Project Summary

Title: A framework for integrating multi-omics data for biomarker discovery to improve resilience in aquaculture

PD: Shelly Wanamaker **Institution:** Gloucester Marine Genomics Institute

Program Area: Animal Breeding, Genetics, and Genomics

Economic challenges imposed by climate change and disease on the aquaculture industry necessitate advances for improved animal welfare and resiliency. Biomarkers associated with environmental and disease resilience traits can be leveraged in breeding and management strategies. However, their discovery has been limited in part by the complexity of molecular systems and the cost of genomics tools used to understand them. Advances in computational approaches including machine learning algorithms, together with the wealth of genomic data that has amassed, enable powerful meta-analyses for improved biomarker discovery in aquaculture species. The proposed project aims to advance the discovery and characterization of biomarkers through mining publicly available shellfish genomic datasets from resilient populations. The objectives are to 1) develop standardized open-access, user-friendly, reproducible bioinformatics pipelines for resilience biomarker discovery through systematic reanalysis, data integration and meta-analysis and 2) build a user-friendly open-access comprehensive database of candidate resilience biomarkers that is widely available for use by the aquaculture community. The resulting database will enable improved molecular tool development for more efficient phenotype selection and health monitoring, relevant to the AFRI animal genomics program area priority goal of increasing animal fitness and improving animal welfare as well as the priority of implementing selection methods that use a systems biology approach for simultaneous improvement of multiple traits.

Project Narrative

1. SPECIFIC APPLICATION TYPE: New application not being submitted to other agencies
2. INTRODUCTION: The wealth of publicly available data generated from omics applications in aquacultured species is underutilized because of limited infrastructure and tools for mining these data. Leveraging these data to understand common molecular drivers of commercially important phenotypes could significantly advance broodstock improvement programs. The long-term goal of this project is to improve resiliency in aquaculture by creating computational resources for genomics integration. The main objective is to create a multidimensional bioinformatics framework for resilience biomarker discovery that could be used for breeding programs, disease monitoring, and physiological assessment in aquaculture. Using shellfish genomic datasets as a proof-of-concept, the following supporting objectives are proposed:

i. Develop standardized open-access, user-friendly, reproducible bioinformatics pipelines for resilience biomarker discovery through systematic reanalysis, data integration and meta-analysis
ii. Build a user-friendly open-access comprehensive database of candidate resilience biomarkers that is widely available for use by the aquaculture community

The PD is uniquely qualified to carry out these objectives, with expertise using genomic, proteomic, and metabolic analyses to characterize molecular features underlying resilience traits in aquatic species[1–4]. She also has expertise in network analysis and data integration having generated and analyzed plant and human interactomes[5–7], and has extensive experience building open-access reproducible bioinformatics pipelines and webpages evidenced on her GitHub and professional websites[8,9]. This seed grant will help the PD be competitive for future funding and to become established as a leading genomics researcher in the East Coast shellfish community **by producing computationally validated preliminary data for a USDA standard grant proposal to experimentally validate candidate biomarkers of resilience** (e.g. comparing candidate biomarkers in outplanted wild and genomically selected populations) **from which to develop rapid molecular tools for resilience screening**, similar to those she developed for shrimp disease monitoring [10].

Genomics applications in aquaculture have lagged behind other agricultural industries particularly in underutilizing biomarkers for health monitoring and for selecting commercially important traits. Molecular markers have had a long-standing utility in selective crop breeding[11], but there are few examples of their use in improving fish farming with less than 10% of aquaculture production derived from selectively bred stocks[12]. Variants for genomic selection for disease resistance in rainbow trout have been identified through SNP array and RAD-seq[13]. However, genomic and marker assisted selection in shellfish aquaculture have been challenging because of high genetic diversity in molluscan species, genotyping costs, larval mortality rates, incidence of marker segregation distortion, and limited genomic resources[14]. **SNP arrays exist for selective breeding but are not cost-effective requiring hundreds to >10,000 SNPs to be surveyed for genotyping associated with few traits**[15,16]. The low economic value of individual animals has led to brute force overproducing to account for losses, yet unexplained larval and adult mortality and disease outbreak are still industry-wide problems that should not be accepted as unavoidable costs[17]. Reducing the number or improving the predictive value of markers needed for trait selection would make molecular screening more cost effective and compatible with simplified diagnostic technology, but improved biomarker identification is needed.

Biomarkers can be defined as molecular features that can be reproducibly and objectively measured across individuals[18] and can serve as an indicator of performance or health. They can be identified through genomics data-driven discovery where the presence or relative abundance

of specific molecules are tightly and statistically associated with a phenotype[19]. Although phenotype-associated molecular features identified through omics studies can be used as candidate biomarkers with applications in selective breeding programs, disease diagnosis, biomonitoring, and welfare certification[18,20], **there is only sparse overlap between molecular features identified by similar studies that observe the same phenotype**[21]. Other times, the same phenotype is observed across studies, but different omics approaches were used, making biomarker discovery a challenge. Complex traits are typically polygenic and traditional selection based on a few *genetic* markers is generally ineffective[22], often requiring many variants to be surveyed at high costs. However, this could be overcome by **incorporating multi-omic data to reduce bias of a single method, increase marker search space to identify unknown interactions, and redefine resilience biomarkers for precision breeding and culturing**. Disparate omics studies have identified resilience associated-molecular features in aquatic species for instance in Pacific oyster, sets of genomic sequence variants, transcripts, and proteins have been found to be associated with larger size and heat stress resilience[23], or decreased viral susceptibility[24]. In epigenomic, proteomic, and metabolomic studies previously done by the PD, sets of genes, proteins, and metabolites were found to be associated with environmental stress resilience in corals, Pacific geoduck, Pacific oyster, and Dungeness crab[1–4]. As a natural extension of the PD's previous work and to overcome within-study bias, the work proposed here intends to **increase biomarker discovery power through bioinformatics standardization and data stratification**. With more and more omics datasets being generated in different aquaculture species, there is now an avenue for data-driven discovery of novel biomarkers associated with commercially important phenotypes that could be targeted by selective breading or management strategies.

Although there is limited existing computational framework for biomarker discovery in aquaculture species, there are examples in humans. An improved reference genome and optimized sequence alignment pipeline used in the **systematic reanalysis of human genomic data detected a previously missed pathogenic biomarker and that led to a proper diagnosis and improved treatment**[25]. Systematic reanalysis has also improved patient treatment by enabling pathogenic biomarker reclassification based on greater population frequency[26]. The ENCODE consortium enabled the integration of more than 60,000 internationally generated datasets through implementation of clear data standards and guidelines[27]. Cancer biomarkers have been identified and prioritized through systematic analysis and network-based data integration of publicly available genomic, epigenomic, and transcriptomic data from different tumor types[28]. This work proposes to enhance resilience biomarker discovery in aquaculture species by similarly using standardized bioinformatics pipelines and improved reference data to reprocess disparate open access omics datasets from resilient shellfish populations. This alleviates within study bias making resilience markers more comparable across studies. Comprehensive cross-study validation through data integration can overcome the challenge of a single molecular feature not providing a holistic measure of phenotype[29].

3. RATIONALE AND SIGNIFICANCE: While significant progress has been made in applying omics to understand phenotypes in aquaculture species, there is a need to develop validated integrative omics bioinformatics approaches to improve interpretation of data from different omics platforms[20]. Stakeholders at the 2017 USDA Genome to Phenome Meeting identified oyster and clam broodstock genomic analysis as a lacking area, and it was reported that phenotype-associated biomarker identification through omics data integration has rarely occurred[30]. This project is directly related to the AFRI priority area (2) animal health and

production and animal products because the discovery, characterization, and comprehensive compilation of candidate biomarkers would provide aquaculturists with new molecular targets for selective breeding and health monitoring tools. This proposal fills the need for standardized and integrative bioinformatics approaches and fills the gaps of underused genomic datasets and biomarker technology in the aquaculture field. The impact of achieving these objectives would be a major advancement in biomarker development facilitating improvements in animal health monitoring and production. The biomarker discovery platform described here (streamlined bioinformatics pipelines, biomarker characterization, and biomarker database creation) has relevance beyond shellfish aquaculture and could be generalized for biomarker discovery in virtually any species for which genomics data exists across animal and plant sectors. With greater ability to economically characterize more biomarkers across commercially important species using existing datasets, there is greater potential for biomonitoring and selective breeding method improvement at large.

4. APPROACH: *Objective 1. Develop standardized open-access, user-friendly, reproducible bioinformatics pipelines for resilience biomarker discovery through systematic reanalysis, data integration and meta-analysis.* To initially establish bioinformatics pipelines, systematic reanalysis will be done on omics datasets from *Crassostrea gigas* studies that compared phenotypes among different populations in response to environmental conditions and to disease (**Figure 1a** and **Table 1**). Once parameters and pipelines are established, datasets will be processed from other bivalve species (e.g. Eastern oyster[31,32], Pacific geoduck[2,33–37]). Using standardized pipelines to reprocess each data type will eliminate variation from using different software and parameters. Using common reference data will standardize feature names across datasets to facilitate downstream comparisons. Candidate biomarkers that are consistently associated with resilience traits across studies and stress types will be indicative of innate traits drivers. Systematic reanalysis pipeline development: As diagramed in **Figure 1b,** datasets and corresponding metadata will be downloaded from public repositories (e.g. AgData Commons, SRA, GEO, ENA). Any missing metadata (e.g. water chemistry parameters, experiment duration, life stage, tissue type, observed phenotype) will be retrieved from the published study and compiled. The same software specific to each data type will be used. Available nf-core community (https://nf-co.re) software will be used because these have already been standardized to ensure high reproducibility with stable releases, high portability, thorough documentation, and support[38]. New pipelines will be created for analyses not currently available (e.g. proteomics and metabolomics workflows) following the nf-core template and guidelines. Optimal alignment and feature calling parameters will be determined by iteratively testing effects of different modifications on alignment statistics, quality control metrics, and total features identified across studies. Modifications resulting in the greatest preservation of data, comparable to or improving upon published results, with the highest quality control metrics will be selected. To circumvent the challenge posed by coverage variation across samples within a study, which could influence the resulting count data, data would be effectively down-sampled to achieve more uniform coverage. To limit the impact from technical biases introduced by different sequencing methods (e.g. RNAseq and TagSeq), data will be appropriately normalized[39] and passed through a normalizing pipeline (e.g. NormalizerDE[40]) prior to differential and integrative analyses. All data output from standardized bioinformatics pipelines will be in a common format to facilitate downstream differential and integrative analyses. Molecules identified through reanalysis are expected to largely overlap with those previously identified in each respective study, and to be novel owing to updated reference genomes, annotations, and improved bioinformatics pipelines.

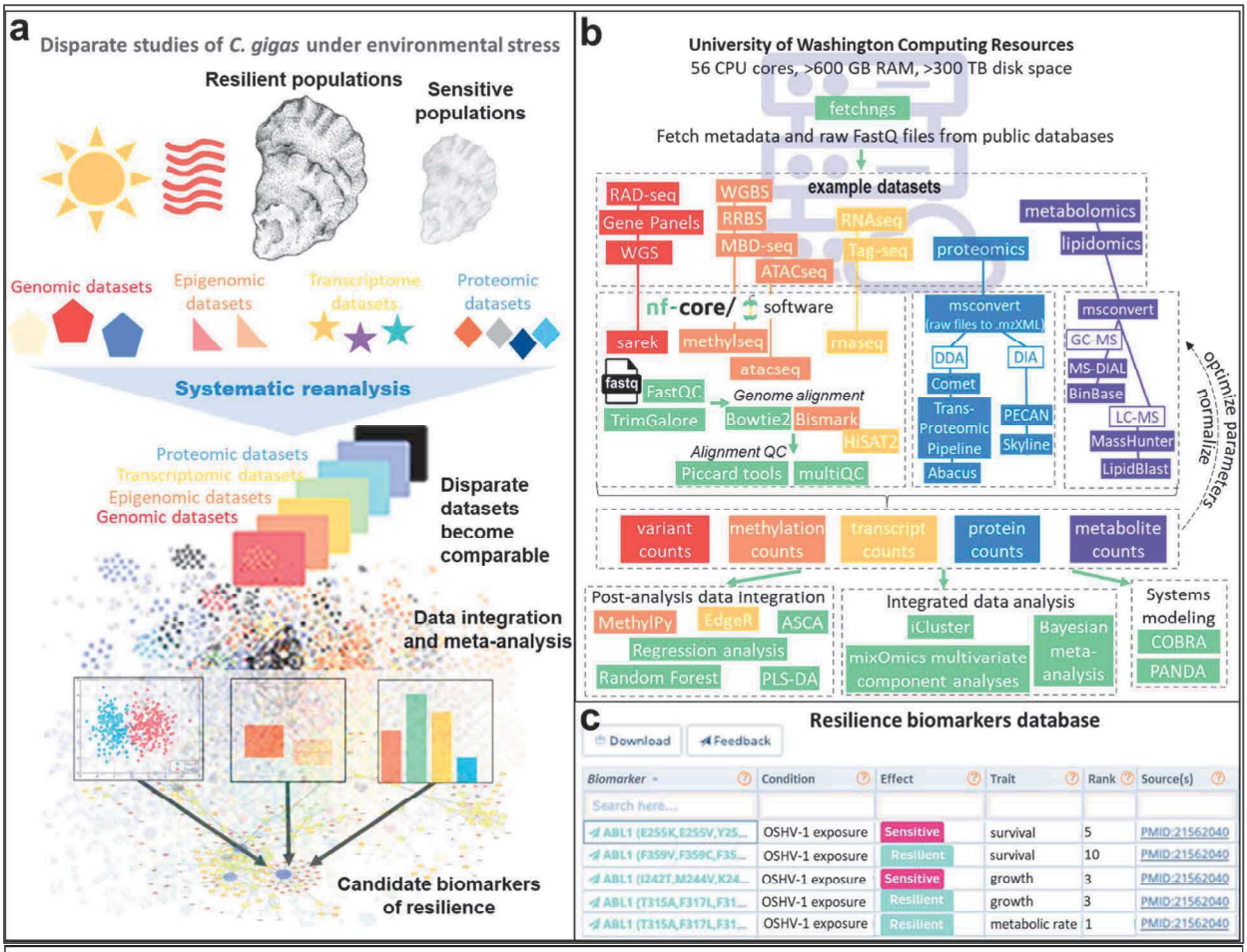

Figure 1. **a)** Graphical summary of Objective 1. **b)** Example pipelines and software for systematic reanalysis and data integration. **c)** Example of resilience biomarker database (modeled after Tamborero *et al.* 2018[46]).

To circumvent the quality limitations posed by any reference genome or protein sequence database, alignment thresholds will be optimized to increase mapping and reduce false positive alignments. Analysis pipelines will be mobilized together on a virtual machine and several options will be examined for open-access hosting including popular cloud-based infrastructure (e.g. nf-core, CyVerse[41], Jetstream[42]). Data integration pipeline development: Both post-analysis data integration and integrated data analysis will be carried out on standardized data generated in Objective 1a, and data integration pipelines will be mobilized to the cloud-based virtual machine. Although standardizing data integration methods could limit the results because some methods may be more sensitive than others to certain datasets, the benefits from enabling broader data integration outweigh this limitation. To circumvent variation from experimental differences in cross-study comparisons, covariates and random effects (e.g. water chemistry parameters) will be included in statistical models for differential analysis. To make data derived from different platforms more comparable, normalization, batch effect correction, and data harmonization methods will be applied. For post-analysis data integration, different omics datasets will be first analyzed in isolation. For each dataset, differential analyses will identify molecular features (e.g. differentially expressed genes) that statistically differ among experimental groups defined in each study. All experimental group comparisons will include statistics that can independently assess (1) group response mean methylation, expression, or protein abundance (e.g. linear

regression and Cohen's d effect size) and (2) group response variance (e.g. multiple linear regression models of relative standard deviation[43]) as they relate to phenotype and experimental condition, similar to that done in the PD's previous studies[1,4]. Molecular features associated with common phenotypes will then be compared across studies for biomarker discovery using meta-analysis methods (e.g. weighted Z score for combining p-values and RankSum method for combining effect sizes across datasets)[44].

Table 1. Publicly available Pacific oyster datasets. Data type denoted as genomic (G), epigenomic (E), transcriptomic (T), and proteomic (P). This table is not exhaustive.

Data Type	stress class	stressor	Phenotype	Phenotpe Summary	Reference
G	environment	thermal	thermotolerance different among lines	resilience	Ding *et al.* 2020
P	environment	thermal	larger size	resilience	Trigg *et al.* 2020
P	environment	thermal	thermotolerance different among *gigas* vs. *angulata*	resilience	Wang *et al.* 2023
T	environment	thermal	thermotolerance different among *gigas* vs. *sikamea*	resilience	Wang *et al.* 2021
T	environment	thermal	thermotolerant diploid vs sensitive triploid	resilience	George *et al.* 2023
T	environment	thermal	thermotolerance different among *gigas* vs. *angulata*	resilience	NCBI BioProject PRJNA253988
E	environment	pH	tolerant diploid vs triploid	resilience	Arredondo-Espinoza *et al.* 2021
E, T	environment	pH	larval survival impacted by parental low pH exposure	sensitivity	Venkataraman *et al.* 2022
P	environment	pH	reduced endpoint size	sensitivity	Dineshram *et al.* 2015
P	environment	pH	reduced shell toughness and hardness	sensitivity	Timmins-Schiffman *et al.* 2014
T	environment	pH	calcification onset time was delayed and rate was reduced	sensitivity	De Wit *et al.* 2018
T	environment	pH	shell parameters and respiration rate impacted	sensitivity	Lutier *et al.* 2022
T	environment	O2	*gigas* vs. *sikamea*	resilience	NCBI BioProject PRJNA587775
P	environment	salinity	salinity stress decreased quality	sensitivity	Chen *et al.* 2022
T	environment	salinity	*gigas* vs. *hongkongensis*	resilience	NCBI BioProject PRJNA266361
E	environment	multifactorial	different families show different tolerances	resilience	Wang *et al.* 2023
G, T, P	environment	multifactorial	size and heat shock response vary with location	resilience	Li *et al.* 2018
T, P	disease	*Vibrio*	induced phagocytosing	sensitivity	Jiang *et al.* 2018
T	disease	*Perkinsus*	different tolerance among gigas vs. verginica	resilience	NCBI BioProject PRJNA778545
G	disease	OSHV-1	resistant vs. susceptible strain	resilience	NCBI BioProject PRJNA828432
G	disease	OSHV-1	different families show different tolerance	resilience	Gutierrez *et al.* 2018
P	disease	OSHV-1	different families show different tolerance	resilience	Lepretre *et al.* 2021
T	disease	OsHV-1	reduced infection at high temperature	resilience	IFREMER, 2019
T	disease	OSHV-1	different families show different tolerance	resilience	Segarra *et al.* 2014

Integrated data analysis involves combining different omics datasets with common phenotypes prior to performing analyses, enabling similarities among datasets to arise from statistics rather than human interpretation[21]. Clustering and dimensionality reduction-based approaches as well as systems-based network modeling approaches will be used to compare standardized data within and across studies to identify biomarkers. The leave-one-out-method will be used to rank biomarkers for reliability (not dependent on one dataset)[45] and perturbation clustering will be used to rank ability to predict phenotype[46]. Biomarkers will be computationally validated using independent datasets that did not undergo integrated data analysis and evaluating their ability to discriminate phenotypes within and across species. To further resolve relationships among molecular features and resilience, genome feature (e.g. gene, transposable element, and promoter annotations), biological process and pathway (e.g. gene and pathway ontology annotations), and inferred protein-protein interaction and metabolic network[40] annotations (e.g. from STRING and IntAct databases[47,48]) will be integrated. Conserved interaction patterns across multiple and highly divergent species[49,50] allow orthologous protein-protein interaction network projection analysis. Unannotated features will be excluded from this functional analysis but will still have a biomarker ranking and can be validated in future experimental studies. Results from differential analyses will be compiled into a large list of

molecular features of diverse classes significantly associated with resilience and will be validated by comparing to features identified in published studies. Features identified across studies associated with resilience will be considered candidate biomarkers. Both species-specific and cross-species biomarkers are expected to be identified from each discovery approach as has been observed in previous studies[51,52].

The work for Objective 1 will be done by Senior Research Associate (SRA) Jennifer Polinski who will be partially funded by this grant. SRA Polinski has extensive experience in high performance computing, data management, bioinformatic pipeline development, and genomic and transcriptomic analysis for diverse species. She will be supervised and mentored by the PD who will be partially funded by this grant. The PD will lead pipeline mobilization and evaluation efforts. Pipeline reproducibility and user-friendliness will be evaluated by personnel at the PD's institution and by collaborators (see letters of support) using independent datasets, pre-established evaluation criteria (ease of access, installation time, run time, ease of setup, etc.) and providing feedback. The pipelines and systematic reanalysis will be presented at national conferences (e.g. the Plant Animal Genome conference) to solicit community-wide feedback, and the framework will be optimized according to suggestions. Attendees will use their own datasets or datasets used in the proposed project and will evaluate the framework in real time. A major Objective 1 outcome is a bioinformatics framework of user-optimized pipelines that can be applied to diverse genomic datasets beyond the proposed work, lending itself to comparative meta-analyses across many species. Another major outcome is the discovery of new biomarkers of resilience in *C. gigas* and other bivalves. The compiled list of candidate biomarkers from within and across diverse studies can be used for developing assays for resilience biomarker selection and health diagnostics.

Objective 2. Build a user-friendly open-access comprehensive database of candidate resilience biomarkers that is widely available for use by the aquaculture community. Study-specific and cross-study discovered biomarkers will be compiled into a user-friendly, interactive, comprehensive database to facilitate their use across the aquaculture field. Database features and build: Modelled after the Cancer Biomarker Database[53], candidate biomarkers will be easily searchable through a column sort feature (e.g. for condition, trait, effect - resilient or sensitive), and a search bar feature for text querying (**Figure 1c**). A 'download' feature will enable users to download the entire or filtered subset of the database. An 'evidence' column will denote species-specific or cross-species evidence. A 'class' column will denote the discovery class (genomic feature, ontology enrichment, or protein interaction network). A 'source' column will list the publication(s) where the original data was generated. A 'rank' column will list a confidence score based on supporting evidence that incorporates criteria like the number of studies, number of species, and number of datatypes. A 'feedback' feature will allow users to solicit feedback and request updates (e.g. biomarker addition) to the table. Instructions will be listed in text and screencast formats describing how the database can be used and how users can contribute or solicit feedback. The table will be generated using the R function 'datatable'[54] and the basis will be a simple CSV file so that no proprietary or complex software is needed. Management, and maintenance: The database and webpage code will be hosted on GitHub and backed up on Open Science Framework for redundancy, which allow the code and data to be maintained in perpetuity through web archiving. In addition to being mirrored on GMGI and University of Washington organizational websites, open access through GitHub allows easy mobility and any user to also mirror the webpage on any site through repository cloning. GitHub hosting allows

users to join as collaborators and contributors initially at the discretion of the PD and proposal collaborators (but extending to others in the future) for data files and webpage format editing and openly soliciting feedback using the publicly accessible 'issues' function. Stable releases will be made annually to ensure data, webpage, and software are up to date. Information from the previous releases will be permanently available and documentation of updates for each release will be available through a link on the home page.

Community testing and introduction: The initial release will be presented and demonstrated in virtual and in-person training meetings with proposal collaborators and lab members who will beta-test independently. Database features and webpage functions will be thoroughly tested following a standardized task list and will be evaluated by criteria ranking (e.g. task difficulty, user-friendliness). Communication for beta testing will be facilitated by GitHub 'discussions' and 'issues' features that allow for tracked, annotated, prioritized, and searchable Q&A and sharing by anyone in the community. Following optimization after beta-testing, a new release will be made and demonstrated in an interactive presentation to the broader community at national conferences (e.g. Plant Animal Genome Aquaculture America, and National Shellfish Association conferences), where community users could engage and contribute to feedback.

Challenges: No challenges are anticipated in building the database webpage given the PD's experience in R and building open access GitHub webpages, or in database management given the low maintenance requirements of GitHub pages. Potential challenges may arise during beta testing and community use in obliging diverse preferences, however this can be remedied by using quantitative evaluation criteria and incorporating changes based on majority rankings and by using up-voting data in suggestions made in GitHub 'discussions'. The Objective 2 outcome is an optimized database webpage available for community use in validating candidate biomarkers and developing molecular tools for biomarker screening to improve aquaculture production. Ultimately, the proposed project would valuably bring clarity to the relationship between genotype and phenotype and improve understanding around how different phenotypes arise within and across species, advancing the development of resilience biomarkers in aquaculture to improve animal health and production.

Project timetable. Colors indicate personnel that will execute activities (blue, PD Wanamaker; Yellow, SRA Polinski; green, both).

	2024				2025			
	Jun-Aug	Sept-Nov	Dec-Feb	Mar-May	Jun-Aug	Sept-Nov	Dec-Feb	Mar-May
Objective 1: Develop standardized open-access, user-friendly, reproducible bioinformatics pipelines for resilience biomarker discovery through systematic reanalysis, data integration and meta-analysis								
Compile and organize omics data and metadata for all *C. gigas* studies and reference genome data	X	X						
Systematic reanalysis of *C. gigas* datasets	X	X	X					
Systematic reanalysis of other shellfish datasets				X	X	X		
Mobilize analysis pipelines and structured datasets onto cloud-based infrastructure	X	X	X					
Post-analysis data integration biomarker discovery			X	X	X	X		
Integrated data analysis biomarker discovery			X	X	X	X		
Systems modeling biomarker discovery			X	X	X	X		
Computational biomarker validation					X	X		
Evaluate reproducibility, user-friendliness, accessibility, and modularity and optimize			X	X	X	X		
Objective 2: Build comprehensive database of candidate resilience biomarkers								
Build and launch biomarker database website					X	X	X	X
Evaluate user-friendliness and accessibility and optimize							X	X

BibliographyReferencesCited

1. Trigg, S. A., Mitchell, K. R., Thompson, R. E., Eudeline, B., Vadopalas, B., Timmins-Schiffman, E. B. & Roberts, S. B. Temporal proteomic profiling reveals insight into critical developmental processes and temperature-influenced physiological response differences in a bivalve mollusc. *BMC Genomics* **21,** (2020).
2. Putnam, H. M., Trigg, S. A., White, S. J., Spencer, L. H., Vadopalas, B., Natarajan, A., Hetzel, J., Jaeger, E., Soohoo, J., Gallardo-Escárate, C., Goetz, F. W. & Roberts, S. B. Dynamic DNA methylation contributes to carryover effects and beneficial acclimatization in geoduck clams. *bioRxiv* (2022).
3. Trigg, S. A., Venkataraman, Y. R., Gavery, M. R., Roberts, S. B., Bhattacharya, D., Downey-Wall, A., Eirin-Lopez, J. M., Johnson, K. M., Lotterhos, K. E., Puritz, J. B. & Putnam, H. M. Invertebrate methylomes provide insight into mechanisms of environmental tolerance and reveal methodological biases. *Mol Ecol Resour* **22,** (2022).
4. Trigg, S. A., McElhany, P., Maher, M., Perez, D., Busch, D. S. & Nichols, K. M. Uncovering mechanisms of global ocean change effects on the Dungeness crab (Cancer magister) through metabolomics analysis. *Scientific Reports 2019 9:1* **9,** 1–12 (2019).
5. Trigg, S. A., Garza, R. M., MacWilliams, A., Nery, J. R., Bartlett, A., Castanon, R., Goubil, A., Feeney, J., O'Malley, R., Huang, S.-S.-C., Zhang, Z. Z., Galli, M. & Ecker, J. R. CrY2H-seq: A massively multiplexed assay for deep-coverage interactome mapping. *Nat Methods* **14,** (2017).
6. Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., Carvunis, A. R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B. J., Hardy, M. F., Jin, M., Kang, S., Kiros, R., Lin, G. N., Luck, K., Macwilliams, A., Menche, J., Murray, R. R., Palagi, A., Poulin, M. M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J. M., Scholz, A., Shah, A. A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A. O., Trigg, S. A., Twizere, J. C., Vega, K., Walsh, J., Cusick, M. E., Xia, Y., Barabási, A. L., Iakoucheva, L. M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M. A., Hill, D. E., Hao, T., Roth, F. P. & Vidal, M. A proteome-scale map of the human interactome network. *Cell* **159,** 1212–1226 (2014).
7. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y. A., Murray, R. R., Spirohn, K., Begg, B. E., Duran-Frigola, M., MacWilliams, A., Pevzner, S. J., Zhong, Q., Trigg, S. A., Tam, S., Ghamsari, L., Sahni, N., Yi, S., Rodriguez, M. D., Balcha, D., Tan, G., Costanzo, M., Andrews, B., Boone, C., Zhou, X. J., Salehi-Ashtiani, K., Charloteaux, B., Chen, A. A., Calderwood, M. A., Aloy, P., Roth, F. P., Hill, D. E., Iakoucheva, L. M., Xia, Y. & Vidal, M. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164,** 805–817 (2016).
8. shellywanamaker (Shelly Wanamaker). at <https://github.com/shellywanamaker>

9. Shelly Wanamaker. at <https://shellywanamaker.github.io/>
10. Major, S. R., Harke, M. J., Cruz-Flores, R., Dhar, A. K., Bodnar, A. G. & Wanamaker, S. A. Rapid Detection of DNA and RNA Shrimp Viruses Using CRISPR-Based Diagnostics. (2023). doi:10.1128/aem.02151-22
11. Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., Yıldız, M., Hatipoğlu, R., Ahmad, F., Alsaleh, A., Labhane, N., Özkan, H., Chung, G. & Baloch, F. S. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnology and Biotechnological Equipment* **32,** Preprint at https://doi.org/10.1080/13102818.2017.1400401 (2018)
12. Houston, R. D. Future directions in breeding for disease resistance in aquaculture species. *Revista Brasileira de Zootecnia* **46,** Preprint at https://doi.org/10.1590/S1806-92902017000600010 (2017)
13. Vallejo, R. L., Gao, G., Liu, S., Fragomeni, B. O., Hernandez, A. G., Parsons, J. E., Martin, K. E., Evenhuis, J. P., Welch, T. J., Leeds, T. D., Wiens, G. D. & Palti, Y. Genome-wide association studies reveal similar genetic architecture with shared and unique QTL for bacterial cold water disease resistance in two rainbow trout breeding populations. *bioRxiv* (2017).
14. Hollenbeck, C. M. & Johnston, I. A. Genomic tools and selective breeding in molluscs. *Front Genet* **9,** Preprint at https://doi.org/10.3389/fgene.2018.00253 (2018)
15. Guo, X., Puritz, J. B., Wang, Z., Proestou, D., Allen, S., Small, J., Verbyla, K., Zhao, H., Haggard, J., Chriss, N., Zeng, D., Lundgren, K., Allam, B., Bushek, D., Gomez-Chiarri, M., Hare, M., Hollenbeck, C., La Peyre, J., Liu, M., Lotterhos, K. E., Plough, L., Rawson, P., Rikard, S., Saillant, E., Varney, R., Wikfors, G. & Wilbur, A. Development and Evaluation of High-Density SNP Arrays for the Eastern Oyster Crassostrea virginica. *Mar Biotechnol (NY)* **25,** 174–191 (2023).
16. Gutierrez, A. P., Turner, F., Gharbi, K., Talbot, R., Lowe, N. R., Peñaloza, C., McCullough, M., Prodöhl, P. A., Bean, T. P. & Houston, R. D. Development of a Medium Density Combined-Species SNP Array for Pacific and European Oysters (Crassostrea gigas and Ostrea edulis). *G3 (Bethesda)* **7,** 2209–2218 (2017).
17. Gray, M. W., Alexander, S. T., Beal, B. F., Bliss, T., Burge, C. A., Cram, J. A., Luca, M. De, Dumhart, J., Glibert, P. M., Gonsior, M., Heyes, A., Huebert, K. B., Lyubchich, V., McFarland, K., Parker, M., Plough, L. V., Schott, E. J., Wainger, L. A., Wikfors, G. H. & Wilbur, A. E. Hatchery crashes among shellfish research hatcheries along the Atlantic coast of the United States: A case study at Horn Point Laboratory oyster research hatchery. *Aquaculture* **546,** (2022).
18. Raposo de Magalhães, C. S. F., Cerqueira, M. A. C., Schrama, D., Moreira, M. J. V., Boonanuntanasarn, S. & Rodrigues, P. M. L. A Proteomics and other Omics approach in the context of farmed fish welfare and biomarker discovery. *Rev Aquac* **12,** Preprint at https://doi.org/10.1111/raq.12308 (2020)

19. McDermott, J. E., Wang, J., Mitchell, H., Webb-Robertson, B. J., Hafen, R., Ramey, J. & Rodland, K. D. Challenges in biomarker discovery: Combining expert insights with statistical analysis of complex omics data. *Expert Opin Med Diagn* **7,** Preprint at https://doi.org/10.1517/17530059.2012.718329 (2013)
20. Nguyen, T. V. & Alfaro, A. C. Applications of omics to investigate responses of bivalve haemocytes to pathogen infections and environmental stress. *Aquaculture* **518,** (2020).
21. Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J. & Wishart, D. Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. *Metabolites* **9,** (2019).
22. Yang, B., Zhai, S., Zhang, F., Wang, H., Ren, L., Li, Y., Li, Q. & Liu, S. Genome-wide association study toward efficient selection breeding of resistance to Vibrio alginolyticus in Pacific oyster, Crassostrea gigas. *Aquaculture* **548,** 737592 (2022).
23. Li, L., Li, A., Song, K., Meng, J., Guo, X., Li, S., Li, C., De Wit, P., Que, H., Wu, F., Wang, W., Qi, H., Xu, F., Cong, R., Huang, B., Li, Y., Wang, T., Tang, X., Liu, S., Li, B., Shi, R., Liu, Y., Bu, C., Zhang, C., He, W., Zhao, S., Li, H., Zhang, S., Zhang, L. & Zhang, G. Divergence and plasticity shape adaptive potential of the Pacific oyster. *Nat Ecol Evol* **2,** (2018).
24. Segarra, A., Mauduit, F., Faury, N., Trancart, S., Dégremont, L., Tourbiez, D., Haffner, P., Barbosa-Solomieu, V., Pépin, J. F., Travers, M. A. & Renault, T. Dual transcriptomics of virus-host interactions: Comparing two Pacific oyster families presenting contrasted susceptibility to ostreid herpesvirus 1. *BMC Genomics* **15,** (2014).
25. Cowley, M. J., Liu, Y. C., Oliver, K. L., Carvill, G., Myers, C. T., Gayevskiy, V., Delatycki, M., Vlaskamp, D. R. M., Zhu, Y., Mefford, H., Buckley, M. F., Bahlo, M., Scheffer, I. E., Dinger, M. E. & Roscioli, T. Reanalysis and optimisation of bioinformatic pipelines is critical for mutation detection. *Hum Mutat* **40,** (2019).
26. Hiatt, S. M., Amaral, M. D., Bowling, K. M., Finnila, C. R., Thompson, M. L., Gray, D. E., Lawlor, J. M. J., Cochran, J. N., Bebin, E. M., Brothers, K. B., East, K. M., Kelley, W. V., Lamb, N. E., Levy, S. E., Lose, E. J., Neu, M. B., Rich, C. A., Simmons, S., Myers, R. M., Barsh, G. S. & Cooper, G. M. Systematic reanalysis of genomic data improves quality of variant interpretation. *Clin Genet* **94,** (2018).
27. Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J. W., Tanaka, F. Y., Adenekan, P., O'Neill, E. & Cherry, J. M. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48,** (2020).
28. Dimitrakopoulos, C., Hindupur, S. K., Hafliger, L., Behr, J., Montazeri, H., Hall, M. N. & Beerenwinkel, N. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34,** (2018).

29. Hook, S. E., Gallagher, E. P. & Batley, G. E. The role of biomarkers in the assessment of aquatic ecosystem health. *Integr Environ Assess Manag* **10,** Preprint at https://doi.org/10.1002/ieam.1530 (2014)
30. Rexroad, C., Vallet, J., Matukumalli, L. K., Reecy, J., Bickhart, D., Blackburn, H., Boggess, M., Cheng, H., Clutter, A., Cockett, N., Ernst, C., Fulton, J. E., Liu, J., Lunney, J., Neibergs, H., Purcell, C., Smith, T. P. L., Sonstegard, T., Taylor, J., Telugu, B., Van Eenennaam, A., Van Tassell, C. P., Wells, K., Martin, A., Murdoch, B., Sayre, B., Keel, B., Schmidt, C., Hostetler, C., Seabury, C., Tuggle, C., Elsik, C., Gill, C., Ciobanu, D., Bailey, D., Hamernik, D., Grings, E., Connor, E., Rohrer, G., Plastow, G., Rosa, G., Zhou, H., Koltes, J., Decker, J., Weller, J., Woodward-Greene, J., Steibel, J., Long, J., Lee, K., Kuehn, L., Worku, M., Salem, M., McCue, M., Serao, N., Riggs, P., Sponenberg, P., Schnabel, R., Brooks, S., Fernando, S., McKay, S., Schmitz-Esser, S., White, S., Lamont, S., Kurt, T., Palti, Y., Moser, D., Wiggans, G., Van Orsouw, E., Anderson, J., Adetula, A. A., Dechow, C., MacHugh, D., McCarthy, F., Parker-Gaddis, K., MacNeil, M., Motroni, R., Shanower, T., Giannakopoulos, E., Nugent, M., Weaver-Missick, T., Krieg, A., Pujara, R., Campbell, T., Kebede, S., Rowan, T., Nilson, S., Kramer, L., Daza, K., Krehbiel, B., Kiana, P., Ujcic, J., Boarman, J. & Coleman, L. Genome to phenome: Improving animal health, production, and well-being - A new USDA blueprint for animal genome research 2018-2027. *Front Genet* **10,** Preprint at https://doi.org/10.3389/fgene.2019.00327 (2019)
31. Proestou, D. A., Sullivan, M. E., Lundgren, K. M., Ben-Horin, T., Witkop, E. M. & Hart, K. M. Understanding Crassostrea virginica tolerance of Perkinsus marinus through global gene expression analysis. *Front Genet* **14,** (2023).
32. Johnson, K. M., Sirovy, K. A., Casas, S. M., La Peyre, J. F. & Kelly, M. W. Characterizing the Epigenetic and Transcriptomic Responses to Perkinsus marinus Infection in the Eastern Oyster Crassostrea virginica. *Front Mar Sci* **7,** (2020).
33. Timmins-Schiffman, E., Guzmán, J. M., Elliott, R., Vadopalas, B. & Roberts, S. B. Dynamic response in the larval geoduck clam proteome to elevated pCO2. *bioRxiv* (2019).
34. Timmins-Schiffman, E., Guzmán, J. M., Elliott Thompson, R., Vadopalas, B., Eudeline, B. & Roberts, S. B. Larval Geoduck (Panopea generosa) Proteomic Response to Ciliates. *Sci Rep* **10,** (2020).
35. Timmins-Schiffman, E., Guzmán, J. M., Elliott Thompson, R., Vadopalas, B., Eudeline, B. & Roberts, S. B. Dynamic response in the larval geoduck (Panopea generosa) proteome to elevated pCO2. *Ecol Evol* **10,** (2020).
36. Timmins-Schiffman, E., Coffey, W. D., Hua, W., Nunn, B. L., Dickinson, G. H. & Roberts, S. B. Shotgun proteomics reveals physiological response to ocean acidification in Crassostrea gigas. *BMC Genomics* **15,** (2014).
37. Spencer, L. H., Horwith, M., Lowe, A. T., Venkataraman, Y. R., Timmins-Schiffman, E., Nunn, B. L. & Roberts, S. B. Pacific geoduck (Panopea generosa) resilience to natural pH variation. *Comp Biochem Physiol Part D Genomics Proteomics* **30,** (2019).

38. Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P. & Nahnsen, S. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* **38,** Preprint at https://doi.org/10.1038/s41587-020-0439-x (2020)
39. van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics* **7,** (2006).
40. Willforss, J., Chawade, A. & Levander, F. NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis. *J Proteome Res* **18,** (2019).
41. Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D. & Antin, P. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol* **14,** (2016).
42. Stewart, C. A., Hancock, D., Stanzioneb, D., Turnerd, G., Cockerill, T. M., Merchant, N., Taylor, J., Vaughn, M., Foster, I., Skidmore, E., Tuecke, S. & Gaffney, N. I. Jetstream: A self-provisioned, scalable science and engineering cloud environment. in *ACM International Conference Proceeding Series* **2015-July,** (2015).
43. Huh, I., Zeng, J., Park, T. & Yi, S. V. DNA methylation and transcriptional noise. *Epigenetics Chromatin* **6,** (2013).
44. Makinde, F. L., Tchamga, M. S. S., Jafali, J., Fatumo, S., Chimusa, E. R., Mulder, N. & Mazandu, G. K. Reviewing and assessing existing meta-analysis models and tools. *Brief Bioinform* **22,** Preprint at https://doi.org/10.1093/bib/bbab324 (2021)
45. Pellegrini, M., Pulvirenti, A., Bostan, H., Draghici, S., Shafi, A., Nguyen, T., Peyvandipour, A. & Nguyen, H. A Multi-Cohort and Multi-Omics Meta-Analysis Framework to Identify Network-Based Gene Signatures. *Frontiers in Genetics | www.frontiersin.org* **1,** 159 (2019).
46. Nguyen, T., Tagett, R., Diaz, D. & Draghici, S. A novel approach for data integration and disease subtyping. (2017). doi:10.1101/gr.215129.116
47. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., Van Roey, K., Cesareni, G. & Hermjakob, H. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42,** (2014).
48. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J. & Von Mering, C. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47,** (2019).

49. Zhong, Q., Pevzner, S. J., Hao, T., Wang, Y., Mosca, R., Menche, J., Taipale, M., Taşan, M., Fan, C., Yang, X., Haley, P., Murray, R. R., Mer, F., Gebreab, F., Tam, S., MacWilliams, A., Dricot, A., Reichert, P., Santhanam, B., Ghamsari, L., Calderwood, M. A., Rolland, T., Charloteaux, B., Lindquist, S., Barabási, A., Hill, D. E., Aloy, P., Cusick, M. E., Xia, Y., Roth, F. P. & Vidal, M. An inter-species protein–protein interaction network across vast evolutionary distance. *Mol Syst Biol* **12,** (2016).
50. Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M. & Ideker, T. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* **102,** (2005).
51. Clark, M. S., Sommer, U., Sihra, J. K., Thorne, M. A. S., Morley, S. A., King, M., Viant, M. R. & Peck, L. S. Biodiversity in marine invertebrate responses to acute warming revealed by a comparative multi-omics approach. *Glob Chang Biol* **23,** (2017).
52. Panahi, B., Frahadian, M., Dums, J. T. & Hejazi, M. A. Integration of cross species RNA-seq meta-analysis and machine-learning models identifies the most important salt stress–responsive pathways in microalga Dunaliella. *Front Genet* **10,** (2019).
53. Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M. P., Vivancos, A., Rovira, A., Tusquets, I., Albanell, J., Rodon, J., Tabernero, J., de Torres, C., Dienstmann, R., Gonzalez-Perez, A. & Lopez-Bigas, N. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* **10,** 1–8 (2018).
54. Xie, Y., Cheng, J. & Tan, X. DT: A Wrapper of the JavaScript Library 'DataTables'. Preprint at (2023)
55. Zhang, F., Hu, B., Fu, H., Jiao, Z., Li, Q. & Liu, S. Comparative transcriptome analysis reveals molecular basis underlying fast growth of the selectively bred pacific oyster, Crassostrea gigas. *Front Genet* **10,** (2019).
56. Jiang, S., Qiu, L., Wang, L., Jia, Z., Lv, Z., Wang, M., Liu, C., Xu, J. & Song, L. Transcriptomic and quantitative proteomic analyses provide insights into the phagocytic killing of hemocytes in the Oyster Crassostrea gigas. *Front Immunol* **9,** (2018).
57. De Wit, P., Durland, E., Ventura, A. & Langdon, C. J. Gene expression correlated with delay in shell formation in larval Pacific oysters (Crassostrea gigas) exposed to experimental ocean acidification provides insights into shell formation mechanisms. *BMC Genomics* **19,** (2018).
58. Dineshram, R., Quan, Q., Sharma, R., Chandramouli, K., Yalamanchili, H. K., Chu, I. & Thiyagarajan, V. Comparative and quantitative proteomics reveal the adaptive strategies of oyster larvae to ocean acidification. *Proteomics* **15,** (2015).
59. Ding, F., Li, A., Cong, R., Wang, X., Wang, W., Que, H., Zhang, G. & Li, L. The Phenotypic and the Genetic Response to the Extreme High Temperature Provides New Insight Into Thermal Tolerance for the Pacific Oyster Crassostrea gigas. *Front Mar Sci* **7,** 539202 (2020).

60. Wang, C., Du, M., Jiang, Z., Cong, R., Wang, W., Zhang, G. & Li, L. Comparative proteomic and phosphoproteomic analysis reveals differential heat response mechanism in two congeneric oyster species. *Ecotoxicol Environ Saf* **263,** (2023).
61. Wang, C., Li, A., Wang, W., Cong, R., Wang, L., Zhang, G. & Li, L. Integrated Application of Transcriptomics and Metabolomics Reveals the Energy Allocation-Mediated Mechanisms of Growth-Defense Trade-Offs in Crassostrea gigas and Crassostrea angulata. *Front Mar Sci* **8,** 744626 (2021).
62. George, M. N., Cattau, O., Middleton, M. A., Lawson, D., Vadopalas, B., Gavery, M., Roberts, S. B. & Matthew George, C. N. Triploid Pacific oysters exhibit stress response dysregulation and elevated mortality following heatwaves. *Glob Chang Biol* **00,** 1–19 (2023).
63. Venkataraman, Y. R., White, S. J. & Roberts, S. B. Differential DNA methylation in Pacific oyster reproductive tissue in response to ocean acidification. *BMC Genomics* **23,** 1–16 (2022).
64. Lutier, M., Di Poi, C., Gazeau, F., Appolis, A., Le Luyer, J. & Pernet, F. Revisiting tolerance to ocean acidification: Insights from a new framework combining physiological and molecular tipping points of Pacific oyster. *Glob Chang Biol* **28,** (2022).
65. Chen, L., Shi, H., Li, Z., Yang, F., Zhang, X., Xue, Y., Zhang, H. & Xue, C. Molecular mechanism of protein dynamic change in Pacific oyster (Crassostrea gigas) during depuration at different salinities uncovered by mass spectrometry-based proteomics combined with bioinformatics. *Food Chem* **394,** 133454 (2022).
66. Wang, X., Cong, R., Li, A., Wang, W., Zhang, G., Li, L. & Blasco, J. Transgenerational effects of intertidal environment on physiological phenotypes and DNA methylation in Pacific oysters. *Science of the Total Environment* **871,** 162112 (2023).
67. Roberto, A. E., Ana M., I., Steven, R. B., Maria Teresa, S. G. & Cristina, E. F. Differentially methylated gene regions between resistant and susceptible heat-phenotypes of the Pacific oyster Crassostrea gigas. *Aquaculture* **543,** (2021).
68. Gutierrez, A. P., Bean, T. P., Hooper, C., Stenton, C. A., Sanders, M. B., Paley, R. K., Rastas, P., Bryrom, M., Matika, O. & Houston, R. D. A genome-wide association study for host resistance to ostreid herpesvirus in Pacific oysters (Crassostrea gigas). *G3: Genes, Genomes, Genetics* **8,** 1273–1280 (2018).
69. Leprêtre, M., Faury, N., Segarra, A., Claverol, S., Degremont, L., Palos-Ladeiro, M., Armengaud, J., Renault, T. & Morga, B. Comparative Proteomics of Ostreid Herpesvirus 1 and Pacific Oyster Interactions With Two Families Exhibiting Contrasted Susceptibility to Viral Infection. *Front Immunol* **11,** 621994 (2021).
70. Segarra, A., Faury, N., Pépin, J. F. & Renault, T. Transcriptomic study of 39 ostreid herpesvirus 1 genes during an experimental infection. *J Invertebr Pathol* **119,** 5–11 (2014).