Test

A common standard for the representation of genomic variation is the **Variant Call Format**, or **VCF**. The formal specification of VCF is described here:

http://samtools.github.io/hts-specs/VCFv4.1.pdf

VCF files contain information on the alternative alleles that exist in a genome (the variants), and which alleles are carried by which individuals or sub-populations (the genotypes).

Your task

The following file is in VCF:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/exon/snps/CEU.exon.2010_09.genotypes.vcf.gz

Please download the file and answer the following questions:

1. How many variant records does the file contain?

2. How many genotype calls are there per variant record?

3. Are the genotype calls *phased* or *unphased* (hint: look at section 1.4.2 of the spec)?

4. Write code or pseudo code (in any language of your choosing) to calculate *allele frequencies* for each variant in the file

5. Design a **relational database schema** to store the following information:

    - variant ID
    - chromosomal location of the variant
    - the alleles and their corresponding frequencies

6. Write code or pseudo code to populate your database schema from the VCF file

7. How might you store the genotypes such that they could be retrieved quickly, for a project that has produced genotypes for ~1200 individuals across ~80 million sites?