UMI Report

**Overview**

We conducted three sequencing runs in order to establish which methods gave the best performance when processed through our pipeline built to collapse UMI groups.

Here I processed the data through the UMI pipeline using both the full datasets and a set downsampled to 25 million read pairs. The results and some observations follow.

**UMI Full Data**

Graphed below are the total reads in the initial BAM file versus the mean target coverage and percentage of bait on a per run basis. Each graph displays trends for the initial BAM and the final BAM (post collapse). These graphs are intended to show the in-run performance pre and post collapse relative to sample total read count.

Run 263R (Standard Prep on NextSeq)

266R (1/5-1/50 ligations on HiSeq)

280R (1/5 ligations and 20ul into PCR)

**Initial Observations**

Runs 266R and 280R appear to perform similarly in terms of the relationships between number of reads per sample and the resulting accuracy and coverage. For those two runs as read count increases in the final bam the percent off bait decreases and the mean target coverage increases.

However, for run 263R with standard prep as the reads per sample increases the mean target coverage does increase but the PCT off base also increases. **Most notable though is that the percent off bait is lower post collapse for run 263R, where it is higher for runs 266R and 280R.**

Now that we’ve seen how each run performs pre and post collapse it is time to compare the runs against each other. But first, some basic relationships for these runs pre and post collapse.

Observation – As expected mean target coverage scales with total reads in both the initial BAM and final BAM. All runs perform similarly pre-collapse. Post-collapse run 263R scales best while 280R scales worst.

Observation – Here we compare total reads in the initial BAM to the final BAM. As expected they are highly correlated with 263R scaling at a higher rate than 266R and 280R. 266R and 280R perform very similarly in regards to total read counts. Let’s explore more.

Observation – similar to the previous graph comparing initial and final total reads, final mean target coverage can be expressed as a function of initial total reads.

Observation – above we see that the PCT off bait of the final BAM performs differently for each run. Overall run 263R consistently has the least percent off bait though it increases slightly as the number of initial total reads increases. Runs 266R and 280R both increase their efficiency as the number of reads in the initial BAM increases suggesting at some point they would meet and possibly exceed performance of 263R in this metric.

**UMI Family Analysis**

Next we look at the UMI histogram data.

Observations – 263R clearly has a larger percentage of UMI family sizes of 1 than 263R and 280R. However, just viewing the data in this context can be misleading. Next we analyze the UMI family size data in terms of UMI family size totals relative to total reads in the initial BAMs.

Observation – Here we see at family size 1 run 263R has a lot more reads than 266R and 280R. Following is the same analysis at subsequent family sizes.

Observations – Relative to initial total reads, at UMI family size 1-4 263R appears to outperform 266R and 280R. Next we skip ahead to UMI family sizes of 14 and 15.

Observation – Even when we get to UMI family sizes of 14 and 15 run 163R is slightly outperforming 266R and 280R though they appear to be getting much closer than family sizes 2-4.

**Downsampling to 25M read pairs experiment data**

Next we take a look at the target coverage metrics when normalizing the samples in each run to 25 million read pairs.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample ID | TOTAL\_READS | INITIAL\_MEAN\_TARGET\_COVERAGE | FINAL\_MEAN\_TARGET\_COVERAGE |
| 263R02-B01-MONCv1 | 50000000 | 10750.33759 | 2569.253806 |
| 263R03-C01-MONCv1 | 50000000 | 10517.86905 | 1778.976339 |
| 263R04-D01-MONCv1 | 50000000 | 10633.04974 | 2097.805717 |
| 263R05-E01-MONCv1 | 50000000 | 10238.41065 | 2072.593673 |
| 263R06-F01-MONCv1 | 50000000 | 10536.08569 | 1872.869898 |
| 266R01-A01-MONCv1 | 50000000 | 11736.86827 | 414.7239 |
| 266R02-B01-MONCv1 | 50000000 | 11852.92601 | 223.270431 |
| 266R03-C01-MONCv1 | 50000000 | 11908.09849 | 172.490129 |
| 266R04-D01-MONCv1 | 50000000 | 11972.40371 | 88.931574 |
| 266R05-E01-MONCv1 | 50000000 | 11978.10736 | 55.291139 |
| 280R02-B01-MONCv1 | 50000000 | 11370.63229 | 11.678011 |
| 280R03-C01-MONCv1 | 50000000 | 11588.08339 | 18.907173 |
| 280R04-D01-MONCv1 | 50000000 | 11468.539 | 55.383332 |
| 280R05-E01-MONCv1 | 50000000 | 11364.17187 | 201.459816 |
| 280R06-F01-MONCv1 | 50000000 | 11737.80006 | 70.678348 |

Here we see that after normalizing for read count going into the pipeline run 263R has a higher mean target coverage than 266R and 280R.

**Summary**

Our previous analytical efforts had been focused on determining which wet lab methods on 266R and 280R had been more efficient. Here we looked at how these three runs performed on a run level.

Run 263R appears to outperform in most metrics after running through the UMI pipeline, including mean target coverage, percent off bait, and total UMI reads with family size two or greater.

Some of our previous analyses had relied on histograms and fractional measurements that ignored total read counts. When normalizing against total read count run 263R appears to outperform 266R and 280R.