UNIVERSITY OF MIAMI
ROSENSTIEL SCHOOL of MARINE & ATMOSPHERIC SCIENCE

Using Bioinformatics to Analyze DNA Methylation in *Fundulus heteroclitus*

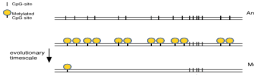
Emily Speciale
University of Miami, evs26@miami.edu

Marine Genomics Lab

Background

Summary: We are using bioinformatics to analyze levels of DNA methylation within the Atlantic killifish. We identified several genes that have been conserved over evolutionary time, mostly related to metabolic processes, giving us better insight into understanding temperature adaptations and evolution.

- *Fundulus heteroclitus* (Atlantic killifish) is a small teleost fish with a broad geographic range (southern Canada to Florida), often used in evolutionary studies because difference populations experience different environments.
- DNA methylation is an epigenetic mark that can alter CpG sites (cytosine next to guanine) within an organism's lifetime, especially in response to environmental stressors.
- Methylation causes CpG sites to mutate at a higher rate than unmethylated CpG sites, so heavily methylated regions will have fewer CpG sites than expected (CpG O/E ratio).
- We use CpG O/E ratios to identify CpG islands, which are defined as regions with at least 50% CG content and a CpG O/E ratio greater than 0.6. By identifying CpG islands, we can pinpoint evolutionarily conserved regions, which may be under epigenetic regulation.



Methods

1.) Use e!Ensembl to download data from the flank region of *F. heteroclitus*, taking a sample of 500 base pairs from each gene.
2.) Use Python to calculate CpG O/E Ratio for each gene using the equation:

$$CpG\ O/E = \frac{number\ of\ CpG}{number\ of\ C \times number\ of\ G} \times \frac{L}{l} - 1$$

in which *l* is equal to the number of nucleotides in the sample.
3.) Use the package biomaRt in R to create an overall list with each gene's ensembl gene ID, gene ontology (GO) term, transcript length, GO domain, GO term name, description, percentage of GC content, and CpG O/E ratio.
4.) Create a new list with all CpG islands from the overall list (>50% GC content and > 0.6 CpG O/E ratio).
5.) Use Cytoscape BiNGO to determine which GO categories are significantly overrepresented in the CpG island data set. Use this information to identify genes of evolutionary significance.

Results

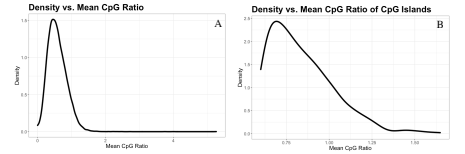


Figure 1. *(A)* Density plot of the mean CpG O/E ratio for each gene description in the data set. *(B)* Density plot of the mean CpG O/E ratio for only gene descriptions that are CpG islands (>50% GC content and > 0.6 CpG O/E ratio).

Cytoscape BINGO GO Description	Corrected P-Value	Number of Genes
Regulation of RNA Metabolic Process	2.8121E-16	89/355
Regulation of Macromolecule Metabolic Process	1.4006E-14	94/355
Regulation of Cellular Metabolic Process	3.1197E-14	105/355
Regulation of Nitrogen Compound Metabolic Process	4.6982E-14	94 /355
Regulation of Primary Metabolic Process	7.7609E-14	98/355
Regulation of Metabolic Process	4.9726E-13	105/355

Table 1. Overrepresented GO descriptions involving metabolic processes and their corresponding p-values (FDR-corrected) and number of genes in a GO category compared to the reference set.

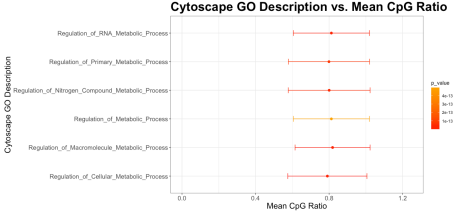


Figure 2. The mean CpG O/E ratio and standard deviation of each GO description in Table 1. Darker coloration represents a lower p-value and thus more significance of the GO term overrepresentation.

Discussion & Future Directions

- We found **47 biological-process GO descriptions that were significantly overrepresented** in the set of regions identified as CpG islands.
- Out of the top 15 most statistically significant descriptions, **six involved metabolic processes.**
- From these descriptions, we identified **105 genes that are related to metabolic processes** and have most likely been evolutionarily conserved over time, including multiple transcription factors, forkhead box proteins, homeoboxes, and GTPase activators.
- To confirm our bioinformatic analysis, we plan to use **methylation-sensitive high-resolution melting** to profile levels of DNA methylation in some of these genes.
- **As shown in the figure below,** this can be done by creating methylated and unmethylated controls and comparing their melting profiles to our sample to determine the percentage of methylation.

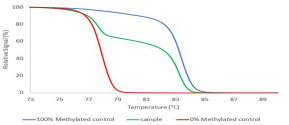


Figure 3. Methylation sensitive high-resolution melting uses the variation in melting temperature among differentially methylated samples to quantify methylation levels.

Acknowledgements

I would like to thank Melissa Drown, Dr. Marjorie Oleksiak, and Dr. Douglas Crawford for their contributions to this project. Furthermore, I would like to thank the RSMAS committee for the SURGE) Award, which will be used to present this research at a national conference.