

Whole genome sequencing, assembly and annotation of *Halomonas smyrnensis*, a levan producing halophilic bacterium

E. Sogutcu¹✉, Z. Emrence², M. Arikan², A. Cakiris², N. Abaci², E. Toksoy Oner¹, D. Ustek², K.Y. Arga¹

¹Department of Bioengineering, Marmara University, Goztepe, Istanbul, Turkey

²Department of Genetics, Institute for Experimental Medicine, Istanbul University, Istanbul, Turkey

Motivations

Next-generation sequencing technologies have been available in the relatively short time frame providing a wealth of data and specific information that cannot be obtained by other experimental approaches. In systems biology research, the microbial genome sequence is the starting point for detailed analysis of identifying gene-protein associations and metabolic reconstruction. For this purpose, we performed whole genome sequence analysis of *Halomonas smyrnensis*, which has been reported as a high level levan-producer halophilic bacterium for the first time by our research group. Levan is a linear fructose polymer and has many potential uses in foods, feeds, cosmetics, and the pharmaceutical and chemical industries.

Methods

A hybrid strategy is performed both in sequencing and assembly process. The genome sequence of *Halomonas smyrnensis* has obtained using two independent experimentation: Roche

454 GS FLX+ System and Ion Torrent Sequencer. Reads have de novo assembled into scaffolds using the 454 Newbler, CLC Genomics Workbench 5.0.1 and Geneious Pro 5.5.6 assembler software. The draft genome has structured with assembled contigs that have ordered by CONTIGuator bacterial genome finishing tool. Thereafter we have performed draft genome annotation via RAST annotation server.

Results

The Roche 454 GS FLX+ System (1,442,441 reads with an average length of 495.14 bp) and Ion Torrent Sequencer System (359,558 reads with an average length of 122.42 bp) resulted with 260x coverage. As a result of the de novo assembly process, 105 high-quality contigs with size greater than 8000 bp were obtained. Among those, 42 contigs were ordered and structured with a genome size of 4,242,280 bp (G+C content 67.8%). The annotation of draft genome ended up with 3845 coding sequences and 96 RNAs.