Genomic Exploration of the Regenerative Non-model Organism *Botrylloides violaceus*

Liane Wong[1], Kamran Bastani[2], Sofia Muñiz[1], Borislav Hristov[2], Jean Davidson[1], Elena Keeling[1]

[1] Department of Biological Sciences, California Polytechnic State University, San Luis Obispo
[2] Department of Computer Science and Software Engineering, California Polytechnic State University, San Luis Obispo

Abstract

Advances in genome sequencing have made it possible to study the genetic landscape of a wide range of animals, including those with unique biological traits. In this project, we explore the genome of Botrylloides violaceus, a marine colonial ascidian capable of whole-body regeneration. The first subproject examines short conserved DNA sequences, known as motifs, that serve as binding sites for transcription factors, proteins that help turn genes on or off. We scanned the B. violaceus genome for motifs linked to known regeneration-related genes and pathways, including the Wnt signaling pathway, and mapped their locations relative to the annotated genome. This approach allowed us to identify candidate genes near these regulatory sequences that may play a role during regeneration. The second subproject looks for very small genes that may have been missed in earlier analyses, some as short as a few dozen amino acids. While tiny proteins like these have recently been found in humans, they remain largely unexplored in other animals. We used a simple gene-finding approach that looks for patterns indicating possible coding regions, and applied machine learning and protein-folding tools to predict which sequences are likely to produce real, functional proteins. Together, these two approaches help us investigate the genome of an understudied animal.

Figure 1. A *B. violaceus* colony fragment collected from Morro Bay, CA, and imaged under a dissecting microscope. An individual body, called a zooid, is circled in black.

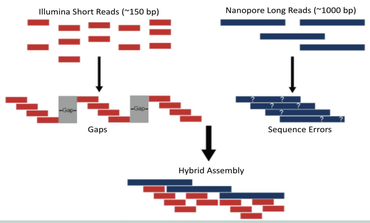

Figure 2. A diagram of a hybrid genome assembly. Image obtained from Jack Sumner.

Workflow and Goals

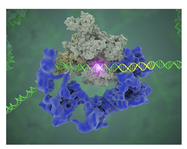

Figure 3. A diagram of a transcription factor binding to a TF binding site on a gene. Image screenshotted from NIH Division of Intramural Research.

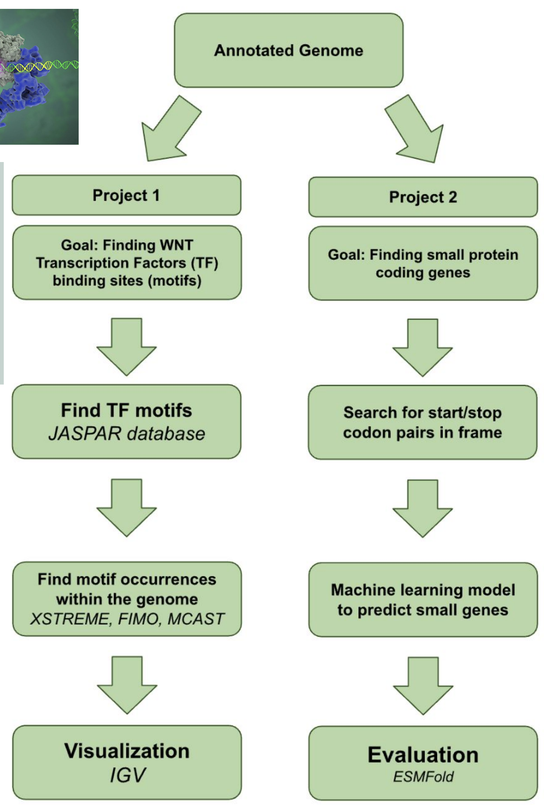

Figure 4. A summary of our workflow for projects 1 and 2 for the genomic analysis of the non-model organism *B. violaceus*.

Project #1

Finding TF Binding Site Motifs and Candidate Target Genes

MOTIF Finding Software	Advantages	Disadvantages
FIMO	- Scans motifs within the genome - Output individual motifs - Available on Galaxy	- Needs more stringent parameters to combat false positives - Does not account for spatial relationship with other motifs
XSTREME	- Output individual motif (de novo or known) - Perform motif discovery and enrichment - Can handle unaligned DNA	- Slower - Not available on Galaxy
MCAST	- Output motif clusters - Adjustable cluster size and scoring	- Not designed for scanning of individual motifs - Requires the usage of multiple motifs as input - Not available on Galaxy
MOTIF Search	- Uses P-fam database - Outputs an E-value for each motif	- Limited to protein motifs

Table 1. Moti finding softwares with advantages and disadvantages.

Transcription Factor	Motif Sequence	Number Found
LEF1		62
FOXK1		117

Table 2: Analysis of LEF1 and FOXK1 TF consensus motifs sequences (sourced from the JASPAR database) and the number of occurrences identified in the genome using FIMO (p = 1e-07).

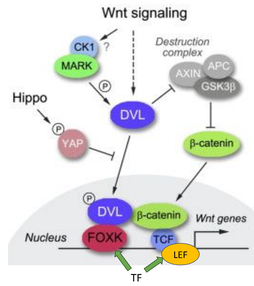

Figure 5. WNT signaling can lead to activation of both LEF and FOXK1 TFs (green arrow). Image taken from Developmental Cell[13].

Approach to Visualizing TF Binding Site Motifs in IGV

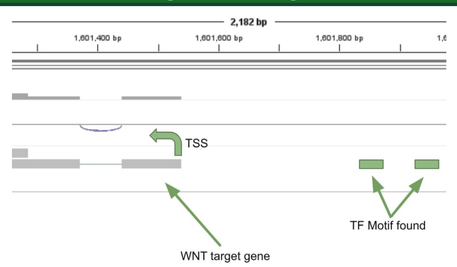

Figure 6. IGV visualization showing found TF motif upstream of the transcription start site (TSS) of WNT target genes.

Current Work:

- Visualization of TF binding site motifs found in the genome using IGV to identify nearby (within 10kb) genes.
- Identification of genes found within 10kb of TF binding site motifs in IGV.
- Associate TF binding sites to closest genes using bedtools.

Next Steps:

- Functional annotation of candidate target genes using KEGG.
- Aligning motif sites found in *B. violaceus* to related ascidian species to compare differences in candidate target genes.
- Identification of other TF motifs outside of the WNT signaling pathway.

Project #2

Pipeline to Find Potential Small Genes

Use of ESMFold to predict protein folding based on amino acid sequence

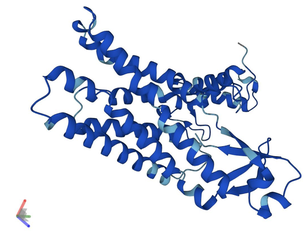

Figure 7a. ESMFold Predicted structure for amino acid sequence of a large protein coding gene (~380 amino acids).

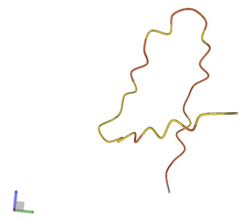

Figure 7b. ESMFold predicted protein structure for a small gene encoding a 39- amino acid sequence: KLYPVELMTRISLKKNPRSFYLIISPDICIESTLTTNR.

Validation of Methods

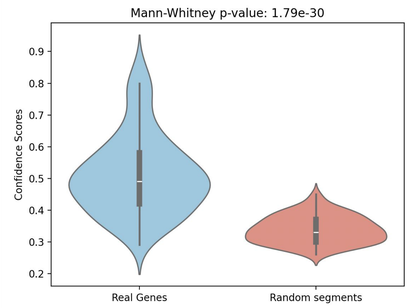

Figure 8. Distribution of confidence scores generated by ESMFold on annotated *B. Violaceus* genes under 100 amino acids and random segments selected from the genome. The p-value indicates that the real proteins have higher confidence scores than random sequences.

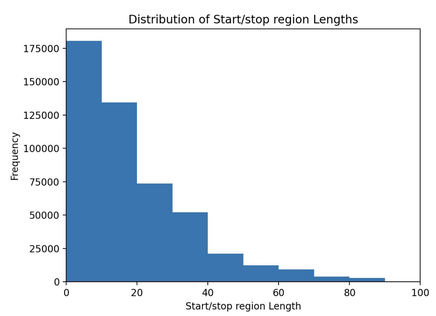

Figure 9. Distribution of in frame start/stop codon region lengths in assembled *B. Violaceus* genome. These lengths are much shorter than expected

Current Issues:

- This is resulting in distances between in frame start/stop codon regions being far too short
  - This is a draft genome, we expect random sequence errors
  - This likely results in misidentified start and stop codons which would cause short start/stop regions

Work in Progress/Next Steps:

- Complete distribution of ESMFold scores for all real genes up to 400 amino acids (as in Fig. 8)
- Switch to analysis with the *Drosophila* genome as a proof of concept
- Develop machine learning model to predict candidate genes for micro-proteins
- Run ESMFold on predicted candidate genes and compare to baseline distribution

Acknowledgements

We would like to thank the Baker/Koob Endowments for their generous funding and support in sequencing of the draft genome. Additionally, we give our thanks to Dr. Elena Keeling, Dr. Jean Davidson, and Dr. Borislav Hristov for their continuous support and guidance on this project.

References

1. Jack T. Sumner, Cassidy L. Andrasz, Christine A. Johnson, Sarah Wax, Paul Anderson, Elena L. Keeling, Jean M. Davidson. (2023). DOI: 10.1093/g3journal/jkad181.
2. Deutsch et al. (2024) DOI: 10.1101/2024.09.09.612016
3. Wanniarachchi, D.V., Viswakula, S. & Wickramasuriya, A.M. The evaluation of transcription factor binding site prediction tools in human and Arabidopsis genomes. *BMC Bioinformatics* 25, 371 (2024). https://doi.org/10.1186/s12859-024-05995-0
4.Jayaram N, Usvyat D, R Martin AC. Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics. 2016 Nov 2;17(1):547. doi: 10.1186/s12859-016-1298-9. PMID: 27806697; PMCID: PMC6889335.
5. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K, Kumar V, Lemma RB, Lucas J, Chèneby J, Baranasic D, Khan A, Fornes O, Gundersen S, Johansen M, Hovig E, Lenhard B, Sandelin A, Wasserman WW, Parcy F, Mathelier A JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles *Nucleic Acids Res.* 2024 Jan 5;52(D1):D174-D182.; doi: 10.1093/nar/gkad1059
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). PMID: 2231712. DOI: 10.1016/S0022-2836(05)80360-2.
7. Li, H. (2021). DOI: 10.1093/bioinformatics/btab705. v2.28-r1209.
8. Nip, K.M., Hafezqorani, S., Gagalova, K.K. et al. (2023). DOI: 10.1038/s41467-023-38553-y.
9. Mosè Manni, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, Evgeny M Zdobnov. (2021). DOI: 10.1093/molbev/msab199. v5.7.1.
10. https://www.ncbi.nlm.nih.gov/books/NBK279690/
11. https://www.blast2go.com/
12. http://eggnog-mapper.embl.de/
13. Zhang, Y., Wang, X. Targeting the Wnt/β-catenin signaling pathway in cancer. *J Hematol Oncol* **13,** 165 (2020). https://doi.org/10.1186/s13045-020-00990-3
14. https://esmatlas.com/resources?action=fold