Homework 3

Use the following files SNPs and annotation

You will likely need to re-use your solutions from Homework 2.

1. Using the data which present the SNPs and the genes, we would like to find genes which have the most number of changes. This will require counting the number of SNPs in each gene and then dividing the number of SNPs by the length of the gene in kilobases.

   Generate a report that has four columns like this - you will likely need to use BEDTools to generate the information about which genes have SNPs and you will need to distill that down into the count of SNPs per gene. You will need to calculate the length of each gene feature. Finally you will need to compute the 4th column by dividing the SNP count by the gene length and adjust that to kilobases instead of bases. Here is example of expected output.

```
gene_name       length  SNP     SNPs_per_kb
OS06G0510200    1391    50      35.9453630482
OS06G0487620    553     20      36.1663652803
OS06G0120200    2568    93      36.214953271
```

There are several ways to solve this. You should try your best to do as much of it as you can in python. In particular you will need to use some of the string functions to split out some of the parts you don't want from that last column.

2.