

## 1. Input Reads

### a. General Stats

General stats summary table for the raw reads, displaying by default the percentage of duplicate reads, average read length and total number of sequences in K. Average read length should be ~250bp, and a high (>90%) duplicate percentage is normal for HMAS data. Total sequence counts are generally acceptable between 100K and 1000K.

### b. Sequence Counts

Unique read and estimated duplicate read counts for each sample. A high percentage of duplicate reads (>90%) is expected for HMAS data. Total sequence counts are generally acceptable between 100K and 1000K.

### c. Seq quality histogram

Showing the mean quality value of each bp for each reads. Samples in general will have a higher quality score than negative controls. R1 typically has higher quality than R2. Samples and positive controls generally have a higher quality score than negative controls.

### d. Per seq quality scores

Showing the number of reads with mean quality scores. Most reads should have a phred score >30. Two peaks, one each for R1 and R2, are usually observed.

### e. Per base N content

Showing the percentage of base calls at each position a 'N' is called. A normal HMAS run does not have any observable peaks.

## 2. Reads Processing

### a. Primer removal: step trims primer sequence from each read.

#### i. Cutadapt filtered reads

Number of reads with primers successfully trimmed (passing filters) and those failed. ~90% of reads typically pass this filter.

#### ii. Trimmed seq lengths (5')

Abundance of sequences (count) in each sample with a given length trimmed from the 5' end of each read. For the Salmonella HMAS assay, the range of trimmed primer lengths is 20-22bp.

#### iii. Trimmed seq lengths (3')

Abundance of sequences in each sample with a given length trimmed from the 3' end of each read. This includes the reverse complement of the 3' primer and any read through sequence beyond that primer, which varies with the length of the target amplicon. A normal Salmonella HMAS assay run will have a range of trimmed sequences between 21-100 bp.

### b. Pair merging: step assembles R1 and R2 for each PE read and performs initial quality checking.

#### i. PEAR assembled vs unassembled reads

Number of successfully assembled (passing) and un-assembled (failed) reads. Typical samples have <1% of reads failing.

### 3. Amplicon Filter

- a. Quality filtering: assess the quality of the assembly (from the last step) and remove reads that fail (with one or more expected errors, based on combined phred scores of the entire read).
  - i. Qfilter

Number of total reads and removed reads. Typical samples have <0.1% of reads failing.

- b. Dereplication: step extracts unique sequences for all reads in each sample.
  - i. Dereplication total seq vs unique

**Every read count after this step is in terms of unique sequences.** Showing the number of total reads and the unique reads. A ratio of 5 or larger typically indicates a good mean read depth for this sample.

- c. Denoising: step removes unique reads with likely sequencing errors inferred from read frequency and sequence composition.
  - i. Total vs removed reads

Number of total and removed unique sequences. It is normal for most unique sequences to be removed as most of the remaining reads have high read frequency hence only a small percentage of total (non-unique) reads were removed.

### 4. Reporting: summary statistics for high quality sequences.

- a. Primer performance report (table)

Average, minimum, and maximum read count per primer pair across all samples in the run, with a low average read count (<30x red; 30x~50x yellow; >50x green) indicating bad primer performance in this run.

- b. Sample read length report (table)

For the final high-quality sequences, showing the total read count, average read length, minimum and maximum read length for each sample.

- c. HMAS run report (table)

Combined summary statistics for all the samples in the run, showing the mean read depth and the number (and percentage) of successful primer pairs (out of total 2461 in the Salmonella HMAS primer panel).

For negative control samples, mean read depth is usually n/a or 2; anything  $\leq 10x$  is acceptable. Percent successful primer pairs should be extremely low, n/a or 0-5% commonly seen, and number of successful primer pairs should be <10.

For positive and passing samples, mean read depth is at least 30x (<30x red; 30x~50x yellow; >50x green) and percent successful primer pairs is at least 90%. (<80% red; 80%~90% yellow; 90%~98% green; >98% blue)