QAA Lab notebook

PART 1

Project Location:

```
/projects/bgmp/camk/bioinfo/Bi623/QAA
```

SSH to talapas, create conda environment called QAA, installed FASTQC, and checked with files are mine.

```
conda create -n QAA
conda activate QAA
conda install fastqc
```

```
(QAA) [camk@n0350 demultiplexed]$ fastqc --version
FastQC v0.12.1
```

File assignments located:

```
/projects/bgmp/shared/Bi623/QAA_data_assignments.txt
```

My files:

```
21_3G_both_S15_L008
16_3D_mbnl_S12_L008
```

files here:

```
/projects/bgmp/shared/2017_sequencing/demultiplexed/21_3G_both_S15_L008_R1_0
01.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/21_3G_both_S15_L008_R2_0
01.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/16_3D_mbnl_S12_L008_R1_0
01.fastq.gz
```

```
/projects/bgmp/shared/2017_sequencing/demultiplexed/16_3D_mbnl_S12_L008_R2_0
01.fastq.gz
```

Running Fastqc:

```
fastqc
/projects/bgmp/shared/2017_sequencing/demultiplexed/21_3G_both_S15_L008_R1_0
01.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/21_3G_both_S15_L008_R2_0
01.fastq.gz --outdir .
```

```
fastqc
/projects/bgmp/shared/2017_sequencing/demultiplexed/16_3D_mbnl_S12_L008_R1_0
01.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/16_3D_mbnl_S12_L008_R2_0
01.fastq.gz --outdir .
```

SCP'd the output fastqc analysis files to my own computer:

Command:

```
scp tlp1:/projects/bgmp/camk/bioinfo/Bi623/QAA/* .
```

Location:

```
/Users/cameronkunstadt/bioinfo/Bi623
```

Running Quality Score plotting script:

Location:

```
/projects/bgmp/camk/bioinfo/Bi622/Demultiplex
```

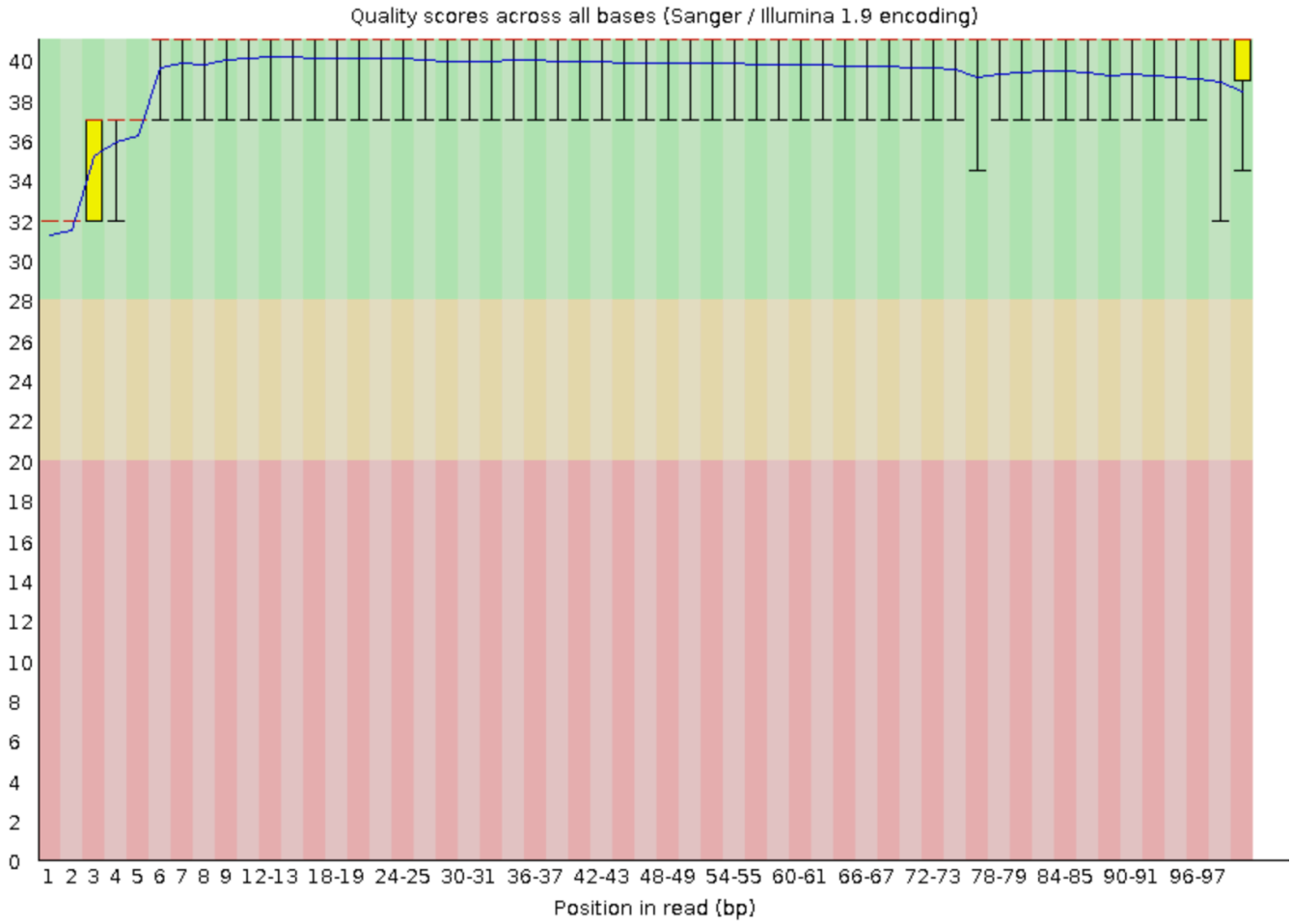
```
run_plot.QAA.sh
```

This has my code for running the plots using my old demultiplexing script. Once it finished I moved the jpgs to my computer here:

```
/Users/cameronkunstadt/bioinfo/Bi623
```

16_3D_mbnl_S12_L008_R1_001.jpg

fastqc version



All in all they look very similar, fastqc has the added error bars

ADD MORE HERE

PART 2

Install cutadapt and timmomatic:

```
conda install cutadapt
conda install trimmomatic
```

version check:

```
(QAA) [camk@n0350 QAA]$ cutadapt --version
4.9
```

```
(QAA) [camk@n0350 QAA]$ trimmomatic -version
0.39
```

I'm assuming for identifying the adapter sequences, we can just look at the fastq report for each file and look at the over represented sequences, and trust that thats the adapter for each one. By visual they all look good.

21_3G_both_S15_L008_R1_001.fastq.gz overrepresented sequence:

```
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCCTAAGATCTCGTAT
```

21_3G_both_S15_L008_R2_001.fastq.gz overrepresented sequence:

```
GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCTTAGGACGTGTAGATCT
```

16_3D_mbnl_S12_L008_R1_001.fastq.gz overrepresented sequence:

```
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACGATCAGATCTCGTAT
```

16_3D_mbnl_S12_L008_R2_001.fastq.gz overrepresented sequence:

```
GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCTGATCGTGTGTAGATCT
```

These are the real ones:

```
R1: `AGATCGGAAGAGCACACGTCTGAACTCCAGTCA`

R2: `AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT`
```

So they're close but have this extra A at the beginning, and my overrepresented sequences are too long. Don't know why, continuing with trimming.

So I used this to check how many of my sequences had adapters in them:

```
(base) [camk@login1 demultiplexed]$ zcat 21_3G_both_S15_L008_R1_001.fastq.gz
| grep "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | wc -l
66732

(base) [camk@login1 demultiplexed]$ zcat 21_3G_both_S15_L008_R1_001.fastq.gz
```

```
| wc -l
36949196
```

(66732 / 36949196) * 100 = 0.18% adapter content

```
(base) [camk@login1 demultiplexed]$ zcat 16_3D_mbnl_S12_L008_R1_001.fastq.gz
| grep "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | wc -l
115556

(base) [camk@login1 demultiplexed]$ zcat 16_3D_mbnl_S12_L008_R1_001.fastq.gz
| wc -l
32940788
```

(115556 / 32940788) * 100 = 0.35% adapter content

```
(base) [camk@login1 demultiplexed]$ zcat 21_3G_both_S15_L008_R2_001.fastq.gz
| grep "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" | wc -l
67707

(base) [camk@login1 demultiplexed]$ zcat 21_3G_both_S15_L008_R2_001.fastq.gz
| wc -l
36949196
```

(67707 / 36949196) * 100 = 0.18% adapter content

```
(base) [camk@login1 demultiplexed]$ zcat 16_3D_mbnl_S12_L008_R2_001.fastq.gz
| grep "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" | wc -l
115921

(base) [camk@login1 demultiplexed]$ zcat 16_3D_mbnl_S12_L008_R2_001.fastq.gz
| wc -l
32940788
```

(115921 / 32940788) * 100 = 0.35% adapter content

Cutting adapters:

```
cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -o
21_3G_both_S15_L008_R1_001.cut.fastq.gz
```

```
/projects/bgmp/shared/2017_sequencing/demultiplexed/21_3G_both_S15_L008_R1_0
01.fastq.gz
```

```
cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -o
16_3D_mbnl_S12_L008_R1_001.cut.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/16_3D_mbnl_S12_L008_R1_0
01.fastq.gz
```

```
cutadapt -a AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o
21_3G_both_S15_L008_R2_001.cut.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/21_3G_both_S15_L008_R2_0
01.fastq.gz
```

```
cutadapt -a AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o
16_3D_mbnl_S12_L008_R2_001.cut.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/16_3D_mbnl_S12_L008_R2_0
01.fastq.gz
```

Huge Trimmomatic commands:

```
trimmomatic PE 16_3D_mbnl_S12_L008_R1_001.cut.fastq.gz \
16_3D_mbnl_S12_L008_R2_001.cut.fastq.gz \
trimmed/16_3D_mbnl_S12_L008_R1_001_paired.trim.cut.fastq.gz \
trimmed/16_3D_mbnl_S12_L008_R1_001_unpaired.trim.cut.fastq.gz \
trimmed/16_3D_mbnl_S12_L008_R2_001_paired.trim.cut.fastq.gz \
trimmed/16_3D_mbnl_S12_L008_R2_001_unpaired.trim.cut.fastq.gz \
LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35
```

```
trimmomatic PE 21_3G_both_S15_L008_R1_001.cut.fastq.gz \
21_3G_both_S15_L008_R2_001.cut.fastq.gz \
trimmed/21_3G_both_S15_L008_R1_001_paired.trim.cut.fastq.gz \
trimmed/21_3G_both_S15_L008_R1_001_unpaired.trim.cut.fastq.gz \
trimmed/21_3G_both_S15_L008_R2_001_paired.trim.cut.fastq.gz \
trimmed/21_3G_both_S15_L008_R2_001_unpaired.trim.cut.fastq.gz \
LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35
```

Saved into a file called run_trim.sh, takes a bit under 20 minutes I think. I forgot to add the time stuff, but the slurm.out file shows this:

```
Quality encoding detected as phred33
Input Read Pairs: 9237299 Both Surviving: 8853312 (95.84%) Forward Only
Surviving: 335058 (3.63%) Reverse Only Surviving: 6909 (0.07%) Dropped:
42020 (0.45%)
TrimmomaticPE: Completed successfully
```

Getting read length distributions for plotting

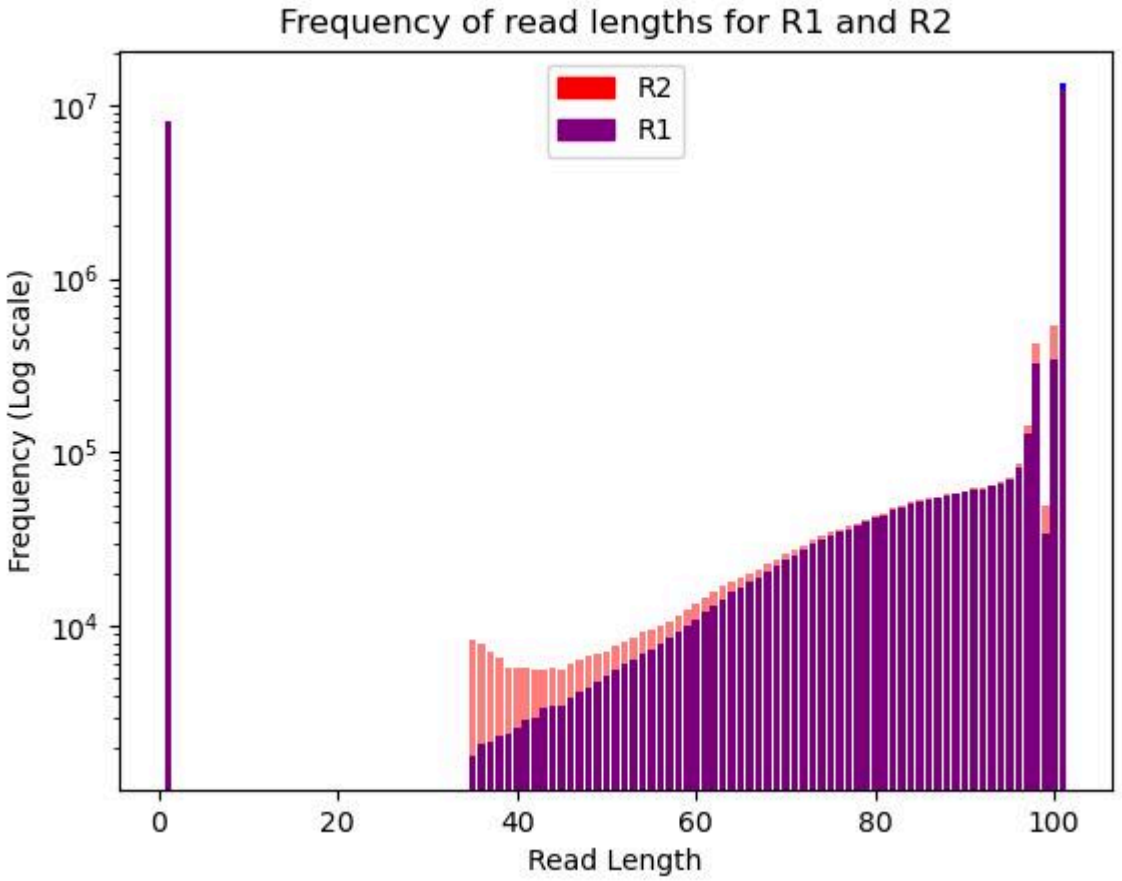
```
(QAA) [camk@n0349 trimmed]$ zcat
16_3D_mbnl_S12_L008_R1_001_paired.trim.cut.fastq.gz | grep -v -e '^@' -e '--
' | awk '{print length}' | sort | uniq -c | sort >
16_3D_mbnl_S12_L008_R1_dist.txt
```

```
(QAA) [camk@n0349 trimmed]$ zcat
16_3D_mbnl_S12_L008_R2_001_paired.trim.cut.fastq.gz | grep -v -e '^@' -e '--
' | awk '{print length}' | sort | uniq -c | sort >
16_3D_mbnl_S12_L008_R2_dist.txt
```

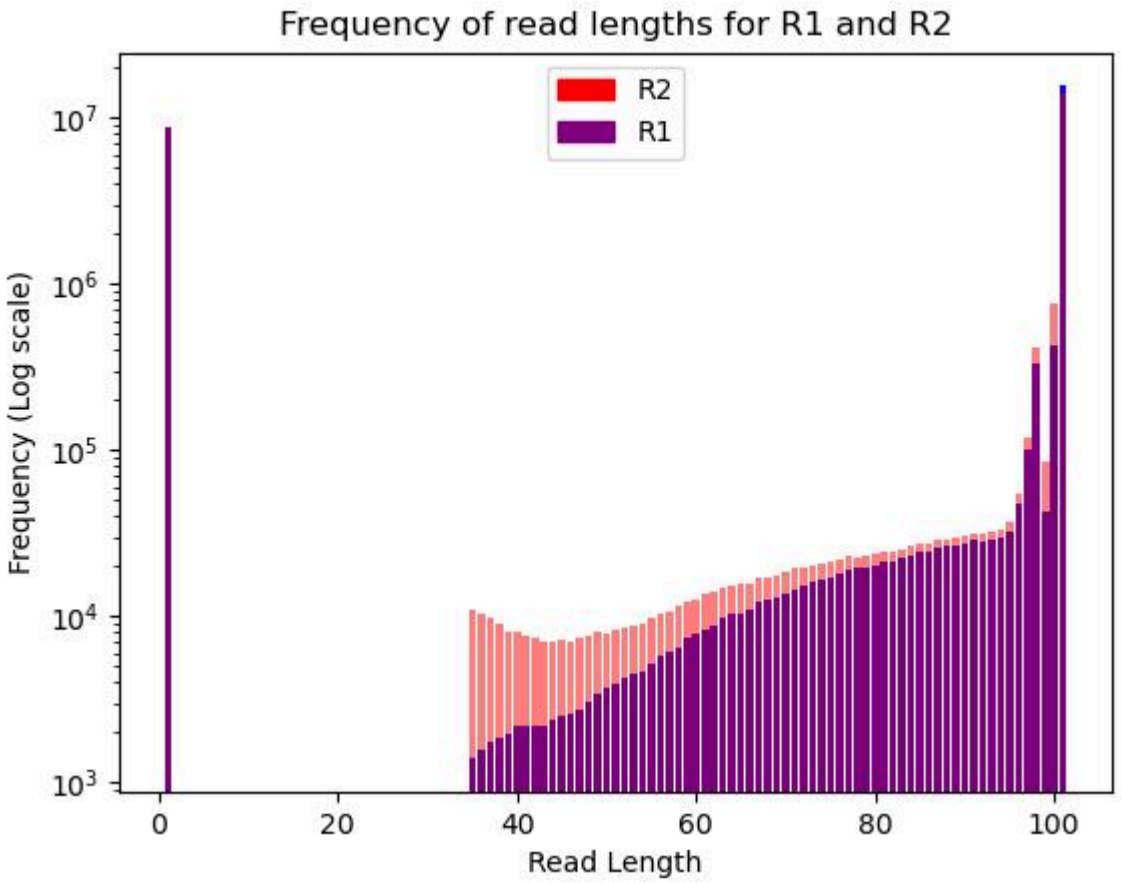
```
(QAA) [camk@n0349 trimmed]$ zcat
21_3G_both_S15_L008_R1_001_paired.trim.cut.fastq.gz | grep -v -e '^@' -e '--
' | awk '{print length}' | sort | uniq -c | sort >
21_3G_both_S15_L008_R1_dist.txt
```

```
(QAA) [camk@n0349 trimmed]$ zcat
21_3G_both_S15_L008_R2_001_paired.trim.cut.fastq.gz | grep -v -e '^@' -e '--
' | awk '{print length}' | sort | uniq -c | sort >
21_3G_both_S15_L008_R2_dist.txt
```

python plot_dist.py -f1 16_3D_mbnl_S12_L008_R1_dist.txt -f2
16_3D_mbnl_S12_L008_R2_dist.txt



python plot_dist.py -f1 21_3G_both_S15_L008_R1_dist.txt -f2 21_3G_both_S15_L008_R2_dist.txt



Part 3

Installing software to environment:

```
(QAA) [camk@n0349 trimmed]$ mamba install star
Preparing transaction: done
Verifying transaction: done
Executing transaction: done

(QAA) [camk@n0349 trimmed]$ mamba install numpy
Preparing transaction: done
Verifying transaction: done
Executing transaction: done

(QAA) [camk@n0349 trimmed]$ mamba install matplotlib
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
```

```
(QAA) [camk@n0349 trimmed]$ mamba install htseq
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
```

genome:

```
/projects/bgmp/camk/bioinfo/Bi623/QAA/Mus_musculus.GRCm39.dna.primary_assembly.fa
```

output folder:

```
/projects/bgmp/camk/bioinfo/Bi623/QAA/Mus_musculus.GRCm39.dna.ens112.STAR_2.7.11b
```

```
/projects/bgmp/camk/bioinfo/Bi623/QAA/trimmed/16_3D_mbnl_S12_L008_R1_001_paired.trim.cut.fastq.gz

/projects/bgmp/camk/bioinfo/Bi623/QAA/trimmed/16_3D_mbnl_S12_L008_R2_001_paired.trim.cut.fastq.gz


/projects/bgmp/camk/bioinfo/Bi623/QAA/trimmed/21_3G_both_S15_L008_R1_001_paired.trim.cut.fastq.gz

/projects/bgmp/camk/bioinfo/Bi623/QAA/trimmed/21_3G_both_S15_L008_R2_001_paired.trim.cut.fastq.gz
```

Database creation script:

This should have been named something else

```
run_star.sh
```

looks like:

```
#!/bin/bash
#SBATCH --account=bgmp
#SBATCH --partition=bgmp
#SBATCH -c 1
#SBATCH --nodes=1

mamba activate QAA

/usr/bin/time **-v** STAR **--runThreadN** 8 **--runMode** genomeGenerate \
--genomeDir
/projects/bgmp/camk/bioinfo/Bi623/QAA/Mus_musculus.GRCm39.dna.ens112.STAR_2.
7.11b \
--genomeFastaFiles
/projects/bgmp/camk/bioinfo/Bi623/QAA/Mus_musculus.GRCm39.dna.primary_assemb
ly.fa
--sjdbGTFfile
/projects/bgmp/camk/bioinfo/Bi623/QAA/Mus_musculus.GRCm39.112.gtf
```

Ran star

Script:

```
align_reads.sh
```

output files:

```
16_3D_mbnl_S12_L008_Aligned.out.sam

21_3G_both_S15_L008_Aligned.out.sam
```

Ran is_mapped.py script from PS8

```
python is_mapped.py -f 16_3D_mbnl_S12_L008_Aligned.out.sam -g K00337:83:

(base) [camk@n0349 QAA]$ python is_mapped.py -f
16_3D_mbnl_S12_L008_Aligned.out.sam -g K00337:83:
```

```
Mapped Reads: 15662583
Unmapped Reads: 365733
```

```
(base) [camk@n0349 QAA]$ python is_mapped.py -f
21_3G_both_S15_L008_Aligned.out.sam -g K00337:83:

Mapped Reads: 17061180
Unmapped Reads: 645444
```

htseq -count

```
(QAA) [camk@n0349 QAA]$ htseq-count 16_3D_mbnl_S12_L008_Aligned.out.sam
Mus_musculus.GRCm39.112.gtf --stranded=yes
```

```
cat 16_3D_mbnl_S12_L008.strand.txt

__no_feature 7,107,777
__ambiguous 6384
__too_low_aQual 7398
__not_aligned 178950
__alignment_not_unique 388074
```

```
cat 16_3D_mbnl_S12_L008.rev.txt

__no_feature 429,966
__ambiguous 136,863
__too_low_aQual 7398
__not_aligned 178950
__alignment_not_unique 388074
```

```
cat 21_3G_both_S15_L008.strand.txt

__no_feature 7,811,301
__ambiguous 7301
__too_low_aQual 12685
```

```
__not_aligned 315910
__alignment_not_unique 383496
```

```
cat 21_3G_both_S15_L008.rev.txt

__no_feature 819,790
__ambiguous 141100
__too_low_aQual 12685
__not_aligned 315910
__alignment_not_unique 383496
```