Using the subsets of 1,000 Genomes Project chromosome 22 data (*ALL.1kg.chr22.subset.map* and ALL.1kg.chr22.subset.ped) given in Slack, you need to calculate:

- **Minor allele frequency (MAF):** This is the frequency of the least often occurring allele at a specific location. Most studies are underpowered to detect associations with SNPs with a low MAF and therefore we exclude these SNPs.
- **Missingness rates for each variant**, i.e., the decimal value representing the proportion of individuals missing data for each variant
- **Missingness rates for each individual**, i.e., the decimal value representing the proportion of variants missing data for each individual
- *(Bonus question, optional)* p-value resulting from χ^2 test for deviation from Hardy-Weinberg equilibrium for each variant

Prompt the user for exclusion thresholds for each of the above and generate four output files that list the variants/individuals that fail for each criterion in one column and the relevant value in another column. *You may hard-code the following values; however, bonus points will be assigned for anyone that permits a user-prompt for the following criterion*. These should be inclusive, so if the user inputs 0.05 for individual missingness, you should list all variants/individuals >=5% of variants missing. For submission, use these thresholds for calculations:

- 0.01 for MAF
- 0.12 for variant missingness
- 0.01 for individual missingness
- p-value of 1e-10 for Hardy-Weinberg

Use these file names to write your results:

- MAF- "variant_maf_excluded.txt"
- variant missingness- "variant_miss_excluded.txt"
- individual missingness- "indiv_miss_excluded.txt"
- Hardy-Weinberg- "hardy_weinberg_excluded.txt"

Some reference information:

- These files are in PLINK ped/map format. If you haven't worked with this format before, the PLINK documentation has explanations of each: ped, map.
- Missing values are denoted in PLINK as two zeros: "0 0". You'll need to count the number of missing genotypes for each variant and individual and convert it to a decimal value.
- Wikipedia has a good walkthrough of testing for deviation from Hardy-Weinberg equilibrium. Take a look at chi2.sf in scipy.stats to get the p-value.

More information for bonus question

Hardy-Weinberg equilibrium "states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences." Take a look at the linked wiki article to better understand the assumptions and sources of possible deviation. It is commonly employed as a quality control filtering step prior to genome-wide association testing, to identify loci where genotype frequency deviates from HWE,

indicating that either an allele is under selection or where (more likely) you have a consistent genotyping error.

We will use a χ^2 test for deviation for this assignment. Assuming *p* is the allele frequency of one allele at a given SNP (A in the example below) and *q* is the frequency of the other allele (a in the example below), these values can be calculated as follows:

Table 3: Example Hardy–Weinberg principle calculation

Phenotype	White-spotted (AA)	Intermediate (Aa)	Little spotting (aa)	**Total**
Number	1469	138	5	1612

From this, allele frequencies can be calculated:

$$\begin{aligned} p &= \frac{2 \times \mathrm{obs(AA)} + \mathrm{obs(Aa)}}{2 \times (\mathrm{obs(AA)} + \mathrm{obs(Aa)} + \mathrm{obs(aa)})} \\ &= \frac{1469 \times 2 + 138}{2 \times (1469 + 138 + 5)} \\ &= \frac{3076}{3224} \\ &= 0.954 \end{aligned}$$

and

$$\begin{aligned} q &= 1 - p \\ &= 1 - 0.954 \\ &= 0.046 \end{aligned}$$

Now we use *p* and *q* to calculate expected values for each genotype, like this:

$$\begin{aligned} \mathrm{Exp(AA)} &= p^2 n = 0.954^2 \times 1612 = 1467.4 \\ \mathrm{Exp(Aa)} &= 2pqn = 2 \times 0.954 \times 0.046 \times 1612 = 141.2 \\ \mathrm{Exp(aa)} &= q^2 n = 0.046^2 \times 1612 = 3.4 \end{aligned}$$

Then get the χ^2 value:

$$\begin{aligned} \chi^2 &= \sum \frac{(O-E)^2}{E} \\ &= \frac{(1469 - 1467.4)^2}{1467.4} + \frac{(138 - 141.2)^2}{141.2} + \frac{(5 - 3.4)^2}{3.4} \\ &= 0.001 + 0.073 + 0.756 \\ &= 0.83 \end{aligned}$$

Finally, calculate the p-value for each SNP, assuming 1 degree of freedom and likely using a built-in function like *chisqprob* in *scipy.stats*. You should do this for each SNP in the provided data, and as output, generate a file with a list of SNPs (you can give just the rs* ID in the file) that have p ≤ a user-defined threshold (ask the user for this value using input() or raw_input()). Typically a value like p = 1e-10 is used as the cut-off, but in this case we want your code to prompt the user for the desired threshold.

There are a number of working parts to this assignment, and points will be assigned for each part you get to work.