Advancing reef coral diagnostic capabilities using molecular biotechnology and artificial intelligence

Advancing reef coral diagnostic capabilities using molecular biotechnology and artificial intelligence

A B Mayfield

National Museum of Marine Biology and Aquarium, 2 Houwan Rd., Checheng, Pingtung 944, Taiwan

andersonblairmayfield@gmail.com

Abstract. Coral reef ecosystems around the planet are threatened by an onslaught of anthropogenic stressors, most notably global climate change (GCC); indeed, no regions have been spared from our wide-ranging human impact. Consequently, there has been an urgent push to 1) model how marine organisms will respond to future changes in their environments and 2) make data-driven predictions as to which populations are most stress sensitive. Given our recently elevated level of understanding of the responses of reef-building corals to environmental change and GCC, we are now in a position in which it may be possible to make projections as to which corals are most susceptible to GCC, as well as which will likely demonstrate resilience. Herein I explore the potential for data-trained predictive modeling approaches based on artificial intelligence to generate models that can accurately predict coral stress susceptibility (CSS). Specifically, I advocate that coral reef-focused partial least squares and neural networking algorithms (trained with either molecular or environmental data) should be developed, with their prognostic capability then field-tested at sites that span a gradient of human impact and ecological resilience in the high-biodiversity "Coral Triangle." If the developed predictive models are characterized by the analytical capacity to forecast CSS, we will possess one means of identifying reefs that should be prioritized for conservation in this era of rapidly changing global climate.

1. Introduction

Coral reefs harbor immense biodiversity and provide a wealth of benefits to humankind (e.g., as nurseries for numerous commercial fish species). Unfortunately, these ecosystems are threatened by myriad anthropogenic stressors [1], from global-scale impacts like climate change [2] to local ones like seawater pollution [3]. The elevated temperatures associated with global climate change (GCC) are especially concerning since most corals live near the upper threshold of their thermotolerance [4]. Even increases in temperature of only 1ºC above the summer mean can cause a collapse of the mutualistic relationship between reef-building scleractinians and the photosynthetically active dinoflagellates (genus *Symbiodinium*) that inhabit their gastroderms; this phenomenon is known as "bleaching" due to the paling of the coral tissues [5]. Since corals rely on the energy obtained from *Symbiodinium*-fixed carbon to not only meet their metabolic needs, but also to accrete the calcium carbonate skeletons that serve as the foundation of coral reefs, bleaching can lead to both coral death and reef degradation.

Upon having carried out a plethora of controlled laboratory studies with several Indo-Pacific coral species, including *Seriatopora hystrix* (e.g., [6]) and *Pocillopora acuta* (Table 1; e.g., [7-8]), we have

developed a relatively sophisticated understanding of the responses of these scleractinians to GCC scenarios (though see deficiencies below.). In fact, our knowledge is finally to the point where it would be fruitful to attempt to use the explanatory data acquired during such tank studies to predict coral stress susceptibility (CSS), or, more generally, behavior/physiological status *in situ*; if we could use data from laboratory exposures (*sensu* Table 1) and/or published field datasets on coral physiology (e.g., [9-10]) to make predictions as to how conspecifics would respond to environmental heterogeneity in other locations (or in the same site at later dates), then we would likewise possess the capacity to determine which reefs (and/or coral populations) are most likely to persist in the face of GCC.

Although it is true that many corals bleach when exposed to elevated temperatures over prolonged durations, others have demonstrated a marked capacity for resilience [11-12], and numerous investigators around the globe, myself included, are currently attempting to elucidate the genetic basis of such thermotolerance. Indeed, we do not yet have a grasp of the cellular pathways underlying the coral bleaching response [13]. In an ideal world, we would carefully uncover the genetic underpinnings of bleaching resilience, or lack thereof, in replicated coral populations from diverse locations across the globe. Alongside such studies, we would molecularly model the physiological mechanisms of coral bleaching at the single-cell level, and such experiments would be repeated with corals sampled from reef sites spanning large-scale latitudinal gradients that are either 1) bleaching-prone or 2) bleaching resistant. Several years may be required to acquire such data, and, unfortunately, time is not on our side given the rate at which seawater temperature is rising. I therefore propose that we not only continue to remedy these deficiencies in our knowledge of coral bleaching and the coral stress response, but also attempt to use the data in hand to make preliminary predictions about future coral behavior/CSS. I present one such approach for doing so in the following paragraph, with details found further on.

Upon analyzing the "proteomes" (population of all synthesized proteins) of corals that resisted bleaching (Table 1-experiment#1), as well as those that instead succumbed to high-temperature stress and bleached (Table 1-experiment#2), I propose to develop both "bottom-up" (molecules=>physiology) and "top-down" (environment=>physiology) predictive models for gauging future coral performance *in situ* using data from not only the aforementioned GCC manipulation studies (Table 1) carried out at the National Museum of Marine Biology and Aquarium's (NMMBA) state-of-the-art coral reef mesocosm facility in Southern Taiwan, but also from the most wide-ranging coral reef survey ever undertaken: the Khaled bin Sultan Living Oceans Foundation's "Global Reef Expedition" (GRE; see [14-16] for details.). Upon incorporating all environmental and coral molecular-physiological data acquired during such studies into the statistical software package JMP® Pro (ver. 14; Cary, NC, USA; www.jmp.com), artificial intelligence (AI)-based partial least squares (PLS) and neural networking (NN) models will be developed, and the predictive capacity of the resulting bottom-up and top-down algorithms will be field tested at four well-studied reef sites in Southern Taiwan that differ dramatically in 1) oceanography and 2) the environmental resilience of the resident coral communities (Figure 1a-b). If AI can be used to identify coral colonies, or coral reefs in the case of the top-down model, that are markedly stress sensitive prior to visible, late-stage manifestations of coral health decline, then managers could be alerted such that they could mitigate local-scale stressors (e.g., overfishing) in order to promote coral resilience and potentially thwart bleaching. Such data-based adaptive management holds particular promise in small island nations like Taiwan, where the populace values both scientific research and marine conservation.

2. Approach

To gauge the coral response to environmental change using AI, two different models will be built using a NN/machine learning approach in conjunction with PLS. In the first model (Aim I-referred to as "bottom-up" since it will utilize molecular information ["bottom"] to make physiological ["up"] conjectures), the predictors will be biomarker proteins uncovered from proteomic analysis of experimentally stressed aquarium samples (Table 1-experiments# 1-2; described in detail below). The

coral response will be binomial: stressed (i.e., bleaching-prone) or unstressed (bleaching-resistant). PLS and NN analyses are particularly well suited for these sorts of "omics" analyses, in which the concentrations of thousands of proteins are quantified simultaneously, and both analytical techniques permit response variable reduction to where only those proteins whose concentrations are tightly linked with coral health decline will be incorporated into the final AI training model.

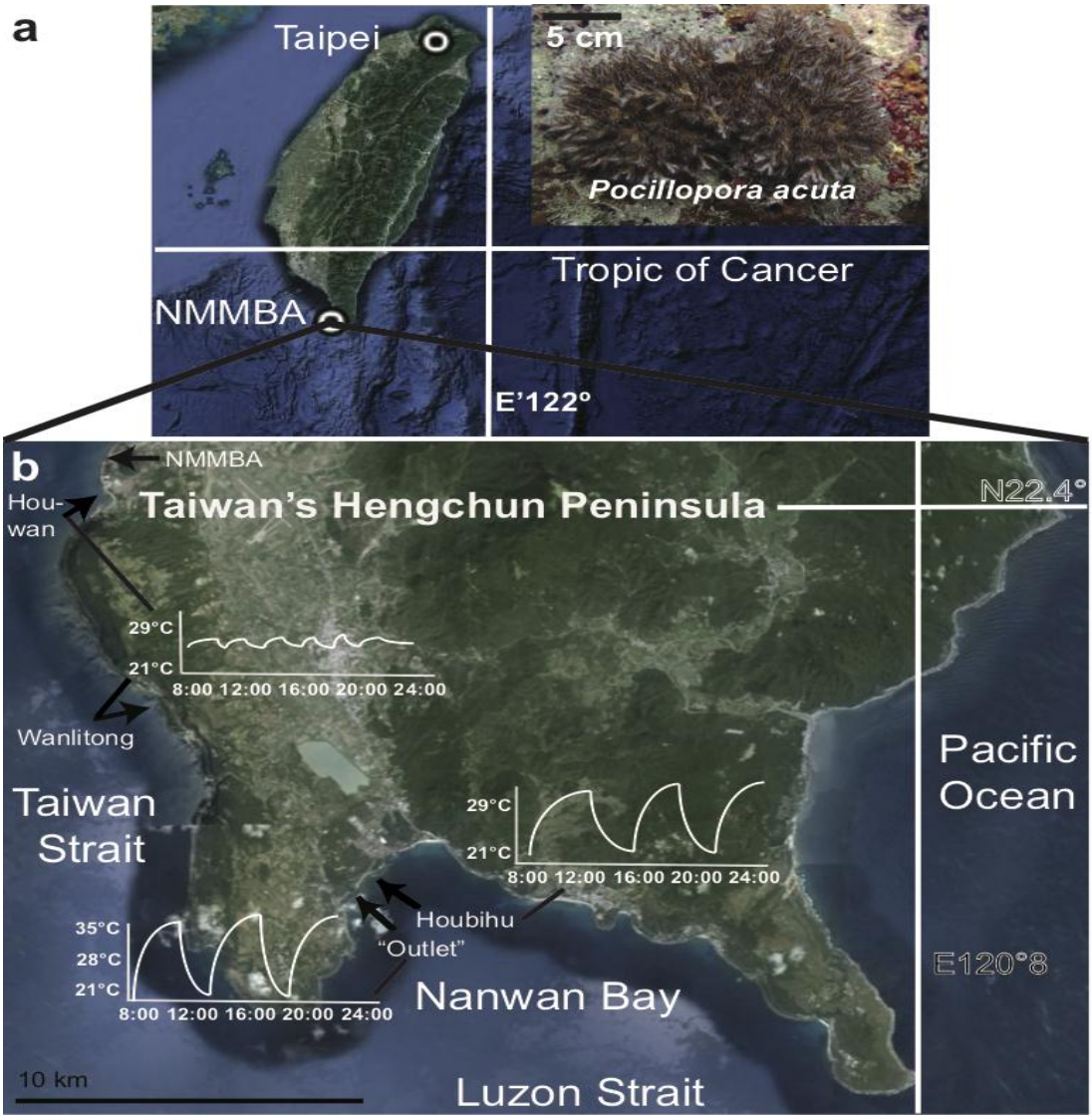
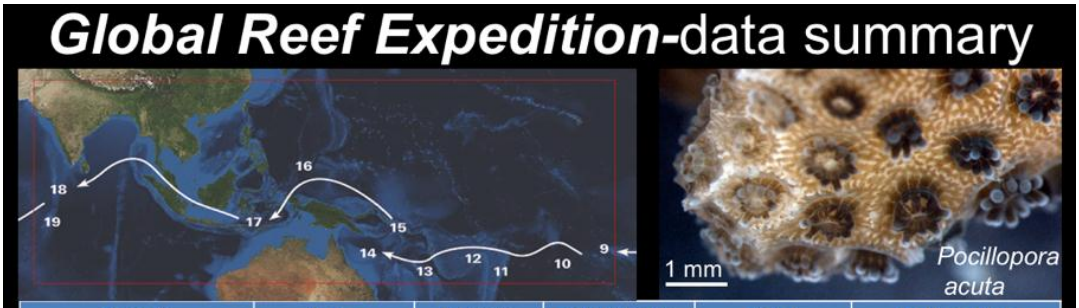


Figure 1. Map of study sites (a-b) and image of the model coral for research, *Pocillopora acuta* (inset within [a]). Taiwan's location at the northern-most region of the "Coral Triangle" (a) has been expanded in (b) to show the study area in more detail.

Nanwan Bay, where the "Outlet" and Houbihu are located, is characterized by summer upwelling of cold, deep-ocean seawater [33]; corals there consequently have special adaptations for dealing with not only elevated (see in text description.), but also highly variable, temperature regimes (22-29ºC over a day; see temperature profile insets.). In contrast, the two Taiwan Strait sites are characterized by a more typical coral reef temperature profile (22-29ºC over a year), nor are these corals routinely exposed to dramatically elevated seawater temperature; they may, then, be more susceptible to environmental change.

I will also build PLS+NN models in an "ecosystem ('top') to physiology ('down')" direction (Aim II); instead of using molecular data to make inferences about future declines in coral performance (*sensu* Aim I), I will exploit reef coral and coral reef datasets to verify if we can use aquarium simulation and environmental data, respectively, to predict which corals and reefs, respectively, will be most stress prone. The first data source will be my extensive series of GCC simulation studies carried out with *P. acuta* at NMMBA (Table 1). Secondly, data from the GRE (Figure 2) will be incorporated; as *P. acuta* was sampled at hundreds of reefs across the Indo-Pacific as part of this monumental marine survey, I have the opportunity to identify the environmental parameters (EP) that are most influential in driving variation in coral physiology when combining these GRE data with those of the controlled GCC simulations carried out in Taiwan. These EP (i.e., the models' predictors) include, but are not limited to, biotic parameters like coral cover, coral diversity, and algal abundance, and abiotic parameters like temperature and salinity (see Figure 3 and [17] for details.). The population densities of the settlements nearest the GRE field sites will also be included in the models to determine whether reefs nearby human habitations are more stress-sensitive than remote ones. Although a variety of both physiological and molecular response variables were assessed in corals of these experiments and field surveys (Fig. 4), there will be a particular focus on those PLS+NN models that best predict, more simply, coral survival.



Country	Notes	Environ-mental data	Coral physio-logical data	Coral molecular data	Deliverables (to date)
Taiwan-mainland	Efforts ongoing	✓	✓	✓	Manuscripts[a-b]
Taiwan-South China Sea	Efforts ongoing	✓	✓	in prep.	Field report only
French Polynesia		✓	✓	✓	Manuscripts[a]
Cook Islands		✓	✓	✓	Manuscripts[a]
Fiji		✓	✓	✓	Manuscript[a]
Tonga		✓	✓	✓	Manuscript[a]
New Caledonia		✓	✓	✓	Manuscript[a-b]
Australia (GBR)		✓	No samples taken		Manuscript[a]
Solomon Islands		✓	✓	✓	Data not yet analyzed
Palau		✓	✓	✓	Data not yet analyzed
Chagos/BIOT		✓	✓	in prep.	Data not yet acquired
Maldives		✓	✓	in prep.	Field report only
Philippines	Proposal under review				Preliminary data

[a]Data analyzed incompletely. [b]Only a portion of the dataset was analyzed.

Figure 2. A summary of data acquired during the Khaled bin Sultan Living Ocean Foundation's "Global Reef Expedition," alongside an inset featuring a dissecting microscope image of the model reef-building coral *Pocillopora acuta.* I am currently attempting to secure permits to sample corals in the Philippines and Indonesia, the two nations in possession of the world's most biodiverse coral reefs. BIOT=British Indian Ocean Territory. in prep.=in preparation (sample processing incomplete).

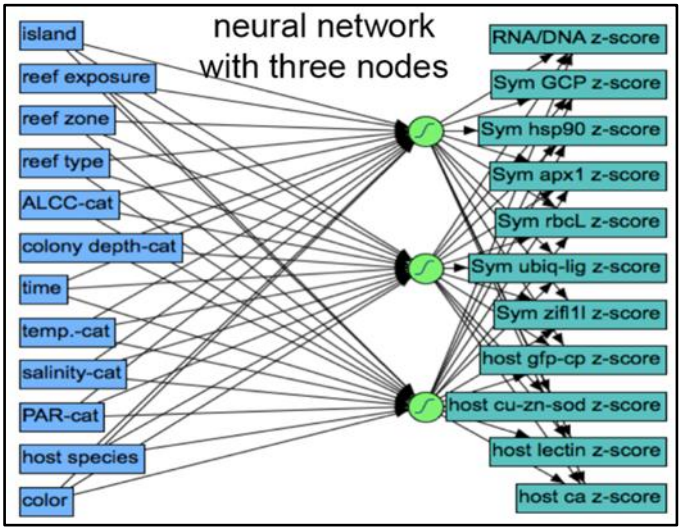


Figure 3. An example of a neural network featuring coral reef environmental data (left side) and reef coral response variables (right side). Gene expression levels, as well as data other response variable data, must be standardized (i.e., converted to *z*-scores) prior to analysis. In the case of this example, the predictive capacity of the model was <30% due to having trained it with a dataset featuring only 70 samples from Fiji. ALCC=average live coral cover. cat=categorical environmental parameter. Please see the works cited in Table 1 for full gene mRNA names.

Table 1. Global climate change manipulation studies carried out at Taiwan's National Museum of Marine Biology and Aquarium. In most cases, *Pocillopora acuta* was inadvertently classified as *P. damicornis* in the published manuscripts. It is worth mentioning that ocean acidification ("OA;" i.e., elevated carbon dioxide partial pressures [pCO_2]) did not adversely affect corals in any experiment; in contrast, temperature generally had a dramatic effect on coral physiology. I propose to undertake proteomic examinations of some samples (experiments# 1-2 [n=12 and 18, respectively] and experiments 3-8 [n=10 for each experiment; 60 samples in total]). NA=not applicable. *S. hystrix*= *Seriatopora hystrix*. Trans-gen=trans-generational (adults=>larvae=>recruits). *experiment instead carried out at the Hawaii Institute of Marine Biology (Hawaii, USA).

Target species	**Life history stage**	**Temperature treatment (°C)**	**High pCO_2 (ppm)**	**Salinity effects tested?**	**Light effects tested?**	**Nutrient effects tested?**	**Time-scale**	**Accli-mation?**	**Refer-ence(s)**
S. hystrix	adult	27 vs. 30	NA	no	no	no	hours	yes	[6], [24]
S. hystrix	adult	26 vs. 23-29 over 6-hr	NA	no	no	no	days	yes	[18], [20], [24], [30], [32-33]
P. acuta	adult	NA	NA	no	yes*	no	hours	yes	[21-22]
P. acuta	adult	NA	NA	yes*	no	no	hours	yes	[23]
P. acuta	larvae	26 vs. 29	415 vs. 635	no	no	no	days	yes	[27]
P. acuta **exp. 1**	adult	26.5 vs. 29.7	NA	no	no	no	months	yes	[18-19], [31]
P. acuta **exp. 2**	adult	31.5-sustained	NA	no	no	no	weeks	no-bleached	[7], [37]
P. acuta **exp. 3**	adult	26 vs. 29	415 vs. 850	no	no	no	weeks	yes	[38]
P. acuta	adult	31.5-return to ambient at night	NA	no	no	no	weeks	yes	[7]

P. acuta **exp. 4**	adult	25	400 vs. 1,000	no	no	no	months	yes	[39]
S. hystrix	adult	25	400 vs. 1,000	no	no	no	months	yes	[39]
P. acuta **exp. 5**	adult	25, 28, or 31	400 vs. 800	no	no	yes	months	yes	[40]
P. acuta **exp. 6**	adult	26 vs. 29.5	NA	no	no	no	months	yes	[41]
P. acuta **exp. 7**	adult	26 vs. 32	NA	no	no	no	hours	yes	[42]
P. acuta **exp. 8**	trans-gen	26 vs. 30	NA	no	no	no	months	yes	[43]

To determine the predictive capacity of the bottom-up and top-down models (Aim III), I will field-test them at four study sites in Southern Taiwan (Fig. 1b): Wanlitong and Houwan in the Taiwan Strait and Houbihu and "Outlet" in Nanwan Bay. The latter site is named such due to its location at the outlet where the thermal effluent from a nearby nuclear power plant enters the ocean; seawater temperatures there can reach 35ºC, yet the coral communities are thriving (see Fig. 1 caption for additional details.). My colleagues and I have been 1) sampling tagged coral colonies and 2) collecting environmental data from these sites for months and years, respectively, and I will use protein biomarker concentration data and field environmental data, respectively, acquired over the course of the first half of 2019 to feed the bottom-up and top-down models, respectively. If the former and latter models can predict which coral colonies and reef sites, respectively, are most bleaching prone prior to the high-temperature, summer 2019 field season, then I will have validated the capacity to use such AI approaches to prioritize reefs for targeted, proactive management. Details of this AI+molecular biotechnological approach for assessing coral health in this era of changing global climate have been outlined below.

3. Methodology details

3.1. Aim I-bottom-up model training.

I will use Thermo Scientific's (USA) "tandem mass tag" (TMT) reagents to sequence the proteomes of the following 30 *P. acuta* samples with a Q Exactive™ mass spectrometer (MS; also from Thermo Scientific): 1) 3 samples maintained at control temperature+3 sub-lethally bleached samples at each of three sampling times (1 [mildly stressed], 7 [significantly stressed], and 14 [onset of bleaching] days; Table 1-experiment #2; n=18 [7]) and 2) 3 samples maintained at control temperature+3 samples that resisted high-temperature bleaching (30°C) at each of two sampling times (2 and 36 weeks; Table 1-experiment #1; n=12 [8]). The protein concentration data from the sub-lethally stressed samples will be compared to experimental controls, and the differentially concentrated proteins uncovered will be used to train the bottom-up PLS+NN model.

3.2. Aim II-top-down model training.

Although the majority of samples from the GCC simulation studies of Table 1, which will be used, in part, to train the top-down PLS+NN model, have been analyzed for an array of physiological- and molecular-scale response variables (Fig. 4), I will profile the proteomes herein of *P. acuta* samples from several other experiments listed in Table 1 (see table caption for details; n=60 proteomes). Furthermore, I have processed the majority of the coral samples from the GRE (Fig. 2), whose data will be used alongside the GCC simulation data for the top-down model training; of note, I have not yet analyzed those corals samples from the remote Chagos Archipelago (Indian Ocean). As corals of this uninhabited region are affected only by GCC, and no other anthropogenic stressors, these samples represent some of the most important coral samples ever collected, and they will be critical for the development of models seeking to elucidate environmental effects on corals. Details on how these

additional data will be acquired and analyzed can be found below under "*Molecular approaches*" (3.4) and "*Modeling approaches*" (3.5), respectively.

3.3. Aim III-field testing of the bottom-up and top-down predictive models.

Upon generating the bottom up (protein biomarker) and top-down (GCC+GRE eco-physiological datasets) PLS+NN models with JMP Pro, I will test their predictive capacity *in situ*. As mentioned above, my colleagues and I have been sampling corals from two of our four study sites (Wanlitong and Outlet; Fig. 1b) since 2017 and will continue to do so at regular intervals over 2018 and 2019. To validate the predictive capacity of the bottom-up model, protein biomarker concentrations will be measured in corals (n=48) sampled from two of the four Southern Taiwanese field sites, Outlet (hypothetically bleaching-tolerant) and Wanlitong (hypothetically bleaching-susceptible), and biomarker signatures will be "fed" into the bottom-up model to predict which colonies will bleach as temperatures rise (summer 2019) and which will not. As a more simplistic, similarity-based approach, I will compare the protein profiles of field-sampled *P. acuta* colonies to those of the GCC aquarium experiments (Table 1) using permutational analysis of variance (PERMANOVA; PRIMER ver. 7). This multivariate statistical approach will allow me to visualize, using multi-dimensional scaling, whether the overall protein profiles of the field colonies are more similar to those of bleaching-*tolerant* or bleaching-*prone* corals.

Simultaneously, the predictive capacity of the top-down approach will be verified by monitoring the EP determined by the PLS+NN model to be most important in influencing the coral response at all four field sites (Fig. 1b); upon inputting the field environmental data into the models, I will rank the four sites in terms of bleaching susceptibility and monitor the condition of the corals on the reefs to determine if those sites predicted to be most bleaching-prone indeed bleach prior to (or more severely than) those predicted to be more resilient. In the event that the predictive capacity of both model types is verified, the cheaper, top-down approach will be promoted given that, unlike the bottom-up approach, it requires neither expensive instrumentation (e.g., MS) nor highly trained personnel (i.e., molecular biologists).

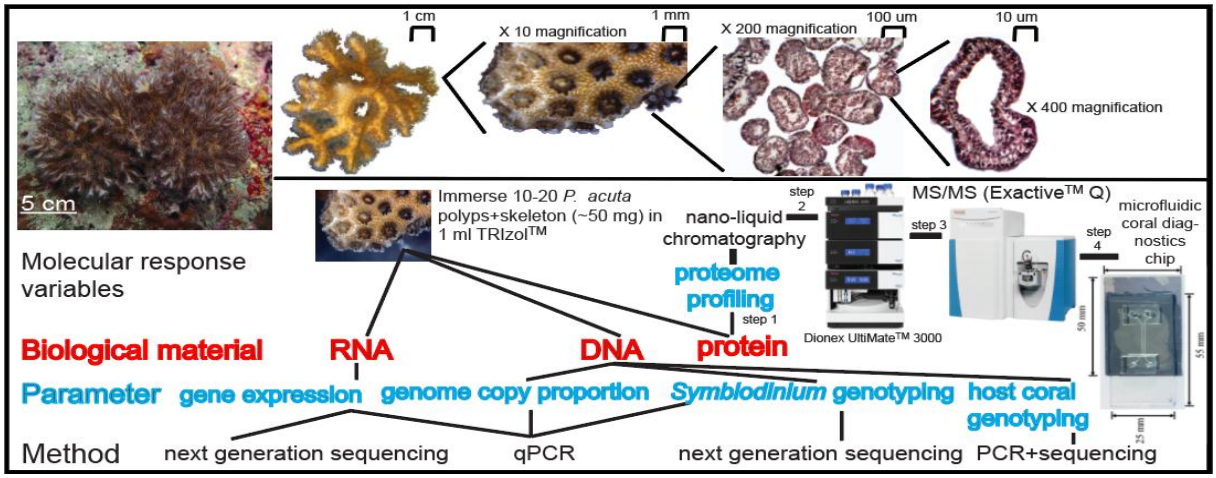


Figure 4. *Pocillopora acuta* over a variety of biological scales and the molecular protocol routinely used to gauge its health in the South Pacific and Coral Triangle. Protein biomarkers uncovered from analysis of samples of Aim I will be incorporated into a microfluidic chip such that coral health can be ascertained moments after corals are sampled (i.e., while still on the diving vessel), not days later in the laboratory.

3.4. Molecular approaches.

There is no correlation between gene expression and protein concentration in corals or their endosymbiotic dinoflagellates [18-20]. As such, widespread transcriptome profiling efforts by coral biologists (myself included) have *not* greatly improved our understanding of the coral response to GCC. Are we to develop proactive (pre-bleaching) coral health biomarkers and computer models, we instead require quantitative data on coral protein concentrations. By using TMT protein labeling approaches in conjunction with MS and customized bioinformatic scripts (known as "MS-SCAN") implemented on my *P. acuta* transcriptome+proteome server (http://symbiont.iis.sinica.edu.tw/coral_pdltte/static/html/index.html#home), we could acquire such proteomic data and determine which proteins are involved in coral bleaching (experiment #2 of Table 1); those proteins only synthesized by corals that ultimately bleached will be used to train the bottom-up predictive models discussed at length elsewhere.

3.4.1. Molecular approaches-detailed. For all aims, I will use a series of molecular protocols I have developed over the past decade for work with reef-building coral-*Symbiodinium* endosymbioses [21-22]. RNAs, DNAs, and proteins will be extracted with TRIzol™ from 50-mg biopsies from each sampled *P. acuta* nubbin (in the GCC experiments) or field colony (GRE samples; Fig. 4) at the National Oceanographic and Atmospheric Association's (NOAA) Atlantic Oceanographic and Meteorological Laboratory's (AOML) recently constructed molecular research laboratory (housed at the University of Miami's Rosenstiel School of Marine and Atmospheric Science [RSMAS] and funded through NOAA's "Omics' Initiative"). The RNAs will be assayed for expression of several target genes with real-time quantitative PCR (qPCR) as described in my prior works [23-24]. Furthermore, qPCR assays for genes expressed only by temperature-stressed samples are currently being designed [25] and will be ready for use by 2019. Next generation sequencing (NGS)-based transcriptome profiling (i.e., RNA-Seq) will also be undertaken with a subset of samples from Aims I-II using an Illumina platform. Although, as mentioned above, gene expression cannot be used to make physiological inferences about coral-*Symbiodinium* "holobionts" (host+endosymbionts), mRNA-level data *can* be used to identify corals displaying aberrant behavior (discussed in the next paragraph).

Those coral colonies identified as outliers based on a suite of multivariate statistical approaches [26], most importantly the Mahalanobis distance (a multivariate equivalent of the standard deviation), will be targeted for proteomics analysis as described below; given the prohibitive expense of profiling the proteomes (~$500/sample) of all 150 coral biopsies proposed to be generated across all three aims (90 GCC simulation samples [Taiwan]+48 field samples from Southern Taiwan+12 samples from Chagos), I intend to first employ this previously developed aberrancy detection approach to screen the collective sample set for those coral biopsies that are most likely to be stressed. This mRNA focus is due in part to the ease of carrying out qPCR, as well as its low expense, relative to protein profiling.

The DNAs co-extracted from the same coral biopsies from which RNAs were isolated will be used for four different purposes. First, an RNA/DNA ratio will be calculated to estimate levels of total gene transcription [6]. The DNAs will also be used to 1) estimate *Symbiodinium* density by calculating the "genome copy proportion" (GCP) [27] and 2) genotype the host corals [10] and 3) their *in hospite Symbiodinium* populations [28].

Next, proteins will be 1) extracted from the same coral biopsies from which RNAs and DNAs were isolated [29], 2) purified [30], and 3) labeled with TMT reagents such that up to 10 coral protein samples can be analyzed simultaneously by nano-liquid chromatography (LC) followed by MS. Instead of identifying differentially concentrated protein spots from 2-dimensional protein gels (the original proteomics approach), which is somewhat subjective and results in semi-quantitative data at best, this TMT-LC-MS/MS approach sequences all proteins synthesized by each sample in a manner analogous to RNA-Seq and provides spectral data that can be used to extract quantitative information on their cellular concentrations. TMT-LC-MS/MS will be undertaken at-cost at the proteomics facility housed within the University of South Florida's Center for Drug Discovery and Innovation.

Bioinformatics approaches will then be used to uncover those proteins uniquely expressed by samples of one treatment and not those of the other. For Aim I there will be a particular focus on proteins that are synthesized only by samples experiencing temperature stress (i.e., "core bleaching proteins" [CBPs]). A coral-customized proteomic script known as "MS-SCAN" will be used to characterize the sequenced proteins upon uploading the MS data files (.MGF) onto my coral transcriptome+proteome server (http://symbiont.iis.sinica.edu.tw/coral_pdltte/static/html/index.html#home); this interactive coral bioinformatics resource/repository has been described in detail in a prior work [31]. I can identify the CBPs the same day I receive the data files from the proteomics facility, allowing for the their rapid characterization. In addition to these RNA-, DNA-, and protein-level response variables (Fig. 4), several physiological indices of performance will also be assessed in experimental coral specimens from all aims. Please see my prior publications for examples, but these include (non-exhaustively) growth (for GCC simulation experiment samples only given the difficulty of assessing growth *in situ*), *Symbiodinium* photosynthetic efficiency [32] and chlorophyll content [33], and reproductive output.

3.5. Modeling approaches.

Although multiple regression and its various incarnations (e.g., stepwise regression) can theoretically handle large, complex environmental datasets, such as those proposed to be analyzed herein, the high collinearity among response variables (a hallmark of coral molecular datasets; see [34] for details.) leads to models that lack robustness. For this reason, statisticians have been making a strong case for physiologists to instead employ PLS [35], a predictive modeling approach that can be used when there are more response variables than samples (an issue with all "omic" analyses), as well as when there is a high degree of collinearity between response variables. Given these attributes, PLS appears to lend itself well to the assessment of coral reef data; not only does this algorithm attempt to model the response variable data (y) in such datasets, but it also designs parsimonious, best-fit models for the predictor variables (x; i.e., the experimental treatment in the case of the GCC simulation studies [Table 1] or the EP in the case of the GRE dataset [Fig. 2]).

However, issues that plague all modeling approaches (namely tradeoffs between over- and under-fitting) nevertheless persist with PLS, even with large (and growing) datasets such as the two featured herein. For these reasons, AI is quickly being exploited by biologists because of its capacity to learn from mistakes made from past simulations when building predictive models, and I propose to use NN with JMP Pro to build adaptable models capable of forecasting the likelihood of coral bleaching. Details of how NN works can be provided upon request, but it is based on the computer program learning from non-optimal projections in which data were over- or under-fit such that ensuing iterations strike a better balance in fitting the training and validation datasets; this balancing process is termed "boosting" by JMP. By integrating data from controlled tank studies (Table 1) and the field (Fig. 1a-b and Fig. 2), I hypothesize that these AI-optimized models (e.g., Fig. 3) will possess an elevated capacity to identify reefs, and reef corals, of compromised resilience within a timespan in which management intervention could seek to mitigate local-scale stressors and therefore thwart coral bleaching. In contrast, the popular bleaching model developed by NOAA ("Coral Watch"), which is based on "degree-heating weeks," suffers from predicting bleaching likelihood only days before this phenomenon is likely to begin; at such a point, it may be too late to enact any legislative changes.

4. Expected outcomes

If the molecular biomarker- and environmental dataset-based machine learning AI models are able to predict which coral colonies and reef sites, respectively, are most susceptible to environmental change, then the developed analytical system will represent the first proactive means of assessing sub-lethal levels of stress in corals and will consequently aid us in determining which reefs are most stress-prone prior to more visible, late-stage manifestations of severe health decline (e.g., bleaching). This novel technology seeks to replace the retroactive, vision-based manner in which we currently assess coral reef health (i.e., by instead documenting death [36]). In the event that the former, bottom-up PLS-NN

model is characterized by such predictive capacity, the underlying protein biomarkers will be integrated into a small, portable, relatively cheap, microfluidic diagnostic chip (Fig. 4); this device (discussed in detail in the next section) would allow for us to make predictions about CSS within minutes of sampling (i.e., while still at sea). Using the current approach outlined herein, at least several days are required to generate coral health-indicative proteomic data.

If the bottom-up *and* environmental data-based, top-down PLS-NN models are able to accurately forecast coral bleaching events with confidence, the algorithms will be published on open-access websites (e.g., coralreefdiagnostics.com and www.nationalgeographic.org) prior to publication in open-access journals (e.g., *PLoS ONE*). Although publication in the peer-reviewed literature will be critical for my future job security, it is the models themselves that are most important for conservation, and I aim to interact with Kenting National Park (KNP) officials over the course of the project. Furthermore, I gave a keynote speech on predictive modeling in the marine environment in Taitung, Taiwan in September 2018 (http://www.icsmtet.com/en-Untitled-2.html) and made numerous connections with Taiwanese citizens working in the ecosystem management, tourism, and conservation sectors. It will therefore be entirely feasible to ensure that my data reach those capable of doing the most good with them over the course of this project. As part of the **"First Maluku International Conference On Marine Science and Technology,"** I aim to further alert managers, scientists, and concerned members of the general public alike, of our growing capacity to begin to monitor Earth's coral reef ecosystems in a proactive (i.e., pre-death), data-driven manner.

Although corals and the reefs they construct will be the direct beneficiaries of this project, those millions of seafood-reliant Taiwanese, Filipino, and Indonesia nationals whose livelihoods depend on these Coral Triangle reefs for sustenance (and tourism revenue) will undoubtedly benefit from a proactive management plan for the targeted protection of their local reefs. In addition to the societal and conservation impacts of this project, it would give me significant opportunities to interact with local scientists and students in Taiwan and elsewhere in the Coral Triangle, namely the Philippines and Indonesia (where I currently have collaborators). Not only do I intend to give both scientific and general public seminars in Taiwan, as well as in the Miami area upon my move to South Florida's NOAA-AOML in May of 2019, but I also intend to maintain contact with my Taiwanese, Filipino, and Indonesian collaborators in the coming years to further explore ways we can work together to understand how corals will respond to the changes in their abiotic milieu that will come to pass over the coming decades as a result of GCC. Indeed, the majority of my time over the past two years has been dedicated to unpaid coral reef conservation and marine biology-focused public outreach enterprises, and I intend to uphold this service-oriented mentality over the course of my career.

5. Data dissemination

As mentioned above, I will publish all results in open-access journals and on my personal website (coralreefdiagnostics.com) such that any interested individual has access to all data generated (e.g., gene expression levels, protein profiling results, images of the sampled coral colonies, etc.), as well as both the bottom-up and top-down models described herein. In fact, I will alert local marine managers at KNP (described above) of such findings well before the respective manuscripts are published. The reason for doing so is because it can take months or even years to publish a scientific manuscript; during that time period, the bleaching-prone coral colonies likely to be identified herein will have already bleached and died. Data will be shared with coral reef managers through email or Microsoft's OneDrive data cloud, which I use regularly to share coral imagery data, presentations, and other large files. I will also archive data (at cost) on the open-access website "dryad.org," which I have used previously as a repository for coral health and image data from the South Pacific. Finally, I am an avid advocate of data transparency and am currently working with developers at JMP to make interactive data plots both on my website and in manuscripts so that those interested can recreate the same figures that I made (and therefore gain greater confidence in the analysis and interpretation of results).

It should be noted that, although software developed by a for-profit software company, JMP (a subsidiary of the much larger SAS Institutes, Inc.), will be used to build the PLS and NN-based

predictive models, the model codes themselves are likely to be quite simple, especially for PLS; several lines of universal scripting code is all that will likely need to be exported to a website, manuscript, email, etc., and JMP allows for the exporting of all relevant code in a format known as "JMP scripting language" (JSL), which is convertible to more common languages like Python or Java. In other words, one does not need to purchase an expensive JMP license (~$4,000USD) to interpret the code. In the case of the inherently more complex NN algorithms, it is likely that a JSL-based NN code could nevertheless be written such that an individual could use freeware such as R to read, interpret, and execute the program using their own protein biomarker (bottom-up) or environmental (top-down) data. Regardless of the approach, or the number of model terms, both PLS and NN take only seconds to run on a standard personal laptop, despite the elegant nature of the latter, in particular. As such, although AI has the potential to revolutionize coral health diagnostics, the actual amount of computing power needed is minimal. I therefore anticipate that a plethora of scientists, even in the least developed nations, will have the potential to utilize these technologies; anyone with internet connection and the capability of executing R (3-4 MB) will potentially benefit from the developed models in the event that the cheaper, top-down alternative is found to have predictive capacity.

If, on the other hand, only the more expensive, biotechnology-driven, bottom-up AI/PLS+NN model is found to have a high predictive capacity with respect to coral health, I will work with a Taiwanese microchip development laboratory (that of Dr. Gwo-Bin Lee at the state-run National Tsing-Hua University; the second highest-rated university in Taiwan) to manufacture small (5-6 cm) microfluidic chips (Fig. 4) featuring custom probes or antibodies that target the proteins found to be indicative of bleaching sensitivity (the "bleaching susceptibility chip") after validating the efficacy of the associated bottom-up "bleaching susceptibility index" (BSI) model. Although I do not intend to profit from this endeavor, which will enable me to assess coral health while still in the field, I may ultimately need to sell the chips at-cost since the microchip lab is non-profit.

6. Evaluation of results

An entire third of this project (Aim III) will be dedicated to evaluating the experimental models. Four to five *P. acuta* colonies will be tagged along each of two transects at each of two depths at each of four sites in Southern Taiwan (each of which being characterized by differing degrees of human impact and coral bleaching susceptibility; Fig. 1b) at each of four sampling times: January, March, May, and July of 2019. Then, TMT protein labeling+nano-LC+MS (i.e., TMT-LC-MS/MS) protein biomarker profiling will be carried out as described above with a subset of colonies from Wanlitong (Taiwan Strait; hypothetically stress-sensitive) and Outlet (Nanwan Bay; hypothetically bleaching-resilient). Two corals from each transect will be randomly selected for TMT-LC-MS/MS (n=48 total). When combined with the 30 GCC simulation samples to be analyzed from Aim I and those of other GCC simulations (n=60 across five experiments; Table 1) and GRE field sites (n=12 from the Chagos Archipelago) proposed to be assessed from Aim II, the proteomes of 150 corals will be profiled. The aforementioned, qPCR-based aberrancy detection system will be used to select a subset of these 150 samples for proteome profiling in the event that my current proposals are not funded in full.

Upon characterizing the proteomes of the field-sampled Taiwanese corals, the resulting data will be input into the bottom-up PLS+NN model developed in Aim I; those found to be over-expressing proteins known to be associated with bleaching will be given high BSI scores and would be expected to bleach at the high-temperature sampling time (September 2019). If, instead, the PLS-NN model yields low BSI scores (i.e., proteomes characterized by high concentrations of proteins associated with bleaching resistance upon comparison to those identified in experiment #1 of Aim I) in samples collected during the cooler sampling times (January, March, and May), such samples will instead be given low BSI scores and would not be predicted to bleach during periods of elevated temperatures.

Similarly, environmental data will be gathered from the four field sites bimonthly (January, March, and May, and July of 2019). Some such data, such as temperature and light will be measured by deployed loggers (HOBO Pendant™), which will be left at the sites to log data at 10-min intervals. Other such data, such as coral cover, will instead be acquired by dedicated surveyors. Upon inputting

the environmental data into the PLS+NN top-down models developed from assessment of the GCC simulation and GRE datasets (Aim II), I will make predictions as to which reefs are most likely to bleach as temperatures rise over the course of the summer. Then, I will return to the field sites in September (when temperatures typically peak) to determine the diagnostic capacity of my models.

If my models (either or both) are found to have predictive capacity with respective to coral health and bleaching, I can proceed to not only publish the associated manuscripts, but, more importantly, alert government officials (e.g., KNP officials) of my findings. For instance, if I find that a high percentage of corals on a particular reef are bleaching-prone based on their biomarker signatures (bottom-up model), yet they have not yet bleached, managers could be alerted to attempt to promote coral resilience by, for instance, closing down the reef to fishing; doing so would diminish the likelihood of algal overgrowth of corals (since herbivorous fish would become more abundant). Upon validation of the BSI, I will work with a microchip lab (discussed above) to miniaturize and expedite the coral diagnostics process to where it can be employed *in situ*, with data derived only moments later. Ideally, though, the cheaper, environmental data-based top-down model will likewise be characterized by a high prognostic capability, as this simpler approach better lends itself to developing nations lacking in the funds or infrastructure for proteomic analyses.

7. External capacity development

In addition to the societal and conservation impacts of this proposed work (addressed above), it would give me significant opportunities to interact with local scientists in Taiwan and elsewhere in the Coral Triangle. While at NMMBA for the first third of the project, I will collaborate not only with Taiwanese scientists, but also NMMBA graduate students and postdoctoral researchers. Over the past 6-7 years, I have mentored a Taiwanese Ph.D. student interested in coral molecular biology, Dr. Hung-Kai Chen, who is now a postdoctoral researcher at NMMBA, and I will continue to mentor Dr. Chen over the course of this project. Furthermore, I will co-supervise a Canadian Ph.D. student, Crystal McRae, who is currently based at NMMBA; many of the GCC simulations discussed throughout this proposal (Table 1), in fact, feature in her dissertation. As such, this project will result in the training of a late-stage graduate student, one who will hopefully take these skills back to her homeland (Canada), or, more likely, KNP (where her husband works) upon graduation (late 2019). Indeed, Ms. McRae's familial connection with KNP will aid in 1) permit acquisition (all coral sampling will take place within KNP.), and 2) effective dissemination of findings to park authorities. To belabour the point, it will be critical to convey my findings to not only other marine biologists, but also to students, the general public, government officials, and marine managers, such as these KNP park officials.

Not only do I intend to give both scientific and general public seminars in both Taiwan and South Florida (where I will move to start a new position in May of 2019), but I will maintain contact with my Taiwanese, Canadian, and American collaborators in the coming years to further explore ways we can work together to understand how corals will respond to the changes in their abiotic milieu that are now occurring. Additionally, I routinely collaborate with researchers in the countries in possession of the world's most beautiful (arguably) and high-biodiversity coral reefs, the Philippines and Indonesia, and I will seek to share findings from this project to my Coral Triangle collaborators (e.g., Dr. Gino Limmon of Pattimura University, Indonesia and Dr. Victor Tizcon of the University of the Philippines-Los Banos) over the course of 2019 and into the future. As mentioned elsewhere, the simpler, cheaper, top-down model featuring environmental data-based predictions of coral reef health will ideally prove to have high predictive capacity; in this case, the computer scripts for the resulting algorithms could simply be emailed to my collaborators. If, on the other hand, only the more expensive and complex bottom-up model proves effective, then I will distribute the aforementioned, protein biomarker-based CSS diagnostic microchips throughout Taiwan, the Philippines, Indonesia, and elsewhere in the Coral Triangle in 2020-2021.

Finally, I will collaborate with scientists at LOF (where I was once a research fellow), the institute that funded the world's largest coral reef survey ever undertaken (the GRE discussed elsewhere in this proposal), as these scientists routinely develop educational tools used in classroom training of students

about coral reef ecosystems. As such, findings to emerge from this proposed work will not only result in the development of esoteric computer scripts that may only be interpretable by a few, highly trained scientists, but the more general biological findings of interest (e.g., HOW corals bleach in the instance of those samples of experiment#2 of Table 1) will be distilled into more basic, fundamental principles that will be integrated into the open-source marine biology teaching materials under constant development by LOF. See www.livingoceansfoundation.org/educationaltools for details and examples.

8. Acknowledgements

I greatly benefited from discussions with Javed Inamdar (SAS Institutes, India), John Powell (JMP, USA), Vincent Chen (CarbonZeroToo, Taiwan), and JMP's statistical software development team. This proposal does not represent an endorsement of SAS, JMP, or any of their software; all opinions are my own. This work was funded, in part, by research grants from the National Science Foundation (USA).

9. References

[1] Mayfield AB and Gates RD 2007 *Comp. Biochem. Physiol*. **147A** 1-10
[2] Hoegh-Guldberg O, Mumby PJ, Hooten AJ, Steneck RS, Greenfield P, Gomez E, et al. 2007 *Science* **318** 1737-42
[3] Huang YCA, Hsieh HJ, Huang SC, Meng PJ, Chen YS, Keshavmurthy S, et al. 2011 *Mar. Ecol. Prog. Ser.* **423** 83-93
[4] Brown BE 1997 *Coral Reefs* **16** S129-38
[5] Gates RD 1990 *Coral Reefs* **8** 193-7
[6] Mayfield AB, Wang LH, Tang PC, Hsiao YY, Fan TY, Tsai CL, et al. 2011 *PLoS ONE* **e26529**
[7] Mayfield AB, Chen M, Meng PJ, Lin HJ, Chen CS and Liu PJ 2013 *Mar. Environ. Res*. **86** 1-11
[8] Mayfield AB, Fan TY and Chen CS 2013 *Coral Reefs* **32** 909-21
[9] Mayfield AB, Bruckner AW, Chen CH and Chen CS 2015 *Platax* **12** 1-17
[10] Mayfield AB, Chen CS, Dempsey AC and Bruckner AW 2016 *Platax* **13** 1-25
[11] Barshis DJ, Ladner JT, Oliver TA, Seneca FO, Traylor-Knowles N and Palumbi SR 2013 *Proc. Natl. Acad. Sci. USA* **110** 1387-92
[12] Krueger T, Horwitz N, Bodin J, Giovani ME, Escrig S, Meibom A, et al. 2017 *Royal Soc. Open Sci.* **4** 170038
[13] Jones RJ, Hoegh-Guldberg O, Larkum AWD and Schreiber U 1998 *Plant Cell Environ*. **21** 1219-30
[14] Mayfield AB, Chen CS and Dempsey AC 2017 *PLoS ONE* **e0185857**
[15] Mayfield AB, Chen CS and Dempsey AC 2017 *PLoS ONE* **e0177267**
[16] Mayfield AB, Chen CS and Dempsey AC 2017 *Platax* **14** 1-45
[17] Mayfield AB, Dempsey AC, Inamdar J and Chen CS in press *Platax*
[18] Mayfield AB, Wang YB, Chen CS, Chen SH and Lin CY 2016 *Mol. Ecol.* **25** 5944-58
[19] Mayfield AB, Chen YJ, Lu CY and Chen CS 2018 *PLoS ONE* **e0192001**
[20] Mayfield AB, Chen YJ, Lu CY and Chen CS 2018 *Open J. Mar. Sci.* **8** 223-52
[21] Mayfield AB, Hsiao YY, Fan TY, Chen CS and Gates RD 2010 *J. Exp. Mar. Biol. Ecol*. **395** 215-22
[22] Mayfield AB, Hsiao, YY, Fan TY and Chen CS 2012b *Platax* **9** 1-24
[23] Mayfield AB, Fan TY and Chen CS 2013 *Platax* **10** 1-29
[24] Mayfield AB, Chen YH, Dai CF and Chen CS 2014 *Int. J. Mar. Sci.* **4**(50) 1-23
[25] Mayfield AB and Chen CS in prep.
[26] Mayfield AB 2016 *J. Mar. Sci. Eng.* **4:63**
[27] Putnam HP, Mayfield AB, Fan TY, Chen CS and Gates RD 2013 *Mar. Biol.* **160** 2157-73
[28] Correa AMS, McDonald MD and Baker AC 2009 *Mar. Biol.* **156** 2403-11
[29] Mayfield AB, Hsiao YY, Chen HK and Chen CS 2014 *Mar. Biotech.* **16** 371-84
[30] Mayfield AB, Chen YJ, Lu CY and Chen CS 2016 Proteins responsive to variable temperature

exposure in the reef-building coral *Seriatopora hystrix Coral Reefs: Ecosystems, Environmental Impact and Current Threats* ed S Ortiz (New York: NOVA Publishing) chapter 1 pp 13-72
[31] Mayfield AB, Wang YB, Chen CS, Chen SH and Lin CY 2014 *Mol. Ecol.* **23** 5816-30
[32] Mayfield AB, Fan TY and Chen CS 2013 *J. Mar. Biol.* **Article ID 569369**
[33] Mayfield AB, Chan PH, Putnam HP, Chen CS and Fan TY 2012 *J. Exp. Biol.* **215** 4183-95
[34] Mayfield AB, Chen CS and Dempsey AC in review *J. Sea Res.*
[35] Cox I and Gaudard M 2013 *Discovering Partial Least Squares with JMP®.* SAS Institute, Inc., Cary, NC, USA
[36] Liu PJ, Meng PJ, Liu LL, Wang JT and Leu MY 2012 *Mar. Pollut. Bull*. **64** 1129-35
[37] Mayfield AB, Chen CS and Liu PJ 2014 *Platax* **11** 1-23
[38] Putnam HP, Mayfield AB, Fan TY and Gates RD in prep.
[39] Liu PJ, Meng PJ, Lin HJ and Mayfield AB in prep.
[40] Mayfield AB and Liu PJ in prep.
[41] McRae CM, Fan TY, Cote I, Chen CA, and Mayfield AB in prep.
[42] McRae CM, Cote I, Fan TY, and Mayfield AB in prep.
[43] McRae CM, Huang WB, Fan TY, Chen CA, and Mayfield AB in prep.