Table 1. Description of datasets used to evaluate the OptiClust algorithm and compare its performance to other algorithms. Each dataset contains sequences from the V4 region of the 16S rRNA gene. The even and staggered datasets were generated by extracting the V4 region from full length reference sequences and the datasets from the natural communities were generated by sequencing the V4 region using a Illumina MiSeq with either paired 150 or 250 nt reads.

Dataset (Ref.)	**Read Length**	**Samples (N)**	**Total Seqs. (N)**	**Unique Seqs. (N)**
Soil (XX)	150	18	948,243	143,677
Marine (XX)	250	7	1,384,988	75,923
Mice (XX)	250	360	2,825,495	32,447
Human (XX)	250	489	20,951,841	121,281
Even (XX,YY)	NA	NA	1,155,800	11,558
Staggered (XX,YY)	NA	NA	1,156,550	11,558