First Semester Report

Bernáth Szabolcs (bernathsz@student.elte.hu)
Doctoral School of Physics
Statistical Physics, Biological Physics
and Physics of Quantum Systems Program

Supervisors: Dr. Imre Derényi, Dr. Gergely J. Szöllősi
Thesis title: Understanding somatic evolution and development based on genome-scale sequence data

1. Introduction

As technology advances, we have more and more genomic data available. Processing these data allows us to quantify somatic mutations in cells. Somatic mutations are genetic changes characteristic of somatic cells that are not passed on to subsequent generations. The study of such mutations is fundamental and necessary in various fields, including cancer research.

If whole-genome sequencing (WGS) data originate from plant samples, specifically from different branches of trees (from leaf, fruit tissues), the extracted somatic mutations can be tracked. It can be stated that if the same mutation is found in multiple branches, it most likely originated before the divergence of those branches — unless, of course, we consider the rare possibility that the same mutation arose independently in two cells at the exact same site. New branches emerge from stem cells in the apical meristems of plants, but the mechanisms occurring within these stem cell populations are not yet universally described. However, tracking somatic mutations can provide insights into these fundamental processes.

2. Description of research carried out in the current semester

The apical meristem can be divided into three layers: L1, L2, and L3. These layers are also composed of plant stem cells; however, each layer has a different function and gives rise to different tissues. There has been no previous study investigating whether these layers mix. This question is particularly significant, because based on the current but limited literature, reproductive cells in plants originate from the L2 layer. If a somatic mutation occurs in an L2 stem cell, there is a possibility that the offspring will inherit it. In a recent study [1] Goel et al. sequenced the leaves, fruit flesh, and skin of peach trees. They found that significantly more somatic mutations accumulate in the skin of the fruit, which originates from the L1 layer, than in the flesh of the fruit, which derives from the L2 layer. If stem cells from the L1 population invade the L2 population in the meristem, mutations from L1 could also appear in the offspring. Indications of this phenomenon were found in [1], but due to strict filtering criteria in mutation analysis, no definitive conclusion was drawn.

To draw my own conclusions from the data, I had to start from the beginning to evaluate the raw data set from [1]. Thus, during my first semester of research, I learned how to handle whole genome sequencing (WGS) data. Raw sequencing data consist of millions of $\sim$150 base pair long sequence reads, varying in length depending on the sequencing method. The first step in processing is filtering out unreadable short reads (skewer), followed by aligning the remaining reads to the reference genome using alignment tools (bowtie2, minimap2). The aligned files are then indexed and filtered to remove duplicate reads (samtools), which may result from sequencing artifacts. The last step involves identifying mutations, known as mutation or variant calling (lofreq, bcftools, GATK HaplotypeCaller, GATK Mutect2). Each of the mentioned tools has various input parameters that influence the results. To determine the optimal parameters, I had to understand their underlying principles. At the end of the semester, I had successfully extracted somatic mutations from the dataset. The next step is to learn how to determine the accuracy of somatic mutations and how to distinguish true mutations from sequencing artifacts, especially in the case of low allele frequency mutations (allele frequency is the number of reads containing the mutation divided by the total number of reads at a given site).

3. Studies in the current semester

During the semester, I completed the following classes:

- Sensory biophysics (FIZ/3/010E)
- Preclinical models in cancer research (FIZ/3/082)

4. Teaching Activities

I was one of the instructors for the Modern Physics Laboratory course, teaching the ”biolfiz” measurements for 4×45 minutes per week.

References

[1] Goel, M., Campoy, J.A., Krause, K. et al. The vast majority of somatic mutations in plants are layer-specific. Genome Biol 25, 194 (2024).