

Supplementary Information: Evolutionary dynamics and selection modeling of climate-driven *Vibrio vulnificus* expansion into high-latitude environments

Hevar N. Barznji^{1*}

¹MLS Department, Komar University of Science and Technology, Sulaymaniyah, Iraq

*Corresponding author email: hevar.nyazf17@komar.edu.com

July 23, 2026

Overview & Computational Repository

All computational workflows, raw analytical scripts, configuration files, environment definitions, and supplementary data generated for this study are hosted in the project's official GitHub repository:

https://github.com/Hevar1099/vibrio_evolution

This repository provides a fully reproducible, containerized pipeline engineered in **Snakemake** and **Docker** to execute end-to-end genome quality control, pangenome assembly, machine learning feature discovery, and downstream evolutionary selection modeling of *Vibrio vulnificus* isolates from high-latitude brackish environments.

1 Repository Structure & Artifacts

The repository is structured to separate pipeline logic, execution scripts, data processing stages, and final visualization outputs:

```
vibrio_evolution/
|-- Dockerfile           # Container environment specification
|-- Snakefile            # Primary Snakemake workflow
    orchestrator
|-- config/
|   '-- config.yaml      # Pipeline parameter definitions
|-- scripts/             # Modular Python and R analytical
    scripts
|-- demo_outputs/        # Pre-rendered figures and
    supplementary tables
|   |-- figures/         # High-resolution vector and PNG
    figures
|   '-- tables/          # Delimited tabular summary datasets
'-- results/            # Workflow execution output targets
```

2 Pipeline Workflow & Analytical Stages

The workflow processes genomic datasets through four sequential execution stages:

1. **Metadata Acquisition and Cohort Filtering:** Public genome metadata integration, geographical/niche classification, and cohort downsampling (`scripts/s1_*.py`).
2. **Quality Control and Taxonomic Validation:** Assembly evaluation via `CheckM2` and taxonomic assignment via `GTDB-Tk` (`scripts/s2_*.py`).
3. **Annotation and Pangenome Assembly:** Structural gene annotation via `Prokka`, pangenome construction via `Panaroo`, and machine learning consensus feature discovery on the accessory genome (`scripts/s3_*.py`).
4. **Selection Modeling and Evolutionary Dynamics:** Core phylogenomic inference, selection candidate annotation, mobile genetic element (MGE) permutation testing, and publication figure generation (`scripts/s4_*.py`, `scripts/s6_*.py`, `scripts/s7_*.R`).

—

3 Supplementary Figures and Tables Index

Pre-computed preview files, complete tabular summaries, and publication-ready figures are accessible directly within the repository under `demo_outputs/`:

Supplementary Figures

- **CheckM2 Quality Control Scatter Plot:** `demo_outputs/figures/checkm2_qc_scatter.png`
- **CheckM2 Quality Metric Distributions:** `demo_outputs/figures/checkm2_qc_distributions.png`
- **Pangenome Matrix Distribution:** `demo_outputs/figures/pangenome_matrix_distribution.pdf`
- **ML Consensus Feature Summary:** `demo_outputs/figures/figure1b_ml_consensus.pdf`
- **Master Pipeline Layout Canvas:** `demo_outputs/figures/layout_a4_box_matrix.pdf`

Supplementary Tables

- **Table S1 — Classifier Performance Summary:** `demo_outputs/tables/table1b_ml_performance.tsv`
- **Table S2 — ML Consensus Gene Annotations:** `demo_outputs/tables/ml_consensus_annotations.tsv`
- **Table S3 — Candidate Selection Gene List:** `demo_outputs/tables/selection_evolution_annotations.tsv`
- **Table S4 — Master Supplementary Result Summary:** `demo_outputs/tables/supplementary_table_1.tsv`

—

4 Execution & Reproducibility Guide

4.1 Environment Setup

To ensure complete computational reproducibility across platforms, a pre-configured execution container is provided.

```
# 1. Clone the repository
git clone https://github.com/Hevar1099/vibrio_evolution.git
cd vibrio_evolution

# 2. Build the Docker image
docker build -t vibrio_evolution .

# 3. Launch interactive container session
docker run --rm -it -v "$(pwd):/workspace" -w /workspace
vibrio_evolution bash
```

4.2 Workflow Execution

Inside the containerized environment, individual analysis steps or full pipelines can be executed via Snakemake:

```
# Validate pipeline graph (Dry-run)
snakemake --snakefile Snakefile --cores 1 --dryrun

# Execute complete analytical workflow
snakemake --snakefile Snakefile --cores 16
```

4.3 Hardware Requirements

Due to memory-intensive CheckM2 evaluation, GTDB-Tk taxonomic alignment, and Panaroo graph assembly stages, full execution from raw genomes requires a high-performance compute (HPC) cluster or server equipped with at least **16 CPU cores** and **128 GB RAM**.