Computational Genomics 2025
Homework 3: VarNet

PLEASE NOTE THE FOLLOWING INSTRUCTIONS:

*1. You are to complete this assignment alone. The assignment is open book, so you are allowed to use any books or information available online, your own notes and your previously constructed code, etc. HOWEVER YOU ARE NOT ALLOWED TO COMMUNICATE OR IN ANY
WAY ASK ANYONE FOR ASSISTANCE WITH THIS EXAM IN ANY FORM.*

2. Please pay attention to instructions and complete ALL requirements for ALL questions, e.g. some questions ask for code, plots, AND written answers.

3. A complete answer to this assignment will include a single pdf file named in this format: 'your_name'.compG_HW2.pdf. This file should include all of your commands in linux from installation to the end, your codes, and your resulting plots.

4. The exam must be ***sent to mehrmohamadi@ut.ac.ir by 11pm on the 5th of Teer, 1404.*** *It is your responsibility to make sure that it is received by then and no excuses will be accepted.*

In this project, we explore the application of VarNet, a deep learning model for detecting somatic variants, on targeted sequencing data from the HCC1395 breast cancer cell line, aiming to evaluate its performance on real tumor-normal samples.

Data description:

The HCC1395 AmpliSeq dataset is a high-depth targeted sequencing dataset generated from the HCC1395 breast cancer cell line and its matched normal sample. It is based on a custom AmpliSeq panel designed to capture 1368 genomic regions (~275-bp amplicons) surrounding 2477 carefully selected somatic and germline variants across all chromosomes, with balanced representation of variant allele frequencies and confidence levels. The target regions were defined using a BED file derived from the **Illumina Comprehensive Cancer Panel**, which covers key cancer-related genes and genomic intervals.

Download *AmpliSeq.bwa.HCC1395BL_1.bam* and *AmpliSeq.bwa.HCC1395_1.bam* from [here].

Model Description:

VarNet is a deep learning-based somatic variant caller that works on paired tumor-normal BAM files. It first filters genomic positions to identify candidate mutation sites with a higher likelihood of somatic variants, then encodes aligned reads at each candidate site into an image-like representation, capturing features such as base identity, base and mapping quality, strand bias, and reference base information. This image is then processed by a convolutional neural network (CNN) to classify whether the site contains a somatic variant.

Project Goal

VarNet, trained on a diverse set of whole-genome sequencing tumor-normal samples across seven cancer types, has demonstrated high performance in variant calling. To evaluate its generalizability, we aim to test VarNet on the HCC1395 breast cancer dataset (limit your analysis to **chromosome 1** to reduce processing time and computational load).

Instructions:

1. Extract chromosome 1 from both tumor and normal BAM files, as well as the reference genome, using *samtools*.
2. Run the provided notebook **VarNet_inference.ipynb** in Google Colab to perform variant calling. Before installing the required packages, replace the **requirements.txt** file inside the VarNet directory with the provided one.
3. The output VCF file will be saved under: **VarNet/output/<sample_name>/.**

Reporting Task

Select one variant from the resulting VCF file and visualize it in *IGV*. Include the following in your report:

- A screenshot of the aligned reads at the variant position.

- A brief interpretation of the variant—for example, discuss whether it lies within a gene, a coding or regulatory region, whether it affects protein production (e.g., missense or nonsense mutation), or if it is associated with disease based on databases such as dbSNP or COSMIC.

- Review the original paper to extract detailed information about the CNN architecture used in VarNet—for example, the number of convolutional layers, presence of activation functions such as ReLU, batch normalization, and pooling layers—and provide a concise summary of its components and their functions.