Genomics Assignment 5
Xiaojun Gao

Part 1:

First, copy all data files to my own directory:

```
cd /opt/ccb/data/RNAseq_project/parathyroid_tumor_samples/fastq_data
cp * homework5
```

Next, run hisat2 on all of the files. I named the samples from 1 to 11, so below is an example of code for sample1, SRR479052:

```
hisat2 -x /opt/ccb/data/grch38/indexes/chr17 -1
SRR479052_chr17_1.fastq.gz -2 SRR479052_chr17_2.fastq.gz -S sample1.sam
```

Converting the result from hista2 to bam format by samtools.

```
samtools sort -@ 8 -o sample1.bam sample1.sam
```

Obtaining the mapped read and unmapped read count by Samtools:

```
samtools view -c -F 260 sample1.bam
samtools view -c –f 4 sample1.bam
```

Using awk to put the count of N in the 6th field of each line to the 23rd field of each line (last field) and outputting to file ending in .count:

```
awk -F '\t' '{ $23 = gsub("N","N",$6) }; 1' sample1.sam > sample1.count
```

Use the following program to count the number of lines that has non-zero number of N at column 6.

```
g++ -o myProgram 1.cpp
./myProgram sample1.count
```

Result:
Sample 1 SRR479052
Mapped reads 1124466
Unmapped reads 838
Spliced alignments 435463

Sample 2 SRR479054
Mapped reads 637263
Unmapped reads 511
Spliced alignments 247180

Sample 3 SRR479056
Mapped reads 693578
Unmapped reads 560
Spliced alignments 268856

Sample 4 SRR479058
Mapped reads 1016791
Unmapped reads 787
Spliced alignments 402472

```
Sample 5 SRR479061
Mapped reads 2192064
Unmapped reads 1144
Spliced alignments 860578

Sample 6 SRR479064
Mapped reads 2017209
Unmapped reads 1029
Spliced alignments 786478

Sample 7 SRR479066
Mapped reads 1016750
Unmapped reads 866
Spliced alignments 386525

Sample 8 SRR479068
Mapped reads 1720424
Unmapped reads 1408
Spliced alignments 662447

Sample 9 SRR479070
Mapped reads 3249566
Unmapped reads 2820
Spliced alignments 1281385

Sample 10 SRR479073
Mapped reads 991012
Unmapped reads 864
Spliced alignments 371517

Sample 11 SRR479076
Mapped reads 843254
Unmapped reads 744
Spliced alignments 311827
```

Part 2:

Building the index file named indexfile:

```
hisat2-build /opt/ccb/data/RNAseq_project/
Schizosaccharomyces_pombe.ASM294v2.30.dna.genome.fa indexfile
```

Running hisat2 default settings and generating output .sam file.

```
hisat2 -x indexfile -1 /opt/ccb/data/RNAseq_project/
S_pombe_SRR2833398_1.fastq.gz -2 /opt/ccb/data/RNAseq_project/
S_pombe_SRR2833398_2.fastq.gz -S part2.sam
```

Converting the result from hista2 to bam format by samtools:

```
samtools sort -@ 8 -o part2.bam part2.sam
```

Obtaining the mapped read and unmapped read count by Samtools:

```
samtools view -c -F 260 part2.bam
samtools view -c -f 4 part2.bam
```

Using command and program to obtain number of spliced alignments:

```
awk -F '\t' '{ $23 = gsub("N","N",$6) }; 1' sample1.sam > sample1.count
g++ -o myProgram 1.cpp
./myProgram part2.count
```

Results with default settings:
Mapped reads: 4685135
Unmapped reads: 307715
Spliced alignments: 207838

The setting I am planning to change is ––pen-noncansplice which has default value 12.
--pen-noncansplice <int> penalty for a non-canonical splice site

The first value I changed into is 20. However, I discovered that the **results does not change from the default settings.** This might be because 12 is high enough for the penalty and a larger penalty would not have an effect on the mapping.

The second value I changed into is 5.
The results are the following:
Mapped reads: 4690751
Unmapped reads: 302099
Spliced alignments: 193862
We can see that under this new settings the mapped reads is more than the number in default setting, while the unmapped reads is less. Moreover, there is also an increase in the number of spiced alignments. This is because the less the penalty for non-canonical splice sites, the more likely it recognize a match with a lower score. That is, the total number of mapped reads should increase.

To further verify, I changed the value to 0.
The results are the following:
Mapped reads: 4695537
Unmapped reads: 297313
Spliced alignments: 285702
Since the mapped reads further increase and unmapped reads decrease compared to when the value is 5, the above explanation is valid.