Kiley Huffman

QBIO-490x: Directed Research

PI: Dr. Julia Schwartzman

July 18, 2025

Beyond 16S rRNA: Identifying Alternative Marker Genes for Strain-Level Differentiation in *Vibrio splendidus*

Abstract

Strain-level resolution is essential for understanding the role of bacterial populations in both natural ecosystems and laboratory environments. Scientists have previously relied on the *16S rRNA* gene for bacterial identification and classification (Hug et al., 2016); however, *16S rRNA* often lacks the resolution needed to distinguish closely related strains for some genera of bacteria (Leunda‑Esnaola et al., 2017). In this project, I address this limitation by attempting to identify an alternative strain-level molecular marker for the marine bacterium *Vibrio splendidus* that is capable of differentiating closely-related strains.

Specifically, I sought out molecular markers with at least 5 nucleotide polymorphisms over a 300-400 bp region of DNA. Using a dataset of aligned gene sequences from 17 *V. splendidus* genomes, I analyzed sequence conservation, variability, and primer suitability to identify candidate molecular markers. Ultimately, two genes were selected for their ability to distinguish closely-related strains. The first candidate gene selected was *ftsY*, which encodes a signaling protein involved in the expression of integral membrane proteins. The second candidate gene was *metC0, a* cystathionine beta lyase that is involved in methionine biosynthesis. These two genes contained variable regions that were wedged in between two perfectly conserved 20-mers, making them ideal candidates for primer design.

Introduction

Populations shape the ecological function, evolutionary dynamics, and disease progression of bacteria. Within bacterial populations, small genetic differences in strains can lead to major changes in phenotypes—affecting growth rates, pathogenicity, metabolic capabilities, and responses to environmental stressors (Van Rossum et al, 2020). Traditional molecular markers like *16S rRNA* have long served as the go-to method for bacterial phylogeny, but their limited resolution at the strain-level poses challenges for fine-scale differentiation (Liu et al., 2012). Species within the *Vibrio* genus, in particular, exhibit high sequence conservation in the *16S rRNA* gene, despite genome content and conservation diverging greatly, making it difficult to resolve intra-species relationships using standard molecular markers. Protein-coding genes such as *gyrB*, *rpoB*, *hsp60*, and *recA* have demonstrated greater variability than *16S rRNA* in bacterial groups, making them promising candidates for strain-level analysis in other taxa. For instance, *gyrB* and *recA* have been shown to sequence divergence, effectively separating closely related *V. splendidus*-related strains into distinct clades, and matching DNA–DNA hybridization results that are not resolved by 16S rRNA alone (Pascual et al., 2010). Furthermore, multilocus sequence analysis (MLSA), which utilizes several housekeeping genes, has been shown to significantly outperform 16S rRNA alone in discriminating *Vibrio* species (Thompson et al., 2004).

In this study, I sought to identify alternative marker genes capable of resolving strain-level diversity within the *Vibrio splendidus* clade. *V. splendidus* is a species of marine bacteria that is widely found in coastal ecosystems. Even minimal genetic variation in *V. splendidus* can produce major differences in ecological impact. For example, several *V. splendidus* strains isolated from diseased turbot larvae shared a high genomic similarity, but only a subset caused high larval mortality (Toranzo et al., 1999). Furthermore, a study conducted by Miranda et al. found that three *V. splendidus* strains, isolated from scallop hatcheries in Chile, all caused severe larval mortality, but with different symptoms (Miranda et al., 2014). To develop a molecular marker for *V. splendidus* strains, I identified two candidate genes—*FtsY* and *MetC0*—from a dataset of aligned gene sequences across 17 *V. splendidus* genomes. My analysis of nucleotide sequence alignments for orthologs of these genes demonstrate that they could serve as reliable markers for strain-level resolution of *V. splendidus* populations. These genes were selected based on various criteria, including sequence conservation, pairwise identity, inter-strain variability, and primer design suitability. These genes were then evaluated for their discriminatory power using PCR amplification across multiple strains.

Materials

Aligned FASTA files were generated by Eesha Rangani, a previous PhD student in the Schwartzman lab, using the comparative genomics software Anvi'o to identify all orthologous genes present in single copy across 17 *V. splendidus* genomes. These 2947 nucleotide alignment files were stored in a single directory (aligned_files) and analyzed in Jupyter Notebook (Version 7.0.8; Jupyter Notebook). Each FASTA file contained one orthologous gene alignment across all strains. Prior to analysis, these alignments were parsed using **Biopython (**Version 1.85; Biopython) and all of the sequences were trimmed to the length of the shortest sequence within each file to ensure consistent comparisons. Specifically, the **SeqIO** module was used to read multiple sequence alignment (MSA) files in the FASTA format. First, the *SeqIO.parse()* function was employed to load each aligned sequence into **Biopython's SeqRecord** objects, which were then converted to strings for downstream analysis of conservation, variability, and pairwise identity. Thus, all analyses in this study were conducted using sequence data parsed directly from aligned FASTA files via **SeqIO**, allowing for straightforward integration into custom consensus-matrix and k-mer scanning functions.

Methods: Primer Identification

To evaluate sequence conservation, variability, and primer design suitability across orthologous genes in *Vibrio splendidus*, I developed a set of Python functions (defined in **msa_functions.py**) for processing multiple sequence alignments (MSAs), identifying conserved and variable regions, and computing molecular properties relevant to marker selection and primer-design. All of the coding files for this project can be found on my GitHub.

Sliding Window Analysis and Perfect Match Detection

Each alignment was processed using a sliding-window approach with a block size (k-mer length) of 20 base pairs. For each window, a consensus matrix was constructed to count the presence of each nucleotide (A, T, C, G) at each position across all sequences. Then, the number of positions with zero representation for at least one nucleotide were tabulated using the *count_zeroes_in_consensus()* function. These zero-counts were used as a measure of sequence conservation, with higher zero counts indicating more conserved regions and lower zero counts indicating greater nucleotide diversity. Perfect matches were identified using the *get_perfect_match_indices()* function which returns indices of windows where all sequences in the block are identical (location of the perfectly conserved 20-mers). These results were saved to *all_perfect_matches.csv*.

Variable Region Identification

Variable regions were defined as the sequences between the perfectly conserved 20-mers. The function *find_variable_regions()* identified all variable regions with lengths between 200–400 bp. This size range was chosen to reflect typical design constraints for molecular markers used for high-throughput sequencing. Using the function *collect_all_variable_regions(),* variable regions across all alignments were extracted and exported to *all_variable_regions.csv*. Each entry included the filename, start and end coordinates, and region length.

Consensus Matrix Analysis and Zero Counts

To quantify conservation within 20-bp windows, a consensus matrix was generated for each alignment using the *count_zeroes_in_consensus()* function which counts how many bases (A, T, C, G) are absent at each position across sequences. The number of zeroes reflects how conserved a column is: highly conserved positions result in more zeroes (due to lack of diversity), whereas variable positions show fewer zeroes. For each alignment, several statistics were computed, including the maximum, mean, and variance of zeroes across all 20-bp blocks, the count of 76-zero blocks (regions with complete conservation), the number of perfect matches, the distance between perfect matches, and the normalized perfect match rate (number of perfect 20-mers divided by the total number of possible 20-mers in the alignment). These metrics were all calculated programmatically, using functions defined in **msa_functions.py**, and exported to *alignment_stats.csv*.

Pairwise Identity Analysis

Sequence similarity between strains was assessed using the *calculate_pairwise_identity()* function which computes the minimum, maximum, and average pairwise percent identity while ignoring ambiguous bases (N) and gaps. These metrics were recorded for each gene alignment in *alignment_stats.csv.*

Primer Candidate Identification and Evaluation

To support potential PCR-based validation of marker regions, conserved 20-mers were extracted using the *get_conserved_20mer()* function which scans the upstream and downstream 20-mers of variable regions to identify ideal primer sites. These results were then filtered by specific criteria (detailed below) and recorded in *best_strain_differentiaion_primers.csv*, which is ranked by variability (calculated as Shannon entropy). From *best_strain_differentiaion_primers.csv*, we selected the two regions with the highest variability, which were located within the *ftsY* and *metC0* ortholog alignments. We then ordered the primers for each of these regions for PCR Analysis.

Filtering Criteria

1. GC Content (40-60%): This GC range provides a balance between primer stability and the risk of secondary structure formation (Thermo Fisher Scientific, n.d.; Bitesize Bio, 2015).

2. Temperature (55–60°C): This temperature range ensures that primers anneal efficiently and specifically to the template DNA during the annealing phase of PCR.

3. Presence of GC Clamp: The presence of a GC clamp, typically defined as having one or more G or C nucleotides within the last five bases at the 3′ end of the primer, is also recommended for ideal primer design (Thermo Fisher Scientific, n.d.).

4. Amplicon Length (200-400bp): This amplicon length range is short enough to amplify and resolve easily on an agarose gel, yet long enough to contain informative sequence data for downstream applications (Bitesize Bio, 2015).

Results

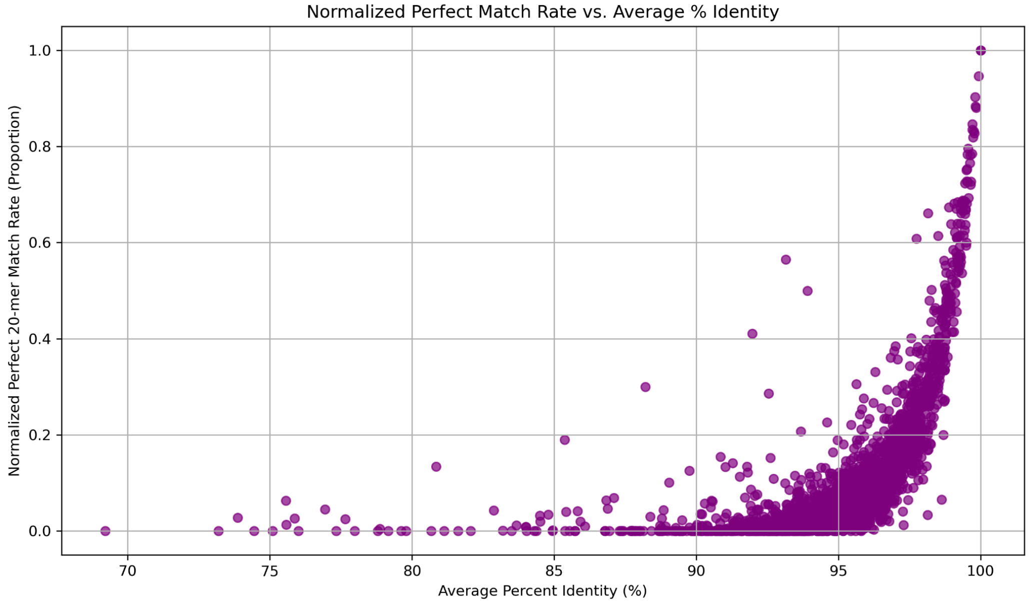


Figure 1. Each dot represents the average percent identity vs. the normalized perfect match rate for every perfectly conserved 20-mer across each gene alignment.

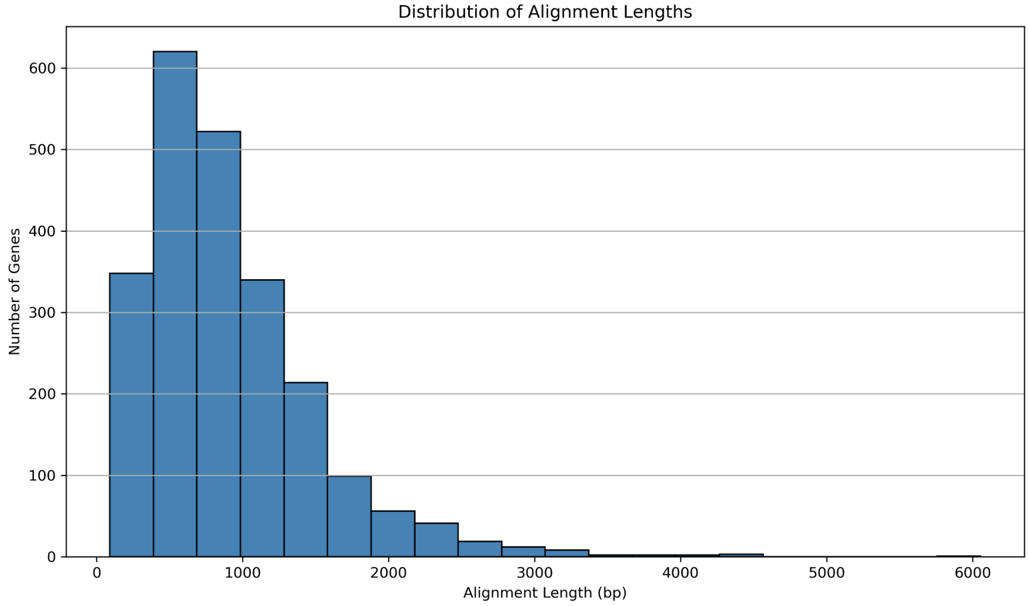


Figure 2. This histogram illustrates the distribution of alignment lengths of the genes in the dataset of FASTA files where each individual file contains one orthologous gene alignment across all 17 strains of *V. splendidus.*

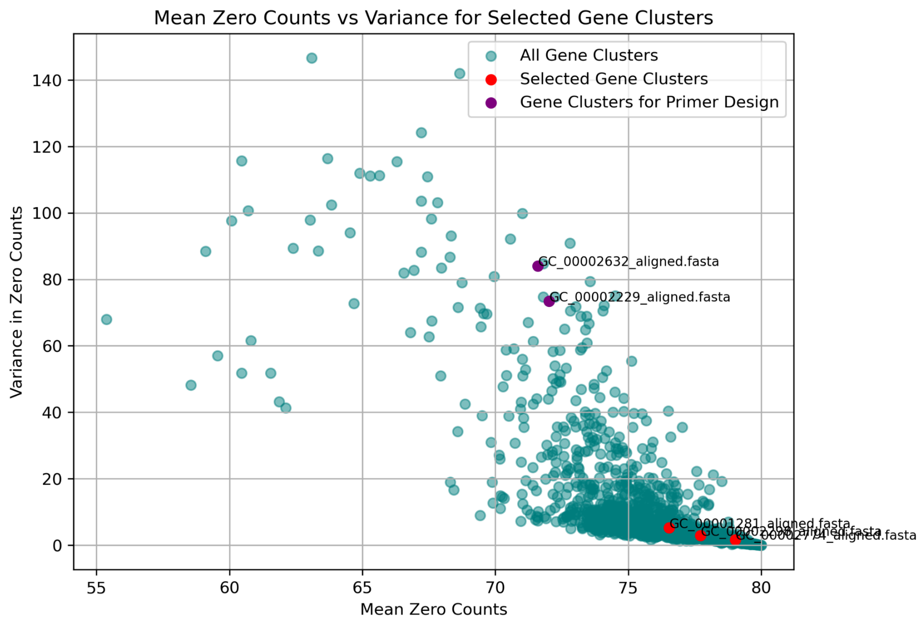


Figure 3. Each dot represents the average percent identity vs. the mean of the zero counts for the consensus matrices calculated for all of the 20-mer windows in each alignment. The red dots represent genes that have previously been used as molecular markers *gyrB* (GC_00001281), *hsp60* (GC_00002298), and *rpoB* (GC_00002774). The purple dots represent the two gene clusters selected for primer design: *metC0* (GC_00002229) and *ftsY* (GC_00002632)

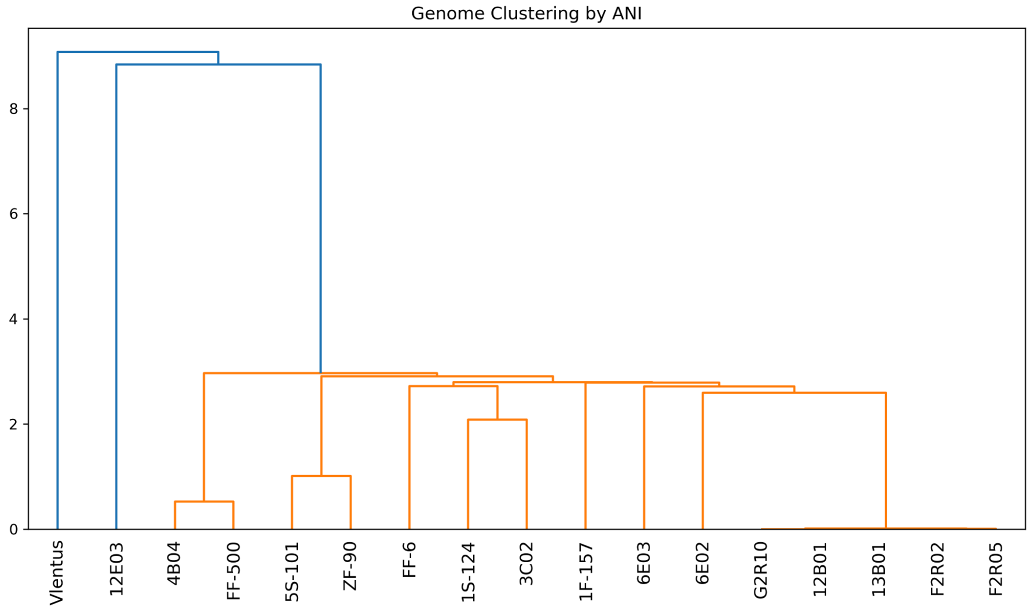
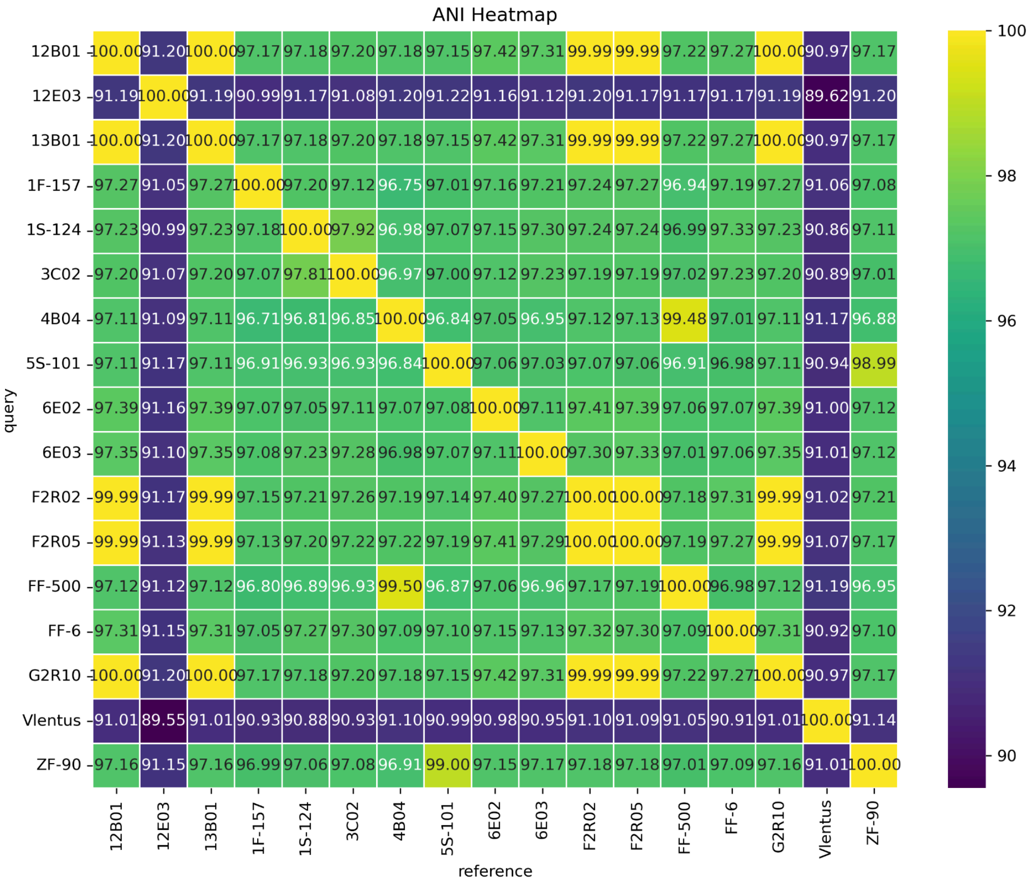


Figure 4. Genome clustering of the 17 *V. Splendidus* genomes done by Average nucleotide identity (ANI).



	12B01	12E03	13B01	1F-157	1S-124	3C02	4B04	5S-101	6E02	6E03	F2R02	F2R05	FF-500	FF-6	G2R10	Vlentus	ZF-90
12B01	100.00	91.20	100.00	97.17	97.18	97.20	97.18	97.15	97.42	97.31	99.99	99.99	97.22	97.27	100.00	90.97	97.17
12E03	91.19	100.00	91.19	90.99	91.17	91.08	91.20	91.22	91.16	91.12	91.20	91.17	91.17	91.17	91.19	89.62	91.20
13B01	100.00	91.20	100.00	97.17	97.18	97.20	97.18	97.15	97.42	97.31	99.99	99.99	97.22	97.27	100.00	90.97	97.17
1F-157	97.27	91.05	97.27	100.00	97.20	97.12	96.75	97.01	97.16	97.21	97.24	97.27	96.94	97.19	97.27	91.06	97.08
1S-124	97.23	90.99	97.23	97.18	100.00	97.92	96.98	97.07	97.15	97.30	97.24	97.24	96.99	97.33	97.23	90.86	97.11
3C02	97.20	91.07	97.20	97.07	97.81	100.00	96.97	97.00	97.12	97.23	97.19	97.19	97.02	97.23	97.20	90.89	97.01
4B04	97.11	91.09	97.11	96.71	96.81	96.85	100.00	96.84	97.05	96.95	97.12	97.13	99.48	97.01	97.11	91.17	96.88
5S-101	97.11	91.17	97.11	96.91	96.93	96.93	96.84	100.00	97.06	97.03	97.07	97.06	96.91	96.98	97.11	90.94	98.99
6E02	97.39	91.16	97.39	97.07	97.05	97.11	97.07	97.08	100.00	97.11	97.41	97.39	97.06	97.07	97.39	91.00	97.12
6E03	97.35	91.10	97.35	97.08	97.23	97.28	96.98	97.07	97.11	100.00	97.30	97.33	97.01	97.06	97.35	91.01	97.12
F2R02	99.99	91.17	99.99	97.15	97.21	97.26	97.19	97.14	97.40	97.27	100.00	100.00	97.18	97.31	99.99	91.02	97.21
F2R05	99.99	91.13	99.99	97.13	97.20	97.22	97.22	97.19	97.41	97.29	100.00	100.00	97.19	97.27	99.99	91.07	97.17
FF-500	97.12	91.12	97.12	96.80	96.89	96.93	99.50	96.87	97.06	96.96	97.17	97.19	100.00	96.98	97.12	91.19	96.95
FF-6	97.31	91.15	97.31	97.05	97.27	97.30	97.09	97.10	97.15	97.13	97.32	97.30	97.09	100.00	97.31	90.92	97.10
G2R10	100.00	91.20	100.00	97.17	97.18	97.20	97.18	97.15	97.42	97.31	99.99	99.99	97.22	97.27	100.00	90.97	97.17
Vlentus	91.01	89.55	91.01	90.93	90.88	90.93	91.10	90.99	90.98	90.95	91.10	91.09	91.05	90.91	91.01	100.00	91.14
ZF-90	97.16	91.15	97.16	96.99	97.06	97.08	96.91	99.00	97.15	97.17	97.18	97.18	97.01	97.09	97.16	91.01	100.00

Figure 5. The heatmap represents the percent conservation across all 17 of the closely-related V. splendidus strains.

Conclusion

Overall, this study sought to identify and evaluate molecular markers for strain-level differentiation within closely related bacterial clades, as an alternative to 16S rRNA sequencing. By analyzing a comprehensive set of orthologous gene alignments from 17 *V. splendidus* genomes, we identified two alternative genetic markers with the potential for a significant discriminatory power. Two genes, *FtsY* and *MetC0*, were found to be strong candidates for strain-level resolution due to the presence of variable regions flanked by conserved 20-mer sequences within each gene. This nucleotide composition made these two genes well-suited for primer design and PCR. Currently, the Schwartzman Lab and I have been in the process of testing our primers through PCR analysis techniques on several *V. splendidus* strains. Our findings have the potential to enable researchers to better study and understand fine-scale diversity among bacterial populations, providing a valuable framework for future microbial ecology studies.

Appendix: Code Repository

The code for this study can be found on GitHub at this link:

https://github.com/kileyhuffman21/VibrioSplendidus_MSA_Project.git

References

Bitesize Bio. (2015). *Designing for luck: 8 basic concepts for designing primers for a standard PCR*. https://bitesizebio.com/24608/designing-luck-8-basic-concepts-for-designing-primers-for-a-standard-pcr/

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology, 1*(5), 16048. https://doi.org/10.1038/nmicrobiol.2016.48

Leunda-Esnaola, A., Bunin, E., Arrufat, P., Pearman, P. B., & Kaberdin, V. R. (2024). Harnessing the intragenomic variability of rRNA operons to improve differentiation of *Vibrio* species. *Frontiers in Microbiology, 15*, 11061105. https://doi.org/10.3389/fmicb.2024.11061105

Liu, W., Li, L., Khan, M. A., & Zhu, F. (2012). Popular molecular markers in bacteria. *Molecular Genetics, Microbiology and Virology, 27*(3), 103–107. https://doi.org/10.3103/S0891416812030056

Miranda, C. D., Rojas, R., Geisse, J., & Romero, J. (2014). Pathogenicity of *Vibrio splendidus*-related strains isolated from diseased commercial scallop larvae. *Journal of Invertebrate Pathology*, 119, 38–43. https://doi.org/10.1016/j.jip.2014.04.004

Pascual, J., Macián, M. C., Arahal, D. R., Garay, E., & Pujalte, M. J. (2010). Multilocus sequence analysis of the central clade of the genus *Vibrio* by using the 16S rRNA, recA, pyrH, rpoD, gyrB, rctB and toxR genes. *International Journal of Systematic and Evolutionary Microbiology*, 60(1), 154–165. https://doi.org/10.1099/ijs.0.010702-0

Premier Biosoft. (n.d.). *PCR primer design guidelines*. https://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html

Thermo Fisher Scientific. (n.d.). *PCR primer design tips*. https://www.thermofisher.com/blog/behindthebench/pcr-primer-design-tips/

Thompson, F. L., Iida, T., & Swings, J. (2004). Biodiversity of vibrios. *Microbiology and Molecular Biology Reviews*, 68(3), 403–431. https://doi.org/10.1128/MMBR.68.3.403-431.2004

Toranzo, A. E., Magariños, B., & Romalde, J. L. (1999). *Vibrio* species causing diseases in aquaculture. *Journal of Applied Microbiology*, 85(S1), 101S–107S. https://doi.org/10.1111/j.1365-2672.1999.tb04786.x

Van Rossum, Thea, et al. "Diversity within species: interpreting strains in microbiomes." *Nature Reviews Microbiology* 18.9 (2020): 491-506. https://doi.org/10.1038/s41579-020-0368-1