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

GLOUCESTER MARINE GENOMICS INSTITUTE

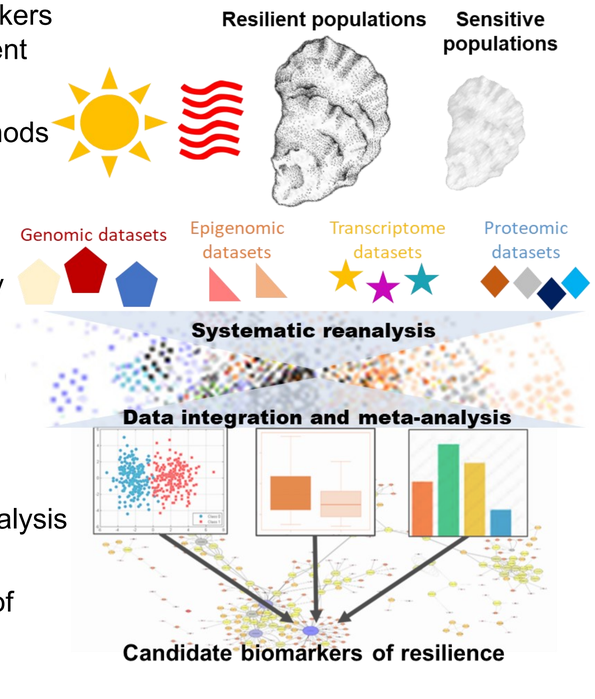
Shelly A. Wanamaker[1], Emma Strand[1], Steve Yost[1], Steven B. Roberts

1. Gloucester Marine Genomics Institute, 417 Main Street Gloucester, MA
2. University of Washington School of Aquatic and Fishery Sciences, 1122 NE Boat Street Seattle, WA

This work is supported by the Health and Production and Animal Products program under the Animal Breeding, Genetics, and Genomics section, project award no. 2024-67015-41794 from the U.S. Department of Agriculture's National Institute of Food and Agriculture.

BACKGROUND

- Disease and environmental resilience biomarkers can be leveraged in breeding and management strategies
- Advances in AI and other computational methods enable powerful meta-analyses of existing genomic data and can improve biomarker discovery
- Project goal: To advance biomarker discovery through mining publicly available genomic datasets from resilient shellfish populations
- Outcomes:
  1. Standardized, easily reproducible bioinformatics pipelines for systematic reanalysis, data integration and meta-analysis of diverse omics datasets
  2. Open-access comprehensive database of candidate biomarkers for use by the aquaculture community

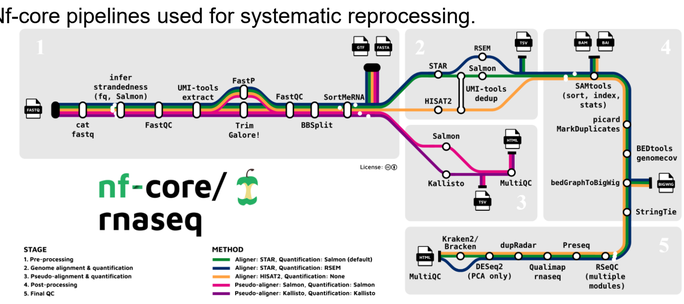


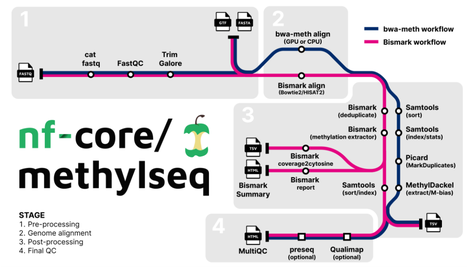
METHODS

Datasets systematically reprocessed to-date. Screenshot taken from about page on project website.

Species	Data Type	stress class	stressor	Phenotype	Phenotpe Summary	Reference	DOI	meta data file	counts table
C. gigas	T	environment	thermal	thermotolerance different among lines	resilience	Arredondo-Espinoza *et al.* 2023	https://doi.org/10.1016/j.cbd.2023.101089	SraRunTable.csv	salmon.merged.gene_counts.tsv
C. virginica	T	disease	Perkinsus	infection tolerance	resilience	Proestou *et al.* 2023	https://doi.org/10.3389/fgene.2023.1054558	SraRunTable.csv	salmon.merged.gene_counts.tsv
C. virginica	E,T	disease	Perkinsus	infected	sensitivity	Johnson *et al.* 2020	https://doi.org/10.3389/fmars.2020.00598	SraRunTable(1).csv	salmon.merged.gene_counts.tsv
C. gigas & virginica	T	disease	Perkinsus	infection tolerance	resilience	Chan *et al.* 2021	10.3389/fgene.2021.795706	SraRunTable(2).csv	salmon.merged.gene_counts.tsv *Cgigas data here
C. virginica	T	disease	Perkinsus	infection	sensitivity	Sullivan and Proestou 2021	https://doi.org/10.1016/j.aquaculture.2021.736831	SraRunTable(3).csv	salmon.merged.gene_counts.tsv

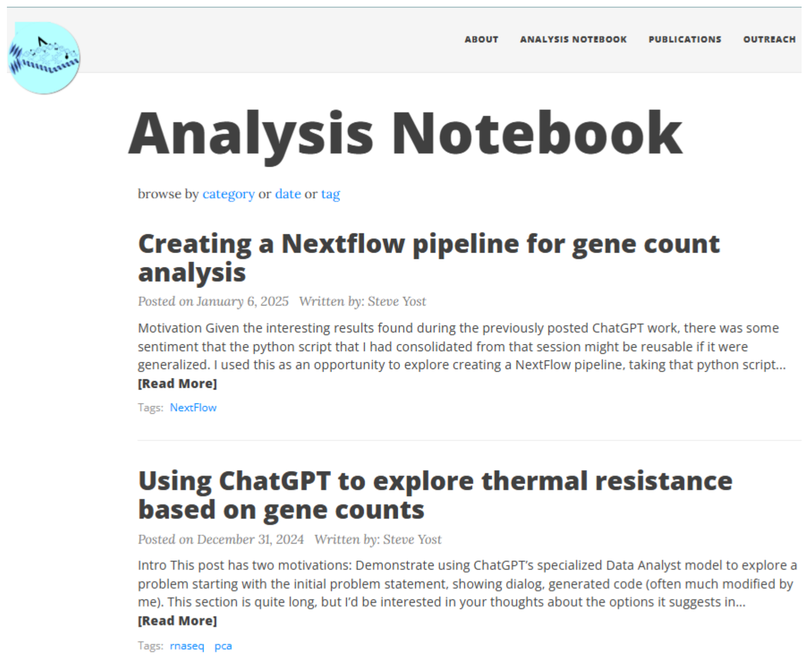
Nf-core pipelines used for systematic reprocessing.





RESULTS

Standardized, easily reproducible bioinformatics pipelines



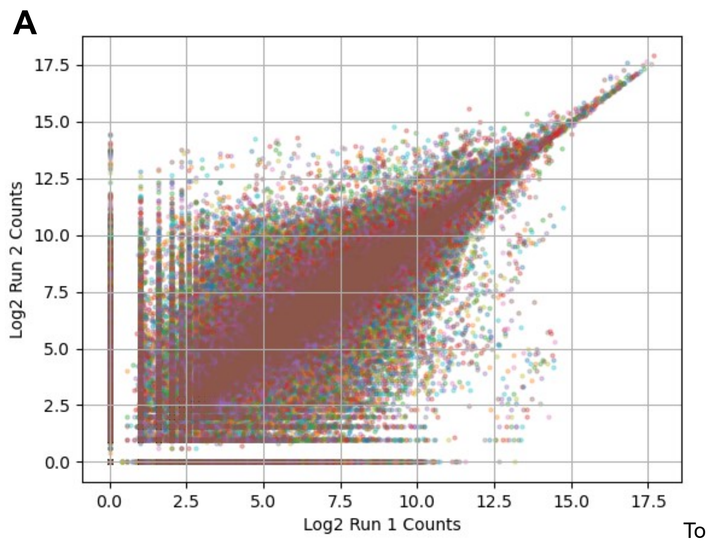
Reproducible analyses are regularly posted to an open-access analysis notebook hosted on the project website (https://resilience-biomarkers-for-aquaculture.github.io/) through GitHub. Above is a screenshot of two example notebook entries describing different data integration strategies.

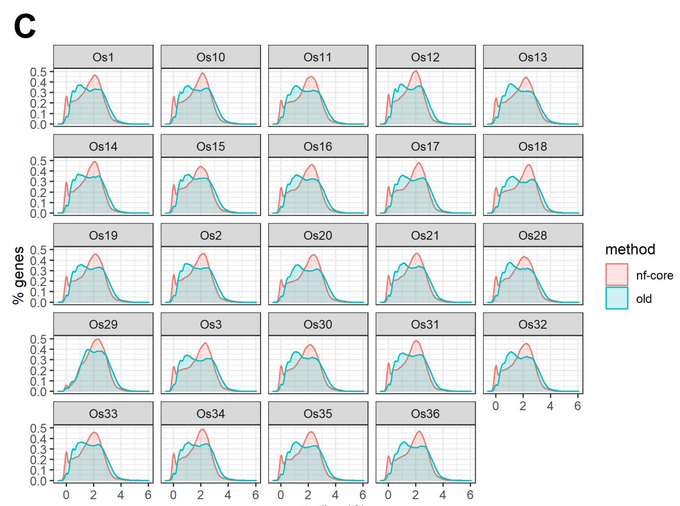
Pipelines enable rapid processing in university and public access HPCs

Dataset	Sequencing	Samples	Pipeline	Size (Gbases)	Run Time (HH:MM:SS)
Arredondo-Espinoza et al	PE 2x100	36	RNAseq	93	10:28:16
Proestou et al	PE 2x150	61	RNAseq	818	5:41:09
Johnson et al	SE 1x100	40	RNAseq	155	0:32:12
Chan et al	SE 1x50	10	RNAseq	93	2:09:25
Sullivan and Proestou	PE 2x125	44	RNAseq	110	0:47:55
Roberts et al	PE 2x150	32	Methylseq	882	21:21:19

- >2.2 Tbases of data processed in 41 hours on University of Washington HPC cluster
- To demonstrate an alternative to a local HPC, the RNAseq pipeline was run with the Arredondo-Espinoza dataset on the seqera platform a centralized command center for managing and executing pipelines at scale and completed in 5h 13m 9s at an estimated cost of $6.52.

Systematic reanalysis using updated genome and nf-core pipeline show consistency with initial study





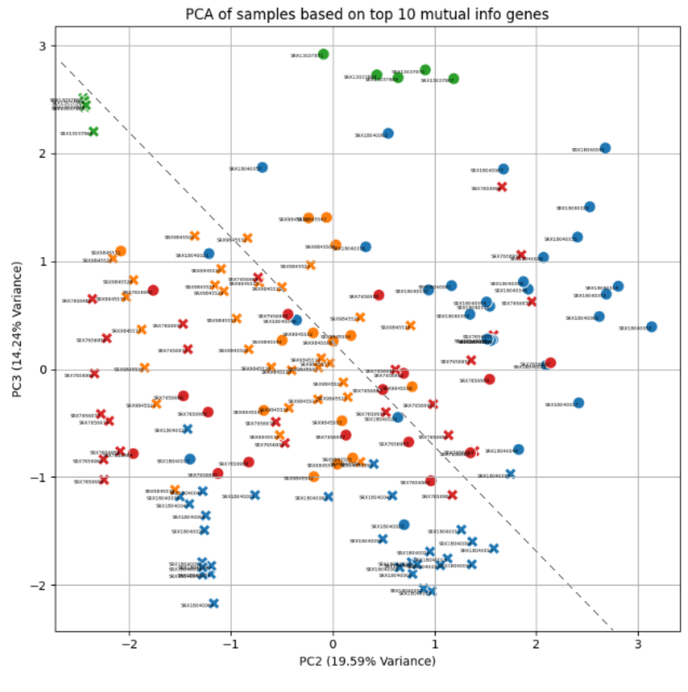
To understand the impact of using a different bioinformatics pipeline and reference genomes, for simplicity we first tested one RNAseq dataset from a thermotolerant and thermosensitive *C. gigas* family comparison (Arredondo-Espinoza et al 2023 *Comp Biochem and Phys D*).

Hypothesis: results wil be mostly similar with some new genes identified and some previously genes lost.

We found this to be true and genes with less than 10 counts varying the most between approaches. This gave us confidence that using the nf-core pipeline and current reference genome would capture most of the previously discovered genes and wouldn't wholly bias our analyses.

(A) Scatter plot showing the relationship between gene counts using the 2012 genome used in the original study ('Run 1') and the current reference genome ('Run 2'). Each dot indicates a gene and colors indicate different samples. **(B)** PCA plot showing consistent variance among samples regardless of reference genome used for analysis (Data 1, 2012 genome; Data 2, current reference). Both A and B are results from the nf-core RNAseq pipeline which is a different from the pipeline used in the original analysis. **(C)** Density plots comparing the distribution of counts across genes facetted by sample with red indicating the original analysis pipeline and 2012 genome ('old') and blue indicating the results from the nf-core RNA-seq pipeline with the current reference genome ('nf-core').

Integrated data analysis identifies differentially expressed genes that mostly distinguish disease-tolerant and sensitive phenotypes across independent studies



GeneID	PC2	PC3
LOC111101693	0.681087	0.089209
LOC111100685	0.589123	0.006932
LOC111128064	0.269451	0.38607
LOC111109914	0.164107	-0.30206
LOC111102626	0.14706	-0.25441
LOC111119295	0.139126	-0.60061
LOC111108291	-0.0496	0.326621
LOC111135887	-0.10578	0.179414
LOC111101867	-0.10985	0.19556
LOC111118025	-0.15107	-0.38617

RNAseq data from 4 independent studies of *P. marinus* tolerance and sensitivity in *C. virginica* were systematically reprocessed using the nf-core RNAseq pipeline and the current reference genome. Resulting gene counts files were merged and normalized to counts per million.

Initial PCA of the merged data showed experiment to be the greatest contributor to variance in the data. Mutual information analysis was then performed in attempt to distinguish variance from experiment from trait.

The left plot shows a PCA (PC2 and PC3) of the top 10 genes identified by the mutual information analysis which mostly distinguishes samples by trait (Xs and Os) along the diagonal regardless of experiment (color). The genes are listed in the table along with their PC2 and PC3 loadings.

CONCLUSIONS

- Preliminary systematic reanalysis with RNAseq datasets shows pipelines can process data rapidly
- Nf-core pipeline and using the current reference genome produces results consistent with original study and shouldn't bias these analyses
- Preliminary integrated data analysis appears to identify biomarkers of a disease tolerance trait

FUTURE DIRECTIONS

- Perform post-analysis data integration (assess variation in data sets independently) and then combine results for *Perkinsus* – infected *C. virginica* datasets
- Test additional normalization and integrated data analysis methods
- Include additional disease and environmental RNAseq studies in the meta-analysis
- Integrate other 'omics datasets to broaden biomarker discovery
- Host trainings on using pipelines and performing analyses
- Compile comprehensive resilience biomarker database
- Disseminate biomarker discovery resources to the community