2nd Progress Update on the Bioinformatics Master's Thesis Project

Author: Salomon Elieser Marquez Villalobos
Advisors: Fernando Pastor Rodriguez and Igor Ruiz de los Mozos
Tutor: Diego Garrido Martn
Date: December 2, 2025

1. Project status description

In the previous report, we mentioned that the focus of the Master's Thesis was on the evaluation of the RiNALMo model, an RNA LLM, to predict secondary (2D) structures associated with colorectal cancer. As an initial step, we contacted the authors of the RiNALMo model (GitHub issue #2) to identify the specific infrastructure required to run this model.

After reviewing the infrastructure of the University of Navarra's cluster, we determined that the GPU-equipped nodes did not offer the technical specifications needed to execute RiNALMo. Specifically, older GPU architectures such as the P100 or T4 typically used for proof-of-concept (PoC) tests on platforms like Colab or Kaggle, were not compatible with RiNALMo, since the model requires at least a GPU with 24 GB of memory and a more recent GPU architecture such as L4, A10, or A100.

Given the need for appropriate infrastructure to evaluate the RiNALMo model, we submitted an application to the EuroCC Spain Testbed program. This program offers "free access to high-performance computing (HPC) resources and specialized support within the EUROCC SPAIN TESTBED proof-of-concept (PoC) framework, aimed at companies, public administrations, and universities." As of today, the proposal is still pending approval.

These circumstances required us to redirect the scope of the project to consolidate a deliverable for the thesis. For this reason, over the past 4 weeks, we have focused on the analysis of differentially expressed genes (DESeq) and differential transcript usage (DTU) on the main dataset, corresponding to the RNA-seq study SRP479528, which includes 44 samples: 22 with colorectal cancer and 22 normal samples from patients over 50 years old. This was possible thanks to the earlier development of a functional Nextflow pipeline using nf-core/rnaseq and Salmon pseudo-alignment for processing and quantifying these samples.

Based on the DESeq and DTU analyses, we reoriented the central objective of the Thesis toward identifying differentially expressed genes, switch genes, and their corresponding isoforms from the 44 samples. The sequences identified in these analyses will serve as input for RNA 2D structure prediction models. As a proof of concept for this project, we will use the classical RNAstructure tool, reserving RiNALMo-based predictions for future collaboration after the thesis defense, once suitable computing infrastructure becomes available.

It is also important to note that we will extend the DESeq and DTU analyses to 42 additional samples: 21 with colorectal cancer and 21 normal samples from the RNA-seq study SRP357925, corresponding to patients under 50 years old. In total, we will be processing 86 samples. Expanding the dataset will allow us to widen the range of candidate isoforms with potential functional impact on colorectal cancer pathogenesis and meaningful biological relevance for predicting and studying their 2D structure.

2. Degree of achievement of objectives and expected results

Although we have not been able to begin evaluating RiNALMo due to infrastructure limitations, the project has progressed well in relation to the revised objectives. We have completed the necessary tasks to ensure a robust deliverable, including the transcriptomic analysis of the SRP479528 dataset, development of the functional nf-core/rnaseq pipeline, and generation of preliminary DESeq and DTU results.

3. Justification of changes (if applicable)

The lack of computational resources made us explore new horizons. Our EuroCC Testbed application is still under review, but cannot be considered within the immediate scope of the project. This prompted us to shift toward an analysis centered on expression and identification of relevant isoforms as the basis for 2D structural prediction using conventional methods.

4. List of completed activities

During this period, we completed the following activities:

- Configuration, execution, and validation of the nf-core/rnaseq pipeline for full processing of the SRP479528 and SRP357925 datasets.

- Initial analyses of differential expression (DESeq2) and differential transcript usage (DTU) for the SRP479528 dataset, along with the initial curation of results to identify genes, transcripts, and isoforms of interest.

- Use of RNAstructure for 2D prediction of RNA sequences and extraction of structural motif metrics using the bpRNA toolkit.

5. Planned activities in the work plan

In the coming weeks, we will focus on:

- Extending the DESeq and DTU analyses to the second dataset (SRP357925) to complete the analysis of all 86 samples and consolidate a definitive list of candidate genes and isoforms.
- Applying RNAstructure to predict the 2D structure of identified candidates.
- Writing the Thesis document.

6. Unplanned activities completed or scheduled

We incorporated several unplanned activities, including the formal application to the EuroCC program for access to advanced GPUs and adaptation of the analysis pipeline to handle a larger sample volume.

7. Schedule deviations and mitigation actions

The main deviation arose from the technical limitation preventing the execution of RiNALMo. We mitigated this by redirecting the project toward full transcriptomic analysis and structural prediction using accessible tools such as RNAstructure. This prevented delays that could compromise the Thesis defense.

8. Partial results obtained (include deliverables)

The master-bioinformatics GitHub repository includes the following deliverables:

- Execution reports of the nf-core/rnaseq pipeline for the SRP479528 and SRP357925 datasets.
- HTML reports of the DESeq analysis for SRP479528 and the DTU analyses for both datasets.