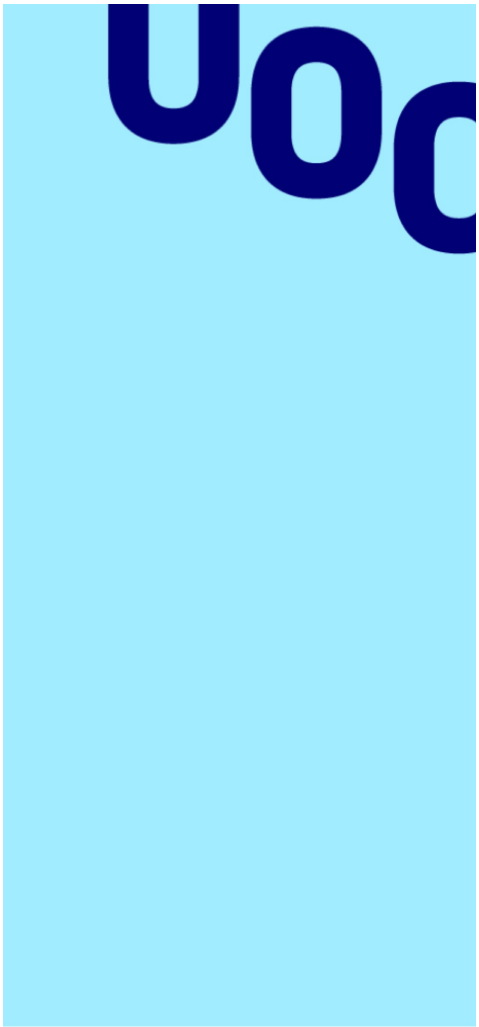
Functional Characterization of lncRNA Biomarkers in Colorectal Cancer Using RiNALMo



Salomon Elieser Marquez Villalobos

M0.207 Omics Data Analysis

Master in Bioinformatics and Bioestadistics

Name of the tutor:
Fernando Pastor and Igor Ruiz

Name of the SRP:
Diego Garrido Martin

October 8, 2025

Universitat Oberta de Catalunya

Final Work Card

Title of the work:	Functional Characterization of lncRNA Biomarkers in Colorectal Cancer Using RiNALMo
Name of the author:	Salomon Elieser Marquez Villalobos
Name of the tutor:	Fernando Pastor and Igor Ruiz
Name of the SRP:	Diego Garrido Martin
Date of delivery:	October 8, 2025
Studies or Prgram:	Master in Bioinformatics and Bioestadistics
Area or the Final Work:	M0.207 Omics Data Analysis
Language of the work:	English
Keywords:	CRC, biomarkers, AI

Abstract

Chapter 1

Introduction

1.1 Context and justification of the project

Colorectal cancer (CRC) is caused by a sequence of somatic genomic alterations affecting driver genes in core cancer pathways [1]. CRC is the third most common malignancy and a leading cause of cancer-related death globally, with a rising incidence in younger individuals, early-onset colorectal cancer (EOCRC <50 years old) patients versus later-onset colorectal cancer (LOCRC >50 years old) patients [2]. The development of CRC is driven by the accumulation of genomic alterations in key oncogenic and tumor-suppressive pathways, including WNT, RAS-MAPK, PI3K, and $TGF-\beta$ [3]. While large-scale sequencing efforts have successfully identified numerous protein-coding driver genes, the functional role of the non-coding genome remains a critical and largely unexplored frontier in CRC biology.

Long non-coding RNAs (lncRNAs) are regulatory RNA molecules longer than 200 nucleotides that are not translated into proteins but exert their function as RNA. They are key regulators of gene expression, participating in chromatin architecture, transcription, and splicing. Their highly cell- and tissue-specific expression patterns make them excellent candidates for cancer biomarkers and therapeutic targets. However, some traditional bioinformatics pipelines for biomarker discovery have biased against lncRNAs. For instance, methods that rely on protein-protein interaction (PPI) networks, such as the one used by [4], exclude non-coding genes from their analysis.

This limitation is being overcome by comprehensive whole-genome sequencing (WGS) studies, which analyze the entire genome without a bias toward coding regions. Recent landmark studies by [5] and [3] have cataloged thousands of genomic alterations in CRC to identify novel drivers. These efforts have successfully identified non-coding candidates with high-confidence, *involving the lincRNA LINC00673 (also known as LINC00511), a transcript that interacts with the CRC driver genes EZH2 and PTPN11* [5], highlighting them as promising but functionally uncharacterized players in CRC.

The critical gap in the field is no longer just the discovery of these lncRNAs, but the understanding of their molecular function. This project is justified by the opportunity to address this gap directly by applying a state-of-the-art computational tool. We will leverage RiNALMo (RiboNucleic Acid Language Model), a new 650M-parameter language model pre-trained on 36 million non-coding RNA sequences [6]. RiNALMo has achieved state-of-the-art results on downstream tasks like secondary structure prediction, non-coding RNA (ncRNA)

family classification, splice-site prediction, mean ribosome loading (MRL), translation efficiency (TE), and expression level (EL) prediction.

1.2 Work objectives

1.2.1 General objective

The primary objective of this thesis is to characterize the structure and function of lncRNA candidates previously identified in colorectal cancer, using the state-of-the-art RiNALMo language model.

1.2.2 Specific objectives

1. To perform a literature review to select and curate a list of lncRNA candidates with strong evidence of association with CRC (e.g., LINC00673, PWRN1 from [5]).
2. To establish a robust computational environment on a high-performance computing (HPC) cluster or a cloud-based platform for application of the RiNALMo model.
3. To apply RiNALMo model to the prioritized CRC-associated lncRNA candidates to generate predictions about their secondary structure, ncRNA family, and splice-site.
4. To analyze and interpret the predicted structures to formulate functional hypotheses regarding their molecular mechanisms in CRC.

1.3 Focus and method followed

Given the reduced 3-month timeframe, this project will focus on the functional characterization of known candidates using RiNALMo.

- **Phase 1: Candidate Selection and Pipeline Preparation (October)**. This phase will focus on rapidly establishing the project's foundation.
    1. Candidate Selection: High-confidence lncRNA candidates (LINC00673, PWRN1) will be selected based on their identification as putative non-coding drivers in the [5] study. Their sequences will be retrieved from genomic databases (e.g., Ensembl).
    2. Computational Environment Setup: Access to a suitable HPC cluster (e.g., at University of Navarra) or a cloud platform (e.g., Modal) will be secured. A reproducible software environment will be built using Conda or Singularity containers, including Python, PyTorch, and specialized libraries like FlashAttention-2. The lightweight version of RiNALMo (35M params) codebase and pre-trained weights will be downloaded from RiNALMo GitHub Repository and installed, as specified by the authors.
    3. Pipeline Validation: The quickstart guide of the lightweighted version of RiNALMo will be run for inference of a few RNA sequences to validate the entire computational pipeline.

- **Phase 2: Downstream tasks and Functional Prediction (November)** This phase will execute the core computational work of the thesis.

    1. Secondary Structure Prediction Task: The 35M-parameter RiNALMo model will be executed on the HPC platform to predict the secondary structures of the selected CRC-associated lncRNAs.
    2. Analysis and Interpretation: The predicted structures will be visualized and analyzed for conserved or notable motifs (e.g., hairpin loops, bulges) that could serve as binding sites for proteins or other RNAs, providing mechanistic hypotheses for their function.
    3. Complete Downstream Tasks: The other downstream tasks will be implemented prioritizing the structure, ncRNA family and splice site prediction ones.
    4. Increment Model Size: RiNALMo comprises the following sizes: 35M params (micro), 150M params (mega), and 650M params (giga). For the mega and micro sizes, a Tesla P100 node with 12 GB VRAM is required whereas for the giga model, a nice-to-have option would be an L40S node with 48 GB VRAM. Thus, if the UNAV cluster permits, other versions of the model will be explored.

1.4 Working Plan

1.4.1 Tasks

1. Project Setup (Week 1-2: Oct 1-15): Finalize literature review, confirm lncRNA candidates, acquire all necessary data, and apply for HPC access.

2. Environment and Pipeline Validation (Week 3-4: Oct 16-31): Set up the complete computational environment on the UNAV HPC cluster. Validate that the lightweighted version of RiNALMo is up and running.

3. Secondary Structure Prediction (Week 5-6: Nov 1-15): Apply the model to generate secondary structure predictions for the selected CRC lncRNAs.

4. Implement Other Downstream Tasks (Week 7-8: Nov 16-30): Continue the analysis, visualization, and interpretation of results with other downstream tasks and if possible scale up the RiNALMo model size to compare tasks' performance.

5. Thesis Writing and Finalization (Week 9-11: Dec 1-15): Complete the analysis, discuss relevant results and write the full thesis manuscript for submission to the advisors.

1.4.2 Milestones

1. M1 (by October 31): Computational pipeline fully validated on a small scale on the UNAV HPC cluster.

2. M2 (by November 30): All primary computational results, including RiNALMo downstream tasks for lncRNA predictions, have been generated and analyzed.

3. M3 (by December 15): Submission of the final thesis manuscript draft to advisors.

1.4.3 Risk Analysis

Table 1.1: Risk Assessment and Mitigation Strategies

Risk	Mitigation Strategy
1. Time Constraints: The 3-month timeline leaves little room for unexpected delays.	The project scope has been narrowed to focus solely on the functional characterization of lncRNAs, eliminating the discovery phase and reusing transcripts already associated with CRC.
2. Delays in HPC Access or Execution: Technical issues on the HPC could delay the execution of the model.	Development and pipeline validation will be done on the UNAV HPC cluster. Initially, the secondary structure task will be submitted and monitored closely to resolve any issues promptly.
3. RiNALMo Model/Code Issues: The public release of the model and scripts may have bugs, be delayed, or be difficult to use.	Although the [6] provides the code and weights of the RiNALMo Model, a backup plan will be prepared to characterize the same lncRNAs using established, less novel tools (e.g., RNAstructure, CONTRAfold) to ensure a completed project.
4. Uninformative Results: The predicted structures may not reveal obvious functional motifs.	The primary contribution of this thesis is the novel application of a state-of-the-art methodology to a critical biological problem. The thesis will be framed around the workflow itself, with the interpretation of results forming the discussion, even if they are not immediately conclusive.

1.5 Brief summary of products obtained

This thesis will deliver a focused, high-impact contribution by being one of the first studies to apply the RiNALMo language model for CRC.

- A complete Master's thesis manuscript detailing the workflow, results, and interpretation.
- Generation of the first predicted secondary structures for high-confidence CRC-associated lncRNAs like LINC00673 and PWRN1, providing a foundation for future experimental validation.

- Documentation of a rapid and powerful workflow for applying large RNA language models to functionally characterize newly discovered non-coding genes from large genomic studies.
- A practical demonstration of the feasibility and utility of a foundational RNA model to gain mechanistic insights into cancer biology, moving beyond expression-based biomarker discovery.
- All developed scripts and analysis notebooks will be organized and made available in a public repository (e.g., GitHub) to ensure reproducibility.

Bibliography

[1] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127):1546–1558, March 2013.

[2] Olivia M. Marx, Marc M. Mankarious, Walter A. Koltun, and Gregory S. Yochum. Identification of differentially expressed genes and splicing events in early-onset colorectal cancer. *Frontiers in Oncology*, 14, April 2024.

[3] Luís Nunes, Fuqiang Li, Meizhen Wu, Tian Luo, Klara Hammarström, Emma Torell, Ingrid Ljuslinder, Artur Mezheyeuski, Per-Henrik Edqvist, Anna Löfgren-Burström, Carl Zingmark, Sofia Edin, Chatarina Larsson, Lucy Mathot, Erik Osterman, Emerik Osterlund, Viktor Ljungström, Inês Neves, Nicole Yacoub, Unnur Gudnadóttir, Helgi Birgisson, Malin Enblad, Fredrik Ponten, Richard Palmqvist, Xun Xu, Mathias Uhlén, Kui Wu, Bengt Glimelius, Cong Lin, and Tobias Sjöblom. Prognostic genome and transcriptome signatures in colorectal cancers. *Nature*, 633(8028):137–146, September 2024.

[4] Ayoub Vaziri-Moghadam and Mohammad-Hadi Foroughmand-Araabi. Integrating machine learning and bioinformatics approaches for identifying novel diagnostic gene biomarkers in colorectal cancer. *Scientific Reports*, 14(1):24786, October 2024.

[5] Alex J. Cornish, Andreas J. Gruber, Ben Kinnersley, Daniel Chubb, Anna Frangou, Giulio Caravagna, Boris Noyvert, Eszter Lakatos, Henry M. Wood, Steve Thorn, Richard Culliford, Claudia Arnedo-Pac, Jacob Househam, William Cross, Amit Sud, Philip Law, Maire Ni Leathlobhair, Aliah Hawari, Connor Woolley, Kitty Sherwood, Nathalie Feeley, Güler Gül, Juan Fernandez-Tajes, Luis Zapata, Ludmil B. Alexandrov, Nirupa Murugaesu, Alona Sosinsky, Jonathan Mitchell, Nuria Lopez-Bigas, Philip Quirke, David N. Church, Ian P. M. Tomlinson, Andrea Sottoriva, Trevor A. Graham, David C. Wedge, and Richard S. Houlston. The genomic landscape of 2,023 colorectal cancers. *Nature*, 633(8028):127–136, September 2024.

[6] Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić. RiNALMo: General-purpose RNA language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1):5671, July 2025.