# Genome-wide association study

**EXERCISE**: Researchers are interested in detecting SNPs associated with colorectal cancer (variable `cascon`). To this end, they performed a GWAS using DNA information of 2312 individuals. Data are available in the Biocloud virtual machine. Genotype information is available in PLINK format (files 'colorectal.bed', 'colorectal.bim', 'colorectal.fam'). Phenotypic information can be found in the file 'colorectal.txt' that includes these variables:

```
id: identification number
cascon: case-control status (0: control, 1:case)
sex: gender status (male, female)
age: age in years
smoke: smoking status
bmi: body mass index
```

1. Perform a Genome-wide association study including:

   - Quality control (QC) at both individual and SNP level. NOTE: Skip those QC steps that cannot be performed due to memory space problems in Biocloud or try to figure out how to addressed them (if possible).
   - Get p-values assessing association between SNPs and colorectal cancer (e.g. GWAS analysis).
   - Create a Manhattan plot and highlight those SNP that are statistically significant after Bonferroni correction.
   - Create a Locus Zoom plot for those SNPs that are significantly associated with colon cancer after Bonferroni correction. Use LocusZoom tool that is available here `http://locuszoom.org/`.

2. **TO DELIVER**: A single pdf containing three sections: Methods, Results and Appendix. The first two sections should mimic the sections that are normally written in a manuscript (3 pages as maximum - I will not evaluate anything in other pages than the first three). Appendix should contain R code, figures and tables. The pdf canb be created R Markdown (or `knitr`). Here you can find an introduction to Markdown:

   `https://github.com/isglobal-brge/TeachingMaterials/blob/master/Longitudinal_data_analysis/Reproducible_Research/Reproducible_Research.pdf`

   NOTE: Only 1 pdf file should be uploaded - anything else will be evaluated.