Universidad de Navarra
Facultad de Ciencias

Propuesta de Trabajo Fin de Máster
Año académico 2026-2027
MÁSTER EN CIENCIA DE DATOS PARA CIENCIAS EXPERIMENTALES

Proyecto Nº 35

Título: Decoding RNA dynamics and structure in colon cancer through integrative high-throughput transcriptomics

Departamento/ Laboratorio: Dirección de Medicina Personalizada y Laboratorio, NASERTIC - Polo de Innovación IRIS (TFM en empresa) & Laboratorio de Aptámeros,
Programa Innovación Terapéutica CIMA/UNAV

Director: Igor Ruiz de los Mozos
Correo electrónico: igor.ruiz@unav.es
Codirector: Fernando Pastor
Correo electrónico: fpasrodri@unav.es

Resumen:

RNA regulation cannot be understood from steady-state gene expression alone. Within this TFM project, the student will analyse real high-throughput datasets that capture complementary layers of RNA biology: nascent transcription, molecular immunoprecipitation/pull-down enrichment and RNA structure measured with PARIS2. The aim is to discover regulatory patterns, candidate biomarkers and structure-associated features relevant to colon cancer and personalised medicine.
The student will build an end-to-end, reproducible analysis from raw reads to biological interpretation: quality control, preprocessing, alignment and quantification; statistical modelling of transcription and enrichment; reconstruction of RNA-RNA duplexes and structural contacts; and integration of the resulting data layers. Multivariate statistics, dimensionality reduction, clustering, network analysis and machine learning will be used to extract interpretable biological signals from high-dimensional data. Depending on the dataset, methods may include feature selection, regularised models, tree-based approaches and supervised or unsupervised learning.
This is not a pipeline-running exercise. The student will compare methods, quantify technical biases, make defensible analytical decisions and interpret results with experimental researchers. Models will be evaluated with rigorous cross-validation and attention to interpretability and data leakage. As an optional advanced extension, deep-learning approaches may be explored for RNA sequence/structure representation or multimodal data integration when scientifically justified. The project therefore combines data science, AI, transcriptomics, structural RNA biology and precision medicine.

Objetivos específicos:

1. Build and benchmark reproducible quality-control and preprocessing workflows for the different sequencing modalities, including adapter/UMI handling where required.
2. Quantify transcriptional dynamics and molecular enrichment with statistically sound models, accounting for replicates, batch effects and experimental design.
3. Reconstruct RNA-RNA interactions from PARIS2 chimeric reads, generating interpretable duplex groups, contact maps and interaction networks.
4. Integrate complementary RNA layers using multivariate analysis, networks and machine learning to identify coherent regulatory patterns, candidate biomarkers and structure-associated features.
5. Develop and critically evaluate interpretable predictive/classification models where appropriate, using robust feature selection, cross-validation and performance assessment; deliver a reusable workflow and publication-quality visualisations.

Universidad de Navarra

Facultad de Ciencias

Perfil del alumno:
Students of the Master in Data Science for Experimental Sciences interested in bioinformatics, machine learning, RNA biology, cancer or precision medicine. The project is ideal for a student who enjoys combining quantitative reasoning, programming and biological interpretation and wants to work with real research-grade omics data.

Datos y recursos disponibles:
• Real NGS datasets: raw and processed nascent RNA, molecular immunoprecipitation/pull-down and RNA-structure sequencing, with complementary transcriptomics where useful.
• Access to NASERTIC's high-performance computing infrastructure - the largest HPC environment in Navarra - for large-scale sequencing analysis, machine learning and data integration.
• Direct integration in NASERTIC's Personalised Medicine and Laboratory team. NASERTIC is one of the three public high-capacity sequencing hubs supporting personalised medicine in Spain, offering exposure to real large-scale genomics workflows and an active multidisciplinary research environment.

Qué aprenderá el estudiante:
• End-to-end analysis of high-throughput transcriptomics, from FASTQ files to biological conclusions.
• Experimental design, statistical modelling and critical assessment of technical bias.
• Specialised analysis of nascent RNA, enrichment-based assays and RNA-RNA structural interactions.
• Reproducible bioinformatics in R/Python, Bash/Linux, Git and workflow-oriented analysis on HPC.
• Machine learning for biological discovery: feature engineering/selection, dimensionality reduction, supervised and unsupervised learning, validation and model interpretation; optional deep learning when justified.

OPTATIVAS RECOMENDADAS

1. Análisis e interpretación de datos de alto rendimiento
2. Análisis de secuencias y bioinformática estructural
3. Deep Learning
4. Advanced Topics in Machine Learning