Title

Assembly and Sequence Analysis of the Single Cell Genome of an Uncultured, Marine Flavobacterium

Authors

Copeland, A.
Woyke, T.
Xie, G.
et al.

Publication Date

2009-02-02

Assembly and sequence analysis of the single cell genome of an uncultured, marine Flavobacterium

Alex Copeland[1], Tanja Woyke[1], Gary Xie[3], Cliff Han[3], Jan-Fang Cheng[1], Hajnalka Kiss[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

The complexity of microbial communities and intraspecies variation hinders the assembly of individual genomes from metagenomic shotgun libraries. Here we report on the assembly and analysis of the genome of an uncultured Flavobacterium from Gulf of Maine bacterioplankton.

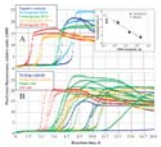
A single cell was isolated from a coastal water sample using fluorescence-activated cell sorting (FACS). Whole genome amplification via multiple strand displacement amplification (MDA) was used to prepare sufficient DNA for constructing Sanger and 454 (GS20) libraries of the single-cell amplified genomes (SAG). We will discuss assembly and analysis of the sequence data with an emphasis on challenges unique to this type of data, limitations of the existing methods and implications for future investigations of rare and unculturable microorganisms.

Background & Methods

Costal water (Maine)

Cell sorting

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



Flavobacteria sp. MS024-2A and MS024-2C

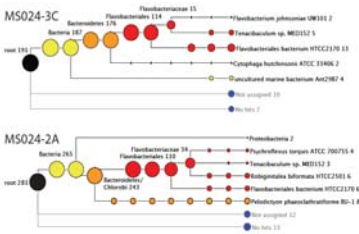
16S rRNA gene and proteorhodopsin gene PCR of single cells

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**

DNA validation



The taxonomic contents of the blastx output for the unassembled reads of the *Flavobacteria* sp. MS024-2A and MS024-3C estimated and visualized using the Metagenome Analyzer (MEGAN) (Huson, Genome Res 2007).

Assembly & Closure

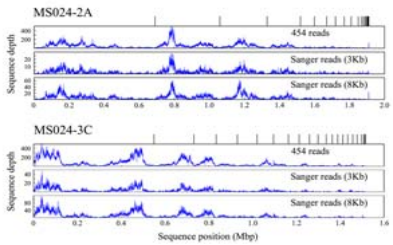
General features of the flavobacterial SAG assemblies.

	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**

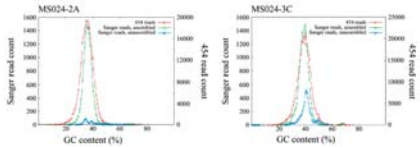
	MS024-2A chimeric reads/ clones (%)	MS024-2A overall chimerism	MS024-3C chimeric reads/ clones (%)	MS024-3C overall chimerism
Read-based chimerism				
3Kb library reads (untreated MDA DNA)	1.9	1 chimer/ 28 Kbp	NA	NA
3Kb library reads (S1 treated MDA DNA)	2.0	1 chimer/ 25 Kbp	1.9	1 chimer/ 33 Kbp
8Kb 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**

After performing some closure work, including primer walking on shotgun clones, and PCR/adapter PCR on the diluted MDA products, rates of chimerism in the Sanger libraries were assessed.

MDA bias

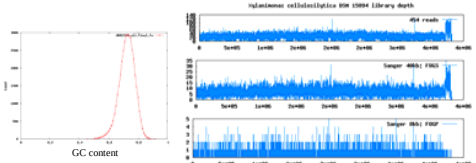


MDA bias evaluated by sequence depth distribution. The contigs for the SAG were aligned by length and contig breaks are indicated by the tic marks along the top. The mean sequence depth is 56 (± 63) for MS024-2A and 83 (±110) for MS024-3C.

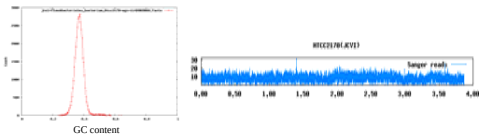


GC content histogram of the unassembled and assembled Sanger and 454 sequence reads for the two genomes. As hoped the samples were free of obvious contamination.

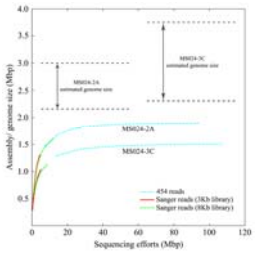
To highlight the degree of coverage bias in the MDA amplified samples, depth distribution plots were produced for two unrelated, unamplified projects which were not expected to be biased.



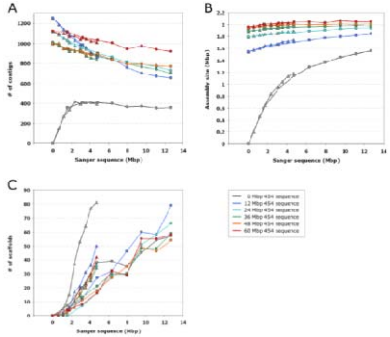
Xylanimonas cellulosilytica was sequenced with 454+Sanger by JGI. As expected the coverage was reasonably uniform, but this genome has distinctly different GC content than our *Flavobacteria*.



Flavobacteria HTCC2170 is a Sanger-only project sequenced by JCVI. Again, coverage is quite uniform.



Genome coverage as function of the genome sequencing effort. The curve displays saturation at high coverage suggesting that additional sequencing would result in repeated sampling of the over-amplified genomic regions, not targeting the yet missing part of the genome.



While 454 sequencing offers higher throughput at lower cost, paired-end Sanger sequence data was produced for high-quality sequence coverage of homopolymeric regions and improved assembly of low coverage areas and repeats. The beneficial effect on scaffolding of adding the paired-end Sanger sequence data to the 454 data are evident.

Conclusion

We demonstrate how a combination of single cell FACS and amplification via MDA can be used to isolate and sequence the genomes of uncultured microorganisms. While bias and chimerism complicate the analysis, method optimization should improve coverage and reduce bias. Notwithstanding the bias, we believe we were able to recover the majority of the gene content of these genomes.

Acknowledgements

We thank Lynne Goodwin (Los Alamos National Laboratory) for project management and coordination 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.