Application of Machine Learning Algorithms for Coral Disease Fate in Caribbean Corals

Emily Van Buren[1], Kelsey Beavers[1], Nicholas MacKnight,[1] Li Wang[2], Laura Mydlarz[1]
[1] University of Texas at Arlington, Department of Biology
[2] University of Texas at Arlington, Department of Mathamatics

Background

- The genotype fragment matching design provides a baseline, exposure, and infection rates for four different species of corals exposed to two different coral diseases [1,2].
- Disease resistance follows the same pattern in both diseases.
- Through a layering of machine learning methods, various unbiased and supervised methods will help indicate any genes or biological functions that could act as a marker for disease fate.
- Unsupervised machine learning approaches include PCA, IPCA, and sIPCA.
- Supervised machine learning approaches include PLS-DA, Logistic Regression, and SVM.

Methods

- 38 healthy coral fragments were used in this study
- Transcriptome reads were trimmed, filtered, sorted, indexed, and quantified using fastp, BBSplit, and salmon [3,4,5].
- Normalized read counts were generated using tximport and DESeq2 for each individual [6,7].
- Annotations with an e-value of < 1e-5 were kept in the master filtered list. All read counts were normalized by rlog by Treatment and Species Host.
- PCA was performed on all read counts using package PCAtools [8].
- IPCA and PLSDA on all species using package mixOmics [9].
- Logistic regression and Support Vector Machine Learning-RFE (SVM-RFE) were performed using RegParallel and sigFeature with biological enrichments identified through STRINGv11 [10, 11, 12].
- All signficant genes were analyzed using ggvenn [13]

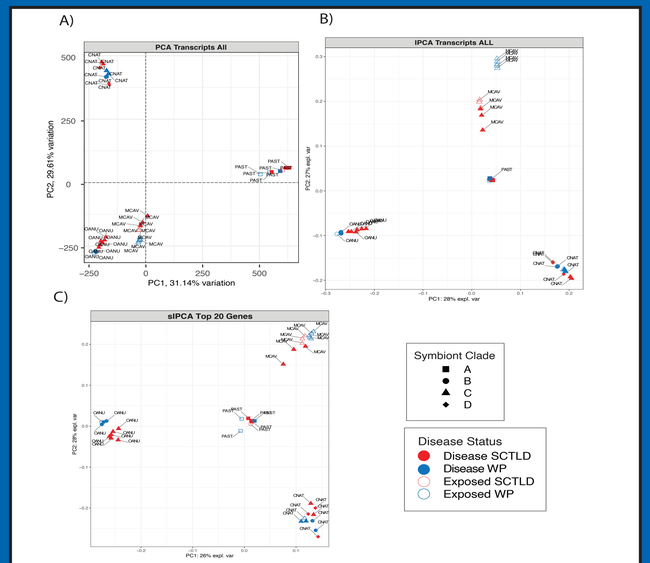


Figure 2: All unsupervised approaches include information regarding their disease status; Diseased SCTLD or Diseased WP and Exposed SCTLD and Exposed WP. Filled samples indicated diseased fate, and outlined samples indicated disease exposure only. Red indicates exposure to SCTLD pathogen, and Blue indicates exposure to WP. A) PCA has a 31.14% variation along the x-axis and a 29.61% variation along the y-axis. Groupings follow lineage, with Mcav and Oanu grouped. B) IPCA has a variation of 28% along the x-axis and 27% along the y-axis. Species separate, and Past and Mcav are close together following disease outcome. C) sIPCA looking at the top 20 genes that influence variance, show greater separation within species regarding disease outcome. There is a 26% variation on the x-axis and a 28% variation on the y-axis.

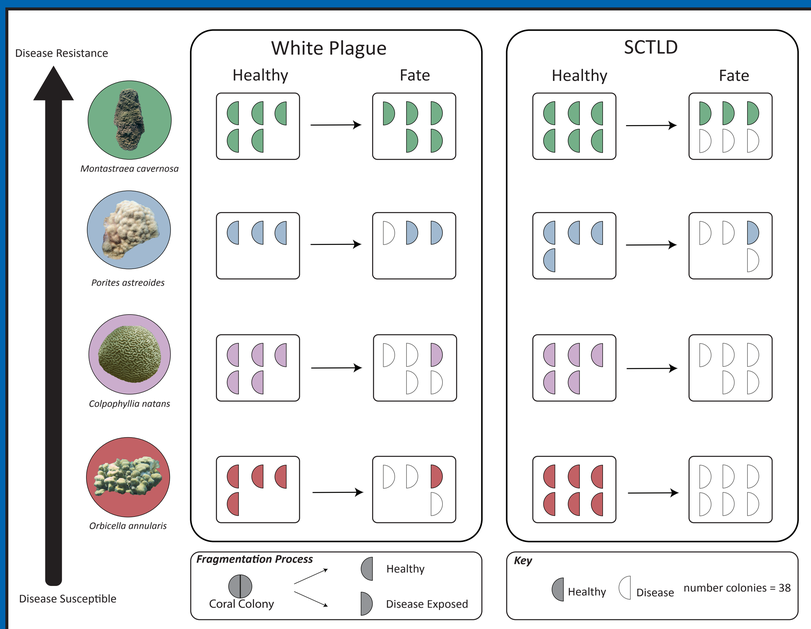


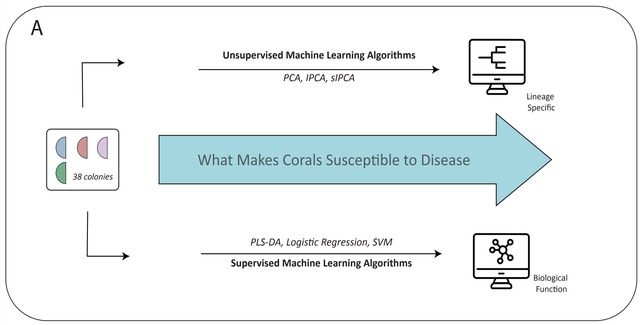
Figure 1: Samples provided in this study come from genotype match studies. Four coral species overlapped in these two studies, and their disease resistance and susceptibility were the same in both pathogen challenges. Utilizing the fragmentation process, which includes a coral colony being split and then maintained either in a healthy environment or exposed to a pathogen, we can identify the fate of the healthy samples. Using the expression and known fate, we can identify differences in genetic expression correlated to disease fate.

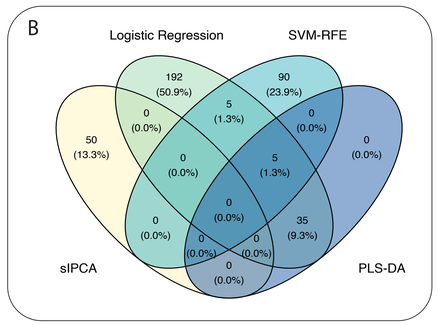
Results

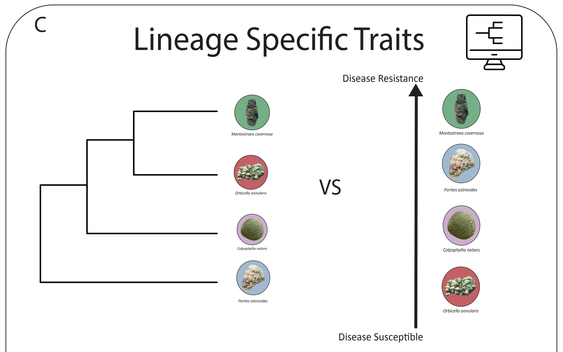
- Unsupervised approaches highlight the lineage-specific differences between healthy species.
- Introducing non-linear approaches in unsupervised decreases the noise of lineage-based genetic expression
- PLS-DA does not provide additional information already obtained from Logistic Regression
- 237 genes were significantly correlated to disease outcome through a logistic regression study
- 100 genes were significantly correlated to disease outcome through the SVM-RFE study
- Supervised approaches like SVM-RFE and Logistic Regression can be used to identify biological functions correlated to disease outcome

Conclusions

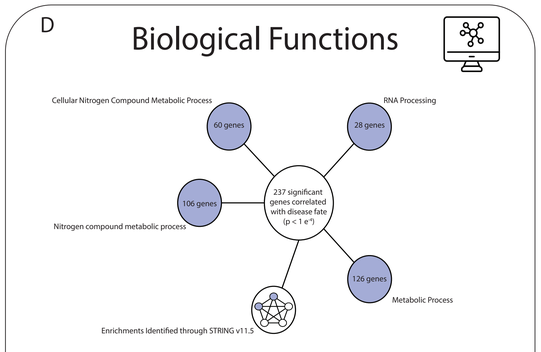
- Unsupervised approaches that decrease lineage-specific noise can help us identify species' unique characteristics that influence their disease susceptibility.
- The combined information of unsupervised and supervised approaches can help us identify a variety of genes, both lineage-specific and biologically relevant to disease outcomes.
- Binary classifiers such as SVM-RFE and Logistic Regression have more significant information than a supervised approach with multiple variables (disease assignment in PLS-DA)
- Focusing on approaches like Logistic Regression and SVM-RFE will provide greater insight into the biological processes present/absent or genetic expression variation that plays into disease outcome.







Linear approaches like PCA may mask information regarding disease outcomes due to evolutionary lineage. Introducing independence into the unsupervised algorithms through IPCA helps decrease lineage-related noise. By running sparse on IPCA, we can identify which genes are driving the differentiation.



Significantly correlated diseases identified through supervised approaches can be anlayzed for biological enrichment through STRING v11. We see enrichments in Logistic Regression through four significant biological enrichmentsand SVM-RFE having no enrichments. PLS-DA is encompases by Logistic Regression and SVM-RFE overlap.

Figure 3: All machine learning needed to be compared to see what information became valuable to the project. Looking at the 38 coral colonies that remained healthy during this study, we can organize the two significant factors that make corals susceptible to disease; lineage-specific qualities and biological functions present/absent or frontloaded by an individual. This divide in information follows the unsupervised and supervised approaches. B) Taking the top 20 significant variable genes from sIPCA, the top 40 significant genes from PLS-DA, the top 100 significantly correlated genes from SVM-RFE, and the top 237 significantly correlated genes from logistic regression, we identified overlaps of significant gene outcomes between PLS-DA and logistic regression and SVM-RFE. Logistic regression, sIPCA, and SVM-RFE provide significant genes associated with disease outcomes or lineage-specific qualities. C) When identifying lineage-specific traits, selecting the least amount of noise is crucial. All unsupervised results showed groupings based on species. The introduction of independence variation through IPCA decreased lineage-based noise to allow for identifying lineage-based traits relevant to disease outcome. D) Biological functions were found through STRING v11 analysis of 237 Logistic regression genes and 100 SVM-RFE genes. For logistic regression, four enrichments were found. For SVM-RFE no enrichments were found. These biological functions can be further examed through presence/absence and heatmap expression to see how these functions relate to disease outcome.

Citations

1. MacKnight NJ, Cobleigh K, Lasseigne D, Chaves-Fonnegra A, Gutting A, Dimos B, Antoine J, Fuess L, Ricci C, Butler C, Muller EM, Mydlarz LD, and Brandt M. 2021. Microbial dysbiosis reflects disease resistance in diverse coral species. Communications Biology, 4,679. https://doi.org/10.1038/s42003-021-02163-5
2. Meiling SS, Muller EM, Lasseigne D, Rossin A, Veglia AJ, MacKnight N, Dimos B, Huntley N, Correa AMS, Smith TB, Holstein DM, Mydlarz LD, Apprill A and Brandt ME. 2021. Variable species responses to experimental stony coral tissue loss disease (SCTLD) exposure. Front. Mar. Sci., 8: 464. https://doi.org/10.3389/fmars.2021.670829 (2021)
3. Chen S, Zhou Y, Chen Y, Gu J; fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics, Volume 34, Issue 17, 1 September 2018, Pages i884–i890, https://doi.org/10.1093/bioinformatics/bty560
4. B. Bushnell, J. Rood, and E. Singer. 2017. BBMerge – Accurate paired shotgun read merging via overlap. PLoS ONE. 12(10): e0185056.
5. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods.
6. Charlotte Soneson, Michael I. Love, Mark D. Robinson (2015): Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research
7. Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)
8. Kevin Blighe and Aaron Lun (2020). PCAtools: PCAtools: Everything Principal Components Analysis. R package version 2.0.0. https://github.com/kevinblighe/PCAtools
9. Rohart F, Gautier B, Singh A, and Le Cao K-A (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS computational biology 13(11):e1005752
10. Blighe K, Lasky-Su J (2022). _RegParallel: Standard regression functions in R enabled for parallel processing over large data-frames_. R package version 1.14.0, <https://github.com/kevinblighe/RegParallel>.
11. Das, P., Roychowdhury, A., Das, S., Roychoudhury, S. and Tripathy, S., 2020. sigFeature: novel significant feature selection method for classification of gene expression data using support vector machine and t statistic. Frontiers in genetics, 11, p.247.
12. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P‡, Jensen LJ‡, von Mering C‡. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan; 47:D607-613.
13. Yan L (2021). _ggvenn: Draw Venn Diagram by 'ggplot2'_. R package version 0.1.9, <https://CRAN.R-project.org/package=ggvenn>.