# Single cell genome reconstruction of two uncultured, proteorhodopsin-containing Flavobacteria


Tanja Woyke[1], Alex Copeland[1], Gary Xie[3], Cliff Han[3], Jan-Fang Cheng[1], Hajnalka Kiss[3], Jimmy Saw[3], Pavel Senin[3], Michael E. Sieracki[2] & Ramunas Stepanauskas[2]

*[1] DOE Joint Genome Institute, Walnut Creek, California, [2] Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine, [3] Los Alamos National Laboratory, Los Alamos, New Mexico*



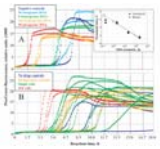
## Abstract

**Determining the genetic makeup of predominant microbial taxa with specific metabolic capabilities remains one the major challenges in microbial ecology and bioprospecting, due to the limitations of current cell culturing and metagenomic methods. The complexity of microbial communities and intraspecies variations hinders the assembly of individual genomes from metagenomic shotgun libraries. Here we report the use of single cell genomics to access the genome of two proteorhodopsin-encoding flavobacteria from Gulf of Maine bacterioplankton. We use high throughput fluorescence-activated sorting of single cells, whole genome amplification via multiple displacement amplification, PCR-screening and subsequent shotgun sequencing of these single amplified genomes (SAGs), allowing the genomic analysis of their novel photometabolic system and the sequence comparison to environmental marine sequence data.**


## Background & Methods

Costal water (Maine)

Cell sorting



Real-time multiple displacement amplification (MDA) of single cells

16S rRNA gene and proteorhodopsin gene PCR of single cells

Stepanauskas, PNAS 2007.

Sequencing

| | MS024-2A | MS024-3C |
|---|---|---|
| 454 data | 73 Mb | 90 Mb |
| 3Kb Sanger | 4.58 Mb | 5 Mb |
| 8Kb Sanger | 12.87 Mb | 17.67 Mb |

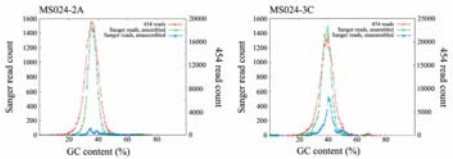
## Sequence assemblies

General features of the flavobacterial SAG assemblies.

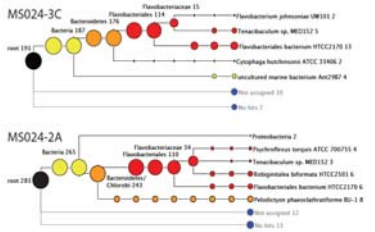
| | MS024-2A | MS024-3C |
|---|---|---|
| **Assembly statistics** | | |
| **Assembly size [bp]** | **1,905,484** | **1,515,248** |
| Estimated genome size [bp] | 2,156,286 - 3,004,105 | 2,307,484 - 3,726,020 |
| **Number of contigs** | **17** | **21** |
| Largest contig [bp] | 684,032 | 549,383 |
| **GC content [%]** | **36** | **39** |
| Mean read depth (± sd) | 56 (± 63) | 83 (± 110) |
| **454 reads** | **47** | **68** |
| **Sanger reads** | **9** | **14.3** |
| **Gene predictions** | | |
| **Total genes** | **1,824** | **1,426** |
| Protein coding genes | 1,785 | 1,400 |
| **with function prediction** | **1,205** | **960** |
| **w/o function prediction** | **580** | **440** |
| Number of rRNA operons | 2 | 1 |
| **Number of tRNA genes** | **33** | **24** |

The sequence data of the SAGs was Phrap assembled, followed by primer walking on shotgun clones, and PCR/adapter PCR on the diluted MDA products.

## Data QC

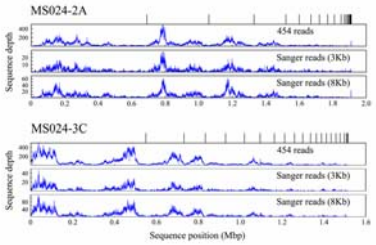



GC contents histogram of the unassembled and assembled Sanger and pyrosequence reads for the two SAG exhibits a tight uni-modal distribution.

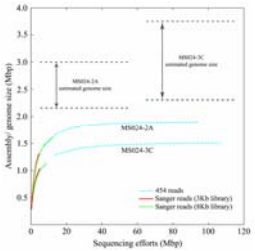



Taxonomic assignment analysis. The taxonomic contents of the blastx output for the unassembled reads of the Flavobacteria sp. MS024-2A was estimated and visualized using the Metagenome Analyzer (MEGAN) (Huson, Genome Res 2007).

## The MDA bias




MDA bias as evaluated by sequence depth distribution. The contigs for the SAG are aligned by length and contig breaks are indicated by the tic marks along the top. The mean sequence depth is 56 (± 63) for MS024-2A and 83 (±110) for MS024-3C.



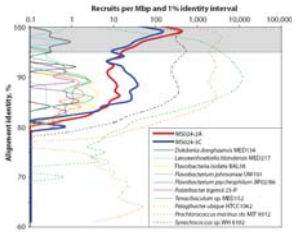

Genome coverage as function of the genome sequencing effort for the flavobacterial SAG. The curve displays near-saturation indicating that additional sequencing would mostly result in repeated sampling of the over-amplified genomic regions, not targeting the yet missing part of the genome.

## Chimeras

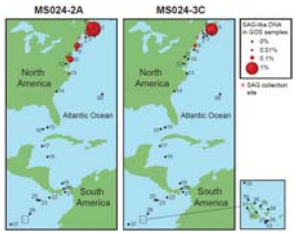
Table S1. Chimeric rearrangments in the SAG DNA.

| | MS024-2A | | MS024-3C | |
|---|---|---|---|---|
| | chimeric reads/ clones (%) | overall chimerism | chimeric reads/ clones (%) | overall chimerism |
| **Read-based chimerism** | | | | |
| 3Kb library reads (untreated MDA DNA) | 1.9 | 1 chimer/ 28 Kbp | NA | NA |
| 3Kb library reads (S1 treated MDA DNA) | 2.0 | 1 chimer/ 25 Kbp | 1.9 | 1 chimer/ 33 Kbp |
| 8Kb library reads (S1 treated MDA DNA) | 2.1 | 1 chimer/ 30 Kbp | 1.6 | 1 chimer/ 40 Kbp |
| 454 reads (S1 treated MDA DNA) | | 1 chimer/ 19 Kbp | | 1 chimer/ 25Kbp |
| **Average (all reads)** | | **1 chimer/ 21 Kbp** | | **1 chimer/ 27Kbp** |
| **Clone-based chimerism** | | | | |
| 3Kb clones (untreated MDA DNA) | **14.5** | 1 chimer/ 20 Kbp | **NA** | NA |
| paired reads facing into the same direction | 8.6 | | | |
| paired reads facing away from insert | 3.8 | | | |
| paired reads outside the insert size range | 0.2 | | | |
| paired reads in different contigs | 0.8 | | | |
| paired reads contained in each other | 1.1 | | | |
| **3Kb clones (S1 treated MDA DNA)** | **16.8** | **1 chimer/ 15 Kbp** | **16.4** | **1 chimer/ 20 Kbp** |
| **paired reads facing into the same direction** | **7.8** | | **9.6** | |
| **paired reads facing away from insert** | **3.0** | | **0.9** | |
| **paired reads outside the insert size range** | **3.2** | | **4.5** | |
| **paired reads in different contigs** | **2.0** | | **0.8** | |
| **paired reads contained in each other** | **0.8** | | **0.6** | |
| 8Kb clones (S1 treated MDA DNA) | **36.4** | 1 chimer/ 17Kbp | **29.5** | 1 chimer/ 27Kbp |
| paired reads facing into the same direction | 22.6 | | 17.4 | |
| paired reads facing away from insert | 2.5 | | 2.7 | |
| paired reads outside the insert size range | 7.4 | | 3.9 | |
| paired reads in different contigs | 3.3 | | 3.8 | |
| paired reads contained in each other | 0.6 | | 1.8 | |
| **Average (all clones)** | | **1 chimer/ 17Kbp** | | **1 chimer/ 23Kbp** |

## Fragment recruitment




Global Ocean Sampling (GOS) (Rusch, PLoS Biol 2007) metagenome fragment recruitment by the SAGs MS024-2A and MS024-3C, the currently sequenced marine Flavobacteria isolate genomes, the non-marine *F. johnsoniae*, and the three best GOS fragment recruiters *Pelagibacter*, *Prochlorococcus* and *Synechocuccus*.




Geographic distribution of the GOS (Rusch, PLoS Biol 2007) metagenome fragments with >95% nucleotide identity to MS024-2A and MS024-3C.

## Conclusion

**Using the single cell approach, we demonstrate how a combination of single cell FACS and amplification via MDA can be used to access the genomes of uncultured environmental microorganisms, representative of their given environment.**


## Acknowledgements

**We would like to thank PGF for the sequencing efforts and Lynne Goodwin (Los Alamos National Laboratory) for the coordination of the efforts involved in this project. We also thank H. Tu and M. Zhang for their help with the chimera detection analysis.**

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.