Fourth Semester Report

Szabolcs Bernáth (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

Advances in whole-genome sequencing technologies have enabled the genome-wide detection of somatic mutations, providing new opportunities to investigate developmental processes in long-lived organisms. Previous studies have shown that long-lived trees accumulate surprisingly few somatic mutations. This suggests that these perennial trees have highly efficient mechanisms that preserve genome stability within stem cell populations.

In my PhD research, I started to reanalyze an existing dataset from a single apricot tree, in which different branches were sequenced separately, in order to identify high-confidence somatic mutations. These mutations serve as natural markers that can be used to reconstruct the developmental history of the tree. We are especially interested in mutations shared between branches: if the same mutation is detected in several branches, it most likely arose before those branches diverged.

By tracing these lineages, we aim to gain insight into the cellular processes occurring within the shoot apical meristem during branch formation. In addition, the observed somatic mutational patterns and frequencies may reveal stem cell dynamics that enable the observed genomic stability in these trees.

2. Description of research work carried out in the past three semesters

2024/25 fall semester

During the first semester of my PhD research, I became familiar with the basic bioinformatics methods that are essential for my project. I started by exploring the sequencing dataset, organizing the samples, formatting the files, and aligning the raw sequencing reads to the reference genome. I tested different alignment programs that perform either local or end-to-end alignment, and I learned about the advantages and disadvantages of each approach.

After the alignment step, I studied the information stored in the aligned files. I learned what base quality scores are and how they are assigned by the sequencing machine, what mapping quality means, and why it is not enough to rely only on this value. I also learned the necessary preprocessing steps for aligned files, including marking and removing duplicate reads, sorting and indexing the alignment files, and filtering low-quality alignments.

The next step was somatic mutation detection. There are many available software tools for this purpose, and different parameter settings can produce different results. I tested several mutation callers, but I found that in many cases the output was difficult to interpret and compare. For this reason, I decided to develop my own analysis pipeline that is specifically designed for my dataset and research question, with the aim of obtaining a reliable set of somatic mutations.

2024/25 spring semester

During the second semester, I worked on the extension of an existing mathematical model describing the accumulation of driver mutations in stem cell lineages. The original model considered only a single stem cell. We generalized it to describe an entire stem cell niche and also extended the model by allowing non-Poisson mutation rates, making it more flexible and biologically realistic.

We conducted stochastic simulations to generate stem cell lineage trees based on known data to model human colonic crypts. We then applied our mathematical framework on these lineages to investigate whether symmetric or asymmetric stem cell divisions provide an evolutionary advantage by reducing the probability of accumulating cancer-causing mutations. We found that symmetric divisions do not make a significant difference in the accumulation of driver mutations under biologically relevant mutation rates and probabilities.

From this project, we wrote a manuscript that is now almost complete and will be submitted for publication soon.

2025/26 fall semester

In the third semester, the main goal was to filter the raw variants detected by my previous highly permissive variant-calling pipelines, which resulted in a high number of false positives. We realized that in genomic regions where sequencing reads map to multiple locations, it is not possible to reliably estimate allele frequencies, which is essential for accurate mutation detection. For this reason, we focused on removing or filtering these problematic regions.

These issues are mainly caused by repetitive elements, duplicated regions, and structural differences in diploid genomes. To investigate this, I tested several existing annotation tools, but they did not provide a complete solution for separating these complex regions.

To address the problem, I developed a new method based on fragmenting both parental genomes into short artificial reads and mapping them back to the genomes. Based on how many times these fragments align, genomic regions can be classified as hemizygous, homozygous, or duplicated. This approach allowed us to effectively identify and filter unreliable regions.

3. Description of research work carried out in the fourth semester

In the fourth semester, I continued to work on the data analysis project to identify somatic mutations in an apricot tree.

First, I revisited and refined my original variant-calling pipeline. Inspecting the alignment files in IGV, we observed that the beginning and end of reads are not fully reliable, and we also excluded sites with low base quality. We therefore applied stricter variant-calling rules than before and only called variants in homozygous and hemizygous regions identified by our method. Even with these stricter rules, we obtained more than 600,000 variants in a 238 Gb genome. Most of these variants are likely false positives caused by sequencing errors and noisy or misalignments. Our goal was to build a set of independent filters that minimize the loss of true variants while removing as many false positives as possible.

First, we identified heterozygous variants, where the parental genomes do not match, using our simulated read and alignment approach. This step removed around 70% of the variants, but we later found that the reference genomes were not fully reliable. In several positions, the reads clearly supported heterozygosity, but this was not reflected in the reference. Therefore, we introduced an additional filter to remove the variants that are present in all samples.

Later, we observed that in highly repetitive regions, especially short tandem repeats, alignment errors occur frequently. We therefore filtered out these regions identified using Tandem Repeat Finder (TRF).

Next, we added filters to remove variants showing strand bias in paired-end reads, as well as variants biased towards low mapping quality.

We also observed that 4 out of 21 samples behaved strangely different from the others, so we separated these samples for further analysis.

At the end of the pipeline, we obtained around 150 candidate variants, which still include some false positives, so manual curation is still required. However, the preliminary results suggest that the examined tree does not accumulate low-frequency somatic mutations and the detected mutations are all present with high allele frequency. This is consistent with findings from a 238-year-old oak tree, where researchers identified only around 50 variants. Overall, these results support the hypothesis that bottleneck processes occur during branching, meaning that only a small number of stem cells contribute to the formation of new branches.

4. Publications

- The work carried out during the second semester resulted in a manuscript titled 'Symmetric Divisions Are Ineffective in Reducing the Accumulation of Driver Mutations in Stem Cell Niches', which is now approximately 99% ready for submission.
- A popular science article titled 'A fák hosszú életének titka' was published in *Fizikai Szemle*. This publication was accepted as 6 elective academic credits.

5. Conferences during the four semesters

- In the third semester, I attended the XXX. Congress in Biophysics organized by the Hungarian Biophysical Society in Szeged, where I gave a 15-minute oral presentation on the topic of modeling stem cell behavior in trees using mutational patterns.

6. Teaching activities during the four semesters

In every semester I was one of the instructors for the Modern Physics Laboratory course, teaching the "biolfiz" measurements in 4×45-minute sessions over a period of six-seven weeks.

7. Professional activities during the four semesters

Since April 2026, I have been co-supervising an MSc student.