## Q1a

The five output files are,

lab_beagle.dataset.chr15_region.bgl.r2 - This file contains the R^2 for each SNP, this tells us the imputation quality for that SNP. This file has two columns, a SNP identifier and an R^2 value for that SNP.

lab_beagle.dataset.chr15_region.bgl.phased.gz - This file tells us about the phase for each individual, i.e it separates the two alleles into the chromosomes that they belong. This lets us infer the haplotypes.

lab_beagle.dataset.chr15_region.bgl.gprobs.gz - This file contains the genotype probabilities for each individual, the first three columns are the marker information/alleles followed by the probabilities for genotypes A1A1, A1A2 and A2A2.

lab_beagle.dataset.chr15_region.bgl.dose.gz - This file contains the dosage inforamtion, this is formatted as rsid, alleles at that location and dosage for each individual.

lab_beagle.log - This is a summary of the analysis that BEAGLE ran, this contains the command line parameters and running time etc. If the verbose flag is set more information is recorded.

## Q1b

The gprobs file contains the genotype probabilities from Beagle, the dosage file contains the dosage for each individual. The dosage is calculated using the genotype probabilities as dosage = 0*prob(0 alternate alleles) + 1 * prob(1 alternate allele) + 2 * prob(2 alternate alleles)

## Q1c

I would store the genotype probabilities file since the dosage information can be regenerated using the probabilities.

## Q1d

Using the dosage file the probabilities of 0,1,2 T alleles for HG00106 and HG00116 would be, (0, 0, 1) and (0, 0.0386, 0.9614)

## Q1e

For rs1700006, the dosage values for HG00106 and HG00116 are 0 and 0.8684 The alleles are A, G. My guess for the genotypes would be AA and AG. The phased alleles at this location for these individuals are A A, G A hence my guess appears to be correct.

## Q1f

The SNP that imputed most poorly is chr15:78884553 with an R^2 of 0.157 From the BEAGLE documentation "the allelic R^2 cannot be computed for a marker when the most likely genotype is the same for each individual (i.e. only one genotype is observed), or when there is no data for the marker in any Beagle input file. " This results in the NA value.

## Q2a

The format=3 implies there are three genotype probabilities in the dosage file, for A1A1, A1A2 and A2A2 genotypes.

## Q2b

One of the SNPs that was imputed was rs11633585

```
         SNP  A1  A2     FRQ    INFO      OR      SE       P
  rs11633585   A   C  0.9624  0.9534  1.3096  0.8658  0.7554
```

Its allele frequency is 0.9624, odds ratio is 1.3096 and p-value is 0.7554

## Q2c

```
[ramua@fuggle lab_week9]$ awk '$NF <= 0.05' lab_plink.assoc.dosage
         SNP  A1  A2     FRQ    INFO      OR      SE       P
    rs555018   A   G  0.6596  1.0050  0.5200  0.3310 0.04816
    rs660652   A   G  0.3085  1.0027  1.9969  0.3392 0.04145
```

There are two SNPs with p-value less than 0.05 these are both genotyped since these are present in the dataset.chr15_region.bgl file. The p-values and odds ratio are indicated above, the most significant SNP has an odds ratio of 1.99 and a p-value of 0.04145, this SNP was genotyped.