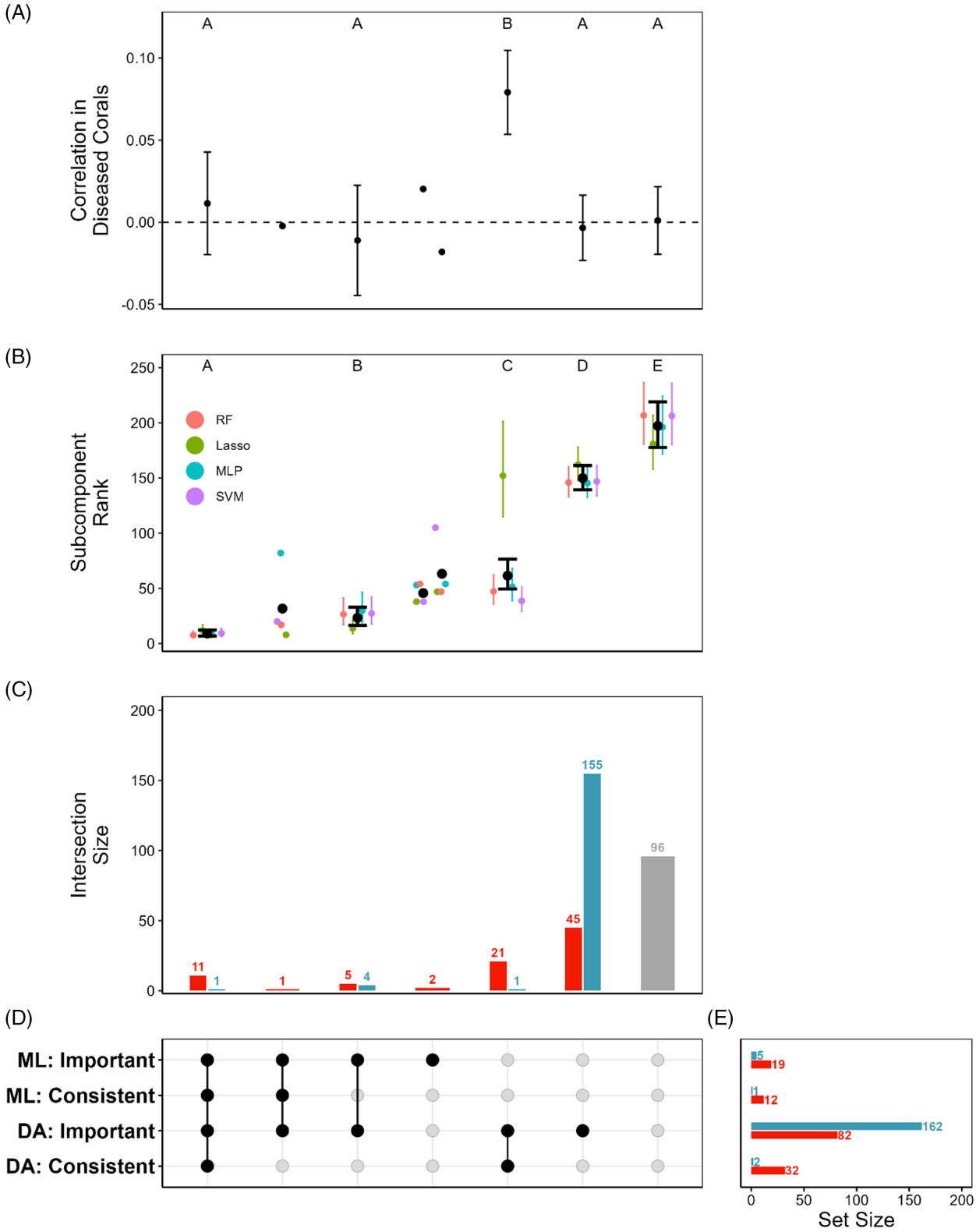

FIGURE 5 (A) Average correlation of the ASVs in each category with the ASVs identified as important and consistent by both ML and DA analyses in diseased coral fragments. (B) ML subcomponent model ranking across each of the four subcomponents (colours) along with the average ranking. In panels A and B points represent the mean value with error bars showing the 95% confidence intervals. Points without error bars show individual ASV values in groups with too few ASVs to model. Different letters at the top of the panels indicate significant differences among groupings. (C) The number of ASVs identified as healthy (blue), disease (red), or neither (grey) associated in (D) each combination of important/consistent ASVs identified by ML or DA models. (E) Shows the number of healthy (blue) and disease (red) associated ASVs depending on if the ASV is identified as important or consistent by either DA or ML models.

and consistent by DA alone (61.5 ± 6.9 SE, $t_{(334)} = 7.5$, $p < 0.0001$).

We then took the top 12 ASVs that were consistent and important across both ML and DA and looked at the average correlation in the disease samples within and between these top ASVs abundances against the ASVs in the other groupings (Figure 5A) to identify if any of the ASV groupings are correlated with the top ASVs and thus likely to be down-weighted in the ML feature selection. The average correlation within the top 12 ASVs was 0.01 (±0.02 SE). All other categories had similarly low correlations (all $p > 0.59$) except the ASVs identified as both important and consistent by DA alone which were significantly positively correlated with the top twelve ASVs (0.08 ± 0.01 SE, $t_{(372)} = 4.2$, $p = 0.0001$). This group of ASVs is also the only grouping that showed significant differences within its subcomponent ranking across models. Specifically, lasso logistic regression, which is particularly sensitive to feature correlations compared with other subcomponent models, ranked these ASVs 106 (±19 SE) ranks higher (i.e., poorer) than the other subcomponent models ($t_{(786)} = 5.6$, $p < 0.0001$), thus further down-weighting the relative importance of these correlated ASVs.

Tank validation of ASV importance

We used 16S rRNA amplicon sequencing data from our July 2017 tank-based transmission experiment to verify if the 12 top disease-associated ASVs and one healthy-associated ASV met the expectations of a pathogen or opportunist. 16S rRNA amplicon sequencing was obtained for six healthy coral genotypes before and after exposure (days 2 and 8) to disease or control (healthy) slurries. In the disease-exposed tanks, four out of the six disease-exposed genotypes developed WBD, allowing us to compare ASV differential abundances due to disease exposure and disease transmission outcomes. Our a priori expectations were that WBD pathogens would increase significantly in the disease exposure tanks and differ significantly due to disease exposure as well as disease outcome, whereas opportunistic bacteria would differ due to disease exposure and not disease outcome.

Pathogens

Three out of the 12 top disease-associated ASVs met our expectations for WBD pathogens (Figures 4D and 6A). Our top-ranked candidate, ASV25—*Cysteiniphilum litorale*, increased by 80-fold over time in the disease exposure tanks (±58.1 SE, $t_{(20.0)} = 6.0$, $p_{fdr} = 0.00005$), differed 16-fold (±8 SE, $t_{(6.9)} = 5.5$, $p_{fdr} = 0.006$) between the exposure treatments, and 34-fold (±23 SE, $t_{(67.5)} = 5.3$, $p_{fdr} = 0.000009$) due to disease outcome in disease exposed corals. Second ranked ASV8—*Vibrio* sp. increased 27-fold (±31 SE, $t_{(23)} = 2.8$, $p_{fdr} = 0.013$), had a ninefold exposure difference (±7 SE, $t_{(7.0)} = 2.8$, $p_{fdr} = 0.035$), and 19-fold outcome difference (±21 SE, $t_{(68.6)} = 2.7$, $p_{fdr} = 0.021$). Seventeenth ranked ASV38—*Neptuniibacter* sp. increased 4.2-fold (±1.6 SE, $t_{(16.9)} = 3.6$, $p_{fdr} = 0.007$), had a 2.5-fold exposure difference (±0.7 SE, $t_{(6.7)} = 3.0$, $p_{fdr} = 0.035$) and a 2.8-fold outcome difference (±1.0 SE, $t_{(63.2)} = 3.0$, $p_{fdr} = 0.014$).

The field prevalences of these three ASVs are also consistent with pathogens. Prevalences of all three ASVs differed due to disease state ($\chi^2_{(1)} = 203.3$, $p < 0.0001$, Figure S2, Table S3) with ASV8—*Vibrio* sp. having the highest average prevalence on diseased corals (95.2% ± 3.1% SE, $p_{ASV25} = 0.022$, $p_{ASV38} = 0.0001$) followed by ASV25—*Cysteiniphilum litorale* (77.5% ± 5.1% SE, $p_{ASV38} = 0.0004$), and then ASV38—*Neptuniibacter* sp. (43.0% ± 5.6% SE). ASV25—*Cysteiniphilum litorale* and ASV38—*Neptuniibacter* sp. had low prevalences on healthy corals (2.9% ± 1.2% SE and 1.4% ± 0.9% SE, respectively, $p = 0.59$) whereas the prevalence of ASV8—*Vibrio* sp. on healthy corals was significantly more elevated at 20.2% ± 7.9% SE ($p_{ASV25} = 0.0005$, $p_{ASV38} = 0.0024$). The prevalence of ASV25—*Cysteiniphilum litorale* has declined significantly in healthy corals ($p_{linear} = 0.024$) across the four sampling times while remaining relatively constant in diseased samples ($p_{linear} = 0.23$). Conversely, the prevalence of ASV8—*Vibrio* sp. increased significantly over time in diseased samples ($p_{linear} = 0.01$) but has been consistent in healthy samples ($p_{linear} = 0.63$). The prevalence of ASV38—*Neptuniibacter* sp. has been consistent through time in healthy ($p_{linear} = 0.17$) and diseased ($p_{linear} = 0.47$) corals.

Opportunists

Two ASVs differed significantly only by disease exposure, as is expected for opportunists (Figures 4D and 6B). A *Thalassotalea* sp. (ASV30) showed a 10.7-fold exposure effect (±7.5 SE, $t_{(6.6)} = 3.4$, $p_{fdr} = 0.035$) followed by the Oceanospirillaceae (ASV26) with a 5.0-fold difference (±2.8 SE, $t_{(6.7)} = 2.8$, $p_{fdr} = 0.035$). Unlike the putative pathogens, these ASVs were not significantly differentially abundant by diseased versus healthy outcomes on fragments exposed to the disease slurry (ASV26: $p_{fdr} = 0.058$, ASV30: $p_{fdr} = 0.17$).

DISCUSSION

WBD is endemic in Bocas del Toro, Panama with consistent prevalence above 28.0% (±2.9% SE) and semi-regular, cyclical flare-ups correlated with warm periods

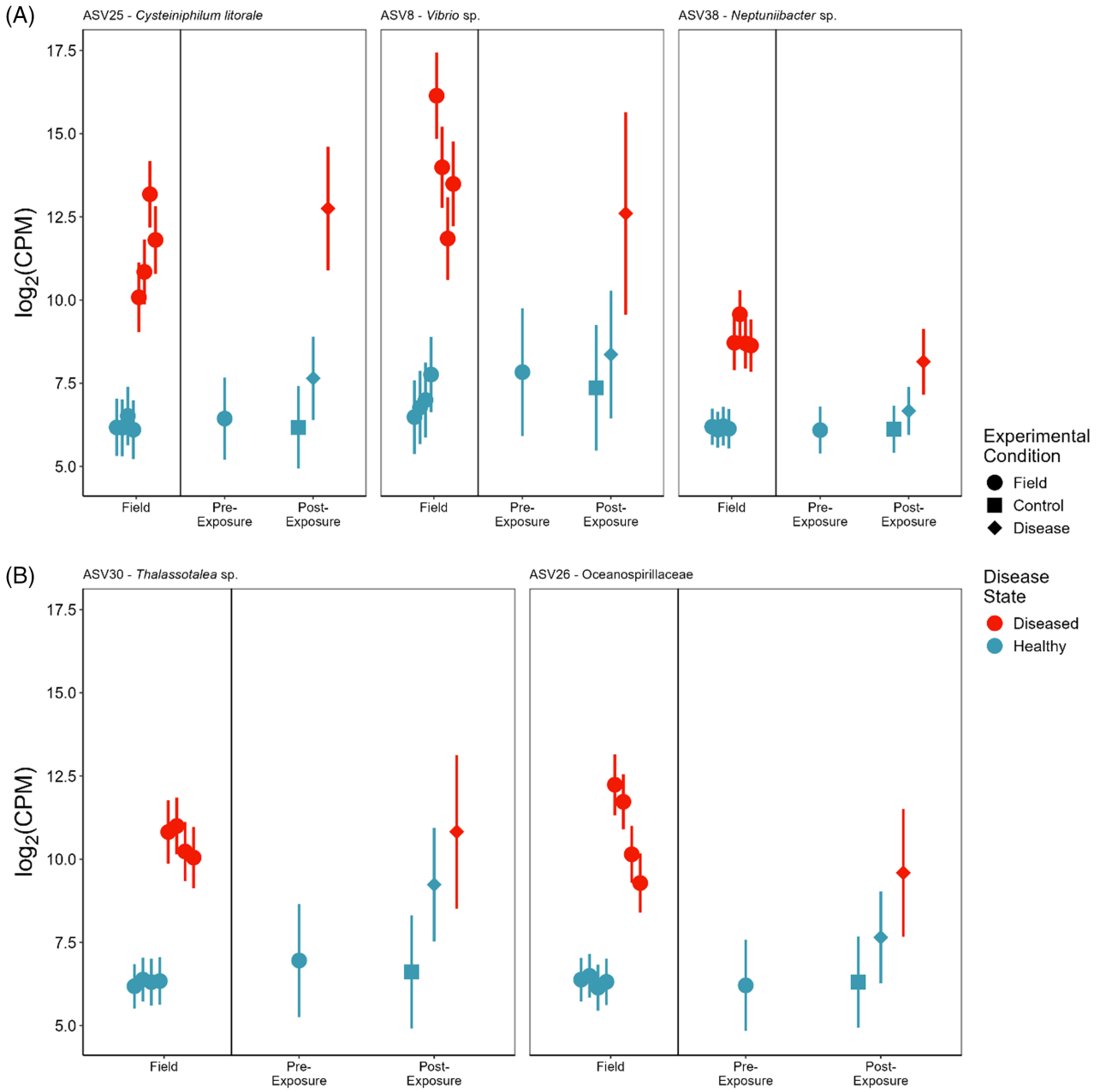

FIGURE 6 (A) ASVs classified as likely pathogens and (B) ASVs classified as likely opportunists. Each panel shows the ASV abundance $\log_2$ (CPM) from field samples and from each of the three timepoint and exposure type combinations in the tank experiment: pre-exposure, post-exposure to the diseased slurry and post-exposure to the control (healthy) slurry. Error bars indicate the 95% confidence interval.

(days above 30°C). Unlike the devastating Caribbean-wide losses during the early stages of the WBD epizootic (Aronson & Precht, 2001), we documented a gradual decline in *A. cervicornis* abundances of 14.5% over 2 years in Bocas del Toro with the most significant decline (7.1%) occurring following a bleaching event between July 2015 and January 2016. Differential abundance analyses of the 16S rRNA gene data identified 32 disease-associated ASVs from 11 bacterial genera, including a significant over-representation of the genera *Thalassotalea*, *Shimia* and *Neptuniibacter* on diseased corals. In contrast, our ML approach predicted *A. cervicornis* disease state with high accuracy (97% or higher) and identified 24 ASVs (19 disease-associated and five healthy-associated) as key features. Twelve disease-associated ASVs from the ML models were highly consistent across time, and three ASVs (i.e., strains)—ASV25 *Cysteiniphilum litorale*, ASV8 *Vibrio* sp. and ASV38 *Neptuniibacter* sp.—met the expectations of pathogens in our tank-based transmission experiment and should be targeted for isolation, cultivation and transmission assays. Moreover, the 97%+ prediction accuracy of the ML models for disease state across 143 disease and 269 healthy

samples suggests that asymptomatic 'apparently healthy' individuals are rare in the field and that WBD is a single disease syndrome that should not be divided into subcategories based on disease signs (e.g., WBD vs. RTL).

WBD pathogen candidates

Our top-ranked pathogen, ASV25, was classified as *Cysteiniphilum litorale*, a recently described species in a novel genus within Fastidiosibacteraceae isolated from coastal Chinese seawater. ASV25 has a 97% sequence match (402/415 bp) with multiple *Cysteiniphilum* strains including strain WZ-4 which caused a skin infection in a human working at a shrimp farm (Liu et al., 2017; Xu et al., 2021). In the field, ASV25 was found at detectable levels on 77.5% ± 5.1% SE of diseased *A. cervicornis* and only 2.9% ± 1.2% SE of healthy corals. *Cysteiniphilum litorale*—previously described as *Francisella*-like (Liu et al., 2017; Qian et al., 2023)—have been associated with WBD (Gignoux-Wolfsohn et al., 2017; Klinges et al., 2022; Walton, 2017) and similar *Cysteiniphilum litorale* ASVs (ASV 5b79cf6d5a5a9bf0bb866aed449eff44, 6 out of 232 nucleotide differences: 2.6%) have been observed on *A. cervicornis* disease grafts in Florida (Rosales et al., 2019). In addition to causing human skin infections (Xu et al., 2021), *Cysteiniphilum* genomes contain a range of different virulence factors, including a partial copy of the *Francisella* pathogenicity island, and closely related *Francisellas* are well-known pathogens across a broad taxonomic range (Cowley & Elkins, 2011; Nano & Schmerk, 2007; Qian et al., 2023), including in marine fishes and molluscs (Birkbeck et al., 2011; Colquhoun & Duodu, 2011).

Our second-ranked pathogen, ASV8 classified as a *Vibrio* species (family: Vibrionaceae) with 98% sequence matches to multiple *Vibrio* species—including *V. harveyi* and the coral pathogen *V. corallilyticus* (Ben-Haim et al., 2003; Luna et al., 2010)—but additional multi-gene sequencing data would be needed to accurately identify this ASV to species (Thompson et al., 2005). In the field, ASV8 was found on 95.2% ± 3.1% SE of diseased corals and 20.2% ± 7.9% SE of healthy corals. Historically, *Vibrio charcharia* (now synonymized with *V. harveyi*) was identified as the causal agent of WBD (Ritchie & Smith, 1998) and elicited WBD/RTL disease signs in in situ grafting assays on *A. cervicornis* (Gil-Agudelo et al., 2006), but 16S rRNA data or cultures were not submitted with this research. Rosales et al. (2019) identified an identical *Vibrio* strain (ASV: 7eb68c2ff12bb8a0a46d036c37f8f26e) as associated with WBD/RTL in *A. cervicornis* from Florida. *Vibrios* are well-known marine pathogens (Farmer III et al., 2015) and have been implicated in numerous coral diseases (Bourne et al., 2009) including WBD (Gil-Agudelo et al., 2006; Ritchie & Smith, 1998; Sweet et al., 2014) and coral bleaching with *Vibrio shiloi* on *Oculina patagonica* (Kushmaro et al., 2001; Rosenberg & Falkovitz, 2004). *Vibrio coralliilyticus* has been implicated in several diseases including tissue lysis in *Pocillopora damicornis* (Ben-Haim et al., 2003), White Syndrome in *Montipora* (Ushijima et al., 2014), and was recently identified as a co-infecting pathogen associated with SCTLD in the Caribbean (Ushijima et al., 2020). *Vibrio coralliilyticus* is usually a coral commensalist that becomes pathogenic at higher ambient temperatures (Ben-Haim et al., 2003; Ushijima et al., 2018) by disrupting nutrient exchange between the coral host and symbionts, causing subsequent destruction of tissue integrity (Gibbin et al., 2019) and outcompeting other commensal bacteria through active prophage induction (Wang et al., 2022).

The final ASV implicated as a potential pathogen is ASV38, classified as a *Neptuniibacter* species (family: Oceanospirillaceae) with a 98% sequence match to an unpublished HIMB1269 strain isolated from the coral *Porites compressa* in Hawaii. *Neptuniibacters* are not generally associated with marine diseases and instead are found in low carbon and nutrient surface waters (*N. caesariensis*; Arahal et al., 2007), salt pans (*N. halophilus*; Chen et al., 2012), associated with sea cucumber larvae (*N. victor*; Kudo et al., 2023) or scallop hatcheries (*N. pectenicola* and *N. marinus*; Diéguez et al., 2017). In the field, ASV38—*Neptuniibacter* sp. was found on only 43.0% ± 5.6% SE of diseased corals and 1.4% ± 0.9% SE of healthy corals; its relatively low prevalence on diseased corals suggests it is a disease-associated opportunist rather than a pathogen that may be reacting to quorum sensing molecules (Rezzonico & Duffy, 2008) produced by a quorum sensing pathogen (Certner & Vollmer, 2015, 2018).

Rosales et al. (2019) previously identified the most likely WBD pathogen in Florida as *Sphingobium yanoikuyae* (family: Sphingomonadacea). Two *Sphingobium* sp. ASVs were present in our dataset but were at low prevalence (1.9% ± 1.7% SE; 15/270 healthy and 4/143 diseased corals) and were removed with our low prevalence (10%) filter prior to analysis. Parasitic *Aquarickettsia* sp. have also been associated with increased disease susceptibility in nursery-raised *A. cervicornis* genotypes in Florida (Klinges et al., 2020). Twenty-one Aquaricketssia ASVs (*MD3-55* sp.) were detected in our dataset, but each was present at low prevalence in our data (1.3% ± 0.3% SE) and was also removed by the low prevalence filter.

Out of the top candidate pathogens identified in our ML models and tank experiments, we believe that ASV25, *C. litorale* and ASV8, *Vibrio* sp., represent the most likely primary WBD pathogens given that related ASV strains from both genera have previously been associated with WBD, and include pathogenic marine strains. Both ASV25, *C. litorale*, and ASV8, *Vibrio* sp., had similarly high prevalences on diseased corals

(77.5% and 95.2%, respectively), but ASV25, *C. litorale*, had much lower prevalence on healthy corals than ASV8, *Vibrio* sp. (2.9% vs. 20.2%, respectively). The low prevalence of ASV25 *C. litorale* on healthy corals suggests it is most likely an extrinsic pathogen, whereas the high prevalence of ASV8, *Vibrio* sp., on healthy corals coupled with *Vibrios* being well-known opportunistic coral pathogens (Munn, 2015) suggest it is a primary intrinsic pathogen or secondary opportunistic commensal (Vega Thurber et al., 2020), which could be directly tested through strain-based infection and co-infection experiments after both ASV strains are brought into pure culture.

Machine learning versus differential abundance analyses

Research into coral diseases has relied heavily on 16S rRNA amplicon sequencing and differential abundance analyses to identify ASVs that are associated with diseased corals, which typically identifies hundreds of disease-associated ASVs as candidate pathogens (e.g., Gignoux-Wolfsohn et al., 2017). Previous research using ML techniques to identify coral disease microbiomes has been broadly successful (70%+ accuracy) but hampered by a small sample size (<100), reducing the advantage of ML over traditional statistical methods (Barque et al., 2024). Differential abundance analyses of our 16S rRNA gene data identified 87 disease-associated ASVs, 32 of which were consistent in the field, whereas our ML approach narrowed this list to 19 disease-associated microbial ASVs, 12 of which were consistent in the field. Both ML and DA identified a greater proportion of healthy-associated ASVs as important which were not also consistent than they did with disease-associated ASVs. This likely reflects the high degree of temporal variability in healthy microbiomes.

ML approaches have several advantages over differential abundance-based methods to identify ASVs associated with disease states. ML classification models more fully utilize the features of the filtered ASV data to predict disease states as well as identify the top ASV features contributing to those predictions. For example, ML models can incorporate interactions among ASVs and use the simple presence/absence of an ASV in a diseased/healthy state rather than only using the ASV abundance as in DA methods. By comparing top ASV features across subcomponent models using an ensemble approach, we were able to identify highly consistent features across a range of ML classifiers, each of which has different approaches for handling interactions and collinearity among features, feature selection and prioritization (Pes, 2020). High predictive accuracy and correspondence across top ASV features (mean feature rank correlation = 0.73) suggest that the biological signatures of the disease associations in the 16S rRNA gene data are strong. Moreover, in comparing the ML versus DA analyses, 12 of the top ASVs identified by our ML analysis as important and consistent were also identified as important and consistent in DA analyses (average rank low correlation), whereas the 22 ASVs identified as important and consistent by the DA analyses alone were poorly ranked by the ML models and showed significant positive correlations the with top 12 ASVs. This demonstrates one of the key advantages of the ML approach—namely, that it can account for intercorrelated features during model selection and deprioritizes lower-ranked ASVs in our ensemble rank test. In our case, these 22 ASVs had significantly poorer ranks in the lasso subcomponent model (Figure 5B). Biologically, these 22 differentially abundant ASVs are likely opportunists associated with the disease microbiome, which the ML methods excluded as 'follower' ASVs during model selection.

Integrating ML and differential abundance approaches in a hypothesis-driven manner allowed us to limit the number of consistently disease-associated ASVs to 12, from 32 based on solely differential abundance analyses, which was further reduced to a tractable number of putative pathogens by utilizing an additional tank-based experiment and a priori assumptions about the behaviour of a pathogenic microbe when infecting a host. The importance of identifying a relatively small number of pathogenic ASV sequence strains is that their sequence identities can be used to isolate and grow these ASVs in culture to test their transmissibility in controlled, tank-based experiments to fulfil the Henle-Koch postulates (Evans, 1976; Fredricks & Relman, 1996; Henle, 1938; Koch, 1893).

Conclusions

Our ML approach demonstrates that ML-based modelling of 16S rRNA gene-based ASV abundances can be used to accurately predict coral disease states and identify top pathogens from hundreds of disease-associated ASVs. Using ML coupled with tank-based transmission data, we identified two ASVs as the most likely pathogens—ASV25 *Cysteiniphilum litorale* and ASV8 *Vibrio* sp. Previous work has identified identical or related *Cysteiniphilum* and *Vibrio* ASVs as potential WBD pathogens (Gignoux-Wolfsohn et al., 2017; Gil-Agudelo et al., 2006; Klinges et al., 2022; Ritchie & Smith, 1998; Rosales et al., 2019; Sweet et al., 2014; Walton, 2017) and thus both ASV strains should be the top targets for isolation, cultivation, genetic characterization and confirmation of Henle-Koch's postulate via transmission assays.

AUTHOR CONTRIBUTIONS

Jason D. Selwyn: Investigation; methodology; validation; visualization; writing – review and editing; writing – original

draft; formal analysis; data curation; software. **Brecia A. Despard:** Investigation; methodology; formal analysis; writing – review and editing. **Miles V. Vollmer:** Writing – review and editing; software; methodology; investigation; data curation. **Emily C. Trytten:** Writing – review and editing; investigation; methodology. **Steven V. Vollmer:** Conceptualization; funding acquisition; project administration; supervision; resources; writing – review and editing; investigation; methodology; formal analysis; data curation.

ACKNOWLEDGEMENTS

We would like to thank the staff of the Smithsonian Tropical Research Institute staff in Bocas del Toro, Panama for their help with this project. Sample collections were permitted with approval of Autoridad Nacional del Ambiente, Panama CITES permits (SEX/A-116-16 and SEX/A-98-19). Grant funding was provided to SVV by National Science Foundation Division of Ocean Sciences grants (NSF OCE-1458158 and OCE-1924145).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in NCBI at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1106053. With code and metadata available on GitHub: https://github.com/VollmerLab/WBD_ML_pathogen. Additional data used in the analysis can be found in Zenodo: https://zenodo.org/doi/10.5281/zenodo.13485942.

ORCID

Jason D. Selwyn https://orcid.org/0000-0002-9100-217X
Brecia A. Despard https://orcid.org/0000-0003-0560-958X
Miles V. Vollmer https://orcid.org/0009-0007-2914-4998
Emily C. Trytten https://orcid.org/0000-0001-9173-8155
Steven V. Vollmer https://orcid.org/0000-0002-1123-8706

REFERENCES

Alvarez-Filip, L., González-Barrios, F.J., Pérez-Cervantes, E., Molina-Hernández, A. & Estrada-Saldívar, N. (2022) Stony coral tissue loss disease decimated Caribbean coral populations and reshaped reef functionality. *Communications Biology*, 5, 1–10.

Arahal, D.R., Lekunberri, I., González, J.M., Pascual, J., Pujalte, M.J., Pedrós-Alió, C. et al. (2007) *Neptuniibacter caesariensis* gen. nov., sp. nov., a novel marine genome-sequenced gammaproteobacterium. *International Journal of Systematic and Evolutionary Microbiology*, 57, 1000–1006.

Aronson, R.B. & Precht, W.F. (2001) White-band disease and the changing face of Caribbean coral reefs. In: Porter, J.W. (Ed.) *The ecology and etiology of newly emerging marine diseases. Developments in hydrobiology*. Dordrecht: Springer Netherlands, pp. 25–38.

Barque, B.M., Rodrigues, P.J.S., de Paula Filho, P.L., Peixoto, R.S. & de Assis Leite, D.C. (2024) Prediction of health of corals *Mussismilia hispida* based on the microorganisms present in their microbiome. In: Pereira, A.I., Mendes, A., Fernandes, F.P., Pacheco, M.F., Coelho, J.P. & Lima, J. (Eds.) *Optimization, learning algorithms and applications*. Cham: Springer Nature Switzerland, pp. 409–423.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.

Ben-Haim, Y., Thompson, F.L., Thompson, C.C., Cnockaert, M.C., Hoste, B., Swings, J. et al. (2003) *Vibrio coralliilyticus* sp. nov., a temperature-dependent pathogen of the coral *Pocillopora damicornis*. *International Journal of Systematic and Evolutionary Microbiology*, 53, 309–315.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 289–300.

Birkbeck, T.H., Feist, S.W. & Verner–Jeffreys, D.W. (2011) *Francisella* infections in fish and shellfish. *Journal of Fish Diseases*, 34, 173–187.

Bolón-Canedo, V. & Alonso-Betanzos, A. (2019) Ensembles for feature selection: a review and future trends. *Information Fusion*, 52, 1–12.

Bourne, D.G., Garren, M., Work, T.M., Rosenberg, E., Smith, G.W. & Harvell, C.D. (2009) Microbial disease and the coral holobiont. *Trends in Microbiology*, 17, 554–562.

Bray, J.R. & Curtis, J.T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27, 325–349.

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A. et al. (2017) glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9, 378–400.

Brown, B.E. (1997) Coral bleaching: causes and consequences. *Coral Reefs*, 16, S129–S138.

Bruno, J., Selig, E., Casey, K., Page, C., Willis, B., Harvell, C. et al. (2007) Thermal stress and coral cover as drivers of coral disease outbreaks. *PLoS Biology*, 5, e124.

Burge, C.A., Mark Eakin, C., Friedman, C.S., Froelich, B., Hershberger, P.K., Hofmann, E.E. et al. (2014) Climate change influences on marine infectious diseases: implications for management and society. *Annual Review of Marine Science*, 6, 249–277.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. & Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13, 581–583.

Casas, V., Kline, D.I., Wegley, L., Yu, Y., Breitbart, M. & Rohwer, F. (2004) Widespread association of a *Rickettsiales*-like bacterium with reef-building corals. *Environmental Microbiology*, 6, 1137–1148.

Certner, R.H. & Vollmer, S.V. (2015) Evidence for autoinduction and quorum sensing in white band disease-causing microbes on *Acropora cervicornis*. *Scientific Reports*, 5, 11134.

Certner, R.H. & Vollmer, S.V. (2018) Inhibiting bacterial quorum sensing arrests coral disease development and disease-associated microbes. *Environmental Microbiology*, 20, 645–657.

Chen, M.-H., Sheu, S.-Y., Chiu, T.-F. & Chen, W.-M. (2012) *Neptuniibacter halophilus* sp. nov., isolated from a salt pan and emended description of the genus *Neptuniibacter*. *International Journal of Systematic and Evolutionary Microbiology*, 62, 1104–1109.

Collobert, R., Kavukcuoglu, K. & Farabet, C. (2011) Torch7: a Matlab-like environment for machine learning. *BigLearn*. NIPS Workshop.

Colquhoun, D.J. & Duodu, S. (2011) *Francisella* infections in farmed and wild aquatic organisms. *Veterinary Research*, 42, 47.

Cortes, C. & Vapnik, V. (1995) Support-vector networks. *Machine Learning*, 20, 273–297.

Cover, T. & Hart, P. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13, 21–27.

Cowley, S. & Elkins, K. (2011) Immunity to *Francisella*. *Frontiers in Microbiology*, 2, 1–21.

Derringer, G. & Suich, R. (1980) Simultaneous optimization of several response variables. *Journal of Quality Technology*, 12, 214–219.

Diéguez, A.L., Balboa, S., Magnesen, T. & Romalde, J.L. (2017) *Neptuniibacter pectenicola* sp. nov. and *Neptuniibacter marinus* sp. nov., two novel species isolated from a great scallop (*Pecten maximus*) hatchery in Norway and emended description of the genus *Neptuniibacter*. *Systematic and Applied Microbiology*, 40, 80–85.

Evans, A.S. (1976) Causation and disease: the Henle-Koch postulates revisited. *The Yale Journal of Biology and Medicine*, 49, 175–195.

Falbel, D. & Luraschi, J. (2023) torch: Tensors and Neural Networks with "GPU" Acceleration.

Farmer, J.J., III, Michael Janda, J., Brenner, F.W., Cameron, D.N. & Birkhead, K.M. (2015) Vibrio, The proteobacteria, part B: the gammaproteobacteria. In: Brenner, D.J., Krieg, N.R. & Staley, J.-R. (Eds.) *Bergey's manual of systematics of archaea and bacteria*. New York: John Wiley & Sons, Ltd, pp. 1–79.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874.

Fix, E. & Hodges, J.L. (1989) Discriminatory analysis. Nonparametric discrimination: consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57, 238–247.

Fredricks, D.N. & Relman, D.A. (1996) Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clinical Microbiology Reviews*, 9, 18–33.

Friedman, J., Hastie, T. & Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Fukami, H., Budd, A.F., Levitan, D.R., Jara, J., Kersanach, R. & Knowlton, N. (2004) Geographic differences in species boundaries among members of the *Montastraea annularis* complex based on molecular and morphological markers. *Evolution*, 58, 324–337.

Gao, X., Lin, H., Revanna, K. & Dong, Q. (2017) A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. *BMC Bioinformatics*, 18, 247.

Gibbin, E., Gavish, A., Krueger, T., Kramarsky-Winter, E., Shapiro, O., Guiet, R. et al. (2019) *Vibrio coralliilyticus* infection triggers a behavioural response and perturbs nutritional exchange and tissue integrity in a symbiotic coral. *The ISME Journal*, 13, 989–1003.

Gignoux-Wolfsohn, S.A. & Vollmer, S.V. (2015) Identification of candidate coral pathogens on white band disease-infected staghorn coral. *PLoS One*, 10, e0134416.

Gignoux-Wolfsohn, S.A., Marks, C.J. & Vollmer, S.V. (2012) White band disease transmission in the threatened coral, *Acropora cervicornis*. *Scientific Reports*, 2, 804.

Gignoux-Wolfsohn, S.A., Aronson, F.M. & Vollmer, S.V. (2017) Complex interactions between potentially pathogenic, opportunistic, and resident bacteria emerge during infection on a reef-building coral. *FEMS Microbiology Ecology*, 93, 1–10.

Gil-Agudelo, D.L., Smith, G.W. & Weil, E. (2006) The white band disease type II pathogen in Puerto Rico. *Revista de Biología Tropical*, 54, 59–67.

Gladfelter, W.B. (1982) White-band disease in *Acropora palmata*: implications for the structure and growth of shallow reefs. *Bulletin of Marine Science*, 32, 639–643.

Greenwell, B. (2023) fastshap: fast approximate Shapley values.

Harvell, C.D., Kim, K., Burkholder, J.M., Colwell, R.R., Epstein, P.R., Grimes, D.J. et al. (1999) Emerging marine diseases – climate links and anthropogenic factors. *Science*, 285, 1505–1510.

Henle, J. (1938) *On Miasmata and Contagia*. Baltimore, MD: Johns Hopkins Press.

Ho, T.K. (1995) Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1, pp. 278–282. Montreal, QC: IEEE.

Hothorn, T., Bretz, F. & Westfall, P. (2008) Simultaneous inference in general parametric models. *Biometrical Journal*, 50, 346–363.

Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. (2004) kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11, 1–20.

Karatzoglou, A., Smola, A. & Hornik, K. (2022) kernlab: kernel-based machine learning lab.

Kaufmann, K. & Thompson, R.C. (2005) Water temperature variation and the meteorological and hydrographic environment of Bocas del Toro, Panama. *Caribbean Journal of Science*, 41, 392–413.

Kenward, M.G. & Roger, J.H. (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983–997.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. et al. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, 41, e1.

Kline, D.I. & Vollmer, S.V. (2011) White band disease (type I) of endangered Caribbean Acroporid corals is caused by pathogenic bacteria. *Scientific Reports*, 1, 7.

Klinges, G., Maher, R.L., Vega Thurber, R.L. & Muller, E.M. (2020) Parasitic "Candidatus Aquarickettsia rohweri" is a marker of disease susceptibility in *Acropora cervicornis* but is lost during thermal stress. *Environmental Microbiology*, 22, 5341–5355.

Klinges, J.G., Patel, S.H., Duke, W.C., Muller, E.M. & Vega Thurber, R.L. (2022) Phosphate enrichment induces increased dominance of the parasite *Aquarickettsia* in the coral *Acropora cervicornis*. *FEMS Microbiology Ecology*, 98, fiac013.

Koch, R. (1893) Ueber den augenblicklichen Stand der bakteriologischen Choleradiagnose. *Zeitschr f Hygiene*, 14, 319–338.

Kruppa, J., Liu, Y., Diener, H.-C., Holste, T., Weimar, C., König, I.R. et al. (2014) Probability estimation with machine learning methods for dichotomous and multicategory outcome: applications. *Biometrical Journal*, 56, 564–583.

Kudo, R., Yamano, R., Yu, J., Koike, S., Haditomo, A.H.C., de Freitas, M.A.M. et al. (2023) Genome taxonomy of the genus *Neptuniibacter* and proposal of *Neptuniibacter victor* sp. nov. isolated from sea cucumber larvae. *PLoS One*, 18, e0290060.

Kuhn, M. & Falbel, D. (2022) brulee: high-level modeling functions with "torch".

Kuhn, M. & Silge, J. (2022) *Tidy modeling with R: a framework for modeling in the Tidyverse*, 1st edition. Sebastopol, CA: O'Reilly Media.

Kushmaro, A., Banin, E., Loya, Y., Stackebrandt, E. & Rosenberg, E. (2001) *Vibrio shiloi* sp. nov., the causative agent of bleaching of the coral *Oculina patagonica*. *International Journal of Systematic and Evolutionary Microbiology*, 51, 1383–1388.

Le, P.T., Pontarotti, P. & Raoult, D. (2014) Alphaproteobacteria species as a source and target of lateral sequence transfers. *Trends in Microbiology*, 22, 147–156.

Legendre, P. & Legendre, L.F.J. (2012) *Numerical ecology*, Vol. 24, 3rd edition. Amsterdam: Elsevier.

Lin, H. & Peddada, S.D. (2020) Analysis of microbial compositions: a review of normalization and differential abundance analysis. *Npj Biofilms and Microbiomes*, 6, 1–13.

Lin, H. & Peddada, S.D. (2024) Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. *Nature Methods*, 21, 83–91.

Liu, L., Salam, N., Jiao, J.-Y., Shun-Mei, E., Chen, C., Fang, B.-Z. et al. (2017) *Cysteiniphilum litorale* gen. nov., sp. nov., isolated from coastal seawater. *International Journal of Systematic and Evolutionary Microbiology*, 67, 2178–2183.

Luna, G.M., Bongiorni, L., Gili, C., Biavasco, F. & Danovaro, R. (2010) *Vibrio harveyi* as a causative agent of the white syndrome in tropical stony corals. *Environmental Microbiology Reports*, 2, 120–127.

McMurdie, P.J. & Holmes, S. (2013) Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, 8, e61217.

Miller, M.W., Lohr, K.E., Cameron, C.M., Williams, D.E. & Peters, E.C. (2014) Disease dynamics and potential mitigation among restored and wild staghorn coral, *Acropora cervicornis*. *PeerJ*, 2, e541.

Molnar, C. (2022) *Interpretable machine learning: a guide for making black box models explainable*. Munich: Independently Published.

Muñiz-Castillo, A.I., Rivera-Sosa, A., Chollett, I., Eakin, C.M., Andrade-Gómez, L., McField, M. et al. (2019) Three decades of heat stress exposure in Caribbean coral reefs: a new regional delineation to enhance conservation. *Scientific Reports*, 9, 11013.

Munn, C.B. (2015) The role of *Vibrios* in diseases of corals. *Microbiology Spectrum*, 3, 1–12. Available from: https://doi.org/10.1128/microbiolspec.ve-0006–2014

Nano, F.E. & Schmerk, C. (2007) The *Francisella* pathogenicity island. *Annals of the new York Academy of Sciences*, 1105, 122–137.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B. et al. (2013) vegan: community ecology package.

van Oppen, M.J.H., Willis, B.L., van Vugt, H.W.J.A. & Miller, D.J. (2000) Examination of species boundaries in the *Acropora cervicornis* group (Scleractinia, Cnidaria) using nuclear DNA sequence analyses. *Molecular Ecology*, 9, 1363–1373.

Paton, S. (2019) Bocas del Toro, Platform Tower_Water Temperature.

Pes, B. (2020) Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Computing and Applications*, 32, 5951–5973.

Precht, W.F., Gintert, B.E., Robbart, M.L., Fura, R. & van Woesik, R. (2016) Unprecedented disease-related coral mortality in southeastern Florida. *Scientific Reports*, 6, 31374.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W. & O'Sullivan, J.M. (2022) A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2, 927312.

Qian, C., Xu, M., Huang, Z., Tan, M., Fu, C., Zhou, T. et al. (2023) Complete genome sequence of the emerging pathogen *Cysteiniphilum* spp. and comparative genomic analysis with genus *Francisella*: insights into its genetic diversity and potential virulence traits. *Virulence*, 14, 2214416.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41, D590–D596.

R Core Team. (2022) R: a language and environment for statistical computing.

Rezzonico, F. & Duffy, B. (2008) Lack of genomic evidence of AI-2 receptors suggests a non-quorum sensing role for luxS in most bacteria. *BMC Microbiology*, 8, 154.

Ritchie, K.B. & Smith, G.W. (1998) Type II white-band disease. *Revista De Biologia Tropical*, 46, 199–203.

Robinson, M.D. & Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11, R25.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139–140.

Rohart, F., Gautier, B., Singh, A. & Cao, K.-A.L. (2017) mixOmics: an R package for'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13, e1005752.

Rosales, S.M., Miller, M.W., Williams, D.E., Traylor-Knowles, N., Young, B. & Serrano, X.M. (2019) Microbiome differences in disease-resistant vs. susceptible *Acropora* corals subjected to disease challenge assays. *Scientific Reports*, 9, 18279.

Rosenberg, E. & Falkovitz, L. (2004) The *vibrio shiloi*/*Oculina patagonica* model system of coral bleaching. *Annual Review of Microbiology*, 58, 143–159.

Schliep, K.P. (2011) Phangorn: phylogenetic analysis in R. *Bioinformatics*, 27, 592–593.

Schliep, K. & Hechenbichler, K. (2016) kknn: weighted k-nearest neighbors.

Shapley, L.S. (1953) A value for n-person games. In: *Contributions to the theory of games (AM-28)*. 1–12: Princeton University Press, p. 409.

Štrumbelj, E. & Kononenko, I. (2014) Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 647–665.

Sutherland, K.P., Porter, J.W. & Torres, C. (2004) Disease and immunity in Caribbean and Indo-Pacific zooxanthellate corals. *Marine Ecology Progress Series*, 266, 273–302.

Sweet, M.J., Croquer, A. & Bythell, J.C. (2014) Experimental antibiotic treatment identifies potential pathogens of white band disease in the endangered Caribbean coral *Acropora cervicornis*. *Proceedings of the Royal Society B: Biological Sciences*, 281, 20140094.

Thompson, F.L., Gevers, D., Thompson, C.C., Dawyndt, P., Naser, S., Hoste, B. et al. (2005) Phylogeny and molecular identification of Vibrios on the basis of multilocus sequence analysis. *Applied and Environmental Microbiology*, 71, 5107–5115.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, 58, 267–288.

Ushijima, B., Videau, P., Burger, A.H., Shore-Maggio, A., Runyon, C.M., Sudek, M. et al. (2014) *Vibrio coralliilyticus* strain OCN008 is an etiological agent of acute *Montipora* white syndrome. *Applied and Environmental Microbiology*, 80, 2102–2109.

Ushijima, B., Richards, G.P., Watson, M.A., Schubiger, C.B. & Häse, C.C. (2018) Factors affecting infection of corals and larval oysters by *Vibrio coralliilyticus*. *PLoS One*, 13, e0199475.

Ushijima, B., Meyer, J.L., Thompson, S., Pitts, K., Marusich, M.F., Tittl, J. et al. (2020) Disease diagnostics and potential coinfections by *Vibrio coralliilyticus* during an ongoing coral disease outbreak in Florida. *Frontiers in Microbiology*, 11, 569354.

Vega Thurber, R., Mydlarz, L.D., Brandt, M., Harvell, D., Weil, E., Raymundo, L. et al. (2020) Deciphering coral disease dynamics: integrating host, microbiome, and the changing environment. *Frontiers in Ecology and Evolution*, 8, 575927.

Vollmer, S.V. & Kline, D.I. (2008) Natural disease resistance in threatened staghorn corals. *PLoS One*, 3, e3718.

Vollmer, S.V. & Palumbi, S.R. (2002) Hybridization and the evolution of reef coral diversity. *Science*, 296, 2023–2025.

Voolstra, C.R., Raina, J.-B., Dörr, M., Cárdenas, A., Pogoreutz, C., Silveira, C.B. et al. (2024) The coral microbiome in sickness, in health and in a changing world. *Nature Reviews. Microbiology*, 22, 1–16.

Walton, C. (2017) Bacterial communities associated with healthy and diseased *Acropora cervicornis* (Staghorn coral) using high-throughput sequencing.

Wang, W., Tang, K., Wang, P., Zeng, Z., Xu, T., Zhan, W. et al. (2022) The coral pathogen *Vibrio coralliilyticus* kills non-pathogenic holobiont competitors by triggering prophage induction. *Nature Ecology & Evolution*, 6, 1132–1144.

Webber, W., Moffat, A. & Zobel, J. (2010) A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28, 1–38.
Westfall, P.H. (1997) Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, 92, 299–306.
Williams, D.E. & Miller, M.W. (2005) Coral disease outbreak: pattern, prevalence and transmission in *Acropora cervicornis*. *Marine Ecology Progress Series*, 301, 119–128.
Wright, E.S. (2016) Using DECIPHER v2.0 to analyze big biological sequence data in R. *The R Journal*, 8, 352.
Wright, M.N. & Ziegler, A. (2017) ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77, 1–17.
Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H. & Deng, S.-H. (2019) Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, 17, 26–40.
Xu, C., Zhang, X., Wu, Q., Chen, L., Qu, P., Zhang, Y. et al. (2021) Skin and soft tissue infection caused by *Cysteiniphilum litorale* in an immunocompetent patient: a case report. *Indian Journal of Medical Microbiology*, 39, 545–547.
Yeo, I.-K. & Johnson, R.A. (2000) A new family of power transformations to improve normality or symmetry. *Biometrika*, 87, 954–959.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Selwyn, J.D., Despard, B.A., Vollmer, M.V., Trytten, E.C. & Vollmer, S.V. (2024) Identification of putative coral pathogens in endangered Caribbean staghorn coral using machine learning. *Environmental Microbiology*, 26(9), e16700. Available from: https://doi.org/10.1111/1462-2920.16700