Homework #2 - Sequence Alignment and Variant Calling

Introduction

Patient-derived xenografts (PDXs) are powerful experimental models in which human tumors are implanted into immunodeficient mice. They provide a valuable resource for studying tumor biology and therapeutic responses in a setting that closely mimics the in vivo human tumor microenvironment. Importantly, sequencing data from PDXs are often made publicly available, enabling researchers and students to practice bioinformatics pipelines on real-world datasets.

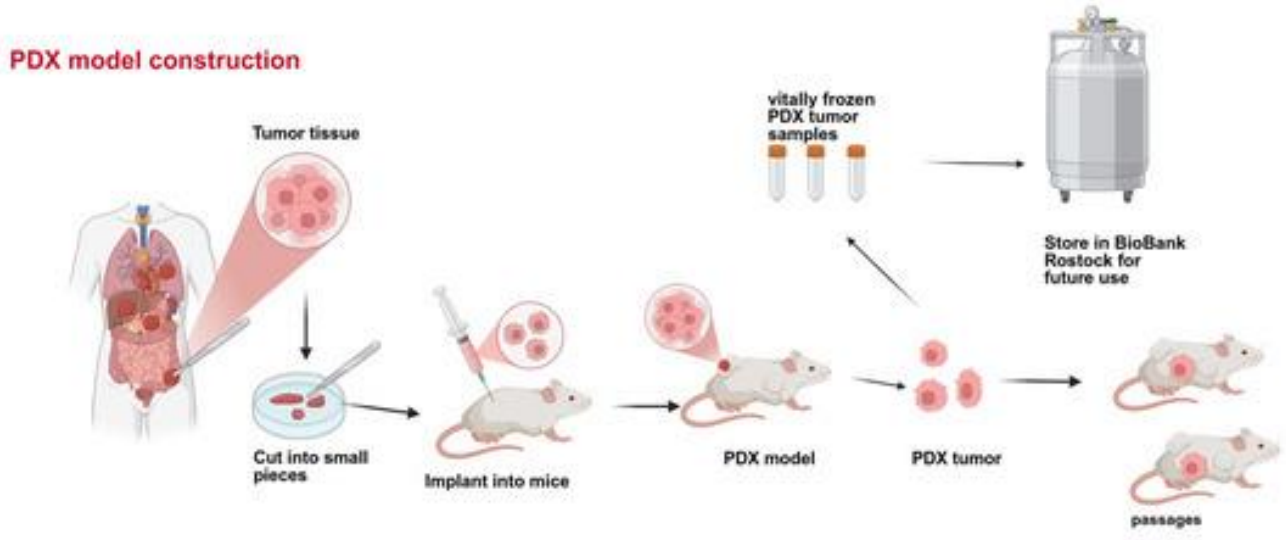

Fig.1- Construction and application of the PDX models. PDX models were constructed by transplanting patient-derived tumor tissue into immunodeficient mice. They offer the advantage of delivering expandable tumor tissue for a variety of subsequent research applications. (Adopted from El Hage, M.; Su, Z.; Linnebacher, M. Mutational Patterns in Colorectal Cancer: Do PDX Models Retain the Heterogeneity of the Original Tumor? *Int. J. Mol. Sci.* **2025**, *26*, 5111. https://doi.org/10.3390/ijms26115111)

When a tumor is first implanted into a mouse, this initial engraftment is called the **0th passage (P0)**. Each time the tumor is harvested from one mouse and re-implanted into another, the model goes through a new **passage (P1, P2, P3, …)**. Over successive passages, the tumor may accumulate new mutations, adapt to the mouse microenvironment, or undergo clonal selection. Because of this, early passages (like P0) are generally considered to be closer to the patient's original tumor, whereas later passages (like P2) may carry additional changes.

In this exercise, we will use sequencing data from two passages of the same PDX to simulate a basic tumor–normal analysis workflow:

- The **0th passage (P0)** will serve as the "normal" sample, representing the baseline state.
- The **2nd passage (P2)** will serve as the "tumor" sample, representing a more evolved state with potential new mutations.

For convenience purposes, the data was subsampled to contain only the 17th chromosome.

Prerequisites

A. Do all the pre-requirements necessary for tutorials 5-7.
B. Copy all the files from Stav's homework directory. There should be 6 files. Type:
```
cp /home/stavnaky/HW1_Input_Files_2025_2026/* $HOME
```

Alignment

1. **Align** the two fastq files of each passage to the fasta reference, to create 2 sam files (0th passage, 2nd passage) **(12 points)**
2. **Convert** the sam files to bam files. How much storage was saved? **(4 points)**
3. **Sort** the bam files. **(7 points)**
4. **Index** the sorted bam files. **(4 points)**

Variant Calling

5. **Make a pileup** of the two passages. **(10 points)**
6. **Call variants** to include only SNPs. **(12 points)**
7. **Filter** to include only germline variants (Think carefully! Use the VCF format fields) **(15 points)**
8. **Filter** to include only:
    a. Base Quality Bias Z-score (BQBZ) is 0
    b. Depth (DP) above 8 copies
    c. Mapping Quality (MQ) of at least 60
    d. Allele number (AN) is 4

    Use **statistics** to see if the filtration succeeded.

    Prove that the filtration succeeded by saving screenshots of the statistics.

(15 points)

9. **Loss Of Heterozygosity (LOH)** is when a cell goes from having two different gene copies to only one, and in cancer that often means losing the protective copy. It's a common hallmark of cancer (Demonstration in fig.2). **Filter the file**

from Q8 to contain only LOH events (normal is heterozygous, tumor is either alt-homozygous or ref-homozygous). **(13 points)**

10. What is the mutation at position 78392462 of chromosome 17? **(4 points)**
11. Using **bcftools stats-** what is the most common kind of substitution? **(4 points)**

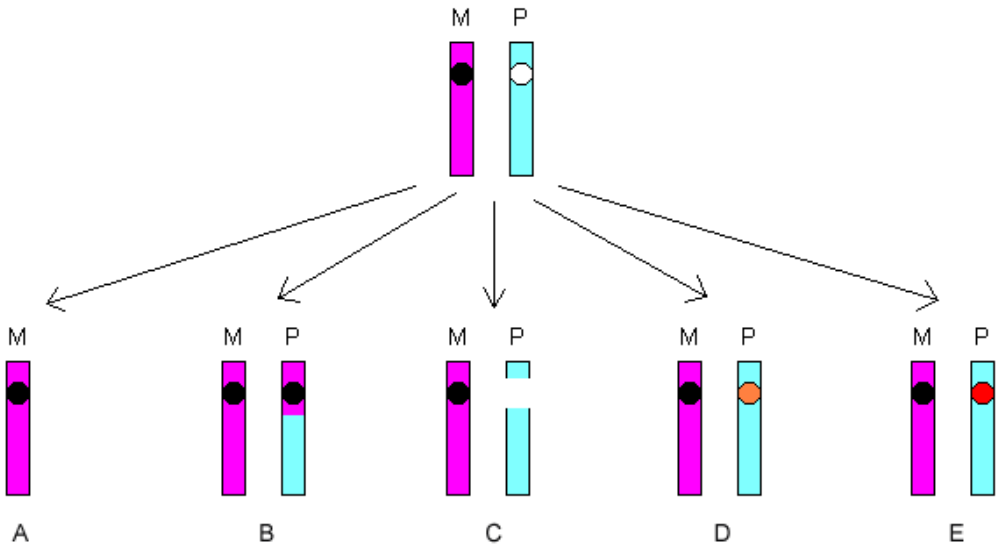

Fig. 2- Loss of heterozygosity means that the other allele of the gene is lost. Then only already mutated allele is left. M=maternal chromosome, P=paternal chromosome. White circle=normal allele, black circle=mutated allele, orange circle=point mutation in the allele, red circle=allele is inactivated due to the epigenetic reason. Arrows show different malicious genetic happenings. In A-C cases the loss of heterozygosity of the gene has happened and mechanisms are different. A=The other chromosome has been lost, B=The malicious recombination has happened and C=The deletion of the normal allele has happened. In D and E cases both alleles have been inactivated, but the heterozygosity remains. (Figure adapted from Wikipedia page of Loss Of Heterozygosity)