**Title:** Draft Genome Assemblies of Three Halophilic Bacteria From Soil, Salt Lick, and a Fish Tank

# Abstract:

Halophiles represent an industrially important but still understudied group of organisms. Here, we present a draft genome sequence of three different salt tolerant bacterial strains from the genera Halomonas, Bacillus, and Staphylococcus. These organisms were isolated from soil, a cow salt lick, and a fish tank, and then cultured at 200 g/L NaCl, sequenced, and assembled into draft genome assemblies.

# Background:

Halophiles have been identified as significant sources of stable enzymes that work in extremely high salinity and have a broad range of biotechnological applications (1). There is therefore a pressing need to expand the currently limited genome database of halophilic microorganisms. To help address this need, we cultured three environmental samples and isolated DNA sequences as part of a Course-based Undergraduate Research Experience. From these, we identified three different salt tolerant strains from our assembled draft genomes. Our work can be utilized in studying halophilic mechanisms by providing a broader scope of genomes for future reference (2).

# Methods:

## Specimen Collection and Culturing

In order to search for halophilic microbes, we collected 17 samples from a range of environments from around Davis, California, USA. After one week of incubation at 37°C, three samples successfully resulted in bacterial growth in high salt growth medium (200 g/L NaCL). These three samples were sourced in late September or early October 2021 from: 1) soil from the Davis Arboretum, 2) a cow salt lick, and 3) the remains of evaporated water on the glass top of a saltwater fish tank containing a single rabbitfish (*Siganus Guttatas*) held at room temperature (~22°C) under ambient light. To isolate single colonies, we streaked each liquid sample onto high salt agar plates (200g/L NaCl). Plates were incubated for 5-6 days at 37°C. We then picked single colonies from each sample and cultured them in growth medium for one week at 37°C. Samples where then pelleted and DNA was extracted. For liquid culture, Halobacterium 372 (<https://www.dsmz.de/microorganisms/medium/pdf/DSMZ_Medium372.pdf>) was used. Agar was added to this media to make bacterial growth plates. The soil sample was streaked once, the salt lick sample three times, and the tank sample twice, for a total of six samples.

## Genomic DNA Purification, Sequencing, and Bioinformatics

We purified genomic DNA from each successfully cultured sample using the Monarch DNA extraction kit (New England Biolabs) according to their Quick Protocol. The sample was pelleted and lysed with phosphate buffered saline, lysosomes, tissue lysis buffer, proteinase, and RNAase. DNA was then purified using gDNA binding buffer, a gDNA Purification column elution buffer and wash buffer. To prepare the samples for sequencing, DNA was further purified via ethanol precipitation. DNA sequencing and bioinformatics analysis was performed by MicrobesNG (<https://microbesng.com/>, see <https://microbesng.com/documents/5/MicrobesNG_Methods_Document_-_PDF.pdf> for details). Briefly, each sample was lysed, purified, and then sequenced on an Illumina sequencer using 250bp paired-end reads. The resulting reads were trimmed via Trimmomatic 0.30 (3), assembled via SPAdes 3.7 (4), and annotated using Prokka 1.11 (5).

## Genome Assembly and Quality Assessment

We first used the ContEst16S tool assess contamination in our samples by comparison against known 16S rRNA sequences (6). We considered samples contaminated if the results returned more than one 16S match. Three out of our six samples had contamination and were not further analyzed. We then cleaned and assessed the remaining sequences using the BioStrings R package by filtering out any contigs that were shorter than 500 bp, and then calculated the number of contigs, length of each contig, total genome size, N50, and L50 values, and GC content for the resulting final assemblies (Table 1, 7). To determine the possible species identification for each sample, we then ran the largest 16S sequences obtained by the ContEst16S tool through the NCBI’s online BLASTn (Nucleotide Basic Local Alignment Search Tool) tool (8).

# Results:

Blast results against our largest 16S sequences from each uncontaminated samples was found to be similar to previously sequenced samples, each with very high percent identity (>99%): the fish tank sample matched to a one to a sample from a known halophilic *Halomonas* species, the Arboretum soil sample to an unidentified *Bacillus* species, and the salt lick sample equally well (identical e-values with 100% query coverage) to many (20+) *Staphylococcus* species (see Table 1).

# Data Availability:

Assemblies and annotations from all six (contaminated and uncontaminated) samples are available from GenBank with the accession number (will be inserted when available). The rmarkdown script used to analyze the assemblies and produce this manuscript is also available at <https://github.com/hemstrow/BIS23_manuscript_2022>.

# References:

1. Liu C, Baffoe DK, Zhan Y, Zhang M, Li Y, Zhang G. 2019. [Halophile, an essential platform for bioproduction](https://doi.org/10.1016/j.mimet.2019.105704). Journal of Microbiological Methods 166:105704.

2. Becker EA, Seitzer PM, Tritt A, Larsen D, Krusor M, Yao AI, Wu D, Madern D, Eisen JA, Darling AE, Facciotti MT. 2014. [Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response](https://doi.org/10.1371/journal.pgen.1004784). PLoS genetics 10:e1004784–e1004784.

3. Bolger AM, Lohse M, Usadel B. 2014. [Trimmomatic: a flexible trimmer for Illumina sequence data](https://doi.org/10.1093/bioinformatics/btu170). Bioinformatics 30:2114–2120.

4. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. [SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing](https://doi.org/10.1089/cmb.2012.0021). Journal of Computational Biology 19:455–477.

5. Seemann T. 2014. [Prokka: rapid prokaryotic genome annotation](https://doi.org/10.1093/bioinformatics/btu153). Bioinformatics 30:2068–2069.

6. Lee I, Chalita M, Ha S-M, Na S-I, Yoon S-H, Chun J. 2017. [ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences](https://doi.org/10.1099/ijsem.0.001872). International Journal of Systematic and Evolutionary Microbiology 67:2053–2057.

7. Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2019. Biostrings: Efficient manipulation of biological strings. R package version 2:10–18129.

8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. [Basic local alignment search tool](https://doi.org/10.1016/S0022-2836(05)80360-2). Journal of Molecular Biology 215:403–410.

# Tables:

**Table** **1**: Sample origin, BLASTn results (Top 16S Hit and Accession Number, % Identity, % Query Coverage, and 16S Fragment Length), total assembled genome size with all contigs, total assembled size only only including contigs larger than 500bp, N50, L50, and % GC for each sample. N50, L50, and % GC are reported for the assembly without contigs smaller than 500bp.

|  | 231232\_2ArWaCKT | 231237\_7MaSpsaltlick | 231240\_10ClHs |
| --- | --- | --- | --- |
| Sample Origin | Saltwater Fish Tank | Cow Salt Lick | Arboretum Soil |
| Top BLASTn Hit | Halomonas sp. THAF12 | Staphylococcus sp. | Bacillus sp. |
| Accession Number | CP045399.1 | MT539740.1 | MN227356.1 |
| BLASTn % Identity | 100.0 | 100.0 | 99.8 |
| BLASTn % Query Coverage | 100.0 | 100.0 | 99.0 |
| 16S Fragment Length | 984 | 727 | 1477 |
| Total Size | 8827618 | 6414125 | 8795180 |
| Size >= 500 | 8197885 | 6346283 | 8419700 |
| Number of Contigs | 1443 | 326 | 1069 |
| N50 | 18693 | 302755 | 46163 |
| L50 | 46 | 7 | 39 |
| % GC | 61.81 | 32.69 | 45.56 |