Post Genomics – Fall 2025
Homework 2
Due: Thursday, September 11, 2025

Part 1 (Non-Programming)
Get familiar with VCF file: download example_VCF.vcf file on Bb and summarize information in it including but not limited to:

1.1 What are the three types of data line(s) found in a VCF file?

- **Meta-information Lines:** *These lines, which begin with a double pound sign (##), provide a description of the data within the file.*
- **Header Line:** *The header line, identified by a single pound sign (#), serves as a column header for the data that follows. It contains a set of eight mandatory, tab-separated fields:* **CHROM**, **POS**, **ID**, **REF**, **ALT**, **QUAL**, **FILTER**, *and* **INFO**.
- **Data Lines:** *Each data line represents a single genetic variant. The information on each line is organized into columns that correspond to the fields defined in the header line. These lines contain the actual variant calls, providing specific details such as the chromosome (***CHROM***), the position (***POS***), the reference allele (***REF***), and the alternate allele (***ALT***). The final columns on each data line, if present, contain genotype information for each sample, as defined by the* **FORMAT** *column.*

1.2 Date of creation?

```
##fileDate=20170506
```

So, the file was created in 2017/05/06

1.3 How many samples and variants are there in the file?

There are 2 samples and 751 variants in this VCF file.

1.4 Name the 11 columns found in this file, write a brief description (hint: refer to the header line).

1. #CHROM: Refers to the chromosome. [string (required), not white spaces allowed]
2. POS: Refers to the position. [integer (required)]
3. ID: Identification of type UUID. [string (required), not white spaces nor semicolon allowed.]
4. REF: Reference bases (A, G, T, C, N). [string (required)]
5. ALT: Alternate bases. [string (required)]
6. QUAL: Quality score if the given variant. [Numeric (required), a float/double number]
7. FILTER: Filter status, PASS indicating that the variant pass all the filters. [string (required), not white spaces nor semicolon allowed.]
8. INFO: Additional information. [string (required), not white spaces nor semicolon allowed.]
9. FORMAT: The format of the samples. [colon-separated alphanumerical string (required if samples exist)]
10. NORMAL: The "NORMAL" column corresponds to the first sample data. [Formatted according to "FORMAT"]
11. TUMOR: The "TUMOR" column corresponds to the second sample data. [Formatted according to "FORMAT"]

1.5 Which version of human genome reference is being used?

```
##reference=GRCh38.d1.vd1.fa
```

So, the used version of the human genome is: GRCh38.d1.vd1.fa

1.6 How many attributes are found in the CSQ section of the INFO column in this file, in your opinion, which 5 do you think would be most relevant/useful?

There are 66 attributes. From which the 5 more important are:

1. Allele: Which alternate allele it refers to.
2. Consequence: Describes the effect of a change.
3. Impact: Qualitative assessment of the severity of the change in the variant.
4. Symbol/Gene: The gene name/symbol/ID.
5. Protein change (or also HGVSp): Shows the specific amino acid substitution.

1.7 Given the FORMAT column example, GT:AD:BQ:DP:SS, for each colon separated entry, what is the data type and provide a brief description.

- **GT (Genotype):** Shows the genotype for the sample.
- **AD (Allelic Depths):** Provides the number of reads supporting the reference and alternate alleles.
- **BQ (Average Base Quality):** Gives the average quality score of the bases supporting the alleles.
- **DP (Read Depth):** The total number of reads at the position.
- **SS (Variant Status):** Indicates if the variant is germline, somatic, or wild-type relative to a normal sample.

1.8 For the first 5 variants, using their AD fractional values and a 0.1 cutoff, determine if they are kept or rejected for both the normal and tumor samples.

$$0.1 \leq \frac{AD1}{AD1+AD2} \text{ \&\& } 0.1 \leq \frac{AD2}{AD1+AD2}$$

Set M = $\frac{AD1}{AD1+AD2}$ and N = $\frac{AD2}{AD1+AD2}$

Now, using logical language:

$$0.1 \leq \ M \ \wedge \ 0.1 \leq \ N$$

But before using it, may be useful to better simplify to only one condition, for instance since M is the compliment of N, M+N will always be one, so we only need to test if M or N are in [0.1, 0.9], so:

$$M \in [0.1, 0.9]$$

Now, solving by hand:

Normal AD	Tumor AD	NORMAL AD1	NORMAL AD2	TUMOR AD1	TUMOR AD2	M normal	M tumor	FILTER
8,3	29,24	8	3	29	24	0.73	0.55	PASS
12,0	89,8	12	0	89	8	1.00	0.92	FAIL
12,0	7,1	12	0	7	1	1.00	0.88	FAIL
10,0	24,3	10	0	24	3	1.00	0.89	FAIL
15,0	93,9	15	0	93	9	1.00	0.91	FAIL

Part 2 (Python Programming)

Take the 5 VCF files from the folder on the bioinformatics server: /bioinfo/applications/course_ref/post_genomics/Fall_2023/HW2_VCF_Files/. Transform the VCF files into CSV files using the python script provided writeCSV_2023.py. The main structure of code is provided, but you need to fill some key blocks, these areas are highlighted in green commented lines. (Hint: Recall allele depth, AD, parameter practice in class on 9/2/2025)

Part 3 (Non-Programming)

Solve the Systems of equations.

Ax=b

x=A^{-1}b

Using the formula below solve the vector, x, using matrix multiplication. For 3.2, you may use https://matrix.reshish.com/inverse.php to calculate the inverse. All other work must be shown. To inverse a 2x2 matrix follow the steps below

$$A = \begin{vmatrix} a & b \\ c & d \end{vmatrix} \qquad A^{-1} = \frac{1}{a*d-b*c} \begin{vmatrix} d & -b \\ -c & a \end{vmatrix}$$

Part 3 (Non-Programming) cont'd.

3.1:

$-3x + 6y = 0$

$1x + 2y = 1$

$$A = \begin{bmatrix} -3 & 6 \\ 1 & 2 \end{bmatrix} \rightarrow A^{-1} = \frac{1}{(-3)(2) - 6 \times 1} \begin{bmatrix} 2 & -6 \\ -1 & -3 \end{bmatrix}$$

$$x_0 = x$$

$$x_1 = y$$

$$= \frac{1}{-12} \begin{bmatrix} 2 & -6 \\ -1 & -3 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} -\frac{1}{6} & \frac{1}{2} \\ \frac{1}{12} & \frac{1}{4} \end{bmatrix}$$

$$A\vec{x} = \vec{b}$$

$$\begin{bmatrix} -3 & 6 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\vec{x} = A^{-1}\vec{b}$$

$$\vec{x} = \begin{bmatrix} -\frac{1}{6} & \frac{1}{2} \\ \frac{1}{12} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{4} \end{bmatrix}$$

$$x = x_0 = \frac{1}{2} ; \quad x_1 = \frac{1}{4} = y$$

3.2:
x + 2y - z = 10
2x + y + 2z = 5
-x + 2y + z = 6

<u>Using RREF</u>

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix} \rightarrow A^{-1} = \text{RREF}[A|I]$$

$$A^{-1} = \begin{bmatrix} \frac{3}{16} & \frac{1}{4} & -\frac{5}{16} \\ \frac{1}{4} & 0 & \frac{1}{4} \\ -\frac{5}{16} & \frac{1}{4} & \frac{3}{16} \end{bmatrix}$$

$x_0 = x$
$x_1 = y$
$x_2 = z$

$$A\vec{x} = \vec{b}$$

$$\begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 5 \\ 6 \end{bmatrix}$$

$$\vec{x} = A^{-1} b$$

$$\frac{1}{16}\begin{bmatrix} 3 & 4 & -5 \\ 4 & 0 & 4 \\ -5 & 4 & 3 \end{bmatrix}\begin{bmatrix} 10 \\ 5 \\ 6 \end{bmatrix} = \frac{1}{16}\begin{bmatrix} 3\times10+4\times5-5\times6 \\ 4\times10+0\times5+4\times6 \\ -5\times10+5\times4+6\times3 \end{bmatrix}$$

$$\vec{x} = \frac{1}{16}\begin{bmatrix} 20 \\ 64 \\ -12 \end{bmatrix} = \begin{bmatrix} 5/4 \\ 4 \\ -3/4 \end{bmatrix}$$

$x_0 = x = 5/4$

$x_1 = y = 4$

$x_2 = z = -3/4$

What to Submit:

1) **A single PDF with Image of handwritten solutions to Part 1 and Part 3.**
2) **Python file with code in Part 2.**

*** Name the files (Your Last Name)_HW2.pdf and (Your Last Name)_HW2.py ***