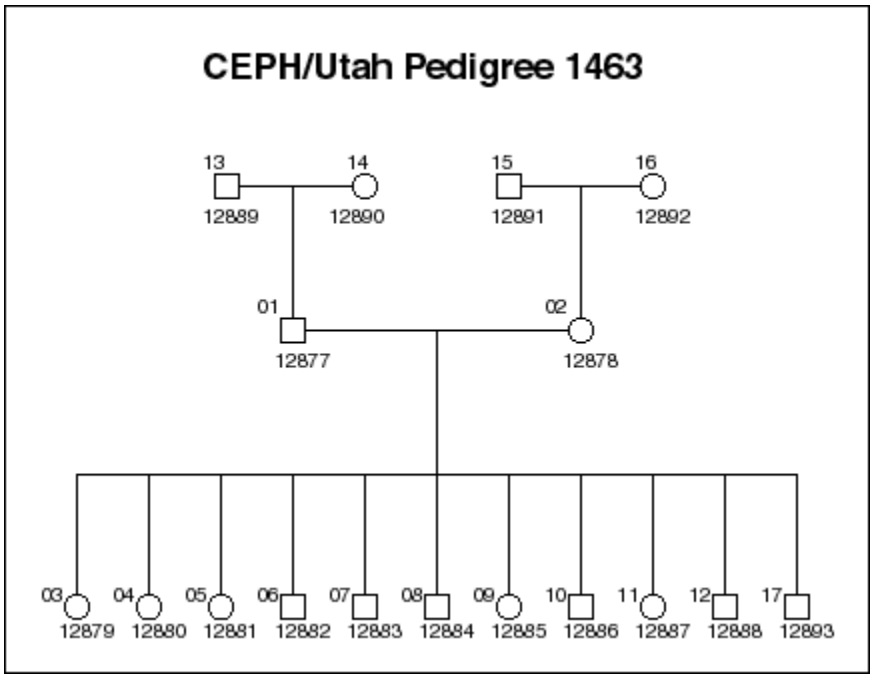
Due Date: Friday, March 22nd, 11:59 pm

In this assignment, you will profile genome variation information and attempt to answer biologically relevant questions. The **variant call format** (VCF) is a generic text file format for storing genome variation data such as SNVs, indels, and structural variants, together with rich annotations. It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information of samples for each position. A VCF file is usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome. For more information, please read the VCF guide available on the GATK website:
https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format

You've been provided with two VCF files where low quality variants and variants affecting the scaffolds and the mitochondrial genome have been filtered out. The VCFs contain:

1. Small genome variation (SNVs and indels) in `SNV_indel.biallelic.vcf` and
2. Structural variation (SV) in `sv.reclassed.filtered.vcf`. Variants have been classified as large deletions (DEL), tandem duplications (DUP), inversions (INV), mobile element insertions (MEIs) and "unknown" variants that are described by two or more "breakend" (BNDs) records (one for each breakpoint).

The variant calls are from a large family, one of whom (NA12878) has been sequenced and analyzed countless times and has been awarded the notorious title of "The Most well studied Genome in the World". The SNV/indel VCF file was generated using the Genome Analysis Toolkit (GATK) and the SV VCF file was generated using the Hall Lab pipeline (Lumpy & SpeedSeq).



DO NOT copy the VCF files to your assignment folder; instead, use a symbolic link.

```
ln -s source_file myfile
```

The pedigree is attached for your perusal. Pedigrees use a standardized set of symbols—squares represent males and circles represent females. These samples are found in the two VCF files provided with 'NA' in front of their 5-digit ID numbers. Individuals 15, 16, and 02 form a trio (dad, mom, and child) that you will be analyzing in-depth later.

For simplicity, this assignment is divided into two parts:

1. In **Part I**, you will profile the various types of genomic variation in the NA12878 individual, including SNVs, indels, and SVs.
2. In **Part II**, you will focus on small genome variations like SNVs and indels of a trio.

Part I

With the provided VCF files `SNV_indel.biallelic.vcf` and `sv.reclassed.filtered.vcf`

Step 1: Write a python script, `count_gv.py`, to count the different classes of genome variation (SNVs, small indels, as well as larger structural variants) **for the individual NA12878**. The script should read in both files and output a table with the names and numbers of **non-reference alleles** belonging to different variant classes, i.e. SNVs, indels, large deletions, tandem duplications, inversions, MEIs and BNDs. (For BND type variant, 2 BND calls represent one variant) [*You might want to read in the files line by line, to avoid disasters*.] Make sure to write doc string and usage. Exit the script if wrong number of parameters were given.

The usage of the script will be:

```
$ python3 count_gv.py <SNV_indel VCF> <SV VCF>
```

Question 1

What is your output from the above run?

Question 2

Using the numbers of variants reported in Question 1 for each variant class for NA12878, what proportion of genomic variants are SVs?

Question 3

Describe the spectrum of genome variation for the individual NA12878.

Step 2: Modify `count_gv.py` to plot the length distribution of small indels (output file: `histogram_indels.png`), large deletions (output file: `histogram_deletions.png`) and MEIs (output file: `histogram_meis.png`) observed in NA12878. Label your axes (include units), choose appropriate scaling, and give your plots descriptive titles. **Remember to place these plots in your submission folder!**

The usage of the script will continue to be:

```
$ python3 count_gv.py <SNV_indel VCF> <SV VCF>
```

Question 4

a) Describe the distributions.
b) Speculate how the length distribution might differ if we limit the data to exonic indels?

Part II

Homozygous SNVs are those where both copies of the chromosomes of the sample genome have the same allele which is different from the reference genome. If the two copies of a sample's chromosomes have different alleles and one matches the reference and one does not, this would an example of heterozygous SNV. Refer to Dr. Jin's lectures for more information on genotype nomenclature.

Step 1: Write a python script, `quantify_genotype.py` , to quantify the number of homozygous and heterozygous SNVs and indels, **for the individual NA12878**. Count the total number of genotype calls that are homozygous reference, heterozygous, and homozygous alternate for that individual. Also count the number of genotype calls with one or more alleles missing. Report it in the form of a table with the four possibilities and their corresponding counts. Remember to write doc string and exit script if wrong number of parameters were given.

The usage of the script will be:

```
$ python3 quantify_genotype.py <SNV_indel VCF>
```

Question 5

What is the output from running Part II's Step 1?

Question 6

Given the number of homozygous alternate (or non-reference homozygous) and heterozygous SNVs and indels you found, does the difference in the numbers make biological sense? Why, or why not?

You shall now focus your analysis on the NA12891-NA12892-NA12878 trio.

Step 2: Write a python script, `violate_MS.py` , to count the number of variants that clearly violate the rules of Mendelian segregation, given the trio's relationships to one another. (For simplicity, only consider autosome in this assignment, but in reality, autosomes and sex chromosomes are all important.) The usage of the script will be:

```
$ python3 quantify_genotype.py <SNV_indel VCF>
```

Question 7

How many variants clearly violate the rules of Mendelian segregation?

Question 8

Describe four potential reasons that could explain these violations.

The above analysis does not consider the quality of the genotype. Generally, it's best to filter variants based on genotype quality score to reduce the number of false positive variants. (Your answer for question 7 hopefully describes some of the justifications for this filtering step.) You will attempt to make your calls stringent by filtering based on genotype quality scores.

Each line, or record, of a VCF corresponds to one variant. Column 9 of the VCF gives the format for how the genotypes of each individual are reported in that record. For example, genotypes may be reported as a set of fields in the format `GT:AD:DP:GQ`. Other formats in other records may include other fields, but fields are always separated by a colon. The `GT` field refers to the genotypes called as 0/0, 0/1, or 1/1 for homozygous reference, heterozygous, or homozygous alternate. `AD` gives the depth at which each allele was called (one depth for homozygous calls, two depths for heterozygous calls). `DP` gives the total read depth, i.e. the sum of the `AD` values. Finally, `GQ` refers to genotype quality, which is reported on the Phred scale.

Unfortunately, **some individuals may be missing certain values that are given in the format column. Missing values may be represented by a period (e.g. '.' for a single value or './.' for missing alleles in the GT field)** or may not appear at all if they are at the right end of the format fields. However, if `GQ` is the fourth field in the format for a certain record, then you can be sure that any individual who has a `GQ` value will have it reported in the fourth field. (Note that `GQ` may not be the fourth field in the format for all records.)

Step 3: Modify `violate_MS.py` to filter variants such that you only keep records in which all three individuals in the trio have `GQ` scores at or above a given threshold. If any of the three are missing `GQ` values or if any value is below the threshold, that record should not be kept. **If the user does not provide a threshold, the script should not filter the variants.** (For simplicity again, only consider autosome) The usage of the script will now be:

```
$ python3 violate_MS.py <SNV_indel VCF> [GQ threshold*]
*default: no thresholding based on genotype quality score
```

Run your modified `violate_MS.py` on the trio data with a `GQ` threshold of 20.

Question 9

How many variants **now** violate mendelian segregation?

Please turn in:

A completed `README.txt`

Commented scripts:

- `count_gv.py`
- `quantify_genotype.py`
- `violate_MS.py`

Figures appropriately scaled with labelled axes and informative titles:

- `histogram_indels.png`
- `histogram_deletions.png`
- `histogram_meis.png`