Bioinformatic Pipeline Characterization Results

Nate Olson

2017-07-06

Contents

The sequencing dataset was processed using three bioinformatic pipelines. The resulting count tables were characterized for number of features, sparsity, and filter rate (Table 1). The expectation is that this mixture dataset will be more sparse relative to other datasets due to the redundant nature of the samples where 35 of the 45 samples are derived directly from the other 10 samples and that there are four PCR replicates for each sample. Sparsity was lower for De-novo clustering (QIIME) than sequence inference (DADA2) even though DADA2 has fewer total features. Different pipelines have different approaches for handling low quality reads. See individual pipeline reports for which steps reads are excluded from the datasets. QIIME pipeline has the highest filter rate while the highest number of features per sample.

Table 1: Summary statistics for the different bioinformatic pipeliens. DADA2 is a denoising sequence inference pipeline, QIIME is a open-reference clustering pipeline, and mothur is a de-novo clustering pipeline. No template controls were excluded from summary statistics. Sparsity is the proportion of 0's in the count table. Features is the total number of OTUs (QIIME and mothur) or SVs (DADA2) in the count. Sample coverage is the median and range (minimum - maximum) per sample total feature abundance. Filter rate is the proportion of reads that were removed while processing the sequencing data for each bioinformatic pipeline.

Pipelines	Features	Sparsity	Sample Coverage	Filter Rate
dada2	3144	0.93	68649 (1661-112058)	0.24 (0.18-0.59)
mothur	38358	0.98	53775 (1265-87806)	0.4 (0.35-0.62)
qiime	11385	0.94	25254 (517-46897)	0.7 (0.62-0.97)