Homework 3

Task: You have been given a set of RNA sequences that are assembled. You would like to investigate how many of them have poly-A tails and if they have the

Here is a file of assembled mRNA transcripts from an RNAseq experiment in FASTA format.

- Write a script to read in the data, and count the polyA sites
- Generate a distribution of polyA lengths (distance between the site and the end of the tail/contig)
- Compute summary statistics for these lengths (mean, median)
- Plot histogram of this distribution - using R

2. A restriction enzyme cuts DNA at specific locations. Identify the number of cut sites in the genome of Bacillus subtilis of the EcoRI (GAATTC) motif. Your program should simply print out

```
Genome file: [MY FILE]
Number of EcoRI (GAATTC) cut sites: [Number of sites]
...
```

Here is a list of RE sites - you will need to re-write some of these to convert to a regular expression.

```
EcoRI   = "GAATTC"
Bsu15I  = "ATCGAT"
Bsu36I  = "CCTNAGG"
BsuRI   = "GGCC"
EcoRII  = "CCWGG"
```

The abbreviation for DNA patterns is as follows

Only worry about the genome not the plasmids. The paper describing was published in 1997 but the most recent version of the genome B. subtilis genome for strain 168. See all the strains available here and you can find a direct link to the strain 168 FTP folder of the assembly

The goal of this is to write generic code so you are welcome to run this on any genome really. There are many sequenced strains, it would be interesting to compare if the number of cut sites (or their size) varied among strains. Think about (or try) to make your program handle a folder of sequence files to read and provide a report.

3.