RNA-Seq Quality Assessment Assignment

SJ Kim

2022-09-07

Objectives

Use existing tools for quality assessment and adaptor trimming, compare the quality assessments to those from my software, and to demonstrate some ability to summarize other important information about this RNA-Seq data set.

Data

Library Assignments: /projects/bgmp/shared/Bi622/QAA_data_assignments.txt

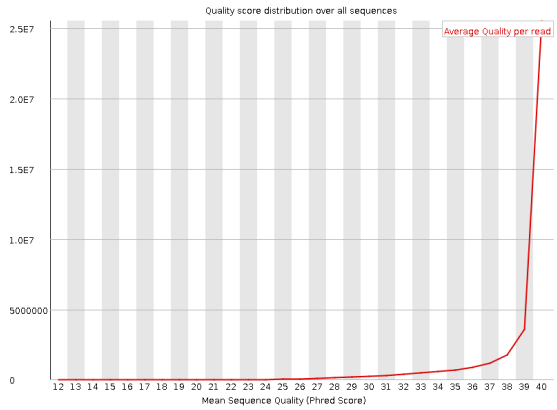
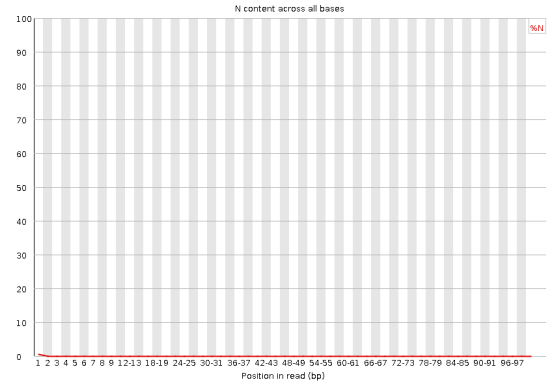
SJ: 22_3H_both_S16_L008 8_2F_fox_S7_L008

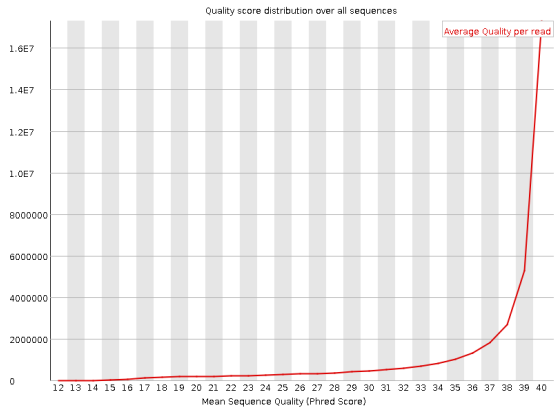
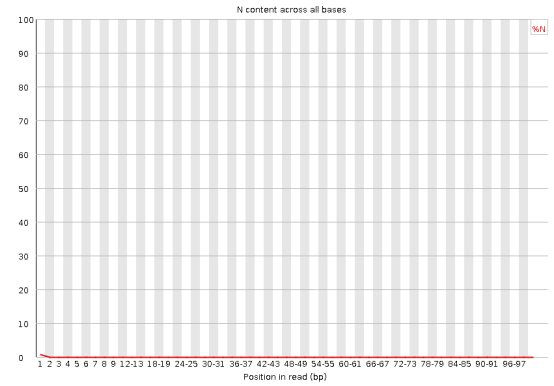
The demultiplexed, gzipped .fastq files are here: /projects/bgmp/shared/2017_sequencing/demultiplexed/

Read Quality Score Distributions

Used FastQC to generate quality score distribution plots and per-base N content plots.

Fig. 1: Quality Score Distribution and per-base N content for sample 22, read 1 and read 2. Low N content is one of many indicators of high quality reads. The per-base N content plots show low amounts of N content in the reads, which is consistent with the quality score distributions.

Sample	Quality Score Distribution	Per-base N Content
Sample 22, Read 1		

Sample	Quality Score Distribution	Per-base N Content
Sample 22, Read 2		

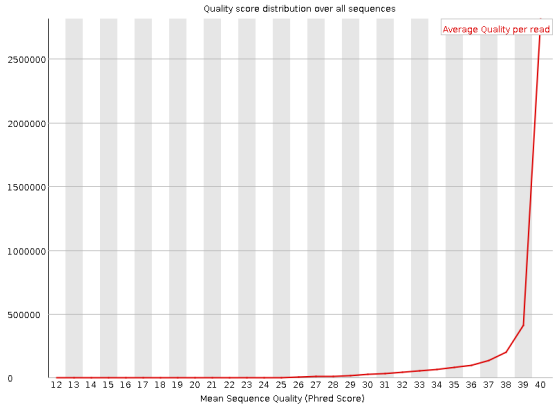
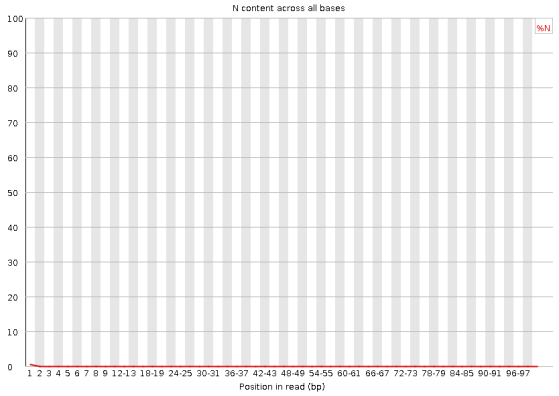
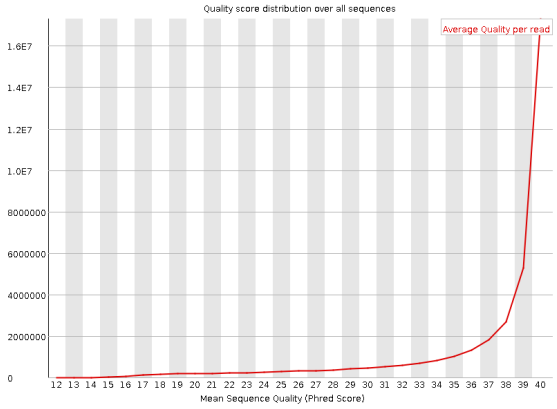
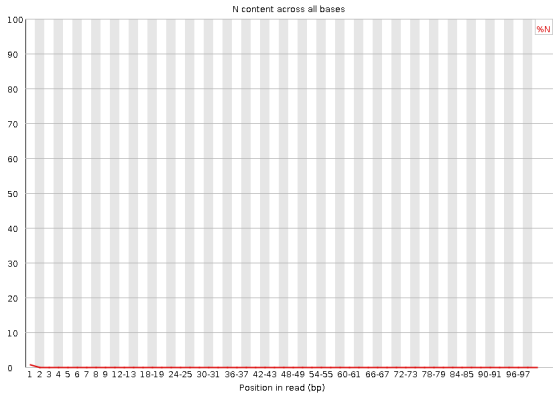
Sample	Quality Score Distribution	Per-base N Content
Sample 8, Read 1		
Sample 8, Read 2		

Fig. 2: Average per-base Quality Distribution: Sample 22 Read 1

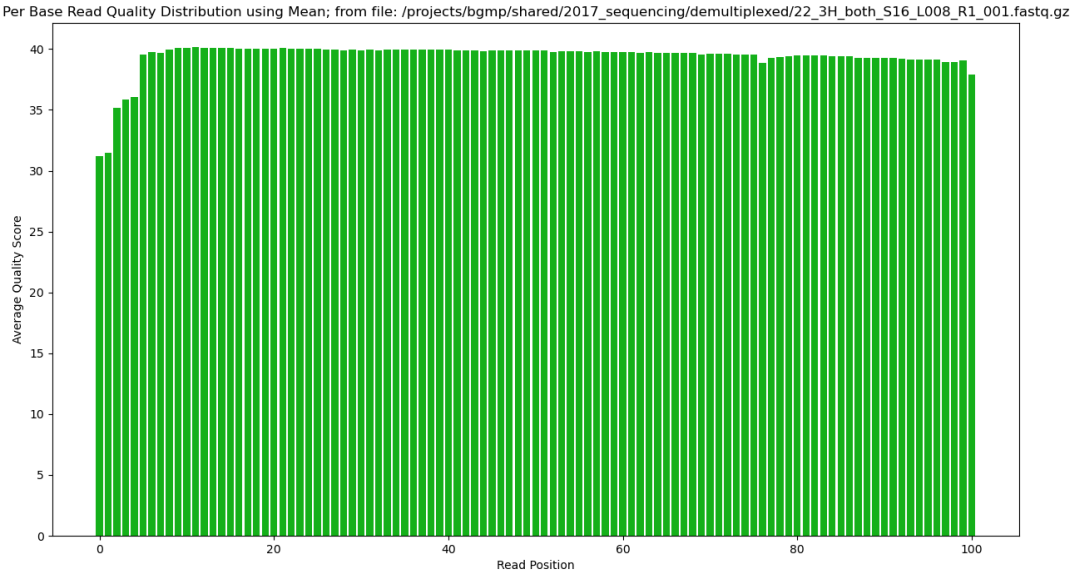


Fig. 3: Average per-base Quality Distribution: Sample 22 Read 2

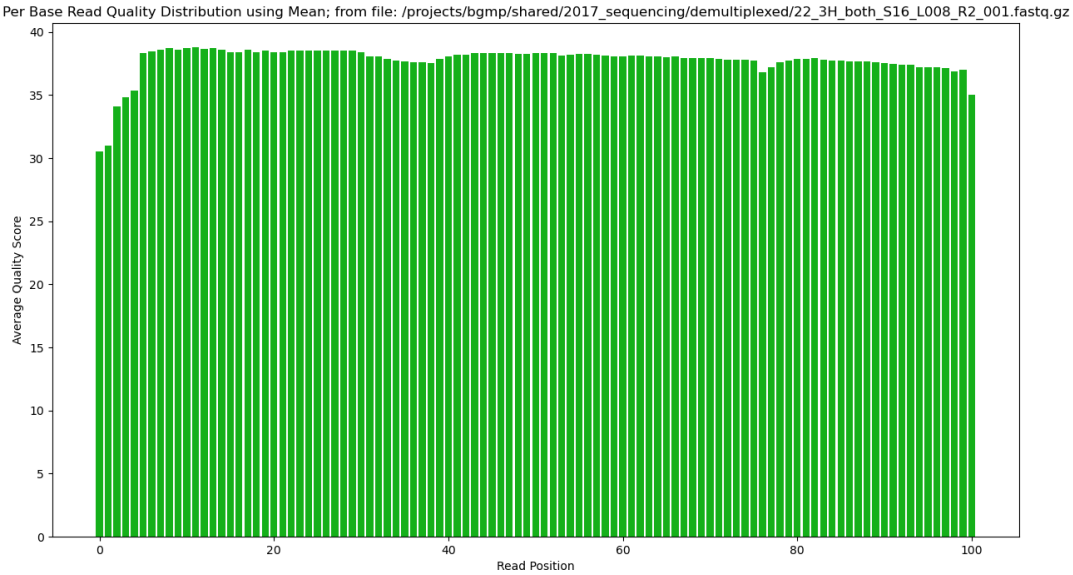


Fig. 4: Average per-base Quality Distribution: Sample 8 Read 1

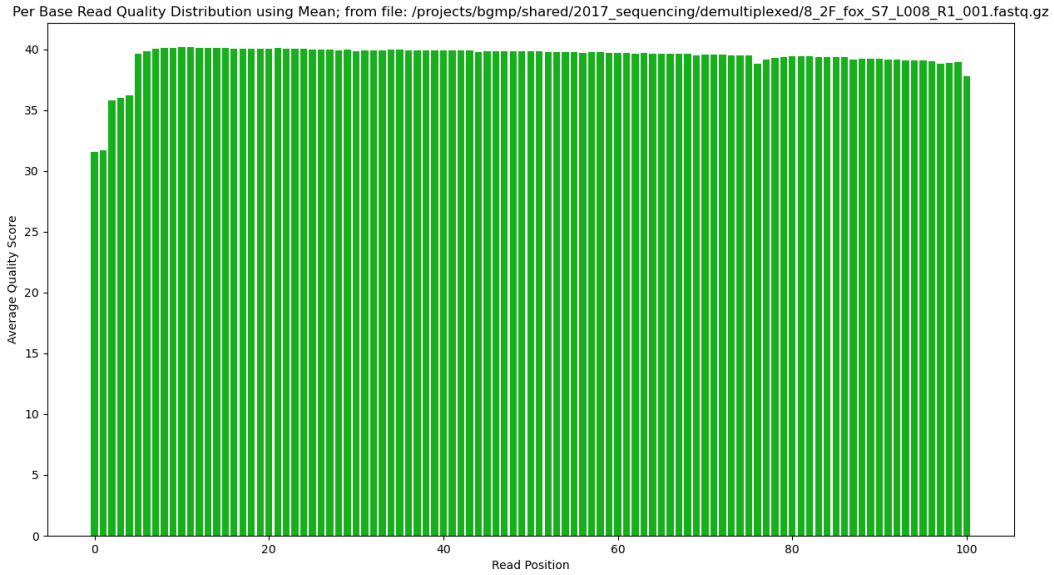
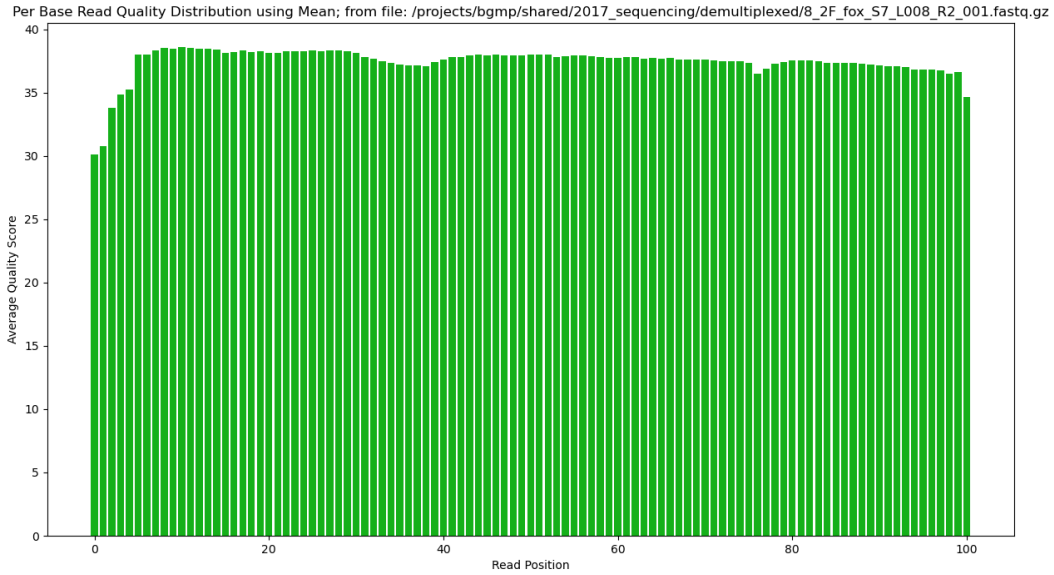


Fig. 5: Average per-base Quality Distribution: Sample 8 Read 2



The overall quality of libraries from sample 8 and sample 22 are very good since the average per-base quality is higher than 30 for most positions and there is low N content. The first ~10 positions have lower quality scores but are still high quality (higher than Phred score of 30) on average. The FastQC plots show very similar information to my script's plots. However, the runtime was very different. FastQC was finished running after 4 minutes, whereas my plotting script took 24 minutes total. FastQC was approximately 6 times faster than my script. FastQC's algorithm has been under development by experts for many years so is likely to be optimized in all possible ways.

Adapter Trimming Comparison

Software versions used:

cutadapt --version (4.1), trimmomatic -version (0.39)

The adapters are Illumina TruSeq adapters:

R1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA

R2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

https://support-docs.illumina.com/SHARE/AdapterSeq/illumina-adapter-sequences.pdf pg 53

```
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/22_3H_both_S16_L008_R1_001.fastq.gz
| grep -c "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA"
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/22_3H_both_S16_L008_R2_001.fastq.gz
 | grep -c "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT"
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/8_2F_fox_S7_L008_R1_001.fastq.gz
 | grep -c "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA"
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/8_2F_fox_S7_L008_R2_001.fastq.gz
 | grep -c "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT"
```

For sample 22, read 1: there were 7,563 sequences with the adapter out of 4,050,899 total sequences.

For sample 22, read 2: there were 7,848 sequences with the adapter out of 4,050,899 total sequences.

For sample 8, read 1: there were 161,695 sequences with the adapter out of 36,482,601 total sequences.

For sample 8, read 2: there were 164,539 sequences with the adapter out of 36,482,601 total sequences.

```
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/22_3H_both_S16_L008_R1_001.fastq.gz
| grep --color=always "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | head
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/22_3H_both_S16_L008_R1_001.fastq.gz
| grep --color=always "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | tail
```

For every sample, there was just a small proportion of sequences with adapters. Adapter content was found using command-line tool 'grep' with option —color=always. The adapters were usually found towards the end of the sequences, and the low proportion of of sequences with adapters indicates longer inserts.

SANITY CHECK! grepping R1 with R2 adapter should result in 0 matches, and vice versa:

```
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/22_3H_both_S16_L008_R1_001.fastq.gz
 | grep -c "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT"
0 #whewf
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/22_3H_both_S16_L008_R2_001.fastq.gz
 | grep -c "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA"
0 #yay
```

Sample 22, cutadapt results:

Total read pairs processed: 4,050,899

Read 1 with adapter: 153,089 (3.8%)

Read 2 with adapter: 186,534 (4.6%)

For sample 22, 3.8% of read 1 had adapters which were trimmed and read 2 had 4.6% adapters which were trimmed.

Sample 8, cutadapt results:

Total read pairs processed: 36,482,601

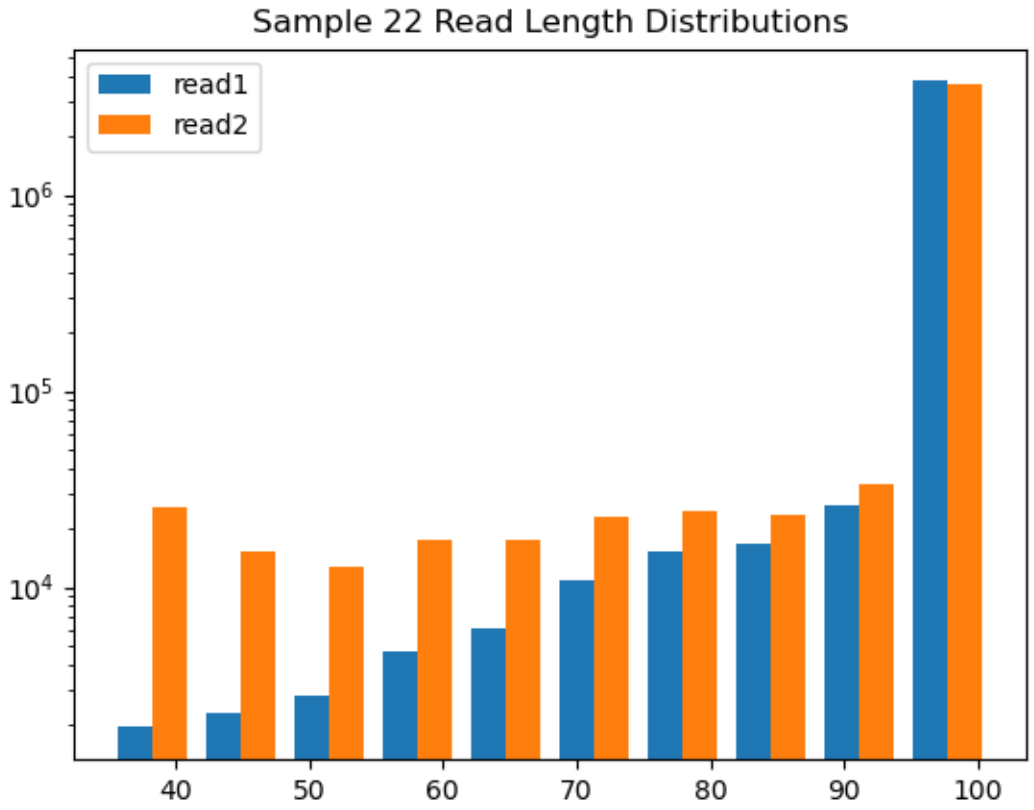
Read 1 with adapter: 2,145,600 (5.9%)

Read 2 with adapter: 2,403,490 (6.6%)

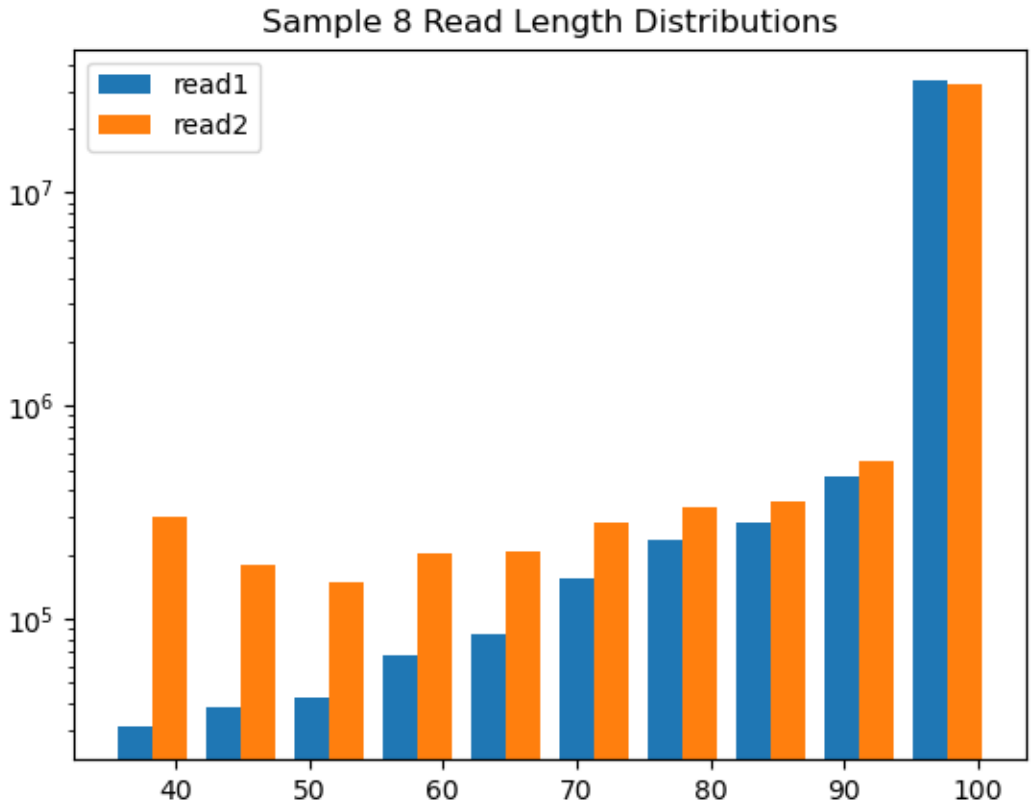
For sample 8, 5.9% of read 1 had adapters that were trimmed and read 2 had 6.6% of adapters that were trimmed.

Trimmed read length distributions, y-axis log scale:

Sample 22's read 1 was trimmed more extensively than read 2.



Sample 8's read 1 was trimmed more extensively than read 2.



Both samples had Read 1 trimmed more than Read 2. This could be due to degradation during sequencing since Read 2 is acquired a while after Read 1 is acquired. In general, Read 2 tends to be slightly lower quality so I would expect read 2 to be trimmed more often.

Alignment and Strand-Specificity

Alignment software:

STAR --version 2.7.10a

Ensembl Release 107

Mouse genome fasta: http://ftp.ensembl.org/pub/release-107/fasta/mus_musculus/dna/Mus_musculus.GRCm39.dna.primary_assembly.fa.gz

Mouse gtf: http://ftp.ensembl.org/pub/release-107/gtf/mus_musculus/Mus_musculus.GRCm39.107.gtf.gz

```
cd /projects/bgmp/skim6/bioinfo/Bi622/QAA/part3-10/
./mapcount.py -f /projects/bgmp/skim6/bioinfo/Bi622/QAA/part3-9_star/Mus_musculus.GRCm39.dna.ens107.STA
#sample 22:
mappedCount:7,677,920
notMapped:124,334
```

Sample 22 had 7,677,920 mapped reads and 124,334 unmapped reads.

```
./mapcount.py -f /projects/bgmp/skim6/bioinfo/Bi622/QAA/part3-9_star/Mus_musculus.GRCm39.dna.ens107.STA
#sample 8:
mappedCount:67,070,995
notMapped:2,511,319
```

Sample 8 had 67,070,995 mapped reads and 2,511,319 unmapped reads.

htseq-count --version: 2.0.2

Is this data from strand-specific RNA-seq libraries?

sample 22

Sum the number of reads that mapped to a feature, calculate the total number of reads. determine percentage.

```
cat fw_22_htseq22205030.genecount | awk '$1~"ENSMUSG" {mapped+=$2} END {print mapped}'
```

fw_22 mapped reads: 148,785

```
cat fw_22_htseq22205030.genecount | awk '{mapped+=$2} END {print mapped}'
```

fw_22 total reads: 3,901,127

```
cat rv_22_htseq22205225.genecount | awk '$1~"ENSMUSG" {mapped+=$2} END {print mapped}'
```

rv_22 mapped reads: 3,394,239

```
cat rv_22_htseq22205225.genecount | awk '{mapped+=$2} END {print mapped}'
```

rv_22 total reads:3,901,127

fw_22 % mapped: 148785/3901127*100 = 3.81%

rv_22 % mapped: 3394239/3901127*100 = 87.0%

The reverse orientation had a much higher percentage of mapped reads, which means that this is a stranded RNA-seq library, with the reverse strand as the "sense" strand.

sample 8

Sum the number of reads that mapped to a feature, calculate the total number of reads. determine percentage.

```
cat fw_8_htseq22205437.genecount | awk '$1~"ENSMUSG" {mapped+=$2} END {print mapped}'
```

fw_8 mapped reads: 1,282,235

```
cat fw_8_htseq22205437.genecount | awk '{mapped+=$2} END {print mapped}'
```

fw_8 total reads: 34,791,157

```
cat rv_8_htseq22205726.genecount | awk '$1~"ENSMUSG" {mapped+=$2} END {print mapped}'
```

rv_8 mapped reads: 28,041,293

```
cat rv_8_htseq22205726.genecount | awk '{mapped+=$2} END {print mapped}'
```

rv_8 total reads: 34,791,157

fw_8 % mapped: 1282235/34791157 * 100 = 3.69%

rv_8 % mapped: 28041293/34791157 * 100 = 80.6%

The reverse orientation mapped 80.6% and forward mapped 3.69%, therefore this is a stranded RNA-seq library and the reverse strand was the "sense" strand.