# Progress Update on the Bioinformatics Master's Thesis Project

## 1. Project Identification

**Provisional title of the project:**
"**Prediction of Secondary Structures of lncRNAs Associated with Colorectal Cancer Using an RNA Language Model**"

**Author:** Salomón Elieser Marquez Villalobos
**Supervisors:** Fernando Pastor Rodríguez and Igor Ruiz de los Mozos
**Academic Tutor:** Diego Garrido Martín
**Report date:** November 4, 2025

---

## 2. Description of Project Progress

Over the past four weeks, the project's direction has been consolidated around evaluating the ability of **RiNALMo**, an RNA-focused LLM, to predict **secondary (2D) structures** of **long non-coding RNAs (lncRNAs)** linked to colorectal cancer.

The initial literature review has been completed, and the primary dataset has been selected: **RNA-seq study SRP479528**, composed of **44 patient samples**—22 with colorectal cancer and 22 controls.

On the technical side, a functional **Nextflow pipeline** using **nf-core/rnaseq** with **Salmon** pseudo-alignment has been prepared for processing and quantifying the 44 samples, producing the corresponding count matrix. Feasibility of applying **transfer learning** with RiNALMo has also been evaluated, and the model authors have been consulted regarding computational requirements.

---

## 3. Degree of Completion of Objectives and Expected Results

The specific objectives of the TFM are being met, with some adjustments to the technical scope:

- **Literature review:** completed.
- **Dataset selection and problem definition:** completed.
- **RNA-seq pipeline implementation:** executed and successfully validated.

- **Preparation for structural inference with RiNALMo:** ongoing; currently limited by computing infrastructure.

- **Definition of evaluation metrics for RiNALMo compared to classical 2D structure prediction tools (e.g., RNAstructure):** theoretically complete; practical implementation pending.

Upcoming work will focus on **DESeq2 analysis at gene and transcript level**. The resulting transcripts will serve as input for RiNALMo and RNAstructure.

---

## 4. Justification of Changes (if applicable)

The main change concerns the scope of transfer-learning tasks (fine-tuning).
Given the size of RiNALMo (~650M parameters) and the lack of access to high-capacity GPUs (A10/A100), the plan is shifting toward **zero-shot or few-shot inference**, rather than full fine-tuning. This adjustment keeps the project feasible without undermining its core aim: assessing the model's predictive performance.

---

## 5. Completed Activities

- Review of the state of the art in 2D RNA structure prediction using foundational models. The project relies on the following key publications:

    - *Identification of differentially expressed genes and splicing events in early-onset colorectal cancer (https://doi.org/10.3389/fonc.2024.1365762)*

    - *RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks (https://doi.org/10.1038/s41467-025-60872-5)*

    - *Comprehensive benchmarking of large language models for RNA secondary structure prediction (https://doi.org/10.1093/bib/bbaf137)*

- Final definition of the TFM's scope.

- Selection and download of dataset **SRP479528** (RNA-seq).

- Implementation and testing of **nf-core/rnaseq** with Salmon on the University of Navarra HPC cluster.

- Generation of the count matrix.

- Contact with RiNALMo authors to resolve questions about technical requirements (see GitHub issue #2).

- Review of structural comparison metrics between RiNALMo and RNAstructure: base-pair accuracy, motif similarity, and structural similarity.

---

## 6. Planned Activities

- Perform structural inference of lncRNAs using RiNALMo.
- Run equivalent predictions with classical tools such as RNAstructure.
- Compute structural comparison metrics.
- Document results and prepare figures and tables for the final report.
- If time permits, develop an **R/Shiny** application to visualize results.

---

## 7. Unplanned Activities Completed or Scheduled

Execution of a **DESeq2 differential expression analysis**—at gene and transcript level—on dataset SRP479528, following best practices taught in **Advanced Bioinformatics** (UPNA, Igor Ruiz de los Mozos) and the course **Omics Data Analysis (M0-157)** (UOC, Diego Garrido Martín).

---

## 8. Timeline Deviations and Mitigation Actions

There has been a delay in the RiNALMo stage due to computational infrastructure limitations. As a mitigation measure, **zero-shot/few-shot inference** will be used for 2D RNA structure prediction.

---

## 9. Partial Results Obtained (Attached Deliverables)

A fully functional and validated **nf-core/rnaseq** pipeline running on an HPC cluster, including per-sample count matrices, is available in the **master-bioinformatics** GitHub repository.