Title: Case-control design identifies ecological drivers of endemic coral diseases

Jamie M. Caldwell[1,2*], Greta Aeby[3], Scott F. Heron[4,5], and Megan J. Donahue[1]

1. Hawaii Institute of Marine Biology, University of Hawaii, Kaneohe, HI 96744

2. Department of Biology, Stanford University, Stanford, CA 94305

3. Qatar University, Department of Biological & Environmental Sciences, Doha, Qatar

4. Marine Geophysical Laboratory, Physics, College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia

5. NOAA Coral Reef Watch, College Park, MD 20740, USA

* contact: jsziklay@hawaii.edu

Abstract

Endemic disease transmission is an important ecological process that is challenging to study because of low occurrence rates. Here, we investigate the ecological drivers of two coral diseases -- growth anomalies and tissue loss -- affecting five coral species in Hawaii. We first show that a statistical framework called the case-control study design, commonly used in epidemiology but rarely applied to ecology, provided high predictive accuracy (68-84%) and disease detection rates (62-86%) compared with a traditional statistical approach that yielded high accuracy (98-100%) but low disease detection rates (0-13%). Using this framework, we found that larger corals were associated with higher disease risk across all host-disease pairs. We found that human activities in watersheds adjacent to shallow, semi-protected coral reefs were associated with growth anomalies whereas reef structure, water movement, and temperature exposure were associated with tissue loss. Variation in risk factors across host-disease pairs suggests

that either different pathogens cause the same gross lesions in different species or that the same disease may arise in different species under different ecological conditions.

Non-technical summary

Understanding the conditions that drive disease transmission during non-outbreak periods is central for understanding disease ecology but challenging to study because diseased individuals are rare. In this study, we applied a statistical technique commonly used in epidemiology to study disease occurrence for two coral diseases that affect five species of coral in Hawaii. Using this statistical technique, we improved disease detection rates considerably and identified several factors associated with different coral diseases. The most consistent finding was that larger coral colonies were at greater risk of infection across all disease types and host species. For growth anomalies, a disease that is found year-round in Hawaii at low levels, disease was more likely to occur in semi-protected bays near areas with high coastal development or agriculture and golf courses. For tissue loss, a disease that causes large outbreaks but is otherwise very rare, disease risk was associated with coral density, depth, water movement, and water temperature. Differences in the ecological drivers of disease across host species and disease types suggests that what is commonly considered the same disease based on visual disease lesions may in fact be caused by different pathogens, or, that the same pathogen may cause disease in different coral species under different conditions.

Main

Disease is an ecologically and evolutionarily important process in shaping populations, communities, and ecosystems [1–8], but identifying factors that promote disease at low endemic levels or between epidemics is challenging because disease occurrences are rare in space and time. Disease can affect populations, communities, and ecosystems through multiple pathways, such as changing the physiology, behavior, distribution, abundance, and fitness of organisms (e.g., Behringer, Butler, & Shields, 2006; Lafferty & Morris, 1996; Woodworth et al., 2005). For example, when the trematode *Plagioporous* sp. infects the coral *Porites compressa* it reduces coral growth and causes infected polyps to appear as bright swollen nodules that cannot retract into their skeletal cups increasing fish predation, thus affecting coral physiology, behavior, and fitness (Aeby, 1991). As another example, infection in three-spined sticklebacks by the fish parasite *Gyrodactylus* spp. has been shown to alter the zooplankton community structure and nutrient cycling in streams and lakes, which in turn, affect the survival and fitness of the subsequent fish generation [8]. Despite the ecological and evolutionary importance of disease, understanding the factors that increase individual disease risk for endemic diseases with low prevalence or during interepidemic periods for diseases with epidemic cycles can be challenging because of the low probability of observing diseased individuals during non-outbreak periods.

With limited occurrence data relative to nonoccurrence data, commonly used biostatistics such as logistic regression can significantly underestimate the probability of rare events such as disease occurrence [12]. Further, such models are likely to have high

predictive accuracy (proportion of correctly identified event and nonevent observations), which can be misleading: for example, if 3 of 100 individuals are diseased, a model where all observations are predicted healthy would return 97% predictive accuracy but provide very limited information about disease occurrence or causation. Ecologists interested in rare events such as insect outbreaks and invasion of new species could benefit from adopting statistical approaches from fields of inquiry concerned with rare phenomena. One statistical framework that works well with low event-to-nonevent ratio datasets is called the case-control study design [13]. The case-control study design is commonly used in epidemiology where the number of people affected by a condition is very small relative to the potential control population (e.g., few people are HIV+ compared with the general population). In the case-control design, for each subject with some condition, a control subject is selected that does not have the condition but is otherwise similar in terms of factors such as age, sex, and occupation. This approach naturally lends itself to many ecological studies with low event-to-nonevent ratios such as wildlife disease occurrence.

Here, we compare logistic regression models with and without a case-control design to answer three ecological questions for marine wildlife diseases: 1) how accurately can we predict disease occurrence; 2) what are the ecological factors associated with maintaining low levels of disease occurrence; and 3) are risk factors more similar for closely related species? We used coral diseases for this case study because (i) they can be extremely rare (< 0.1% of disease occurrence observations); (ii) lesions can be visually identified; (iii) they can cause high mortality or strongly reduce fecundity with

likely long-term population effects; and (iv) large datasets are publicly available on coral health observations and hypothesized disease drivers.

In this study, we investigate two types of coral diseases – growth anomalies and tissue loss diseases – affecting five species of coral from two widely distributed genera in the Indo-Pacific Ocean. Growth anomalies are chronic, protuberant masses of coral skeletons (i.e., tumors) that reduce growth, fecundity, and overall coral survival [14–17]. Growth anomalies in some coral genera (e.g. *Porites*) have been associated with larger coral colonies [18] located in shallow water (< 3 m) [15] with elevated nutrients (Kaczmarsky & Richardson, 2011) and in regions with high human populations (Aeby et al., 2011). Tissue loss diseases, or white syndromes, usually have very low endemic prevalence levels (<1% in the Indo-Pacific) [21] but can be associated with rapid outbreak events resulting in localized mass mortality (Aeby et al., 2016a; Brodnicke et al., 2019; Hobbs, Frisch, Newman, & Wakefield, 2015; Precht et al., 2016; Roff et al., 2011; Walton, Hayes, & Gilliam, 2018). This suite of diseases is characterized by progressive tissue loss across the coral colony with lesions progressing slowly (chronic to subacute) or rapidly (acute). Bacterial pathogens cause many tissue loss diseases [28–31] and can be associated with high coral cover (i.e., density dependent transmission) and temperature stress (e.g. anomalously warm or cold ocean temperatures) (Bruno et al., 2007; Caldwell, Heron, Eakin, & Donahue, 2016; Heron et al., 2010; Maynard et al., 2011; Randall, Jordan-Garza, Muller, & Van Woesik, 2014). Based on these previous findings, we hypothesize that long-term, chronic stressors are more likely to drive growth anomaly persistence whereas locations with pulses of short-term, acute stressors are

more likely to be associated with interepidemic tissue loss occurrence. Disease lesions indicating growth anomalies and tissue loss appear on numerous coral species, but few pathogens underlying these diseases have been identified and so it is currently unknown whether these disease lesions constitute a single disease or multiple diseases. Thus, we investigate each host-disease pair separately.

Results

Comparison between statistical approaches

Logistic regression models with and without a case-control design had high overall predictive accuracy (proportion of correctly predicted healthy and disease observations) for all host-disease pairs, but only the models with the case-control design accurately predicted disease occurrence (proportion of disease observations the model correctly predicted as diseased; i.e., the true positive rate). Using colony health state as a binary response variable (0 indicating healthy, 1 indicating diseased) and a suite of hypothesized predictor variables (Tables 1-2), we found that mean predictive accuracy ranged from 98-100% for logistic regression without the case-control design and 68-84% for logistic regression with the case-control design (Table 3). The true positive rate, or the proportion of disease observations the model correctly predicted as diseased, is of greatest interest for rare event prediction and ranged from 0-13% for logistic regression without the case-control design and 62-86% for logistic regression with the case-control design (Table 3). Given the moderate-to-high mean predictive accuracy (68-84%) and mean true positive rates (62-86%) for the case-control design, we report only results using the case-control design below.

Table 1. Predictor variables included in candidate models for each disease type.

Growth anomalies	Tissue loss
colony size	
host density	
depth	
hot snap	
cold snap	
winter condition	
human population size	
embayment	
agriculture and golf runoff	stream exposure
coastal development	rainfall anomaly
herbivore fish abundance	herbivore fish biomass
wave energy maximum monthly mean	wave energy maximum anomaly
frequency of irradiance anomalies	chlorophyll-*a* maximum anomaly

Table 2: Description of biological, environmental, physical, and anthropogenic predictor variables used in candidate models.

Predictor variable	Ecological category	Characteristic	Description	Minimum value	Maximum value	Source
Colony size	Biological	Site	Maximum diameter or average size of size class (cm).	<1	300	Survey data[A]
Host density	Biological	Site	Number of colonies divided by survey area (m^{-2}).	0.02	9.49	Survey data[A]
Herbivore fish abundance	Biological	Site	Pooled estimate of herbivorous fish calculated between 2012 and 2013 (g m^{-2}).	0.06	0.42	Heenan et al. 2014 [37]
Chlorophyll-*a*	Environmental	Chronic	Maximum anomaly calculated within a 30-meter boundary of the coast between 2003 and 2013 from NASA MODIS satellite imagery (mg m^{-3}).	0.0004	0.3566	Wedding et al. 2018 [38, B]
Irradiance	Environmental	Chronic	Photosynthetically active radiation (PAR) average annual frequency of anomalies calculated within a 30-meter boundary of the coast between 2003 and 2013 from NASA MODIS satellite imagery (mol m^{-2} day^{-1}).	0.13	0.27	Wedding et al. 2018 [38, B]
Hot Snap	Environmental	Acute	Magnitude and duration of heat stress above a summertime threshold at 5 km resolution (°C-week).	0	3.6	NOAA CRW[C]
Cold Snap	Environmental	Acute	Magnitude and duration of cold stress below a wintertime threshold at 5 km resolution (°C-week).	-4.8	0	NOAA CRW[C]
Winter Condition	Environmental	Acute	Accumulation of positive and negative thermal anomalies about a wintertime threshold at 5 km resolution (°C-week).	-9.4	8.2	NOAA CRW[C]
Rainfall	Environmental	Acute	Monthly rainfall anomaly interpolated from rain gauge data, expert knowledge, radar observations, meteorological model simulations and vegetation data. Anomalies were calculated as the rainfall variation in any given month and year from the 1978-2007 mean.	0.04	4.35	Rainfall Atlas of Hawai'i[39,D]; Frazier et al. 2016 [40]
Depth	Physical	Site	Distance below sea surface (m).	0.6	26.0	Survey data[A]
Embayment	Physical	Site	Binary variable indicating whether survey site is inside or outside of semi-enclosed habitat.	0	1	Calculated in this study
Stream	Physical	Acute	Inverse planar distance between survey site and stream	2.72	3.33	Calculated in

exposure			mouth in nearest watershed (m^{-1}).			this study
Wave energy	Physical	Acute	Average annual maximum anomaly of wave energy flux calculated between 2000 and 2013 (kW m^{-1}).	0.08	285.74	Wedding et al. 2018 [38, B]
Wave energy	Physical	Chronic	Average annual maximum monthly mean of wave energy flux calculated between 2000 and 2013 (kW m^{-1}).	0.04	103.48	Wedding et al. 2018 [38, B]
Coastal development	Anthropogenic	Chronic	Relative level of new development between 2005 and 2011.	0.00	0.42	Wedding et al. 2018 [38, B]
Agricultural and golf runoff	Anthropogenic	Chronic	Proxy for fertilizer and chemicals like pesticides and herbicides discharge from agricultural land and golf courses calculated as area of cultivated land and golf course per watershed (km^2) with a distance-decay function.	0.00	0.53	Lecky 2016[41]
Human population	Anthropogenic	Chornic	Number of people living within 1 km of coastline nearest to survey site.	0	4425	NASA SEDAC GPW v4[E]

Public datasets:

A. Hawai'i Coral Disease database: https://www.sciencedirect.com/science/article/pii/S2352340916304607
B. Ocean Tipping Points: http://www.pacioos.hawaii.edu/projects/oceantippingpoints/#data
C. NOAA Coral Reef Watch CoralTemp: https://coralreefwatch.noaa.gov/satellite/coraltemp.php
D. NASA SEDAC Gridded Population of the world v4: http://sedac.ciesin.columbia.edu/data/collection/gpw-v4
E. Rainfall Atlas of Hawai'i: http://rainfall.geography.hawaii.edu/[39]

Table 3: Logistic regression with case-control design has higher detection rates of disease occurrence compared with logistic regression without case-control design. Each row corresponds to a single host-disease pair; models at the genus level include observations from all hosts species within that genus (e.g., *Montipora* spp includes all observations from *Montipora capitata* and *Montipora patula*). Disease type refers to growth anomalies (GA) or tissue loss (TLS). Total diseased and total healthy refers to the total number of colony observations within the specified host-disease pair after removing observations with missing predictor data. Mean diseased and mean healthy refer to the average number of diseased and healthy observations from the withheld test data across 500 random samples. Mean accuracy is the average predictive accuracy (correctly predicted diseased and healthy observations divided by total number of observations) across the 500 randomly sampled withheld test data sets. Mean TPR is the average true positive rate (correctly predicted diseased observations divided by the total number of diseased observations) across the 500 randomly sampled withheld test data sets.

				Logistic regression without case-control				Logistic regression with case-control			
Host	Disease type	Total diseased (n)	Total healthy (n)	Mean diseased (n)	Mean healthy (n)	Mean accuracy (%)	Mean TPR (%)	Mean diseased (n)	Mean healthy (n)	Mean accuracy (%)	Mean TPR (%)
Montipora capitata	GA	49	35864	10	7173	100	0	10	10	81	75
Montipora patula	GA	35	23118	7	4624	100	12	7	7	79	71
Montipora spp	GA	84	58982	17	11797	100	2	17	17	80	73
Porites compressa	GA	160	17889	31	3579	99	13	32	31	84	86
Porites evermanni	GA	42	1779	8	357	98	6	8	8	72	68
Porites lobata	GA	1396	64312	279	12863	98	5	279	258	76	73
Porites spp	GA	1598	83980	319	16797	98	4	319	298	75	71
Montipora capitata	TLS	39	28669	8	5734	100	3	8	8	83	83
Montipora patula	TLS	12	17480	2	3497	100	0	2	2	68	62
Montipora spp	TLS	51	46149	10	9230	99	2	11	10	79	76
Porites compressa	TLS	89	16915	18	3383	100	0	18	18	73	71
Porites evermanni	TLS	17	1567	3	314	99	7	3	3	81	80
Porites lobata	TLS	351	61489	70	12298	99	0	71	69	74	67
Porites spp.	TLS	457	79971	91	15995	99	0	91	90	74	68

Ecological risk factors for species-specific models

Of the 20 risk factors (predictor variables) we initially examined (see methods), 13 were included in our candidate models (with the 13 risk factors differing by disease type; Tables 1-2), but the most predictive risk factors varied by disease type and host species. The 13 risk factors can be broadly classified into four categories: biological (colony size, host density, and herbivorous fish biomass), environmental (hot snap, cold snap, winter condition, and irradiance), physical (depth, embayment, wave exposure, and stream exposure), and anthropogenic (coastal development, agricultural/golf runoff, and human population size). For growth anomalies in *Montipora capitata* and *Montipora patula* we identified two biological (colony size and herbivorous fish biomass) and one environmental (cold snap) risk factor (Fig. 1). For growth anomalies in *Porites* spp., all hypothesized risk factors except herbivorous fish biomass were associated with disease occurrence across the three *Porites* species, but only colony size, cold snap, and depth were risk factors for all three *Porites* species (Fig. 1). The only risk factors common to at least one species in each genus were colony size and cold snap, and colony size was the only risk factor common to all five host species (Fig. 1) indicating that risk factors for growth anomalies were more similar for species within the same genus compared with species in a different genus.

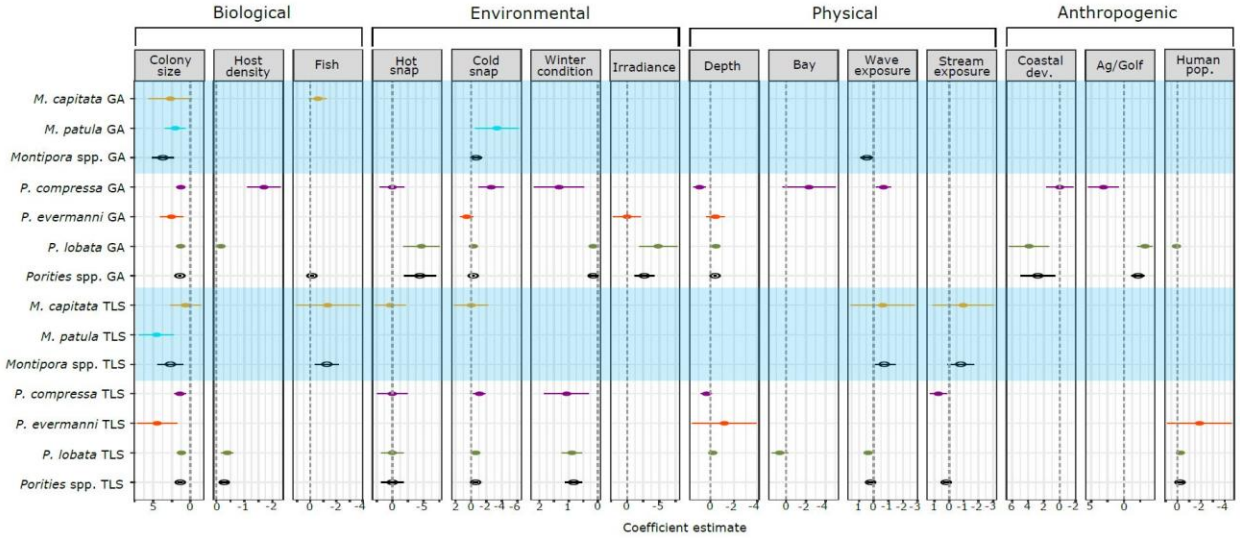

Figure 1: Effect size of predictor variables by host-disease pair. Average coefficient values based on model runs across 500 case-control datasets for all host-disease pairs. Species-specific model coefficients are shown in filled circles while genera model coefficients are shown in black open circles. Lines show the average standard error for the coefficients based on model runs across 500 case-control datasets (very small standard error values may not be visible; very large standard error values indicate detection of an effect and its magnitude depended heavily on the subset of training data). Coefficient values and standard error lines are colored based on host species. GA refers to growth anomalies and TLS refers to tissue loss.

For tissue loss diseases, we identified eleven risk factors associated with disease occurrence (seven associated with *Montipora* species; nine associated with *Porites* species), including biological (colony size and host density), environmental (hot snap, cold snap, and winter condition), physical (depth, embayment, wave exposure, and stream exposure), and anthropogenic (human population size) risk factors (Fig. 1). We identified six risk factors for *Montipora capitata* and one risk factor for *Montipora patula,*

with colony size important for both species (Fig. 1). Six risk factors were associated with *Porites compressa*, three with *Porites evermanni*, and nine with *Porites lobata*; colony size and depth were important for all three *Porites* species (Fig. 1). Of eleven risk factors associated with tissue loss, colony size was the only risk factor associated with all species (Fig. 1); five risk factors (stream exposure, wave exposure, cold snap, hot snap, and colony size) were associated with at least one species in each genus and three factors (host density, herbivorous fish biomass, and embayment) were exclusively associated with one host species (*P. lobata, M. capitata,* and *P. compressa,* respectively) (Fig. 1). For tissue loss, risk factors were not consistent among species in either genus.

Of the 13 risk factors included in our candidate models for each species-specific host-disease pair (Table 1), the only factor that was consistently associated with disease risk was colony size (Fig. 1). Disease risk was higher in larger colonies for all host-disease pairs (Fig. 1-2). Disease risk increased by 3-12%, depending on host-disease pair, for every 10 cm increase in colony size, based on species-specific average colony size (*Montipora capitata* = 16 cm; *Montipora patula* = 19 cm; *Porites compressa* = 23 cm; *Porites evermanni* = 22 cm; *Porites lobata* = 15 cm). Fig. S1 shows individual plots for each host-disease pair with data points overlaid for one randomly sampled case-control training dataset.

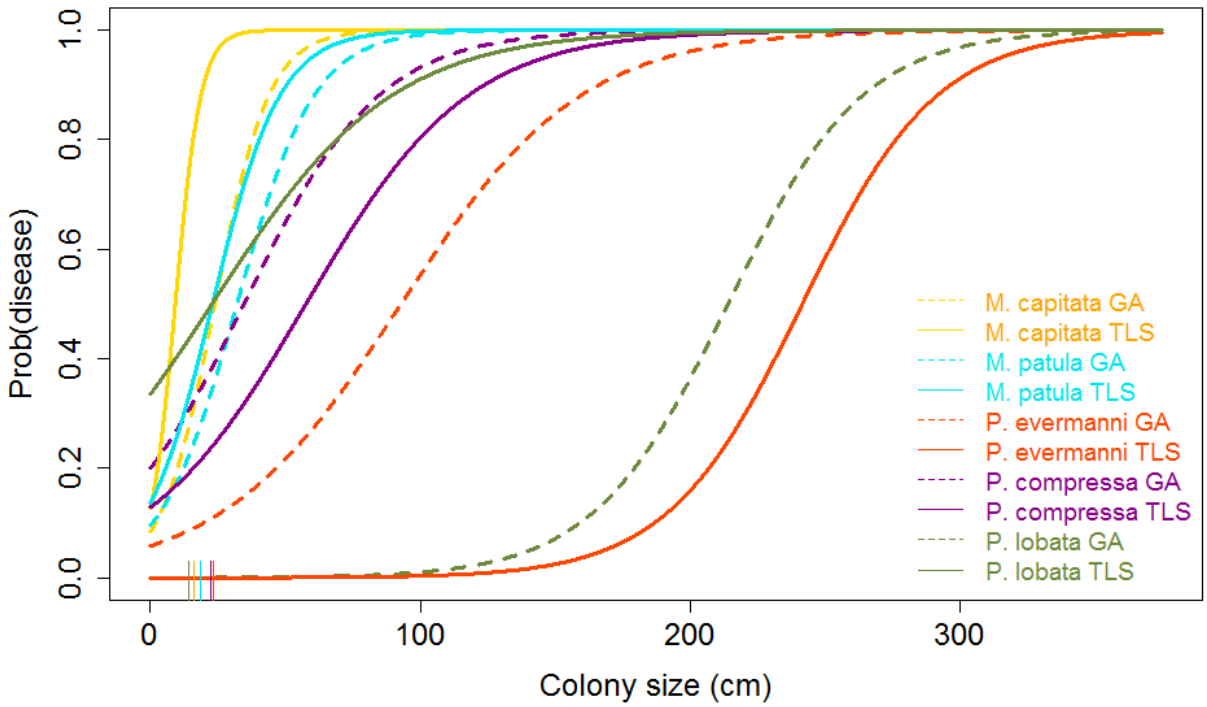

Figure 2: Disease risk increases with colony size for all host-disease pairs. Dashed lines indicate relationships for growth anomalies (GA) and solid lines indicate relationships for tissue loss (TLS). Lines are colored by host species. Average colony size is indicated along the horizontal axis.

Differences between ecological factors associated with species and genera models

While the predictive accuracy between species and genera models did not differ substantially across disease types (Table 3), several of the ecological risk factors identified in the species-specific models differed from those included in the genus level models (Fig. 1). For growth anomalies, the *Montipora* genus model (pooled data from *M. capitata* and *M. patula*) had two risk factors (colony size and cold snap) in common with the species-specific models and one risk factor (wave exposure) that was not

included in either species-specific model (Fig. 1). The *Porites* genus model (pooled data from *P. compressa, P. evermanni,* and *P. lobata*) had eight risk factors (colony size, host snap, cold snap, winter condition, irradiance, depth, coastal development and agricultural and golf runoff) in common with the species-specific models, one risk factor (herbivorous fish abundance) that was included in the genus model but not included in any species-specific model, and one risk factor (host density) that was absent from the genus model but included in species-specific models (Fig. 1). For tissue loss disease, the *Montipora* genus model overlapped for four risk factors (colony size, herbivorous fish abundance, wave exposure, and stream exposure) whereas the *Porites* genus model for tissue loss had eight risk factors (colony size, host density, hot snap, cold snap, winter condition, wave exposure, stream exposure, and human population size) in common with the species-specific models and one risk factor (depth) absent from the genus model but included in the species-specific models (Fig. 1).

Discussion

Understanding the ecological conditions that promote endemic diseases with low prevalence and for epidemic diseases during interepidemic periods is central for understanding disease dynamics but can be difficult to investigate using common biostatistical approaches. In this study, we adopted statistical methods better suited towards highly skewed occurrence data to investigate ecological correlates of several coral diseases. We show that the case-control study design improved our ability to predict diseased and healthy colony observations with relatively high specificity given a suite of ecological conditions. In contrast, a more traditional biostatistical approach

yielded high overall predictive accuracy (heavily biased by successful predictions of healthy state) but often failed to accurately predict the events of interest: disease occurrences. This study highlights the usefulness of the case-control design for ecological questions using infrequent occurrence data. In this study, we identified biological, environmental, physical, and anthropogenic drivers that promote low levels of coral disease and explored how those factors differ across disease types and host species. These results provide insight into factors that should be explored further through experimental and observational studies and that may be favorable for intervention strategies.

Among the suite of ecological risk factors that we included in our study, we found that disease risk was higher for larger colonies across all host-disease pairs. A positive size-disease relationship has been found in other coral disease studies (Caldwell, Donahue, & Harvell, 2018; Couch et al., 2014; Jolles, Sullivan, Alker, & Harvell, 2002) and is likely associated with exposure area or duration of exposure to potential disease causing agents (biotic or abiotic), which would be particularly relevant for non-infectious diseases. Further, a positive relationship between disease prevalence and coral cover has been well established for infectious diseases like tissue loss that are more easily transmitted among a higher density of susceptible hosts [16,32–34], potentially because coral size-frequency distribution is directly related to coral cover (Bak & Meesters, 1998; Meesters et al., 2001). Coral cover was not consistently available in the dataset used for this study. Although we found a positive linear relationship between disease risk and size within the range of 0 to 300 cm, it is possible that disease risk saturates or is

reduced above some species-specific size threshold (Fig. 2). Regardless of what the functional form of this relationship looks like across the entire range of colony sizes, the consistent importance of size across different host-disease pairs provides biological evidence for evaluating disease risk at the colony scale (also supported by simulation models, Jordán-Dahlgren, Jordán-Garza, & Rodríguez-Martínez, 2018) rather than evaluating prevalence at the transect or site scale as is currently common practice. Such probabilistic models of colony health state could easily be integrated into a hierarchical framework to predict local or regional disease prevalence.

In addition to colony size, our findings suggest that human activities in watersheds adjacent to shallow, semi-protected coral reefs promote low levels of chronic disease year-round. Specifically, we found growth anomaly disease risk was associated with coastal development, agriculture and golf runoff, and human population size; however, the direction and magnitude of these relationships differed by host species. Growth anomalies in *Porites compressa* was more common in locations with elevated levels of pesticide and fertilizer runoff from agriculture and golf courses and reduced coastal development, whereas in *Porites lobata,* growth anomalies were more common in locations with higher coastal development and lower impacts from agriculture and golf courses (Fig. 1). The opposing relationships between *P. compressa* and *P. lobata* for agriculture and golf runoff and coastal development likely reflect differences in land use in nearby watersheds rather than nutrient-specific disease responses and support previous research associating disease risk to nutrient enrichment, human populations, and poor water quality (Aeby et al., 2011; Kaczmarsky & Richardson, 2007; Yoshioka,

Kim, Tracy, Most, & Harvell, 2016). In addition, we found that growth anomalies were more common in reefs with lower fish biomass, potentially indicating an indirect relationship to more degraded reefs or fishing intensity and associated impacts. Growth anomalies were also more likely to occur in shallow water, in reefs with low host density (likely because as colony sizes increase, host density decreases), and inside bays and other areas with low wave exposure (possibly because of longer residence times and associated concentration of pollutants). This research indicates that reducing coastal development and high-intensity land use activities such as farming and golf adjacent to semi-protected reefs would be the most effective management approach to break transmission cycles of growth anomalies.

Although growth anomalies are present year-round in Hawaii, our results suggest temperature exposure may explain their dynamic formation and development cycles throughout the year. Previous research showed that the formation of new, small growth anomalies peaks in late summer (September) while growth anomaly development peaks in late spring/early summer (May/June; Stimson, 2010). Our results suggest that growth anomaly risk increases with mild summers and warmer winters, providing further support for these dynamic cycles of formation and growth.

For tissue loss, we found that specific combinations of reef structure, water movement, and temperature exposure differentially promote disease in different host species during interepidemic periods. For example, we found that wave exposure and stream exposure were important risk factors for tissue loss, but the direction and magnitude of these

relationships varied by species. For *Porites lobata,* tissue loss was more common as wave exposure increased whereas we found the opposite relationship for *Montipora capitata,* and there is no evidence that this pattern was driven by differences in habitat associations for each species (Fig. S2-3). One hypothesis that would explain this result is that water movement may be important for transmitting pathogens for *Porites lobata* but more stagnant water where pathogens can proliferate may be more likely to promote disease in *Montipora capitata*. Alternatively, the association with water movement may be related to water diffusion processes as colonies exposed to high water flow minimize the diffusion boundary layer [49] and could therefore limit exposure of pathogens in areas with high water movement and vice versa. Similarly, the probability of tissue loss occurrence increased with higher stream exposure for *Porites compressa* but decreased for *Montipora capitata* (this pattern was also found at the genus level). This result suggests that stream exposure and potentially associated changes in salinity, sedimentation, and/or land-based pollutant runoff may be more important for disease occurrence in *Porites lobata* than for *Montipora capitata*. Contrary to several previous studies on tissue loss in other regions [32–34,50], we did not find a strong relationship between tissue loss and anomalously warm summer temperatures. Instead, for both *Porites compressa* and *Porites lobata*, we found a slightly negative association with anomalously warm thermal stress in the summer but a strong positive association with anomalously warm winter temperatures. Although there are no documented outbreaks of *Porites* tissue loss in Hawaii, this result may shed some light on why several tissue loss outbreaks in *Montipora* in Hawaii occurred in the winter (Aeby et al., 2016b; Caldwell et al., 2018) compared with outbreaks in other regions where outbreaks

occurred in the summer [33,34,53].

The variation in risk factors across host-disease pairs suggests that there may be different pathogens causing the same gross lesions in different species or that the same disease may arise in different species under different ecological conditions. Corals have a limited repertoire of responses to insult; therefore, disease lesions that are grossly identical, such as tissue loss, could be caused by different pathogens of the same general type (e.g., bacterial pathogens), which has been suggested in other studies (e.g., Work, Russell, & Aeby, 2012). Recent research supports this hypothesis as multiple bacterial pathogens have been identified for tissue loss in *Montipora capitata* in Hawaii [55]. Alternatively, the same causative agent(s) could be triggered to infect different species under different conditions. Our results support this hypothesis as we identified important risk factors in the genus models that were not identified in the species-specific models within that genus (and vice versa) for some host-disease pairs. In these instances, we would expect niche differentiation to explain these patterns. For example, we found no relationship between wave exposure and growth anomalies for *Montipora capitata* or *Montipora patula* in the species-specific models, but we found a positive relationship in the *Montipora* genus model. This pattern could arise if growth anomalies in *Montipora capitata* is associated with one wave exposure regime and growth anomalies in *Montipora patula* is associated with a different wave exposure regime: in each case wave exposure would not be identified in the species-specific models because disease risk would be uniform across that specific regime, but in combination in the genus model wave exposure would be identified as important. Our

understanding of coral disease ecology is still limited but our results suggest that there may be specific ecological conditions that support or restrict the spread of specific diseases in each species. Thus, researchers should strongly consider how combining data for multiple species to assess disease risk may impact our understanding of factors that promote transmission.

This study demonstrates the usefulness of the case-control design to study rare events in space and time using existing datasets with low probability of occurrence, but also provides a path forward for more efficiently collecting new data for future investigations. For example, for studies specifically focused on coral disease, we recommend complementing more traditional ecological assessments that characterize overall reef health, such as belt transect surveys and rugosity measurements, with survey efforts that maximize information on disease observations. Using experimental designs to collect data geared towards rare and/or clustered populations would allow us to better investigate the underlying mechanisms that are affecting disease processes. Therefore, for coral diseases, collecting more extensive information such as size, morphology, geographic location, clustering with nearby diseased colonies, severity of infection, area of affected tissue, etc. allows the use of modeling techniques to better explore the ecology of specific diseases. Such survey designs and associated analyses have already been developed and widely used in other fields such as epidemiology, plant pathology, and conservation biology (e.g., case-control design [56], cluster sampling [57,58], line transect surveys [59]) and could be adapted for coral disease and other ecological studies. Rare occurrence or abundance data is very common in ecology and

approaching ecological questions with this statistical design in mind will allow ecologists to more rigorously quantify ecological relationships for infrequent disturbances.

Methods

Coral health observations

We used coral health observations from the Hawaii Coral Disease Database (HICORDIS; Caldwell et al., 2016) and additional surveys collected by the authors following the same methodologies used in HICORDIS (e.g., 10 x 1 m^2 belt transects surveys) resulting in 362,366 coral health observations collected between 2004 and 2015. We separated the data by host (species or genus of interest) and disease type (growth anomalies or tissue loss) and used only observations with available data for all hypothesized predictor variables. The number of diseased and healthy colony observations we used for each host-disease pair are shown in Table 3.

Ecological disease drivers

For each coral health observation, we collected biological, environmental, physical, and anthropogenic predictor data corresponding to the time and location of each observation (Table 2). We selected predictor variables based on relationships described in previous studies and based on personal observations. After removing variables that were highly colinear (see *Testing for multicollinearity* below), our hypothesized predictor variables were: (i) biological variables of colony size, host density, and herbivorous fish biomass; (ii) environmental variables of anomalous temperature (hot snap, cold snap, winter condition), rainfall, chlorophyll-*a* (proxy for water quality/eutrophication), and irradiance;

(iii) physical variables of depth, inside/outside of an embayment, wave exposure, and stream exposure (proxies for potential terrestrial runoff and retention); and (iv) anthropogenic variables of human population size, coastal development, and agriculture and golf runoff. Variables reflect short-term (acute, pulse) or long-term (chronic, press) interactions or site characteristics (Table 2).

For each predictor variable, we used survey data where recorded, or compiled externally-sourced data from the vicinity (and within the watershed) of where the survey was conducted. We used measurements of colony size and depth that were recorded in the coral health surveys and calculated host density from the survey data. We extracted all other predictor variables by survey location and by time (where applicable; i.e., acute interactions) using ArcGIS v 10.4. Measurements of rainfall were land-based, using gridded values from the land location nearest to the survey site that contained data within the appropriate watershed and month; human population size was similarly determined. For chlorophyll-*a*, wave exposure, irradiance, hot snaps, cold snaps, winter condition, and herbivorous fish abundance, we used data from the pixel containing the survey location when possible or else the nearest pixel with available data over the appropriate time period. Coastal development and agricultural and golf runoff were calculated for each ocean pixel as the value from the nearest watershed decayed with distance from shore; therefore, we used the same technique as described above acquiring data from the pixel containing the survey location when possible or else the nearest pixel with available data. We manually determined whether a site was inside or outside of an embayment by visualizing the survey site locations relative to coastal

boundaries downloaded from the State of Hawaii Office of Planning. We calculated stream exposure as the inverse planar distance between the site location and the nearest stream mouth within a given watershed. Higher stream exposure values indicate sites closer to stream mouths. For sites located in watersheds without streams, the stream exposure value was set to zero.

We used a different combination of predictor variables for growth anomalies and tissue loss models because of hypothesized differences in their modes of transmission and etiologies. Thirteen predictor variables were identified for each: eight hypothesized predictor variables were common to both diseases and five predictor variables were unique to a disease type (Table 1).

Variance Components Analysis

Since the coral health data is naturally hierarchical, we conducted a variance components analysis (VCA) to determine the variation associated with the different spatial and temporal scales included in the dataset and to determine if we needed to include such factors in our models. We calculated variance associated with four spatial scales: 1) site, where variation among sites would be associated with local, watershed variation; 2) coast, where variation among coasts would be associated with differences in oceanography and weather (e.g., windward vs. leeward coasts); 3) island, where variation among islands would be associated with island geology, latitude, human population size, and development, and 4) region, where variation would be associated with geology (high versus low islands), marine protection status, and whether or not the

region is inhabited by people (populated Main Hawaiian Islands versus uninhabited Northwestern Hawaiian Islands). In addition, we calculated variance associated with month and year of survey. We conducted this analysis using the anovaVCA function in the VCA package in R statistical software v 3.5.1. Most variation in the coral health data was associated with site and year with negligible variance explained at the other spatial and temporal scales for all host-disease pairs (Table S1). Therefore, we initially included site and year as random effects in our models.

Testing for multicollinearity

To test for multicollinearity among predictor variables, we calculated the variance inflation factor. The variance inflation factor (VIF) is a ratio that quantifies the amount of multicollinearity among covariates in a regression model by dividing the variance of a model with multiple terms by the variance of a model with a single term. A VIF equal to one indicates a variable is not correlated with other variables, a value between one and five indicates moderate correlation, and a value above five indicates high correlation. We initially hypothesized that, in addition to the predictor variables listed in Table 1, total coral density, month, and site should be included as covariates for both growth anomalies and tissue loss models, and that an additional metric of urban runoff should be included in the growth anomalies model. However, these four metrics had VIF values greater than five indicating strong multicollinearity with other covariates (Fig. S4), so we excluded these predictor variables from our candidate models.

Logistic regression without the case-control design

We investigated the ecological drivers of each host-disease pair using logistic regression with all available coral health data (Tables 1-3) and assessed model predictive accuracy and true positive rates based on withheld portions of data (see below). The response variable in our logistic regression models described the colony observation health state as 0 for healthy and 1 for diseased. We included the identified 13 predictor variables listed in Table 1 (combination dependent on disease type) for the full models and conducted backward model selection to find the best fit model based on the lowest Akaike Information Criterion (AIC) value using the stepAIC function in the Mass package in R statistical software v 3.5.1. We used 80% of the data for model development and model selection, and the remaining 20% of data to quantify model predictive accuracy (proportion of total colony observations correctly identified as healthy and diseased) and true positive rate (proportion of diseased colony observations correctly identified as diseased). We split the data using the sample function in base R, which selects data in proportion to the raw data (e.g., if the full data has a zero-to-one ratio of 90%-to-10%, then the sample will also contain 90% zeros and 10% ones).

To ensure that the randomly sampled training and testing data did not bias our results, we repeated our model development and selection and accuracy assessments 500 times. Therefore, for each host-disease pair, we split the full data into 500 paired training and testing datasets, conducted model selection within each training dataset, and then used the most frequently selected best fit model (Table S2) to predict health

state on the 500 withheld test datasets. We compared predicted health state with observed health state to calculate accuracy and true positive rates within each withheld test dataset and report mean values across all 500 withheld test datasets.

Logistic regression with the case-control design

For the logistic regression analysis with the case-control design, we followed the approach described above except instead of creating 500 randomly selected training and testing datasets from all available coral health data for each host-disease pair, we created 500 case-control datasets each with matching numbers of diseased and healthy observations and randomly split those case-control datasets into paired training data and testing data (80% training/20% testing). To create each case-control dataset, we selected all diseased colony observations for a given host-disease pair, and for each diseased colony observation we randomly selected a healthy colony observation of the same host species surveyed in the same month (but not necessarily year) and from the same island as the diseased colony observation. We repeated this process to create 500 case-control datasets per host-disease pair that all had the same disease observations but varied in the paired healthy colony observations. We followed the same process of model development, selection, and testing as described above. Results for the frequency of model selection for the best fit model are available in Table S2.

Acknowledgements: We thank the Fore-C research team for valuable feedback on analyses and the following funding sources for this research: NASA Roses Ecological

Forecasting grant NNX17AI21G (MJD, SFH, JMC), NASA Earth and Space Science Fellowship (JMC, MJD), and Stanford University Woods Institute for the Environment Environmental Ventures Program (JMC). The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author and do not necessarily reflect the views of NOAA or the Department of Commerce.

Author contributions: All authors conceived of project, interpreted analyses, and edited and approved the manuscript. JMC conducted analyses and wrote the manuscript.

Competing interests: The authors declare no competing interests in relation to the work described.

References

1. Anderson, R. M. & May, R. M. Population biology of infectious disease: Part I. *Nature* **280**, 361–367 (1979).
2. May, R. M. & Anderson, R. M. Population biology of infectious diseases: Part II. *Nature* **280**, 455–461 (1979).
3. Hudson, P. & Greenman, J. Competition mediated by parasites: biological and theoretical progress. *Trends Ecol. Evol.* **13**, 387–390 (1998).
4. Kohler, S. L. & Hoiland, W. K. Population regulation in an aquatic insect: the role of disease. *Ecology* **82**, 2294–2305 (2001).
5. MacNeil, C. *et al.* Parasite-mediated predation between native and invasive

amphipods. *Proc. R. Soc. London. Ser. B Biol. Sci.* **270**, 1309–1314 (2003).

6. Lafferty, K. D. & Morris, A. K. Altered behavior of parasitized killifish increases susceptibility to predation by bird final hosts. *Ecology* **77**, 1390–1397 (1996).

7. Holdo, R. M. *et al.* A disease-mediated trophic cascade in the Serengeti and its implications for ecosystem C. *PLoS Biol.* **7**, e1000210 (2009).

8. Brunner, F. S., Anaya-Rojas, J. M., Matthews, B. & Eizaguirre, C. Experimental evidence that parasites drive eco-evolutionary feedbacks. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 3678–3683 (2017).

9. Behringer, D. C., Butler, M. J. & Shields, J. D. Avoidance of disease by social lobsters. *Nature* **441**, 421–421 (2006).

10. Woodworth, B. L. *et al.* Host population persistence in the face of introduced vector-borne diseases: Hawaii amakihi and avian malaria. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1531–6 (2005).

11. Aeby, G. S. Behavioral and ecological relationships of a parasite and its hosts within a coral reef system. *Pacific Sci.* **45**, 263–269 (1991).

12. King, G. & Zeng, L. Logistic Regression in Rare Events Data. *Polit. Anal.* **9**, 137–163 (2001).

13. Lewallen, S. & Courtright, P. Epidemiology in practice: case-control studies. *Community eye Heal.* **11**, 57–8 (1998).

14. Sutherland, K. P., Porter, J. W. & Torres, C. Disease and immunity in Caribbean and Indo-Pacific zooxanthellate corals. *Mar. Ecol. Prog. Ser.* **266**, 273–302 (2004).

15. Stimson, J. Ecological characterization of coral growth anomalies on Porites

compressa in Hawai‘i. *Coral Reefs* **30**, 133–142 (2010).

16. Ruiz-Moreno, D. *et al.* Global coral disease prevalence associated with sea temperature anomalies and local factors. *Dis. Aquat. Organ.* **100**, 249–261 (2012).

17. Palmer, C. V. & Baird, A. H. Coral tumor-like growth anomalies induce an immune response and reduce fecundity. *Dis. Aquat. Organ.* **130**, 77–81 (2018).

18. Couch, C. S. *et al.* Spatial and temporal patterns of coral health and disease along leeward Hawai’i Island. *Coral Reefs* **33**, 693–704 (2014).

19. Kaczmarsky, L. & Richardson, L. L. Do elevated nutrients and organic carbon on Philippine reefs increase the prevalence of coral disease? *Coral Reefs* **30**, 253–257 (2011).

20. Aeby, G. S. *et al.* Growth anomalies on the coral genera Acropora and Porites are strongly associated with host density and human population size across the Indo-Pacific. *PLoS One* **6**, e16887 (2011).

21. Williams, G. J., Aeby, G. S., Cowie, R. O. M. & Davy, S. K. Predictive modeling of coral disease distribution within a reef system. *PLoS One* **5**, e9264 (2010).

22. Aeby, G. S. *et al.* Emerging coral diseases in Kane‘ohe Bay, O‘ahu, Hawai‘i (USA): two major disease outbreaks of acute Montipora white syndrome. *Dis. Aquat. Organ.* **119**, 189–198 (2016).

23. Hobbs, J.-P. A., Frisch, A. J., Newman, S. J. & Wakefield, C. B. Selective Impact of Disease on Coral Communities: Outbreak of White Syndrome Causes Significant Total Mortality of Acropora Plate Corals. *PLoS One* **10**, e0132528 (2015).

24. Precht, W. F. *et al.* Unprecedented Disease-Related Coral Mortality in Southeastern Florida. *Sci. Rep.* **6**, 31374 (2016).

25. Roff, G. *et al.* The ecology of 'Acroporid white syndrome', a coral disease from the southern Great Barrier Reef. *PLoS One* **6**, e26829 (2011).

26. Walton, C. J., Hayes, N. K. & Gilliam, D. S. Impacts of a Regional, Multi-Year, Multi-Species Coral Disease Outbreak in Southeast Florida. *Front. Mar. Sci.* **5**, (2018).

27. Brodnicke, O. B. *et al.* Unravelling the links between heat stress, bleaching and disease: fate of tabular corals following a combined disease and bleaching event. *Coral Reefs* 1–13 (2019). doi:10.1007/s00338-019-01813-9

28. Sussman, M., Willis, B. L., Victor, S. & Bourne, D. G. Coral pathogens identified for White Syndrome (WS) epizootics in the Indo-Pacific. *PLoS One* **3**, e2393 (2008).

29. Ben-Haim, Y. *et al.* Vibrio coralliilyticus sp. nov., a temperature-dependent pathogen of the coral Pocillopora damicornis. *Int. J. Syst. Evol. Microbiol.* **53**, 309–315 (2003).

30. Ushijima, B. *et al.* Vibrio coralliilyticus Strain OCN008 is an etiological agent of acute Montipora White Syndrome. *Appl. Environ. Microbiol.* **80**, (2014).

31. Ushijima, B. *et al.* Mutation of the toxR or mshA genes from Vibrio coralliilyticus strain OCN014 reduces infection of the coral Acropora cytherea. *Environ. Microbiol.* **18**, 4055–4067 (2016).

32. Bruno, J. F. *et al.* Thermal stress and coral cover as drivers of coral disease outbreaks. *PLoS Biol.* **5**, e124 (2007).

33. Heron, S. F. *et al.* Summer hot snaps and winter conditions: modelling white syndrome outbreaks on Great Barrier Reef corals. *PLoS One* **5**, e12210 (2010).

34. Maynard, J. A. *et al.* Predicting outbreaks of a climate-driven coral disease in the Great Barrier Reef. *Coral Reefs* **30**, 485–495 (2011).

35. Randall, C. J., Jordan-Garza, A. G., Muller, E. M. & Van Woesik, R. Relationships between the history of thermal stress and the relative risk of diseases of Caribbean corals. *Ecology* **95**, 1981–1994 (2014).

36. Caldwell, J. M., Heron, S. F., Eakin, C. M. & Donahue, M. J. Satellite SST-Based Coral Disease Outbreak Predictions for the Hawaiian Archipelago. *Remote Sens.* **8**, 93 (2016).

37. Heenan, A. *et al. Ecological Monitoring 2012-2013 - reef fishes and benthic habitats of the main Hawaiian Islands, American Samoa, and Pacific Remote Island Areas*. (2014). doi:10.13140/RG.2.1.4351.9122

38. Wedding, L. M. *et al.* Advancing the integration of spatial data to map human and natural drivers on coral reefs. *PLoS One* **13**, e0189792 (2018).

39. Giambelluca, T.W. *et al.* Online Rainfall Atlas of Hawai‘i. *Bull. Amer. Meteor. Soc.* **94**, 313–316 (2013).

40. Frazier, A. G., Giambelluca, T. W., Diaz, H. F. & Needham, H. L. Comparison of geostatistical approaches to spatially interpolate month-year rainfall for the Hawaiian Islands. *Int. J. Climatol.* **36**, 1459–1470 (2016).

41. Lecky, J. H. Ecosystem Vulnerability and Mapping Cumulative Impacts on Hawaiian Reefs. (2016). doi:10.13140/RG.2.2.33194.52162

42. Jolles, A. E., Sullivan, P., Alker, A. P. & Harvell, C. D. Disease Transmission of

Aspergillosis in Sea Fans: Inferring Process from Spatial Pattern. *Ecology* **83**, 2373–2378 (2002).

43. Caldwell, J. M., Donahue, M. J. & Harvell, C. D. Host size and proximity to diseased neighbours drive the spread of a coral disease outbreak in Hawai'i. doi:10.1098/rspb.2017.2265

44. Bak, R. & Meesters, E. Coral population structure:the hidden information of colony size-frequency distributions. *Mar. Ecol. Prog. Ser.* **162**, 301–306 (1998).

45. Meesters, E. H. *et al.* Colony size-frequency distributions of scleractinian coral populations: spatial and interspecific variation. *Marine Ecology Progress Series* **209**, 43–54

46. Jordán-Dahlgren, E., Jordán-Garza, A. G. & Rodríguez-Martínez, R. E. Coral disease prevalence estimation and sampling design. *PeerJ* **6**, e6006 (2018).

47. Kaczmarsky, L. & Richardson, L. L. Transmission of growth anomalies between Indo-Pacific Porites corals. *J. Invertebr. Pathol.* **94**, 218–221 (2007).

48. Yoshioka, R. M., Kim, C. J. S., Tracy, A. M., Most, R. & Harvell, C. D. Linking sewage pollution and water quality to spatial patterns of Porites lobata growth anomalies in Puako, Hawaii. *Mar. Pollut. Bull.* (2016). doi:10.1016/j.marpolbul.2016.01.002

49. Lesser, M. P., Weis, V. M., Patterson, M. R. & Jokiel, P. L. Effects of morphology and water motion on carbon delivery and productivity in the reef coral, Pocillopora damicornis (Linnaeus): Diffusion barriers, inorganic carbon limitation, and biochemical plasticity. *J. Exp. Mar. Bio. Ecol.* **178**, 153–179 (1994).

50. Randall, C. J. & Van Woesik, R. Contemporary white-band disease in Caribbean

corals driven by climate change. *Nat. Clim. Chang.* **5**, 375–379 (2015).

51. Ross, M., Stender, Y., White, D. & Aeby, G. S. Outbreak of the coral disease, Montipora White Syndrome in Maui, Hawai'i. in 9–13 (2012).

52. Aeby, G. S. *et al.* Emerging coral diseases in Kaneohe Bay, Oahu, Hawaii (USA): two major disease outbreaks of acute Montipora white syndreom. *Dis Aquat Org* **119**, 189–198 (2016).

53. Jones, R., Bowyer, J., Hoegh-Guldberg, O. & Blackall, L. L. Dynamics of a temperature-related coral disease outbreak. *Mar. Ecol. Prog. Ser.* **281**, 63–77 (2004).

54. Work, T. M., Russell, R. & Aeby, G. S. Tissue loss (white syndrome) in the coral Montipora capitata is a dynamic disease with multiple host responses and potential causes. *Proc. R. Soc. B Biol. Sci.* **279**, 4334–4341 (2012).

55. Beurmann, S. *et al.* Pseudoalteromonas piratica strain OCN003 is a coral pathogen that causes a switch from chronic to acute Montipora white syndrome in Montipora capitata. *PLoS One* **12**, e0188319 (2017).

56. Breslow, N. E. Statistics in Epidemiology: The Case-Control Study. *J. Am. Stat. Assoc.* **91**, 14–28 (1996).

57. Venette, R. C., Moon, R. D. & Hutchison, W. D. Strategies and Statistics of Sampling for Rare Individuals. *Annu. Rev. Entomol.* **47**, 143–174 (2002).

58. Balmaseda, A. *et al.* Index Cluster Study of Dengue Virus Infection in Nicaragua. *Am. J. Trop. Med. Hyg.* **83**, 683–689 (2010).

59. Burnham, K., Anderson, D., & Laake, J. Estimation of Density from Line Transect Sampling of Biological Populations. *Wildl. Monogr.* **72**, 3–202 (1980).

60. Caldwell, J. M. *et al.* Hawai‘i Coral Disease database (HICORDIS): species-specific coral health data from across the Hawaiian archipelago. *Data in Brief* **8**, 1054–1058 (2016).