WGS Competency Assessment Graphical Report

Example Report

Analysis by "Analyst name"

August 26, 2019

Contents

1 Overview

This report provides an analytical evaluation of sequencing metrics for bacterial whole genome sequencing data comprising a sequencing competency assessment exercise. Sequence metrics have been compared to expected values based on recommended data submission thresholds or derived from the distribution of isolates sequenced during the 2018 PulseNet-GenomeTrakr harmonized Proficiency Testing (PT) exercise as part of a 16-sample 2X250 run on an Illumina MiSeq machine (V2 chemistry) using the Nextera XT DNA Library Prep Kit.

	Isolate ID	Organism
1	SAP18-0432	Salmonella enterica subsp. enterica serovar Enteritidis
2	SAP18-H9654	Salmonella enterica subsp. enterica serovar Enteritidis
3	SAP18-6199	Salmonella enterica subsp. enterica serovar Typhimurium
4	SAP18-8729	Salmonella enterica subsp. enterica serovar Newport
5	LMP18-H2446	Listeria monocytogenes
6	LMP18-H8393	Listeria monocytogenes

Table 1: The 2018 PT isolates.

	Sample ID	Isolate ID	Machine	Flowcell
1	SEQ000070739	SAP18-H9654	M04590	BKRCT
2	SEQ000070740	SAP18-8729	M04590	BKRCT
3	SEQ000070741	SAP18-6199	M04590	BKRCT
4	SEQ000070742	SAP18-0432	M04590	BKRCT
5	SEQ000070743	LMP18-H8393	M04590	BKRCT
6	SEQ000070744	LMP18-H2446	M04590	BKRCT

Table 2: Isolates submitted for analysis.

2 Results

Of the 6 isolates submitted as part of this exercise, 6 (100%) were determined to be correctly annotated and to have fully met the analytical expectations of this exercise.

The following section contains quantitative and graphical representations of your PT sequencing isolates (colored data points) and the distribution of runs from all participating GenomeTrakr labs (boxplots). The isolates are labeled (x -axis) using the abbreviated Isolate ID. For each boxplot, the box defines the median value as well as the lower and upper quartiles (25% and 75%). The whiskers extend to 1.5 times the interquartile range from the median. Pay particular attention to any figures in which values for your run fall outside of the 25% - 75% interquartile range (the boundaries of the box in the plot).

2.1 Competency Assessment Metrics

The following metrics are under evaluation in this competency assessment (**Figs. 1-6**) -
- • **Taxonomy Check** - was the correct organism sequenced?
1. **Mean depth** - the mean depth of coverage. QC threshold = 20X.
2. **Mean insert size** - the mean length of the sequence between the adapters.
3. **Mean read quality scores - forward reads (R1)** - QC threshold = Q30.
4. **Mean read quality scores - reverse reads (R2)** - QC threshold = Q30.
5. **NG50** - A metric describing the distribution of contig lengths.
6. **Contigs** - The total number of contigs in the assembly.

| SampleID	Isolate ID	Metric	Threshold	Value
SEQ000070739	SAP18-H9654	Taxonomy Check	PASS	**PASS**
SEQ000070739	SAP18-H9654	MeanR1Qual	≥ 30	**35.2**
SEQ000070739	SAP18-H9654	MeanR2Qual	≥ 30	**33**
SEQ000070739	SAP18-H9654	MeanDepth	≥ 20	**141.98**
SEQ000070739	SAP18-H9654	MeanInsert	≥ 203.3	**467.16**
SEQ000070739	SAP18-H9654	NG50	≥ 259138.5	**401093**
SEQ000070739	SAP18-H9654	Contigs	≤ 97.25	**68**
SEQ000070740	SAP18-8729	Taxonomy Check	PASS	**PASS**
SEQ000070740	SAP18-8729	MeanR1Qual	≥ 30	**35.1**
SEQ000070740	SAP18-8729	MeanR2Qual	≥ 30	**32.7**
SEQ000070740	SAP18-8729	MeanDepth	≥ 20	**100.07**
SEQ000070740	SAP18-8729	MeanInsert	≥ 203.1	**480.47**
SEQ000070740	SAP18-8729	NG50	≥ 0	**498666**
SEQ000070740	SAP18-8729	Contigs	≤ 132	**103**
SEQ000070741	SAP18-6199	Taxonomy Check	PASS	**PASS**
SEQ000070741	SAP18-6199	MeanR1Qual	≥ 30	**35.1**
SEQ000070741	SAP18-6199	MeanR2Qual	≥ 30	**32.7**
SEQ000070741	SAP18-6199	MeanDepth	≥ 20	**103.15**
SEQ000070741	SAP18-6199	MeanInsert	≥ 144.9	**478.79**
SEQ000070741	SAP18-6199	NG50	≥ 76894.5	**325078**
SEQ000070741	SAP18-6199	Contigs	≤ 116.5	**69**
SEQ000070742	SAP18-0432	Taxonomy Check	PASS	**PASS**
SEQ000070742	SAP18-0432	MeanR1Qual	≥ 30	**35**
SEQ000070742	SAP18-0432	MeanR2Qual	≥ 30	**32.3**
SEQ000070742	SAP18-0432	MeanDepth	≥ 20	**88.05**
SEQ000070742	SAP18-0432	MeanInsert	≥ 180.65	**467.6**
SEQ000070742	SAP18-0432	NG50	≥ 38077.5	**479246**
SEQ000070742	SAP18-0432	Contigs	≤ 102.5	**66**
SEQ000070743	LMP18-H8393	Taxonomy Check	PASS	**PASS**
SEQ000070743	LMP18-H8393	MeanR1Qual	≥ 30	**36.5**
SEQ000070743	LMP18-H8393	MeanR2Qual	≥ 30	**34**
SEQ000070743	LMP18-H8393	MeanDepth	≥ 20	**130.35**
SEQ000070743	LMP18-H8393	MeanInsert	≥ 160.25	**428.81**
SEQ000070743	LMP18-H8393	NG50	≥ 67986.5	**523238**
SEQ000070743	LMP18-H8393	Contigs	≤ 159	**53**
SEQ000070744	LMP18-H2446	Taxonomy Check	PASS	**PASS**
SEQ000070744	LMP18-H2446	MeanR1Qual	≥ 30	**36.5**
SEQ000070744	LMP18-H2446	MeanR2Qual	≥ 30	**33.9**
SEQ000070744	LMP18-H2446	MeanDepth	≥ 20	**96.87**
SEQ000070744	LMP18-H2446	MeanInsert	≥ 172.6	**434.23**
SEQ000070744	LMP18-H2446	NG50	≥ 347436.5	**497410**
SEQ000070744	LMP18-H2446	Contigs	≤ 128.5	**64**

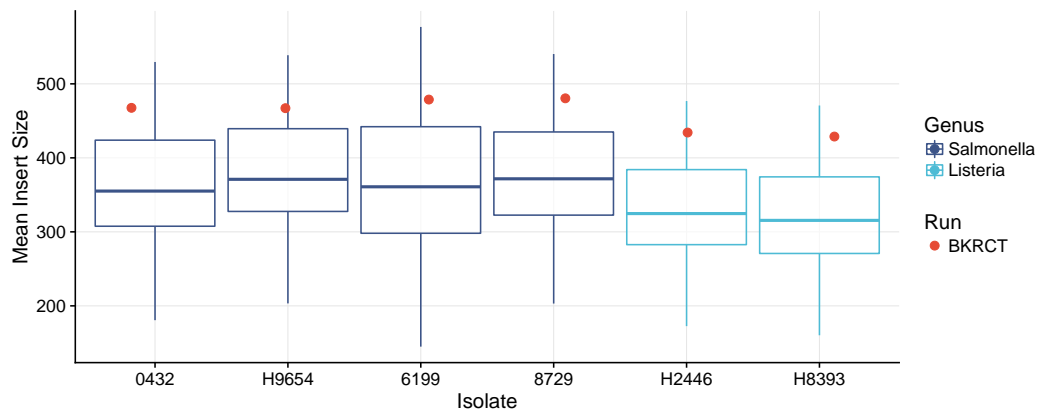**Table 3:** Competency Assessment Metrics - Quantitative Results.
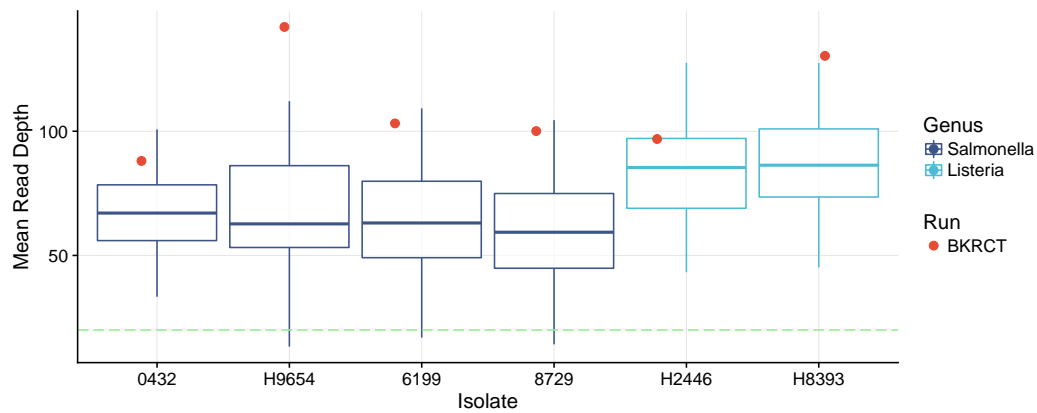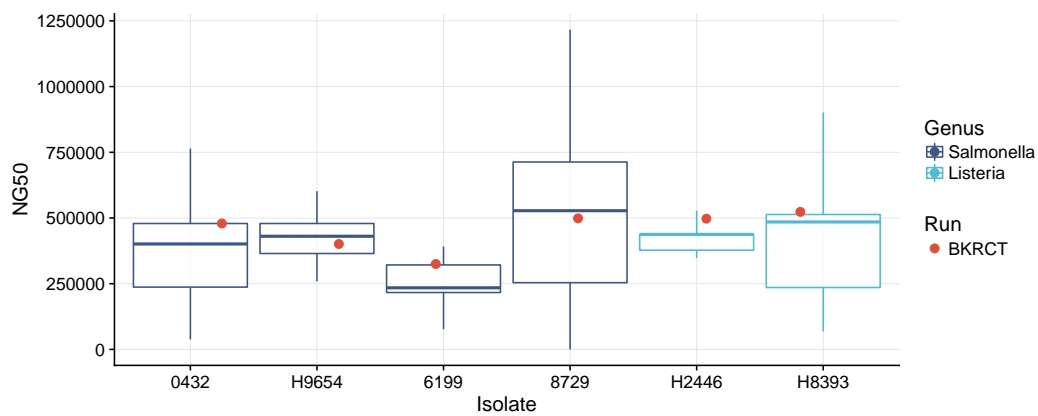
Figure 1: Mean depth
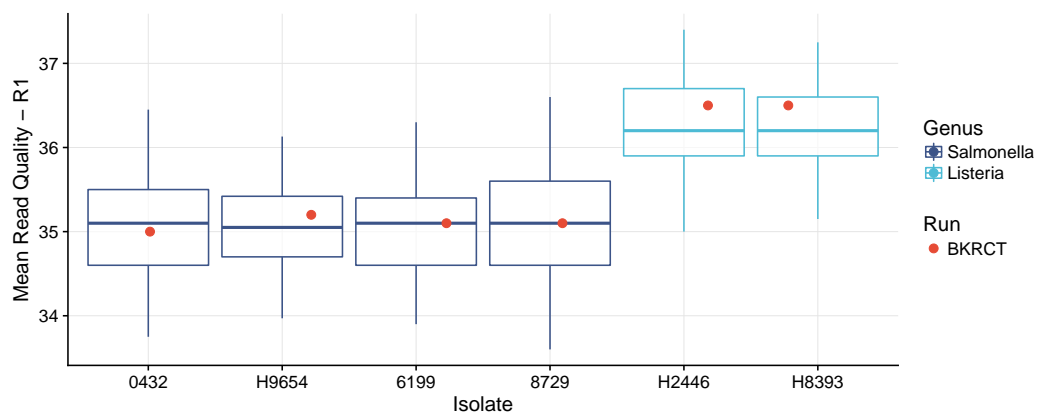
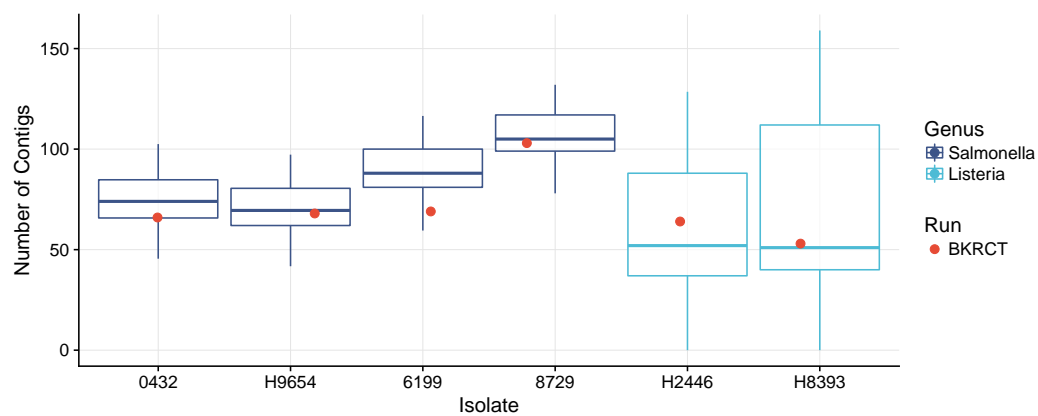**Figure 2:** Mean read quality score - forward reads (R1)
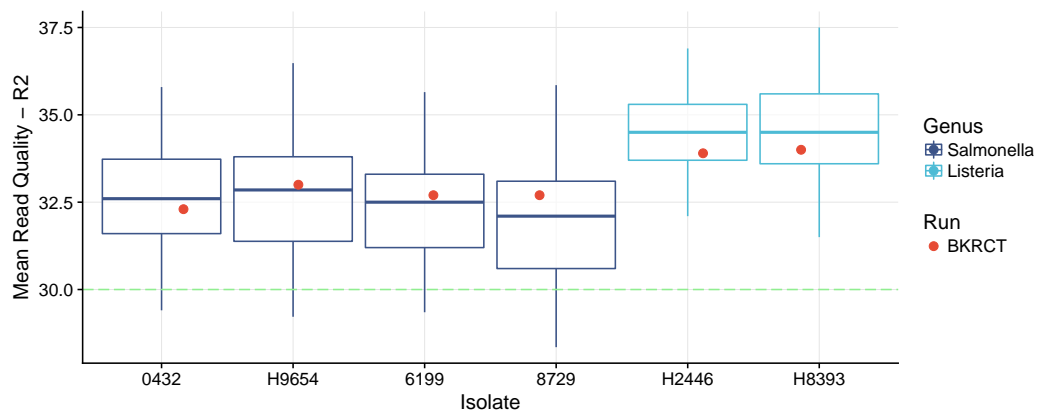
Figure 3: Mean read quality score - reverse reads (R2)

Figure 4: Mean insert size

Figure 5: NG50


Figure 6: Contigs

2.2 Additional Metrics

The following metrics are provided for informational purposes. (**Figs. 7-15**) -

2.2.1 Run Metrics

The following run metrics were evaluated (**Fig. 7**) -

a. Cluster Density (K/mm^2) - density of clusters for each tile.
b. Clusters PF (%) - the percentage of clusters passing filter for each tile.
c. Reads PF (M) - the number of reads/clusters passing filter (millions).
d. Yield (Gb) - the total yield in Gigabases.
e. Bases >= Q30 (%) - the percentage of bases with a quality score >= 30.
f. Reads Identified (%) - the percentage of Reads PF which were assigned to an index.
g. Indexing CV - the coefficient of variation for the number of counts across all indices.
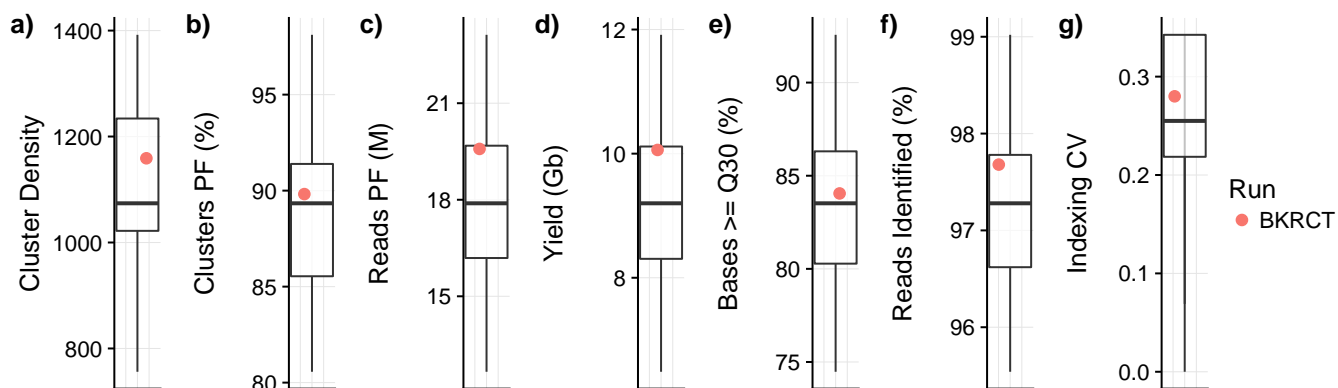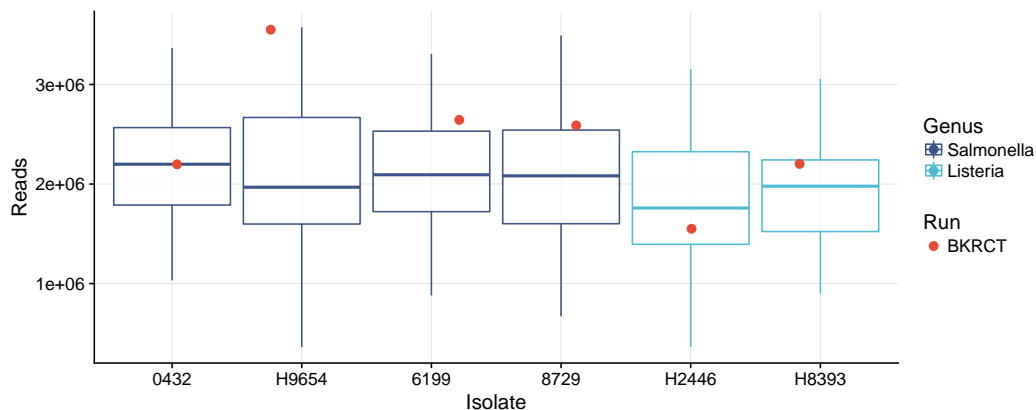

Figure 7: Run Metrics

	BKRCT	Expected
Cluster Density (K/mm2)	1159	800-1100
Clusters PF (%)	89.82	
Reads PF (M)	19.58	12-15
Yield (Gb)	10.06	7.5-8.5
Bases >= Q30 (%)	84.05	>75%
Reads Identified (%)	97.68	
Indexing CV	0.28	

Table 4: Run metrics and manufacturer expected values for MiSeq V2 2x250 runs.

2.2.2   Read Metrics

The following read metrics were evaluated (**Fig. 8**)-

   **8. Reads** - the number of read pairs.

Figure 8: Reads

2.2.3   Alignment Metrics

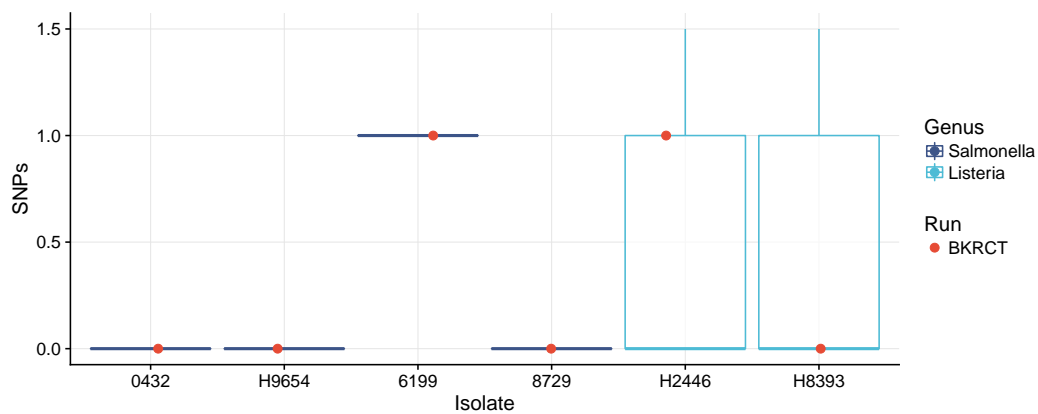The following alignment metrics were evaluated (**Figs. 9-13**) -

   9. Reads mapped (%) - the percentage of reads which could be mapped to the reference.

   10. Genome fraction - the percentage of the reference with $>= 1X$ coverage.

   11. SNPs - the number of SNPs reported.

   12. Low coverage positions - the number of positions with depth of coverage $< 10X$. This metric is highly inversely correlated with assembly quality.

   13.   Coverage Plots - visualizations of coverage depth and variance across each position. Coverage plots were smoothed to 10000 points. Histograms of the coverage depth should resemble a Poisson-like distribution with a small standard deviation.
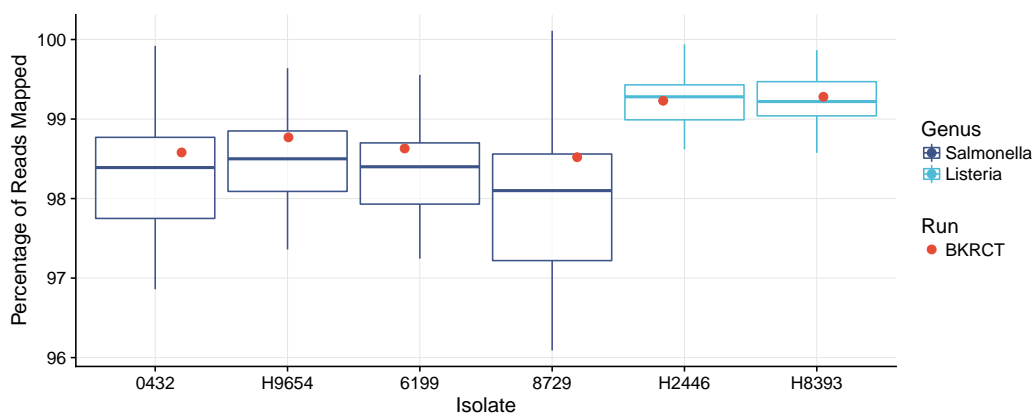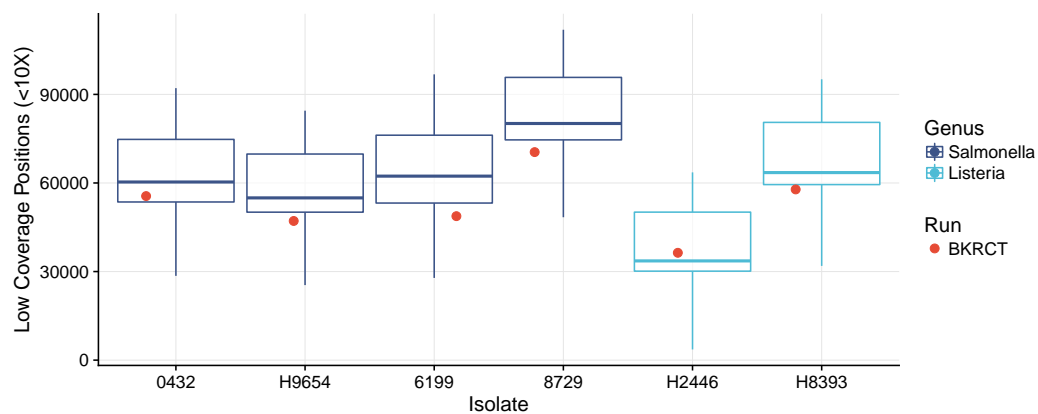

Figure 9: Reads mapped (%)
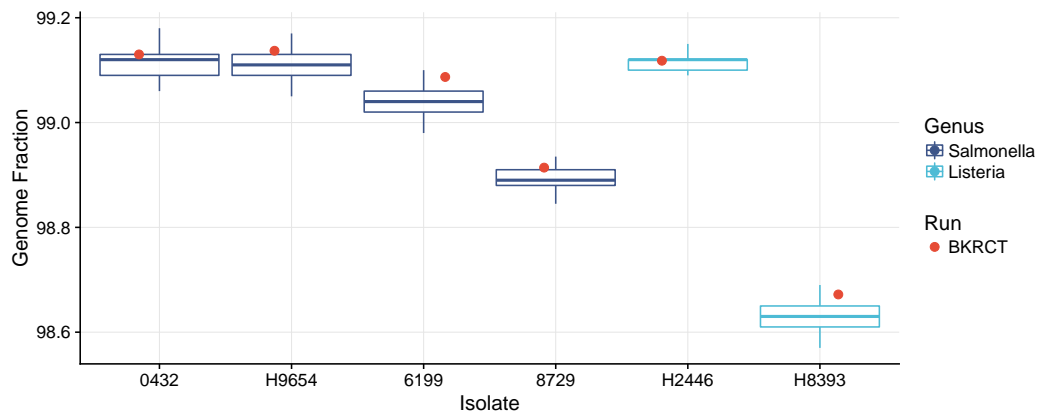
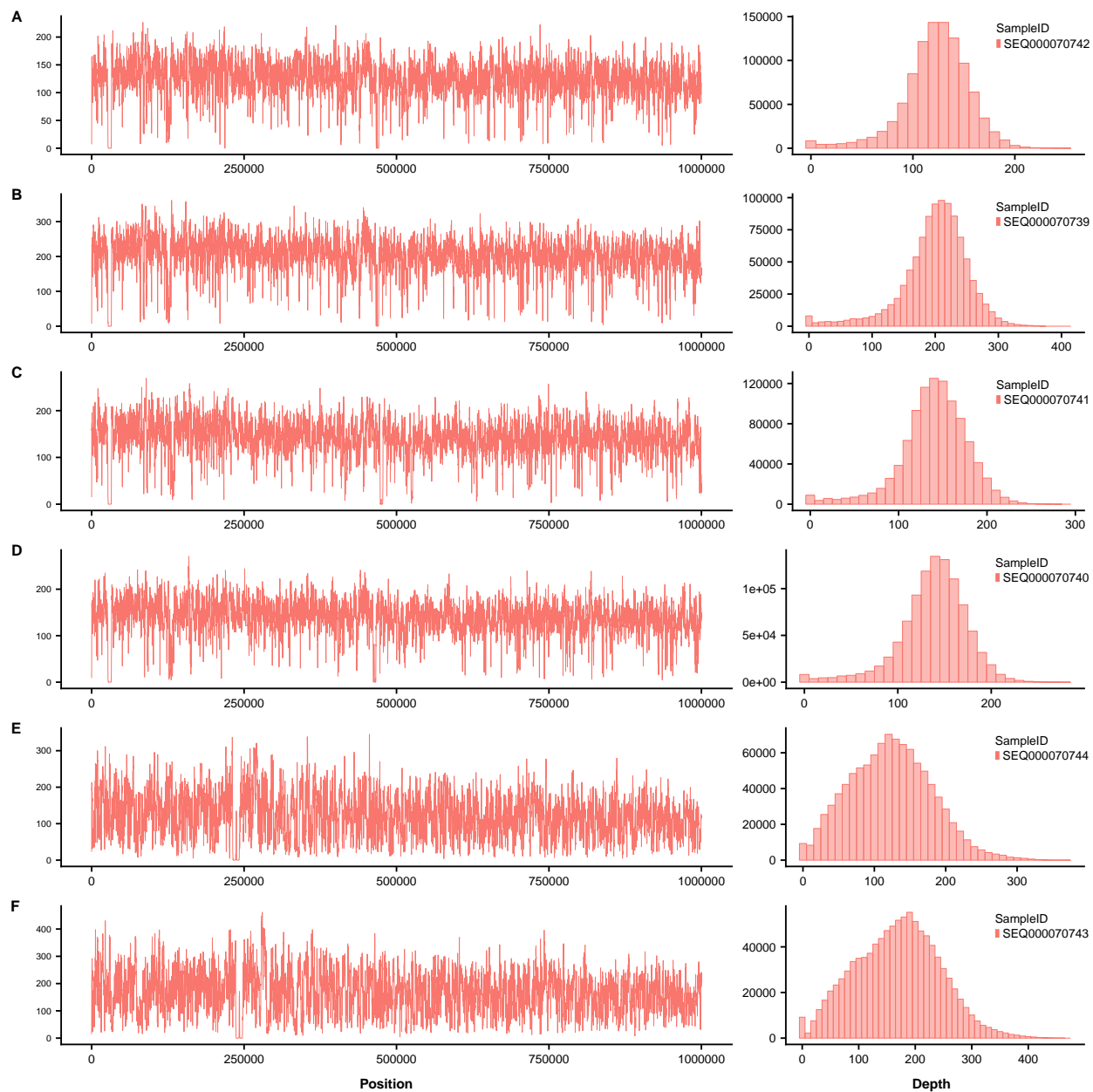**Figure 10:** Genome fraction (%)

Figure 11: SNPs


Figure 12: Low coverage positions ($< 10X$)

Figure 13: Coverage Plots. A.0432 B.H9654 C.6199 D.8729 E.H2446 F.H8393 .

2.2.4 Assembly Metrics

The following assembly metrics were evaluated (**Figs. 14-15**) -

14. Total Length Delta - the difference between the lengths of the assembly and the reference. A perfect assembly would produce a delta of zero. Positive values indicate possible contamination.

15. Unaligned Length - the total length of any assembled contigs which could not be aligned to the reference genome. The presence of unalignable contigs could indicate an unexpected genomic element (plasmid, phage) or contamination.
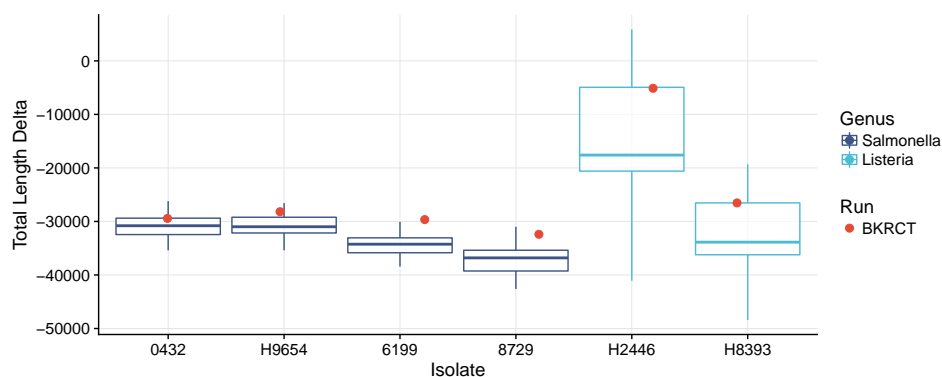
Figure 14: Total Length Delta
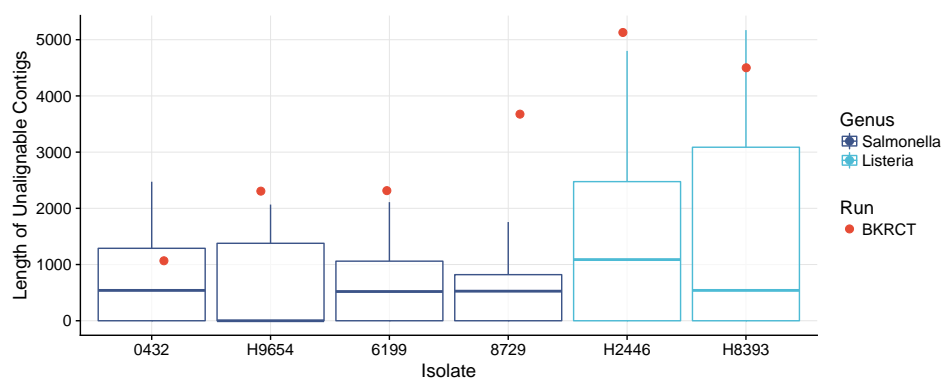
Figure 15: Unaligned Length

A   Appendix

	SampleID	IsolateID	Sequencer	Machine	FlowCell	LibKit	Chemistry	RunDate	SequencedBy	SeqLength
1	SEQ000070739	SAP18-H9654	Illumina MiSeq sequence	M04590	000000000-BKRCT	Nextera XT	V2	2018-02-28	DGS	35-251
2	SEQ000070740	SAP18-8729	Illumina MiSeq sequence	M04590	000000000-BKRCT	Nextera XT	V2	2018-02-28	DGS	35-251
3	SEQ000070741	SAP18-6199	Illumina MiSeq sequence	M04590	000000000-BKRCT	Nextera XT	V2	2018-02-28	DGS	35-251
4	SEQ000070742	SAP18-0432	Illumina MiSeq sequence	M04590	000000000-BKRCT	Nextera XT	V2	2018-02-28	DGS	35-251
5	SEQ000070743	LMP18-H8393	Illumina MiSeq sequence	M04590	000000000-BKRCT	Nextera XT	V2	2018-02-28	DGS	35-251
6	SEQ000070744	LMP18-H2446	Illumina MiSeq sequence	M04590	000000000-BKRCT	Nextera XT	V2	2018-02-28	DGS	35-251

Table 5: Run Metadata.

	SampleID	Reads	MeanR1Qual	MeanR2Qual	PercMapped	MeanDepth	SNPs	MeanInsert	GenomeFraction	CovLT10	NG50	Contigs	LengthDelta	UnalignedLength
1	SEQ000070739	3550542	35.20	33.00	98.77	141.98	0	467.16	99.14	47139	401093	68	-28176	2306
2	SEQ000070740	2588958	35.10	32.70	98.52	100.07	0	480.47	98.91	70458	498666	103	-32396	3676
3	SEQ000070741	2644586	35.10	32.70	98.63	103.15	1	478.79	99.09	48767	325078	69	-29647	2315
4	SEQ000070742	2197876	35.00	32.30	98.58	88.05	0	467.60	99.13	55569	479246	66	-29429	1067
5	SEQ000070743	2202392	36.50	34.00	99.28	130.35	0	428.81	98.67	57823	523238	53	-26534	4501
6	SEQ000070744	1550652	36.50	33.90	99.23	96.87	1	434.23	99.12	36359	497410	64	-5116	5129

Table 6: Run Metadata Continued.