Sequencing Data Quality Assessment

Nate Olson

2017-11-20

Quality assessment of sequencing run summarizing number of reads per sample. Two barcoded experimental samples have less than 35,000 reads (Fig. 1A). The rest of the samples with less than 35,000 reads are no template PCR controls (NTC). Excluding the one failed reaction with 2,700 reads and the NTCs, the total range in the observed number of sequences per samples is 3195 to 152267 reads with a median library size of 8.9548×10^4. For the expected overlap region, based on primer positions and read lengths (16S PCR fig), the forward read has consistently higher base quality scores relative to the reverse read with a narrow overlap region with high base quality scores for both forward and reverse reads (Fig. 1B).

The sequencing dataset was processed using four bioinformatic pipelines. The resulting count tables were characterized for number of features, sparsity, and filter rate (Table 1, Fig. 1C). The pipelines evaluated have different approaches for handling low quality reads resulting in the large variability in filter rate (Table 1). QIIME pipeline has the highest filter rate and highest number of features per sample. The targeted amplicon region has a relatively small overlap region, 136 bp for 300 bp paired end reads. The high filtration rate is due to the drop in base calling accuracy at the ends of the reads especially the reverse reads resulting in a high frequency of unsuccessfully merged reads pairs (Fig. 1B). Additionally, to remove potential sequencing artifacts from the dataset QIIME excludes singletons, OTUs only observed once in the dataset. The expectation is that this mixture dataset will be less sparse relative to other datasets due to the redundant nature of the samples where 35 of the samples are derived directly from the other 10 samples and 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. With sparsity greater than 0.9 for the three pipelines it is unlikely that any of the pipelines successfully filtered out a majority of the sequencing artifacts.

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	Total Abundance	Drop-out Rate
dada2	3144	0.93	68649 (1661-112058)	0.24 (0.18-0.59)
mothur	38469	0.98	53775 (1265-87806)	0.4 (0.35-0.62)
qiime	11385	0.94	25254 (517-46897)	0.7 (0.62-0.97)

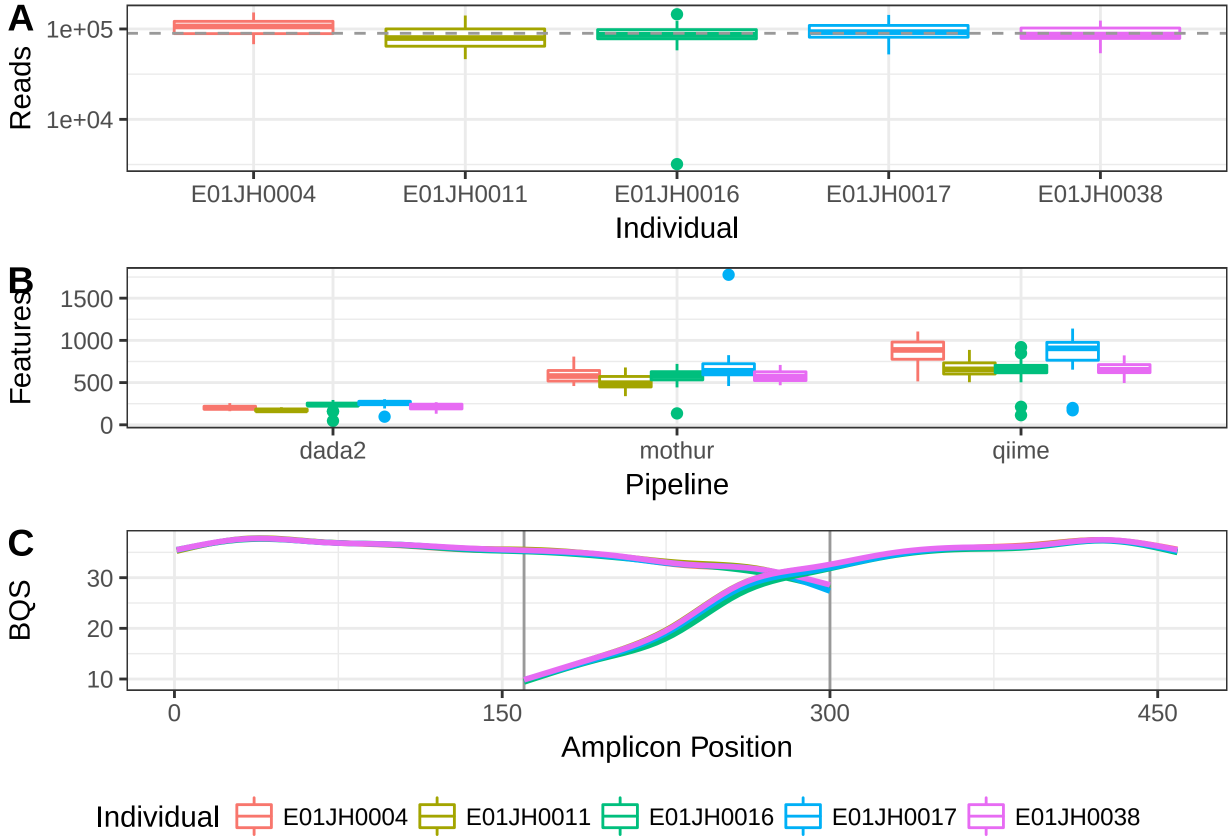

Figure 1: Sequencing dataset summary. (A) Distribution in the number of reads per barcoded sample (Library Size) by individual. Dashed horizontal line indicates overall median library size. (B) Smoothing spline of the base quality score (BQS) by sequencing cycle. Vertical lines indicate approximate overlap region between forward and reverse reads. (C) Distribution of the number of features per sample.

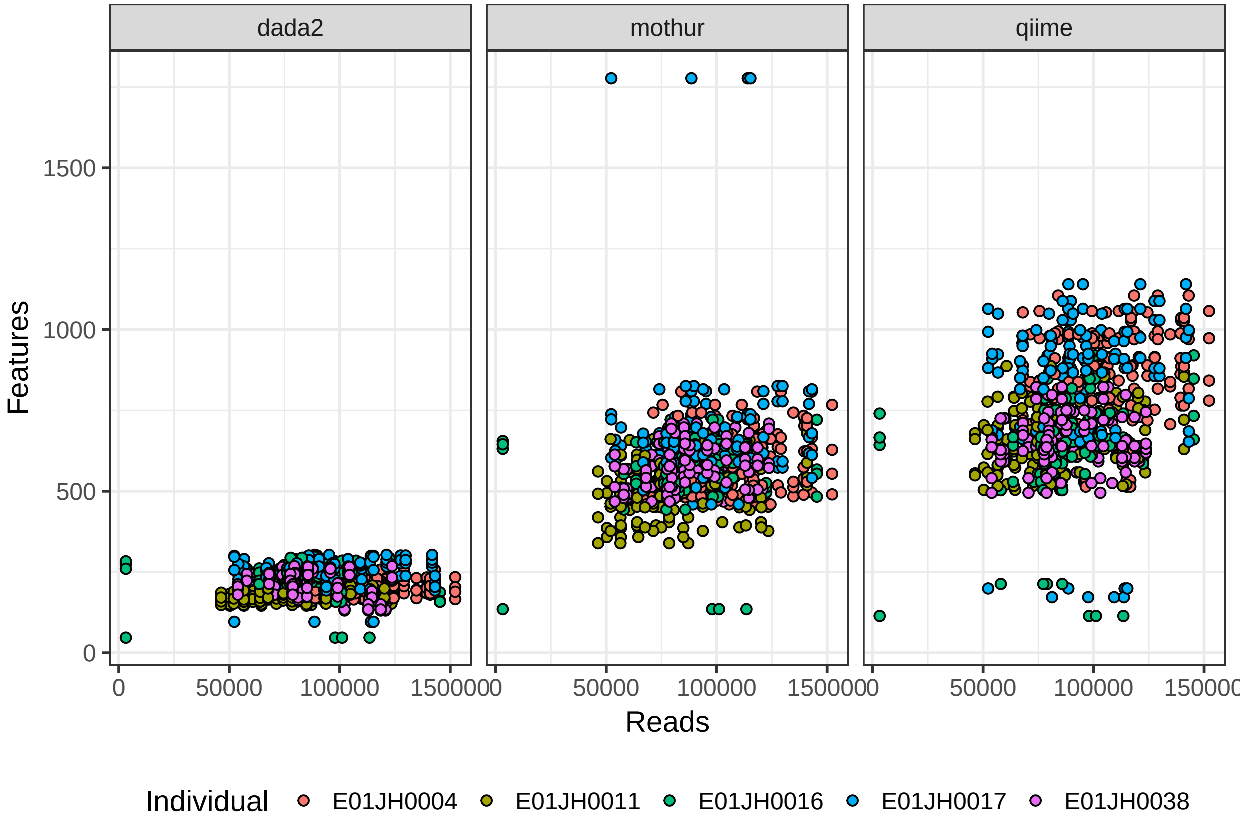

Figure 2: Relationship between the number of reads and features per sample by bioinformatic pipeline.