Bigelow | Single Cell Genomics Center

Updated July 20, 2017

This folder contain results from an experiment that assessed the performance of SCGC's single amplified genome (SAG) generation, sequencing and assembly pipeline. Five SAGs of each bacterial benchmark strain and three SAGs of each algal benchmark strain were selected from a 384-well plate based on their lowest critical point (Cp) value as the selection criterion. The SAGs were sequenced with NextSeq 500 (Illumina) in 2x150 bp mode, quality-trimmed, pre-normalized and assembled, followed by post-assembly trimming, as previously described (Stepanauskas et al. 2017)Due to an elevated frequency of assembly artifacts in short contigs, only contigs larger than 2,000 bp were included in the final assemblies.

The quality of raw reads was assessed using fastqc (http://www.bioinformatics.babraham.ac.uk /projects/fastqc). Assembly quality was assessed with QUAST (Gurevich et al. 2013), assuming that publicly available genome sequences are correct and that these strains have not evolved since their original sequencing. Additional assembly quality control includes a tetramer principal component analysis (PCA). Tetramer frequencies are calculated for 1,600 bp windows using 200 bp steps. Frequencies of reverse-complementary tetramers are combined and represented as a N×136 feature matrix. Principal component analysis (PCA) is then used to extract the most important components of this high-dimensional feature matrix. Plots are generated for the first eight principal components. Two contigs containing the most positive and two contigs containing the most negative values along each principal component are assigned individual colors, and their windows are connected with lines. BLASTN (Altschul et al. 1990) against NCBI's nt database is performed on two windows containing extreme values along each principal component. A significant and consistent tetramer frequency divergence by a contig from the rest of the genome may indicate that that contig contains viral, plasmid, or other horizontally transferred DNA or is a contaminant (Woyke et al. 2009, Swan et al. 2013). To resolve the nature of such contigs, they should be examined for gene content and sequence similarity to other genomes. Please note that contaminant contigs have been extremely rare in SAGs sequenced by SCGC after the establishment of SCGC's all-in-house sequencing and assembly pipeline in early 2014.

References:

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal Of Molecular Biology 215:403-410

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29:1072-1075

Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, Becraft ED, Brown JM, Pachiadaki MG, Povilaitis T, Thompson BP, Mascena CJ, Bellows WK, Lubys A (2017) Improved genome recovery and integrated cell-size analyses of individual, uncultured microbial cells and viral particles. Nat Commun 8:84

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonźalez JM, Luo H, Wright JJ, Landry ZC, Hanson NW, Thompson BP, Poulton NJ, Schwientek P, Acinas SG, Giovannoni SJ, Moran MA, Hallam SJ, Cavicchioli R, Woyke T, Stepanauskas R (2013) Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proceedings of the National Academy of Sciences of the United States of America 110:11463-11468

Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, Kiss H, Saw J, Senin P, Yang C, Chatterji S, Cheng J-F, Eisen JA, Sieracki ME, Stepanauskas R (2009) Assembling the marine metagenome, one cell at a time. Plos ONE 4:e5299