Homework 2

BEDTools

Recently the 3000 Rice genome project released data on resequencing 3000 rice genomes. See the site here and OryzaSNP.

You can access this file in the SNPs link below

I downloaded the Rice GFF annotation from Gramene from this FTP location

Use the following files SNPs and annotation

1. How many **genes** have SNPs that overlap them?
2. How many **SNPs** are in the coding regions of genes (denoted by the CDS feature)?

(note you don't need to DO any of the following to do the homework, I am providing you information on how the files were obtained)

For your reference. I downloaded SNP data with allele frequency info from the Rice 3K project (no need for you to download the whole file). I then converted it to BED format and retained only the Chr6 data.

```
wget http://oryzasnp-atcg-irri-org.s3-website-ap-southeast-1.amazonaws.com/3krg-3k_filt_
wget http://oryzasnp-atcg-irri-org.s3-website-ap-southeast-1.amazonaws.com/3krg-3k_filt_
gunzip *.gz
plink --file 3k_filtered --recode vcf-iid -out 3k_filtered.SNP_filt
grep ^6 3k_filtered.SNP_filt.vcf  > chr6_3kSNP_filt.vcf
awk 'BEGIN{OFS="\t"} {print $1,$2-1,$2}' chr6_3kSNP_filt.vcf > rice_chr6_3kSNPs_filt.bed
```