UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE CIENCIAS EXACTAS

DEPARTAMENTO DE FÍSICA

Trabajo de Tesis Doctoral:

Expresión génica en cáncer de mama: análisis cuantitativo de células individuales y modelos de redes regulatorias

---

Tesista: Daniela Senra

Directora: Nara Guisoni

Director: Luis Diambra

Año: 2025

Lic. Daniela Senra
Expresión génica en cáncer de mama: análisis cuantitativo de células individuales y modelos de redes regulatorias
Tesis doctoral, mayo 2025
Dirección: Dr. Luis Diambra y Dra. Nara Guisoni

Doctorado de la Facultad de Ciencias Exactas, Área Física
Departamento de Física
Facultad de Ciencias Exactas
Universidad Nacional de La Plata

Resumen

En la presente tesis doctoral, se explora la complejidad de la expresión génica en el cáncer de mama mediante un enfoque interdisciplinario que integra el análisis de datos de secuenciación de ARN de células individuales (scRNA-seq) y el modelado matemático de dos redes de regulación génica relevantes en cáncer.

El cáncer de mama es una de las principales causas de mortalidad por cáncer en Argentina y en el mundo, caracterizándose por la heterogeneidad tanto entre pacientes como dentro de un mismo tumor. Esta heterogeneidad puede manifestarse a distintos niveles, incluyendo variaciones a nivel genético, epigenético, transcriptómico, proteómico y en el microambiente tumoral. Esta complejidad contribuye significativamente a la capacidad de las células tumorales para resistir a las terapias, evadir la respuesta inmune y adaptarse a cambios en su entorno.

En la primera parte de la tesis, se estudia la composición celular de muestras de mama sana en *H. sapiens* y se infiere la trayectoria de diferenciación desde las células madre mamarias hasta las células diferenciadas. Para ello, se desarrolla una metodología que utiliza la red de interacción proteína-proteína asociada a la diferenciación celular, con el fin de calcular un índice que cuantifica la pluripotencia a partir de datos de transcriptoma de célula única (scRNA-seq). Luego se generaliza esta herramienta para calcular la actividad de redes asociadas a diversos procesos biológicos de relevancia en cáncer y se aplica esta metodología a muestras de cáncer de mama. Además, se definen otros parámetros para cuantificar la cantidad de mutaciones, la entropía y la heterogeneidad transcriptómica. Se analiza también la variabilidad y la correlación entre estas cantidades en relación al subtipo de tumor (ER+, HER2+ y cáncer de mama triple negativo) y su agresividad.

La segunda parte de esta tesis se enfoca en dos procesos biológicos centrales en la progresión del cáncer: la transición epitelio-mesénquima (EMT), un fenómeno vinculado con la capacidad de invasión y migración en cáncer, y la pluripotencia, asociada a la capacidad de regeneración y proliferación tumoral. Se emplean modelos matemáticos basados en ecuaciones diferenciales para simular la dinámica de las redes de regulación génica de la pluripotencia y de la EMT. Estos modelos permiten analizar cómo los mecanismos de regulación, mediados por factores de transcripción y microARNs, definen el estado celular. Además, se modela el acoplamiento de ambas redes, basándose en evidencia experimental y en predicciones bioinformáticas sobre la interacción entre ellas. Se evalúa el efecto de la integración en una única red, en particular los cambios en la reversibilidad de la EMT al modificar la influencia de la red de pluripotencia sobre la de EMT.

Los resultados obtenidos aportan una comprensión más profunda de los mecanismos de regulación en el cáncer de mama, enfocándose en subpoblaciones celulares clave que se consideran responsables de la resistencia a las terapias y de la recaída después del tratamiento.

Abstract

This PhD thesis addresses the complexity of gene expression in breast cancer through an interdisciplinary approach that combines single-cell RNA sequencing (scRNA-seq) data analysis with mathematical modeling of two gene regulatory networks relevant to cancer biology.

Breast cancer is one of the leading causes of cancer-related mortality in Argentina and worldwide. It is characterized by heterogeneity both between patients and within a single tumor. This heterogeneity can manifest at various levels, including genetic, epigenetic, transcriptomic, proteomic, and variations in the tumor microenvironment. Such complexity significantly contributes to the ability of tumor cells to resist therapy, evade immune responses, and adapt to changes in their environment.

In the first part of the thesis, the cellular composition of healthy breast tissue samples in *H. sapiens* is studied, and the differentiation trajectory from mammary stem cells to differentiated cells is inferred. To achieve this, a methodology is developed that utilizes the protein–protein interaction network associated with cell differentiation to calculate an index that quantifies pluripotency from single-cell transcriptomic data (scRNA-seq). This tool is then generalized to measure the activity of networks associated with various biological processes relevant to cancer and is applied to breast cancer samples. Additionally, other parameters are defined to quantify the number of mutations, entropy, and transcriptomic heterogeneity. The variability and correlation among these measures are analyzed in relation to tumor subtype (ER+, HER2+, and triple-negative breast cancer) and aggressiveness.

The second part of this thesis focuses on two central biological processes in cancer progression: the epithelial–mesenchymal transition (EMT), which is linked to the invasive and migratory capabilities of cancer cells, and pluripotency, which is associated with tumor regeneration and proliferative capacity. Mathematical models based on differential equations are employed to simulate the dynamics of the gene regulatory networks governing pluripotency and EMT. These models allow for the analysis of how regulatory mechanisms mediated by transcription factors and microRNAs determine the cell states. Additionally, the coupling of both networks is modeled, based on experimental evidence and bioinformatic predictions regarding their interaction. The effect of integrating them into a single network is evaluated, particularly focusing on changes in EMT reversibility when modifying the influence of the pluripotency network on the EMT network.

The results provide deeper insights into regulatory mechanisms in breast cancer, with a focus on key cellular subpopulations considered responsible for therapy resistance and post-treatment relapse.

Agradecimientos

A mis directores, Nara y Luis, por su paciencia, guía y calidez humana a lo largo de todo el desarrollo de esta tesis.

A mis colegas del INIFTA, Marisa, Paula, Liorén, Gustavo, Raúl y Claudio. A mis compañeros del CREG, Lucila, Natalia, Ivana, Carla, Gonzalo, Micaela, Rocío, Mariano, Carolina, Nicolás y Sheila. Gracias por el compañerismo, las charlas en los pasillos, los almuerzos y hacer que la rutina sea divertida.

A mis amistades, Julieta, Thiago, Iriel, Inés, Malena, Natalí, Benjamín, Lorenzo, Camila, Gloria, Nahuel, Lucas y especialmente a Mara.

A mi psicóloga Ivonne, por su enorme ayuda este último año.

A mi familia, a mi mamá y papá que siempre me alientan y apoyan en todas mis decisiones. A mi hermana y a mi hermano, que, aunque sean menores, aprendo de ellos todos los días.

A mis familiares y amistades de Mendoza, en especial a Ivana, que siempre me reciben con los brazos abiertos.

A mi compañero de vida Martín y a nuestra gata, mi hogar y mi lugar de paz en el mundo.

A la Universidad Nacional de La Plata, por la educación pública y de calidad que me brindó desde la escuela secundaria, y a todos los docentes que encontré en el camino, especialmente a Martita, que sembró en mí la semilla de la curiosidad por la ciencia.

A CONICET por darme la posibilidad de hacer esta tesis. A todos los organismos que aportaron financiamiento para el desarrollo de la tesis y actividades científicas, AGENCIA, Open Bioinformatics Foundation, Human Cell Atlas, Asociación de Física Argentina, TREFEMAC, International Centre for Theoretical Physics y CABANA.

Y a todas las personas que aportaron su granito de arena en este proceso.

Gracias

Índice general

Glosario

ARN Ácido ribonucleico.

ARNm ARN mensajero, ácido nucleico monocatenario obtenido después de la transcripción que actúa como plantilla para la síntesis de las proteínas.

CBS Por sus siglas en inglés *Cascading Bistable Switches*, es un modelo de red de regulación génica de la EMT.

CSC Célula madre cancerosa, por sus siglas en inglés *Cancer Stem Cell*.

E/M Estado celular híbrido con características epiteliales y mesenquimales.

EMT Transición epitelio-mesénquima, por sus siglas en inglés *Epithelial-Mesenchymal Transition*.

GRN Red de regulación génica, por sus siglas en inglés *Gene Regulatory Network*.

MET Transición mesénquima-epitelial, por sus siglas en inglés *Mesenchymal-Epithelial Transition*.

miARN También conocidos como microARNs, son pequeños ARNs no codificantes que regulan la expresión génica.

PCA Análisis de componentes principales, por sus siglas en inglés *Principal Component Analysis*. Técnica de reducción de dimensionalidad.

PPIN Red de interacción proteína-proteína, por sus siglas en inglés *Protein-Protein Interaction Network*.

scRNA-seq Secuenciación de ARN de células individuales, por sus siglas en inglés *single-cell RNA sequencing*.

TCS Por sus siglas en inglés *Ternary Chimera Switch*, es un modelo de red de regulación génica de la EMT.

UMAP Proyección y aproximación de variedad uniforme, por sus siglas en inglés *Uniform Manifold Approximation and Projection*. Técnica de reducción de dimensionalidad.

1 Introducción general

El cáncer de mama es una de las enfermedades oncológicas con mayor incidencia e impacto en la salud pública en Argentina y a nivel mundial. A pesar de los avances en las técnicas de diagnóstico y tratamiento, el cáncer de mama sigue siendo una de las principales causas de mortalidad por cáncer. Su complejidad biológica y heterogeneidad presentan grandes desafíos para el diagnóstico y la decisión sobre el tipo de tratamiento y se considera que esta heterogeneidad es un aspecto crítico que limita la efectividad de las terapias convencionales. El cáncer de mama no es una enfermedad única, sino que abarca una variedad de subtipos con características moleculares, genéticas y epigenéticas distintas. Esta diversidad se observa tanto entre los distintos pacientes (heterogeneidad intertumoral) como dentro de un mismo tumor (heterogeneidad intratumoral), dentro de un solo tumor pueden coexistir múltiples subpoblaciones de células con características y comportamientos diferentes, incluyendo la respuesta a los tratamientos.

La heterogeneidad intratumoral es particularmente relevante, ya que permite a las células cancerosas adaptarse y evolucionar frente a cambios en su microambiente y a presiones selectivas, como las generadas por los tratamientos médicos. Esta diversidad celular da origen a la aparición de subpoblaciones con características distintivas, tales como la capacidad de resistir a tratamientos o de iniciar metástasis. Entre estas subpoblaciones, las células madre cancerosas (CSC, del inglés *Cancer Stem cells*) han captado una atención creciente en el ámbito de la investigación en cáncer. Estas células han sido propuestas como uno de los principales factores en la recaída y progresión de la enfermedad, ya que son capaces de regenerar el tumor después de un tratamiento aparentemente exitoso.

En este contexto, en los últimos años, la secuenciación de ARN de células individuales (scRNA-seq, del inglés *Single-cell RNA sequencing*) se ha consolidado como una herramienta fundamental para explorar la heterogeneidad. A diferencia de los métodos de secuenciación convencionales, que proporcionan el perfil de expresión génica promedio de una muestra, la técnica de scRNA-seq permite capturar el transcriptoma de cada célula individualmente. Esto no solo posibilita la identificación de tipos celulares raros y la caracterización de subpoblaciones, sino que también permite estudiar las interacciones entre células y los mecanismos de regulación génica específicos de cada subpoblación. En el caso del cáncer de mama, el análisis de datos de scRNA-seq ha emergido recientemente como una herramienta clave para estudiar la diversidad celular que caracteriza a estos tumores, identificando poblaciones de células tumorales y del microambiente tumoral que interactúan entre sí.

Uno de los objetivos centrales de esta tesis doctoral es desarrollar métodos computacionales para cuantificar propiedades biológicamente relevantes en cáncer a partir de datos de scRNA-seq. En primer lugar, se propone un marco metodológico novedoso para identificar células con características de pluripotencia. Este enfoque se basa en la integración de redes de interacción de proteínas asociadas al proceso de pluripotencia. Para ello, se propone la definición de un parámetro, llamado actividad, que mide el grado de activación de dicho proceso en células individuales. También se propone extender la aplicabilidad de esta herramienta para cuantificar la actividad de procesos biológicos en general, asociados a mecanismos relevantes en cáncer. Adicionalmente, se proponen otros parámetros para cuantificar otras características

de interés en cáncer, como la alteración en el número de copias de genes, la entropía y la heterogeneidad transcriptómica. Se aplicarán estas herramientas para analizar cómo estas propiedades varían entre subtipos de cáncer de mama, con el fin de identificar patrones asociados a agresividad y resistencia terapéutica asociados a los distintos subtipos.

Por otro lado, existen redes de regulación génica que controlan procesos críticos en la progresión tumoral, como la transición epitelio-mesénquima (EMT, del inglés *Epithelial-Mesenchymal Transition*) y la pluripotencia. La EMT es un proceso mediante el cual las células epiteliales, que normalmente forman tejidos estructurados, pierden sus características de adhesión y adquieren propiedades mesenquimales, lo cual les confiere una mayor capacidad de migración e invasión. Este proceso es fundamental en la progresión del cáncer y en el desarrollo de metástasis, ya que permite a las células tumorales migrar del tumor primario e invadir tejidos circundantes y órganos distantes. La pluripotencia es una característica de las células madre y progenitoras, las cuales tienen la capacidad de diferenciarse en múltiples tipos celulares. En el contexto del cáncer, la pluripotencia está asociada con una alta tasa de proliferación y de generación de nuevas células tumorales, lo que contribuye a la agresividad del tumor y a su capacidad para regenerarse después de un tratamiento. La interacción entre la EMT y la pluripotencia es particularmente relevante en el cáncer de mama, ya que ambos procesos se encuentran frecuentemente activados, en particular en los subtipos más agresivos.

El estudio de estas redes de regulación génica complejas requiere un enfoque multidisciplinario que combina observaciones experimentales con el modelado. El modelado matemático permite simular y analizar la dinámica de factores de transcripción activadores o inhibitorios clave, proporcionando una herramienta útil para realizar predicciones y para la comprensión de los mecanismos subyacentes que definen un dado fenotipo o estado celular. En particular, el uso de ecuaciones diferenciales ordinarias ha demostrado ser efectivo para describir la dinámica de las redes génicas, ya que permite modelar la interacción entre múltiples genes y sus productos de una manera que captura el comportamiento no lineal y la interdependencia de estos sistemas. En esta tesis, se propone el desarrollo de modelos matemáticos basados en ecuaciones diferenciales para explorar la dinámica de las redes de regulación génica de la pluripotencia (como el circuito de OCT4, SOX2 y NANOG) y de la EMT (como el circuito de miR-200/ZEB/miR-34/SNAIL). La formulación matemática de estos circuitos permite estudiar cómo la regulación de la expresión génica mediada por factores de transcripción y microARNs influye sobre el estado celular. A su vez, se propone investigar cómo la interacción y el acoplamiento de las redes de pluripotencia y EMT pueden explicar algunas observaciones experimentales.

En conjunto, esta tesis tiene como objetivo abordar la complejidad de la expresión génica en el cáncer de mama mediante un enfoque integrador que combina el estudio de datos de scRNA-seq y modelos matemáticos de redes de regulación génica. Se busca así profundizar en la comprensión de la heterogeneidad tumoral y de los procesos biológicos que contribuyen a la progresión de la enfermedad, con énfasis en la pluripotencia y en la EMT, y en la interacción entre ambos procesos.

Estructura de la tesis

Esta tesis se organiza en dos partes fundamentales en base al marco teórico, las herramientas metodológicas utilizadas y la disciplina en la que se encuadra. En términos generales, en la

Parte I se emplean métodos bioinformáticos para analizar datos públicos de expresión génica de célula única. La Parte II se basa en un enfoque de biología de sistemas utilizando modelos matemáticos basados en ecuaciones diferenciales de dos redes de regulación génica relevantes en cáncer.

Capítulo 1: Introducción general

En este capítulo se presenta una introducción general a la tesis, donde se abordan los conceptos fundamentales que orientan el desarrollo del trabajo. Se plantean los objetivos generales y se establece un hilo conductor que articula los distintos capítulos. Además, se brinda una descripción del marco teórico y de las principales herramientas que se utilizan a lo largo del trabajo. Finalmente, se describe la estructura de la tesis, acompañada de una breve descripción del contenido de cada capítulo.

Parte I: Análisis de datos de scRNA-seq

Capítulo 2: Introducción

En este capítulo se presenta una introducción al sistema biológico de interés: el cáncer de mama. Además, se describe la técnica de scRNA-seq y se detallan los aspectos metodológicos comunes a la primera parte de la tesis, incluyendo las principales etapas de preprocesamiento y análisis de los datos. Se resumen las estrategias y herramientas computacionales empleadas en el análisis de los datos (*downstream analysis*), proporcionando el marco metodológico general utilizado en los capítulos siguientes.

Capítulo 3: Pluripotencia e inferencia de la trayectoria de diferenciación en mama sana

Se propone una herramienta computacional para cuantificar la pluripotencia a partir de datos transcriptómicos de scRNA-seq. Este enfoque utiliza la red de interacción proteína-proteína (PPIN del inglés *protein-protein interaction network*) asociada con el proceso de diferenciación, y la matriz de expresión génica para calcular un parámetro que se denominará actividad de diferenciación. Este *score* refleja cuán activa se encuentra la PPIN asociada a la diferenciación para cada célula. Se evalúa el rendimiento del algoritmo propuesto en comparación con dos herramientas publicadas, en cuatro conjuntos de datos: mama, colon, médula ósea y pulmón. Finalmente, se infiere la trayectoria de diferenciación de la mama sana, describiendo el flujo de trabajo completo desde los datos crudos.

Capítulo 4: Cuantificación de procesos biológicos en mama sana y cáncer de mama

En este capítulo se extiende el enfoque desarrollado en el capítulo anterior para cuantificar la actividad asociada a una PPIN arbitraria. De este modo, se generaliza la metodología para cuantificar la actividad asociada a cualquier proceso biológico, o incluso a cualquier lista de genes. Se aplica el método propuesto analizando diversos procesos biológicos en muestras humanas de mama sana y de cáncer de mama.

Capítulo 5: Heterogeneidad en cáncer de mama

Se analizan datos públicos de scRNA-seq provenientes de muestras de cáncer de mama humano, clasificadas en los tres subtipos más prevalentes: ER+, HER2+ y triple negativo. El objetivo principal es evaluar conceptos fundamentales en la biología del cáncer, tales como las alteraciones en el número de copias de genes (CNAs, del inglés *Copy number alterations*), la entropía, la heterogeneidad transcriptómica y la actividad de distintas redes de interacción de proteínas. Para ello, se propone un marco metodológico para la cuantificación de estos

aspectos. Se exploran los distintos *scores* tanto a nivel de célula individual como a nivel de muestra, y se evalúa el comportamiento de estos en los diferentes subtipos de cáncer de mama.

Parte II: Redes de regulación génica

Capítulo 6: Introducción

En este capítulo se introducen los procesos de transición epitelio-mesénquima (EMT) y pluripotencia, destacando su relevancia en el contexto del cáncer y el estado actual del conocimiento sobre estos fenómenos. Se describe el proceso de expresión génica junto con los principales mecanismos biológicos que lo regulan, proporcionando así el marco conceptual necesario para el posterior modelado matemático. Además, se detallan los aspectos metodológicos a la segunda parte de la tesis, incluyendo el formalismo utilizado para modelar redes de regulación génica y el estudio de sistemas dinámicos. Se presentan las herramientas matemáticas y computacionales que serán aplicadas para modelar las redes de EMT y de pluripotencia en los capítulos siguientes.

Capítulo 7: Modelos de redes de regulación génica de la transición epitelio-mesénquima y la pluripotencia

En este capítulo se revisa el conocimiento actual de las principales moléculas y los mecanismos de regulación específicos de la EMT y la pluripotencia. Se describen los modelos matemáticos desarrollados hasta el momento de la red regulatoria central de la EMT y de la pluripotencia. Se analiza el comportamiento de estas redes regulatorias y se realizan adaptaciones de los modelos para los fines de esta tesis de doctorado.

Capítulo 8: Acoplamiento de los módulos de pluripotencia y transición epitelio-mesénquima

En este capítulo se revisan los principales trabajos que han abordado la integración de modelos matemáticos de las redes de regulación génica asociadas a la EMT y a la pluripotencia. A continuación, se propone la integración de los modelos desarrollados en el capítulo anterior para la EMT (SNAIL/ZEB/miR-200/miR-34) y la pluripotencia (OCT4/SOX2/NANOG), haciendo especial énfasis en la motivación biológica detrás de la selección de estas redes, los modelos matemáticos utilizados y las posibles interacciones entre ambos módulos.

Capítulo 9: Conclusiones generales

Se presentan las conclusiones generales de la presente tesis doctoral, preguntas que aún permanecen abiertas y posibles líneas de investigación futuras que surgen de los resultados obtenidos.

Apéndices

Apéndice A: Tablas de la Parte I. Se proporcionan las tablas complementarias de la Parte I, incluyendo: el listado de procesos biológicos para los cuales se calcula la actividad de la PPIN asociada del Capítulo 3, los detalles técnicos del preprocesamiento de muestras de scRNA-seq, y los resultados del análisis estadístico completos del Capítulo 5.
Apéndice B: Figuras suplementarias.
Apéndice C: Derivación matemática de las expresiones utilizadas para modelar la regulación por factores de transcripción en la Parte II.
Apéndice D: Se presentan las tablas que detallan los valores de los parámetros y de las

condiciones iniciales utilizadas en los modelos de redes de regulación génica de la Parte II.

Trabajos publicados
Lista de trabajos publicados en el marco de esta Tesis Doctoral.

Parte I

Análisis de datos de scRNA-seq

Introducción

2

> *Si las personas no creen que las matemáticas son simples, es solo porque no se dan cuenta de lo complicada que es la vida*
>
> **— John von Neumann**

2.1. Mama sana

La mama es un órgano glandular presente en todos los mamíferos, cuyo rol principal es la producción de leche para alimentar a las crías por parte de la madre, un proceso conocido como lactancia. La mama es un órgano complejo compuesto por tres tipos principales de tejidos: el tejido glandular, el tejido adiposo y el tejido conectivo. En la Figura 2.1 se muestra una ilustración de la mama humana sana adulta.

El tejido glandular es el tejido funcional de la mama y está compuesto por una red de lóbulos y conductos galactóforos, responsables de producir y transportar la leche hacia el pezón. Los lóbulos contienen alvéolos que se agrupan en racimos y están revestidos por células epiteliales. El tejido glandular de la mama adulta está revestido por una bicapa de células epiteliales, conocidas como luminales y basales. Las células luminales son las células más internas y forman el revestimiento de los conductos y alvéolos. Estas células son las responsables de secretar la leche durante la lactancia y expresan receptores hormonales específicos, como los receptores de estrógeno y progesterona, que regulan su proliferación y diferenciación. Las células basales forman la capa externa del tejido glandular y se localizan entre las células luminales y la membrana basal. El epitelio basal consiste en células mioepiteliales, que tienen propiedades contráctiles, lo que permite el transporte de la leche a través de los conductos durante la lactancia, y células madre mamarias, una pequeña población de células pluripotentes. Por esto, se dice que el epitelio basal es bipotente, ya que puede producir tanto células basales como luminales [1, 2]. También se ha reportado ampliamente la existencia de células progenitoras luminales capaces de generar todo el epitelio luminal [3] y otros trabajos reportan que bajo ciertas condiciones las células luminales unipotentes presentan plasticidad y pueden desdiferenciarse a basales [4].

El tejido adiposo es el tejido que rodea y separa las unidades glandulares y es importante tanto para la protección como para el soporte estructural de la mama. También tiene funciones endócrinas, regulando el crecimiento epitelial y su función, así como los factores de crecimiento que afectan el desarrollo y funcionamiento del tejido mamario [5].

Tejido conectivo, también conocido como estroma, sostiene y organiza el tejido glandular y adiposo. Está compuesto principalmente por fibras de colágeno y elastina, proporcionando soporte estructural y contribuyendo a la elasticidad de la mama. Otros tipos celulares que componen el tejido mamario son células inmunes, endoteliales y fibroblastos.

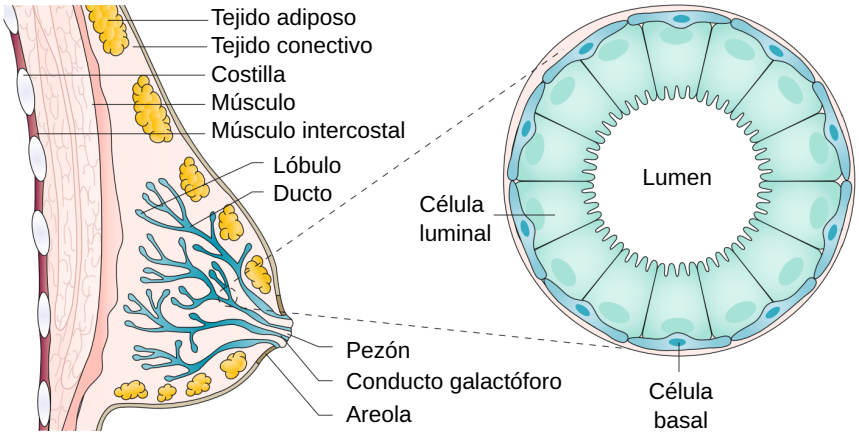

Figura 2.1.: Esquema de la mama humana adulta. A la izquierda se ilustra una vista lateral donde se distinguen los lóbulos y conductos galactóforos. A la derecha, se muestra un corte transversal de un conducto, donde se aprecia que está formado por dos tipos principales de células: células luminales, que recubren el *lumen* del conducto, y células basales, ubicadas entre las luminales y la membrana basal. Adaptación de [6].

2.2. Cáncer de mama

El cáncer de mama es el tipo de cáncer más frecuentemente diagnosticado y la quinta causa principal de mortalidad por cáncer a nivel mundial [7, 8]. El cáncer de mama se origina en el epitelio mamario. Generalmente comienza en las células que recubren los conductos galactóforos, los conductos encargados de transportar la leche desde las glándulas hasta el pezón. Este tipo de cáncer recibe el nombre de carcinoma ductal. Asimismo, las células de los lóbulos, estructuras que agrupan las glándulas productoras de leche, también pueden ser origen de tumores, conocido como carcinoma lobular [9]. Tanto los carcinomas ductales como los lobulares pueden presentarse *in situ*, es decir, confinados al área donde surgieron inicialmente, sin invasión a los tejidos circundantes.

El cáncer de mama abarca un conjunto heterogéneo de subtipos, lo que ha impulsado el desarrollo de diversos sistemas de clasificación. Actualmente, uno de los enfoques más empleados para su clasificación es la inmunohistoquímica de receptores hormonales de las muestras de cáncer, basada en la expresión del receptor de estrógeno (ER), el receptor de progesterona (PR) y el receptor 2 del factor de crecimiento epidérmico humano (HER2). Este sistema permite la identificación de cuatro subtipos principales de cáncer de mama: luminal A, luminal B, HER2 y triple negativo [10-12].

Los subtipos luminal A y B se definen por la presencia de receptores de estrógeno (ER+). El subtipo luminal B se caracteriza por niveles menores de ER y PR, tasas de proliferación más elevadas y una expresión variable de HER2 en comparación con el subtipo luminal A, que generalmente muestra niveles bajos de HER2 [13, 14]. Juntos, los subtipos luminales constituyen aproximadamente el $70\,\%$ de todos los cánceres de mama y, en general, se asocian con un pronóstico favorable. Por otro lado, el cáncer de mama HER2 (HER2+) se distingue por la sobreexpresión del receptor HER2 y la ausencia de ER, representando entre el $15-20\,\%$ de los casos. Este subtipo manifiesta un comportamiento clínico más agresivo en comparación con los subtipos luminales [11]. Finalmente, el cáncer de mama triple negativo (TN), que constituye hasta el $15\,\%$ de los diagnósticos, se caracteriza por la ausencia de expresión de

ER, PR y HER2, siendo el subtipo más agresivo y difícil de tratar, asociado a bajos niveles de diferenciación celular y altas tasas de proliferación.

2.3. Células madre cancerosas

En cáncer, se ha identificado una pequeña subpoblación de células, conocidas como células madre cancerosas (CSC, por sus siglas en inglés *Cancer Stem Cells*), que poseen propiedades similares a las de las células madre normales, incluyendo la autorrenovación y la multipotencia. La existencia de las CSCs se propuso inicialmente al observar que solo una minoría de las células cancerosas era capaz de iniciar tumores tanto *in vitro* como *in vivo* en leucemia y mieloma múltiple [15]. Estudios posteriores en malignidades hematológicas y tumores sólidos, como cáncer de mama, cerebro y colon, han reforzado el concepto de CSCs y su papel en la progresión tumoral, la resistencia a terapias y la recaída. Actualmente se sabe que la proporción de CSCs en tumores es muy baja y generalmente constituyen entre el $0{,}01\,\% - 2\,\%$ de la masa tumoral [16], lo cual presenta un desafío para identificarlas y estudiarlas.

El origen de las CSCs sigue siendo objeto de debate. Hay evidencia que demuestra que las CSCs surgen de células madre normales que adquieren mutaciones y dan inicio al tumor. Mientras que también hay evidencia de que las CSCs pueden originarse a partir de células diferenciadas con capacidad de desdiferenciarse y reactivar programas similares a los de las células madre bajo ciertas condiciones, como la inflamación crónica o la influencia del microambiente tumoral.

Independientemente de su origen, las CSCs exhiben perfiles moleculares y fenotipos que comparten con las células madre normales y pueden promover la carcinogénesis. Las CSCs se caracterizan por su capacidad de autorrenovación y diferenciación en las diversas poblaciones celulares que componen el tumor. Esta plasticidad les permite adaptarse a las condiciones dinámicas del microambiente, evadir respuestas inmunes y resistir a intervenciones terapéuticas. Además, se ha postulado que las CSCs se presentan como un paradigma para comprender la heterogeneidad tumoral, dado su papel en la generación de subpoblaciones celulares con características moleculares y funcionales distintas y la capacidad de mantener estas poblaciones [17].

El desafío de dirigir terapias contra las CSCs radica en su resistencia a los tratamientos convencionales, como la quimioterapia y la radioterapia. El estudio de las CSCs es esencial para comprender los procesos fundamentales de la iniciación, progresión y recaída del cáncer. Este conocimiento también es crucial para el desarrollo de terapias dirigidas a eliminar las CSCs, reduciendo así la metástasis, superando la resistencia terapéutica y mejorando los resultados en los pacientes. Por estos motivos, las CSCs son actualmente objeto de estudio para el desarrollo de terapias contra el cáncer y en los últimos años se han comenzado ensayos clínicos de distintas drogas dirigidas a ellas [16, 18].

2.4. Heterogeneidad tumoral

La heterogeneidad tumoral en el cáncer de mama es una característica central que se refiere a la variabilidad genética, epigenética y fenotípica observada tanto entre pacientes (heterogeneidad intertumoral) como dentro de un mismo tumor (heterogeneidad intratumoral) [19, 20].

Esta diversidad celular permite a las células tumorales adaptarse y evolucionar, generando subpoblaciones con características distintas en respuesta a cambios en el microambiente tumoral y a las presiones selectivas, como la terapia.

En cáncer de mama, la heterogeneidad tumoral se manifiesta en diferencias en la expresión de receptores hormonales y de factores de crecimiento, así como en variaciones en el perfil mutacional entre distintas regiones del tumor. Por ejemplo, ciertas subpoblaciones de células pueden presentar mutaciones específicas o estados epigenéticos que les confieren ventajas proliferativas o resistencia a la apoptosis, favoreciendo la aparición de clones agresivos y resistentes a la terapia. Estas subpoblaciones dificultan el tratamiento, ya que diferentes áreas del tumor pueden responder de manera distinta a los tratamientos.

Estudios basados en análisis transcriptómico de células individuales han proporcionado información valiosa sobre esta heterogeneidad en el cáncer de mama, revelando la presencia de diferentes tipos celulares. Esta variabilidad está asociada con un peor pronóstico, ya que facilita la evasión de la respuesta inmune y permite que las células tumorales se adapten rápidamente a ambientes cambiantes. Así, la comprensión de la heterogeneidad tumoral en el cáncer de mama resulta clave para el desarrollo de terapias personalizadas.

2.5. Secuenciación de ARN de células individuales

La secuenciación de ARN de células individuales (scRNA-seq) es un conjunto de técnicas de secuenciación de nueva generación que permiten determinar las secuencias de nucleótidos y, mediante su mapeo bioinformático, se puede estimar la abundancia de los transcriptos de células individuales. A diferencia de los métodos de secuenciación en masa (*bulk*) mediante los cuales se obtienen niveles de expresión génica promedio dentro de un tejido. Esta técnica fue introducida por primera vez en el año 2009 [21] y, con el advenimiento de nuevas tecnologías, hoy se realizan análisis transcriptómicos de alta resolución y alto rendimiento de células individuales, pudiéndose secuenciar el transcriptoma de hasta un millón de células en un proyecto. Esta técnica proporciona información que no se puede obtener a través de los métodos tradicionales de secuenciamiento (*bulk*): mayor resolución de las diferencias celulares, composición de diferentes tipos de células dentro de un tejido, descubrir poblaciones de células raras, encontrar relaciones de regulación entre genes y rastrear las trayectorias de distintos linajes celulares en desarrollo [22].

Como se menciona previamente, los tumores de mama contienen un espectro heterogéneo de células que incluye células cancerosas, vasculares, inmunes y de fibroblastos. Mediante scRNA-seq se puede comprender con mayor profundidad la heterogeneidad de las poblaciones celulares y la diversidad de estados celulares, lo que la convierte en una herramienta útil para diseccionar las propiedades de los múltiples tipos de células dentro y alrededor de los tumores de mama. En trabajos previos se ha utilizado scRNA-seq para analizar la heterogeneidad tumoral en el cáncer de mama y se han identificado grupos de células relacionadas con un mal pronóstico o una respuesta terapéutica. Los perfiles de expresión génica de las células inmunes del microambiente tumoral de cáncer de mama han revelado subpoblaciones de células inmunitarias específicas que pueden ser posibles dianas de inmunoterapia. A partir de datos de scRNA-seq de cáncer de mama también se han llevado a cabo estudios centrados en las comunicaciones célula-célula, los estados reguladores de una sola célula y la distribución de las células inmunitarias. Además, se ha utilizado scRNA-seq para analizar la asociación

entre la respuesta terapéutica y las células inmunes infiltradas específicas en el entorno tumoral [23, 24].

El proceso de la técnica de scRNA-seq se puede dividir, en líneas generales, en dos partes: la adquisición de datos y el procesamiento de los mismos. La primera parte implica la implementación de protocolos experimentales mediante los cuales se determinan las secuencias de nucleótidos, que varían según la plataforma utilizada. Por otro lado, el procesamiento de los datos comprende el análisis computacional desde los datos crudos hasta lo que se conoce como *downstream analysis*, el cual incluye una amplia gama de herramientas matemáticas y computacionales para estudiar el sistema biológico de interés. En esta tesis, nos centraremos en este último aspecto, pero antes haremos una introducción general a la técnica.

2.5.1. Adquisición de los datos

En comparación con el secuenciamiento tradicional de ARN (*bulk*), el aislamiento de las células es el primer paso para obtener información del transcriptoma de una célula individual. Recientemente se han desarrollado una gran variedad de protocolos y enfoques para scRNA-seq, aunque en aspectos generales todos estos protocolos comparten los mismos principios básicos. Existen distintos métodos de aislamiento como la micromanipulación, la microdisección por captura láser (LCM, por sus siglas en inglés *Laser Capture Microdissection*), técnicas de microfluídica y la clasificación de células activadas por fluorescencia (FACS, por sus siglas en inglés *Fluorescence-Activated Cell Sorting*). En segundo lugar, se necesitan protocolos específicos para realizar la transcripción reversa de ARNm y la amplificación de ADNc (ADN complementario) con alta eficiencia, ya que la cantidad de ARN en una sola célula es menor que en los análisis de ARN *bulk*. En la Figura 2.2 se esquematizan las etapas experimentales de la técnica scRNA-seq.

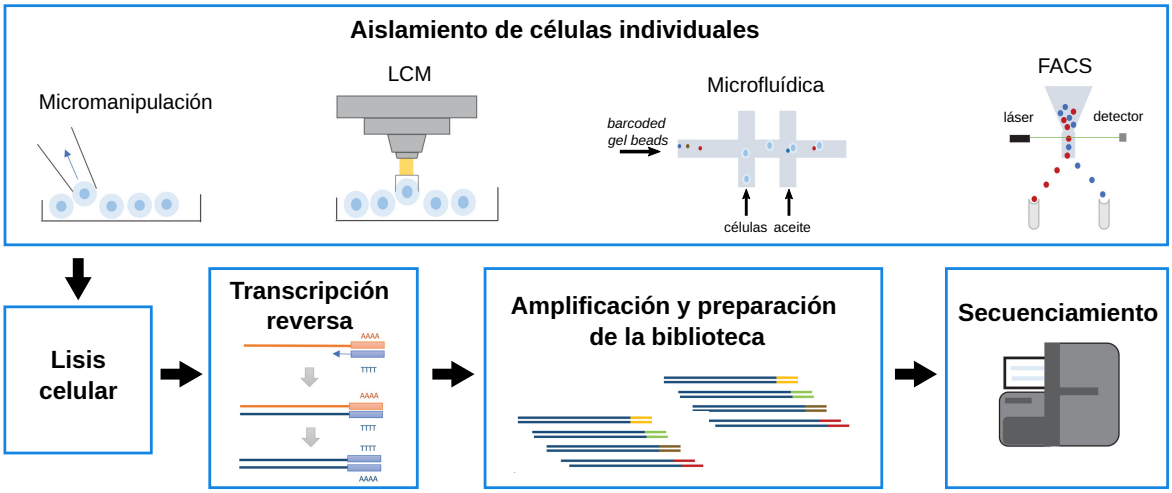

Figura 2.2.: Pasos para la adquisición de datos scRNA-seq. Adaptación de [25].

Aislamiento de las células individuales

La micromanipulación implica la recolección manual de células individuales utilizando una pipeta de vidrio bajo un microscopio. La supervisión microscópica asegura que cada muestra

corresponda efectivamente a una sola célula; sin embargo, este método lleva mucho tiempo y tiene un bajo rendimiento, además de que el cizallamiento mecánico durante el proceso puede dañar las células. Por estos motivos, la micromanipulación raramente se utiliza. La microdisección por captura láser (LCM) es otra técnica para capturar células individuales de tejido sólido. Este método se realiza bajo el microscopio, donde se emplea un rayo láser para unir las células seleccionadas a un polímero colocado sobre un portaobjetos de vidrio. Las células individuales de la película se transfieren a un tubo para microcentrífuga. Aunque LCM es un proceso trabajoso y técnicamente desafiante, especialmente en lo que respecta a evitar la contaminación de células circundantes y proteger la integridad del ARN celular, presenta ventajas importantes. La técnica permite el aislamiento de células individuales a partir de muestras sólidas y puede proporcionar información espacial valiosa, por estas razones todavía es utilizada en la práctica.

En general, la microfluídica se refiere a tecnologías que utilizan estructuras de microescala para manipular fluidos con precisión en volúmenes ultrabajos, típicamente en la escala de nanolitros a femtolitros. Entre las plataformas de microfluídica se destacan la plataforma robótica Fluidigm C1 y los métodos basados en microgotas. En las plataformas de microfluidos, el aislamiento automatizado de células individuales es seguido por transcripción reversa automatizada y la preamplificación en chip, lo que permite reducir los volúmenes de reacción a nanolitros o incluso picolitros, optimizando los costos. Los métodos basados en microgotas, ofrecen la capacidad de capturar miles de células en un solo experimento, lo que ha impulsado su popularidad debido a su alto rendimiento y bajo costo. Sin embargo, su principal desventaja es que la eficiencia de captura puede ser baja para células con alta adhesividad.

La clasificación de células activadas por fluorescencia es un tipo especializado de citometría de flujo. Este método se basa en el uso de anticuerpos específicos marcados con fluorescencia para detectar y clasificar células de interés dentro de poblaciones celulares heterogéneas. Debido a su alto rendimiento, bajo costo y flexibilidad de uso con distintos tipos de muestras, FACS se ha convertido en un método extensamente utilizado para aislar células individuales. Sin embargo, los inconvenientes de este método incluyen el riesgo de dañar la viabilidad celular durante la clasificación y la dificultad para diferenciar subpoblaciones de células que expresan marcadores similares.

Cada método de aislamiento de células individuales tiene sus propias ventajas y desventajas. La micromanipulación permite una recolección precisa con un bajo riesgo de daño celular. LCM permite el aislamiento de células individuales en muestras sólidas y preserva la información espacial. Los métodos de microfluídica disminuyen el volumen de reactivos, haciéndolos rentables. FACS permite capturar células de interés específicas dentro de una población heterogénea.

Transcripción reversa y amplificación del ADNc

Los pasos usualmente requeridos para la generación de bibliotecas (comúnmente conocidas como *libraries*) de scRNA-seq incluyen la lisis celular, la transcripción reversa en ADNc de la primera hebra, la síntesis de la segunda hebra y la amplificación de ADNc [22]. Típicamente, las células se lisan en una solución tampón hipotónica y se realiza un filtrado para capturar solo ARN mensajeros (ARNm) seleccionando las colas poliadeniladas mediante cebadores Oligo(dT). Para la transcripción reversa (de ARNm a ADNc), el ARNm se utiliza como molde para que la transcriptasa reversa sintetice la primera hebra de ADNc utilizando como cebadores Oligo(dT), que hibridan con las colas poliadeniladas. La transcriptasa reversa

genera una molécula híbrida de ADN/ARN. Las segundas hebras se pueden generar utilizando colas de poli(A) (*poly(A) tailing*) como lo hace Quartz-seq o mediante un mecanismo de cambio de plantilla (*template switching*) empleado por Smart-seq, Smart-seq2, STRT-seq y Drop-seq. El ADNc debe ser amplificado para generar cantidades suficientes de material para preparar las bibliotecas de secuenciación. Esto se realiza utilizando PCR convencional o transcripción [26].

Entre la amplia gama de protocolos de scRNA-seq existentes, un aspecto clave es que algunos proporcionan datos de transcripción completos, como Smart-seq y Smart-seq2, mientras que otros cuentan solo los extremos $3'$ o $5'$ de las transcripciones, como CEL-seq y MARS-seq. La selección de un protocolo específico depende de la naturaleza de la pregunta de investigación. En comparación con los métodos que solo capturan y secuencian los extremos $3'$ o $5'$ de los ADNc, los protocolos capaces de transcripción completa son más adecuados para análisis de patrones de *splicing* alternativos, detección de expresión alélica e identificación de edición de ARN. La ventaja de los protocolos que secuencian el extremo $3'$ o $5'$ de la transcripción es que pueden incorporar identificadores moleculares únicos (UMI) en el proceso de transcripción reversa. Estas etiquetas permiten la identificación y cuantificación de transcripciones individuales, lo que mejora la cuantificación a nivel de genes. Por lo tanto, en algunos casos en los que no se requieren datos de transcripción completa, estos protocolos que utilizan *barcodes* se han convertido en los más utilizados para fines de cuantificación, especialmente para grandes poblaciones de células.

2.5.2. Procesamiento de los datos

Preprocesamiento

Control de calidad. La presencia de datos de baja calidad en scRNA-seq es problemática, ya que puede llevar a resultados erróneos en los análisis posteriores (*downstream analysis*). La baja calidad en los datos de scRNA-seq puede originarse por varias razones, como daños celulares durante el aislamiento o fallas en la preparación de la biblioteca (por ejemplo, transcripción reversa o amplificación por PCR ineficientes). Estas deficiencias suelen manifestarse en "células" con un bajo número total de cuentas, pocos genes expresados y proporciones elevadas de expresión de genes mitocondriales o de *spike-in* [27]. Para identificar las células de baja calidad en base a los perfiles de expresión, se utilizan métricas de control de calidad (*QC metrics*).

- Tamaño de la biblioteca: definido como la suma total de las cuentas de todos los genes de cada célula. Las células con tamaños de biblioteca pequeños son de baja calidad, ya que el ARN se ha perdido en alguna fase de la preparación de la biblioteca.
- Número de genes expresados: representa el número de genes endógenos con cuentas no nulas en cada célula. Las células con muy pocos genes expresados son de baja calidad, lo que indica que no se ha capturado adecuadamente la diversidad del transcriptoma celular, limitando la capacidad para representar fielmente el perfil de expresión de cada célula.
- Proporción de lecturas mitocondriales: el porcentaje de lecturas mapeadas a genes mitocondriales suele emplearse como una métrica de calidad. Una alta proporción de lecturas mitocondriales es indicativa de células de baja calidad debido a la pérdida de ARN citoplasmático en células dañadas. En estos casos, la membrana celular dañada

permite la salida de los transcriptos en general pero la salida de los mitocondriales se encuentra impedida, lo que resulta en un enriquecimiento relativo de transcriptos mitocondriales.

Normalmente, en base a estas medidas de control de calidad, se definen valores de corte que varían entre experimentos y tipos celulares para filtrar células dañadas, dobles o múltiples, *empties* y *debris*.

Normalización y factor de escala. Las diferencias sistemáticas en la cobertura de secuenciación entre bibliotecas son comunes en los datos de RNA-seq de células individuales. Estas diferencias suelen surgir por variaciones técnicas en la captura de ADNc o en la eficiencia de amplificación por PCR entre células. La normalización busca eliminar estas diferencias para que no interfieran en la comparación de perfiles de expresión entre células, asegurando que la heterogeneidad observada o la expresión diferencial dentro de la población celular sea de origen biológico y no debida a sesgos técnicos.

La normalización por el tamaño de biblioteca es la estrategia más simple. Esta consiste en dividir el perfil de cuentas de cada célula por el tamaño de la biblioteca, que se define como la suma total de las cuentas de todos los genes en cada célula. Así, los valores de expresión normalizados mantienen la misma escala que las cuentas originales, para obtener datos comparables entre células. Este tipo de normalización, suele ser suficiente en aplicaciones donde el objetivo es identificar clústers y marcadores relevantes para cada uno. Además, típicamente se multiplican todos los valores de la matriz de expresión génica por un factor de escala (valores típicos alrededor de 10000) para no tener valores extremadamente pequeños en el siguiente paso.

Transformación logarítmica. El siguiente paso consiste en aplicar una transformación logarítmica a los perfiles de expresión normalizados. Dicha transformación se aplica porque las diferencias en valores log-transformados representan cambios logarítmicos en la expresión, lo cual es esencial para procedimientos de análisis que se basan en distancias euclídeas. Trabajar con datos log-transformados asegura que estos procedimientos midan distancias entre células en términos de cambios logarítmicos en la expresión. Esto permite centrarse en genes con diferencias relativas significativas en lugar de grandes valores absolutos de conteo. Por ejemplo, es más interesante un gen cuya expresión es $0{,}1$ en una célula y $0{,}2$ en otra que un gen cuya expresión es $0{,}8$ en una célula y $0{,}9$ en otra.

Genes más variables (HVF, por sus siglas en inglés *Highly variable features*) y escaleo. Para ciertos tipos de análisis, como la reducción de dimensiones o el *clustering*, se aplica una transformación adicional que se ha demostrado que mejora la eficiencia y lleva a resultados más significativos. La idea principal es seleccionar los genes que presentan mayor variabilidad en la población de células. Así se utilizan los genes que presentan variaciones biológicas y se descartan aquellos cuyas pequeñas variaciones se deben a ruido, ya sea técnico o ruido basal. Para ello, uno de los enfoques más utilizados es aplicar métodos de estabilización de la varianza y modelar la relación varianza-media. Esto permite estimar cuánto de la señal se debe a ruido, y de esta forma restarle esta estimación para utilizar solo la componente biológica. En base a esto, se pueden estimar los genes que presentan mayor variabilidad reduciendo la cantidad de variables y en consecuencia, la variabilidad causada por el ruido. Así, se pueden seleccionar la cantidad de genes más variables, siendo los valores típicos entre 1000 y 2000. Para aplicar técnicas de reducción de dimensiones, un procedimiento estándar adicional consiste en escalear los genes de tal forma que la expresión media sea 0 y la varianza sea 1. De este modo, al aplicar técnicas de reducción de dimensiones, los genes con valores más elevados de expresión no enmascaran a otros genes con menores niveles de expresión.

Downstream analysis

Reducción de dimensiones. Como se menciona previamente, no todos los genes tienen información relevante, en parte debido al ruido, aunque también hay genes que tienen información redundante ya que se coexpresan. Por este motivo, y por la "maldición de la dimensión" (en inglés *curse of dimensionality*) que ocurre al analizar datos de alta dimensionalidad, se suelen utilizar técnicas de reducción de dimensiones. Estos algoritmos permiten representar los datos en un espacio de menos dimensiones a las originales, lo cual permite aplicar técnicas de modelado y visualización. Existe un amplio espectro de herramientas, siendo una de las más utilizadas PCA (del inglés *Principal Component Analysis*), una transformación lineal que transforma los datos a un espacio que maximiza la varianza. Otros métodos comunes incluyen UMAP (por sus siglas en inglés *Uniform Manifold Approximation and Projection*), t-SNE (por sus siglas en inglés *t-Distributed Stochastic Neighbor Embedding*) y mapas de difusión. Las técnicas no lineales t-SNE y UMAP son herramientas sumamente utilizadas en la actualidad y presentan la ventaja de preservar la estructura global y local de los datos, siendo UMAP significativamente más eficiente, particularmente importante al trabajar con grandes conjuntos de datos.

***Clustering*.** El agrupamiento (*clustering*) es un conjunto de herramientas de aprendizaje no supervisado que en este contexto se utiliza para agrupar células en conjuntos que presentan características similares. Una gran cantidad de métodos y criterios se han desarrollado para agrupar datos en general. En particular, los métodos más utilizados en scRNA-seq son KNN (del inglés *k-nearest neighbors*) y k-means [28].

Análisis de expresión diferencial de genes y anotación. Son técnicas que se utilizan para identificar los genes que se expresan de manera diferente entre grupos. Usualmente se utilizan los grupos del paso anterior, y los genes obtenidos se suelen utilizar para anotar las células, es decir, identificar el tipo o estado celular. Existen numerosos métodos para llevar a cabo el análisis de expresión diferencial de genes [29].

Inferencia de trayectorias. Son enfoques desarrollados exclusivamente para datos scRNA-seq, a diferencia de los anteriores. Permiten reconstruir las trayectorias de diferenciación desde las células pluripotentes a células más especializadas e identificar la jerarquía de linajes. Existen numerosas herramientas disponibles, como Slingshot y Monocle3 [30].

Integración. Con frecuencia se dispone de múltiples conjuntos de datos que provienen de distintos experimentos, los cuales pueden haberse generado utilizando diferentes protocolos experimentales o plataformas de secuenciación. En los últimos años, se han desarrollado herramientas que permiten corregir estas diferencias técnicas, posibilitando el análisis conjunto de los datos. Entre los enfoques más utilizados para la integración de datos de scRNA-seq se encuentran Harmony, *canonical correlation analysis* (CCA) y funcionalidades del paquete Seurat, entre otros [31, 32].

Estas son algunas técnicas comúnmente utilizadas al analizar datos de scRNA-seq que se utilizan en los siguientes capítulos, aunque existen muchas más que no son descritas, como la inferencia de redes regulatorias, inferencia de ganancias y deleciones de genes *(Copy Number Alterations)*, inferencia de comunicación célula-célula, entre otras.

3 Pluripotencia e inferencia de la trayectoria de diferenciación en mama sana

3.1. Introducción

Los avances recientes en la secuenciación de ARN de una sola célula (scRNA-seq), que permiten obtener perfiles transcriptómicos de células individuales, ofrecen una capacidad prometedora para explicar los procesos de desarrollo. La capacidad de cuantificar la pluripotencia es relevante para comprender la diferenciación, los linajes celulares y la jerarquía de estos. Además, esta tarea tiene importancia en la investigación en cáncer para identificar células madre cancerosas. Asimismo, la identificación de células pluripotentes es un paso crucial para realizar la inferencia de trayectorias, una aplicación ampliamente utilizada al trabajar con datos de scRNA-seq para comprender los procesos de diferenciación.

Se han desarrollado varios algoritmos para reconstruir trayectorias de diferenciación utilizando scRNA-seq. En 2019, Saelens *et al.* reportaron la existencia de más de 70 técnicas de análisis de trayectorias [30], y en los últimos años han surgido muchas más [33-35]. Muchas de estas técnicas requieren información previa para inferir la trayectoria, como por ejemplo la célula de origen [36, 37]. La información biológica previa puede ayudar al método a encontrar la trayectoria correcta; sin embargo, un conocimiento previo incorrecto puede conducir a trayectorias incorrectas. Tradicionalmente, se utilizan marcadores de pluripotencia previamente conocidos para identificar la célula de origen, lo cual no siempre es factible debido a la alta tasa de pérdida de datos (conocida como *dropout*) asociada con la técnica de scRNA-seq. Además, los marcadores de pluripotencia dependen del tejido y del estadio de desarrollo, y no siempre están disponibles para todos los casos.

En este sentido, para cuantificar la pluripotencia, los miembros del laboratorio de Teschendorf propusieron una metodología basada en un enfoque de sistemas. En este trabajo, calcularon un parámetro llamado entropía de la red utilizando una red de interacción proteína-proteína (PPIN) [38]. En trabajos posteriores, el grupo profundizó en la investigación sobre la cuantificación de la pluripotencia, proponiendo diferentes alternativas para calcular la entropía [39-41]. Los autores afirman que las células diferenciadas tienen activadas ciertas vías específicas que conducen a niveles bajos de entropía, mientras que las células pluripotentes presentan un patrón amplio de vías de señalización activadas y no muestran preferencia por una línea particular. En términos de datos transcriptómicos de células individuales, esto se traduce en perfiles de expresión génica uniformes, lo que resulta en niveles altos de entropía.

Otras publicaciones abordan la entropía sin utilizar una PPIN como base. Por ejemplo, StemID busca identificar células madre mediante un *score* que combina la entropía media del clúster y el número de enlaces interclúster que definen la topología del árbol de linaje [42]. Otra variación es SLICE, un algoritmo basado en la entropía de Shannon con algunas modificaciones en su implementación [43].

Existen un número limitado de técnicas disponibles para cuantificar la pluripotencia que no se basan en el uso de la entropía. Un ejemplo es el trabajo de Palmer *et al.*, quienes derivaron una *signature* de expresión génica específica de la pluripotencia y la emplearon para calcular un índice de pluripotencia a partir de muestras de microarrays de expresión génica [44]. Para ello, proyectaron las coordenadas de un perfil de expresión en la primer componente principal del espacio génico definido por la *signature* de pluripotencia, utilizando esta proyección como una medida relativa de pluripotencia.

En 2020, Gulati *et al.* desarrollaron CytoTRACE, una herramienta computacional para identificar células madre a partir de datos de scRNA-seq [45]. Los autores observaron que el número total de genes expresados se correlaciona generalmente con el estado de diferenciación celular. Dado que la técnica de scRNA-seq está diseñada para capturar la expresión génica, propusieron identificar aquellos genes que se correlacionan con el número total de genes expresados y, a partir de ellos, generar una *signature* específica para el conjunto de datos. Sin embargo, señalaron una limitación importante: CytoTRACE no es adecuado para identificar células madre en estado de quiescencia, como las células madre hematopoyéticas, debido a su baja actividad metabólica y su bajo contenido de ARN.

En los últimos años, ha aumentado considerablemente la información disponible sobre las interacciones entre proteínas y su papel en los procesos biológicos. Estas interacciones son fundamentales para definir los fenotipos celulares [46]. Con el auge de las tecnologías de secuenciación de alto rendimiento, se han desarrollado diversas metodologías para integrar estos datos mediante enfoques basados en redes biológicas [39].

El objetivo general de este capítulo es desarrollar una herramienta computacional para cuantificar la pluripotencia a partir de datos de scRNA-seq. Con este fin, se propone utilizar la PPIN asociada al proceso de diferenciación celular, junto con la matriz de expresión génica, para calcular un parámetro que denominamos actividad de diferenciación. Este *score* cuantifica cuán activa se encuentra la PPIN asociada a la diferenciación en cada célula. El método propuesto se llama ORIGINS, ya que el objetivo es identificar las células pluripotentes o de origen en procesos de diferenciación.

Se propone utilizar esta medida para identificar células madre y progenitoras. Se evaluará el rendimiento del algoritmo propuesto en distintos conjuntos de datos, y se lo comparará con herramientas previamente publicadas. De este modo, se buscará evaluar distintas características técnicas relevantes de la metodología, como el manejo de la memoria RAM, el tiempo de cómputo y la flexibilidad del formato de los datos de entrada. Además, se presentará el flujo de trabajo completo, desde la matriz de cuentas hasta la inferencia de la trayectoria de diferenciación, utilizando un conjunto de datos de mama humana.

3.2. Metodología

3.2.1. Actividad de diferenciación

La Ontología Genética (GO, por sus siglas en inglés, Gene Ontology) proporciona vocabularios controlados y estructurados, así como clasificaciones detalladas de funciones moleculares, procesos biológicos y compartimentos celulares [47]. El análisis de las anotaciones de GO asociadas a un conjunto de Genes Diferencialmente Expresados (DEG, Differentially Expressed Genes) es una práctica habitual para extraer información sobre los posibles significados

biológicos de los resultados experimentales. Alternativamente, para identificar procesos biológicos diferenciales (BP, del inglés *Biological Processes*) en una población celular, en esta parte de la tesis se propuso una estrategia que utiliza el conjunto de reacciones bioquímicas involucradas en la función biológica de interés, en lugar de varios DEG. En este contexto, se construyó una PPIN asociada a los productos génicos involucrados en el proceso biológico de diferenciación celular (BP-GO:0030154) como reacciones bioquímicas putativas.

Para ello, se seleccionó un conjunto de $11\,582$ proteínas de *Homo sapiens* asociadas al BP-GO:0030154, registradas en la base de datos QuickGO [48]. Estas proteínas incluyen 87 términos hijos del proceso de diferenciación celular y constituyen los nodos de la red. Además, se consideraron las interacciones bioquímicas reportadas en Pathways Commons (versión 12), que integran $2\,424\,055$ interacciones recopiladas a partir de 22 bases de datos [49]. De este conjunto de interacciones, se seleccionaron $191\,072$ interacciones entre proteínas humanas asociadas al BP-GO:0030154, excluyendo aquellas interacciones que involucran compuestos químicos u otras moléculas que no son proteínas. Estas interacciones entre proteínas constituyen las aristas de la red que se muestra en la Figura 3.1, la cual puede ser descrita mediante una matriz de adyacencia A.

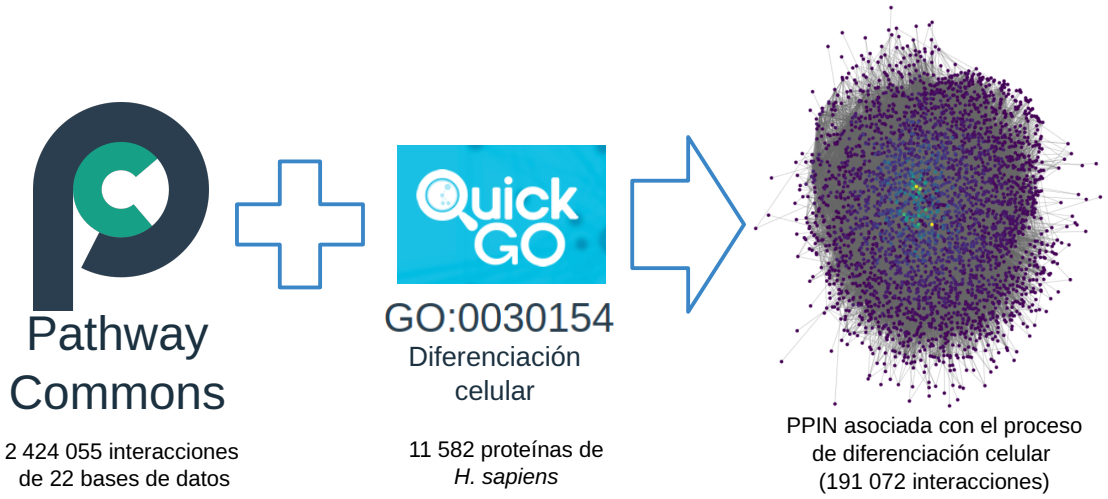

Figura 3.1.: Diagrama del procedimiento para construir la PPIN asociada al proceso de diferenciación celular a partir de información previa obtenida de Pathway Commons y Gene Ontology.

Una vez construida la PPIN asociada al proceso de diferenciación celular, se estableció un método para calcular su nivel de actividad, basado en el perfil de expresión génica de cada célula. Se asumió que el nivel de actividad de la red es el resultado de la acumulación de las reacciones bioquímicas que ocurren en ella. Siguiendo la Ley de Acción de Masas, se estimó la probabilidad de que ocurra una interacción como el producto de las concentraciones de las proteínas, sin incorporar los detalles estequiométricos. Esta aproximación permitió cuantificar la contribución de una arista de la PPIN entre los nodos a y b, $a \leftrightarrow b$, únicamente a partir del perfil de expresión como: $x_a \times x_b$, donde x_a y x_b son los niveles de expresión asociados a las proteínas A y B, respectivamente. Por lo tanto, para un perfil de expresión dado x_i, se define una matriz de los pesos de las aristas (*weighted edge matrix*) $W_{ij} = A_{ij}x_ix_j$, donde i y $j = 1, 2, ..., N_g$ y N_g es el número de genes presentes en la vía.

Así, se definió el nivel de actividad asociado al BP de diferenciación celular de la siguiente forma:

$$P = \sum_{i,j=1}^{N_g} W_{ij} \tag{3.1}$$

Para datos de scRNA-seq, cada célula k tiene un perfil de expresión asociado y la correspondiente matriz de aristas ponderadas W^k, a partir de la cual se calculó el nivel de actividad P^k asociado al BP de diferenciación celular de la k-ésima célula. Finalmente, los niveles de actividad se escalearon de manera que los valores de actividad se encuentren normalizados entre 0 y 1 de la siguiente forma: $P^k_{scaled} = \frac{P^k - min\left(P^k\right)}{max\left(P^k\right) - min\left(P^k\right)}$, donde $min\left(P^k\right)$ y $max\left(P^k\right)$ son los niveles mínimos y máximos de actividad entre todas las células. En la Figura 3.2 se esquematiza el procedimiento para calcular la actividad de la PPIN asociada al proceso de diferenciación celular, donde el *input* del algoritmo es la matriz de expresión y el *output* es un vector de actividades, donde cada elemento corresponde a una célula.

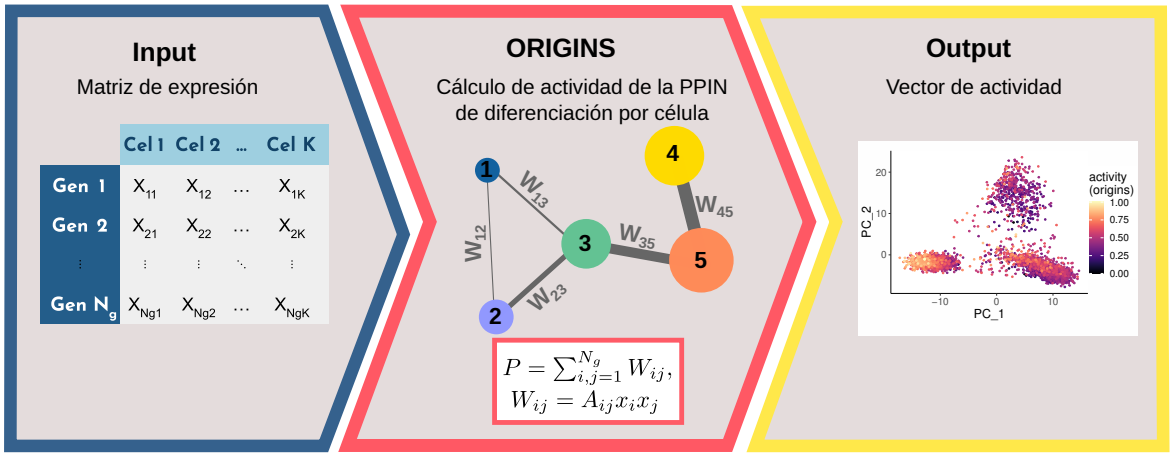

Figura 3.2.: Esquema del flujo de trabajo de ORIGINS para calcular la actividad de la PPIN asociada a la diferenciación celular a partir de datos de scRNA-seq.

3.2.2. Validación

Se evaluó el rendimiento de la metodología propuesta para cuantificar la pluripotencia utilizando cuatro conjuntos de datos de scRNA-seq de origen humano: epitelio mamario [9], epitelio de colon [50], médula ósea (células hematopoyéticas) [51] y pulmón [52]. Los tipos celulares en estos datos estaban previamente etiquetados (proceso conocido como anotación) como parte de los metadatos, exceptuando el caso del epitelio mamario. En el caso del epitelio mamario, se describe en detalle el proceso de anotación en la Sección 3.3.2: *Aplicación al epitelio de mama humano*.

Para evaluar el rendimiento de ORIGINS, se lo comparó con dos metodologías propuestas en la literatura, desarrolladas específicamente para datos de scRNA-seq: el *score* SR (del inglés *Signaling Entropy Rate*) implementado en LandSCENT [41] y el índice de pluripotencia de CytoTRACE [45]. Ambos algoritmos están disponibles como paquetes de R y fueron instalados desde sus repositorios oficiales. Además, se propuso una estrategia rápida para estimar la actividad de diferenciación. Para ello, se seleccionaron las 2000 características más variables

(HVF) de las matrices de expresión utilizando la función `FindVariableFeatures` de Seurat con el parámetro `selection.method = vst`. Esto permitió reducir la dimensionalidad de la matriz de expresión para cada muestra y calcular la actividad sobre la matriz reducida y normalizada, la cual se denominó actividad HVF (ORIGINS).

La totalidad del código se implementó en el lenguaje de programación R (versión 4.1.2). Los principales paquetes utilizados para llevar a cabo el trabajo fueron LandSCENT (versión 0.99.5), CytoTRACE (versión 0.3.3) y Seurat (versión 4.1.0). Las especificaciones del hardware utilizado fueron: Kernel versión 5.13.0-30-generic, un procesador Intel® Core™ i7-8700 de 12 núcleos a 3.20 GHz y 16 GB de RAM.

3.2.3. Descripción de los datos

- **Epitelio mamario.** Se utilizaron datos de células epiteliales mamarias humanas disponibles públicamente en la base de datos GEO (GSE113197) [9]. Se empleó una muestra de mamoplastía de reducción proveniente de una adulta sana (Ind4).
- **Epitelio del colon.** Los datos se descargaron de la base de datos GEO (GSE125970) y se utilizó una muestra de colon humano adulto (Colon-2).
- **Médula ósea.** Se utilizaron datos de scRNA-seq de progenitores hematopoyéticos de médula ósea humana, disponibles en la base de datos GEO bajo código de acceso GSE117498 [51]. La cuantificación de pluripotencia se realizó utilizando los datos correspondientes al Donante A.
- **Pulmón.** Los datos scRNA-seq incluyen 19 muestras de pulmón, en esta tesis se utilizó uno de los cinco donantes adultos (D122), un hombre sano de 32 años. Los datos se descargaron del Portal de Datos cellxgene. Los datos crudos también se encuentran disponibles en la base de datos GEO (GSE161383) [52].

3.2.4. Disponibilidad de los datos y del código

Los datos analizados en este trabajo están disponibles públicamente en la base de datos GEO bajo los siguientes códigos de acceso: GSE113197 para células epiteliales mamarias, GSE125970 para células epiteliales del colon, GSE117498 para células madre hematopoyéticas de médula ósea y GSE161383 para células pulmonares.

ORIGINS se disponibilizó como un paquete de R de código abierto en el repositorio de GitHub: https://github.com/danielasenraoka/ORIGINS.

3.3. Resultados y discusiones

3.3.1. Comparación con otros métodos

Los índices de pluripotencia (LandSCENT, CytoTRACE, ORIGINS y ORIGINS-HVF) calculados utilizando la muestra de mama se visualizan en el espacio UMAP en las Figuras 3.3A-D. Se investiga cómo varían estos parámetros según los diferentes tipos celulares: luminales (L1

y L2) y basales. Las células basales presentaron los niveles promedio más altos tanto de SR (LandSCENT) como de actividad (ORIGINS) y su aproximación (Figuras 3.3E, G y H). Este resultado era esperado, dado que en la publicación original de los datos, los autores sugirieron la presencia de células madre mamarias dentro de la población basal [9]. De manera similar, el estudio de LandSCENT también identificó a la mayoría de las células multipotentes en el epitelio mamario como células basales [41].

De la misma manera, se esperaba que el clúster L1 tuviera un nivel de pluripotencia superior al de L2, dado que estas son células más inmaduras; sin embargo, este comportamiento solo se evidenció para la actividad (ORIGINS) y su aproximación. De hecho, los resultados de CytoTRACE no coincidieron con este orden: el clúster L1 mostró, en promedio, los valores más altos, seguido de L2 y las células basales (Figuras 3.3B y F).

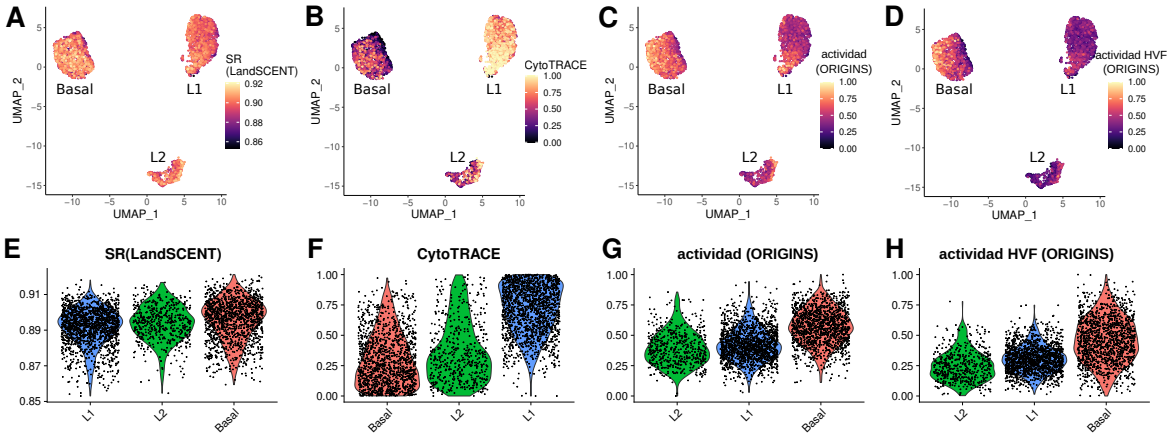

Figura 3.3.: **(A-D)**: Representación UMAP de los datos de scRNA-seq de la muestra de mama sana coloreada según los puntajes calculados por LandSCENT, CytoTRACE, ORIGINS utilizando todos los genes y los más variables (HVF). **(E-H)**: Gráficos de violín de los puntajes de pluripotencia por tipo celular ordenados según los valores crecientes de los puntajes promedio.

En cuanto a la muestra de colon, cuyas células ya estaban anotadas (Figura 3.4A), los cuatro índices de pluripotencia se representaron en el espacio UMAP en las Figuras 3.4B-E. Todos los métodos presentaron limitaciones para identificar las células madre, dado que las células de amplificación transitoria (TA, del inglés *transit amplifying*) mostraron los valores promedio más altos en lugar de las células madre (Figuras 3.4F-I). En el intestino, las células madre se dividen de manera asimétrica, dando origen a otra célula madre y a una célula hija progenitora de amplificación transitoria. Estas células TA no diferenciadas son altamente proliferativas, pasan por un número limitado de divisiones celulares y eventualmente se diferencian en linajes de absorción (enterocitos) o de secreción (células mucosas, enteroendócrinas, células de Paneth) [53-55].

En el caso de las células de médula ósea, se analizaron varios tipos celulares: células madre hematopoyéticas (HSC, del inglés *hematopoietic stem cells*), progenitoras multipotentes (MPP, del inglés *multipotent progenitors*), progenitoras multilinfoide (MLP, del inglés *multilymphoid progenitors*), linfocitos pre-B / células Natural Killer (PREB/NK), progenitoras de megacariocitos y eritrocitos (MEP, del inglés *megakaryocyte-erythroid progenitors*), progenitoras mieloides comunes (CMP, del inglés *common myeloid progenitors*) y progenitoras de granulocitos y monocitos (GMP, del inglés *granulocyte-monocyte progenitors*) [51], como se muestra en la Figura 3.5A. Según el modelo hematopoyético clásico, las HSC son las precursoras de todos

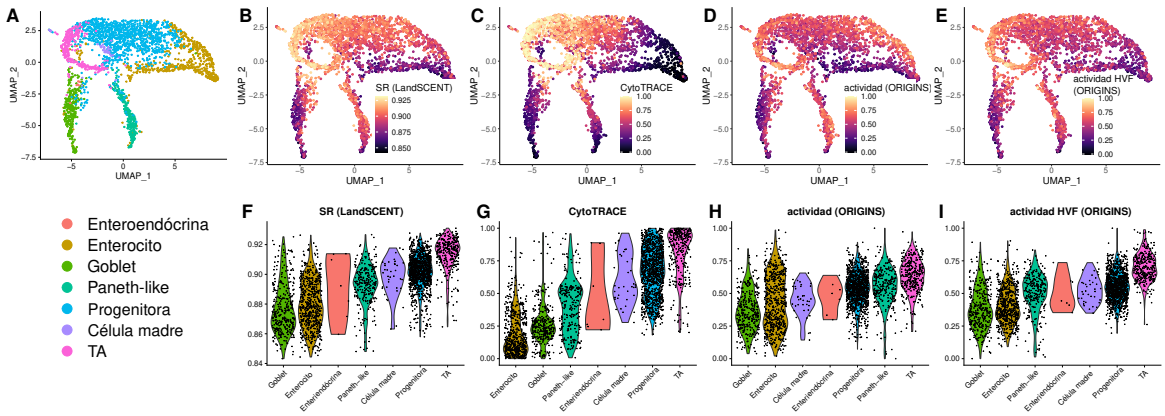

Figura 3.4.: Representación en el espacio UMAP de los datos de scRNA-seq de la muestra de colon. **(A)** Código de colores según el tipo celular. **(B-E)** Código de colores según el *score* calculado mediante LandSCENT, CytoTRACE, y ORIGINS utilizando todos los genes y los más variables (HVF). **(F-I)** Diagramas de violín de los *scores* de pluripotencia por tipo celular en orden creciente de los promedios.

los tipos de células sanguíneas [56, 57]. Estas células suelen encontrarse en un estado de quiescencia [58], pero pueden activarse en respuesta a las demandas del organismo [59]. Las HSC ocupan la cima de la jerarquía y originan diferentes progenitores. Las Figuras 3.5B-E muestran los índices de pluripotencia calculados en la representación UMAP para varios tipos de células progenitoras. Cabe destacar que todos los métodos coincidieron en que la mayor potencia promedio correspondió a las GMP, como se observa en las Figuras 3.5F-I. Aunque sería esperable que las HSC fueran las más pluripotentes, seguidas por las MPP, y luego las MLP y CMP, ninguno de los métodos parece ser adecuado para ordenar las células según su pluripotencia basándose en el modelo hematopoyético. Esto podría deberse a la variabilidad en los niveles de quiescencia y compromiso celular en las células madre y progenitoras hematopoyéticas [45, 60, 61].

En el caso del pulmón, un órgano con una gran diversidad de tipos celulares, se analizaron datos de una muestra que incluye alrededor de 30 tipos celulares distintos (Figura 3.6A). En la jerarquía celular del pulmón, las células alveolares tipo II (AT2, por sus siglas en inglés *alveolar type II*) son las células madre mejor caracterizadas, capaces de diferenciarse en células alveolares tipo I (AT1, por sus siglas en inglés alveolar type I) [62]. Las células club también actúan como células progenitoras, diferenciándose en células ciliadas [63, 64], mientras que las células basales, *club-like* y neuroendócrinas también han sido identificadas como progenitoras [65, 66]. Las Figuras 3.6B-E muestran los índices de pluripotencia calculados en el espacio UMAP. Las células AT2 y *club-like* presentaron los valores promedio más altos de SR (LandSCENT) y actividad (ORIGINS), mientras que las AT2 también mostraron el valor más alto de actividad en general, aunque no el promedio más alto. Las células basales y club se ubicaron consistentemente entre los ocho *scores* promedios más altos para SR, CytoTRACE y actividad (ORIGINS), como se muestra en las Figuras 3.6F-H). Sin embargo, la aproximación actividad HVF (ORIGINS) no logró clasificar adecuadamente los tipos celulares en términos de pluripotencia, lo que podría deberse a que las 2000 características más variables (HVF) no son suficientes para capturar la complejidad de la diversidad celular en el pulmón (Figura 3.6I).

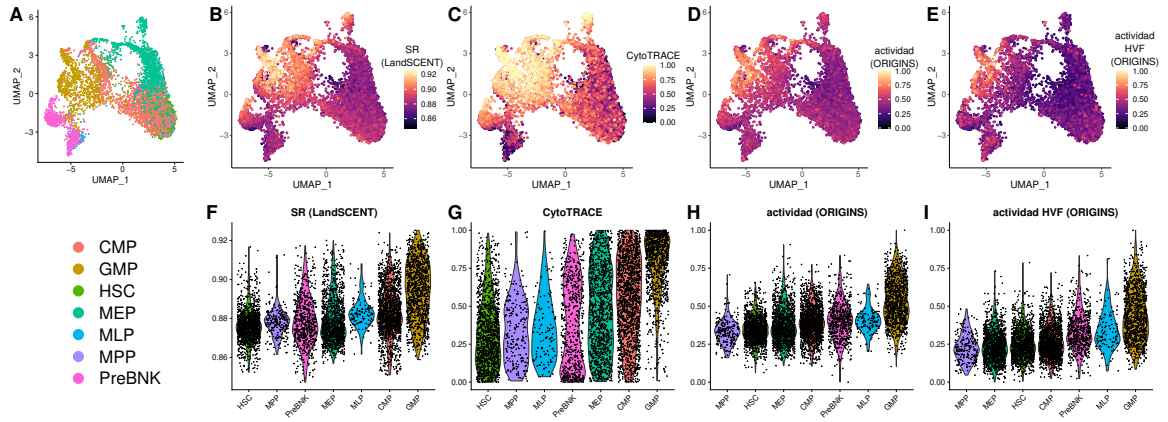

Figura 3.5.: Representación en el espacio UMAP de los datos de scRNA-seq de la muestra de médula ósea. **(A)** Código de colores según el tipo celular. Los tipos celulares se abrevian según sus siglas en inglés: *hematopoietic stem cells* (HSC), *multipotent progenitors* (MPP), *multilymphoid progenitors* (MLP), *pre-B lymphocytes / Natural Killer cells* (PreBNK), *megakaryocyte-erythroid progenitors* (MEP), *common myeloid progenitors* (CMP) y *granulocyte-monocyte progenitors* (GMP) **(B-E)**. Código de colores según el *score* calculado mediante LandSCENT, CytoTRACE, y ORIGINS utilizando todos los genes y los más variables (HVF). **(F-I)** Diagramas de violín de los *scores* de pluripotencia por tipo celular en orden creciente de los promedios.

Correlación con otros métodos

Se calculó el coeficiente de correlación de Pearson entre las metodologías para todos los conjuntos de datos, como se muestra en la Figura 3.7. Se encontró que todas las cantidades estaban correlacionadas positivamente. Considerando los cuatro conjuntos de datos analizados, el coeficiente de correlación promedio entre la actividad (ORIGINS) y SR (LandSCENT) fue aproximadamente $0{,}77$, entre la actividad (ORIGINS) y CytoTRACE $0{,}44$, entre el SR (LandSCENT) y CytoTRACE $0{,}63$, y entre la actividad (ORIGINS) y su aproximación de actividad HVF (ORIGINS) $0{,}67$.

Eficiencia

El tiempo de cómputo de todos los algoritmos y entre todas las muestras se reporta en la Tabla 3.1. En promedio, el SR (LandSCENT) tardó aproximadamente un 25 % más que la actividad (ORIGINS) y CytoTRACE tardó menos del $1\,\%$. Como era de esperar, la actividad HVF (ORIGINS) tomó menos tiempo ($\leq 2\,\%$) que la actividad (ORIGINS).

Uso de RAM

LandSCENT fue el programa que requirió mayor espacio de memoria RAM. Por ejemplo, se necesitó una memoria adicional de $6{,}4$ Gb de RAM para el conjunto de datos de mama, mientras que CytoTRACE requirió $5{,}8$ Gb adicionales. Esto representa un desafío al cuantificar la pluripotencia en muestras de scRNA-seq, que típicamente incluyen miles de células, utilizando computadoras personales estándar. En este aspecto, ORIGINS ofrece una ventaja

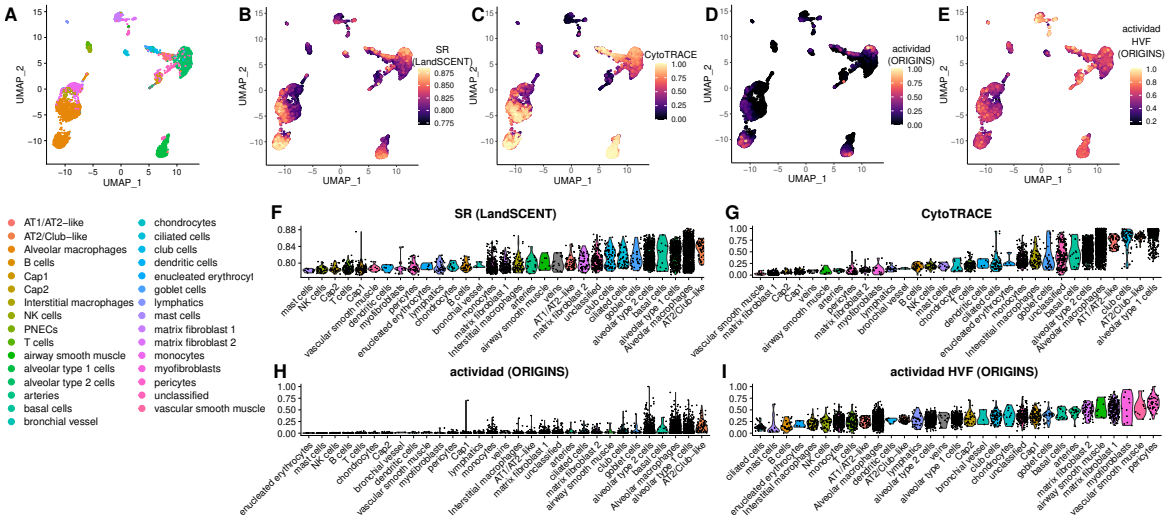

Figura 3.6.: Representación en el espacio UMAP de los datos de scRNA-seq de la muestra de pulmón. **(A)** Código de colores según el tipo celular. **(B-E)** Código de colores según el *score* calculado mediante LandSCENT, CytoTRACE, y ORIGINS utilizando todos los genes y los más variables (HVF). **(F-I)** Diagramas de violín de los *scores* de pluripotencia por tipo celular en orden creciente de los promedios. Se mantuvieron los nombres de los tipos celulares proporcionados por [52].

	SR (LandSCENT)	CytoTRACE	actividad (ORIGINS)	actividad HVF (ORIGINS)
Mama	3,43 h	28,27 s	2,95 h	3,00 min
Colon	4,11 h	44,19 s	3,11 h	3,11 min
Médula ósea	8,12 h	59,95 s	6,24 h	4,09 min
Pulmón	5,37 h	35,00 s	4,71 h	3,19 min

Tabla 3.1.: Tiempo de cómputo de los *scores* SR (LandSCENT), CytoTRACE, actividad (ORIGINS) and actividad HVF (ORIGINS) para los distintos conjuntos de datos.

significativa, ya que no requiere memoria RAM adicional, más allá del vector en el que se almacena la actividad. Para el mismo conjunto de datos, la memoria utilizada por este vector fue 26,6 KB.

Simplicidad y base biológica

El concepto central que subyace al cálculo de la actividad de la PPIN de diferenciación es relativamente sencillo. En resumen, la actividad de diferenciación de una célula es proporcional a la suma de los pesos de las aristas en la PPIN asociada al proceso de diferenciación celular. Estos pesos, que conectan dos transcriptos (nodos), se aproximan mediante la multiplicación de los niveles de expresión de las proteínas asociadas, de acuerdo con la ley de acción de masas. Así, una arista tiene un peso mayor cuando ambos nodos están altamente expresados y, a la inversa, un peso menor si los niveles de expresión son bajos. Al sumar los pesos de todas las aristas de la red, es posible cuantificar la actividad de diferenciación de la célula.

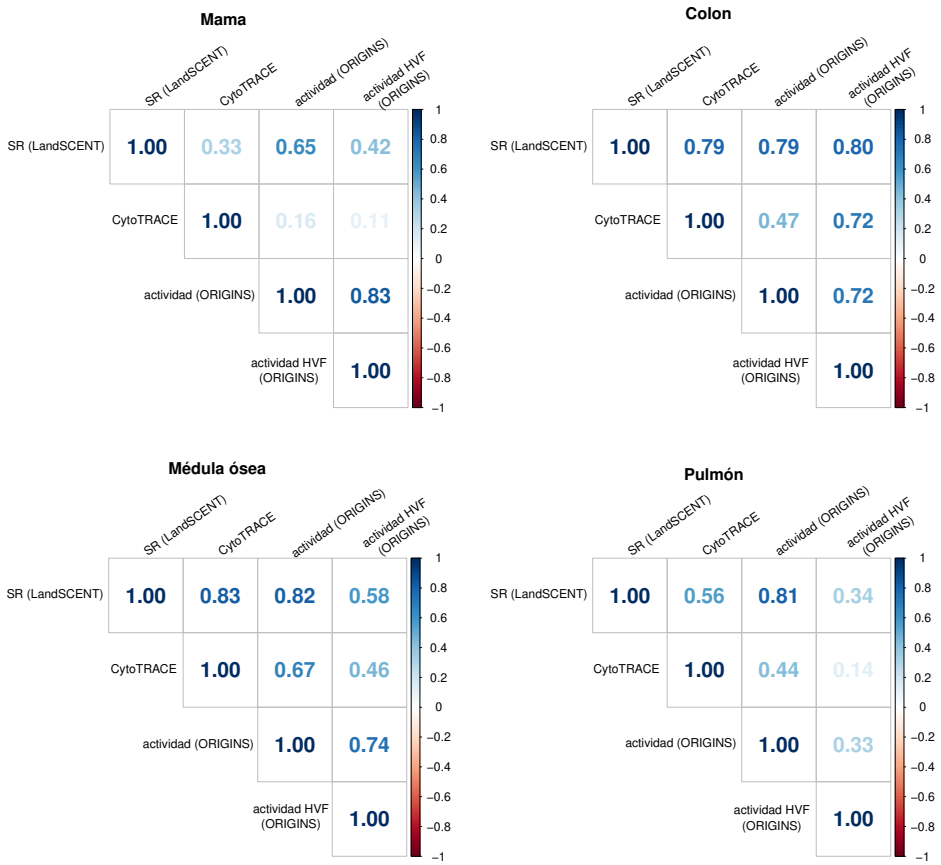

Figura 3.7.: Matrices de correlación entre todas las metodologías utilizadas, SR (LandSCENT), CytoTRACE, actividad (ORIGINS) and actividad HVF (ORIGINS) para todos los conjuntos de datos.

ORIGINS puede manejar matrices de expresión esparsas

Las matrices de expresión normalizadas suelen contener una gran cantidad de elementos nulos (ceros), como es el caso de aquellas generadas mediante la normalización del paquete R Seurat. A diferencia del *score* SR [41], ORIGINS no presenta inconvenientes al manejar matrices que contienen ceros en sus valores. El usuario puede proporcionar cualquier matriz de expresión normalizada no negativa.

User-friendly

El algoritmo de ORIGINS es fácil de utilizar. Con solo escribir cuatro líneas de código en R, se puede calcular la actividad de diferenciación:

```
install.packages("remotes") #if remotes package not installed
remotes::install_github("danielasenraoka/ORIGINS")
library(ORIGINS)
diff_activity <- activity(expression_matrix, differentiation_edges)
```

3.3.2. Aplicación al epitelio de mama: inferencia de la trayectoria de diferenciación

Se aplicó la metodología propuesta, ORIGINS, para identificar células madre en la glándula mamaria humana. A continuación, se describen todos los pasos realizados, desde el procesamiento de los datos crudos hasta la inferencia de la trayectoria de diferenciación. Se utilizó un conjunto de datos de scRNA-seq de células epiteliales mamarias humanas que está disponible públicamente en la base de datos GEO (GSE113197) [9]. Este conjunto de datos fue adquirido utilizando la plataforma 10× Genomics Chromium que comprende aproximadamente 25000 células de cuatro mujeres nulíparas de entre 17 y 36 años, denotadas como Individuos 4 a 7 (Ind4-7). En un trabajo previo, se utilizó la muestra del Individuo 4 (Ind4) para cuantificar la pluripotencia utilizando LandSCENT [41], por lo que, con fines comparativos, se describe el análisis en detalle para esta donante.

Flujo de trabajo del análisis de datos

Se realizó el análisis de datos de scRNA-seq utilizando el pipeline de Seurat. Se descargó la matriz de conteo de UMI y, tras aplicar un filtrado de células y genes, se redujo el ruido y las redundancias. Luego, se normalizaron los datos y se realizó una reducción de dimensionalidad (Figuras 3.9A y B). El análisis de *clustering* y la evaluación de la expresión diferencial permitieron anotar las células (Figuras 3.8A y B). Se identificaron tres clústers principales, que corresponden a células basales mioepiteliales, luminales inmaduras (secretoras y relacionadas con el sistema inmune) y luminales maduras (sensibles a hormonas). Estos tipos celulares coinciden con los descritos en el trabajo original en el cual se publicaron los datos [9] y en estudios posteriores [41]. Siguiendo la notación original, estos grupos se anotaron como Basal, Luminal1 (L1) y Luminal2 (L2). A continuación, se describe en detalle el flujo de trabajo utilizado.

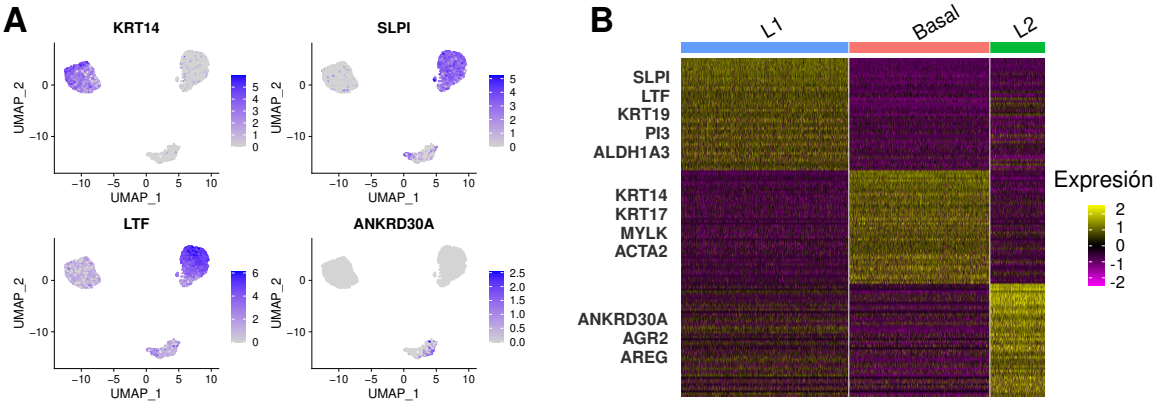

Figura 3.8.: **(A)** Representación UMAP de los datos de scRNA-seq de la muestra de mama. Código de colores según los marcadores típicos. **(B)** Mapa de calor de los genes diferencialmente expresados para los tres tipos celulares.

Filtrado. Se eliminaron células con cantidades atípicas de genes, ya sea excesivas o insuficientes, para descartar células *doublets*, *multiplets*, *debris* y *empty droplets*. También se

filtraron los genes expresados en un número reducido de células, eliminando aquellos que se observan en un número menor o igual a tres células. Además, se descartaron células con alto contenido mitocondrial ($> 5\,\%$) para asegurar la calidad de las muestras. Por último, se filtraron tres pequeños clústers no epiteliales (estromales, endoteliales y valores atípicos), como se describió en el estudio original [9].

Normalización. Para comparar las células entre sí, fue necesario realizar un proceso de normalización. Las cuentas de los genes se dividieron por el total de cuentas de cada célula individual, se multiplicaron por un factor de escala y se transformaron mediante logaritmo natural. Se empleó la función *NormalizeData()* del paquete R Seurat. Posteriormente, se realizó un paso para identificar los genes que presentan una alta variación entre las células (HVF).

Reducción de dimensionalidad. Se escalearon los datos antes de reducir la dimensionalidad del conjunto de datos, de modo que la media de expresión sea 0 y la varianza sea 1 entre las células. Posteriormente, se aplicaron los métodos de reducción de dimensionalidad PCA y UMAP, cuyos resultados se presentan en las Figuras 3.9 A y B. Estos análisis revelaron tres grupos principales de células.

Agrupamiento. Se realizó el *clustering* utilizando las funciones FindNeighbors() y FindClusters() de Seurat. En resumen, esto consiste en la construcción de un grafo KNN (por sus siglas en inglés *k-nearest neighbors*) utilizando la distancia euclídea en el espacio PCA y la aplicación del algoritmo Louvain que optimiza la función de modularidad estándar.

Expresión diferencial. Se realizó un análisis de expresión diferencial de genes para obtener las *signatures* génicas de cada clúster, aplicando la prueba no paramétrica de suma de rangos de Wilcoxon y un *test* ROC (por sus siglas en inglés *Receiver Operating Characteristic*) para evaluar el poder predictivo de los marcadores. En la Figura 3.8A se muestran algunos de los marcadores en el espacio UMAP, y en la Figura 3.8B se presenta un mapa de calor que muestra la expresión diferencial entre clústers.

Anotación celular. El análisis de expresión diferencial proporcionó las *signatures* génicas de cada clúster. Se identificaron tres tipos celulares epiteliales. El clúster naranja/rosa en las Figuras 3.9A y B presentó la expresión diferencial de genes asociados a queratinas como KRT14 (Figura 3.8A KRT14), KRT5 y KRT17. Este clúster también mostró la expresión de genes relacionados con el músculo liso, como ACTA2 y MYLK. ACTA2 es un gen que codifica una proteína de actina, la cual está involucrada en la función contráctil del músculo liso. MYLK codifica una quinasa que fosforila cadenas ligeras reguladoras de miosina para facilitar la interacción de la miosina con los filamentos de actina y producir actividad contráctil en el músculo liso. Considerando esto, este clúster fue etiquetado como un clúster basal mioepitelial.

Los dos clústers celulares restantes fueron positivos para la expresión génica de KRT18 y se identificaron como tipos celulares luminales. El clúster azul en las Figuras 3.9A y B mostró altos niveles de expresión de SLPI y LTF (Figuras 3.8A SLPI y LTF), que son los marcadores típicos de Progenitores Luminales. Así, se etiquetó a este clúster como Luminal 1

(L1). Realizando un análisis de enriquecimiento de conjuntos de genes (Enrichr) sobre los 20 principales marcadores génicos, se sugirió una respuesta inmune y función secretora de este tipo celular luminal inmaduro, que se considera como un tipo celular alveolar. El clúster verde en las Figuras 3.9A y B se pudo identificar como un tipo celular luminal maduro debido a la expresión diferencial del gen ANKRD30A (Figura 3.8A ANKRD30A). Otro marcador génico altamente expresado fue AREG, un factor central en la acción de los estrógenos y en el desarrollo ductal de las glándulas mamarias. AGR2, que es un gen sensible a hormonas, también estaba sobreexpresado. En general, este clúster fue etiquetado como Luminal 2 (L2) y está asociado con una función sensible a hormonas.

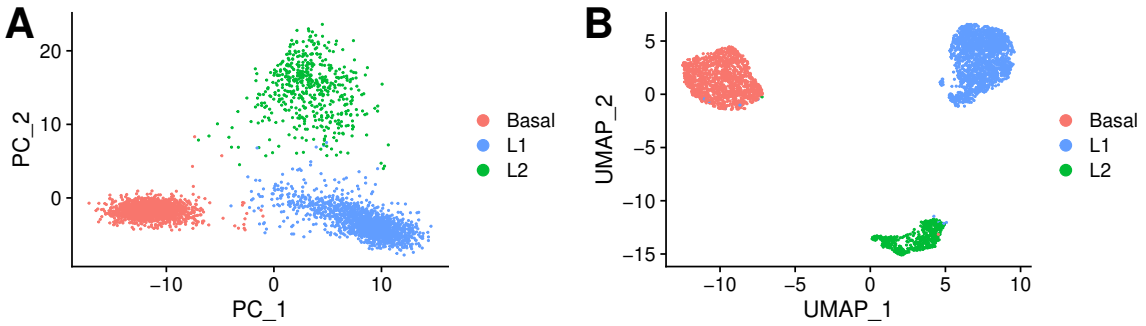

Figura 3.9.: Datos de scRNA-seq de mama en las primeras dos componentes del espacio de PCA **(A)** y UMAP **(B)** coloreadas según el tipo celular.

Cuantificación de la pluripotencia usando ORIGINS

La actividad de la PPIN asociada a la diferenciación se calculó sobre la matriz de expresión génica. La matriz de expresión normalizada utilizada para calcular la actividad no fue la proporcionada por Seurat, ya que este procedimiento devuelve una matriz que contiene elementos nulos. Para poder comparar el rendimiento de ORIGINS con LandSCENT, que no acepta matrices de expresión con elementos nulos como entrada, se aplicó una normalización diferente. Se siguió el proceso de normalización del tutorial de LandSCENT, el cual establece un valor de *offset* de 1,1 antes de la transformación logarítmica para evitar tener valores nulos.

La actividad de diferenciación puede visualizarse en los espacios PCA y UMAP en las Figuras 3.10A y B. Los niveles más altos de actividad se encontraron dentro del grupo basal. Esto está en consonancia con los resultados del trabajo en el que se publicaron originalmente los datos, ya que los autores encontraron un grupo de células basales con capacidad de stemness [9]. De manera similar, en la publicación de LandSCENT, los autores encontraron un mayor porcentaje de células pluripotentes dentro del clúster basal, aunque también observaron altos niveles de SR en células luminales cercanas al clúster basal [41].

Inferencia de trayectoria

La inferencia de trayectorias es un conjunto de técnicas que se utilizan en el análisis de datos de scRNA-seq para inferir jerarquías de linaje. Esencialmente, estos métodos computacionales ordenan las células en función de sus similitudes de expresión a lo largo de una variable

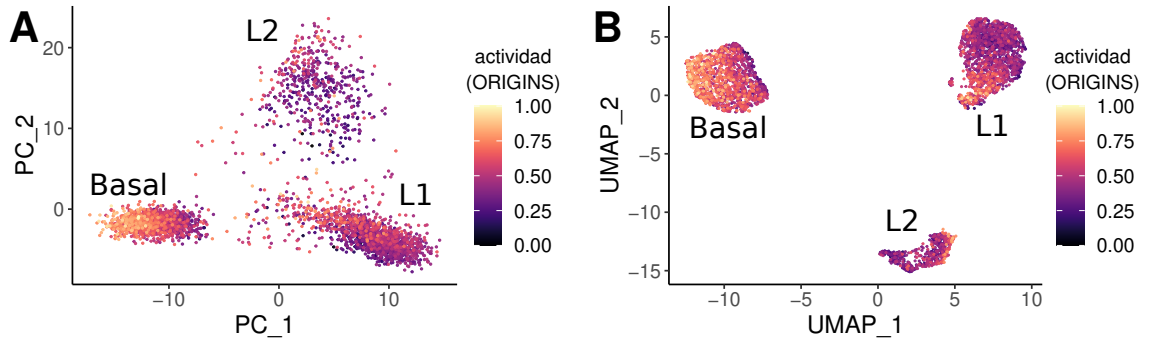

Figura 3.10.: Representación de datos de scRNA-seq de mama sana en los espacios PCA (**A**) y UMAP (**B**) coloreados por la actividad de la PPIN de diferenciación (ORIGINS).

"temporal" llamada pseudotiempo. De acuerdo con Saelens *et al.* [30], entre la gran cantidad de herramientas disponibles, Slingshot [67] se destaca como una de las más simples, robustas y mejor documentadas. Por esta razón, se utiliza Slingshot, un paquete de R, en el conjunto de datos de mama. Slingshot funciona en dos pasos. Primero, infiere la estructura de linaje global utilizando un árbol de expansión mínimo (*Minimum Spanning Tree*) basado en clústers. En una segunda instancia, infiere el pseudotiempo para cada linaje utilizando curvas principales simultáneas. Como muchos métodos, Slingshot requiere la definición previa del origen de la trayectoria, es decir, de células pluripotentes (células madre o progenitoras).

Se estableció el origen como las células con la mayor actividad de la PPIN asociada al proceso de diferenciación, ubicadas dentro del clúster basal. Se encuentra que la trayectoria progresa desde el punto de partida basal y se bifurca en los tipos celulares L1 y L2, pasando por un estado intermedio (Figuras 3.11A y B). Además, este grupo de células precursoras luminales ubicado dentro del clúster L1 tiene niveles de actividad moderadamente altos (Figura 3.10A). En este punto, la trayectoria se ramifica y se dirige hacia las células terminales L1 o L2 en dos linajes separados. Los resultados obtenidos al realizar la inferencia de trayectoria respaldan trabajos previos que sugieren que las células madre/progenitoras de mama son células bipotentes que pueden originar tipos celulares diferenciados basales y luminales [9, 41]. Se denomina Linaje 1 al que comienza en el clúster basal y finaliza en el clúster L1, y Linaje 2 al que termina en el clúster L2.

A continuación, se realizó un análisis de expresión diferencial a lo largo de las trayectorias inferidas en ambos linajes. Se identificaron los genes cuya expresión varió significativamente a lo largo de estas trayectorias, los cuales se visualizan en los mapas de calor presentados en las Figuras 3.12A y B. Para este análisis, se utilizó el paquete de Bioconductor de R, tradeSeq [68]. Brevemente, tradeSeq ajusta un modelo aditivo generalizado negativo binomial (*Generalized Additive Model*) para modelar la relación entre la expresión génica y el pseudotiempo, y luego evalúa relaciones significativas entre la expresión génica y el pseudotiempo.

3.4. Conclusiones

En este capítulo, se ha desarrollado una herramienta para la identificación de células pluripotentes llamada ORIGINS, que se basa en la actividad de redes de interacción proteína-proteína. Se evaluó la *performance* de la metodología comparándola con dos algoritmos disponibles en la literatura en 4 conjuntos de datos humanos. ORIGINS demostró ser eficaz en comparación

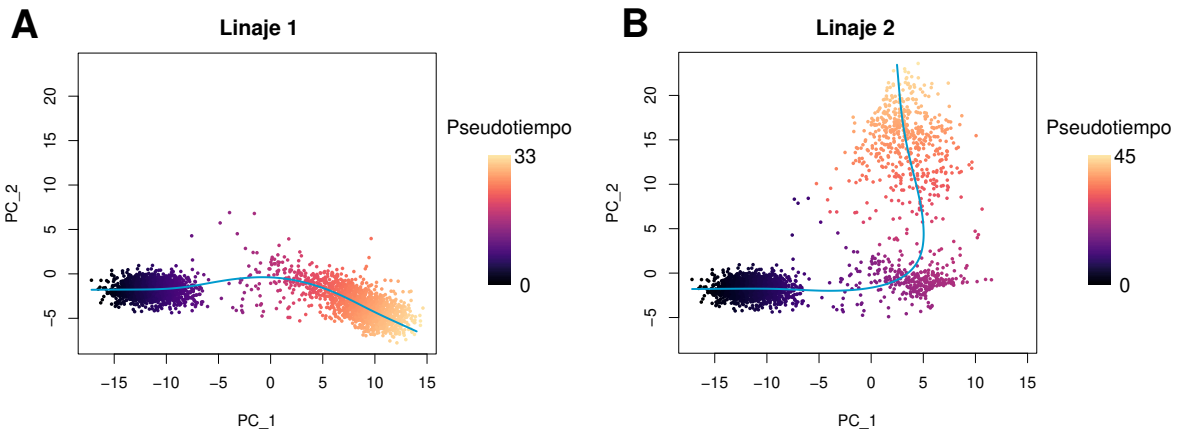

Figura 3.11.: Representación de datos de scRNA-seq de mama en el espacio PCA coloreado por pseudotiempo. La realización de la inferencia de trayectoria revela un camino bifurcado. **(A)** La primera rama, linaje 1, conduce al tipo celular L1. **(B)** La segunda rama, linaje 2, termina en el clúster celular L2.

con otros algoritmos, como LandSCENT y CytoTRACE, ya que no requiere grandes cantidades de memoria y es capaz de manejar matrices de expresión esparsas.

Utilizando un conjunto de datos de scRNA-seq del epitelio mamario, y mediante técnicas de *clustering* y análisis de expresión diferencial de genes, se identificaron tres tipos celulares principales: basal mioepitelial, luminal inmaduro (L1) y luminal maduro (L2). Utilizando ORIGINS, se identificaron células madre en este conjunto de datos. Posteriormente, se aplicó el algoritmo SLINGSHOT para inferir la trayectoria de diferenciación a partir de los datos de expresión celular. SLINGSHOT permitió identificar una bifurcación en la trayectoria que respalda la hipótesis de que las células madre del epitelio mamario son bipotentes; es decir, tienen la capacidad de diferenciarse tanto en tipos celulares basales como luminales. Finalmente, se identificaron los genes que impulsan el proceso de diferenciación a lo largo de las trayectorias inferidas, proporcionando una base para futuros estudios sobre la regulación de la pluripotencia y la diferenciación en el contexto de la biología de la mama humana.

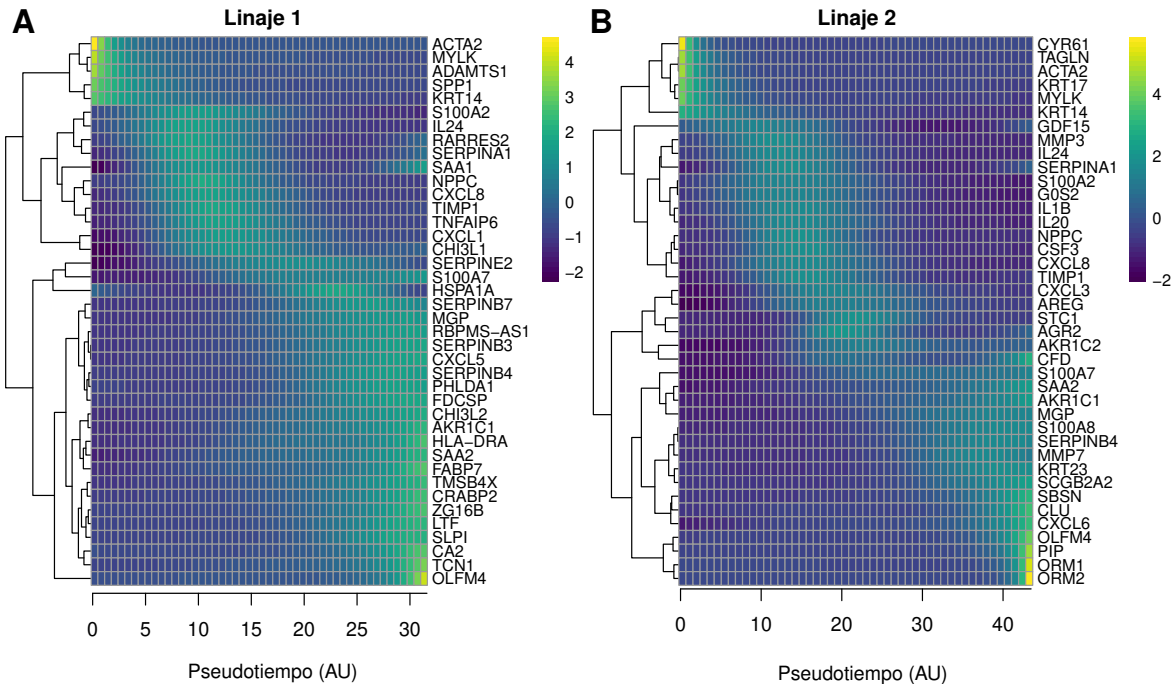

Figura 3.12.: Los 40 genes expresados diferencialmente más significativos a lo largo de la trayectoria para el Linaje 1 **(A)** y el Linaje 2 **(B)** de la muestra de mama.

4

Cuantificación de procesos biológicos en mama sana y cáncer de mama

4.1. Introducción

La técnica de scRNA-seq proporciona datos transcriptómicos de alta resolución, con la capacidad de caracterizar múltiples fenotipos y estados celulares. Sin embargo, identificar diferencias biológicas verdaderas entre células presenta desafíos debido a limitaciones técnicas. En este contexto, el flujo de trabajo típico consiste en identificar clústers celulares en función de la similitud entre los perfiles de transcripción, mediante algoritmos de *clustering*. A continuación, se suele llevar a cabo un análisis de expresión diferencial de genes entre los clústers de células obtenidos previamente. Además, el análisis de enriquecimiento de conjuntos de genes (GSEA, por sus siglas en inglés *Gene Set Enrichment Analysis*), desarrollado originalmente para datos *bulk*, se emplea para obtener un perfil funcional de los conjuntos de genes obtenidos tras el *clustering* y el análisis de expresión diferencial, con el fin de obtener información acerca de los procesos biológicos subyacentes [22, 69, 70].

Debido a la heterogeneidad celular, este enfoque utilizado con frecuencia, puede llevar a una partición excesiva de una población celular uniforme o captar variaciones sutiles, sin un límite claro entre los tipos o estados celulares. Además, debido a la naturaleza esparsa y ruidosa de los datos de scRNA-seq, el análisis de expresión diferencial de genes puede resultar en una caracterización imprecisa de la población celular o producir conclusiones erróneas, incluso utilizando herramientas bien establecidas [71, 72].

Los procesos biológicos que caracterizan a un fenotipo celular específico emergen de la interacción compleja entre distintas clases de moléculas, como el ADN, el ARN, las proteínas y los metabolitos [46]. En los últimos años, se ha incrementado considerablemente la información disponible sobre estas interacciones, lo que ha mejorado la identificación de los elementos que conforman módulos funcionales en diversas condiciones biológicas. Este avance ha permitido el desarrollo de enfoques computacionales basados en redes, capaces de detectar *signatures* transcriptómicas en células individuales, representativas de un estado celular o proceso biológico particular, definidos por la actividad de múltiples genes.

En el capítulo anterior se presentó una herramienta computacional llamada *ORIGINS* [73] que cuantifica la pluripotencia a partir de datos transcriptómicos de células individuales (scRNA-seq), utilizando información de la red de interacción proteína-proteína (PPIN) asociada a la diferenciación celular. Este enfoque permitió estimar la actividad de diferenciación a nivel de célula individual, definida por la dinámica de la PPIN, e identificar poblaciones de células madre y progenitoras.

En este capítulo, con el objetivo de explorar procesos biológicos relevantes en cáncer de mama, se propone una generalización de ORIGINS, orientada a cuantificar no solo la diferenciación celular, sino también otros procesos biológicos de interés, como la proliferación, la migración y el ciclo celular, entre otros. Esta generalización da lugar a una nueva herramienta denominada

ORIGINS2, que permite evaluar la actividad de cualquier proceso biológico o conjunto de genes a partir de datos de expresión génica de células individuales, y utilizando conocimiento de interacciones entre proteínas. Para ello, se plantearán mejoras técnicas, incluyendo la exclusión de reguladores negativos dentro de las PPINs para eliminar falsas contribuciones a la estimación de actividad, y la depuración de interacciones redundantes para optimizar el tiempo de cómputo. Además, se buscará incorporar una funcionalidad que permita construir nuevas PPINs a partir de procesos anotados en la ontología génica o listas personalizadas de genes, permitiendo mayor flexibilidad y ampliando su aplicación a diversos procesos.

Utilizando la herramienta propuesta, se evaluará la actividad de múltiples procesos biológicos en datos de scRNA-seq provenientes de tejido mamario sano y de cáncer de mama triple negativo. Si bien no se identifican clústers bien definidos de células tumorales, la metodología propuesta permite observar una heterogeneidad funcional significativa, con distintos programas biológicos activos en subpoblaciones celulares específicas. Esta capacidad de caracterizar vías activas a nivel celular brinda una herramienta útil y ampliamente aplicable para la exploración de funciones celulares, tanto en condiciones normales como patológicas.

4.2. Metodología

4.2.1. Actividad de una PPIN

Una PPIN es una colección de nodos (proteínas) interconectados mediante aristas (interacciones). Las aristas contienen información sobre los vínculos entre los nodos (pesos) y, en este caso, las aristas no tienen dirección, por lo que la PPIN es un grafo no dirigido. Todas las PPIN utilizadas se construyeron a partir de las interacciones bioquímicas listadas en Pathways Commons (versión 12), que integran $2\,424\,055$ interacciones de 22 bases de datos [49]. Dentro de este conjunto de interacciones, se descartaron aquellas que involucran moléculas no proteicas, como los compuestos químicos. A partir de la lista completa de interacciones proteína-proteína completa, se seleccionaron aquellas correspondientes a productos génicos involucrados en un determinado proceso biológico (BP). Se excluyeron aquellos genes que actúan como reguladores negativos de ese BP y se construyó la matriz de adyacencia asociada a este conjunto de interacciones.

Las listas de proteínas de *H. sapiens* involucradas en cada proceso biológico de interés se descargaron de la base de datos QuickGO [74]. El peso de las aristas se calculó de manera similar a lo realizado en el capítulo para la diferenciación celular, y el nivel de actividad del proceso biológico se define:

$$P = \sum_{i,j=1}^{N_g} A_{ij} x_i x_j. \tag{4.1}$$

x_i y x_j representan los niveles de expresión de los genes i y j, respectivamente. A Es la matriz de adyacencia triangular superior (es decir, A_{ij} es 1 si las proteínas i y j interactúan, y 0 en caso contrario) y N_g es el número de genes en el proceso biológico. A diferencia del algoritmo anterior, aquí se utilizó la matriz de adyacencia triangular superior en lugar de la matriz de adyacencia completa, ya que solo se consideró una interacción entre cada par de

proteínas. Esto evitó redundancias en las aristas y redujo el tiempo de cómputo sin afectar el resultado.

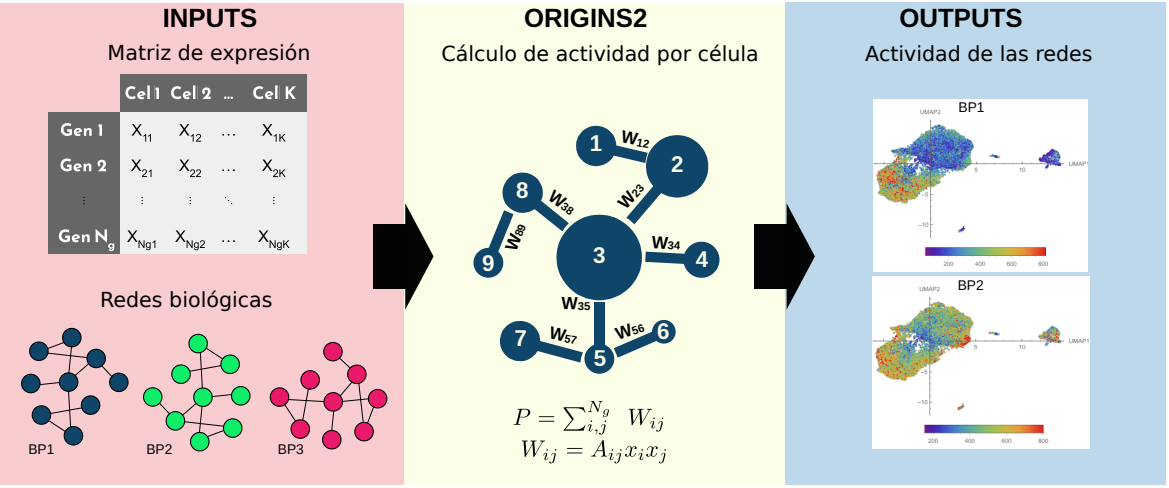

Figura 4.1.: Esquema de la herramienta ORIGINS2.

En la metodología original, se incluyeron todas las proteínas involucradas en la diferenciación celular, el proceso biológico para el cual se diseñó la versión anterior del método. En esta versión, se eliminaron las proteínas que actúan exclusivamente como reguladores negativos del módulo, ya que su sobreexpresión incrementaría erróneamente la actividad del proceso biológico. Estos reguladores negativos fueron identificados seleccionando el proceso de regulación negativa correspondiente en QuickGO [74]. Algunos reguladores negativos actúan simultáneamente como reguladores positivos, por lo que solo se eliminaron de la red las proteínas que actúan exclusivamente como reguladores negativos.

La versión anterior del programa fue diseñada para cuantificar la actividad de la PPIN asociada al proceso biológico de la diferenciación celular, lo cual es particularmente útil para identificar células pluripotentes que pueden ser utilizadas como origen al realizar inferencia de trayectorias. La principal contribución de ORIGINS2 es la capacidad para evaluar la actividad de la PPIN asociada con cualquier proceso biológico o conjunto de genes. La inclusión de PPINs preconstruidas en el paquete de R, asociadas con una amplia gama de procesos biológicos, representó una ventaja importante de ORIGINS2, que amplía el alcance de los posibles análisis. En este contexto, como parte del paquete ORIGINS2 se proporcionaron las PPINs asociadas a la diferenciación celular, el ciclo celular, la proliferación celular, la reparación y replicación del ADN, la respuesta inmune e inflamatoria, la migración celular, entre otros. En la Tabla A.1 del Apéndice A se listan todos los procesos biológicos, junto con las denominaciones originales (Gene Ontology BP, del inglés *Gene Ontology Biological Process*) y el número identificador asociado a cada proceso (*Gene Ontology ID*) tal como se encuentran en la base de datos QuickGO. También se brinda una breve descripción de cada proceso biológico.

Asimismo, dentro de ORIGINS2 se incluyó una funcionalidad que permite construir una PPIN específica para cualquier módulo de interés. Para crear una PPIN personalizada, la función `build_ppin`, incluida en el paquete de R, requiere como entradas una lista de genes asociada con el proceso biológico de interés y la PPIN humana completa, también provista junto con el paquete. Esencialmente, esta función superpone la lista de genes en la PPIN humana completa, seleccionando únicamente aquellas interacciones proteína-proteína cuyos nodos se encuentran en la lista de genes proporcionada por el usuario. El resultado es una lista de adyacencia que contiene un subconjunto de la PPIN humana completa específica para el

proceso biológico en estudio. Esta funcionalidad permite determinar la actividad de la PPIN asociada con cualquier proceso biológico de interés. Para lograr esto, la PPIN previamente creada y la matriz de expresión deben ingresarse como entradas en la función de actividad. Se puede encontrar más información en la documentación detallada provista en el paquete de R.

4.2.2. Descripción de los datos

Se utilizaron datos públicos de scRNA-seq [9, 75], obtenidos en ambos casos mediante la plataforma 10x Genomics Chromium (*droplet-based*).

Del primer estudio, se seleccionó una muestra de mamoplastía de reducción proveniente de una donante caucásica, nulípara y de 36 años de edad. Esta muestra, también utilizada en el capítulo anterior, contiene exclusivamente células epiteliales de tejido mamario normal [9] y se encuentra etiquetada como “ind4“ en la base de datos correspondiente. Las células epiteliales fueron clasificadas en tres categorías principales: Basal, Luminal 1 (L1) y Luminal 2 (L2), de acuerdo con el procedimiento detallado en el capítulo previo.

Del segundo conjunto de datos, se utilizó una muestra de cáncer de mama de grado 3 del subtipo triple negativo (TN), obtenida de una donante de 65 años. Esta muestra está registrada en la base de datos bajo el identificador “TN-0135“ e incluye tanto células tumorales como células del microambiente tumoral [75], además de contar con anotaciones previas.

4.2.3. Disponibilidad de los datos y del código

Los datos analizados en este trabajo se encuentran disponibles públicamente en la base de datos Gene Expression Omnibus (GEO) bajo los códigos de acceso: GSE113197 para células epiteliales de mama y GSE161529 para la muestra de TN.

El método ORIGINS2 fue desarrollado y está disponible como un paquete de R de código abierto y uso gratuito, depositado en el repositorio de GitHub: https://github.com/danielasenraoka/ORIGINS2. En dicho repositorio se incluyó una guía de usuario con instrucciones para la descarga, instalación y uso del programa.

4.3. Resultados y discusiones

4.3.1. Aplicación al epitelio mamario sano

En el capítulo anterior, no se realizó la distinción entre los nodos que actúan como reguladores positivos y negativos. Sin embargo, considerar la expresión de reguladores negativos en la ec. 4.1 conllevaría a sobreestimar la actividad de un módulo biológico específico. Para solucionar este problema, se excluyeron aquellos genes que actúan como reguladores negativos del proceso biológico en estudio. Para ilustrar el efecto de esta modificación, se compararon los resultados de ambos métodos usando datos de scRNA-seq de epitelio mamario sano [9] considerando dos procesos biológicos: diferenciación celular (Figura 4.2) y proliferación celular (Figura 4.3). En el espacio UMAP se muestran los tres tipos celulares previamente

caracterizados: luminal 1 (L1), luminal 2 (L2) y células basales. La visualización de la actividad en el espacio UMAP, tanto con como sin reguladores negativos, presenta resultados relativos similares en la escala de colores. No obstante, como era de esperar, se observan valores absolutos más bajos al excluir los reguladores negativos, dado que el cálculo utiliza una red más reducida en el número de genes. Sin embargo, al comparar las distribuciones de las actividades asociadas a la diferenciación celular y proliferación celular en cada población, las diferencias entre ellas pueden o no ser estadísticamente significativas. Por ejemplo, la Figura 4.3 muestra que las distribuciones de actividad asociada a la PPIN de proliferación celular en células basales y L2 son significativamente diferentes al incluir reguladores negativos, pero no lo son al excluirlos (con un nivel de significancia de $0{,}01$).

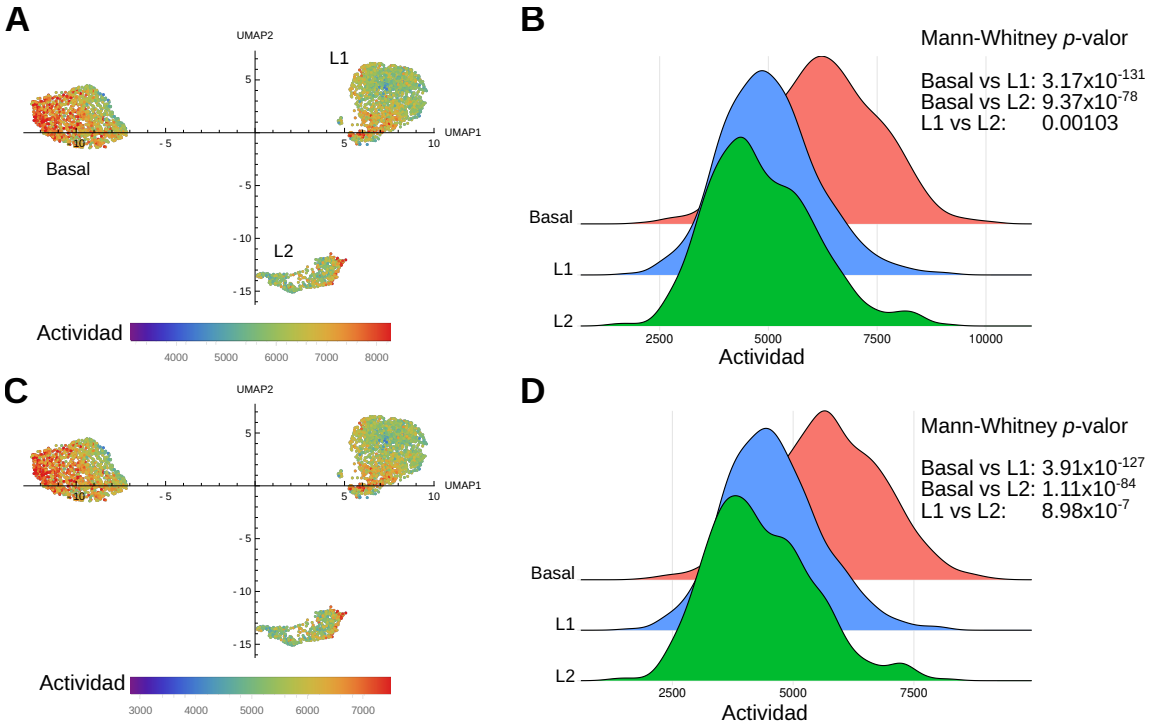

Figura 4.2.: **(A)** Representación UMAP de los datos de scRNA-seq de la muestra de mama sana, código de color según la actividad de la PPIN asociada con todos los genes involucrados en el proceso biológico de diferenciación celular (GO:0030154). **(B)** Estimación de la densidad de probabilidad de las actividades de diferenciación celular por tipo celular, calculada utilizando todos los genes como en el panel A. **(C)** Representación UMAP de la muestra de mama sana, código de color según la actividad calculada excluyendo los genes que regulan negativamente el proceso. **(D)** Estimación de la densidad de probabilidad de la actividad de diferenciación celular excluyendo los genes que regulan negativamente el proceso, como en el panel B. Los resultados de la prueba de Mann-Whitney entre diferentes distribuciones se muestran a la derecha. Las etiquetas L1 y L2 hacen referencia a las células luminales 1 y células luminales 2, respectivamente.

Teniendo en cuenta esta modificación en el cálculo de la actividad de la PPIN, también se evaluaron las actividades asociadas a otros procesos biológicos en el mismo conjunto de datos. La Figura 4.4A muestra la representación UMAP de una muestra de mama sana, coloreada según la actividad asociada con la respuesta inflamatoria aguda. En este caso, los niveles de actividad más altos se concentran en la región superior de las células L1 en el espacio UMAP. En menor medida, algunas células L2 muestran respuesta inflamatoria aguda en el lado izquierdo del grupo L2, lo cual también es visible en la distribución (gráfico verde en la Figura

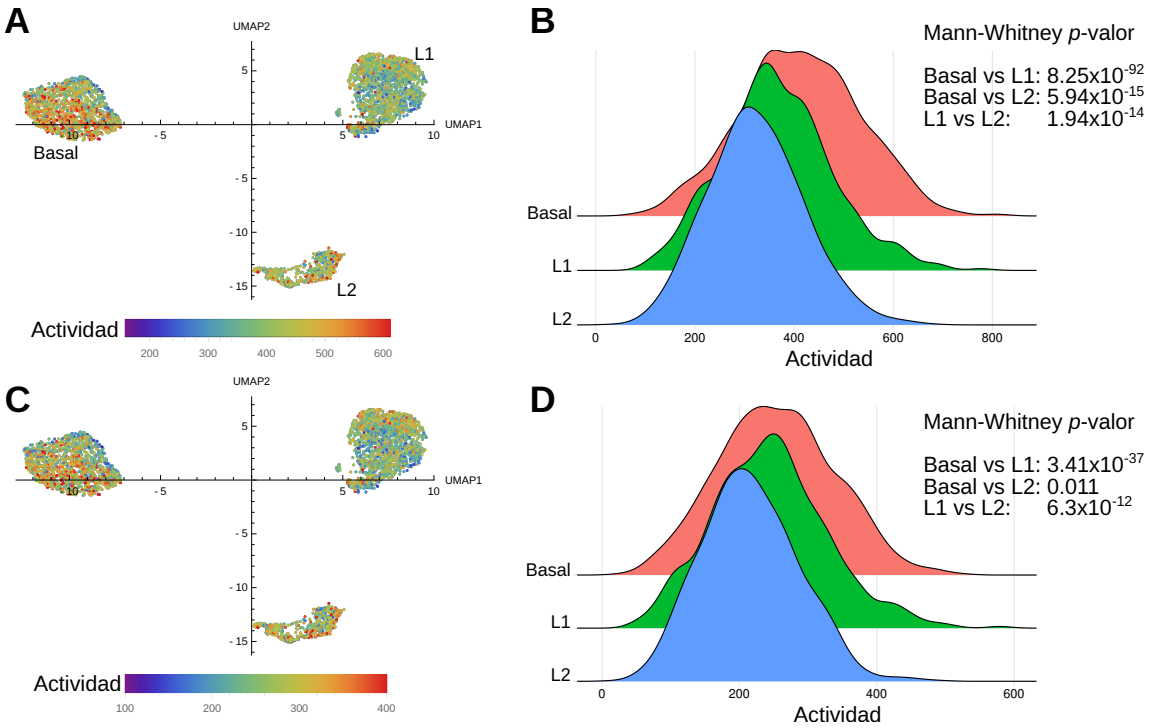

Figura 4.3.: **(A)** Representación UMAP de los datos de scRNA-seq de la muestra de mama sana, código de color según la actividad asociada con todos los genes involucrados en el proceso biológico de proliferación celular (GO:0008283). **(B)** Estimación de la densidad de probabilidad de las actividades de proliferación celular por tipo celular, calculada con todos los genes, al igual que en el panel A. **(C)** Representación UMAP de la muestra de mama sana, código de color según la actividad calculada excluyendo los genes que regulan negativamente el proceso anterior. **(D)** Estimación de la densidad de probabilidad de la actividad de proliferación celular excluyendo los genes que regulan negativamente el proceso, al igual que en el panel B. Los resultados de la prueba de Mann-Whitney entre diferentes distribuciones se muestran a la derecha. Las etiquetas L1 y L2 hacen referencia a las células luminales 1 y células luminales 2, respectivamente.

4.4B). En este ejemplo, es evidente que incluso dentro de un clúster de células, el nivel de actividad de algún módulo biológico puede estar distribuido de manera heterogénea dentro de la población; y que el método propuesto es útil para identificar células o subpoblaciones que llevan a cabo un proceso biológico específico dentro de un clúster. La Figura 4.4C muestra los coeficientes de correlación de Pearson entre actividades de las PPIN asociadas a distintos procesos biológicos calculados en todas las células de este conjunto de datos. Los procesos biológicos analizados pueden agruparse principalmente en dos grupos: uno que incluye respuestas inflamatorias e inmunitarias y otro que incluye la replicación celular y la diferenciación. Estos grupos de procesos biológicos parecen estar anticorrelacionados en esta muestra de mama sana.

4.3.2. Aplicación al cáncer de mama triple negativo

La mama sana contiene una mezcla de tipos celulares que pueden identificarse mediante sus perfiles transcriptómicos distintivos. En el contexto del cáncer, sin embargo, las características de la *signature* tumoral tienden a enmascarar la heterogeneidad intratumoral. Esto dificulta

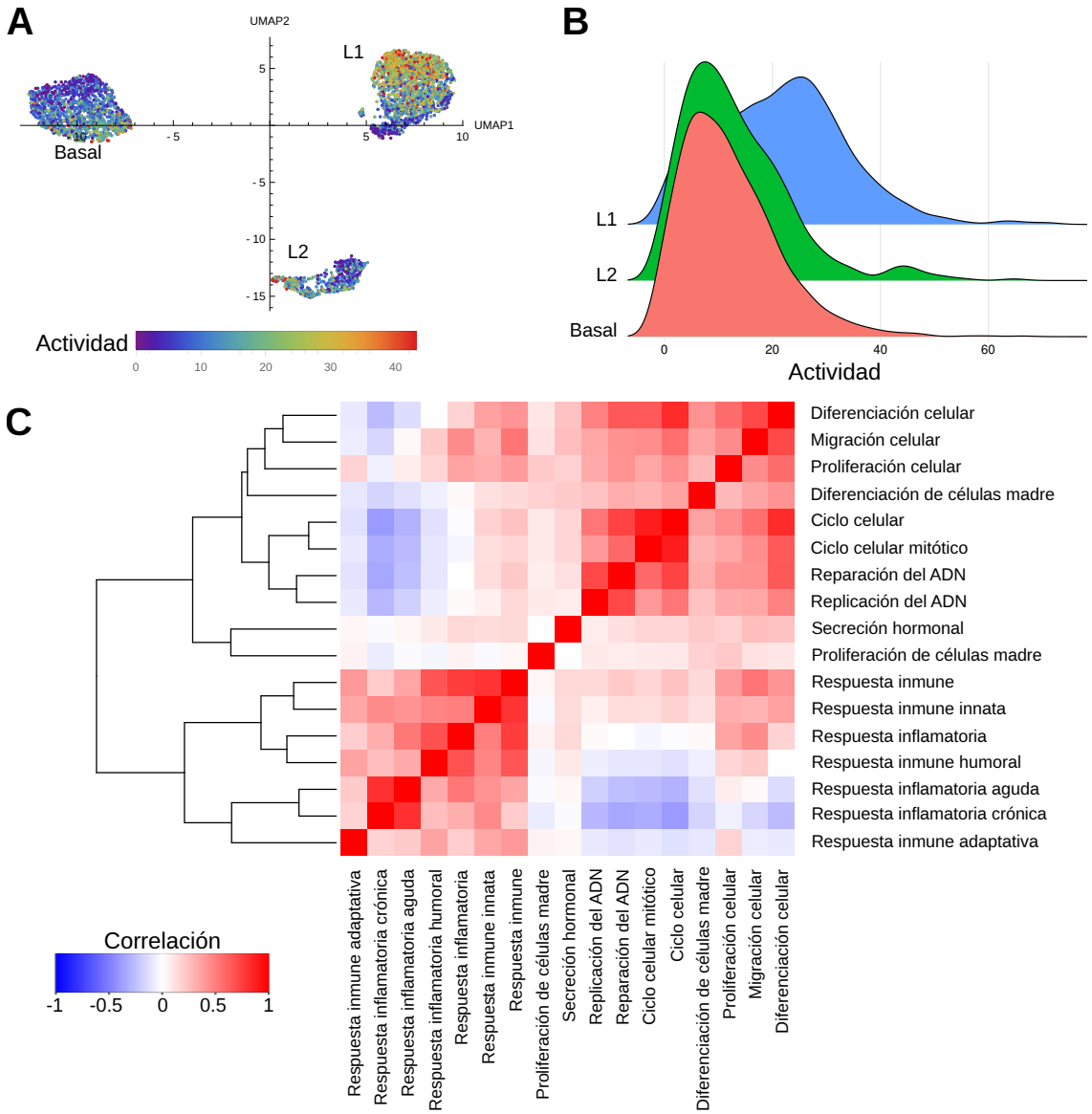

Figura 4.4.: **(A)** Representación UMAP de los datos de scRNA-seq de la muestra de mama sana coloreada según la actividad asociada con el proceso biológico de respuesta inflamatoria aguda, calculada excluyendo los genes que regulan negativamente este proceso. **(B)** Estimación de la densidad de probabilidad de la actividad de la respuesta inflamatoria aguda por tipo celular. **(C)** Coeficiente de correlación de Pearson entre las actividades asociadas con diferentes procesos biológicos calculadas en la misma muestra de mama sana. Las etiquetas L1 y L2 hacen referencia a las células luminales 1 y células luminales 2, respectivamente.

la identificación de subpoblaciones dentro del grupo de células tumorales. Para abordar esta limitación, la herramienta propuesta permitió identificar módulos biológicos que operan en distintas subpoblaciones del clúster de células tumorales.

En la Figura 4.5 se ilustran las actividades asociadas a nueve procesos biológicos en una muestra de cáncer de mama triple negativo (TN) representada en el espacio UMAP. Este conjunto de datos contiene cuatro tipos celulares: células cancerosas (cáncer), fibroblastos asociados al cáncer (FAC), macrófagos asociados al tumor (MAT) y células T (celT). Dentro del

grupo de células tumorales se observa una marcada heterogeneidad, donde ciertos módulos biológicos se encuentran activos en algunos grupos de células. Por ejemplo, mientras que la actividad de diferenciación celular está presente en una amplia proporción de células tumorales, el proceso de migración celular está activo en una pequeña fracción de células. Por otro lado, procesos como el ciclo celular mitótico, la replicación y la reparación del ADN presentan una fuerte correlación y alta actividad en las células que en el espacio UMAP se encuentran en la región inferior izquierda de las células tumorales. Estos hallazgos son consistentes con estudios previos, como el realizado por Pal *et. al* [75], en el que los autores identificaron una subpoblación de células tumorales ciclando en todos los subtipos de cáncer de mama, especialmente en TN. Asimismo, se identifica un grupo pequeño de células tumorales y fibroblastos con alta actividad en la diferenciación de células madre. Finalmente, se observa que, mientras la respuesta inmune innata está activa en células tumorales y macrófagos, la respuesta inmune adaptativa se limita casi exclusivamente a los macrófagos.

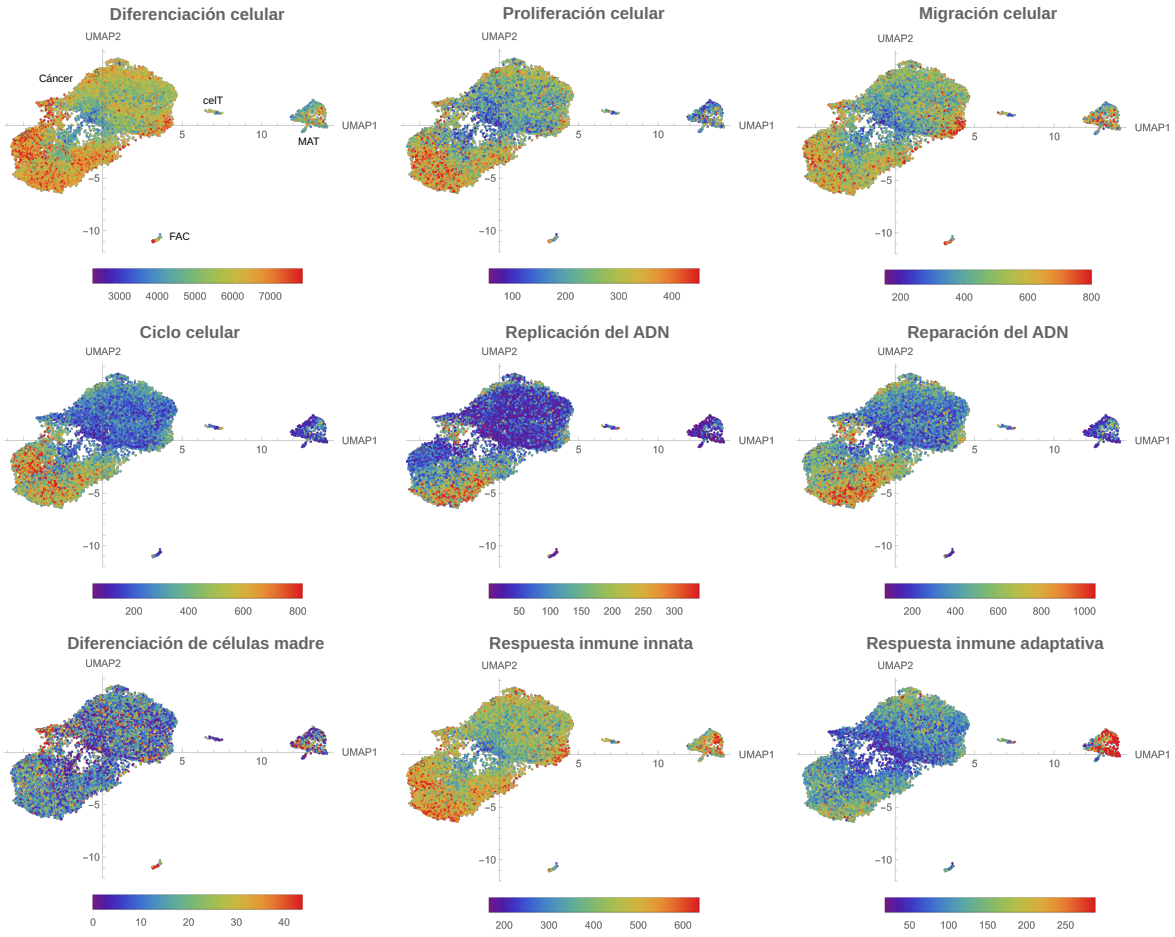

Figura 4.5.: Muestra de scRNA-seq de cáncer de mama triple negativo representada en las primeras dos componentes del espacio UMAP. Código de color según la actividad asociada con nueve procesos biológicos: diferenciación celular, proliferación celular, migración celular, ciclo celular mitótico, replicación de ADN, reparación de ADN, diferenciación de células madre, respuesta inmune innata y respuesta inmune adaptativa. Cáncer: célula tumoral, MAT: macrófago asociado a tumor, FAC: fibroblasto asociado a cáncer y celT: célula T.

La correlación entre las actividades de las PPINs asociadas a los procesos calculados en todas las células de la muestra de cáncer de mama se presenta en la Figura 4.6. En contraste con el

tejido sano (ver Figura 4.4B), la mayoría de los procesos biológicos aparecen correlacionados o poco correlacionados, pero no anticorrelacionados.

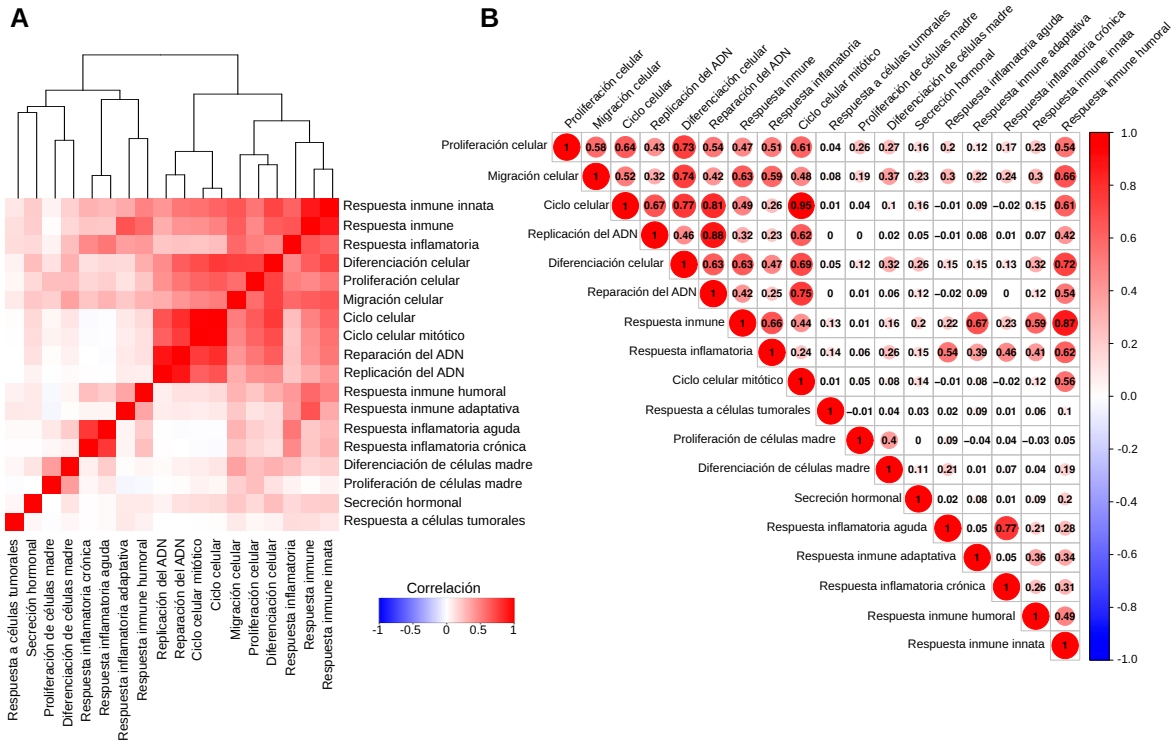

Figura 4.6.: **(A)** Mapa de calor de las correlaciones entre las actividades asociadas con diferentes procesos biológicos calculadas en la misma muestra tumoral de mama. **(B)** Representación alternativa de la matriz de correlación entre las mismas actividades.

Este enfoque resulta especialmente útil para el análisis de datos de scRNA-seq, particularmente en la interpretación de funciones biológicas. Se busca evaluar si las características de relevancia pueden identificarse mediante un análisis directo de redes génicas específicas. El método propuesto permite traducir de manera eficiente los transcriptomas de scRNA-seq en hallazgos biológicos vinculados a una red funcional de interés. El objetivo final es evaluar si las características biológicas más significativas pueden ser elucidadas a nivel célula mediante esta herramienta.

En un flujo de trabajo convencional, la identificación de subpoblaciones a partir de datos transcriptómicos involucra varios pasos en serie: primero, se realiza *clustering* no supervisado para definir los grupos de células, seguido de un análisis de expresión diferencial para obtener los genes marcadores de cada clúster. Finalmente, se lleva a cabo un enriquecimiento de estos marcadores en tipos celulares conocidos y vías biológicas para interpretar el significado biológico de cada grupo. Este enfoque depende críticamente del *clustering* inicial y no permite identificar heterogeneidad en vías específicas a nivel de célula individual, un aspecto relevante en muestras con tipos celulares poco abundantes. ORIGINS2 permite cuantificar la actividad de una PPIN asociada a un proceso biológico en cada célula, sin asumir el número predeterminado de tipos o estados celulares en la muestra.

Proceso biológico	$\rho_{ORIGINS,AUCell}$	$\rho_{ORIGINS,media}$	$\rho_{AUCell,media}$	N_g
Proliferación celular	0.53	0.84	0.49	656
Diferenciación celular	0.39	0.94	0.35	3704
Ciclo celular mitótico	0.68	0.93	0.70	494
Migración celular	0.60	0.83	0.62	914
Respuesta inmune	0.66	0.89	0.63	2414
Diferenciación de células madre	0.46	0.54	0.50	179
Ciclo celular	0.68	0.96	0.67	1367
Replicación del ADN	0.69	0.90	0.73	332
Reparación del ADN	0.62	0.94	0.64	824
Respuesta inflamatoria	0.75	0.84	0.75	581
Proliferación de células madre	0.44	0.52	0.61	64
Secreción hormonal	0.31	0.55	0.61	100
Respuesta inflamatoria aguda	0.65	0.66	0.76	103
Respuesta inflamatoria crónica	0.70	0.77	0.90	13
Respuesta inmune adaptativa	0.84	0.81	0.80	867
Respuesta inmune humoral	0.54	0.62	0.88	415
Respuesta inmune innata	0.53	0.86	0.60	1005

Tabla 4.1.: Coeficiente de correlación de Pearson entre ORIGINS2, AUCell y el promedio de la expresión génica, y número de genes (N_g) de los distintos procesos biológicos estudiados de la base de datos Gene Ontology.

4.3.3. Comparación con otros métodos

Para comparar ORIGINS2 con otras metodologías que realizan enriquecimiento de vías biológicas, se evaluó la expresión promedio a nivel célula del conjunto de genes asociados con el proceso biológico de interés, una práctica común al puntuar los niveles de expresión de un conjunto de genes [41, 76]. También se utilizó AUCell [77], un método computacional que determina el enriquecimiento de conjuntos de genes definidos por el usuario dentro de cada célula. Para evaluar el enriquecimiento, AUCell utiliza el concepto del Área Bajo la Curva (AUC) para determinar si un subconjunto crítico del conjunto de genes ingresado está sobrerrepresentado entre los genes expresados en cada célula. Específicamente, se utilizaron los conjuntos de genes que se encuentran asociados con los procesos biológicos estudiados mediante ORIGINS2. Al proporcionar la matriz de expresión de scRNA-seq y los conjuntos de genes de interés a AUCell, el software calculó un parámetro que indica el nivel de actividad de la vía de interés en cada célula individual. Se computaron estos tres *scores* utilizando la muestra de TN empleada en la Sección 4.3.2.

La Tabla 4.1 muestra el coeficiente de correlación de Pearson entre ORIGINS2, AUCell y la media de todos los conjuntos de genes utilizados en la Figura 4.6A. Los tres métodos presentan correlaciones positivas, con coeficientes de correlación de Pearson que oscilan entre $0{,}31$ y $0{,}96$. Los coeficientes de correlación promedio son $\langle \rho_{ORIGINS,AUCell} \rangle = 0{,}59$, $\langle \rho_{ORIGINS,media} \rangle = 0{,}79$ y $\langle \rho_{AUCell,media} \rangle = 0{,}66$. Los diferentes *scores* de los conjuntos de genes que contienen un número reducido de genes pueden ser sensibles a los *dropouts* y menos estables. Por lo tanto, se calcularon los coeficientes de correlación promedio excluyendo los procesos biológicos con menos de 200 genes, $\left\langle \rho^{N_g>200}_{ORIGINS,AUCell} \right\rangle = 0{,}63$, $\left\langle \rho^{N_g>200}_{ORIGINS,media} \right\rangle = 0{,}86$ y $\left\langle \rho^{N_g>200}_{AUCell,media} \right\rangle = 0{,}66$.

En la Figura 4.7 se comparan los índices AUCell y ORIGINS2 (Actividad) en la muestra de cáncer de mama para tres procesos biológicos: diferenciación celular, proliferación celular y migración celular. En general, ambos *scores* son comparables, aunque AUCell muestra una mayor concentración de actividad en ciertos grupos celulares. En esta figura, se observa que a medida que los coeficientes de correlación entre ORIGINS2 y AUCell aumentan, las representaciones UMAP de los diferentes procesos biológicos se asemejan más entre sí.

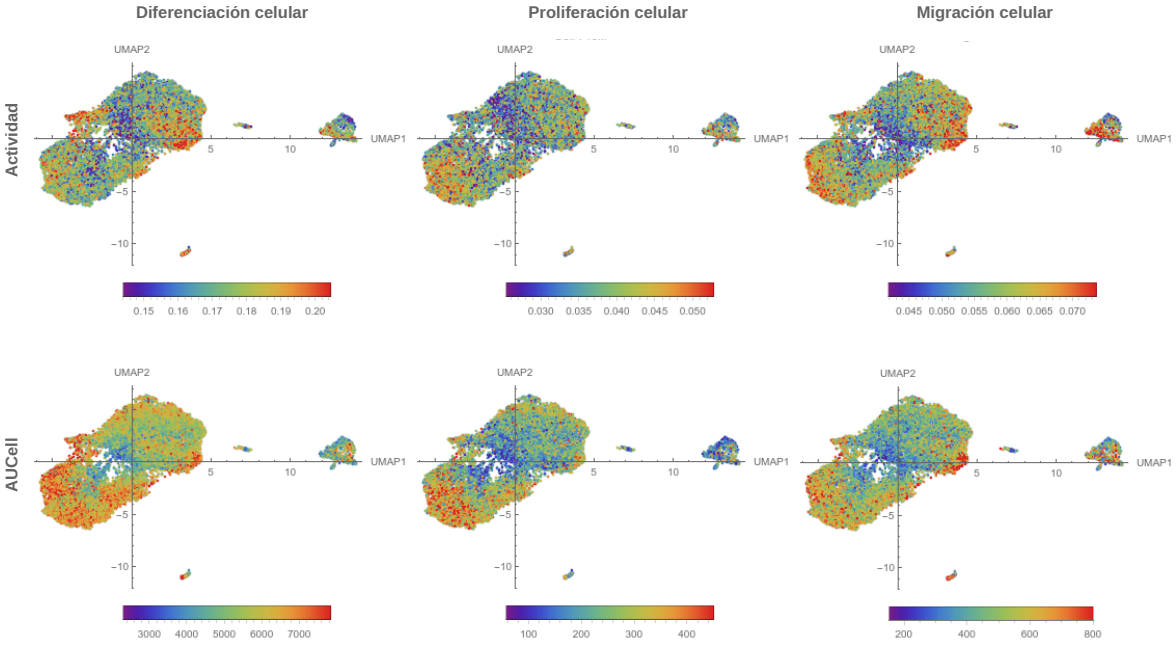

Figura 4.7.: Representación UMAP de los datos de scRNA-seq de la muestra de TN (TN-0135) evaluando tres procesos biológicos: diferenciación celular, proliferación de la población celular y migración celular. **(Actividad)** Código de color según la actividad de la PPIN (ORIGINS2) asociada al proceso biológico correspondiente. **(AUCell)** Código de color según el *score* de AUCell correspondiente al conjunto de genes del proceso biológico en cuestión.

Aunque se encuentra que estos parámetros están correlacionados, es importante reconocer que las metodologías subyacentes para el cálculo son notablemente diferentes. AUCell y el método del promedio utilizan exclusivamente la información de expresión génica (matriz de expresión) para proporcionar un parámetro que es elevado cuando el conjunto de genes de interés está sobrerrepresentado y viceversa. En el caso del algoritmo propuesto en este capítulo, además de ingresar la matriz de expresión, se utiliza información a nivel de proteínas, proporcionada por las PPINs. Por lo tanto, el parámetro calculado con ORIGINS2 refleja el nivel de actividad de la red regulatoria.

ORIGINS2 y los otros métodos evaluados en este capítulo muestran menor robustez en conjuntos de genes pequeños. Esta limitación es atribuible a las pérdidas de datos (*dropouts*) características del scRNA-seq.

4.4. Conclusiones

En este capítulo se presentó ORIGINS2, una generalización del método desarrollado en el capítulo anterior que permite evaluar la actividad de cualquier proceso biológico o conjunto de

genes a partir de datos de expresión génica de células individuales. Además, se incorporaron mejoras como la exclusión de reguladores negativos dentro de las PPINs, la depuración de interacciones redundantes y la posibilidad de construir nuevas PPINs. El método demostró ser eficaz para cuantificar la actividad de redes de interacción de proteínas relacionadas con procesos biológicos como la proliferación, la migración y el ciclo celular, entre otros.

Se aplicó ORIGINS a muestras de scRNA-seq de tejido mamario sano y de cáncer de mama triple negativo. Esto permitió identificar módulos biológicos específicos en distintas poblaciones celulares que no fueron captados mediante el enfoque tradicional basado en *clustering*, análisis de expresión génica diferencial y análisis de enriquecimiento. De esta manera, el método propuesto permitió observar la heterogeneidad dentro de un grupo de células, proporcionando así una herramienta alternativa para el estudio de funciones biológicas a nivel celular.

Comparado con métodos como AUCell y el promedio de expresión génica, ORIGINS2 mostró correlaciones positivas. A pesar de que puede presentar limitaciones en conjuntos génicos pequeños, al igual que los otros métodos, ORIGINS2 ofrece una herramienta potente y adaptable. Esta herramienta es útil no solo en investigaciones de cáncer, sino también en otros estudios de heterogeneidad celular.

Para ampliar la aplicabilidad de la herramienta, se pueden seguir varias líneas de investigación en el futuro, como extender las PPINs a un mayor número de procesos biológicos y calcular actividades relacionadas con procesos biológicos adicionales que puedan ser de interés. En el siguiente capítulo, se explorará esta posibilidad, aplicándola a un mayor número de muestras y a otros procesos biológicos.

5 Heterogeneidad en cáncer de mama

5.1. Introducción

Los tumores son sistemas complejos que se caracterizan por variabilidad genética, transcriptómica, fenotípica y del microambiente. Esta heterogeneidad desempeña un papel fundamental en la metástasis, la progresión tumoral y la recurrencia. El secuenciamiento de ARN a nivel célula, scRNA-seq, que permite obtener el perfil transcriptómico de células individuales, ha introducido nuevas perspectivas en la investigación en cáncer. Esta herramienta ofrece la ventaja de brindar información sobre la composición de poblaciones dentro de una muestra, en comparación con el análisis tradicional conocido como datos *bulk*. En la actualidad, se dispone de diversos conjuntos de datos de cáncer de mama humano, que abarcan desde unos pocos cientos hasta cientos de miles de células [75, 78-80]. Se han realizado una gran cantidad de trabajos para comprender la heterogeneidad celular e identificar tipos y estados celulares mediante análisis de *clustering* a partir de datos de scRNA-seq [25, 81, 82].

El objetivo de este capítulo es evaluar características asociadas con el cáncer a partir de datos de scRNA-seq de cáncer de mama, como la heterogeneidad transcriptómica intratumoral, las alteraciones en el número de copias (CNAs), la entropía y la actividad de redes de interacción proteína-proteína (PPINs) relevantes en cáncer. Se propone cuantificar estas características mediante la definición de *scores* utilizando datos públicos de scRNA-seq de muestras de cáncer de mama humano (ER+, HER2+ y triple-negativo). Aunque estudios previos han examinado estos conceptos por separado en diversos tipos de cáncer, dichos estudios se han basado principalmente en evaluaciones cualitativas o semi-cuantitativas. A continuación, se describen en detalle las características que se abordarán en este capítulo y se brinda una breve revisión general del estado actual de la investigación en cada una de estas áreas.

Heterogeneidad transcriptómica intratumoral. Los tumores poseen una amplia variedad de características fenotípicas y moleculares, tanto a nivel intertumoral como intratumoral. La heterogeneidad intertumoral abarca las variaciones observadas entre distintos tumores, ya sean de diferentes pacientes o del mismo paciente. Además, un solo tumor contiene una mezcla de poblaciones de células tumorales con perfiles genéticos, transcriptómicos y fenotípicos diversos. Este fenómeno, conocido como heterogeneidad intratumoral, es consecuencia de la evolución clonal y de la influencia de diversos factores microambientales [83, 84]. La heterogeneidad intratumoral presenta una ventaja selectiva, aumentando la probabilidad tanto de la preexistencia de poblaciones resistentes como de la capacidad de adaptación [85].

En los últimos años, varios estudios han evaluado la heterogeneidad a nivel transcriptómico en el cáncer de mama utilizando datos de scRNA-seq [75, 78-80, 86]. Estas investigaciones se han centrado en explorar la heterogeneidad mediante marcadores asociados con el cáncer de mama, así como en identificar clústers celulares distintos y sus correspondientes *signatures* de expresión génica. Sin embargo, aunque se han realizado análisis cuantitativos de la

heterogeneidad intratumoral en otros tipos de cáncer [78, 87], en el caso del cáncer de mama aún falta una exploración cuantitativa de esta heterogeneidad.

CNA. Las alteraciones en el número de copias de genes (*Copy Number Alterations*) en células tumorales se consideran un mecanismo que confiere ventaja selectiva dentro de los tumores, resultando en un aumento de la expresión de ciertos genes y en una reducción de la expresión de otros. Las CNAs promueven la progresión tumoral al activar oncogenes mediante ganancias de copias y al inactivar genes supresores tumorales mediante pérdidas de copias. Estas alteraciones en el ADN son una característica distintiva del cáncer en seres humanos, reflejando la extensión y el tipo de inestabilidad genómica única en cada caso [88]. En particular, en cáncer de mama se ha reportado que las CNAs están vinculadas con la progresión del cáncer y un mal pronóstico [89, 90]. Estudios previos han definido índices para cuantificar la extensión de las CNAs a nivel celular a partir de datos de scRNA-seq de cáncer de páncreas y glioblastoma [91, 92]. Sin embargo, en el contexto del cáncer de mama, la inferencia de las CNAs se ha centrado principalmente en discriminar células cancerosas de células no tumorales, sin abordar una cuantificación de su extensión [75, 78].

Entropía. La entropía describe el grado de desorden o incertidumbre en sistemas físicos, y este concepto se ha extendido a diferentes disciplinas. La entropía y otros índices basados en la entropía han sido empleados para cuantificar la heterogeneidad celular y otras características, incluyendo la pluripotencia [38, 93]. En datos de scRNA-seq, la entropía refleja la diversidad o variabilidad en la expresión génica entre células individuales. Las células altamente especializadas suelen expresar un conjunto reducido de genes asociados con su función biológica específica, lo que conduce a niveles de entropía más bajos. En contraste, las células poco diferenciadas, con roles menos definidos, muestran una mayor diversidad de genes activos en simultáneo, dando lugar a niveles de entropía más elevados. Banerji *et al.* calcularon un índice de entropía a partir de datos de tipo bulk en cáncer de mama y observaron que los pacientes con un peor pronóstico presentaban valores más elevados de dicho índice. Además, el *score* resultó ser un predictor más robusto del pronóstico en comparación con otros indicadores pronósticos comúnmente utilizados [94].

Actividad de PPIN. El análisis de enriquecimiento de conjuntos de genes permite explorar la expresión de vías relevantes en datos de scRNA-seq [77, 95]. Procesos como el ciclo celular y la proliferación están ligados a la malignidad en el cáncer de mama, y marcadores asociados a ellos, como MKI67, se emplean como indicadores de pronóstico [96, 97]. Los conjuntos de datos de scRNA-seq en cáncer de mama han revelado clústers de células altamente proliferativas, y la proporción de estas células se correlaciona con la agresividad del cáncer [75, 78]. Por otro lado, en cáncer de mama se ha informado que la transición epitelio-mesénquima (EMT) está asociada con la progresión del cáncer [98]. A diferencia de la proliferación, Pal *et al.* no observaron un clúster de células con expresión aumentada de marcadores de EMT en conjuntos de datos de scRNA-seq de cáncer de mama [75]. Sin embargo, en otro trabajo en el que se utilizaron datos transcriptómicos espaciales de células individuales, se ha identificado un grupo de células contiguas con enriquecimiento en marcadores de EMT [78]. Estas observaciones sugieren que aún resta mucho por explorar sobre el programa de EMT en el cáncer de mama y mejorar los marcadores disponibles [99].

Debido a la alta proporción de valores nulos en las matrices de expresión de scRNA-seq (matrices esparsas), existen desafíos importantes para cuantificar procesos biológicos. Una opción es emplear PPINs asociadas a conjuntos de genes implicados en procesos específicos, como se hizo en capítulos anteriores. Este enfoque integra información regulatoria y proporciona una medida de la actividad de los PPINs, menos sensible al *dropout* característico de este tipo

de datos. Así, se pueden evaluar procesos biológicos de interés, como EMT o ciclo celular, a partir de datos de scRNA-seq.

En este capítulo, se propone un enfoque para la evaluación cuantitativa de estas características, aplicado específicamente a un conjunto de datos de scRNA-seq de cáncer de mama. Se introducen scores diseñados para cuantificar dichas características, explorar las relaciones entre ellas e investigar su asociación y variabilidad en los tres subtipos principales de cáncer de mama (ER+, HER2+ y TN).

5.2. Metodología

5.2.1. Descripción de los datos

El scBrAtlas es una colección de datos de scRNA-seq provenientes de muestras de tejido mamario en condiciones normales, preneoplásicas y cancerosas. Este atlas abarca aproximadamente $430\,000$ de células individuales, obtenidas a partir de 69 muestras quirúrgicas de 55 pacientes.

Las muestras normales fueron obtenidas de mamoplastías de reducción de donantes sin antecedentes familiares de cáncer de mama. Las muestras preneoplásicas corresponden a individuos portadores de la mutación BRCA1, una predisposición genética asociada con un mayor riesgo de desarrollar cáncer de mama. Las muestras cancerosas provienen de tumores primarios de pacientes no tratados e incluyen los tres subtipos principales: ER+, HER2+ y TN. Estas muestras contienen células epiteliales (normales y cancerosas), células del estroma y células del microambiente inmune.

5.2.2. Análisis de los datos

Se obtuvieron los parámetros de control de calidad para garantizar la calidad de los datos. Las muestras fueron sometidas a un proceso de filtrado basado en el tamaño de la biblioteca, el número de genes por célula y el porcentaje de contenido mitocondrial por célula. Como resultado, aproximadamente $15\,\%$ de las células fueron filtradas en cada muestra, dejando un total de $341\,874$ células para el análisis *downstream*. Una descripción detallada del preprocesamiento y el código R correspondiente se encuentra en [100], en la Tabla A.2 del Apéndice A se resumen los detalles del filtrado. Dado que los datos ya estaban preprocesados, se verificó este paso y se utilizaron directamente los datos preprocesados.

El principal interés de este capítulo de la tesis se centra en el estudio de las células cancerosas, por lo que solo se utilizaron las muestras provenientes de mujeres con cáncer de mama. En consecuencia, se excluyeron del análisis subsecuente las muestras provenientes de hombres, así como aquellas correspondientes a tejido sano, preneoplásico o asociado a ganglios linfáticos. La cohorte seleccionada abarca un rango etario amplio, con donantes de entre 25 y 84 años de edad. En la Tabla 5.1 se proporciona una descripción detallada de las muestras, incluyendo la edad de la paciente, el subtipo de cáncer, el tamaño, el grado y el número de células cancerosas. Todas las muestras están compuestas por una mezcla de células tumorales, células epiteliales normales y células del microambiente, tales como fibroblastos, células endoteliales y células inmunes, entre otras. Para el análisis posterior,

ID de la muestra	Edad de la paciente (años)	Subtipo de cáncer	Tamaño (mm)	Grado	Células cancerosas tras filtrado
TN-0126*	64	TN	64	3	1235
TN-0135*	61	TN	22	3	14333
TN-0106	65	TN	25	3	54
TN-0114-T2	84	TN	17	3	177
TN-B1-4031*	25	TN (BRCA1)	20	3	5129
TN-B1-0131*	84	TN (BRCA1)	25	3	5513
TN-B1-0554*	29	TN (BRCA1)	37	3	2337
TN-B1-0177*	30	TN (BRCA1)	13	3	1575
HER2-0308*	32	HER2+	20	3	3317
HER2-0337*	66	HER2+	67	3	3924
HER2-0031*	47	HER2+	18	3	1606
HER2-0069	71	HER2+	27	3	196
HER2-0161*	80	HER2+	45	3	4124
HER2-0176*	60	HER2+	20	3	4682
ER-0319	58	PR+	27	3	568
ER-0001*	58	ER+	32	3	4559
ER-0125*	45	ER+	48	2	3678
ER-0360*	70	ER+	50	2	1934
ER-0032	55	ER+	90	3	417
ER-0042*	58	ER+	18	2	2899
ER-0025*	52	ER+	23	2	4499
ER-0151	49	ER+	35	2	749
ER-0163*	45	ER+	45	3	5378

Tabla 5.1.: Descripción de las muestras de cáncer de mama. Se omitieron las muestras de los pacientes masculinos y aquellas asociadas a ganglios linfáticos, ya que se excluyen para el análisis. Las muestras seleccionadas para análisis, que cumplieron el criterio de contener más de 1000 células tumorales tras el filtrado, se encuentran señaladas con un asterisco (*) junto a sus respectivos IDs. La tabla incluye el ID de las muestras manteniendo los nombres de los IDs originales, edad de la paciente, subtipo de cáncer, tamaño, grado y número de células cancerosas tras el proceso de filtrado. Los metadatos completos se encuentran en la publicación original asociada con los datos en la sección *Supporting Information* como Tabla EV2, EV3 y EV4 [75].

se excluyeron todas las células del microambiente, manteniendo solo las células epiteliales. Luego, se seleccionaron las células cancerosas, excluyendo las células epiteliales normales. Las células fueron etiquetadas inicialmente por los autores de los datos, donde se distinguieron las células cancerosas y normales utilizando inferCNV [76, 101, 102]. Para validar las etiquetas de las células y confirmar la distinción entre células tumorales y normales, se aplicó inferCNV de manera independiente. Tras el filtrado, se excluyeron las muestras que contienen menos de 1000 células tumorales.

El conjunto de datos completo [75] incorpora un total de 19 muestras de cáncer de mama ER+, 6 muestras HER2+ y 8 muestras TN. Tras aplicar los criterios de selección, 6 muestras ER+ (ER-0001, ER-0125, ER-0360, ER-0042, ER-0025 y ER-0163), 5 muestras HER2+ (HER2-308, HER2-0337, HER2-0161 y HER2-0176) y 6 muestras TN (TN-0126, TN-0135, TN-B1-4031,

TN-B1-0131, TN-B1-0554 y TN-B1-0177) fueron consideradas para el análisis posterior. Se mantuvieron los nombres de las muestras proporcionados por los autores de los datos.

Se utilizó R (versión 4.3.1) y el paquete Seurat (versión 4.4.0) para el preprocesamiento y análisis de los datos. Las muestras se integraron por separado según el subtipo utilizando el *pipeline* de Seurat 4 [70, 103]. En las Figuras 5.1A-C se muestran los conjuntos de datos integrados correspondientes a los tres subtipos ER+, HER2+ y TN, respectivamente, donde las muestras se distinguen según el color de las células.

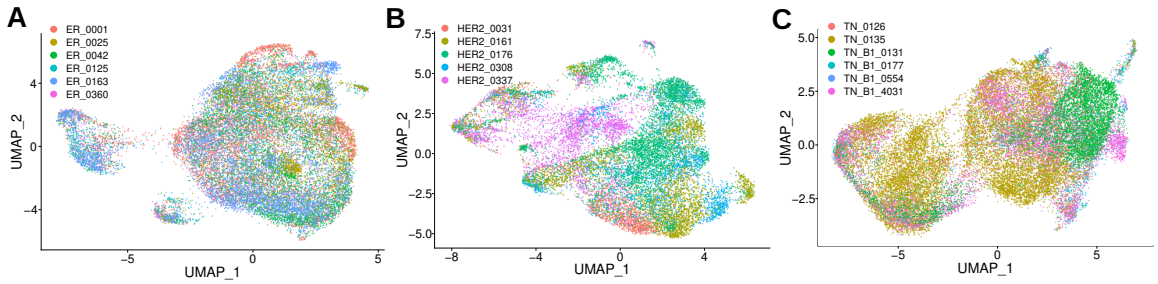

Figura 5.1.: Visualización UMAP de los datos de scRNA-seq correspondientes a las muestras integradas de cada subtipo de cáncer de mama: ER+ **(A)**, HER2+ **(B)** y TN **(C)**. Código de colores de las células según la muestra.

5.2.3. *Scores*

Para investigar la relación entre la agresividad tumoral y ciertas características biológicas específicas, se definieron parámetros para cuantificar la magnitud de estas características. Estos parámetros incluyeron el nivel de CNAs, la heterogeneidad intratumoral, la entropía y la actividad de PPINs de procesos biológicos asociados con la progresión tumoral, como la transición epitelio-mesénquima y el ciclo celular. A continuación, se presentan las definiciones de estas medidas y la justificación de su aplicación.

Grado de CNA

Para inferir las CNAs a partir de datos de scRNA-seq, se utilizó el paquete inferCNV de R (versión 1.16.0), desarrollado por el Trinity CTAT Project [76, 101, 102]. InferCNV permite detectar grandes CNAs somáticos en los cromosomas, como ganancias o pérdidas completas de cromosomas y segmentos cromosómicos considerables. Esto se logra evaluando los niveles de expresión génica relativa de genes contiguos en el genoma y comparándolos con un conjunto de células de referencia "normales". Esta herramienta es ampliamente utilizada para clasificar células tumorales y normales, y generalmente para luego aislar las células tumorales.

En este capítulo, se aplicó inferCNV por separado a las muestras de acuerdo con el subtipo de cáncer. Como población de referencia, se utilizó la muestra de epitelio mamario normal etiquetada como N0372, la cual también fue empleada por los creadores del conjunto de datos [75]. Se utilizó una ventana móvil de 100 genes contiguos. La clasificación preexistente de las células en categorías de cáncer y no cáncer, proporcionada por los autores de los datos, fue validada comparando los perfiles inferCNV obtenidos con las etiquetas preexistentes.

Las herramientas computacionales como inferCNV, copyKAT [104] y CaSpER [105] se emplean comúnmente para estimar CNAs en datos de scRNA-seq. Estas herramientas están diseñadas principalmente para distinguir entre células tumorales y no tumorales, aunque también pueden ser útiles para cuantificar el nivel de CNA dentro de cada célula. Para ello, se definió el grado de CNA para cada célula i en base a la matriz de expresión residual generada por inferCNV:

$$CNA_i = \frac{1}{m} \sum_{j=1}^{m} (X_{ij} - 1)^2 . \tag{5.1}$$

La matriz de expresión residual transpuesta obtenida a partir de inferCNV se denotó como X, conformada por n filas (células) y m columnas (genes). Para simplificar la notación, se evitó el uso explícito del símbolo de transposición. Esta matriz sirve como indicadora de CNA en cada gen j para cada célula i. Un valor de 1 indica neutralidad, valores mayores a 1 representan ganancias y valores menores a 1 denotan pérdidas. Por lo tanto, todos los elementos de la matriz se estandarizaron restando 1. Es importante destacar que, al tratarse de términos al cuadrado, el índice refleja la magnitud sin distinguir entre ganancias y pérdidas. Este parámetro indica la dispersión cuadrática media de los CNAs en relación con las células normales de referencia, y es una variación de índices definidos en trabajos previos [91, 92]. Para cuantificar el grado de las CNAs de una muestra, se promedió el índice *score* CNA en todas las células de la muestra del siguiente modo: $\langle CNA \rangle = \frac{1}{n} \sum_{i=1}^{n} CNA_i$ (5.2).

Heterogeneidad intratumoral

El enfoque clásico para evaluar la heterogeneidad intratumoral en cáncer usando datos de scRNA-seq implica realizar *clustering* para identificar tipos o estados celulares distintos [106, 107]. Posteriormente, estos grupos se caracterizan mediante análisis de expresión diferencial y análisis de enriquecimiento para obtener marcadores específicos para cada grupo. Aunque este método provee información útil sobre la composición de la muestra, no proporciona un índice que mida el grado de heterogeneidad a nivel transcriptómico. Para abordar esto, se propuso una estrategia alternativa para evaluar la heterogeneidad transcriptómica inspirada en trabajos previos [103].

El método consistió en evaluar la variabilidad de la expresión de cada gen en cada muestra de scRNA-seq. El cálculo directo de la varianza a partir de datos log-normalizados no considera la relación entre la media y la varianza típica de este tipo de datos. Para considerar esta dependencia, se aplicó un procedimiento de estabilización de varianza [108, 109]. En particular, se empleó el método implementado en el paquete Seurat, utilizando las funciones *FindVariableFeatures* y *HVFinfo*, las cuales proporcionan un estimador de la varianza ajustado en función de la media de expresión de cada gen [70]. Este enfoque permitió evaluar la variabilidad de los genes teniendo en cuenta la relación entre media y varianza en los datos de scRNA-seq. Para obtener una medida de heterogeneidad en todo el transcriptoma, se definió un parámetro de variabilidad de la expresión génica en una muestra:

$$\langle VAR \rangle = \frac{1}{m} \sum_{j=1}^{m} VAR_j, \tag{5.3}$$

donde VAR_j representa la varianza estabilizada del gen j. $\langle VAR \rangle$ denota la media de varianza entre los genes en una muestra, sirviendo como métrica para evaluar el grado de heterogeneidad intratumoral dentro de la misma.

Este índice cuantifica la heterogeneidad intratumoral calculando la varianza promedio estandarizada en todos los genes expresados. Un *score* bajo indica una alta homogeneidad transcriptómica, lo que significa que las células dentro de una muestra presentan perfiles de expresión génica similares. Por el contrario, valores más elevados indican una mayor heterogeneidad transcriptómica, sugiriendo una mayor diversidad entre las células.

Actividad de redes de interacción de proteínas

En los capítulos anteriores, se desarrolló una metodología para cuantificar la actividad de PPINs a partir de datos de scRNA-seq. En este capítulo, con el objetivo de comparar distintas muestras y PPINs, se introdujeron consideraciones adicionales y se aplicó dicha técnica para calcular la actividad de PPINs asociadas a procesos vinculados con la progresión del cáncer.

Se construyeron las PPINs asociadas a los conjuntos de genes de interés a partir de las interacciones de la PPIN humana completa, que involucran las proteínas codificadas por los genes dentro del conjunto de genes definido x. La actividad de una PPIN para cada célula i se definió de la siguiente manera:

$$ACT_i^x = \frac{1}{e \cdot \mu_i} \sum_{j,k=1}^{m} A_{jk} Y_{ij} Y_{ik}. \tag{5.4}$$

Donde A_{jk} representa la matriz de adyacencia triangular superior, que caracteriza la conectividad entre los genes j y k en la red. Y corresponde a la matriz de expresión normalizada y transpuesta, de dimensiones $n \times m$, e es el número de aristas en la PPIN, y μ_i indica la expresión promedio de todos los genes en la célula i. Cada fila de Y representa el perfil de expresión de la i-ésima célula; en consecuencia, Y constituye la transpuesta de la matriz de expresión convencional, donde las filas corresponden a genes y las columnas a células. Nuevamente, se prioriza simplificar la notación y evitar el uso de símbolos de transposición u órdenes de índices no estándar.

El factor de normalización $e \cdot \mu_i$ se incorporó para corregir las diferencias en el tamaño de los grafos y en la expresión media, permitiendo así realizar comparaciones entre distintas muestras y PPINs. Esta normalización no se implementó en los capítulos anteriores, ya que no era el objetivo en ese contexto. La inclusión de este factor permitió que el *score* de actividad fuera comparable entre redes y células.

Para cuantificar el nivel general de actividad de una PPIN en una muestra determinada, se calculó la actividad promedio considerando todas las células de dicha muestra: $\langle ACT \rangle = \frac{1}{n} \sum_{i=1}^{n} ACT_i$ (5.5). En este capítulo se calculó la actividad de las PPINs asociadas con seis conjuntos de genes de *Homo sapiens*, detallados en la Tabla 5.2. Los genes asociados al proceso biológico del ciclo celular (GO:0007049) se obtuvieron de la base de datos QuickGO [74, 110], y se excluyeron los reguladores negativos de este proceso, tal como se realizó previamente. Los conjuntos de genes restantes se obtuvieron de la Molecular Signatures Database (MSigDB) [116, 117]. Estos incluyeron la regulación positiva de la transición epitelio-mesenquimal (EMT) y cuatro conjuntos de genes asociados a un estudio que caracteriza diferentes líneas celulares de cáncer de mama (basal, luminal y mesenquimal) [112]. Estos conjuntos de genes proporcionaron una visión general de los distintos comportamientos celulares en los diferentes subtipos de cáncer de mama.

ID del conjunto de genes (original)	Símbolo de la actividad	Descripción	Referencia
GOBP *Cell cycle*	ACT^{CC}	GO:0007049. Progresión de eventos bioquímicos y estados morfológicos que ocurren en una célula durante sucesivos eventos de replicación celular o replicación nuclear. El ciclo celular comprende la replicación y segregación del material genético, seguida de la división de la célula.	[74, 110]
GOBP *Positive regulation of epithelial to mesenchymal transition*	ACT^{EMT}	GO:0010718. Cualquier proceso que aumenta la tasa, frecuencia o extensión de la transición epitelio-mesenquimal.	[111]
Charafe breast cancer luminal vs basal dn	$ACT^{LB\text{-}dn}$	Caracterización de líneas celulares de cáncer de mama: Genes diferencialmente expresados en células luminales vs. basales (negativamente).	[112, 113]
Charafe breast cancer luminal vs basal up	$ACT^{LB\text{-}up}$	Caracterización de líneas celulares de cáncer de mama: Genes diferencialmente expresados en células luminales vs. basales (positivamente).	[112, 114]
Charafe breast cancer luminal vs mesenchymal dn	$ACT^{LM\text{-}dn}$	Caracterización de líneas celulares de cáncer de mama: Genes diferencialmente expresados en células luminales vs. mesenquimales (negativamente).	[112, 115]
Charafe breast cancer luminal vs mesenchymal up	$ACT^{LM\text{-}up}$	Caracterización de líneas celulares de cáncer de mama: Genes diferencialmente expresados en células luminales vs. mesenquimales (positivamente).	[112, 114]

Tabla 5.2.: Conjunto de genes para el cual se computa la actividad de la PPIN asociada. ID del conjunto de genes, símbolo que representa la actividad, descripción y referencia a cada conjunto de genes.

Las actividades de las PPINs asociadas con el ciclo celular y la EMT se denotan como ACT^{CC} y ACT^{EMT}, respectivamente. Asimismo, se utilizan los símbolos $ACT^{LB\text{-}up}$, $ACT^{LM\text{-}up}$, $ACT^{LB\text{-}dn}$ y $ACT^{LM\text{-}dn}$ para representar las actividades de las PPINs construidas a partir de los genes diferencialmente expresados, positivamente y negativamente, entre líneas celulares de cáncer de mama luminal vs. basal y luminal vs. mesenquimal.

Entropía

Se computó la entropía de Shannon, una medida bien establecida en teoría de la información que se utiliza para cuantificar el grado de desinformación en la configuración de un sistema. Para ello, se consideró un conjunto de datos de scRNA-seq representado por la matriz de

cuentas transpuesta, denotada como Z. Se calculó la entropía de Shannon para cada célula i de la siguiente manera:

$$H_i = -\sum_{j=1}^{m} p\left(Z_{ij}\right) \cdot \log\ p\left(Z_{ij}\right). \tag{5.6}$$

La entropía está asociada a una función de distribución de probabilidad $p(Z_{ij})$ que representa la probabilidad de que la célula i exprese el gen j. Utilizando datos de scRNA-seq, esta probabilidad se estimó dividiendo la expresión del gen j por la expresión total de esa célula en particular: $p(Z_{ij}) = \frac{Z_{ij}}{\sum_{j=1}^{m} Z_{ij}}$ [93, 118, 119]. Para cuantificar el grado de entropía de una muestra, se calculó el promedio de la entropía de todas las células de una muestra de la siguiente forma: $\langle H \rangle = \frac{1}{n} \sum_{i=1}^{n} H_i$ (5.7).

Tradicionalmente vinculada a la heterogeneidad, en este contexto, la entropía mide la heterogeneidad a nivel de cada célula individual, y no la heterogeneidad intratumoral de la muestra. Un valor alto de H indica que una célula expresa simultáneamente un amplio espectro de genes, una característica típica de células no especializadas, como las células madre o progenitoras [40, 93]. Así, las muestras con valores más altos de $\langle H \rangle$ exhiben características más indiferenciadas asociadas a la pluripotencia.

En resumen, se definieron parámetros para cuantificar los niveles de CNAs, entropía y actividad de PPINs a nivel de célula individual a partir de datos de scRNA-seq. Además, se evaluó la heterogeneidad intratumoral a nivel transcriptómico de la muestra. Los parámetros a nivel de muestra se obtuvieron promediando los valores en todas las células de cada muestra.

5.2.4. Análisis estadístico

Para evaluar la significancia estadística de las diferencias en los valores obtenidos para los subtipos ER, HER2+ y TN, se utilizó un *script* personalizado en el *software* Mathematica Wolfram (versión 13.0), que implementa la prueba de Mann-Whitney sobre las muestras correspondientes a pares de subtipos de cáncer (https://reference.wolfram.com/language/ref/MannWhitneyTest). Esta prueba estadística no paramétrica se empleó para comparar las medias de dos grupos independientes.

Los grupos están conformados por las muestras que pertenecen a un mismo subtipo. Se compararon los valores promedio de las muestras, indicados por $\langle \cdot \rangle$, correspondientes a ocho medidas. Dado que se estudiaron tres tipos de tumores y se analizaron ocho *scores*, se realizaron un total de 24 comparaciones. Por lo tanto, se aplicó el procedimiento de Benjamini-Hochberg [120] para calcular los valores-p ajustados (valores-q) en las pruebas de comparaciones múltiples. Se consideraron estadísticamente significativos aquellos resultados con un $valor - q < 0{,}05$. Los resultados del análisis estadístico se presentan en la Tabla A.3 del Apéndice A.

5.2.5. Disponibilidad de los datos y del código

El conjunto de datos utilizado en este capítulo se obtuvo del scBrAtlas [75], el cual está disponible en dos formatos: matrices de cuentas crudas en la base de datos GEO y objetos de R preprocesados en figshare [121]. Para el análisis que se realizó en este capítulo, se emplearon los objetos de R preprocesados de figshare.

El código desarrollado para calcular los índices y el flujo de trabajo completo está disponible de forma gratuita y como *software* de código abierto en https://github.com/danielasenraoka/R-code-workflow-Unraveling-tumor-heterogeneity, mientras que el código para calcular la actividad de las PPINs fue elaborado en Python y puede encontrarse en https://github.com/danielasenraoka/PyOrigins [122].

5.3. Resultados y discusiones

Se analizaron 17 muestras obtenidas de pacientes con cáncer de mama cuyos detalles se describen en la Tabla 5.1. Las muestras fueron categorizadas según el subtipo de cáncer e integradas (Figuras 5.1 A-C). Se realizaron análisis separados de CNAs para cada subtipo de cáncer. En la Figura 5.2 se muestran los resultados de la inferencia de CNAs mediante mapas de calor, en rojo se indican las ganancias y en azul las pérdidas.

Aunque los patrones de ganancia y deleción no presentan gran sintenia entre los subtipos de cáncer de mama, se pueden destacar algunas características comunes. El chr1 muestra ganancias en chr1q, el cual contiene oncogenes como NRAS, JUN, MYCL, TAL1 y BLYM, y pérdidas en chr1p. Se han reportado ganancias en chr1q en aproximadamente el $60\,\%$ de las pacientes con cáncer de mama, principalmente asociadas al cáncer ER+ [123]. Se observan deleciones en chr2 en casi todas las muestras, independientemente del subtipo. Las amplificaciones en chr8q, que incluyen el proto-oncogén MYC, fueron frecuentes en todos los subtipos, confirmando su papel bien documentado en el cáncer de mama [124-126]. Se observan amplificaciones en chr19 en la mayoría de las muestras, aunque son más pronunciadas en los subtipos ER+ y TN, en concordancia con trabajos previos [127]. En cada subtipo, se notan características comunes entre las donantes. Las muestras HER2+ mostraron amplificaciones consistentes en chr17, particularmente en la banda chr17q12 donde se encuentra el gen HER2, y deleciones frecuentes en chr13. A pesar de que es posible identificar estas alteraciones comunes entre subtipos, se evidencia una notable heterogeneidad dentro de cada uno de ellos, lo que refleja la complejidad y diversidad del cáncer de mama, tal como ha sido previamente reportado [124]. Se destacan las muestras HER2-0308 y TN-B1-0131, que presentan perfiles de CNA que difieren notoriamente de los patrones característicos de sus respectivos subtipos, lo cual será abordado con mayor detalle más adelante.

Para profundizar en la heterogeneidad, se calcularon los índices definidos en la Sección 5.2.3. La distribución de los ocho *scores* a nivel celular puede observarse en los diagramas de violín de la Figura 5.3 para cada muestra. Además, en la Figura 5.4 y B.1 del Apéndice B se visualizan las células representadas en el espacio UMAP integradas por subtipos de cáncer y coloreadas según los *scores*.

Las distribuciones de CNA se ilustran en la Figura 5.3A, donde se pueden apreciar valores más bajos en las muestras ER+ en comparación con las muestras HER2+ y TN (prueba de Mann-Whitney, $valores - q = 0{,}015$ en ambos casos). Además, no se observan agrupaciones

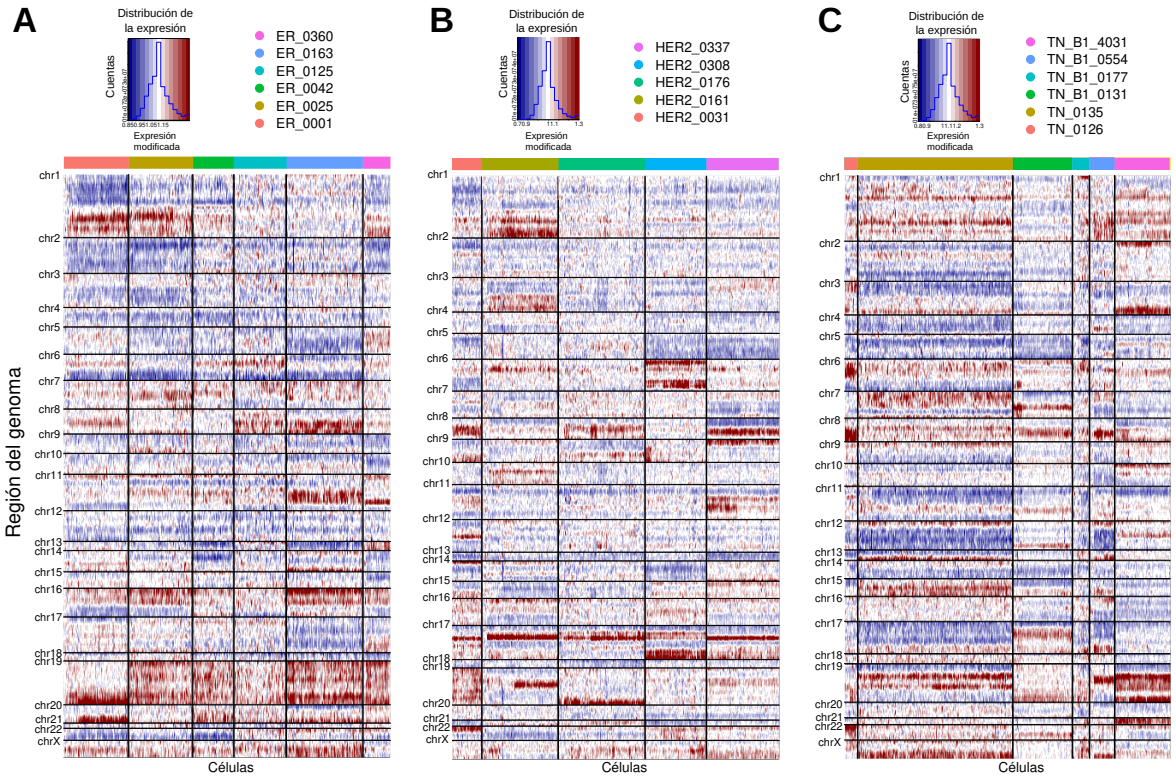

Figura 5.2.: Mapas de calor que muestran las CNAs inferidas, utilizando el paquete inferCNV, para cada subtipo de cáncer, ER+ **(A)**, HER2+ **(B)** y TN **(C)**. Las filas representan los genes ordenados por su ubicación en el genoma y agrupados por cromosomas, mientras que las columnas representan las células individuales agrupadas por muestras y codificadas por color según la muestra. Las amplificaciones están indicadas en rojo y las deleciones en azul. Los histogramas en la parte superior de cada panel muestran la distribución de la modificación en la expresión génica de las células tumorales en relación con las células de referencia normales.

de células en base a este *score* en la visualización de las dos primeras componentes de UMAP de la Figura 5.4. En contraste, como se puede ver en las Figuras 5.3B y C, los índices $\left\langle ACT^{CC} \right\rangle$ y $\langle H \rangle$ muestran un aumento gradual en las muestras ER+, HER2+ y TN. Las muestras HER2+ y TN presentan mayor $\left\langle ACT^{CC} \right\rangle$ en comparación con las muestras ER+ (prueba de Mann-Whitney Test, $valores - q = 0{,}020$). En concordancia con trabajos previos [75, 100], se observan clústers de células con altos valores de ACT^{CC} en todos los subtipos, como se muestra en la Figura 5.4. Los grupos de células tumorales en ciclo, identificados por los autores de los datos mediante el marcador MKI67, coinciden con los grupos de células con niveles elevados de ACT^{CC} en todos los subtipos de cáncer. Además, en consonancia con estos estudios, las muestras TN presentan una mayor proporción de células con niveles elevados de ACT^{CC} en comparación con los subtipos ER+ y HER2+.

La distribución de ACT^{EMT}, representada en la Figura 5.3D, no presenta diferencias significativas entre las muestras. Además, según ACT^{EMT} no se evidencian grupos de células con niveles elevados como se puede ver en la Figura B.1 del Apéndice B, en concordancia con observaciones previas [75].

También se exploró la actividad de las PPINs asociadas a líneas celulares de cáncer de mama [112]. En las Figuras 5.4 y B.1 del Apéndice B se observa que las distribuciones de

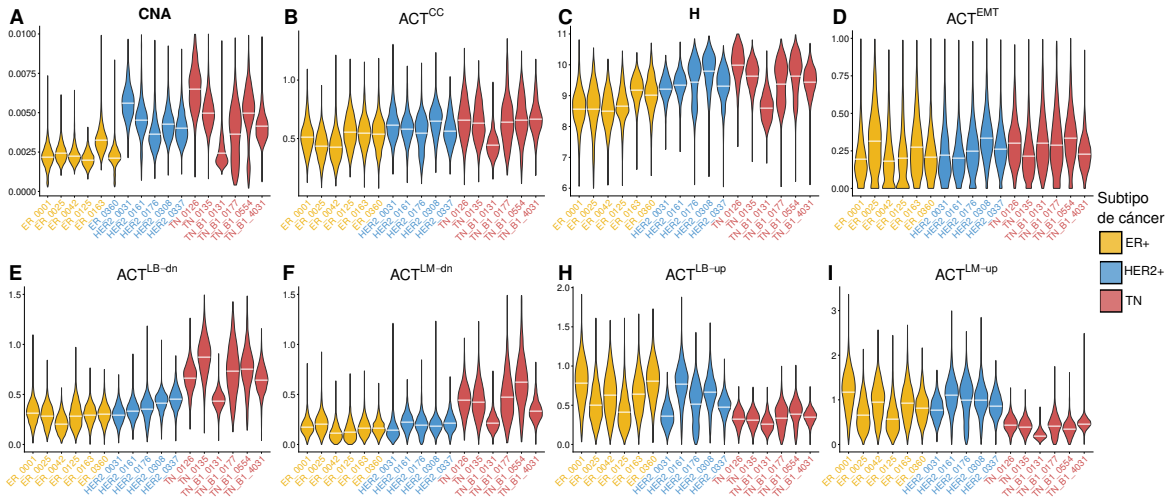

Figura 5.3.: **(A-I)** Distribución celular de los *scores* CNA, ACT^{CC}, H, ACT^{EMT}, $ACT^{LB\text{-}dn}$, $ACT^{LM\text{-}dn}$, $ACT^{LB\text{-}up}$ y $ACT^{LM\text{-}up}$ separados por muestra. Las líneas horizontales bancas representan el valor medio del parámetro correspondiente para cada muestra. Los colores amarillo, azul y rojo corresponden a los subtipos ER+, HER2+, y TN respectivamente.

$ACT^{LM\text{-}up}$ y $ACT^{LM\text{-}dn}$ en el espacio UMAP son similares a las de $ACT^{LB\text{-}up}$ y $ACT^{LB\text{-}dn}$. Como se esperaba, las muestras TN presentan valores más altos de $ACT^{LB\text{-}dn}$ y $ACT^{LM\text{-}dn}$ en comparación con las muestras ER+ (prueba de Mann-Whitney, $valor - q = 0{,}015$ en ambos casos), como se ilustra en las Figuras 5.3E, F y 5.4. Por el contrario, como se puede ver en las Figuras 5.3H, I y 5.4, las muestras ER+ presentan $ACT^{LB\text{-}up}$ y $ACT^{LM\text{-}up}$ significativamente más altos en comparación con las muestras derivadas de tumores basales/mesenquimales, es decir, muestras TN (prueba de Mann-Whitney, $valor - q = 0{,}015$ en ambos casos). De forma similar, las muestras del otro tumor luminal (HER2+) también evidencian valores de $ACT^{LB\text{-}up}$ y $ACT^{LM\text{-}up}$ significativamente más elevados en comparación con las muestras TN (prueba de Mann-Whitney, $valor - q = 0{,}021$ y $0{,}015$, respectivamente). Además, en las muestras ER+ y HER2+ se observan clústers de células con altos niveles de $ACT^{LB\text{-}up}$, mientras que en las muestras TN se encuentran clústers con valores elevados de $ACT^{LB\text{-}dn}$, como se puede apreciar en la Figura 5.4. Curiosamente, las células con características del linaje original se encuentran co-localizadas con células asociadas a una alta entropía H.

Para evaluar la relación entre las distintas características, se obtuvo la matriz de correlación entre los promedios de los *scores* muestrales utilizando el coeficiente de correlación de Pearson, representado en la Figura 5.5A. A nivel de muestra, los promedios de los parámetros $\langle CNA \rangle$, $\left\langle ACT^{CC} \right\rangle$, $\langle H \rangle$, $\left\langle ACT^{LB\text{-}dn} \right\rangle$, $\left\langle ACT^{LM\text{-}dn} \right\rangle$, $\left\langle ACT^{EMT} \right\rangle$ y $\langle VAR \rangle$ revelan correlaciones positivas. Cabe destacar que los primeros cinco parámetros muestran una mayor correlación entre sí, formando un clúster que se obtiene mediante un análisis de agrupamiento jerárquico (hclust del paquete R stats versión 4.3.1), como se observa en la Figura 5.5A. Específicamente, se encuentra que los coeficientes de correlación de Pearson entre $\langle CNA \rangle$, $\langle H \rangle$ y $\left\langle ACT^{CC} \right\rangle$ son los más elevados, con valores entre $0{,}77$ y $0{,}88$. Además, tanto $\langle VAR \rangle$ como $\left\langle ACT^{EMT} \right\rangle$ presentan correlaciones positivas con estos *scores*. En contraste, $\left\langle ACT^{LM\text{-}up} \right\rangle$ y $\left\langle ACT^{LB\text{-}up} \right\rangle$ muestran correlaciones negativas con todos los demás, excepto con $\langle VAR \rangle$.

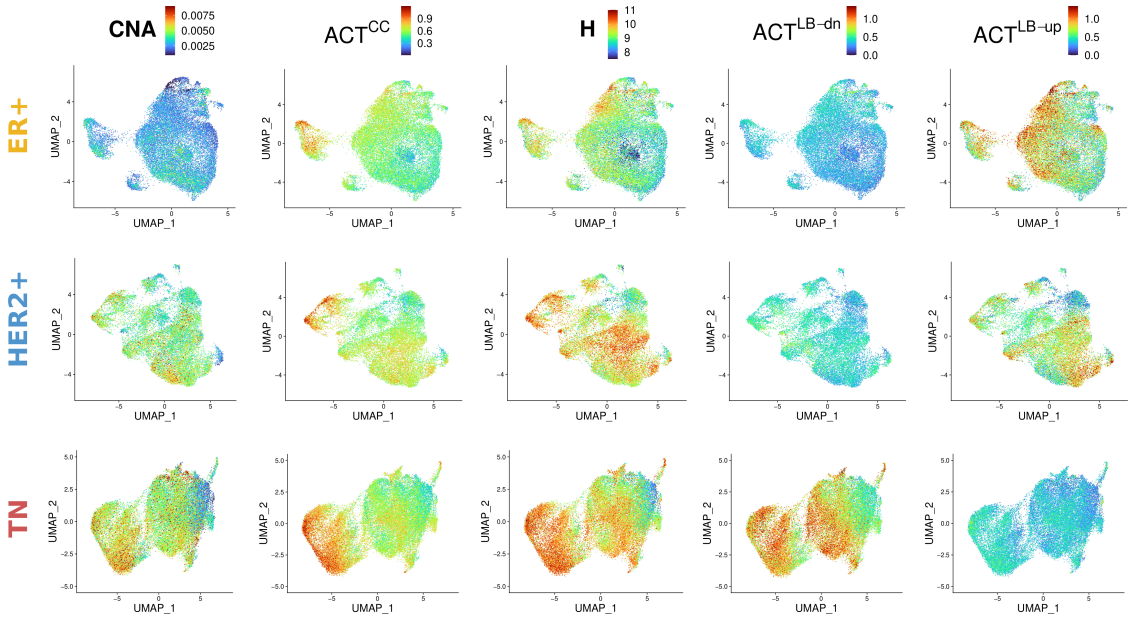

Figura 5.4.: Visualización UMAP de los datos de scRNA-seq correspondientes a las muestras integradas de cada subtipo de cáncer de mama: ER+, HER2+ y TN. Código de colores según los parámetros CNA, ACT^{CC}, H, $ACT^{LB\text{-}dn}$ y $ACT^{LB\text{-}up}$.

Estos resultados sugieren que las muestras con mayor carga de CNAs tienden a mostrar mayor actividad de ciclo celular y entropía. Además, estas muestras manifiestan un fenotipo basal y mesenquimal, lo que podría ser indicativo de la capacidad de las células cancerosas de atravesar una transición epitelio-mesénquima en tumores más agresivos. La heterogeneidad transcriptómica intratumoral $\langle VAR \rangle$ también muestra una correlación positiva con los anteriores; sin embargo, también se observa una correlación positiva con $\left\langle ACT^{LM\text{-}dn} \right\rangle$ y $\left\langle ACT^{LB\text{-}dn} \right\rangle$, aunque con valores menores.

El análisis de la distribución de los nueve índices a nivel de muestra revela patrones distintos en los diferentes subtipos de cáncer de mama. $\langle CNA \rangle$ exhibe mayor heterogeneidad entre las muestras de TN en comparación con las muestras ER+ y HER2+. Además, los niveles de $\langle CNA \rangle$ son más elevados en HER2+ en comparación con ER+, como se visualiza en la Figura 5.5B. Asimismo, $\left\langle ACT^{CC} \right\rangle$, $\langle H \rangle$, $\left\langle ACT^{EMT} \right\rangle$, $\left\langle ACT^{LB\text{-}dn} \right\rangle$ y $\left\langle ACT^{LM\text{-}dn} \right\rangle$ demuestran los niveles más altos en las muestras TN. Las muestras HER2+ presentan niveles intermedios, seguidas por las muestras ER+, como se visualiza en las Figuras 5.5C-G. Esta tendencia coincide con los niveles de malignidad de los subtipos de cáncer, donde los *scores* más altos corresponden a un pronóstico más desfavorable. Teniendo en cuenta los subtipos analizados, existe un consenso general en que el orden pronóstico, desde el más favorable al más desfavorable, es: ER+, HER2+ y TN [128-131]. En las Figuras 5.5H e I se observa el orden opuesto. Este resultado es razonable, ya que TN corresponde a un fenotipo basal/mesenquimal que conduce a una menor actividad relacionada con los tipos luminales, mientras que las muestras ER+ y HER2+ muestran una mayor actividad de estas PPINs. Es importante señalar que estas estimaciones se basan en un número limitado de muestras, lo que lleva a distribuciones estimadas que pueden no representar con precisión las distribuciones reales. Los máximos en las distribuciones estimadas corresponden a muestras que se desvían significativamente del comportamiento general, como se discutirá más adelante.

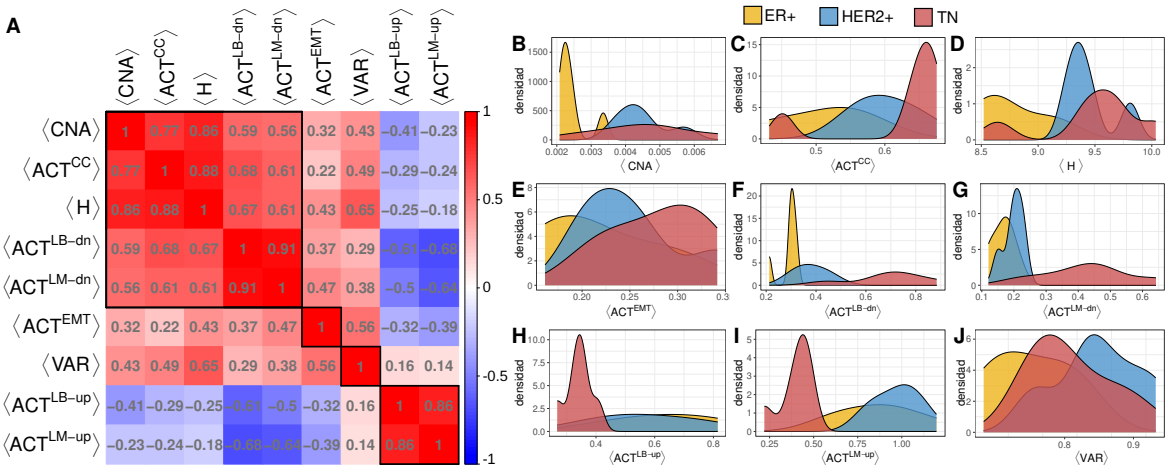

Figura 5.5.: **(A)** Matriz de correlación (coeficiente de correlación de Pearson) entre los nueve parámetros muestrales: $\langle CNA \rangle$, $\langle ACT^{CC} \rangle$, $\langle H \rangle$, $\langle ACT^{LB\text{-}dn} \rangle$, $\langle ACT^{LM\text{-}dn} \rangle$, $\langle ACT^{EMT} \rangle$, $\langle VAR \rangle$, $\langle ACT^{LB\text{-}up} \rangle$ y $\langle ACT^{LM\text{-}up} \rangle$. El color rojo indica una correlación positiva, el azul una correlación negativa, y el blanco representa ausencia de correlación. Los recuadros negros señalan los clústers identificados mediante agrupamiento jerárquico. **(B-J)** Estimación de la distribución de los *scores* muestrales. Código de colores según los subtipos de cáncer: ER+, HER2+ y TN.

En términos de variabilidad de las muestras, ER+ presenta la distribución más sesgada hacia valores más bajos, HER2+ presenta la distribución más sesgada hacia valores más elevados, y la distribución de TN se encuentra en el medio, como se puede ver en la Figura 5.5J. Una posible explicación de esta observación podría ser que los tumores HER2+ presentan tanto características luminales como basales [129], por lo tanto, poseen patrones transcriptómicos más heterogéneos que ER+ y TN. Esta observación sugiere que, aunque la variabilidad se correlaciona positivamente con otros *scores* (aunque en menor medida), este parámetro no necesariamente es un indicador de la agresividad de los tumores. Como se ha mencionado en los capítulos anteriores, en varios tipos de cáncer, incluido el cáncer de mama, se ha identificado un subconjunto de células conocidas como células madre cancerosas (CSCs). Estas células constituyen una pequeña fracción [132, 133], aproximadamente entre el $0{,}1\,\% - 1\,\%$ de la población total de las células tumorales [134] y se ha demostrado que contribuyen a un pronóstico desfavorable. Si bien estas células tienen un rol central, pueden no tener un impacto significativo sobre el *score* de variabilidad transcriptómica debido a su baja abundancia.

En la Figura 5.6 se presentan gráficos de algunos *scores* a nivel de muestra, distinguiéndolos por subtipo de cáncer mediante el uso de diferentes colores y etiquetando las muestras. Estos gráficos permiten visualizar la relación entre los distintos scores, como la correlación positiva reportada en la matriz de correlación de la Figura 5.5A. Además, se aprecian ciertas tendencias generales: las muestras ER+ tienden a agruparse en las regiones inferiores de los gráficos, lo que indica valores de los *scores* más bajos. En contraste, las muestras HER2+ y TN resultan más difíciles de distinguir entre sí, aunque se observa una tendencia de las muestras TN a concentrarse en niveles más altos en comparación con las HER2+.

Al explorar con mayor detalle las muestras individuales, se pueden identificar ciertas particularidades. Por ejemplo, dentro del subtipo TN, la muestra TN-B1-0131 presenta niveles bajos en todos los parámetros evaluados, excepto en $\langle VAR \rangle$, asemejándose a los perfiles típicos de las muestras ER+. Al inspeccionar los metadatos, se observa que esta muestra corresponde a

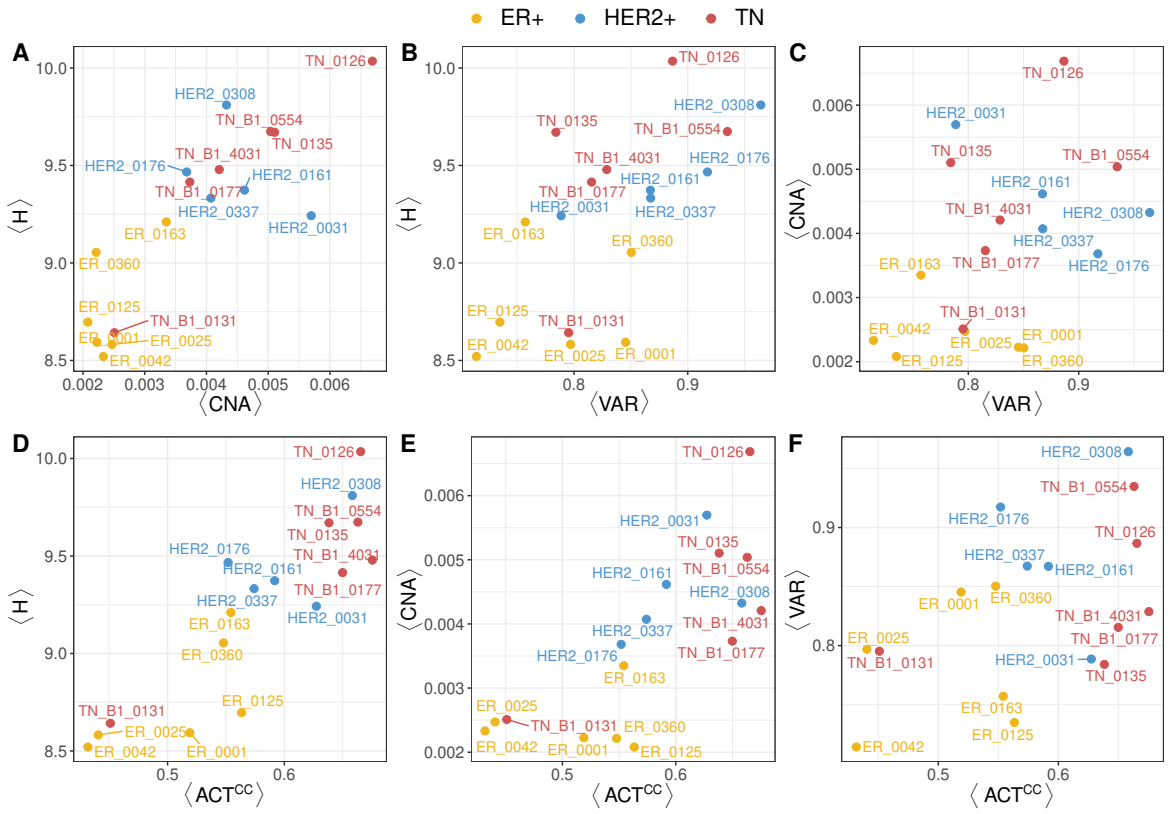

Figura 5.6.: **(A-F)** Relación entre varios parámetros de las muestras donde cada punto representa una muestra. Código de color según el subtipo de cáncer, etiquetas representan el ID de cada muestra.

una paciente de 84 años (ver Tabla 5.1), es decir, 20 años mayor que la siguiente muestra TN al ordenarlas por edad de manera descendente. Además, esta paciente presenta una diferencia de más de 30 años en comparación con la edad promedio de las otras muestras TN, que es de 42 años.

Una observación similar se realiza para la muestra HER2-0308, que presenta *scores* notablemente más altos que el resto de las muestras del subtipo HER2+, con excepción de $\langle CNA \rangle$. Como se detalla en la Tabla 5.1, esta donante tiene 32 años, aproximadamente 30 años menos que la edad promedio de las demás muestras HER2+ (63 años). Estos valores atípicos podrían estar relacionados con la mayor agresividad tumoral en pacientes jóvenes. En este sentido, diversos estudios han reportado que, en pacientes más jóvenes, los tumores tienden a presentar un comportamiento más agresivo, con mayor riesgo de recurrencia, tasas más altas de fracaso de las terapias y mayor mortalidad [135-138].

Las matrices de expresión residual mostradas en la Figura 5.2 proporcionan información complementaria sobre los $CNAs$. A diferencia de otras muestras TN, TN-B1-0131 presenta un valor bajo de $\langle CNA \rangle$. Esto se refleja en su matriz de expresión residual, que se muestra en la Figura 5.2C, y que se caracteriza por un patrón de menores ganancias y pérdidas cromosómicas en comparación con las demás muestras TN. Mientras que HER2-0308 presenta niveles intermedios de $\langle CNA \rangle$ dentro del grupo HER2+, en la Figura 5.2B se puede observar que la matriz de expresión residual se desvía de las otras muestras HER2+. En particular, se observan amplificaciones marcadas en chr6 y chr17, junto con deleciones en chr14.

5.4. Conclusiones

En este capítulo, se presenta un enfoque cuantitativo para evaluar características clave en cáncer. Usando datos de scRNA-seq de muestras humanas de cáncer de mama, se cuantifica el grado de las CNAs, la entropía y la actividad de PPINs asociadas a procesos biológicos específicos (EMT, ciclo celular, líneas celulares luminales, mesenquimales y basales de mama). A nivel de muestra, se calcula el promedio de estas medidas y, además, se introduce un parámetro que cuantifica la heterogeneidad transcriptómica intratumoral de cada muestra. También se explora la asociación entre estas características y los tres subtipos más prevalentes de cáncer de mama: ER+, HER2+ y TN. El trabajo propuesto contrasta con estudios previos que se centraron en aspectos individuales o se basaron en análisis cualitativos basados en marcadores. La novedad de este estudio radica en la evaluación cuantitativa de estas características, un enfoque integral previamente inexplorado en el campo de la transcriptómica de células individuales en cáncer, y en particular en cáncer de mama.

Los resultados a nivel de células individuales revelan patrones interesantes. La actividad de PPIN asociada al ciclo celular y la entropía muestran distintos grados de actividad entre los subtipos de cáncer de mama: ER+, HER2+ y TN, en orden ascendente. Notablemente, se observan clústers de células con actividad mitótica elevada en todos los subtipos, siendo las muestras TN las que presentan la mayor proporción de células mitóticas, indicando una actividad particularmente pronunciada.

Se observan patrones de distribución de CNA distintos entre los tumores ER+ y HER2+/TN. Las muestras provenientes de tumores HER2+/TN presentan valores de CNA significativamente más elevados y una mayor dispersión en comparación con los tumores ER+.

Además, los perfiles de actividad asociados a las líneas celulares basales y luminales utilizados en este trabajo permiten diferenciar entre tumores de tipo basal y luminal. Se identifican algunas tendencias generales, como ganancias en el brazo largo del cromosoma 1 (chr1q), previamente reportadas, especialmente en el subtipo ER+ [123]. Asimismo, las amplificaciones en los cromosomas 8q y 19 respaldan su papel bien documentado en el cáncer de mama [124-126]. A pesar de la presencia de estas alteraciones en el número de copias recurrentes entre distintos subtipos, se observa una heterogeneidad considerable dentro de cada uno de ellos. Este fenómeno, también reportado en estudios previos sobre tumores mamarios [124], refleja la complejidad biológica y la alta variabilidad de esta patología.

No se encuentran grupos celulares con ACT^{EMT} distintivamente elevados en ningún subtipo. Esto podría deberse a la baja proporción de células que experimentan EMT, lo cual puede estar enmascarado dentro de la gran cantidad de células analizadas en total [139]. Además, niveles bajos de expresión de genes relacionados con EMT, como ZEB1, ZEB2 y SNAIL, podrían ser suficientes para desencadenar EMT incluso sin un aumento sustancial en ACT^{EMT}. Otro factor podría ser la ausencia de metástasis en las muestras estudiadas o que las células transicionando hacia un fenotipo mesenquimal están migrando y, por lo tanto, ya no forman parte de las muestras tumorales.

A partir del análisis de los *scores* promedio de las muestras, se observa una correlación positiva entre $\langle CNA \rangle$, $\left\langle ACT^{CC} \right\rangle$, $\langle H \rangle$, $\left\langle ACT^{LB\text{-}dn} \right\rangle$ y $\left\langle ACT^{LM\text{-}dn} \right\rangle$, indicando que las muestras con características basales presentan niveles más altos de estos parámetros. Además, estos índices muestran niveles crecientes en tumores ER+, HER2+ y TN (en orden ascendente), lo cual es consistente con la agresividad conocida de estos subtipos, excepto para $\langle CNA \rangle$. En este caso, $\langle CNA \rangle$ muestra mayores valores en HER2+ y TN en comparación con las muestras

ER+, pero no se aprecia una diferencia significativa entre los tumores HER2+ y TN. Si bien $\left\langle ACT^{EMT} \right\rangle$ correlaciona positivamente con los *scores* anteriores, los coeficientes de correlación son menores. Además, las distribuciones de $\left\langle ACT^{EMT} \right\rangle$ en las muestras ER+, HER2+ y TN mostraron una tendencia de sesgo creciente hacia valores más altos, en ese orden.

Se encuentra una excepción interesante con respecto a $\langle VAR \rangle$, un parámetro que cuantifica la heterogeneidad transcriptómica dentro de cada muestra. En este caso, se observan valores más elevados para las muestras HER2+, seguidos por las muestras TN y, finalmente, por las ER+. Este resultado puede explicarse por el hecho de que los tumores HER2+ poseen propiedades tanto luminales como basales, lo que resulta en una mayor diversidad de perfiles transcriptómicos.

Las medidas cuantitativas revelan un comportamiento peculiar en las muestras HER2-0308 y TN-B1-0131 en comparación con las demás dentro de sus respectivos subtipos. Esta diferencia podría atribuirse a una disparidad en la edad de las pacientes en relación con el rango de edad del resto de las muestras dentro de cada subtipo.

El trabajo de este capítulo presenta varias limitaciones. En primer lugar, el número relativamente pequeño de muestras en la base de datos disponible puede restringir la generalización de los resultados. Si bien la tecnología de scRNA-seq es una herramienta útil para estudiar la heterogeneidad celular, a medida que esta tecnología avance, se disponibilizarán mayores conjuntos de datos de cáncer de mama, lo que permitirá realizar análisis más robustos.

En segundo lugar, la estimación de la actividad de las PPINs depende de bases de datos de interacciones de proteínas existentes, las cuales pueden estar incompletas y no captar todas las interacciones relevantes en la biología tumoral. Esta limitación podría abordarse en el futuro mediante ensayos funcionales que validen el papel de genes específicos en los procesos biológicos de interés.

Por último, el microambiente tumoral (TME, por sus siglas en inglés *Tumoral microenviroment*) ejerce presiones selectivas (por ejemplo, hipoxia, privación de nutrientes, vigilancia inmunológica) que impulsan la evolución de subpoblaciones celulares tumorales [83]. Dado que el TME es espacialmente heterogéneo, estas presiones selectivas generan nichos distintos que pueden favorecer la selección de células tumorales con adaptaciones específicas, lo que contribuye a la heterogeneidad regional dentro del tumor. Además, ciertos componentes del TME, como las citocinas inflamatorias y las especies reactivas de oxígeno, pueden promover inestabilidad genómica en las células tumorales, aumentando la tasa de mutaciones y favoreciendo la generación de subclones [140]. Comprender la compleja interacción entre las células tumorales y el TME es fundamental para obtener una visión integral de la heterogeneidad intratumoral. Si bien el análisis del microambiente puede ofrecer un panorama más completo, esta tesis se enfoca en el estudio de las células tumorales, por lo que dicho análisis queda fuera del alcance del presente trabajo.

Parte II

Redes de regulación génica

Introducción

6

> *Todos los modelos están mal pero algunos son útiles*
>
> **— George E. P. Box**

6.1. Transición epitelio-mesénquima

El concepto de transición epitelio-mesénquima (EMT) fue introducido por primera vez en la década de 1960 por la investigadora Elizabeth Hay, quien observó este proceso experimentalmente y reconoció su importancia durante el desarrollo embrionario [141, 142]. La EMT es un proceso biológico mediante el cual las células epiteliales pierden sus características estructurales y funcionales originales, adquiriendo propiedades y comportamientos típicos de células mesenquimales. Este fenómeno es desencadenado por señales provenientes del microambiente celular y está marcado por una serie de cambios en la expresión génica. Las células epiteliales se caracterizan por su polaridad ápico-basal, morfología poligonal, fuerte cohesión mediante uniones celulares estables, como las uniones estrechas, adherentes y desmosomas y la interacción con la membrana basal. Entre estas características se destaca la expresión de E-cadherina, una proteína clave en las uniones adherentes [143-145].

Durante la EMT, las células epiteliales experimentan la pérdida de estas propiedades debido a la acción de factores de transcripción que reprimen la expresión de genes epiteliales, incluyendo aquellos que codifican proteínas de superficie como la E-cadherina. Como consecuencia, las células reorganizan su citoesqueleto mediante el remodelado de la actina adoptando una morfología fusiforme similar a la de los fibroblastos. Este proceso está acompañado por la expresión de marcadores como la N-cadherina, la vimentina y la fibronectina, que son característicos del estado mesenquimal.

La fase final de la EMT se caracteriza por la degradación de la membrana basal subyacente, lo que permite la migración de las células hacia regiones distantes de la capa epitelial de origen. Este proceso desempeña un papel central en una variedad de contextos fisiológicos y patológicos, incluyendo el desarrollo embrionario, la cicatrización de heridas y la progresión del cáncer.

La EMT se clasifica en tres tipos principales según el contexto biológico en el que ocurre [144].

- Tipo 1: Asociada con procesos fisiológicos como la implantación, formación embrionaria y desarrollo de órganos. Este tipo de EMT da lugar a distintos tipos celulares con fenotipo mesenquimal, los cuales pueden revertir a un estado epitelial y generar epitelio secundario mediante la transición inversa conocida como transición mesénquima-epitelio (MET, por sus siglas en inglés *Mesenchymal-epitelial transition*).

- Tipo 2: Relacionada con la reparación de tejidos en eventos como la cicatrización de heridas, la regeneración y fibrosis de órganos. En estos procesos, la EMT genera fibroblastos y otras células necesarias para la reconstrucción del tejido dañado tras lesiones o inflamaciones. En condiciones normales, esta respuesta cesa cuando la inflamación disminuye; sin embargo, en casos de fibrosis persistente, puede conducir a la destrucción tisular como consecuencia de la inflamación crónica.
- Tipo 3: Ocurre en células neoplásicas que han adquirido alteraciones genéticas y epigenéticas, favoreciendo la formación de tumores y el crecimiento tumoral. Este tipo de EMT está asociado con la invasión y la metástasis en cáncer, siendo un proceso clave en la última etapa de la carcinogénesis, la progresión tumoral. Las células tumorales pueden experimentar una EMT de manera parcial o completa, adquiriendo propiedades mesenquimales que les confieren mayor capacidad migratoria y de diseminación a sitios distantes.

EMT en cáncer

El cáncer es un conjunto de enfermedades complejo y heterogéneo, impulsado por cambios genéticos, epigenéticos y microambientales que permiten a las células evadir los controles normales de crecimiento, proliferar de forma desregulada, invadir tejidos circundantes y formar tumores secundarios en sitios distantes. Entre las características distintivas del cáncer, la metástasis, que es la diseminación de células cancerosas desde el tumor primario hacia órganos distantes, es la principal causa de mortalidad asociada al cáncer, siendo responsable de más del $90\,\%$ de las muertes en pacientes con tumores sólidos [146, 147]. Este proceso comienza cuando algunas células del tumor primario experimentan cambios que les confieren la capacidad de migrar hacia los vasos sanguíneos. A través de un proceso denominado intravasación, las células metastásicas ingresan al torrente sanguíneo. Posteriormente, se diseminan hacia órganos distantes mediante la extravasación. Comprender los mecanismos subyacentes a la metástasis es fundamental para el desarrollo de terapias más efectivas en etapas avanzadas.

Desde que se descubrió que la EMT confiere a las células la capacidad de desprenderse de sus vecinas, migrar e invadir tejidos adyacentes, se ha propuesto que las células cancerosas atraviesan este proceso durante la metástasis [148]. Estudios en múltiples tipos de cáncer han documentado que la pérdida de marcadores epiteliales, como la E-cadherina, junto con el incremento de marcadores mesenquimales, como la N-cadherina, correlacionan con características como la agresividad tumoral, la progresión de la enfermedad y la resistencia a terapias [149].

La existencia y el papel de la EMT en el contexto del cáncer han sido objeto de debate, en particular porque algunos estudios basados en muestras clínicas e imágenes histopatológicas no han identificado células atravesando este proceso. Como se discute en el capítulo 5, el análisis de datos de scRNA-seq realizado en esta tesis tampoco evidenció una presencia significativa de células con niveles elevados de actividad de la EMT en muestras de cáncer de mama. Estos resultados son consistentes con otras investigaciones que han mostrado que el porcentaje de células que experimentan EMT en tumores puede ser extremadamente bajo (aproximadamente $1\,\%$), lo cual dificulta su detección mediante dichas técnicas convencionales [150]. Además, se ha encontrado que esta transición no siempre es completa, ya que en muchos casos las células adquieren estados parciales o híbridos, sin alcanzar un fenotipo

puramente mesenquimal. Estos estudios demostraron que las células que migran y originan la metástasis suelen presentar marcadores de la EMT temprana, mientras que un porcentaje muy pequeño de estas células expresan marcadores de EMT tardía [150].

Evidencia reciente proveniente de análisis de citometría de flujo y de scRNA-seq ha proporcionado pruebas robustas sobre la existencia de estados intermedios a lo largo del eje epitelio-mesénquima en múltiples tipos de tumores, incluidos carcinomas agresivos como los de mama, pulmón y páncreas. Tanto durante la EMT como el proceso inverso (MET), las células pueden ingresar a estados de transición avanzando y retrocediendo de manera reversible a lo largo del eje epitelio-mesénquima, lo que les permite adaptarse y diseminarse en la progresión tumoral, como se esquematiza en la Figura 6.1. Se ha demostrado que las células en estados intermedios en el eje epitelio-mesénquima son plásticas, invasivas y altamente metastásicas [150]. En 2020, la *EMT International Association* (TEMTIA) publicó un consenso para estandarizar el término estado "híbrido E/M" para referirse a un estado a lo largo del eje epitelio-mesénquima, y "plasticidad epitelio-mesénquima" para describir la capacidad de las células de cambiar libremente entre estos estados [143]. En esta tesis, se adoptará esta nomenclatura para referirse a los estados híbridos. Estos trabajos, entre otros, han contribuido a dilucidar el papel de la EMT en la metástasis, destacando la importancia de los estados híbridos epitelio/mesénquima (E/M) como impulsores clave de la heterogeneidad tumoral, la plasticidad celular y la progresión del cáncer.

Comprender cómo las células transitan entre estos estados requiere el estudio de los mecanismos moleculares que controlan la EMT. Este proceso está regulado por una compleja red de factores de transcripción, vías de señalización y ARN no codificantes. Entre los factores de transcripción clave se encuentran las familias SNAIL, ZEB y TWIST, que reprimen la expresión de marcadores epiteliales como la E-cadherina y promueven la expresión de genes mesenquimales como la N-cadherina y la vimentina. Estos factores de transcripción son activados por señales externas presentes en el microambiente tumoral, como el factor de crecimiento transformante β (TGF-β), WNT y Notch. Entre estos, el TGF-β es considerado el inductor más potente de la EMT, el cual puede iniciar una cascada de señalización que refuerza el fenotipo mesenquimal y promueve la invasión [151].

Por su parte, los microARNs (miARNs), pequeñas moléculas de ARN no codificante que regulan la expresión génica de manera postranscripcional, desempeñan un papel fundamental en la regulación de la EMT. La familia miR-200, que inhibe directamente a ZEB1 y ZEB2, actúa como un importante supresor de la EMT, manteniendo características epiteliales. De forma similar, la familia miR-34 modula la EMT mediante la inhibición de SNAIL. En cáncer, menor disponibilidad de estos miARNs conduce a la estabilización del fenotipo mesenquimal.

6.2. Pluripotencia y células madre

Las células madre son células que se caracterizan por su capacidad de autorrenovación y diferenciación en diversos tipos celulares especializados. Estas propiedades son fundamentales en la fase de desarrollo, el mantenimiento de la homeostasis tisular, la regeneración y la reparación a lo largo de la vida de un organismo. En términos generales, las células madre se clasifican en células madre embrionarias, que son pluripotentes y pueden dar lugar a todos los tipos celulares del organismo, y células madre adultas, que son multipotentes y contribuyen al mantenimiento específico de los tejidos. Estas células han sido ampliamente estudiadas en el marco de la biología del desarrollo y de la medicina regenerativa.

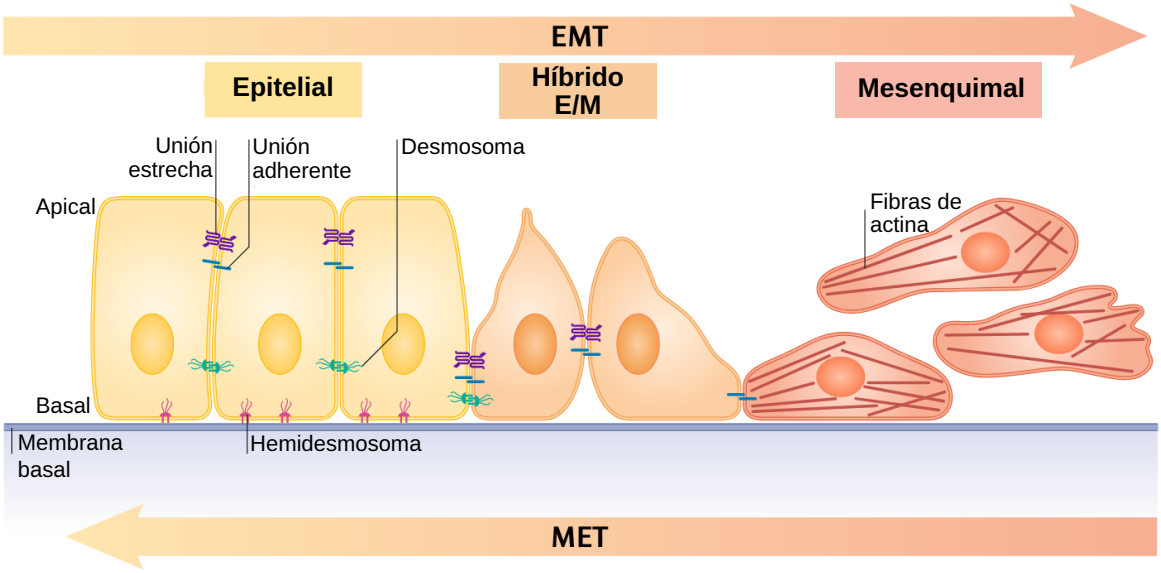

Figura 6.1.: Esquema del proceso de EMT en el que se ilustra la progresión de un fenotipo epitelial a uno mesenquimal atravesando un estado intermedio híbrido E/M. Las células epiteliales presentan polaridad apical-basal, con la región apical en contacto con la luz del tejido y la región basal anclada a la membrana basal mediante hemidesmosomas. Estas células se encuentran unidas entre sí mediante uniones estrechas, uniones adherentes y desmosomas, que otorgan cohesión y organización tisular. A medida que avanza la EMT, estas estructuras de unión se desensamblan, las células pierden su polaridad y adoptan una morfología del tipo huso, junto con una mayor capacidad migratoria y la reorganización del citoesqueleto de actina, características propias del fenotipo mesenquimal. Las flechas horizontales indican la direccionalidad de los procesos de EMT (epitelio a mesénquima) y MET (mesénquima a epitelio). Adaptación de [145].

El comportamiento de las células madre se encuentra regulado por diversos factores de transcripción. Entre estos se encuentran OCT4, SOX2 y NANOG, que forman parte de la red de regulación central que controla la pluripotencia, y señales extrínsecas, que incluyen citoquinas, factores de crecimiento e interacciones con la matriz extracelular. Aunque en menor medida, LIN28 y let-7 han sido asociados a la regulación de la pluripotencia y la diferenciación celular. Alteraciones en la regulación de estos factores pueden conducir a una autorrenovación o diferenciación aberrante, dando lugar a condiciones patológicas.

Pluripotencia en cáncer y CSCs

En la primera parte de esta tesis se abordó el estudio de las células madre cancerosas, una población poco abundante dentro de los tumores que presenta propiedades de autorrenovación y multipotencia, similares a las de las células madre normales. Las CSCs han sido asociadas con la iniciación y progresión tumoral, así como con la resistencia a terapias y la recaída. En esta sección, se retoman algunos conceptos fundamentales sobre su biología, plasticidad y relevancia clínica, que resultan clave para interpretar los resultados y enfoques explorados en los siguientes capítulos.

Las CSCs han sido observadas en leucemia, mieloma múltiple, cáncer de mama, cerebro, colon, entre otras [15]. Se ha reportado que la población de CSCs en tumores es muy baja, constituyendo entre el $0{,}01\,\% - 2\,\%$ de los tumores [16], lo cual representa un desafío para

identificarlas y estudiarlas. El origen de las CSCs continúa siendo motivo de debate, ya que existe evidencia que apoya dos posibilidades no excluyentes entre sí: por un lado, que derivan de células madre normales que adquieren mutaciones y dan inicio al tumor; y por otro, que pueden originarse a partir de células diferenciadas que son capaces de desdiferenciarse y reactivar programas asociados a células madre, bajo ciertas condiciones como la inflamación crónica o la influencia del microambiente tumoral [152, 153].

Más allá de su origen, las CSCs presentan perfiles moleculares y fenotipos similares a los de las células madre normales. Se ha propuesto que las CSCs constituyen un modelo clave para el estudio de la heterogeneidad tumoral, debido a su participación en la generación y mantenimiento de poblaciones celulares con características moleculares y funcionales distintas [17].

Debido a su importancia, se han identificado varios marcadores que caracterizan a las CSCs y se han correlacionado con el diagnóstico, la terapia y el pronóstico. En los tumores sólidos, la aplicación clínica de biomarcadores específicos de CSCs es limitada. Muchos de los marcadores expresados por las CSCs también se encuentran en células madre residentes en tejidos adultos o en células madre embrionarias humanas. Además, la mayoría de los marcadores suelen identificar poblaciones heterogéneas de células madre, lo que indica que su caracterización e identificación requieren del uso de combinaciones de múltiples marcadores.

Con base en evidencia experimental, se ha propuesto que la red central de pluripotencia, compuesta por NANOG, OCT-3/4 y SOX2, actúa como un conjunto de biomarcadores intracelulares característicos de las CSCs en cáncer de mama, colon, hígado y estómago y en tumores hematológicos [154-158]. La sobreexpresión de NANOG se ha detectado en cáncer de riñón, pulmón, ovario, mama, estómago, páncreas, oral, glioma, próstata, útero, entre otros [159-170]. Su alta expresión se correlaciona con un pronóstico desfavorable en carcinomas de ovario, cáncer de colon y de mama [170-173]. NANOG se expresa en niveles más altos en las CSCs que en otras células tumorales en diversos tipos de tumores [159, 169, 173-180]. Por otro lado, algunas observaciones experimentales asocian la pluripotencia con niveles intermedios de OCT4 (niveles muy altos o muy bajos de OCT4 conducen a diferentes trayectorias de diferenciación) [181-183].

Otros marcadores intracelulares como Klf4 y c-MYC, junto con marcadores de superficie tales como CD44, CD24, CD133 y EpCAM se encuentran diferencialmente expresados en las CSCs. También se ha reportado que la sobreexpresión de LIN28 y la subexpresión de let-7 favorece el mantenimiento de CSCs [184]. Asimismo, se ha encontrado evidencia que vincula rutas de señalización como Jak/Stat, Wnt/β-catenina, Hedgehog, Notch, TGF-β y FGF, con la regulación de la pluripotencia y la plasticidad de las CSCs [154].

6.3. EMT y pluripotencia en cáncer

Si bien tanto la EMT como la pluripotencia han sido objeto de numerosos estudios, aún persisten interrogantes respecto de si estos procesos están conectados y mediante qué mecanismos biológicos. Se ha observado una similitud notable entre los perfiles de expresión génica de las células que atraviesan EMT y las células madre, lo que sugiere una conexión funcional entre ambos fenómenos. La evidencia experimental acumulada en la última década indica que la activación del programa de EMT está estrechamente vinculada con la adquisición de propiedades de pluripotencia, tanto en células normales como tumorales [185-187].

Diversos estudios han explorado esta relación. Por ejemplo, Morel *et al*. observaron que la pluripotencia estaba impulsada por la EMT en tumores mamarios [188]. Rhim *et al*., identificaron células diseminadas con fenotipo mesenquimal y características de CSCs en cáncer de páncreas [189]. En cáncer colorrectal, Choi *et al*. demostraron una fuerte correlación entre los marcadores de CSCs y los de EMT, así como con el grado de invasión y metástasis [190]. En líneas celulares de cáncer de mama triple negativo, Vijay *et al*. reportaron que la EMT induce un incremento en la población CD44+/CD24-, típico de CSCs, efecto que puede revertirse inhibiendo GSK3B [191]. Asimismo, Zhou *et al*. mostraron que el silenciamiento de SNAI1 (Snail Family Transcriptional Repressor 1) en cáncer pancreático reduce la formación de esferoides y la expresión de marcadores de pluripotencia [192].

Esta interrelación también ha sido observada al inducir EMT mediante estímulos externos. Múltiples trabajos sugieren que existe una superposición entre los estímulos que inducen EMT y aquellos que promueven características de células madre. TGF-β, Wnt/β-catenina, Hedgehog, Notch y STAT3 son rutas implicadas en ambos procesos [193, 194]. Se ha demostrado que la activación de la EMT mediante factores como TGF-β promueve la expresión de genes de pluripotencia y marcadores de CSCs [195, 196]. De hecho, la EMT inducida por TGF-β en células tumorales hepáticas se asoció con cambios en la expresión de marcadores de pluripotencia como CD133 y SOX2. En carcinomas escamosos de cabeza y cuello, la expresión elevada de factores de transcripción de EMT como SNAIL, SLUG y TWIST fue observada en poblaciones de CSCs [186]. Este patrón también se observó en cáncer oral, donde células con altos niveles de CD44, Vimentina y VE-cadherina mostraban mayor grado tumoral y potencial metastásico [185]. Se ha demostrado que la exposición a TGF-β y CsA no solo induce EMT, sino que también incrementa la expresión de genes relacionados con la pluripotencia y la resistencia a fármacos [195].

Por otro lado, los miARNs juegan un papel crucial en la regulación de la pluripotencia inducida por EMT. En particular, la familia miR-200 regula simultáneamente la EMT y la expresión de genes asociados a células madre. Su inhibición en modelos de cáncer de mama no solo promueve la EMT, sino que también facilita la adquisición de propiedades de CSCs, incluyendo la formación de mamósferas y la tumorigenicidad *in vivo*. La represión de clústeres como miR-200c-141, miR-200b-200a-429 y miR-183-96-182 se ha identificado en células madre mamarias normales, CSCs humanas y células de carcinoma embrionario, lo que sugiere su rol en el mantenimiento del fenotipo epitelial y diferenciado [185].

La acumulación de evidencia sugiere, por tanto, que la EMT y la pluripotencia no son procesos independientes, sino que se encuentran interconectados y son regulados por factores transcripcionales, vías de señalización y miARNs compartidos, formando una única red regulatoria. Dada esta interdependencia, se han desarrollado enfoques terapéuticos que apuntan a inhibir simultáneamente la EMT y las propiedades de CSCs, mostrando resultados prometedores en cáncer de vejiga, mama y líneas celulares de cáncer de pulmón [187].

6.4. Redes de regulación génica

Una red de regulación génica (GRN, por sus siglas en inglés *Gene Regulatory Network*), es un sistema de interacciones moleculares que controla la expresión génica dentro de una célula. En este sistema, genes, proteínas y otras moléculas reguladoras interactúan en procesos que activan o reprimen la transcripción de genes específicos, modulando así la producción de proteínas y determinando las funciones celulares. Estas interacciones suelen representarse

como grafos dirigidos, donde los nodos corresponden a genes, proteínas, complejos de proteínas o ARNm y los enlaces entre los elementos describen procesos bioquímicos como activación, inhibición u otras interacciones moleculares. Las redes de regulación génica desempeñan un papel crucial en la determinación de las funciones celulares y su estructura se puede inferir a partir de experimentos como estudios de mutantes, ensayos de *knockout* o análisis de expresión génica.

En las siguientes secciones, se introducirán los principales conceptos relacionados con la expresión génica, su regulación y el modelado matemático. Estos fundamentos serán empleados en los capítulos posteriores para modelar las redes de regulación génica asociadas a la EMT y a la pluripotencia en el contexto del cáncer.

6.5. Regulación de la expresión génica

Figura 6.2.: Diagrama de la expresión génica (transcripción y traducción) y los distintos niveles en los que actúa la regulación de la expresión génica.

La regulación de la expresión génica abarca una amplia gama de mecanismos que las células emplean para aumentar o disminuir la producción de productos génicos, como las proteínas y el ARNm. Estos mecanismos operan a diferentes niveles, tal como se ilustra en la Figura 6.2. La regulación epigenética incluye modificaciones en el ADN que no alteran la secuencia de nucleótidos, como la metilación o la modificación de histonas. Por otro lado, la regulación de la transcripción ocurre durante la interacción con la ARN-polimerasa y en la elongación del ARNm, donde los factores de transcripción juegan un papel crucial.

La regulación postranscripcional, que controla la cantidad de ARNm disponible después de su transcripción, incluye procesos como el *capping*, el *splicing*, la poliadenilación, la formación de complejos ARNm-microARNs, entre otros. La regulación de la traducción puede ocurrir en la iniciación (unión del ribosoma con el ARNm) y en la elongación de la cadena polipeptídica. Finalmente, la regulación postraduccional se refiere a la modificación de las proteínas ya sintetizadas. Esta incluye la inhibición de enzimas para evitar la unión al sustrato, o la adición de grupos funcionales como metilos, fosfatos o acetilos, que determinan la localización celular de las proteínas. Además, la ubiquitinación marca a las proteínas para su degradación.

El principal punto de control de la regulación de expresión génica ocurre durante la iniciación de la transcripción [197]. Esto ocurre regulando la unión de la ARN-polimerasa al promotor del gen. Las proteínas que se unen al ADN y regulan la asociación de la ARN-polimerasa se conocen como factores de transcripción. Los sitios de unión de los factores de transcripción se llaman operadores y se suelen ubicar próximos a los promotores. Si la presencia de un factor de transcripción incrementa la tasa de unión de la ARN-polimerasa, dicho factor se denomina activador; en cambio, si la reprime, se llama represor o inhibidor. La Figura 6.3 ilustra diferentes escenarios de regulación transcripcional: sin regulación, con un represor y con un activador.

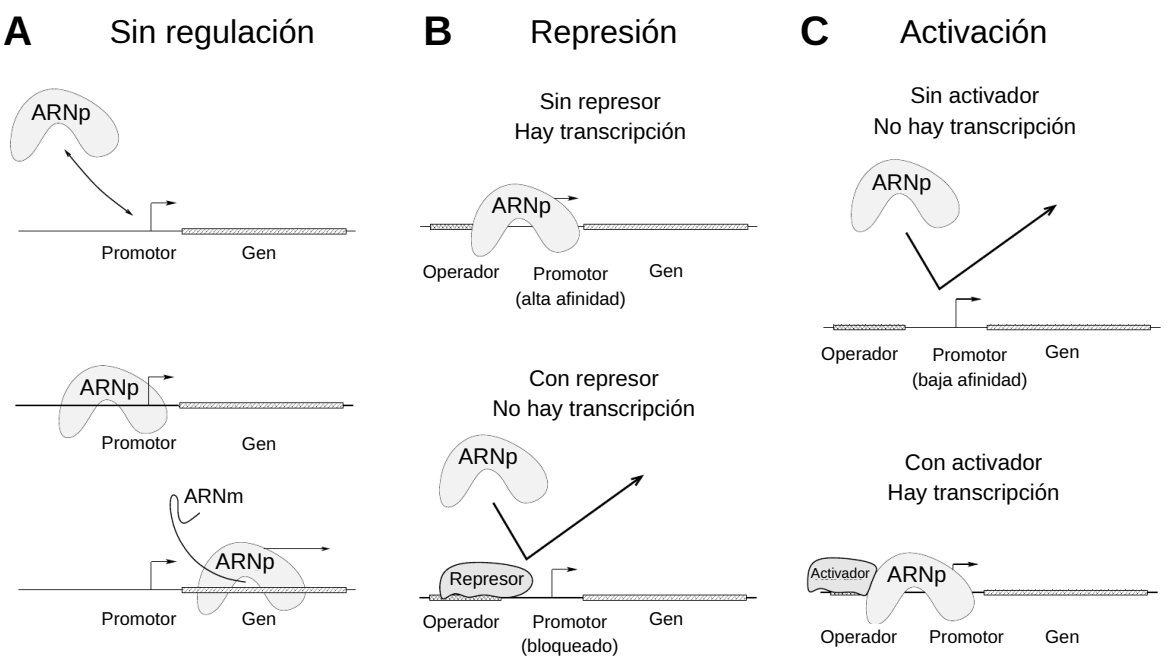

Figura 6.3.: **(A)** Ilustración del proceso de transcripción en ausencia de regulación. La ARN polimerasa (ARNp) se une a la región promotora del gen y luego se desliza a lo largo del ADN hasta la región codificante, donde produce el transcrito de ARNm. **(B)** Regulación de la transcripción por un factor de transcripción, en este caso un represor. Un represor se une a la región operadora y bloquea el acceso de la ARN polimerasa al promotor (ARNp). Mientras el represor esté presente, la transcripción no puede llevarse a cabo. **(C)** Regulación de la transcripción por un factor de transcripción, en este caso un activador. En ausencia de regulación, la afinidad del promotor por la ARN polimerasa (ARNp) es baja y la transcripción no ocurre. Un activador se une al operador y, a su vez, recluta la ARN polimerasa al promotor, favoreciendo así la transcripción. Adaptación de [198].

Además de los factores de transcripción, otros reguladores importantes son los microARNs, también conocidos como miARNs. Estas pequeñas moléculas de ARN no codificante reprimen la expresión génica mediante el silenciamiento del ARNm, ya sea inhibiendo su traducción en proteínas funcionales o promoviendo su degradación [199].

6.6. Modelado de la expresión génica

El estudio de la dinámica de GRNs permite desentrañar los mecanismos que subyacen a procesos biológicos específicos y proporciona información clave para abordar preguntas complejas en biología celular y molecular. En este contexto, el modelado matemático de redes regulatorias constituye una herramienta poderosa, ya que ofrece una alternativa relativamente rápida y económica que es complementaria a los experimentos de laboratorio. Estos modelos han demostrado ser capaces de realizar predicciones novedosas, que se pueden validar experimentalmente, sugiriendo nuevas estrategias a explorar en experimentos. A lo largo de los años, se han desarrollado distintas técnicas para modelar redes de regulación génica, incluyendo ecuaciones diferenciales, redes booleanas, redes bayesianas y técnicas estocásticas.

Esta sección introduce una de las estrategias más utilizadas para modelar GRNs, modelos basados en ecuaciones diferenciales. Se describe un abordaje para modelar la expresión génica y la regulación mediada por factores de transcripción, que luego se extiende a la regulación mediada por miARNs, dado que ambos mecanismos son fundamentales en las redes génicas analizadas en esta tesis.

6.6.1. Modelo de la expresión génica sin regulación

Comprender los mecanismos básicos de la expresión génica es fundamental para el modelado de redes de regulación génica. La expresión génica es el proceso mediante el cual se producen proteínas y ARNs, las moléculas encargadas de llevar a cabo la mayoría de las funciones celulares. La expresión de genes varía en función del tipo de célula y su rol biológico específico. Este proceso de expresión génica puede dividirse en dos etapas fundamentales: la transcripción y la traducción.

La transcripción es el primer paso de la expresión génica y consiste en la producción de ARNm a partir de un gen (ADN). Este proceso es mediado por la enzima ARN-polimerasa, que se une a la región promotora del ADN y cataliza la formación de una cadena de ARN complementaria utilizando nucleótidos precursores. Posteriormente, el ARNm producido se convierte en proteína en un segundo paso, denominado traducción. Durante la traducción, el ribosoma se une al ARNm y cataliza la síntesis de proteínas a partir de aminoácidos, utilizando como molde la secuencia de nucleótidos del ARNm. Aunque la expresión génica es un fenómeno complejo que involucra múltiples componentes y etapas adicionales, para los fines de este modelado nos centraremos en estos dos pasos esenciales.

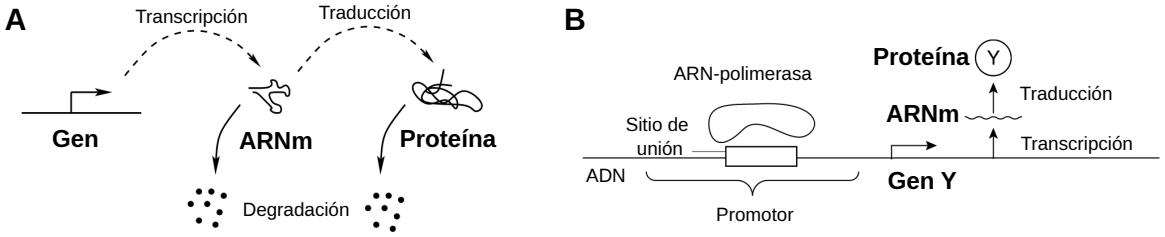

Figura 6.4.: **(A)** Ilustración del proceso de expresión génica general. Se muestran las dos etapas en la que esta ocurre: transcripción y traducción (adaptación de [198]). **(B)** Ilustración del proceso de expresión génica con mayor detalle en la maquinaria celular involucrada en la transcripción (adaptación de [200]).

En la Figura 6.4 se muestra un esquema de la transcripción (ADN a ARNm) y de la traducción (ARNm a proteína). La transcripción y la traducción se encuentran balanceadas por la degradación del ARNm y la proteína y/o el aumento del volumen de la célula. La dinámica del sistema representado en la Figura 6.4 se puede modelar mediante ecuaciones diferenciales:

$$\begin{aligned} \frac{d\,[m]}{dt} &= k_m^0 - d_m\,[m]\,, \\ \frac{d\,[X]}{dt} &= k_X^0\,[m] - d_X\,[X]\,, \end{aligned} \tag{6.1}$$

donde $[m]$ y $[X]$ representan la concentración del ARNm m y de la proteína X codificada por el gen x. A lo largo de esta tesis se utilizan corchetes [] para denotar la concentración de una

molécula. k_m^0 y k_X^0 son las tasas de transcripción y traducción, y d_m y d_X son las tasas de degradación del ARNm y de la proteína, respectivamente. Para el planteo de las ecuaciones diferenciales se utiliza la ley de acción de masas, la cual establece que la velocidad de una reacción es el producto de las concentraciones de los reactivos y la tasa de ocurrencia de esa reacción. De este modo, la producción de la proteína es el producto entre la tasa de traducción y la concentración de ARNm. Asimismo, la degradación se obtiene multiplicando la tasa de degradación de la especie y su concentración.

6.6.2. Modelo de regulación de la expresión génica por factores de transcripción

La regulación de la expresión génica por factores de transcripción se suele modelar incluyendo funciones reguladoras, que dependen de la concentración de los reguladores (factores de transcripción), en la producción del ARN mensajero m. Las funciones reguladoras son una representación simplificada de un proceso muy complejo y sobre el cual, en general, no existe mucha información experimental. De este modo, como se verá a lo largo de la tesis, existe cierta arbitrariedad en la definición de su forma funcional y de sus parámetros. Las funciones reguladoras más utilizadas en la literatura son las llamadas funciones de Hill, denotadas por $H^+([A])$ y $H^-([R])$, para un activador A y un represor R, respectivamente:

$$\begin{aligned} H^+([A]) &= \frac{\left(\frac{[A]}{K_A}\right)^n}{1+\left(\frac{[A]}{K_A}\right)^n}, \\ H^-([R]) &= \frac{1}{1+\left(\frac{[R]}{K_R}\right)^n}. \end{aligned} \tag{6.2}$$

Los parámetros K_A y K_R, denominados constantes de saturación media o de equilibrio (términos que se utilizan indistintamente en esta tesis), representan la concentración del regulador a la cual la tasa de transcripción alcanza la mitad de su valor máximo. n es conocido como coeficiente de Hill o no linealidad y refleja el grado de cooperatividad en la unión del activador o represor. Cuando no hay cooperatividad ($n = 1$), la función se reduce a la ecuación de Michaelis-Menten. Un mayor coeficiente de Hill ($n > 1$) indica una mayor cooperatividad, lo que resulta en una respuesta más pronunciada en función de la concentración del factor de transcripción, ya sea un activador o un represor, como se puede ver en la Figura 6.5. En la Sección C.1 y C.2 del Apéndice C se derivan las expresiones de tales funciones de regulación a partir del modelado de la ocupación del promotor por un factor de transcripción. Con el fin de simplificar y mostrar únicamente el efecto regulador, las funciones de Hill se encuentran normalizadas entre 0 y 1. Sin embargo, el rango de valores que toma la tasa de producción va a depender de la tasa de transcripción basal y la máxima, como se desarrolla en el Apéndice C.

Así, la expresión de una proteína X, regulada por ejemplo por un activador A, se puede describir mediante el siguiente conjunto de ecuaciones diferenciales:

$$\begin{aligned} \frac{d[m]}{dt} &= k_m^0 + \beta_m \frac{\left(\frac{[A]}{K_A}\right)^n}{1+\left(\frac{[A]}{K_A}\right)^n} - d_m[m], \\ \frac{d[X]}{dt} &= k_X^0[m] - d_X[X], \end{aligned} \tag{6.3}$$

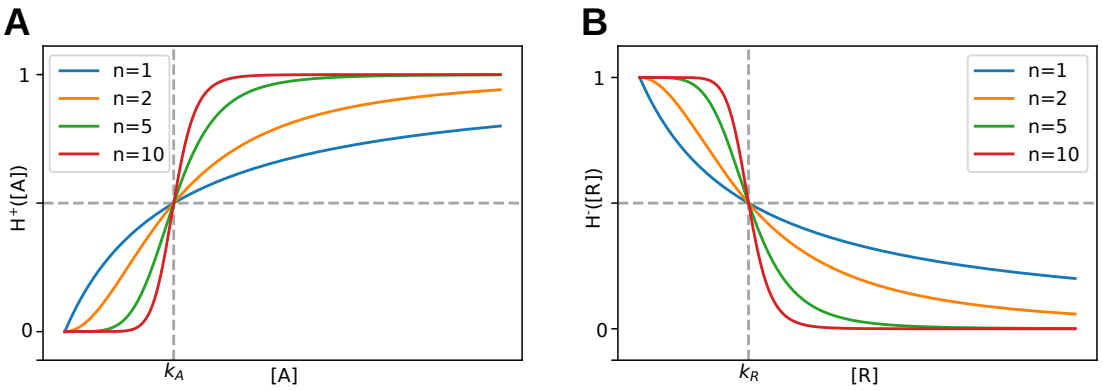

Figura 6.5.: (**A**) Función de Hill para el caso de un activador para distintos valores del coeficiente de Hill. (**B**) Función de Hill para un represor para distintos valores del coeficiente de Hill. La línea de puntos gris indica el valor de la concentración del regulador para el cual la tasa de producción del ARNm alcanza la mitad de su valor máximo (K_A para un activador y K_R para un represor).

donde β_m representa la tasa de transcripción máxima en ausencia de transcripción basal, K_A la constante de saturación media y n el coeficiente de Hill. Los demás parámetros se encuentran definidos en la ec. 6.1. La función de Hill es activadora, ya que a mayor concentración del factor de transcripción A mayor es la producción del mensajero.

En estos ejemplos se considera la regulación de la transcripción por un único factor de transcripción. Sin embargo, la transcripción de los genes es comúnmente regulada por múltiples factores de transcripción. Para ello, es necesario analizar la ocupación del promotor de un gen regulado por múltiples factores de transcripción. En las Secciones C.3 y C.4 del Apéndice C se derivan las expresiones de las funciones de regulación para múltiples factores de transcripción. Allí se distinguen distintos casos, donde los factores de transcripción compiten por el mismo sitio de unión, actúan en forma sinérgica (AND) o son independientes (OR). La elección de qué tipo de regulación conjunta se utiliza en los modelos dependerá de la disponibilidad de evidencia experimental o predicciones bioinformáticas sobre los sitios de unión, que se pueden obtener de bases de datos públicas. Cabe destacar que la selección de los distintos tipos de regulación conjunta por múltiples factores de transcripción pueden llevar a resultados diferentes, incluso utilizando la misma arquitectura de la red de regulación génica.

6.6.3. Modelo de regulación de la expresión génica por microARNs

Los miARNs son pequeñas moléculas de ARN no codificante, de entre 21 y 25 nucleótidos de longitud, transcritos a partir de genes (ADN) pero no traducidas a proteínas. Descubiertos en 1993 [201], han revolucionado el entendimiento de la regulación génica, y actualmente se estima que regulan hasta el $60\,\%$ de los genes en *Homo sapiens* y mamíferos [202]. La importancia de los miARNs fue reconocida con el Premio Nobel de Fisiología o Medicina en 2024, otorgado a Victor Ambros y Gary Ruvkun por el descubrimiento del miARN y su papel en la regulación génica postranscripcional. Estas moléculas desempeñan un papel crucial en la regulación de la expresión génica al silenciar genes diana, ya sea mediante la inhibición de la traducción o la degradación del ARNm.

En el contexto del modelado de redes de regulación génica, se han propuesto distintas estrategias para incorporar la regulación mediada por miARNs. Con el propósito de ilustrar cómo este tipo de interacción puede ser incorporada en un modelo matemático, a continuación se presenta un ejemplo de regulación génica mediada por un miARN. Se considera un gen x que se transcribe en ARNm m, el cual es posteriormente traducido en la proteína X. Se analiza el caso en el que la expresión de x está regulada exclusivamente por un miARN μ, sin considerar otros reguladores como factores de transcripción.

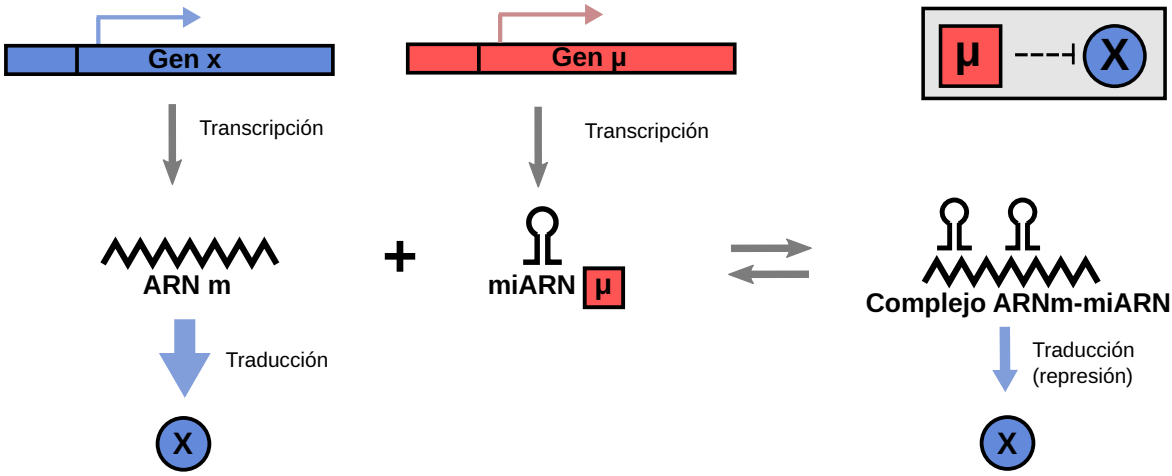

Figura 6.6.: Esquema de la red de regulación génica destacada en el bloque gris ubicado en el extremo superior derecho. La línea punteada indica regulación postranscripcional. El gen x se traduce en la proteína X, y el miARN μ inhibe la expresión de X cuando se forma el complejo ARNm-miARN.

Una estrategia para modelar la regulación mediada por miARNs mediante ecuaciones diferenciales consiste en adaptar el formalismo utilizado para describir la acción de factores de transcripción, como se propone en el trabajo de Tian *et al* [203]. En este caso, se emplean funciones inhibitorias de Hill, pero aplicadas al nivel de traducción de proteínas, en contraste con la regulación por factores de transcripción, que actúan a nivel transcripcional. Bajo esta suposición, los miARNs inhiben la traducción del ARNm sin afectar su síntesis ni degradación directamente. La Figura 6.6 muestra un esquema representativo del sistema, donde se ilustra la formación del complejo ARNm-miARN, en el cual la traducción se encuentra inhibida. A mayor concentración de miARN, mayor es la inhibición de la traducción. De este modo, el sistema puede modelarse de la siguiente manera:

$$\begin{aligned} \frac{d\,[m]}{dt} &= k_m^0 - d_m\,[m]\,, \\ \frac{d\,[X]}{dt} &= \beta_X\,[m]\,\frac{1}{1+\left(\frac{[\mu]}{K_\mu}\right)^{n_\mu}} - d_X\,[X]\,, \\ \frac{d\,[\mu]}{dt} &= k_\mu^0 - d_\mu\,[\mu]\,. \end{aligned} \tag{6.4}$$

En la ec. 6.4, la producción de la proteína X depende tanto de la concentración del ARNm como de una función de Hill represora, la cual refleja la inhibición que depende de la concentración del miARN. Los parámetros k_m^0 y k_μ^0 representan las tasas de transcripción basal del ARNm m y del miARN μ, β_X es la tasa de traducción máxima, n_μ el coeficiente de Hill y K_μ es la constante de equilibrio. El ARNm, el miARN y la proteína son degradados con

tasas d_m, d_μ y d_X, respectivamente. Si bien la degradación no está representada en la Figura 6.4 para evitar sobrecargar el esquema, sí es considerada en el modelo matemático.

Otra alternativa para modelar la regulación por miARNs es la propuesta desarrollada por Lu y colegas [204, 205]. En sus trabajos, dado el pequeño tamaño de los miARNs en comparación con las distancias entre los sitios de unión, se emplean funciones de regulación del tipo Adair (ver ecs. 7.3) en vez de funciones de Hill. Estas funciones regulatorias se pueden derivar de manera análoga a las desarrolladas en la Sección C.2, considerando que un ARNm puede estar ligado a $i = 1, 2, \ldots, n_\mu$ miRNAs, formando un complejo ARNm-miRNA, donde n_μ representa el número de sitios de unión. En este enfoque, además se consideran dos contribuciones de la regulación de la expresión génica por miARNs: la inhibición de la traducción y la degradación de los complejos ARNm-miARN. Para modelar la inhibición de la traducción, se introducen las funciones denotadas por $L([\mu])$. En cuanto a la degradación activa del complejo ARNm-miARN, tanto el ARNm como los miARNs son degradados, lo cual se representa mediante las funciones de regulación $Y_m([\mu])$ y $Y_\mu([\mu])$, asociadas al ARNm y al miARN, respectivamente.

De esta forma, cuando un miARN regula la expresión génica de un gen *x*, el conjunto de ecuaciones diferenciales que caracteriza el sistema es:

$$\begin{aligned} \frac{d\,[m]}{dt} &= k_m^0 - Y_m\left([\mu]\right)[m] - d_m\,[m]\,, \\ \frac{d\,[X]}{dt} &= \beta_X\,[m]\,L\left([\mu]\right) - d_X\,[X]\,, \\ \frac{d\,[\mu]}{dt} &= k_\mu^0 - [m]\,Y_\mu\left([\mu]\right) - d_\mu\,[\mu]\,. \end{aligned} \tag{6.5}$$

En las ecs. 6.5 las definiciones de los parámetros de las ecuaciones diferenciales del ARNm y la proteína (k_m^0, k_μ^0, β_X, d_m, d_X y d_μ) se mantienen respecto a la ec. 6.4. En términos generales, al modelar la degradación del complejo ARNm-miARN para el miARN μ y el ARNm m en las ecs. 6.5, se incorporan nuevos términos respecto a la ec. 6.4 que representan ese fenómeno. En la Sección 7.2.2 del siguiente capítulo se proporciona una descripción más detallada de este enfoque, aplicado al modelado de la red de regulación génica de EMT, donde se describen las funciones de regulación $L([\mu])$, $Y_m([\mu])$ y $Y_\mu([\mu])$.

6.6.4. Modelo de una red de regulación génica (regulación por factores de transcripción y miARNs)

En esta sección se presenta una red de regulación génica simple que integra los mecanismos previamente descritos de forma individual. El objetivo es ilustrar cómo los factores de transcripción y los miARNs pueden actuar conjuntamente para regular la expresión génica. En este modelo, m representa el ARNm transcrito a partir de un gen x, X es la proteína que se traduce a partir de dicho ARNm y μ es un miARN. El factor de transcripción X activa la transcripción de μ, mientras que μ reprime la traducción de X, como se muestra en la Figura 6.7.

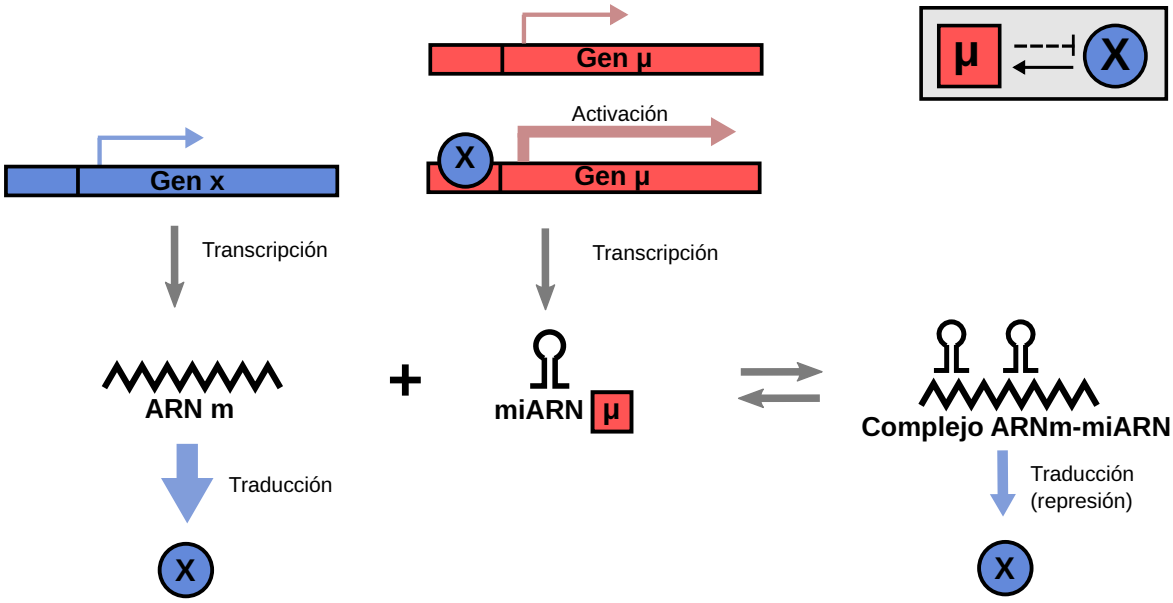

Figura 6.7.: Esquema de la red de regulación génica destacada en el bloque gris ubicado en el extremo superior derecho. La línea punteada indica regulación postranscripcional, mientras que la línea sólida representa la regulación de la transcripción. El factor de transcripción X activa la transcripción del miARN μ, el cual, a su vez, reprime la expresión de X. Se ilustran los mecanismos de regulación dentro de la red, enfatizando las diferencias entre la regulación mediada por factores de transcripción y la regulación llevada a cabo por miARNs.

La dinámica del sistema se puede describir mediante el siguiente conjunto de ecuaciones diferenciales:

$$\begin{aligned} \frac{d\,[m]}{dt} &= k_m^0 - d_m\,[m]\,, \\ \frac{d\,[X]}{dt} &= \beta_X\,[m]\,\frac{1}{1+\left(\frac{[\mu]}{K_{\mu\to X}}\right)^{n_{\mu\to X}}} - d_X\,[X]\,, \\ \frac{d\,[\mu]}{dt} &= k_\mu^0 + \beta_\mu \frac{\left(\frac{[X]}{K_{X\to\mu}}\right)^{n_{X\to\mu}}}{1+\left(\frac{[X]}{K_{X\to\mu}}\right)^{n_{X\to\mu}}} - d_\mu\,[\mu]\,. \end{aligned} \tag{6.6}$$

Este sistema de ecuaciones refleja cómo el factor de transcripción X activa la transcripción del miARN μ, mientras que μ reprime la traducción de la proteína X, utilizando el modelo de regulación por miARNs propuesto por Tian y colaboradores [203]. Se incluyen tasas de transcripción basal constantes, representadas por las constantes k_m^0 y k_μ^0 para el ARNm m y el miARN μ, respectivamente. Nuevamente, aunque en la Figura 6.7 no se esquematiza la degradación para simplificar la visualización y resaltar los mecanismos regulatorios, sí se encuentra incorporada en el modelo. La degradación del ARNm, de la proteína y del miARN es proporcional a sus respectivas concentraciones y a sus tasas de degradación d_m, d_X y d_μ. La traducción es proporcional a la concentración de ARNm, ya que la síntesis de la proteína depende directamente de la cantidad de ARNm disponible. Este proceso es modulado por la acción represora del miARN, representada mediante una función inhibitoria de Hill. En este contexto, β_μ representa la tasa máxima de producción del miARN, $K_{\mu\to X}$ y $K_{X\to\mu}$ son las constantes de equilibrio asociadas a la regulación de la producción de la proteína y el miARN, respectivamente, y $n_{\mu\to X}$ y $n_{X\to\mu}$ son los coeficientes de Hill correspondientes.

Esta red de regulación génica ilustrativa permite integrar los conceptos previamente discutidos, modelando de forma conjunta los procesos de transcripción, traducción y degradación de productos génicos. Asimismo, incorpora dos niveles de regulación: la activación transcripcional y el silenciamiento mediado por miARNs, lo que proporciona una opción para modelar los principales mecanismos de regulación de la expresión génica. A medida que las redes de regulación génica aumentan en tamaño y complejidad, también se incrementa el número de ecuaciones diferenciales necesarias para describirlas, así como el número de parámetros. Esto plantea el desafío de implementar la regulación conjunta ejercida por múltiples reguladores, como se discute en la Sección C.4. Este tipo de modelado es útil para entender cómo diferentes reguladores y mecanismos de regulación interactúan en redes génicas más complejas y definen los fenotipos y estados celulares.

6.7. Sistemas dinámicos

En esta sección, se presenta una breve introducción a los sistemas dinámicos y a algunas herramientas que se emplearán en los próximos capítulos.

El estudio de los sistemas dinámicos proporciona un marco matemático para analizar la evolución temporal de sistemas complejos en diversas disciplinas, como la física, la biología y la ingeniería. En este contexto, los sistemas de ecuaciones diferenciales permiten modelar la dinámica en tiempo continuo de las variables de estado y caracterizar su comportamiento bajo distintas condiciones. En esta tesis, se desarrollan modelos basados en ecuaciones diferenciales ordinarias para describir la dinámica de redes de regulación génica, las cuales son un caso particular de sistemas dinámicos.

6.7.1. Puntos fijos

En el análisis de sistemas dinámicos, un concepto clave es el de punto fijo o estado estacionario, que corresponde a soluciones en las que las derivadas temporales de todas las variables son nulas, es decir, el sistema deja de evolucionar en el tiempo. Matemáticamente, si un sistema está descrito por un conjunto de ecuaciones diferenciales de la forma:

$$\frac{dx_i}{dt} = f_i(x_1, x_2, ..., x_n, \boldsymbol{p}),$$

donde x_i representa las variables dinámicas y $\boldsymbol{p}$ es el conjunto de parámetros, los puntos fijos son aquellos valores $(x_1^*, x_2^*, ..., x_n^*)$ tales que:

$$f_i(x_1^*, x_2^*, ..., x_n^*, \boldsymbol{p}) = 0 \quad \forall i.$$

Para estudiar la naturaleza de estos puntos, es crucial analizar su estabilidad, lo cual indica si el sistema retorna al punto fijo tras pequeñas perturbaciones (estable) o si, por el contrario, se aleja de él (inestable). La estabilidad de un punto fijo puede analizarse mediante la linealización del sistema en torno a dicho punto, considerando la matriz Jacobiana:

$$J_{ij} = \frac{\partial f_i}{\partial x_j}\Big|_{(x^*)}.$$

Si los autovalores de J tienen parte real negativa, pequeñas perturbaciones se amortiguarán con el tiempo. En este caso, las trayectorias cercanas convergen hacia él, y se dice que el punto fijo es estable (atractor). Por otro lado, si al menos un autovalor tiene parte real positiva, el punto fijo es inestable, indicando que pequeñas perturbaciones crecerán con el tiempo, alejando el sistema del punto fijo. Si los autovalores son nulos, el punto fijo es neutro y, si son imaginarios, se presenta un comportamiento oscilatorio. Existen clasificaciones más detalladas según la combinación de partes reales e imaginarias, pero para los fines de esta tesis no se considerarán.

6.7.2. Nulclinas

En sistemas de dos variables, las nulclinas proporcionan información clave sobre la estructura del sistema. Las nulclinas son las curvas donde la derivada temporal de una variable se anula:

$$\frac{dx_1}{dt} = 0 \quad \Rightarrow \quad \text{nulclina de } x_1,$$

$$\frac{dx_2}{dt} = 0 \quad \Rightarrow \quad \text{nulclina de } x_2.$$

Se define la nulclina de x_1 como el conjunto de puntos (x_1, x_2) tales que $\frac{dx_1}{dt} = 0$ y la nulclina de x_2 como el conjunto de puntos (x_1, x_2) tales que $\frac{dx_2}{dt} = 0$. Geométricamente, las nulclinas dividen el plano de fases en regiones con diferentes direcciones del flujo, proporcionando una herramienta gráfica para identificar puntos fijos (intersecciones de nulclinas) y prever el comportamiento cualitativo del sistema. Si las nulclinas de x_1 y x_2 se cruzan en un punto y el flujo en su entorno converge hacia este, el punto fijo es estable.

6.7.3. Diagramas de Bifurcación

Las bifurcaciones ocurren cuando una pequeña variación en un parámetro induce un cambio cualitativo en el comportamiento del sistema, alterando el número o la estabilidad de los puntos fijos. En modelos de sistemas biológicos, esto puede reflejarse en cambios fenotípicos abruptos. Los valores del parámetro para los cuales ocurren estos cambios se denominan puntos de bifurcación, y los diagramas en los cuales se observa cómo varían los puntos fijos respecto a un parámetros se conocen como diagramas de bifurcación.

A continuación, se presenta un ejemplo simple, a modo ilustrativo, de una red de inhibición mutua en la cual dos especies, x_1 y x_2, se reprimen mutuamente. Este sistema se puede describir de la siguiente forma:

$$\begin{aligned}\frac{d}{dt}[x_1] &= \frac{\beta_1}{1+\left(\frac{[x_2]}{K_2}\right)^{n_1}} - d_1[x_1],\\ \frac{d}{dt}[x_2] &= \frac{\beta_2}{1+\left(\frac{[x_1]}{K_1}\right)^{n_2}} - d_2[x_2],\end{aligned} \tag{6.7}$$

donde β_1 y β_2 son las tasas máximas de producción de x_1 y x_2, respectivamente, en ausencia de inhibición; K_1 y K_2 son las constantes de saturación; n_1 y n_2, los coeficientes de Hill; y d_1 y d_2, las constantes de degradación de las especies x_1 y x_2, respectivamente.

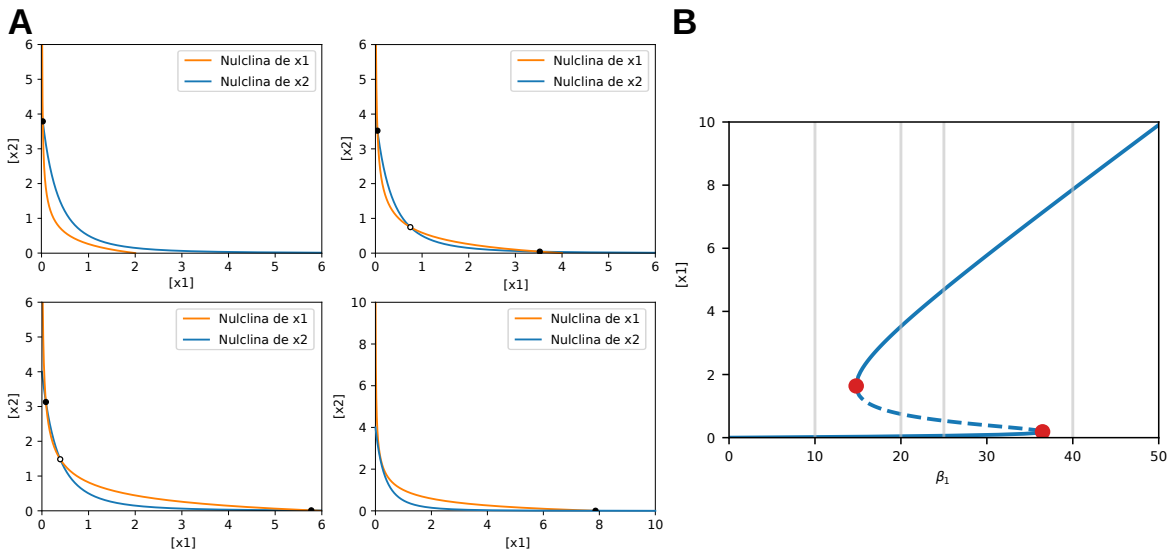

Figura 6.8.: **(A)** Diagramas de fase que muestran las nulclinas para cuatro valores diferentes del parámetro β_1 $(10, 20, 25, 40)$. La intersección de las nulclinas corresponde a puntos fijos. Se utilizan círculos sólidos para los puntos fijos estables y círculos vacíos para los inestables. **(B)** Diagrama de bifurcación que muestra el comportamiento en estado estacionario de $[x_1]$ en función de β_1, las líneas color gris marcan los valores del parámetro β_1 utilizados en la Figura 6.8A. Los puntos rojos indican puntos de bifurcación, las líneas sólidas azules representan estados estables y las líneas de puntos azules son estados inestables. Valores de los parámetros: $\beta_1 = \beta_2 = 20$ $(concentración \cdot tiempo^{-1})$, $K_1 = K_2 = 1$ $(concentración)$, $d_1 = d_2 = 5$ $(tiempo^{-1})$ y $n_1 = n_2 = 3$. Las unidades son arbitrarias.

Los diagramas de fase en la Figura 6.8A representan las nulclinas para cuatro valores diferentes del parámetro β_1 para la red de inhibición mutua modelada por las ecs. 6.7. A medida que β_1 varía, la nulclina de x_1 se desplaza hacia arriba, modificando la cantidad de intersecciones entre las nulclinas y, por lo tanto, la cantidad y estabilidad de los puntos fijos del sistema. El rango de valores de $[x_1]$ y $[x_2]$ en el diagrama del espacio de fases correspondiente al último subgráfico (abajo a la derecha) se ha ampliado con el único propósito de visualizar el punto fijo, que en este caso adquiere un valor más alto en comparación con los escenarios anteriores.

La Figura 6.8B muestra el diagrama de bifurcación para esta red de inhibición mutua. En ella se observa el comportamiento en estado estacionario de la concentración $[x_1]$ en función del parámetro β_1. Los valores de β_1 correspondientes a cada subgráfico del Panel A están señalizados en este diagrama mediante líneas verticales grises. La curva en forma de "S" es característica de sistemas biestables. Los puntos donde ocurren las bifurcaciones se denominan bifurcaciones de tipo *saddle-node*, ya que corresponden a la fusión de un punto

de silla inestable con un nodo estable. En el rango comprendido entre estas bifurcaciones, coexisten tres estados estacionarios.

La Figura 6.8B refleja la capacidad de este sistema biestable para actuar como un interruptor: una variación del parámetro β_1 que lo lleve más allá de los puntos de bifurcación *saddle-node* hará que el sistema cambie entre los estados de baja y alta concentración de $[x_1]$. En el intervalo biestable intermedio, este cambio no ocurre de manera instantánea, sino que introduce un retraso en la respuesta; en esta región, el estado del sistema no está determinado únicamente por el valor actual de β_1, sino también por las condiciones iniciales del sistema.

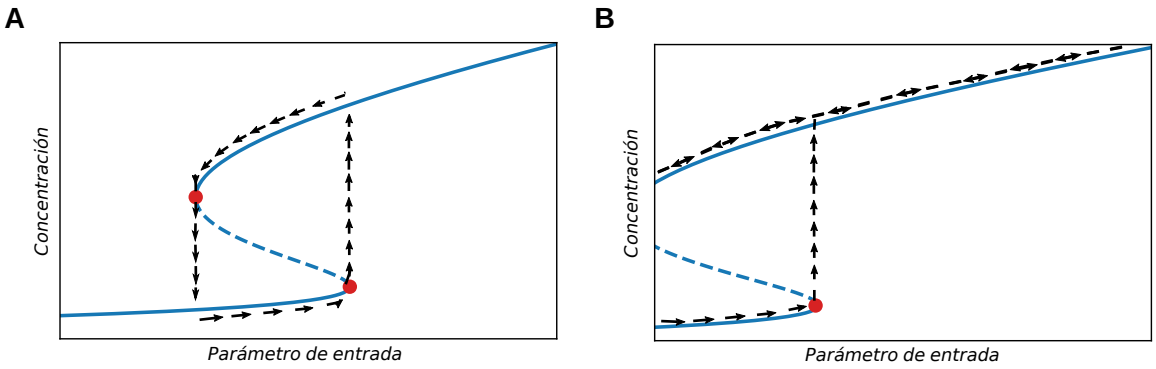

Figura 6.9.: Esquema del comportamiento histerético del sistema. **(A)** En el intervalo biestable, el sistema conserva su estado dependiendo de la trayectoria previa. **(B)** En un sistema con transición irreversible, una de las bifurcaciones *saddle-node* queda fuera del rango de valores relevantes del parámetro de entrada, lo que impide el retorno al estado anterior.

El esquema en la Figura 6.9A ilustra este comportamiento. Si el sistema ingresa en el rango biestable desde un estado de alta concentración, permanecerá en dicho estado dentro de este intervalo. Lo contrario ocurre si el sistema entra desde un estado de baja concentración. Esta dependencia del estado previo se conoce como histéresis. A medida que el sistema alterna entre los dos estados, sigue un ciclo de histéresis, en el que la transición entre ellos ocurre en dos valores umbral distintos del parámetro de entrada (puntos de bifurcación).

Algunos interruptores son irreversibles. Como se muestra en la Figura 6.9B, si una de las bifurcaciones *saddle-node* se encuentra fuera del rango de valores relevantes del parámetro, el sistema realiza una transición unidireccional entre los dos estados, sin posibilidad de retorno.

7 Modelos de redes de regulación génica de la transición epitelio-mesénquima y la pluripotencia

7.1. Introducción

En la Parte I de esta tesis, se definieron y calcularon *scores* para cuantificar características relevantes en cáncer, a partir de datos de scRNA-seq. Entre estas características se incluyeron la EMT, alteraciones cromosómicas (CNAs), la entropía, entre otras. Se observó una correlación positiva entre la entropía, un indicador de pluripotencia, y la actividad de la red de PPIN asociada a la EMT, como se ilustra en la Figura 5.5A, en muestras de cáncer de mama.

Si bien se reconoce que la EMT y la pluripotencia son procesos que desempeñan roles clave en la progresión del cáncer, y hay evidencia experimental que sugiere su interconexión, los mecanismos biológicos subyacentes y la forma en que estos procesos se relacionan aún no se comprenden completamente. Con esta motivación, en este capítulo se analiza, de manera independiente, la red regulatoria central de la EMT y de la pluripotencia mediante modelos matemáticos. Se realizan adaptaciones y se desarrollan modelos de ambas redes en función de los objetivos de esta tesis. Posteriormente, en el capítulo 8, se estudiará la interconexión entre ambas redes.

7.2. Red de regulación génica de la EMT

Las redes de regulación génica se caracterizan por la topología de sus interacciones y por las funciones regulatorias que las gobiernan. En el caso específico del circuito génico asociado a la EMT, se han propuesto modelos con diferentes topologías y funciones regulatorias, lo que refleja la complejidad y la diversidad de enfoques en la modelización de este proceso [203, 205-208]. No obstante, existe un consenso general en la literatura sobre las principales especies que conforman el núcleo de la red reguladora de la EMT: ZEB, SNAIL, miR-200 y miR-34. Asimismo, hay consenso en que esta red es altamente conectada.

En el año 2013, se publicaron dos modelos independientes de la red de regulación génica de la EMT, a partir de los cuales se han derivado múltiples trabajos subsecuentes. Estos modelos son conocidos en la literatura como CBS (por sus siglas en inglés, *Cascading Bistable Switches*), desarrollado por Tian y colaboradores [203], y TCS (por sus siglas en inglés, *Ternary Chimera Switch*), propuesto por Lu y colaboradores [205]. Ambos modelos capturan aspectos clave de la EMT que han sido observados experimentalmente, incluyendo la existencia de tres estados estables que corresponden a los fenotipos epitelial (E), mesenquimal (M) e híbrido epitelial-mesenquimal (E/M). Además, los modelos predicen que la transición presenta histéresis, lo

que implica que la transición inversa, denominada transición mesénquima-epitelio (MET) no ocurre de la misma forma que la EMT.

Aunque los modelos coinciden en que la transición ocurre en dos etapas (de epitelial a híbrido y de híbrido a mesenquimal), difieren en el mecanismo que subyace a estas transiciones, un aspecto que aún no ha sido validado experimentalmente. El modelo CBS propone que la primera transición es mediada por el módulo SNAIL/miR-34, mientras que la segunda es inducida por el módulo ZEB/miR-200. En contraste, el modelo TCS sugiere que el módulo ZEB/miR-200 actúa como un *switch* ternario, regulando ambas transiciones entre los estados E, E/M y M, y que el módulo SNAIL/miR-34 funciona como un amortiguador del ruido.

7.2.1. Modelo CBS

En la Figura 7.1A se muestra un esquema de la arquitectura del modelo CBS de la red regulatoria de la EMT, desarrollado por Tian y colaboradores [203]. El factor de crecimiento transformante β (TGF-β) es un potente inductor de la EMT [209]. En el modelo CBS se considera tanto la contribución del TGF-β endógeno como la del exógeno; este último actúa como una señal externa que puede modular la EMT. El TGF-β, ya sea producido de forma autócrina o como señal exógena, promueve la transcripción del ARNm de snail, que posteriormente se traduce en la proteína SNAIL.

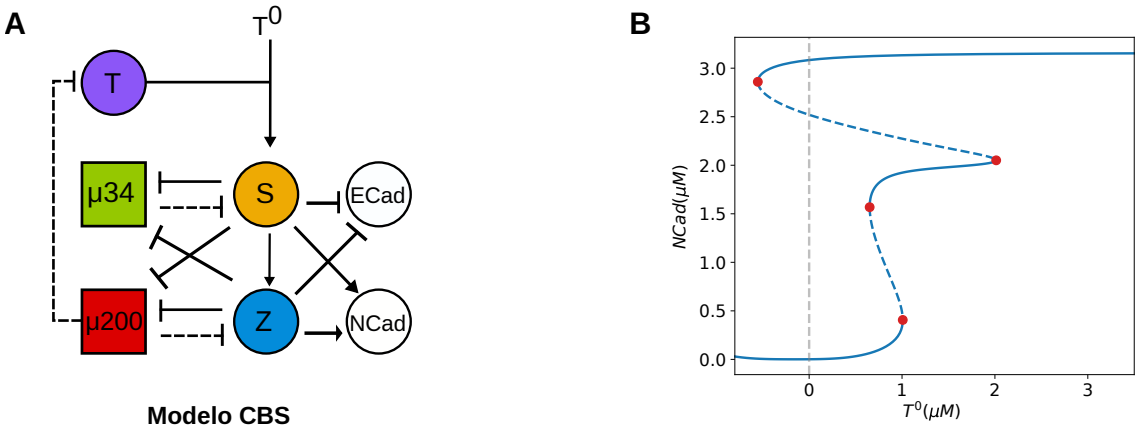

Figura 7.1.: (**A**) Modelo CBS de la red de regulación génica EMT [203]. Las líneas de discontinuas indican regulación por miARNs, mientras que las líneas sólidas representan la regulación por factores de transcripción. (**B**) Diagrama de bifurcación del modelo CBS de la red de regulación génica EMT. Los puntos rojos representan puntos de bifurcación, las líneas sólidas azules representan estados estables y las líneas discontinuas azules son estados inestables. La línea discontinua vertical gris indica la concentración del TGF-β exógeno $T^0 = 0$. Valores menores a 0 no tienen sentido biológico.

La evidencia experimental sugiere la existencia de dos bucles de retroalimentación doble negativa esenciales entre los factores de transcripción y los miARNs. El primero de estos bucles involucra a SNAIL y miR-34 [210, 211]. SNAIL inhibe la transcripción de miR-34, el cual reprime la traducción de snail, estableciendo así un circuito de retroalimentación doble negativa. Entre ZEB y miR-200 hay otro bucle [212-214], ZEB reprime la transcripción de miR-200, y a su vez miR-200 inhibe la traducción de zeb. Además, los dos bucles de inhibición mutua se encuentran conectados entre sí; SNAIL induce la transcripción del gen zeb, dando

lugar a la síntesis de la proteína ZEB. Los factores de transcripción ZEB y SNAIL reprimen la transcripción de miR-34 y miR-200, respectivamente.

Por otro lado, miR-200 inhibe la expresión autócrina de TGF-β [215], creando un bucle de retroalimentación adicional. La E-cadherina, que es un marcador distintivo del fenotipo epitelial, es inhibida por SNAIL y ZEB. La N-cadherina es un marcador de células mesenquimales y es activada por SNAIL y ZEB [216].

El modelo CBS está descrito por el conjunto de ecuaciones diferenciales:

$$
\begin{aligned}
\frac{d\left[\mu 200\right]}{dt} &= k^0_{\mu 200} + \beta_{S,Z\to\mu 200}\frac{1}{1+\left(\frac{[S]}{K_{S\to\mu 200}}\right)^{n_{S\to\mu 200}}+\left(\frac{[Z]}{K_{Z\to\mu 200}}\right)^{n_{Z\to\mu 200}}} - d_{\mu 200}\left[\mu 200\right],\\
\frac{d\left[mz\right]}{dt} &= k^0_{mz} + \beta_{mz}\frac{\left(\frac{[S]}{K_{S\to mz}}\right)^{n_{S\to mz}}}{1+\left(\frac{[S]}{K_{S\to ms}}\right)^{n_{S\to mz}}} - d_{mz}\left[mz\right],\\
\frac{d\left[Z\right]}{dt} &= \beta_Z\left[mz\right]\frac{1}{1+\left(\frac{[\mu 200]}{K_{\mu 200\to Z}}\right)^{n_{\mu 200\to Z}}} - d_Z\left[Z\right],\\
\frac{d\left[\mu 34\right]}{dt} &= k^0_{\mu 34} + \beta_{S,Z\to\mu 34}\frac{1}{1+\left(\frac{[S]}{K_{S\to\mu 34}}\right)^{n_{S\to\mu 34}}+\left(\frac{[Z]}{K_{Z\to\mu 34}}\right)^{n_{Z\to\mu 34}}} - d_{\mu 34}\left[\mu 34\right],\\
\frac{d\left[ms\right]}{dt} &= k^0_{ms} + \beta_{ms}\frac{\left(\frac{[T]+T^0}{K_{T\to ms}}\right)^{n_{T\to ms}}}{1+\left(\frac{[T]+T^0}{K_{T\to ms}}\right)^{n_{T\to ms}}} - d_{ms}\left[ms\right],\\
\frac{d\left[S\right]}{dt} &= \beta_S\left[ms\right]\frac{1}{1+\left(\frac{[\mu 34]}{K_{\mu 34\to S}}\right)^{n_{\mu 34\to S}}} - d_S\left[S\right],\\
\frac{d\left[T\right]}{dt} &= k^0_T + \beta_T\frac{1}{1+\left(\frac{[\mu 200]}{K_{\mu 200\to T}}\right)^{n_{\mu 200\to T}}} - d_T\left[T\right],\\
\frac{d\left[ECad\right]}{dt} &= \beta_{S\to E}\frac{1}{1+\left(\frac{[S]}{K_{S\to E}}\right)^{n_{S\to E}}} + \beta_{Z\to E}\frac{1}{1+\left(\frac{[Z]}{K_{Z\to E}}\right)^{n_{Z\to E}}} - d_E\left[ECad\right],\\
\frac{d\left[NCad\right]}{dt} &= \beta_{S\to N}\frac{\left(\frac{[S]}{K_{S\to N}}\right)^{n_{S\to N}}}{1+\left(\frac{[S]}{K_{S\to N}}\right)^{n_{S\to N}}} + \beta_{Z\to N}\frac{\left(\frac{[Z]}{K_{Z\to N}}\right)^{n_{Z\to N}}}{1+\left(\frac{[Z]}{K_{Z\to N}}\right)^{n_{Z\to N}}} - d_N\left[NCad\right].
\end{aligned}
\tag{7.1}
$$

En estas ecuaciones, las variables $\mu 200$, mz, Z, $\mu 34$, ms, S, T, $ECad$, y $NCad$ representan, respectivamente, el miR-200, el ARNm de zeb y su proteína correspondiente, el miR-34, el ARNm de snail y su proteína correspondiente, el TGF-β endógeno, y los marcadores epitelial y mesenquimal, E-cadherina y N-cadherina. El parámetro T^0 representa el TGF-β exógeno, que actúa como señal externa. La doble inhibición de miR-200 y miR-34 es considerada competitiva mientras que la doble inhibición de la E-cadherina y N-cadherina es considerada como no competitiva sin influencia mutua (OR) (Sección C.3). Las constantes k^0_i, d_i, y β_i corresponden a la tasa basal de producción, la tasa de degradación y la tasa máxima de producción de la especie i. Los parámetros $K_{i\to j}$ y $n_{i\to j}$ denotan las constantes de equilibrio y los coeficientes de Hill que caracterizan la función de regulación ejercida por la especie i sobre la especie j. En las funciones de Hill empleadas, los autores seleccionaron un valor de dos para todos los coeficientes de Hill con el fin de introducir una no linealidad moderada en la

regulación. En la Tabla D.1 del Apéndice D se proporciona una descripción de los parámetros y se detallan los valores que se utilizan.

En el diagrama de bifurcación que se muestra en la Figura 7.1B se observa la existencia de tres estados estables, asociados a los fenotipos epitelial, mesenquimal e híbrido, correspondientes a los niveles de $NCad$ bajo, alto e intermedio, respectivamente. Según este modelo, el proceso de EMT ocurre en dos etapas: la transición reversible del fenotipo epitelial al híbrido y la transición irreversible del fenotipo híbrido E/M al mesenquimal. Dentro del rango de parámetros biológicamente plausibles (concentraciones mayores o iguales a 0, incluida la del TGF-β exógeno T^0), una vez alcanzado el estado mesenquimal, el sistema permanece en él, sin posibilidad de transicionar a otro estado.

Posteriormente, el mismo grupo propone una modificación del modelo CBS, considerando la regulación por miARNs de forma detallada, similar a la planteada en el modelo TCS, que se describirá en la siguiente sección, y adicionando la autorrepresión de SNAIL [208]. Al igual que en la versión original, se encuentran tres posibles fenotipos y la EMT ocurre en dos pasos: la primera transición reversible entre el estado epitelial y el híbrido, y la segunda entre el estado híbrido E/M y el mesenquimal, mayormente irreversible. Las predicciones son equivalentes al modelo original que propone el grupo, sin observarse cambios notables.

7.2.2. Modelo TCS

La determinación del destino celular puede ser modulada por una variedad de señales internas y externas, como HIF-1α, p53, TGF-β, HGF, FGF, EGF, Notch, Wnt y Hedgehog. En el modelo TCS, este conjunto de señales se representa mediante una señal de entrada que actúa sobre una unidad reguladora central de la EMT. La arquitectura de esta unidad central consta de dos módulos quiméricos interconectados: el módulo miR-34/SNAIL y el módulo miR-200/ZEB. Se dice que los módulos son quiméricos porque están compuestos por un miARN y un factor de transcripción.

Cada módulo constituye un circuito de retroalimentación doble, en el que ambos elementos se reprimen mutuamente, de manera análoga al modelo CBS. A diferencia de este último, el modelo TCS incluye además la autorregulación de los factores de transcripción: autoinhibición de SNAIL y autoactivación de ZEB. Aunque no existe evidencia experimental de la autoactivación de ZEB, esta se incluye en el modelo porque permite la existencia de tres estados estables; en su ausencia, el modelo no predice el fenotipo híbrido E/M. Por su parte, la autoinhibición de SNAIL se basa en evidencia experimental que sugiere una represión modesta de su propia expresión [217]. Más recientemente se ha observado que esta autorregulación es más compleja, involucrando una autorregulación negativa a través del potenciador proximal y una positiva mediada por el potenciador distal [218]. En términos de las predicciones, la incorporación de la autoinhibición de SNAIL no afecta los posibles estados estables del sistema.

Los factores de transcripción SNAIL y ZEB reprimen la expresión de genes específicos de células epiteliales, como la E-cadherina, y favorecen la expresión de marcadores mesenquimales como la N-cadherina. Niveles altos de miR-34 y miR-200 se asocian con el fenotipo epitelial, mientras que niveles elevados de SNAIL y ZEB se correlacionan con el fenotipo mesenquimal.

Para analizar la dinámica de esta red de regulación génica, se utiliza el marco teórico desarrollado previamente por el mismo grupo [204]. Este enfoque modela de manera detallada la

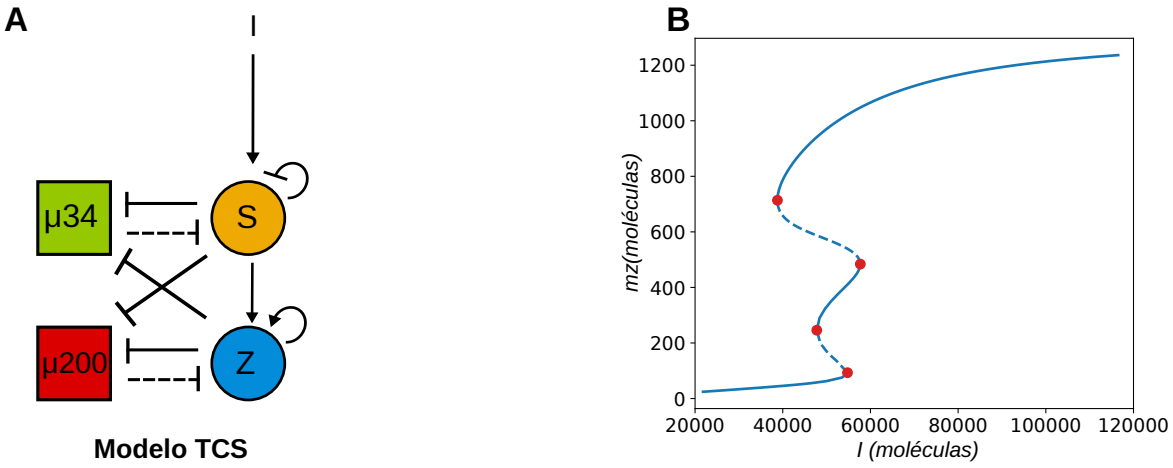

Figura 7.2.: (**A**) Esquema del modelo TCS de la red de regulación génica EMT [205]. Las líneas discontinuas indican regulación por miARNs, mientras que las líneas sólidas representan la regulación por factores de transcripción. (**B**) Diagrama de bifurcación del modelo TCS de la red de regulación génica EMT. Los puntos rojos representan puntos de bifurcación, las líneas sólidas azules representan estados estables y las líneas discontinuas azules son estados inestables.

dinámica de silenciamiento mediada por miARNs, considerando que los miARNs, al unirse al ARNm, pueden inhibir la traducción así como también aumentar la degradación del ARNm, conocida como degradación activa. Consecuentemente, las ecuaciones cinéticas incluyen nuevos tipos de términos que representan funciones reguladoras con efectos inhibidores, como se comenta en la Sección 6.6.3:

$$
\begin{aligned}
\frac{d\left[\mu 200\right]}{dt} &= g_{\mu 200}\, H^g(\left[Z\right], \lambda_{Z\to\mu 200})\, H^g(\left[S\right], \lambda_{S\to\mu 200}) - \left[mz\right] Y_\mu\left(\left[\mu 200\right]\right) - k_{\mu 200}\left[\mu 200\right], \\
\frac{d\left[mz\right]}{dt} &= g_{mz}\, H^g(\left[Z\right], \lambda_{Z\to mz})\, H^g(\left[S\right], \lambda_{[S]\to mz}) - \left[mz\right] Y_m\left(\left[\mu 200\right]\right) - k_{mz}\left[mz\right], \\
\frac{d\left[Z\right]}{dt} &= g_Z\left[mz\right] L\left(\left[\mu 200\right]\right) - k_Z\left[Z\right], \\
\frac{d\left[\mu 34\right]}{dt} &= g_{\mu 34}\, H^g(\left[S\right], \lambda_{S\to\mu 34})\, H^g(\left[Z\right], \lambda_{[Z]\to\mu 34}) - \left[ms\right] Y_\mu\left(\left[\mu 34\right]\right) - k_{\mu 34}\left[\mu 34\right], \\
\frac{d\left[ms\right]}{dt} &= g_{ms}\, H^g(\left[S\right], \lambda_{S\to ms})\, H^g(\left[I\right], \lambda_{I\to ms}) - \left[ms\right] Y_m\left(\left[\mu 34\right]\right) - k_{ms}\left[ms\right], \\
\frac{d\left[S\right]}{dt} &= g_S\left[ms\right] L\left(\left[\mu 34\right]\right) - k_S\left[S\right].
\end{aligned}
\tag{7.2}
$$

I representa la señal externa y las otras variables siguen la misma definición utilizada en el modelo CBS. En el modelo TCS, se diferencian explícitamente las funciones regulatorias cuando la regulación es ejercida por miARNs y factores de transcripción. En las ecuaciones 7.2, las funciones $H^g(\left[i\right], \lambda_{i\to j})$ representan funciones de Hill generalizadas, derivadas en la ecuación C.20 para el caso de dos factores de transcripción y extendidas a la ecuación C.30 para múltiples reguladores, consideradas por los autores como no competitivas con influencia mutua (AND).

En estas funciones, los parámetros $\lambda_{i\to j}$ representan el cambio relativo de la tasa de producción del ARNm j en presencia del factor de transcripción i, respecto a la tasa de transcripción

basal en ausencia de regulación. Los valores de $\lambda_{i\to j} > 1$ corresponden a activadores, mientras que $0 \leq \lambda_{i\to j} < 1$ indica represión. Estos coeficientes reemplazan las tasas máximas de transcripción. Por su parte, las constantes g_j representan las tasas de producción basal del ARNm j, correspondientes a la situación en la que el promotor se encuentra libre. Las funciones $H^g([i], \lambda_{i\to j})$ también dependen de los coeficientes de Hill, aunque los autores no lo describan explícitamente en las ecuaciones.

En el caso de la regulación mediada por miARNs, se consideran dos contribuciones principales: la inhibición de la traducción y la degradación activa de los complejos ARNm-miARN. Para modelar la inhibición de la traducción, se introducen las funciones denotadas por $L([\mu])$. En cuanto a la degradación activa del complejo ARNm-miARN, tanto el ARNm como los miARNs son degradados, lo cual se representa mediante las funciones $Y_m([\mu])$ y $Y_\mu([\mu])$, asociadas al ARNm y al miARN, respectivamente.

Estas funciones de regulación se pueden derivar de manera análoga a las desarrolladas en las Secciones C.2 y C.3, considerando que un ARNm puede tener $i = 1, 2, \ldots, n$ miARNs unidos, formando el complejo ARNm-miARN, donde n es el número de sitios de unión. A diferencia de las funciones de Hill descritas previamente, este enfoque incorpora los estados intermedios, sin realizar suposiciones sobre la cooperatividad. De esta manera, se derivan las funciones de regulación conocidas como ecuaciones de Adair, mencionadas en la Sección C.2, donde la función de Hill aparece como un caso particular. Las funciones de regulación son de la forma:

$$\begin{aligned} L([\mu]) &= \sum_{i=0}^{n} l_i \binom{n}{i} \frac{([\mu]/\mu_0)^i}{(1+[\mu]/\mu_0)^n}, \\ Y_\mu([\mu]) &= \sum_{i=0}^{n} i\,\gamma_{\mu i} \binom{n}{i} \frac{([\mu]/\mu_0)^i}{(1+[\mu]/\mu_0)^n}, \\ Y_m([\mu]) &= \sum_{i=0}^{n} \gamma_{mi} \binom{n}{i} \frac{([\mu]/\mu_0)^i}{(1+[\mu]/\mu_0)^n}, \end{aligned} \tag{7.3}$$

donde las constantes l_i, $\gamma_{\mu i}$ y γ_{mi} representan las tasas de traducción, y de degradación del miARN y del ARNm cuando i miARNs se encuentran formando un complejo con el ARNm, con $i = 0, 1, \ldots, n$. Estas funciones dependen de la concentración del miARN en cuestión (μ). $\binom{n}{i}$ es el número combinatorio de i elementos de un total de n.

Al igual que el modelo CBS, este modelo predice la existencia de tres fenotipos celulares, como se ilustra en la Figura 7.2B. En dicha figura, los estados estables epitelial, mesenquimal e híbrido E/M corresponden a niveles bajos, altos e intermedios del ARNm de ZEB, respectivamente. Además, la EMT ocurre en dos etapas: la primera es la transición reversible del fenotipo epitelial al híbrido E/M, mientras que la segunda es la transición irreversible del fenotipo híbrido E/M al mesenquimal [205]. Si bien la segunda transición es irreversible, sí es posible la transición del fenotipo mesenquimal al epitelial (MET) sin pasar por el estado híbrido, a diferencia de lo que ocurre en el modelo CBS propuesto por Tian y colaboradores [203].

7.3. Red de pluripotencia

En células madre embrionarias la pluripotencia está regulada por una red central de factores de transcripción, donde OCT4, SOX2 y NANOG juegan un papel crucial. Estas células pueden mantenerse en un estado indiferenciado a través de interacciones complejas que incluyen bucles de retroalimentación entre los tres factores. El circuito OCT4-SOX2-NANOG se ha estudiado ampliamente en el contexto de las células madre embrionarias para comprender cómo la dinámica de expresión génica puede influir en las decisiones celulares. La heterogeneidad en NANOG, en particular, ha sido el objeto de estudio de trabajos tanto experimentales como computacionales, ya que se asocia directamente con la capacidad de las células madre para mantener su pluripotencia [219].

El circuito formado por OCT4, SOX2 y NANOG, también cumple un rol esencial en el contexto del cáncer [220-222]. Se ha observado que NANOG se expresa en niveles significativamente más altos en las CSCs en comparación con otras células tumorales en diversos tipos de cáncer [159, 169, 173-180], lo que sugiere que NANOG es un candidato para la identificación de CSCs en cáncer. Por otro lado, estudios experimentales han identificado una relación entre la pluripotencia y los niveles intermedios de OCT4, ya que tanto niveles altos como bajos de OCT4 conducen a diferentes trayectorias de diferenciación [181-183].

En 2006, Chickarmane y colaboradores desarrollaron el primer modelo de la red de regulación génica de la pluripotencia [223]. Este modelo se basa en el formalismo de Shea-Ackers [224], utilizando ecuaciones diferenciales ordinarias para describir las interacciones entre los factores clave de pluripotencia OCT4, SOX2 y NANOG. El modelo incluye bucles de retroalimentación positiva, donde OCT4 y SOX2 forman un heterodímero (OCT4-SOX2) que activa a NANOG, y, a su vez, tanto el dímero como NANOG activan a OCT4 y SOX2. Además, NANOG presenta autoactivación. El circuito también incorpora dos señales externas, así como marcadores de pluripotencia y diferenciación. En el formalismo de Shea Ackers, la producción de proteínas es modelada desde una perspectiva de mecánica estadística. Se basa en la formulación de una función de partición que enumera todos los estados posibles del promotor, es decir, las combinaciones de proteínas que pueden unirse al gen regulado. A cada estado se le asocia una energía libre de unión específica. Este modelo predice un comportamiento biestable, en el que el sistema puede adoptar un estado pluripotente, caracterizado por niveles elevados de NANOG, o un estado diferenciado, asociado a niveles bajos de esta proteína, dependiendo de las señales externas. En un trabajo posterior, publicado en el año 2008, el mismo grupo amplió el modelo para incluir otros genes relevantes, como CDX, GATA-6 y GCNF [225].

En 2009, Kalmar y colegas propusieron un sistema excitable para modelar la pluripotencia en células madre embrionarias, en el cual las fluctuaciones impulsadas por ruido transcripcional permiten que las células cambien de un estado de NANOG alto a uno bajo [226]. Los autores asumen que SOX2 actúa en forma conjunta con OCT4 en la red regulatoria. Este modelo considera regulaciones mutuas y autorreguladoras entre NANOG y OCT4. El modelo asume que OCT4 activa a NANOG a niveles bajos, mientras que niveles altos de OCT4 inhiben a NANOG, y NANOG activa OCT4. Además, se considera que tanto OCT4 como NANOG presentan autorregulación positiva. Este modelo predice la estabilidad de un único estado de NANOG alto, correspondiente al estado pluripotente. Pequeñas fluctuaciones pueden generar saltos en el espacio de fases a niveles bajos de NANOG, en los cuales la célula es más propensa a diferenciarse, que luego de un tiempo retornan al único estado estable.

A continuación, se presenta en detalle un modelo de la red de regulación de la pluripotencia, que se considera relevante para la tesis.

7.3.1. El modelo de Glauche

En 2010, Glauche y colaboradores propusieron un modelo de la red de regulación génica de la pluripotencia [227], el cual reproduce datos experimentales de citometría de flujo para células madre embrionarias [226]. En este trabajo se propone un modelo de bucles de retroalimentación positiva del circuito central OCT4-SOX2-NANOG. En dicho circuito, OCT4 y SOX2 forman un heterodímero que, a su vez, activa la producción de ambos factores de transcripción. Además, el dímero también promueve la activación de NANOG, el cual presenta un mecanismo de autoactivación (ver Figura 7.3A). Al considerar un equilibrio dinámico entre el heterodímero, OCT4 y SOX2, y que el heterodímero es el principal regulador de NANOG, los autores consideran solamente la ecuación para el heterodímero y para NANOG. Así, la dinámica del sistema se describe mediante el siguiente conjunto de ecuaciones diferenciales basadas en la cinética de Hill:

$$\begin{aligned} \frac{d[OS]}{dt} &= u\left(\frac{s_1[OS]^{n_{OS}}}{(k_1+[OS]^{n_{OS}})\cdot d_O}\cdot\frac{s_2[OS]^{n_{OS}}}{(k_2+[OS]^{n_{OS}})\cdot d_S}\right) - d_{OS}\cdot[OS], \\ \frac{d[N]}{dt} &= \frac{s_4[N]^{n_N}}{k_4+[N]^{n_N}} + \frac{s_3[OS]^{n_{OS}}}{k_3+[OS]^{n_{OS}}} - d_N\cdot[N] \end{aligned} \tag{7.4}$$

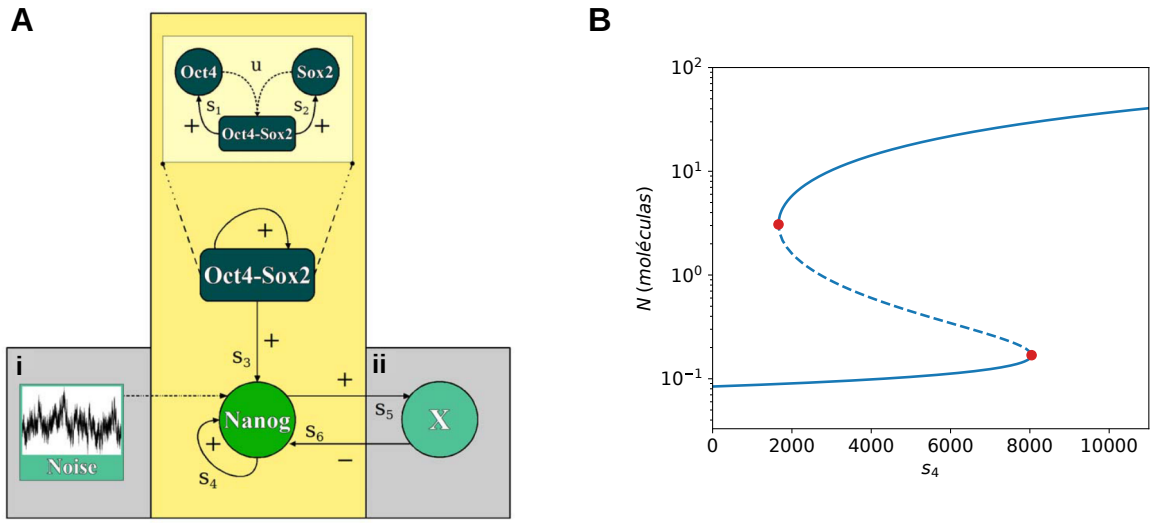

Figura 7.3.: (**A**) Esquema de la red regulatoria del modelo de Glauche *et al.* (adaptación de [227]). (**B**) Diagrama de bifurcación de NANOG (N) variando el parámetro de autorregulación de NANOG (s_4). Los puntos rojos representan puntos de bifurcación, las líneas sólidas azules representan puntos fijos estables y la línea discontinua azul corresponde a puntos fijos inestables.

Las variables OS y N, representan la concentración del heterodímero y de NANOG, respectivamente. El análisis de bifurcación del sistema revela biestabilidad (NANOG alto y NANOG bajo) en función del parámetro que controla la magnitud de la autorregulación de NANOG como se puede apreciar en la Figura 7.3B. Al introducir un término estocástico de ruido blanco en la ecuación de NANOG (cuadro i en la Figura 7.3A), mediante el método de Euler-Maruyama, resuelven el sistema de ecuaciones diferenciales estocásticas, y encuentran que la introducción del ruido blanco permite el salto reversible entre ambos estados estables. A esta implementación del modelo la llaman escenario de fluctuaciones. Por otro lado, prueban un enfoque alternativo, en el cual adicionan un factor de transcripción que es activado por

NANOG y que a su vez lo inhibe (cuadro ii en la Figura 7.3A). En este caso, para ciertos valores de parámetros, encuentran un comportamiento oscilatorio en el espacio de fases de forma periódica. A esta implementación la denominan escenario de oscilación. Ambas alternativas permiten obtener distribuciones de los niveles de NANOG y encuentran que ambos modelos son capaces de reproducir la distribución experimental de NANOG. Posteriormente, se evaluó experimentalmente la plausibilidad de ambas implementaciones [228] y se han realizado otros trabajos basados en este modelo [229, 230].

7.4. Metodología

7.4.1. Metodología computacional

El desarrollo de los modelos de redes de regulación génica se basó en el formalismo matemático detallado en la Sección 6.6 del capítulo introductorio. Para analizar el comportamiento de los modelos se utilizaron las herramientas de sistemas dinámicos mencionadas en la Sección 6.7. Para ello, se empleó la librería *PyDSTool* (versión 1.21.6), una librería de Python que proporciona herramientas computacionales para el desarrollo y análisis de sistemas dinámicos [231]. *PyDSTool*, a su vez, utiliza *scipy* (versión 1.7.3), el análisis fue implementado en Python 3.7.12.

En particular, se utilizó la clase `ContClass` (por *continuation class*) para el análisis de bifurcación. Se utilizó el integrador `VODE` a través de la clase `Generator.Vode_ODEsystem` para la integración numérica, que a su vez emplea `CVODE` de la librería SciPy. Adicionalmente, se emplearon algunas herramientas del módulo `phaseplane` para identificar puntos fijos, analizar su estabilidad (`find_fixedpoints`) y calcular las nulclinas (`find_nullclines`).

7.4.2. Modelo de EMT

Se desarrolló un modelo de red de regulación génica basado en el modelo CBS, el cual se utiliza en el siguiente capítulo para explorar su relación con la red de regulación génica de pluripotencia. La justificación para utilizar este modelo como punto de partida se debe a su simplicidad en comparación con el modelo TCS. En este sentido, el modelo CBS emplea funciones de Hill tradicionales como funciones reguladoras para modelar la regulación por factores de transcripción, mientras que el modelo TCS utiliza funciones de Hill generalizadas.

Por otro lado, el modelo TCS incorpora una descripción detallada de la regulación mediada por miARNs. Este considera que un ARNm puede unirse a múltiples moléculas de miARN, lo que da lugar a funciones de regulación más complejas (ecs. 7.3). Dada la cantidad de ecuaciones diferenciales, estas funciones introducen una cantidad significativamente mayor de parámetros, que actualmente no son accesibles experimentalmente, lo cual añade complejidad al modelo sin necesariamente aportar una mayor claridad en los procesos biológicos. Otro aspecto relevante que desmotivó la utilización de este modelo como punto de partida es la inclusión de la autorregulación de ZEB, para la cual no existe evidencia experimental. Los autores optan por incorporarla debido a que, en su ausencia (y además con alta no linealidad, ya que el coeficiente de Hill debe ser mayor a 3), el modelo no predice triestabilidad.

Aunque el modelo CBS no considera la degradación activa de los ARNm por los miARNs, lo que constituye una limitación del mismo, sus parámetros han sido ajustados para reproducir resultados experimentales. Además, posteriormente se validó la predicción de que el módulo SNAIL/miR-34 regula la transición $E \rightarrow E/M$, mientras que el módulo ZEB/miR-200 controla la transición $E/M \rightarrow M$ [208]. Para los fines de esta tesis, el modelo CBS ofrece un equilibrio más favorable entre simplicidad y capacidad predictiva, lo que lo convierte en una elección adecuada para el análisis propuesto.

En las ecuaciones 7.1 del modelo CBS, se observa una distinción clara entre dos tipos de regulación: la mediada por factores de transcripción, que controlan la producción de ARNm, y la mediada por miARNs, que regulan la producción de proteínas. Esta distinción se aplica a la mayoría de las interacciones, con la excepción del TGF-β. Para incorporar explícitamente la regulación del TGF-β por el miR-200, de manera coherente con el tratamiento de otros elementos del sistema, es necesario modelar tanto la transcripción como la traducción del TGF-β. Por lo tanto, se introdujo una ecuación diferencial adicional al sistema que describe la dinámica del ARNm del TGF-β, denotado como mt. La represión del TGF-β por el miR-200 se implementó inhibiendo la traducción del factor de transcripción, tal como se muestra a continuación:

$$
\begin{aligned}
\frac{d\left[\mu 200\right]}{dt} &= k^0_{\mu 200} + \beta_{S,Z\rightarrow\mu 200}\frac{1}{1+\left(\frac{[S]}{K_{S\rightarrow\mu 200}}\right)^{n_{S\rightarrow\mu 200}}+\left(\frac{[Z]}{K_{Z\rightarrow\mu 200}}\right)^{n_{Z\rightarrow\mu 200}}} - d_{\mu 200}\left[\mu 200\right],\\
\frac{d\left[mz\right]}{dt} &= k^0_{mz} + \beta_{mz}\frac{\left(\frac{[S]}{K_{S\rightarrow mz}}\right)^{n_{S\rightarrow mz}}}{1+\left(\frac{[S]}{K_{S\rightarrow ms}}\right)^{n_{S\rightarrow mz}}} - d_{mz}\left[mz\right],\\
\frac{d\left[Z\right]}{dt} &= \beta_{Z}\left[mz\right]\frac{1}{1+\left(\frac{[\mu 200]}{K_{\mu 200\rightarrow Z}}\right)^{n_{\mu 200\rightarrow Z}}} - d_{Z}\left[Z\right],\\
\frac{d\left[\mu 34\right]}{dt} &= k^0_{\mu 34} + \beta_{S,Z\rightarrow\mu 34}\frac{1}{1+\left(\frac{[S]}{K_{S\rightarrow\mu 34}}\right)^{n_{S\rightarrow\mu 34}}+\left(\frac{[Z]}{K_{Z\rightarrow\mu 34}}\right)^{n_{Z\rightarrow\mu 34}}} - d_{\mu 34}\left[\mu 34\right],\\
\frac{d\left[ms\right]}{dt} &= k^0_{ms} + \beta_{ms}\frac{\left(\frac{[T]+T^0}{K_{T\rightarrow ms}}\right)^{n_{T\rightarrow ms}}}{1+\left(\frac{[T]+T^0}{K_{T\rightarrow ms}}\right)^{n_{T\rightarrow ms}}} - d_{ms}\left[ms\right],\\
\frac{d\left[S\right]}{dt} &= \beta_{S}\left[ms\right]\frac{1}{1+\left(\frac{[\mu 34]}{K_{\mu 34\rightarrow S}}\right)^{n_{\mu 34\rightarrow S}}} - d_{S}\left[S\right],\\
\frac{d\left[mt\right]}{dt} &= k^0_{mt} - d_{mt}\left[mt\right],\\
\frac{d\left[T\right]}{dt} &= \beta_{T}\left[mt\right]\frac{1}{1+\left(\frac{[\mu 200]}{K_{\mu 200\rightarrow T}}\right)^{n_{\mu 200\rightarrow T}}} - d_{T}\left[T\right],\\
\frac{d\left[ECad\right]}{dt} &= \beta_{S\rightarrow E}\frac{1}{1+\left(\frac{[S]}{K_{S\rightarrow E}}\right)^{n_{S\rightarrow E}}} + \beta_{Z\rightarrow E}\frac{1}{1+\left(\frac{[Z]}{K_{Z\rightarrow E}}\right)^{n_{Z\rightarrow E}}} - d_{E}\left[ECad\right],\\
\frac{d\left[NCad\right]}{dt} &= \beta_{S\rightarrow N}\frac{\left(\frac{[S]}{K_{S\rightarrow N}}\right)^{n_{S\rightarrow N}}}{1+\left(\frac{[S]}{K_{S\rightarrow N}}\right)^{n_{S\rightarrow N}}} + \beta_{Z\rightarrow N}\frac{\left(\frac{[Z]}{K_{Z\rightarrow N}}\right)^{n_{Z\rightarrow N}}}{1+\left(\frac{[Z]}{K_{Z\rightarrow N}}\right)^{n_{Z\rightarrow N}}} - d_{N}\left[NCad\right].
\end{aligned}
\tag{7.5}
$$

Donde el valor de todos los parámetros se mantuvo respecto al modelo CBS (ec. 7.1), y se introdujeron nuevos parámetros asociados a la dinámica del ARNm y de la proteína del TGB-β

detallados en la Tabla D.1 del Apéndice D. En la Figura 7.5A se esquematiza la arquitectura de la red.

7.4.3. Modelo de pluripotencia

Hasta el momento, los modelos de la red regulatoria de pluripotencia han sido descritos mediante ecuaciones diferenciales en las cuales las únicas variables son las proteínas. Como se introdujo previamente, el objetivo es integrar los módulos de las redes regulatorias de la EMT y de pluripotencia en un único circuito. Para ello, ambos modelos deben ser comparables y estar enmarcados en el mismo contexto teórico. Los modelos de EMT consideran tanto la regulación por factores de transcripción como por miARNs. Por lo tanto, para incorporar la regulación en ambos niveles, es necesario que el modelo de la red de pluripotencia también contemple explícitamente mensajeros y proteínas como variables independientes.

Para el desarrollo del modelo de la red regulatoria de pluripotencia, se tomó como base el modelo propuesto por Glauche y colaboradores [227], dado que utiliza un formalismo similar al empleado en el modelo de EMT. La principal modificación realizada fue la incorporación explícita de los procesos de transcripción y traducción, modelando la dinámica tanto del ARNm como de las proteínas. En la Figura 7.4A se ilustra un esquema de dicha red regulatoria. Este diagrama muestra únicamente los elementos a nivel proteína.

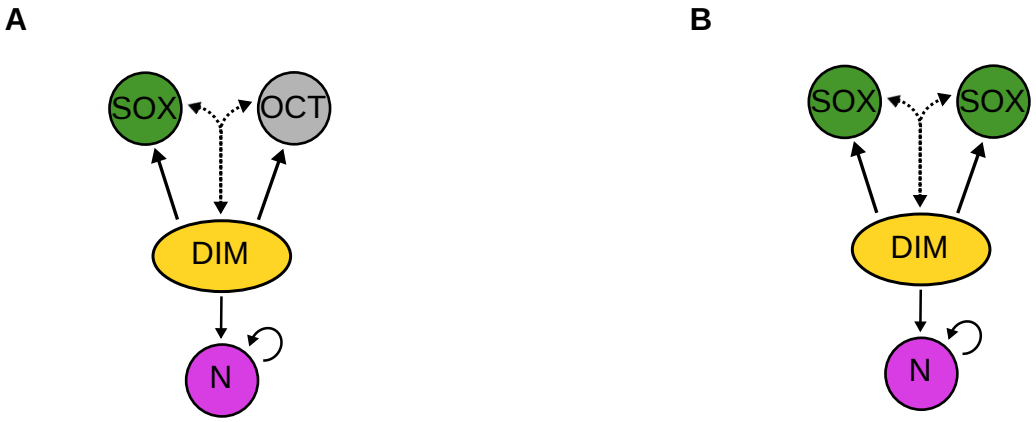

Figura 7.4.: (**A**) Esquema del modelo de la red de regulación génica de pluripotencia con base en la arquitectura propuesta por Glauche y colegas [227]. La línea de puntos representa la dimerización y desdimerización. Las líneas sólidas representan regulación transcripcional. (**B**) Esquema del modelo simplificado de la misma red, con la misma convención para las líneas.

Las proteínas OCT4 y SOX2 forman un dímero, por lo que una mayor concentración de ambos favorece la formación del dímero. Sin embargo, si la concentración de alguna de las dos proteínas es baja, la formación del dímero se verá reducida. El proceso de dimerización se representa mediante una línea de puntos en la Figura 7.4A. El dímero resultante activa la transcripción de SOX2 y NANOG, mientras que NANOG, a su vez, estimula la transcripción de sí mismo. Todas las regulaciones mediante factores de transcripción se consideraron a nivel de ARNm, como se muestra en el sistema de ecuaciones diferenciales que describe la dinámica del circuito:

$$
\begin{aligned}
\frac{d\left[moct\right]}{dt} &= k^0_{moct} + \beta_{moct}\frac{\left(\frac{[DIM]}{K_{DIM\to moct}}\right)^{n_{DIM\to moct}}}{1+\left(\frac{[DIM]}{K_{DIM\to moct}}\right)^{n_{DIM\to moct}}} - d_{moct}\left[moct\right],\\
\frac{d\left[OCT\right]}{dt} &= \beta_{OCT}\left[moct\right] - k_{on}\left[OCT\right]\left[SOX\right] + k_{off}\left[DIM\right] - d_{OCT}\left[OCT\right],\\
\frac{d\left[msox\right]}{dt} &= k^0_{msox} + \beta_{msox}\frac{\left(\frac{[DIM]}{K_{DIM\to msox}}\right)^{n_{DIM\to msox}}}{1+\left(\frac{[DIM]}{K_{DIM\to msox}}\right)^{n_{DIM\to msox}}} - d_{msox}\left[msox\right],\\
\frac{d\left[SOX\right]}{dt} &= \beta_{SOX}\left[msox\right] - k_{on}\left[OCT\right]\left[SOX\right] + k_{off}\left[DIM\right] - d_{SOX}\left[SOX\right],\\
\frac{d\left[DIM\right]}{dt} &= k_{on}\left[OCT\right]\left[SOX\right] - k_{off}\left[DIM\right] - d_{DIM}\left[DIM\right],\\
\frac{d\left[mn\right]}{dt} &= k^0_{mn} + \beta_{DIM\to mn}\frac{\left(\frac{[DIM]}{K_{DIM\to mn}}\right)^{n_{DIM\to mn}}}{1+\left(\frac{[DIM]}{K_{DIM\to mn}}\right)^{n_{DIM\to mn}}} + \beta_{N\to mn}\frac{\left(\frac{[N]}{K_{N\to mn}}\right)^{n_{N\to mn}}}{1+\left(\frac{[N]}{K_{N\to mn}}\right)^{n_{N\to mn}}} - d_{mn}\left[mn\right],\\
\frac{d\left[N\right]}{dt} &= \beta_N\left[mn\right] - d_N\left[N\right].
\end{aligned}
\tag{7.6}
$$

$moct$, $msox$ y mn representan el ARNm de los genes oct4, sox2 y nanog. OCT, SOX y N representan las proteínas OCT4, SOX2 y NANOG, y DIM es el dímero formado por una molécula de OCT4 y una molécula de SOX2. La regulación, tanto por factores de transcripción como por miARNs, se implementó mediante cinética de Hill de forma aditiva no competitiva. Los valores de los parámetros fueron adaptados y seleccionados de forma tal que tengan órdenes de magnitud similares a los utilizados en el modelo de la red de EMT presentado en la sección anterior. A continuación, se describen los parámetros de las ecuaciones.

- k^0_{moct}, k^0_{msox} y k^0_{mn}: tasas de transcripción basal de $moct$, $msox$ y mn. Estas representan la producción constitutiva de ARNm.

- β_{moct}, β_{msox}, $\beta_{DIM\to mn}$ y $\beta_{N\to mn}$: tasas de transcripción máximas de $moct$, $msox$ y mn. Las tasas β_{moct} y β_{msox} son las tasas máximas de transcripción de $moct$ y $msox$ dependientes del dímero, mientras que $\beta_{DIM\to mn}$ y $\beta_{N\to mn}$ corresponden a la transcripción máxima de mn dependiente del dímero y de NANOG, respectivamente.

- $K_{DIM\to moct}$, $K_{DIM\to msox}$, $K_{DIM\to mn}$ y $K_{N\to mn}$: constantes de saturación media de las funciones de Hill para las respectivas regulaciones, que determinan la concentración de reguladores necesarios para activar la mitad de la tasa de transcripción máxima. Por ejemplo, $K_{DIM\to msox}$ representa la constante para la regulación del mensajero de SOX2 por el dímero.

- $n_{DIM\to moct}$, $n_{DIM\to msox}$, $n_{DIM\to mn}$ y $n_{N\to mn}$: coeficientes de Hill de las distintas regulaciones.

- β_{OCT}, β_{SOX} y β_N: tasas de traducción de OCT4, SOX2 y NANOG a partir de sus respectivos ARNm. Indican cuánta proteína se produce por unidad de ARNm.

- d_{moct}, d_{OCT}, d_{msox}, d_{SOX}, d_{DIM}, d_{mn} y d_N, : constantes de degradación de los ARNm y las proteínas. Representan la tasa a la que se degrada cada molécula. Por ejemplo, d_{SOX} es la tasa de degradación de SOX2.

- k_{on} y k_{off}: constantes que representan la tasa de dimerización y disociación del dímero.

Como se menciona previamente, la formación del dímero requiere la disponibilidad de OCT4 y SOX2. En este contexto, fue posible realizar una simplificación del modelo considerando únicamente uno de estos factores y suponiendo que el dímero se genera por la unión de dos moléculas idénticas de dicho reactivo. Esta aproximación se justifica dado que la interacción entre OCT4 y SOX2 es estrictamente necesaria para la formación del dímero, lo que permitió modelar el proceso como dependiente de un único precursor. En la Figura 7.4B se ilustra un esquema de la arquitectura del modelo propuesto, donde dos moléculas del mismo reactivo (SOX2) pueden dimerizarse, el resto de las interacciones del modelo se mantuvieron respecto al modelo esquematizado en la Figura 7.4A.

Además, dado que los valores de los parámetros y las funciones de activación de OCT4 y SOX2 en el modelo original son exactamente iguales, el sistema presenta una simetría completa en relación con estos factores de transcripción. La formulación del modelo mediante ecuaciones diferenciales se puede reescribir de la siguiente forma:

$$
\begin{aligned}
\frac{d\left[msox\right]}{dt} &= k^0_{msox} + \beta_{msox}\frac{\left(\frac{[DIM]}{K_{DIM\to msox}}\right)^{n_{DIM\to msox}}}{1+\left(\frac{[DIM]}{K_{DIM\to msox}}\right)^{n_{DIM\to msox}}} - d_{msox}\left[msox\right],\\
\frac{d\left[SOX\right]}{dt} &= \beta_{SOX}\left[msox\right] - 2k_{on}\left[SOX\right]\left[SOX\right] + 2k_{off}\left[DIM\right] - d_{SOX}\left[SOX\right],\\
\frac{d\left[DIM\right]}{dt} &= k_{on}\left[SOX\right]\left[SOX\right] - k_{off}\left[DIM\right] - d_{DIM}\left[DIM\right],\\
\frac{d\left[mn\right]}{dt} &= k^0_{mn} + \beta_{DIM\to mn}\frac{\left(\frac{[DIM]}{K_{DIM\to mn}}\right)^{n_{DIM\to mn}}}{1+\left(\frac{[DIM]}{K_{DIM\to mn}}\right)^{n_{DIM\to mn}}} + \beta_{N\to mn}\frac{\left(\frac{[N]}{K_{N\to mn}}\right)^{n_{N\to mn}}}{1+\left(\frac{[N]}{K_{N\to mn}}\right)^{n_{N\to mn}}} - d_{mn}\left[mn\right],\\
\frac{d\left[N\right]}{dt} &= \beta_N\left[mn\right] - d_N\left[N\right].
\end{aligned}
\tag{7.7}
$$

De este modo, se redujo el sistema de ecuaciones diferenciales ordinarias al eliminar las ecuaciones correspondientes a la proteína OCT4 y su ARNm. Los valores de todos los parámetros se mantuvieron respecto al modelo previo y son detallados en la Tabla D.3 del Apéndice D. En la ecuación correspondiente a SOX2, se introdujo un factor 2 en los términos asociados a la dimerización y desdimerización, dado que el dímero está compuesto por dos moléculas de SOX2.

7.5. Resultados y discusiones

7.5.1. Modelo de EMT

El análisis de bifurcación del modelo mostrado en la Figura 7.5B revela un comportamiento prácticamente equivalente al modelo original. Sin embargo, es fundamental destacar la importancia de mantener la consistencia en la modelización, ya que la distinción entre la regulación mediada por factores de transcripción y la regulada por miARNs podría tener implicancias al incorporar más elementos al sistema.

En la Figura 7.5C se presentan soluciones del sistema de ecuaciones 7.5 para valores de la señal externa T^0 y distintas condiciones iniciales detalladas en la Tabla D.2 del Apéndice D. Las dos condiciones iniciales corresponden al estado epitelial pero presentan concentraciones

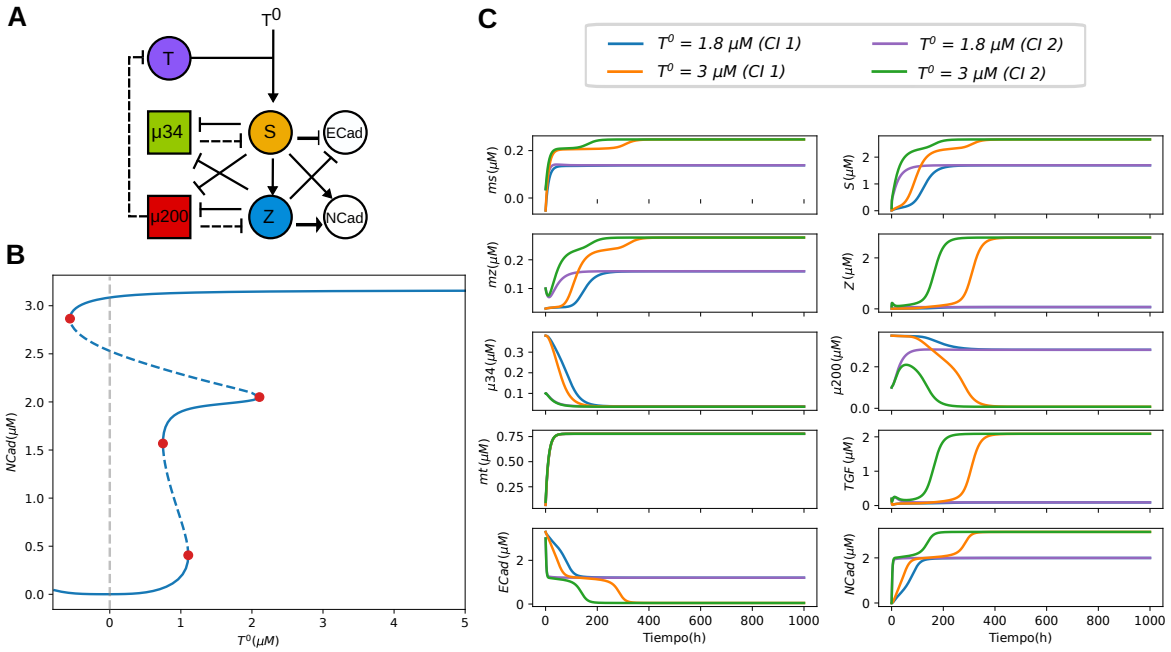

Figura 7.5.: (**A**) Modelo de la red de regulación génica de la EMT basado en el modelo CBS [203]. (**B**) Diagrama de bifurcación de la N-cadherina (marcador de células mesenquimales) en función de la concentración del TGF-β exógeno. Los puntos rojos representan puntos de bifurcación, las líneas sólidas azules representan puntos fijos estables y la discontinua azul corresponde a puntos fijos inestables. La línea discontinua vertical gris indica la concentración del TGF-β exógeno $T^0 = 0$. Valores menores a 0 no tienen sentido biológico. (**C**) Evolución temporal de la concentración de los elementos de la red para distintas condiciones iniciales (CI) y valores de la señal externa T^0 (TGF-β exógeno).

distintas de algunos reguladores. Tales diferencias generan un cambio en el comportamiento transitorio del sistema, pero el estado estacionario es el mismo, para igual valor de la señal externa T^0. Los cambios en la concentración de TGF-β exógeno determinan si el sistema va a un estado final híbrido E/M (T^0 moderado) o mesenquimal (T^0 alto). A partir de las curvas de E-cadherina y N-cadherina se puede observar que la EMT ocurre en dos etapas, primero una transición desde el fenotipo epitelial al híbrido E/M ($E \rightarrow E/M$) y una segunda transición desde el fenotipo híbrido E/M al mesenquimal ($E/M \rightarrow M$).

A partir de las curvas azul y violeta, cuando la concentración de TGF-β exógeno es moderada ($T^0 = 1{,}8\ \mu M$), se observa un aumento rápido en la concentración de ARNm de SNAIL. Sin embargo, el aumento del factor de transcripción SNAIL solo se produce cuando disminuye la represión por parte de miR-34. En el caso de la curva violeta, esto ocurre más rápidamente debido a que la concentración inicial de miR-34 es menor en comparación con la curva azul. En estas condiciones, tanto ZEB como miR-200 no muestran variaciones considerables, y la expresión exógena de TGF-β permanece constante. Este estado corresponde con una represión parcial de la E-cadherina y a una activación parcial de la N-cadherina, características del estado híbrido E/M.

Al incrementar la señal externa de TGF-β exógeno ($T^0 = 3\ \mu M$), como se observa en las curvas amarilla y verde de la Figura 7.5C, se induce una mayor expresión de SNAIL, lo que a su vez activa la transcripción del ARNm de ZEB. La traducción de ZEB ocurre posteriormente, cuando la concentración de miR-200 disminuye. En el caso de la curva verde, esto sucede

antes que en la curva amarilla, ya que la concentración inicial de miR-200 es menor. La expresión conjunta de SNAIL y ZEB promueve la expresión del marcador mesenquimal N-cadherina, mientras que reprime la expresión de E-cadherina, lo que provoca la transición de la célula al fenotipo mesenquimal. Finalmente, la expresión autócrina de TGF-β se activa una vez que la inhibición de su traducción por miR-200 es eliminada.

Para profundizar en el estudio del comportamiento de la red de regulación génica asociada a la EMT, es posible analizar por separado los circuitos miR-34/SNAIL y miR-200/ZEB. El primer módulo, miR-34/SNAIL, constituye un bucle de retroalimentación negativa en el que el factor de transcripción SNAIL reprime la transcripción de miR-34, mientras que miR-34 inhibe la traducción de SNAIL. Además, el TGF-β exógeno desempeña un papel clave al activar la transcripción de SNAIL. Este circuito se representa esquemáticamente en la Figura 7.6A.

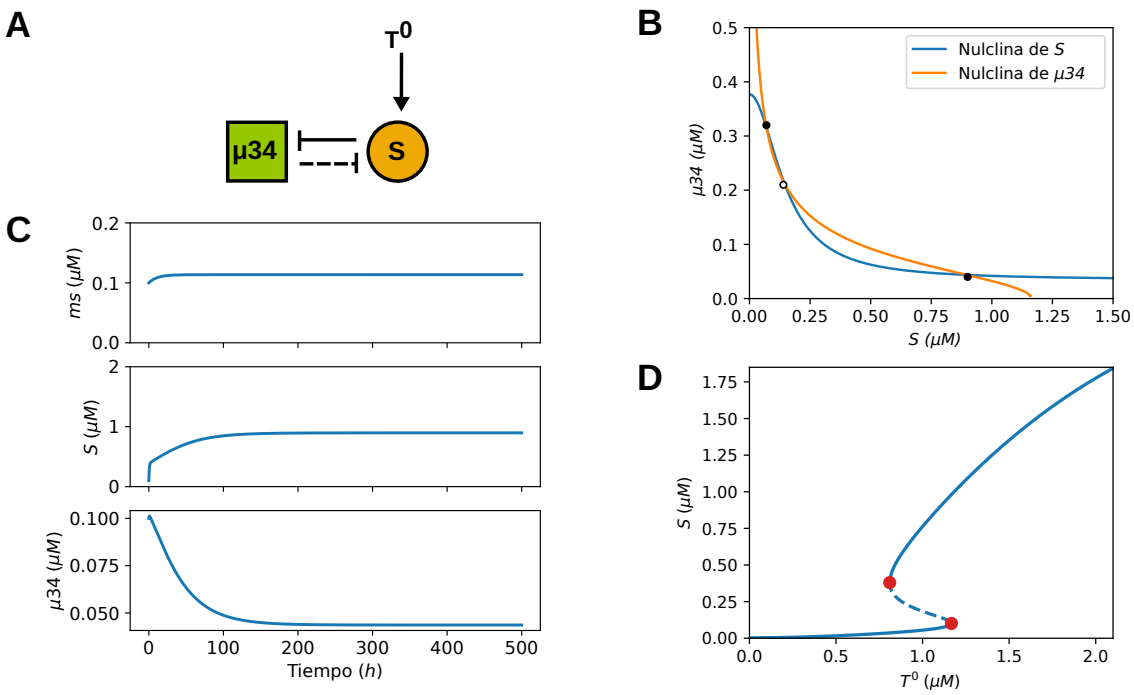

Figura 7.6.: **(A)** Esquema del módulo miR-34/SNAIL de la GRN de EMT. **(B)** Nulclinas de SNAIL y de miR-34 ($T^0 = 1{,}1\ \mu M$). Los puntos negros sólidos y blancos representan los puntos fijos estables e inestables, respectivamente. **(C)** Evolución temporal de la concentración de los elementos de la red ($T^0 = 1{,}1\ \mu M$). **(D)** Diagrama de bifurcación de S en función del parámetro de entrada T^0. Los puntos rojos representan puntos de bifurcación, las líneas sólidas azules representan puntos fijos estables y la línea discontinua azul corresponde a puntos fijos inestables.

El comportamiento dinámico de este sistema se puede describir mediante el siguiente conjunto de ecuaciones diferenciales:

$$
\begin{aligned}
\frac{d\,[\mu 34]}{dt} &= k^0_{\mu 34} + \beta_{S\to\mu 34}\frac{1}{1+\left(\frac{[S]}{K_{S\to\mu 34}}\right)^{n_{S\to\mu 34}}} - d_{\mu 34}\,[\mu 34]\,,\\
\frac{d\,[ms]}{dt} &= k^0_{ms} + \beta_{ms}\frac{\left(\frac{T^0}{K_{T\to ms}}\right)^{n_{T\to ms}}}{1+\left(\frac{T^0}{K_{T\to ms}}\right)^{n_{T\to ms}}} - d_{ms}\,[ms]\,,\\
\frac{d\,[S]}{dt} &= \beta_S\,[ms]\,\frac{1}{1+\left(\frac{[\mu 34]}{K_{\mu 34\to S}}\right)^{n_{\mu 34\to S}}} - d_S\,[S]\,.
\end{aligned}
\tag{7.8}
$$

Se mantuvo el valor de los parámetros utilizados en el circuito EMT completo; en el caso de $\beta_{S\to\mu 34}$ se utilizó el valor de $\beta_{S,Z\to\mu 200}$ que se empleó en el circuito completo. En la Figura 7.6B se muestran las nulclinas de SNAIL (curva azul) y miR-34 (curva amarilla) para un valor fijo del parámetro de entrada ($T^0 = 1{,}1\ \mu M$). La intersección de estas nulclinas define los puntos fijos del sistema, que en este caso son tres: dos estables, indicados con puntos negros sólidos, y uno inestable, representado con un punto blanco con borde negro. El punto fijo estable asociado a valores bajos de SNAIL se interpreta como el estado epitelial, mientras que el punto fijo estable correspondiente a un valor intermedio de SNAIL representa el estado híbrido E/M. En la Figura 7.6C se muestra una solución del sistema de ecuaciones en la cual se observa la evolución temporal de las concentraciones del mensajero de SNAIL, el factor de transcripción SNAIL y el miR-200. Para estas condiciones iniciales el sistema converge al estado híbrido. En la Figura 7.6D se muestra el análisis de bifurcación respecto al parámetro T^0 en el cual se observa biestabilidad, caracterizada por la coexistencia de estados epiteliales (valores bajos de SNAIL) y estados híbridos (valores intermedios de SNAIL).

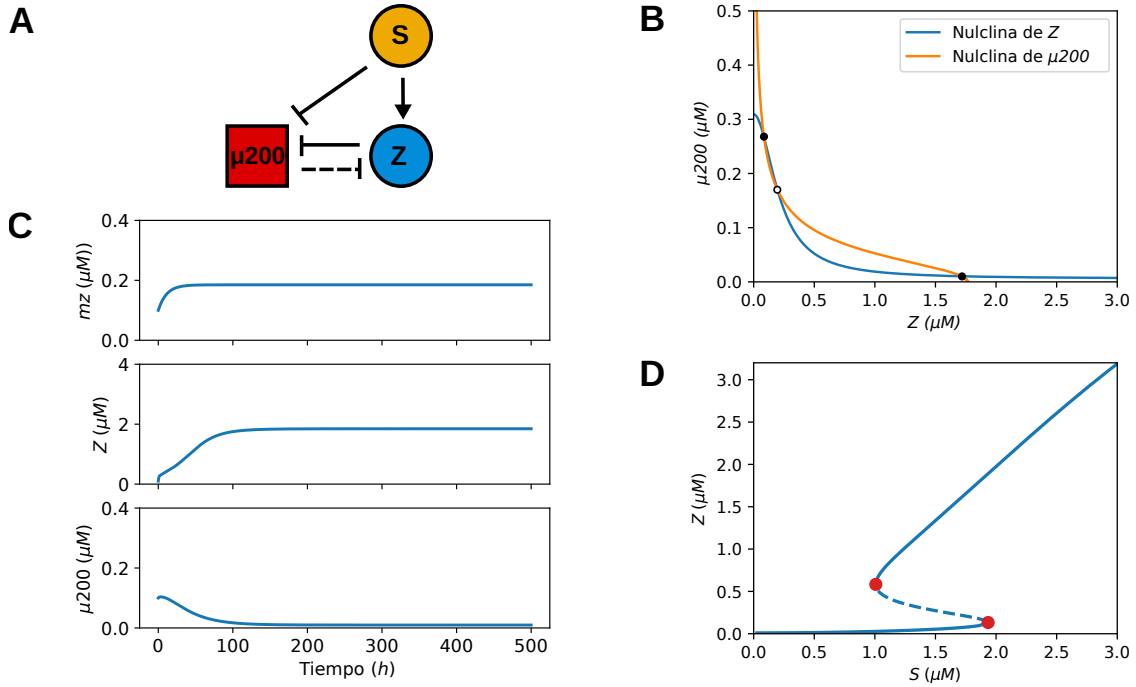

Figura 7.7.: (**A**) Esquema del módulo miR-200/ZEB de la GRN de EMT. (**B**) Nulclinas de ZEB y de miR-200. **(C)** Evolución temporal de la concentración de los elementos de la red. **(D)** Diagrama de bifurcación de Z en función del parámetro de entrada S. Los puntos rojos representan puntos de bifurcación, las líneas sólidas azules representan puntos fijos estables y la línea de puntos azul corresponde a puntos fijos inestables.

Para analizar el segundo módulo, miR-200/ZEB, se consideró a SNAIL como parámetro de entrada, el cual activa la expresión de ZEB y reprime la de miR-200. A su vez, miR-200 y ZEB conforman un bucle de retroalimentación negativa, cuyo esquema se ilustra en la Figura 7.7A. La dinámica de este módulo puede modelarse mediante un sistema de ecuaciones diferenciales ordinarias:

$$\begin{aligned}
\frac{d\,[\mu 200]}{dt} &= k^0_{\mu 200} + \beta_{S,Z\to\mu 200}\frac{1}{1+\left(\frac{[S]}{K_{S\to\mu 200}}\right)^{n_{S\to\mu 200}}+\left(\frac{[Z]}{K_{Z\to\mu 200}}\right)^{n_{Z\to\mu 200}}} - d_{\mu 200}\,[\mu 200]\,,\\
\frac{d\,[mz]}{dt} &= k^0_{mz} + \beta_{mz}\frac{\left(\frac{[S]}{K_{S\to mz}}\right)^{n_{S\to mz}}}{1+\left(\frac{[S]}{K_{S\to ms}}\right)^{n_{S\to mz}}} - d_{mz}\,[mz]\,,\\
\frac{d\,[Z]}{dt} &= \beta_Z\,[mz]\,\frac{1}{1+\left(\frac{[\mu 200]}{K_{\mu 200\to Z}}\right)^{n_{\mu 200\to Z}}} - d_Z\,[Z]\,.
\end{aligned} \tag{7.9}$$

En las ecs. 7.9 se mantuvo el valor de los parámetros utilizados para el circuito EMT completo. Aplicando el método gráfico basado en la intersección de las nulclinas, utilizado previamente en el análisis del módulo anterior, se identifican tres puntos fijos: dos estables y uno inestable, como se muestra en la Figura 7.7B. El punto fijo estable asociado con una baja concentración de ZEB corresponde al estado híbrido E/M (en el sistema completo, el estado epitelial se caracteriza por una concentración aún menor de ZEB). Por otro lado, el punto fijo estable en el que ZEB alcanza su mayor concentración representa el estado mesenquimal M. En la Figura 7.7C se presenta una solución del sistema, donde se observa que la concentración de ZEB evoluciona en el tiempo hasta alcanzar el estado mesenquimal. El diagrama de bifurcación que se obtuvo al variar el parámetro de entrada SNAIL, mostrado en la Figura 7.7D, confirma la biestabilidad del módulo, que da lugar a los estados E/M y M.

Mediante el estudio de los módulos miR-34/SNAIL y miR-200/ZEB se confirmó el rol de cada bloque de la red de regulación de la EMT. El circuito miR-34/SNAIL es responsable de la primera transición entre los fenotipos epitelial e híbrido ($E \to E/M$), mientras que el circuito miR-200/ZEB es responsable de la segunda transición ($E/M \to M$).

7.5.2. Modelo de pluripotencia

Se realizó el análisis de bifurcación del modelo que considera OCT4 y SOX2, esquematizado en la Figura 7.4A. El análisis se efectuó respecto a su constante de autoactivación $\beta_{N\to mn}$, ya que para este circuito no se consideró una señal externa que se pudiera variar. Se encuentra que, al igual que el modelo de Glauche *et al.*, el sistema es biestable, como se puede observar en la Figura 7.8, donde, para un rango de valores del parámetro $\beta_{N\to mn}$, hay dos puntos fijos estables (NANOG alto y NANOG bajo) y uno inestable. Este resultado se obtuvo para los valores de los parámetros reportados en la Tabla D.3 del Apéndice D.

Para el modelo adaptado, en el cual se realizó la simplificación mostrada en la Figura 7.4B, se llevó a cabo el mismo análisis. El diagrama de bifurcación de NANOG en función de su parámetro de autoactivación resultó idéntico al obtenido para el modelo sin simplificaciones, esencialmente reproduciendo la Figura 7.8.

El sistema mantiene el comportamiento biestable, con dos estados estables: uno asociado a altos niveles de NANOG y otro a niveles bajos. Al comparar el diagrama de bifurcación

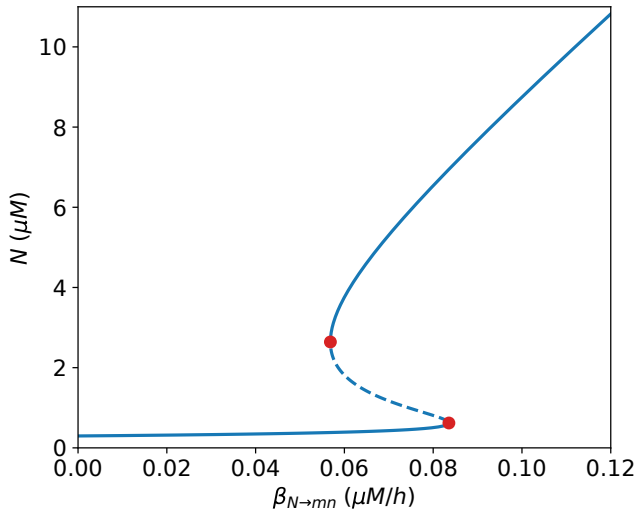

Figura 7.8.: Diagrama de bifurcación de NANOG (N) para los modelos de pluripotencia ilustrados en las Figuras 7.4A y B, variando el parámetro de autorregulación de NANOG ($\beta_{N\to mn}$). Los puntos rojos representan puntos de bifurcación, las líneas sólidas azules representan puntos fijos estables y la línea discontinua azul corresponde a puntos fijos inestables.

con el del modelo previo, que incluye tanto SOX2 como OCT4, se observa que ambos son equivalentes, lo que valida la simplificación aplicada. Este modelo se empleará en el siguiente capítulo para estudiar el acoplamiento de las redes de regulación génica de EMT y pluripotencia.

Integrando numéricamente se obtuvieron algunas soluciones del sistema de ecuaciones diferenciales modelo simplificado. En la Figura 7.9 se muestra la dinámica temporal de la concentración de la proteína SOX2, su ARNm, NANOG y su ARNm, así como del dímero, considerando dos condiciones iniciales distintas, detalladas en la Tabla D.4 del Apéndice D. Para un valor de autoactivación de NANOG que presenta biestabilidad, se observa que tanto SOX2, su ARNm y el dímero convergen al mismo nivel de concentración en el estado estable en ambos casos.

Sin embargo, en el caso de NANOG y su ARNm, las trayectorias dependen de las condiciones iniciales. Bajo las condiciones iniciales representadas en color azul (CI_1), NANOG alcanza un estado estable de baja concentración, asociado a un estado diferenciado. En cambio, bajo las condiciones iniciales representadas en color amarillo (CI_2), el estado estable se caracteriza por concentraciones elevadas de NANOG, correspondientes al estado pluripotente.

Este comportamiento resalta el papel central de la autoactivación de NANOG, ya que permite que, incluso con valores idénticos de SOX2, su ARNm y el dímero, se puedan alcanzar dos estados estables distintos en función del grado de autoactivación de NANOG. Además, es importante señalar que, para otros valores de autoactivación, estos dos estados no coexisten y las transiciones entre ellos no son posibles.

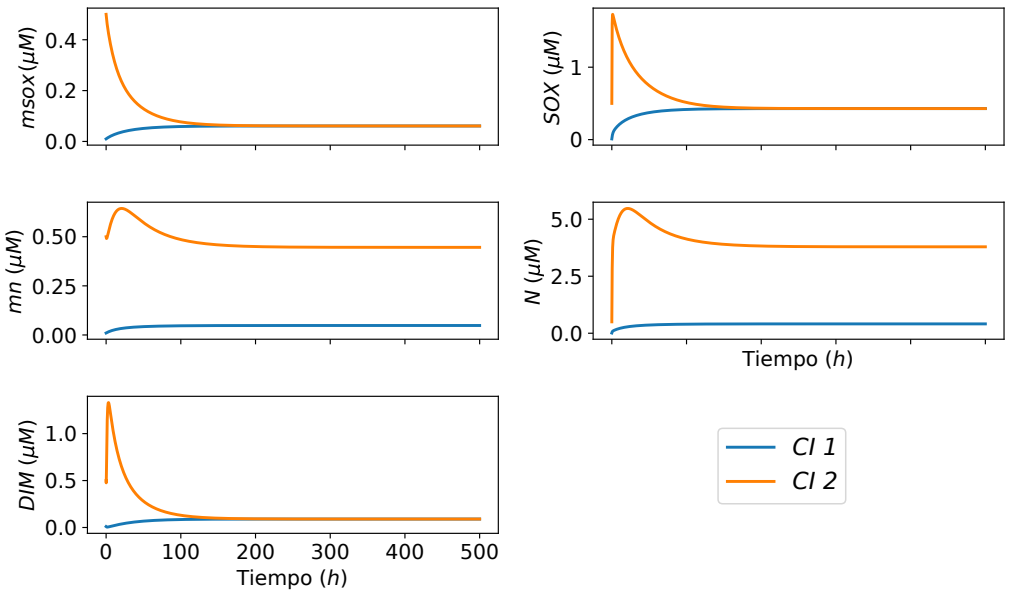

Figura 7.9.: Evolución temporal de la concentración de los elementos de la red para dos condiciones iniciales (CI) diferentes.

7.6. Conclusiones

En este capítulo se desarrolló un modelo de la red de la EMT en base al modelo CBS propuesto por Tian y colaboradores [203]. Se analizó el comportamiento del sistema, principalmente de los fenotipos estables, el epitelial E, el híbrido E/M y el mesenquimal M. A continuación, se analizaron los dos módulos miR-34/SNAIL y miR-200/ZEB por separado, y se encontró que cada módulo es biestable, siendo el primero responsable de la primera transición ($E \rightarrow E/M$) y el segundo de la segunda ($E/M \rightarrow M$).

También se desarrolló un modelo de la red regulatoria génica de pluripotencia que incorpora tanto la transcripción como la traducción, reproduciendo las predicciones de Glauche *et al.* a partir de un modelo reducido [227]. Este modelo considera los factores de transcripción OCT4, SOX2 y NANOG, reconocidos como el circuito fundamental que controla la pluripotencia. El comportamiento de este circuito se estudió mediante ecuaciones diferenciales deterministas y técnicas de análisis de sistemas dinámicos, encontrándose que el sistema presenta biestabilidad.

8 Acoplamiento de los módulos de pluripotencia y transición epitelio-mesénquima

8.1. Introducción

Diversos estudios experimentales han demostrado que las decisiones celulares relacionadas con la transición fenotípica (EMT/MET) y la adquisición de propiedades de pluripotencia están interconectadas a través de redes de regulación génica. No obstante, los mecanismos que gobiernan esta interconexión siguen siendo poco comprendidos. Un enfoque posible para desentrañar la interacción entre la transición epitelio-mesenquimal y la pluripotencia es llevar a cabo un análisis de los circuitos génicos que regulan estos procesos, así como de su acoplamiento. En este capítulo se integran los circuitos centrales de la EMT y la pluripotencia.

8.1.1. miR-200 y el módulo de pluripotencia LIN28/let-7

De acuerdo con la revisión bibliográfica realizada, hasta la fecha, solo un grupo de investigación ha realizado esfuerzos para estudiar la relación entre la EMT y la pluripotencia mediante modelos matemáticos de redes de regulación génica [232-234].

En 2014, Jolly y colaboradores exploraron el circuito LIN28/let-7, asociado a la pluripotencia, modelándolo como un circuito de inhibición mutua. Partiendo del marco teórico propuesto previamente por el mismo grupo [204], incorporaron no solo la inhibición mutua entre LIN28 y let-7, sino también la autoactivación de ambos componentes, como se ilustra en el esquema de la red mostrado en la Figura 8.1A. Además, se incluyó el efecto de miR-200 como señal externa, considerando que miR-200 reprime a LIN28. Sin embargo, el mecanismo de retroalimentación indirecta de LIN28 sobre miR-200 a través de OCT4 no fue considerado.

Este enfoque dio lugar a un modelo con tres posibles estados estables: (i) LIN28 alto y let-7 bajo (1,0), denominado estado *up* (U); (ii) LIN28 bajo y let-7 alto (0,1), conocido como estado *down* (D); y (iii) ambos con niveles intermedios (1/2, 1/2), al que se refieren como *down/up* (D/U) [232].

El modelo de EMT (TCS) previamente desarrollado por el mismo grupo, en el cual miR-200 tiene un rol clave, también predice tres estados estables: epitelial (E), mesenquimal (M) e híbrido (E/M). Los fenotipos epitelial, mesenquimal e híbrido E/M se asociaban a niveles elevados, bajos e intermedios de miR-200. Al analizar el efecto de miR-200 sobre el circuito LIN28/let-7, el diagrama de bifurcación de la Figura 8.1A muestra que, a niveles elevados de miR-200, correspondientes al fenotipo epitelial, solo está presente el estado D. A niveles bajos de miR-200, asociados al estado mesenquimal (M), solo existe el estado U. A niveles intermedios de miR-200, característicos del fenotipo híbrido, pueden coexistir los estados D/U con los estados D y U.

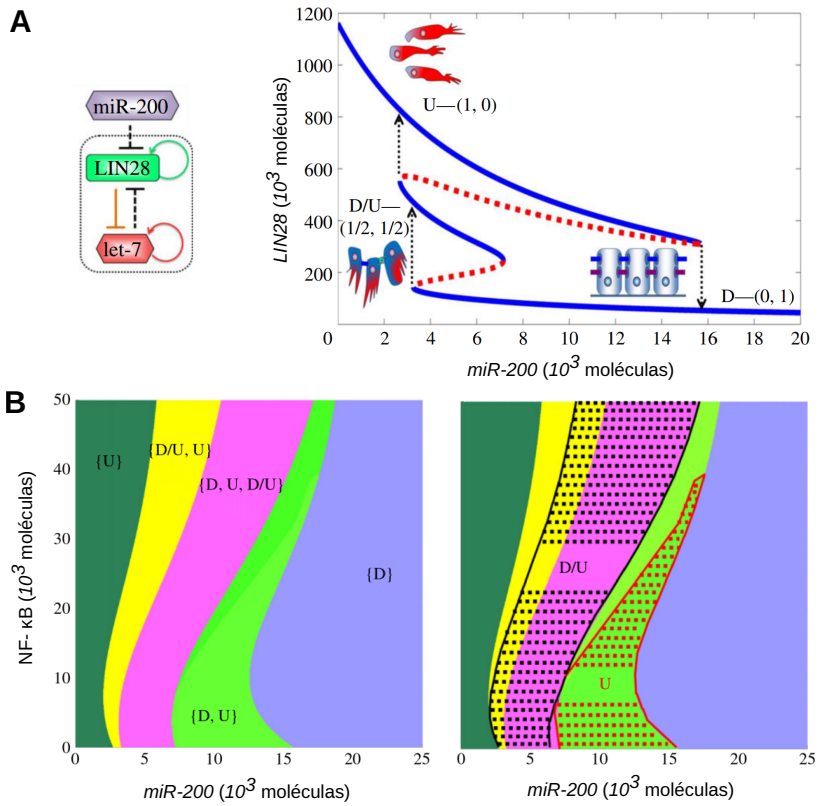

Figura 8.1.: **(A)** Arquitectura del circuito de pluripotencia (LIN28/Let-7) bajo la influencia de miR-200 y diagrama de bifurcación de LIN28 variando la señal de entrada miR-200. **(B)** A la izquierda, diagrama de espacio de fases del circuito LIN28/let-7 en respuesta a las señales miR-200 y NF-kB. Cada fase corresponde a una combinación diferente de fenotipos coexistentes. A la derecha, mapeo de la región con alta probabilidad de pluripotencia, definida por un rango intermedio de niveles de OCT4. Las células en el estado D/U en toda la fase {D, U, D/U} y en parte de la fase {D/U, U} (región de puntos negros) tienen una alta probabilidad de pluripotencia. Lo mismo para las células en el estado U en una parte de la región {D, U} (región de puntos rojos). Adaptación de [232].

Luego incorporan al modelo NF-κB, el cual activa tanto a LIN28 como a let-7. Para investigar aún más la dependencia de la probabilidad de stemness tanto de miR-200 como de NF-κB, construyen el diagrama de espacio de fases mostrado en la Figura 8.1B izquierda. Cada fase corresponde a la coexistencia de uno o más fenotipos. Los estados posibles son: (i) fases con solo un fenotipo {D} y {U}, (ii) fases en las que coexisten dos fenotipos {D, U} y {D/U, U}, y (iii) una fase en la que pueden coexistir los tres fenotipos {D, D/U, U}.

Observaciones experimentales asocian la pluripotencia con niveles intermedios de OCT4, niveles muy altos o muy bajos de OCT4 conducen a diferentes trayectorias de diferenciación [181-183]. En base a estas observaciones, estudiaron la dependencia de los niveles del marcador de pluripotencia OCT4 y su activador LIN28. Los niveles de OCT4 se calcularon como una función de Hill de los niveles de LIN28. Eligieron un rango de niveles de OCT4 que corresponde a una alta probabilidad de pluripotencia y mapearon las regiones de probabilidad de stemness en el diagrama de fases. Así, encontraron que, en el espacio de fases, la probabilidad de mantener la pluripotencia es alta en toda la región D, D/U, U, en parte de la región D/U, U y en parte de la región D, U como se puede observar en la Figura 8.1B.

8.1.2. OVOL y los módulos miR-200/ZEB y LIN28/let-7

En 2015, el mismo grupo de investigación integró los dos módulos de la EMT (miR-200/ZEB) y la pluripotencia (LIN28/let-7) en un solo modelo. En el modelo, la inhibición de LIN28 por miR-200 y de ZEB por let-7 determina qué fenotipo adquiere pluripotencia y dónde se ubica la ventana de pluripotencia en el "eje EMT" [233]. Modificando el valor de los parámetros que conectan ambos módulos (α_1 y α_2) encontraron que una fuerte inhibición de LIN28 desplaza esta ventana hacia el fenotipo mesenquimal (M), mientras que una fuerte inhibición de ZEB la traslada hacia el fenotipo epitelial (E).

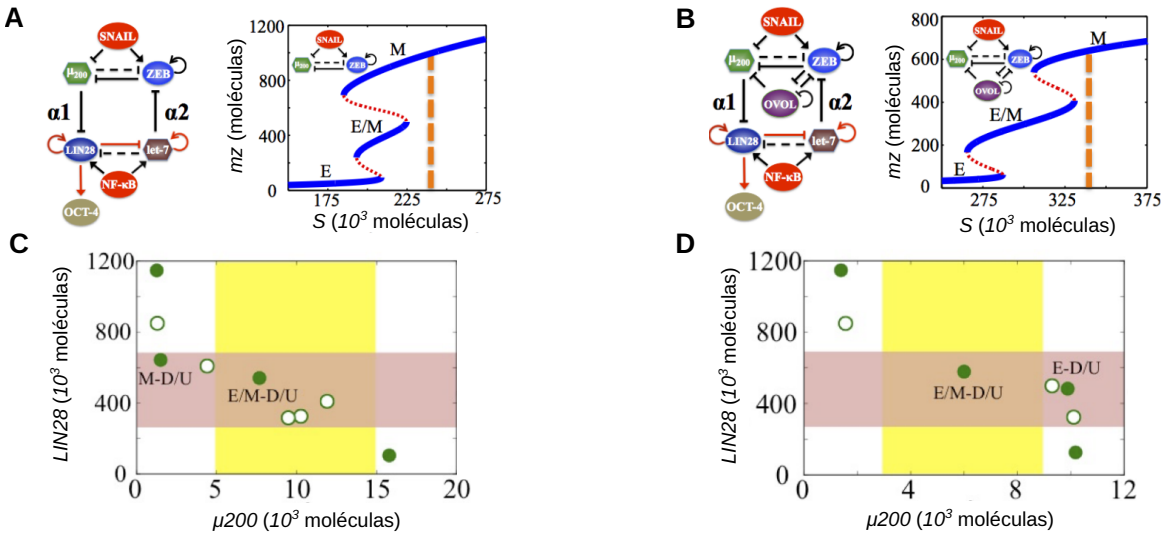

Figura 8.2.: **(A)** Arquitectura del circuito de EMT (miR-200/ZEB/SNAIL) acoplado al circuito de pluripotencia (LIN28/let-7, el factor externo NF-$\kappa\beta$ y el marcador OCT-4) y diagrama de bifurcación del ARNm de zeb (mz) en función de los niveles de la proteína SNAIL (S). **(B)** Arquitectura del circuito de EMT acoplado al circuito de pluripotencia adicionando el factor de transcripción OVOL y diagrama de bifurcación del ARNm de zeb (mz) en función de los niveles de la proteína SNAIL (S). **(C)** Mapa fenotípico del circuito acoplado correspondiente al circuito esquematizado en A. En este caso el fenotipo mesenquimal (M) presenta pluripotencia, además del híbrido E/M. **(D)** Mapa fenotípico del circuito acoplado correspondiente al circuito esquematizado en B. En este caso el fenotipo epitelial (E) adquiere pluripotencia, además del híbrido E/M. El área sombreada en rojo indica la ventana de pluripotencia en base en los niveles de OCT4, mientras que el área sombreada en amarillo representa el rango de niveles de miR-200 que permite la existencia del fenotipo híbrido E/M. Los puntos sólidos verdes representan estados estables mientras que los puntos verdes vacíos son puntos fijos inestables. Adaptación de [233].

Por otro lado, encontraron que OVOL, un factor de transcripción involucrado en la embriogénesis, amplía el rango de parámetros que permiten la existencia del fenotipo híbrido E/M. Sin OVOL, el fenotipo E/M siempre coexiste con el fenotipo M, lo que da lugar a las fases E, E/M, M y E/M, M como se muestra en la Figura 8.2A. Sin embargo, cuando OVOL está presente, el fenotipo E/M puede existir de manera independiente o coexistir con el fenotipo epitelial, lo que resulta en las fases E/M y E, E/M, como se puede apreciar en la Figura 8.2B. Según el modelo, OVOL modula la ubicación de la ventana de pluripotencia, impide que el fenotipo mesenquimal adquiera pluripotencia y favorece que el fenotipo híbrido (E/M) la gane, como se ve en las Figuras 8.2C y D.

Además, el modelo podría explicar resultados experimentales contradictorios sobre la relación entre la EMT y la pluripotencia, como aquellos que indican que una EMT completa lleva a una mayor pluripotencia [188, 235], que la MET está asociada con la pluripotencia [236, 237] y que el fenotipo híbrido presenta la máxima pluripotencia [238, 239], proponiendo que la ventana de pluripotencia es flexible y puede ajustarse a lo largo del eje EMT.

8.2. Metodología

Los modelos de redes de regulación génica que acoplan los procesos de EMT y pluripotencia han abordado las redes de regulación miR-200/LIN28/let-7 y miR-200/ZEB/SNAIL/LIN28/let-7 [232, 233], analizando su impacto en OCT4 para caracterizar el estado de pluripotencia. En estos trabajos, se ha asumido que la pluripotencia se caracteriza por niveles intermedios de OCT4, como se describió previamente. Sin embargo, estos modelos solo contemplan algunos elementos del circuito miR-200/ZEB/miR-34/SNAIL, ya que según los autores este es el módulo responsable de la triestabilidad de la red de la EMT. Además, si bien OCT4 se utiliza para definir la ventana de pluripotencia, su rol dentro de la red no se aborda, limitándose a su consideración como un marcador de pluripotencia. Por otro lado, aunque LIN28 y let-7 tienen un papel relevante en la pluripotencia, no se estudia la red central de pluripotencia conformada por OCT4, SOX2 y NANOG.

En esta sección se plantea el acoplamiento de las redes regulatorias centrales miR-200/ZEB/miR-34/SNAIL (EMT) y OCT4/SOX2/NANOG (pluripotencia) mediante un sistema de ecuaciones diferenciales ordinarias, empleando cinética de Hill para describir la regulación de los factores involucrados. Para ello, se parte del modelo de EMT basado en el modelo CBS [203] desarrollado en la Sección 7.4.2, seleccionado por su menor complejidad, número de parámetros y suposiciones sin evidencia experimental en comparación con el modelo TCS, y del modelo de pluripotencia desarrollado en la Sección 7.4.3 a partir de la arquitectura propuesta por Glauche y colegas [227], que considera explícitamente los mensajeros y proteínas como variables independientes.

Para acoplar ambos modelos, fue necesario determinar las conexiones entre el circuito de EMT (miR-34, miR-200, SNAIL y ZEB) y el circuito de pluripotencia (OCT4/SOX2/NANOG). Para responder a esa pregunta, se realizó una búsqueda bibliográfica exhaustiva y en bases de datos sobre las posibles conexiones entre los elementos de las redes de regulación génica. Se encontró evidencia experimental de la regulación de SOX2 por parte de miR-200 [212, 240-242], así como predicciones en las bases de datos *miRDB* y *TargetScanHuman* que identifican a SOX2 como un objetivo conservado de miR-200 [243, 244]. Por consiguiente, se introdujo miR-200 como una señal que reprime la traducción de SOX2. Ambas fuentes identifican dos sitios de unión para miR-200, lo que justifica la elección de un coeficiente de Hill igual a 2 para esta función de regulación. Esta regulación de miR-200 se implementó a nivel de la proteína SOX2.

Por otro lado, se encontró evidencia de la regulación directa de SOX2 y OCT4 sobre la expresión de miR-200, en este caso como activadores [242, 245]. Con base en esta evidencia experimental y predicciones bioinformáticas, se consideran dos interacciones entre ambos módulos: por un lado, SOX2 (y OCT4) activan la transcripción de miR-200, y por otro, miR-200 reprime la traducción de SOX2. La arquitectura del circuito regulatorio integrado se ilustra en la Figura 8.3A, donde se muestran únicamente las proteínas y los miARNs, omitiendo los transcriptos (ARNm) y los procesos de transcripción, traducción y degradación

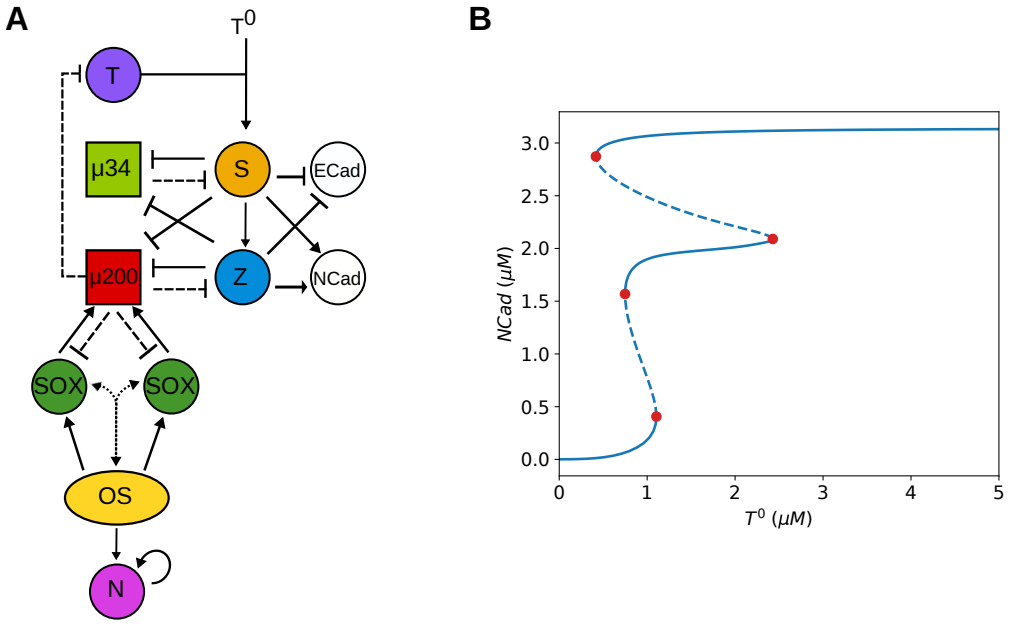

Figura 8.3.: (**A**) Esquema de los modelos de las redes de regulación génica de la transición epitelio-mesénquima (EMT) y pluripotencia acoplados. Las líneas discontinuas indican regulación por miARNs, mientras que las líneas sólidas representa la regulación de la transcripción por factores de transcripción. Las líneas de puntos representan la dimerización y desdimerización. (**B**) Análisis de bifurcación del modelo completo (EMT+pluripotencia) de la N-cadherina ($NCad$), respecto al parámetro de entrada, la concentración exógena de TGF-β (T^0). Los puntos rojos representan puntos de bifurcación, las líneas sólidas azules representan estados estables y las líneas discontinuas azules son estados inestables.

para simplificar el esquema. A partir de los sistemas de ecuaciones descritos previamente para cada circuito (ecs. 7.5 y 7.7), y considerando la conexión entre ambos, se definió el modelo mediante el siguiente sistema de ecuaciones:

$$
\begin{aligned}
\frac{d\left[\mu 200\right]}{dt} &= k^0_{\mu 200} + \beta_{S,Z\to\mu 200}\frac{1}{1+\left(\frac{[S]}{K_{S\to\mu 200}}\right)^{n_{S\to\mu 200}}+\left(\frac{[Z]}{K_{Z\to\mu 200}}\right)^{n_{Z\to\mu 200}}} \\
&+\beta_{SOX\to\mu 200}\frac{\left(\frac{[SOX]}{K_{SOX\to\mu 200}}\right)^{n_{SOX\to\mu 200}}}{1+\left(\frac{[SOX]}{K_{SOX\to\mu 200}}\right)^{n_{SOX\to\mu 200}}} - d_{\mu 200}\left[\mu 200\right], \\
\frac{d\left[mz\right]}{dt} &= k^0_{mz} + \beta_{mz}\frac{\left(\frac{[S]}{K_{S\to mz}}\right)^{n_{S\to mz}}}{1+\left(\frac{[S]}{K_{S\to ms}}\right)^{n_{S\to mz}}} - d_{mz}\left[mz\right], \\
\frac{d\left[Z\right]}{dt} &= \beta_Z\left[mz\right]\frac{1}{1+\left(\frac{[\mu 200]}{K_{\mu 200\to Z}}\right)^{n_{\mu 200\to Z}}} - d_Z\left[Z\right], \\
\frac{d\left[\mu 34\right]}{dt} &= k^0_{\mu 34} + \beta_{S,Z\to\mu 34}\frac{1}{1+\left(\frac{[S]}{K_{S\to\mu 34}}\right)^{n_{S\to\mu 34}}+\left(\frac{[Z]}{K_{Z\to\mu 34}}\right)^{n_{Z\to\mu 34}}} - d_{\mu 34}\left[\mu 34\right], \\
\frac{d\left[ms\right]}{dt} &= k^0_{ms} + \beta_{ms}\frac{\left(\frac{[T]+T^0}{K_{T\to ms}}\right)^{n_{T\to ms}}}{1+\left(\frac{[T]+T^0}{K_{T\to ms}}\right)^{n_{T\to ms}}} - d_{ms}\left[ms\right],
\end{aligned}
\tag{8.1}
$$

$$\frac{d\,[S]}{dt} = \beta_S\,[ms]\,\frac{1}{1+\left(\frac{[\mu 34]}{K_{\mu 34\to S}}\right)^{n_{\mu 34\to S}}} - d_S\,[S]\,,$$

$$\frac{d\,[mt]}{dt} = k^0_{mt} - d_{mt}\,[mt]\,,$$

$$\frac{d\,[T]}{dt} = \beta_T\,[mt]\,\frac{1}{1+\left(\frac{[\mu 200]}{K_{\mu 200\to T}}\right)^{n_{\mu 200\to T}}} - d_T\,[T]\,,$$

$$\frac{d\,[ECad]}{dt} = \beta_{S\to E}\frac{1}{1+\left(\frac{[S]}{K_{S\to E}}\right)^{n_{S\to E}}} + \beta_{Z\to E}\frac{1}{1+\left(\frac{[Z]}{K_{Z\to E}}\right)^{n_{Z\to E}}} - d_E\,[ECad]\,,$$

$$\frac{d\,[NCad]}{dt} = \beta_{S\to N}\frac{\left(\frac{[S]}{K_{S\to N}}\right)^{n_{S\to N}}}{1+\left(\frac{[S]}{K_{S\to N}}\right)^{n_{S\to N}}} + \beta_{Z\to N}\frac{\left(\frac{[Z]}{K_{Z\to N}}\right)^{n_{Z\to N}}}{1+\left(\frac{[Z]}{K_{Z\to N}}\right)^{n_{Z\to N}}} - d_N\,[NCad]\,,$$

$$\frac{d\,[msox]}{dt} = k^0_{msox} + \beta_{msox}\frac{\left(\frac{[DIM]}{K_{DIM\to msox}}\right)^{n_{DIM\to msox}}}{1+\left(\frac{[DIM]}{K_{OS\to msox}}\right)^{n_{DIM\to msox}}} - d_{msox}\,[msox]\,,$$

$$\frac{d\,[SOX]}{dt} = \beta_{SOX}\,[msox]\,\frac{1}{1+\left(\frac{[\mu 200]}{K_{\mu 200\to SOX}}\right)^{n_{\mu 200\to SOX}}} - 2k_{on}\,[SOX]\,[SOX] + 2k_{off}\,[DIM] - d_{SOX}\,[SOX]\,,$$

$$\frac{d\,[DIM]}{dt} = k_{on}\,[SOX]\,[SOX] - k_{off}\,[DIM] - d_{DIM}\,[DIM]\,,$$

$$\frac{d\,[mn]}{dt} = k^0_{mn} + \beta_{DIM\to mn}\frac{\left(\frac{[DIM]}{K_{DIM\to mn}}\right)^{n_{DIM\to mn}}}{1+\left(\frac{[DIM]}{K_{DIM\to mn}}\right)^{n_{DIM\to mn}}} + \beta_{N\to mn}\frac{\left(\frac{[N]}{K_{N\to mn}}\right)^{n_{N\to mn}}}{1+\left(\frac{[N]}{K_{N\to mn}}\right)^{n_{N\to mn}}} - d_{mn}\,[mn]\,,$$

$$\frac{d\,[N]}{dt} = \beta_N\,[mn] - d_N\,[N]\,.$$

Se mantuvieron los valores de todos los parámetros previamente definidos, con excepción de $K_{DIM\to mn}$ y $K_{DIM\to msox}$, los cuales fueron reducidos. Esto se debe a que, al acoplar los dos circuitos génicos, en particular al incluir la inhibición de SOX2 por miR-200, se obtienen concentraciones más bajas de SOX2 y, en consecuencia, del dímero. La adecuación de los valores de las constantes de equilibrio que dependen de la concentración del dímero garantiza que éste regule la transcripción de SOX2 y NANOG.

Además, se introdujeron nuevos parámetros asociados a las interacciones entre los módulos de EMT y pluripotencia. En primer lugar, para modelar la activación de miR-200 por parte de SOX2, se definieron los parámetros $\beta_{SOX\to\mu 200}$, $K_{SOX\to\mu 200}$ y $n_{SOX\to\mu 200}$. Por otro lado, la represión de la traducción de SOX2 mediada por miR-200 implicó la introducción de los parámetros $K_{\mu 200\to SOX}$ y $n_{\mu 200\to SOX}$. Todos los parámetros, junto con sus valores específicos, que mantuvieron sus valores respecto a los modelos aislados de EMT y pluripotencia están detallados en las Tablas D.1 y D.3 del Apéndice D, mientras que los parámetros que se incorporaron al acoplar los circuitos o cuyos valores se modificaron se encuentran en la Tabla D.5 del Apéndice D.

8.3. Resultados y discusiones

Utilizando el conjunto de parámetros mencionado se llevó a cabo el análisis de bifurcación de la N-cadherina respecto al factor externo, el TGF-β exógeno (T^0), como se muestra en la Figura 8.3B. El diagrama de bifurcación del sistema acoplado presenta diferencias moderadas respecto al del sistema EMT aislado (ver Figura 7.5B). La principal diferencia es que el rango de valores del TGF-β exógeno (T^0) en el cual existe el fenotipo mesenquimal es menor en el sistema acoplado respecto al sistema aislado. En el sistema acoplado, dicho estado solo es

viable para $T^0 > 0{,}41\mu M$, mientras que en el sistema aislado es accesible para cualquier valor de $T^0 > 0\mu M$. Además el modelo acoplado predice que es posible la transición del fenotipo mesenquimal al epitelial directamente ($M \rightarrow E$).

En consecuencia, la histéresis que caracteriza al sistema aislado, en el cual se observa la imposibilidad de una transición del estado mesenquimal mediante una MET, difiere en el sistema acoplado. Según el modelo de la red aislada, la única transición MET posible es desde el estado híbrido E/M al epitelial ($E/M \rightarrow E$); si se alcanza el fenotipo mesenquimal, la célula queda atrapada en ese estado. Sin embargo, no es posible la transición del fenotipo mesenquimal al híbrido ($M \nrightarrow E/M$) como se muestra en Figura 7.5B. Motivados por la evidencia que respalda la existencia de la MET, Tian y colegas [203] investigaron escenarios de parámetros en los que la reversibilidad de la EMT es posible. Uno de los casos analizados por los autores consistió en la inclusión de un activador para miR-200, situación que tiene efectos similares a los obtenidos al acoplar los modelos de EMT y pluripotencia, dado que SOX2 activa miR-200.

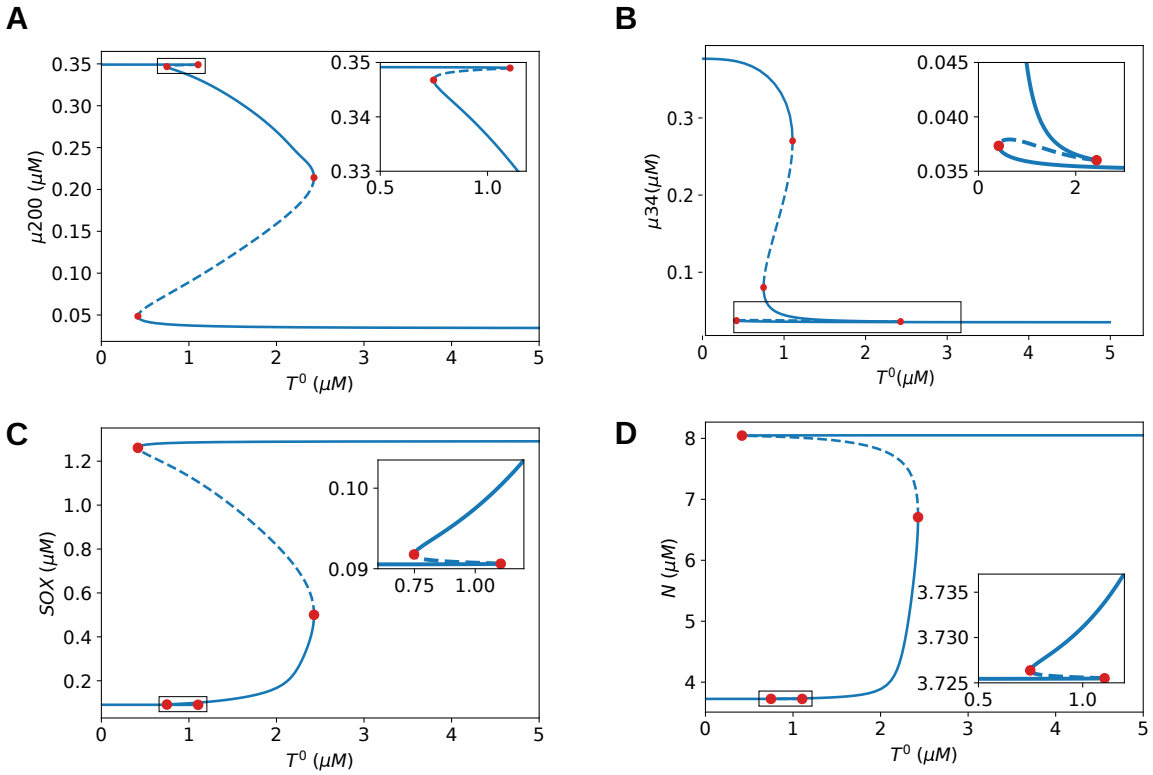

Figura 8.4.: Análisis de bifurcación en función de la concentración exógena de TGF-β (T^0) para las concentraciones de miR-200 (**A**), miR-34 (**B**), SOX2 (**C**) y NANOG (**D**). Los puntos rojos indican puntos de bifurcación, mientras que las líneas sólidas representan estados estacionarios estables y las líneas discontinuas indican estados estacionarios inestables. Se incluyen ampliaciones en las regiones delimitadas por rectángulos para resaltar detalles de los comportamientos locales en torno a las bifurcaciones.

También se analizó el comportamiento de otros elementos del modelo al variar la señal de entrada, el TGF-β exógeno T^0. En la Figura 8.4 se presentan los diagramas de bifurcación de los miARNs miR-34 y miR-200 (Figuras 8.4A y B), así como de las proteínas del modelo de pluripotencia SOX2 y NANOG (Figuras 8.4C y D) en función de la señal externa. En todos los casos se identifican tres estados estables.

En las Figuras 8.4A y B se observa que, a medida que aumenta la señal externa, el fenotipo epitelial (miR-34 y miR-200 elevados) predomina en niveles bajos de TGF-β exógeno, mientras que el fenotipo mesenquimal (miR-34 y miR-200 bajos) se manifiesta en niveles altos de TGF-β exógeno, en concordancia con lo observado en la Figura 8.3B. Se identifican los siguientes estados en función del incremento de TGF-β exógeno: $\{E\}$, $\{E, M\}$, $\{E, E/M, M\}$, $\{E/M, M\}$ y $\{M\}$, tal como se ilustra en la Figura 8.3B. Los fenotipos epitelial y mesenquimal se caracterizan por mantener concentraciones aproximadamente constantes de miR-200 en todo el rango de T^0 estudiado (Figura 8.4A). En contraste, en el estado híbrido E/M, la concentración de miR-200 disminuye conforme aumenta T^0. En la región donde se observa triestabilidad, las concentraciones de miR-200 en los estados epitelial e híbrido E/M son similares. En el caso de miR-34 (Figura 8.4B), la concentración en el estado epitelial presenta variabilidad, mientras que en el estado mesenquimal es aproximadamente constante. Por otro lado, en el estado híbrido E/M, la concentración de miR-34 disminuye con el aumento de T^0.

Tanto para SOX2 como para NANOG, las concentraciones se mantienen aproximadamente constantes en los fenotipos epitelial y mesenquimal, mientras que en el estado híbrido E/M, la concentración de estos factores de transcripción aumenta monótonamente con el incremento de T^0. En la región de coexistencia de los tres estados estables, las diferencias en las concentraciones de SOX2 y NANOG entre el estado epitelial y el híbrido E/M son muy pequeñas, perceptibles únicamente al ampliar considerablemente las escalas (recuadros en las Figuras 8.4C y D). Aunque se identifican tres estados estables para SOX2 y NANOG, los cambios de las concentraciones en la primera transición ($E \rightarrow E/M$) son despreciables en comparación con la segunda transición ($E/M \rightarrow M$). Esto sugiere una variación más gradual en la transición del estado epitelial al híbrido E/M, sin que se produzcan transiciones fenotípicas abruptas.

Al comparar el diagrama de bifurcación de NANOG en función del TGF-β exógeno con la Figura 8.3B, se observa una correspondencia entre el estado epitelial y niveles bajos de NANOG, así como entre el estado mesenquimal y niveles elevados de este factor. En el caso del estado híbrido E/M, NANOG puede adoptar un rango amplio de valores, desde aquellos característicos del estado epitelial hasta valores moderadamente elevados.

Integrando numéricamente las ecs. 8.1 se obtuvieron algunas soluciones del sistema de ecuaciones diferenciales. En la Figura 8.5 se presenta la dinámica de la E-cadherina, la N-cadherina, el dímero y NANOG para las condiciones iniciales detalladas en la Tabla D.6 del Apéndice D, y utilizando distintos valores de la señal externa del TGF-β exógeno. La dinámica de todos los elementos de la red completa se puede observar en la Figura B.2 del Apéndice B. Para analizar las trayectorias en cada condición inicial se observan las curvas del mismo color de los distintos elementos del circuito.

Por ejemplo, para un nivel bajo del TGF-β exógeno ($T^0 = 0{,}2\,\mu M$), se obtiene la curva verde. A partir de la 8.3B se puede predecir que el único fenotipo que existe para ese valor de señal externa es el epitelial. Analizando la evolución temporal, se observa que la concentración de E-cadherina y N-cadherina comienza en valores intermedios, característicos del fenotipo híbrido E/M. A medida que el sistema transiciona hacia el estado epitelial, la concentración de E-cadherina aumenta y la de N-cadherina se disminuye. Inicialmente, la concentración de NANOG incrementa, durante la transición disminuye progresivamente hasta alcanzar su nivel más bajo, asociado con un estado de baja pluripotencia.

Para niveles intermedios del TGF-β exógeno ($T^0 = 1{,}8\,\mu M$) se muestran dos soluciones para distintas condiciones iniciales, correspondientes a la curva azul y violeta. Para $T^0 = 1{,}8\,\mu M$

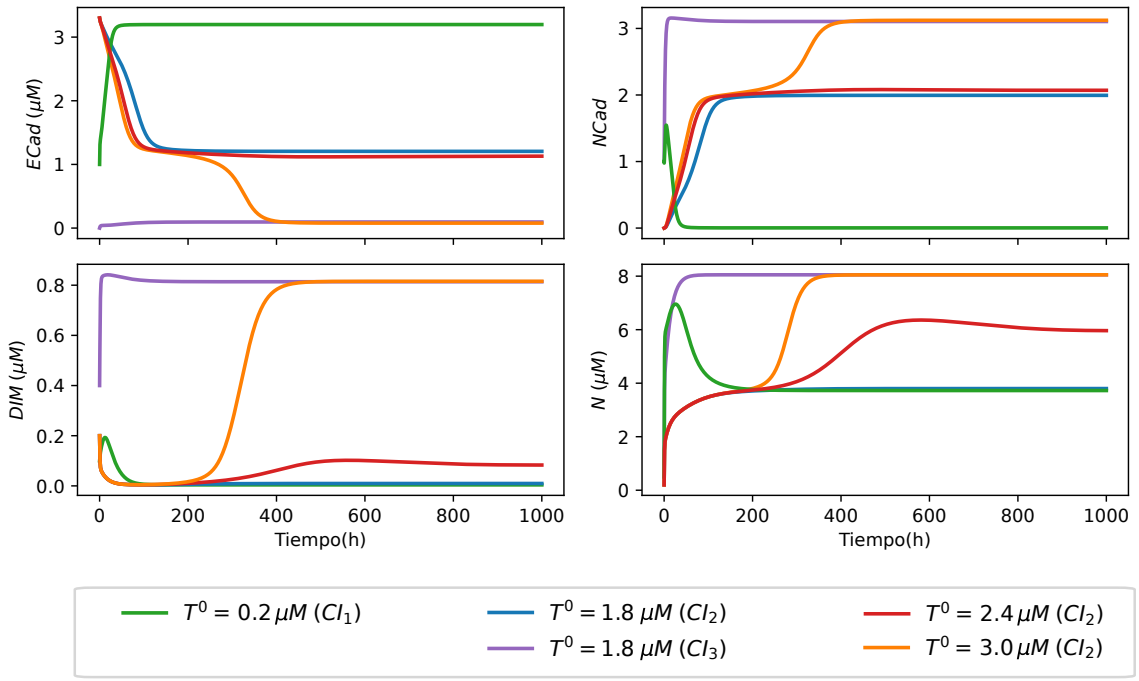

Figura 8.5.: Evolución temporal de la E-cadherina, N-cadherina, el dímero y NANOG para distintas condiciones iniciales y valores TGF-β exógeno T^0. Las unidades de las concentraciones se encuentran en μM.

coexisten el estado híbrido E/M y el estado mesenquimal, como se visualiza en la Figura 8.3B. En el caso de la curva azul, inicialmente las concentraciones de la E-cadherina y la N-cadherina corresponden con el estado epitelial y ocurre una transición hacia el estado híbrido E/M con valores intermedios de la E-cadherina y N-cadherina y niveles bajos de NANOG asociado a un estado no pluripotente. La trayectoria representada por las curvas violetas comienza con un estado bajo de E-cadherina, intermedio de N-cadherina y bajo de NANOG. Rápidamente las concentraciones de la N-cadherina y NANOG aumentan, hasta alcanzar el fenotipo mesenquimal pluripotente.

Para un valor de TGF-β exógeno moderadamente elevado ($T^0 = 2{,}4\,\mu M$) se obtiene la trayectoria representada con color rojo. Inicialmente, el estado es epitelial y luego transiciona al fenotipo híbrido E/M. Se puede notar que NANOG, en este caso, alcanza niveles moderadamente altos.

Al incrementar aún más el valor de TGF-β exógeno ($T^0 = 3\,\mu M$) se obtienen las curvas color naranja. En este caso se observa que el fenotipo comienza epitelial (E-cadherina elevada), transiciona primero al fenotipo híbrido E/M (E-cadherina y N-cadherina intermedias) y luego al mesenquimal (N-cadherina alta). A su vez, cuando ocurre la segunda transición también incrementan los niveles de NANOG, asociados con un estado más pluripotente.

En la sección anterior, al describir el modelo acoplado, se mencionó que la mayoría de los valores de los parámetros se mantuvieron iguales a los utilizados en los modelos aislados, con excepción de $K_{DIM\to mn}$ y $K_{DIM\to msox}$. La necesidad de modificar estos valores se hace evidente al analizar la evolución temporal mostrada en la Figura 8.5. En dicha figura se observa que la concentración del dímero (DIM) nunca alcanza los valores de las constantes de equilibrio en el modelo aislado, los cuales eran $K_{DIM\to mn} = K_{DIM\to msox} = 1\,\mu M$. Si se utilizan esos valores para las constantes de equilibrio en el modelo acoplado, el dímero

prácticamente no regula la transcripción de NANOG y SOX2. Esto motivó la reducción de dichos parámetros, como se detalla en la Tabla D.5 del Apéndice D. Ajustar ligeramente los valores de los parámetros al integrar circuitos que inicialmente fueron modelados de manera independiente es una práctica común. Por ejemplo, Lu y colegas [204] realizaron modificaciones en los valores de los parámetros al acoplar los módulos miR-34/SNAIL y miR-200/ZEB en el modelo TCS. Esta estrategia es razonable, ya que los modelos aislados representan versiones reducidas de sistemas más complejos, y la incorporación de nuevos componentes puede requerir ajustes de los parámetros para capturar adecuadamente la dinámica del sistema integrado.

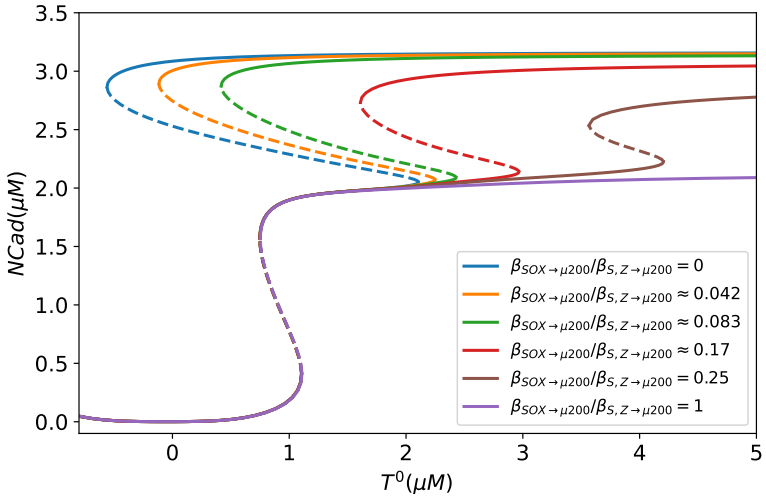

Figura 8.6.: Análisis de bifurcación de los módulos acoplados de la EMT y la pluripotencia, de la N-cadherina ($NCad$), respecto al parámetro de entrada, la concentración exógena de TGF-β (T^0), para distintos valores de la influencia del circuito de pluripotencia sobre el circuito EMT representado mediante el cociente de los parámetros parámetro $\beta_{SOX\to\mu200}$ y $\beta_{S,Z\to\mu200}$. Las líneas sólidas representan puntos fijos estables y las líneas discontinuas representan puntos fijos inestables.

Anteriormente se estudió el comportamiento de algunos elementos del modelo, como la N-cadherina, al variar la señal externa del TGF-β exógeno para un valor fijo del parámetro $\beta_{SOX\to\mu200} = 0{,}001\,\mu M/h$, que es la tasa de producción máxima del miR-200 regulada por SOX2. Este parámetro tiene un papel clave, ya que es el que regula la influencia del circuito de pluripotencia sobre el circuito de la EMT. Mientras más alto sea su valor, mayor será el efecto y para un valor nulo se espera recuperar el comportamiento original de la red de regulación génica de EMT.

Un índice más útil para cuantificar la influencia de la red de pluripotencia sobre la red de EMT es la relación $\beta_{SOX\to\mu200}/\beta_{S,Z\to\mu200}$, ya que refleja el valor máximo de la tasa de producción de miR-200 dependiente de SOX2 respecto al valor máximo de la tasa de producción de miR-200 dependiente SNAIL y ZEB, los tres reguladores de miR-200. Con el fin de explorar el efecto que tiene la red de pluripotencia sobre la red de EMT, se realiza un análisis de bifurcación para distintos niveles de influencia del circuito de pluripotencia sobre el de EMT, variando el valor de $\beta_{SOX\to\mu200}$. En la Figura 8.6 se muestran tales diagramas de bifurcación.

Cuando el módulo de pluripotencia no tiene efecto sobre el módulo de EMT (curva azul) se tiene el diagrama de bifurcación del modelo de la EMT aislado de la Figura 7.5B. A

medida que aumenta el parámetro $\beta_{SOX\to\mu 200}$, la región de parámetros para la que existe el estado híbrido es más amplia y el estado mesenquimal ocurre a partir de valores mayores de T^0. La ampliación en la región de parámetros en el modelo puede ser entendida como un aumento en la región de condiciones fisiológicas en un experimento. Utilizando $\beta_{SOX\to\mu 200} = 0{,}001\,\mu M/h$ (curva verde), se pierde la situación de irreversibilidad para la cual el sistema no puede retornar del estado mesenquimal, ya que no puede transicionar al fenotipo epitelial. Aumentando el parámetro aún más (curvas roja y marrón), el fenotipo mesenquimal realiza la transición inversa mesénquima-epitelio en dos etapas, primero transicionando al estado híbrido E/M y luego al epitelial, además el rango de parámetros para el que ocurre el estado híbrido es mayor. También se observa que con el aumento de $\beta_{SOX\to\mu 200}$ se reduce la región de parámetros de triestabilidad. Finalmente, cuando la influencia del módulo de pluripotencia sobre el módulo de la EMT es aún más elevada se pierde el fenotipo mesenquimal (curva violeta). Estos resultados observados como consecuencia de acoplar el circuito de pluripotencia y de EMT, son similares a los resultados que encuentran Jolly *et al.* al incluir OVOL en su modelo, basado en el modelo TCS.

8.4. Conclusiones

Se acoplaron los modelos de las redes de regulación génica de la pluripotencia y de la EMT desarrollados en el capítulo anterior. Se observó que la influencia del módulo de pluripotencia desempeña un papel crucial sobre el módulo de la EMT. Al incrementar esta influencia, la región de parámetros que permite el estado híbrido E/M se expande, y se observa que la EMT es reversible, a diferencia de lo que ocurría para el módulo aislado. Para valores intermedios de dicha influencia entre los subcircuitos, la MET ocurre en un solo paso, mientras que a valores más elevados, se produce en dos etapas: del fenotipo mesenquimal al híbrido E/M, y del híbrido E/M al epitelial. Este resultado es novedoso, ya que el modelo EMT aislado predecía la imposibilidad de salir del estado mesenquimal una vez alcanzado, aunque sí era posible la transición del estado híbrido E/M al epitelial.

Las predicciones del modelo acoplado podrían explicar observaciones experimentales controvertidas, ya que en algunos trabajos se reporta la observación experimental de la MET en cáncer y en otros no [150]. Se concluye que modificar la influencia del módulo de pluripotencia puede alterar dicha transición, lo cual podría verificarse experimentalmente. Es importante destacar que trabajos previos, como el de Jolly *et al.* [233], han encontrado resultados similares al integrar un circuito EMT que considera solo miR-200/ZEB/SNAIL y elementos relacionados con la pluripotencia, como LIN28/let-7. Sin embargo, el circuito de pluripotencia estudiado en este trabajo se considera la red central reguladora de la EMT.

En términos generales, al acoplar los circuitos de EMT y pluripotencia, el modelo predice que el fenotipo epitelial corresponde a un estado diferenciado y no pluripotente. El fenotipo mesenquimal muestra los niveles más elevados de NANOG, un marcador característico de pluripotencia. Tanto el estado epitelial como el mesenquimal presentan niveles constantes de NANOG, bajos y altos respectivamente, mientras que el fenotipo híbrido E/M puede expresar un amplio rango de valores de NANOG.

Tradicionalmente, se ha reportado una fuerte correlación entre la EMT y la pluripotencia en cáncer, observándose que las células mesenquimales poseen mayores características de pluripotencia [188-192]. Más recientemente, se ha demostrado que las células que migran y originan metástasis presentan predominantemente marcadores de EMT temprana, mientras

que solo un pequeño porcentaje expresa marcadores de EMT tardía. En esta línea, trabajos más actuales han planteado la hipótesis de que las células cancerosas con mayor grado de pluripotencia, es decir, las CSCs, se encuentran en el estado híbrido E/M dentro del eje de EMT [150, 246]. Por lo tanto, hay evidencia que relaciona al fenotipo mesenquimal con la máxima pluripotencia y otras observaciones sugieren que el híbrido es el más pluripotente.

Aunque estas observaciones experimentales pueden parecer contradictorias, en un trabajo de revisión reciente, Bornes y colegas reportan una posible explicación [150]. La mayoría de los estudios sobre la regeneración tumoral a partir de células cancerosas pluripotentes o sometidas a EMT se han realizado *in vitro*, lo que supone una limitación para comprender con precisión los procesos que ocurren en un organismo complejo, donde la interacción con el sistema inmune desempeña un papel crucial. En este contexto, se ha propuesto que la ventaja metastásica de las células híbridas E/M se debe a sus propiedades mixtas epiteliales y mesenquimales, razón por la cual pueden migrar en clústers y evadir el sistema inmune. En contraste, las células mesenquimales, al carecer de propiedades de adhesión celular propias del fenotipo epitelial, migran de manera individual, lo que las hace más susceptibles a ser detectadas y eliminadas por el sistema inmune *in vivo*. Esto puede explicar la razón por la cual se ha observado que la mayoría de los tumores metastásicos han sido iniciados por células híbridas E/M.

El modelo matemático de las redes conjuntas de EMT y pluripotencia, desarrollado en este capítulo, predice que el fenotipo mesenquimal es el más pluripotente dentro del eje EMT. A partir de estas predicciones, se plantea la hipótesis de que, aunque las células mesenquimales presentan mayor pluripotencia, las células con mayor potencial metastásico son aquellas en el estado híbrido E/M, que, si bien presentan niveles heterogéneos de NANOG, tienen una mayor probabilidad de evadir el sistema inmune y colonizar sitios distantes del tumor primario. En este sentido, se propone que las células con mayor pluripotencia no necesariamente representan el mayor riesgo de metástasis. En cambio, son las células híbridas E/M, con niveles moderadamente elevados de NANOG, las que podrían poseer una ventaja adaptativa al combinar capacidad migratoria, adhesión entre células y mecanismos efectivos de evasión inmune, lo que podría explicar su papel central en la progresión tumoral y la diseminación metastásica.

Una perspectiva a explorar en el futuro podría ser evitar la reducción del sistema de ecuaciones diferenciales de la red de regulación génica de pluripotencia. En esta tesis, se eliminó el factor de transcripción OCT4 y su mensajero para simplificar el modelo. Se ha reportado que niveles intermedios de OCT4 están asociados con el máximo nivel de pluripotencia, mientras que concentraciones altas o bajas inducen la diferenciación. Incluir OCT4 explícitamente en el modelo permitiría realizar análisis comparables a los de Jolly *et al.* [232, 233] y estudiar el efecto del acoplamiento sobre la ventana de pluripotencia.

Otra vía posible de trabajo a futuro puede incluir el refinamiento del modelo de la regulación por miARNs; en esta tesis se consideró que los miARNs reprimen la traducción de los factores de transcripción mediante cinética de Hill, como proponen Tian *et al*. Para los fines de este trabajo se considera que el modelo de regulación por miARNs de Lu *et al.* es excesivamente complejo en cuanto al número de parámetros que se introducen, teniendo en cuenta que se realizan otras grandes aproximaciones. No obstante, es factible desarrollar modelos que introduzcan menor cantidad de parámetros que diferencien la degradación pasiva y activa mediada por miRNAs. También se puede ajustar el modelo de EMT en base a publicaciones más recientes, en particular modificando los coeficientes de Hill y evaluando la inclusión de

otras regulaciones, como la autorregulación negativa de SNAIL, tal como se ha planteado en otros modelos de EMT [205, 208].

9 Conclusiones generales

Las conclusiones generales de esta tesis se basan en los resultados obtenidos a partir del análisis de datos de scRNA-seq (Parte I) y en el modelado de redes de regulación génica (Parte II), los cuales abordan el estudio de diferentes aspectos de interés en cáncer. A lo largo de los capítulos, se desarrollaron métodos computacionales que fueron utilizados para estudiar la mama humana sana y el cáncer de mama; y modelos matemáticos que aportaron nuevas perspectivas sobre los circuitos génicos de la EMT y la pluripotencia, y la interconexión de estos procesos.

En los Capítulos 3 y 4 de la Parte I, la metodología ORIGINS y su versión extendida, ORIGINS2, se introdujeron como herramientas para el análisis de scRNA-seq, ambos implementados como paquetes de R de código abierto y posteriormente desarrollado también en python. El objetivo de ORIGINS es cuantificar la pluripotencia mediante el uso de la red de interacción proteína-proteína (PPIN) asociada al proceso de diferenciación celular. Esta PPIN se construyó a partir de bases de datos públicas que contienen información de interacción de proteínas (Pathway Commons) y *Gene Ontology* (QuickGO). Se evaluó el rendimiento del algoritmo y se encontró que es comparativamente menos demandante en términos de memoria que otros algoritmos disponibles en la literatura, como LandSCENT y CytoTRACE. Al comparar ORIGINS con estos métodos en cuatro conjuntos de datos humanos, se observó la capacidad para identificar células pluripotentes, ofreciendo un recurso valioso en la investigación de células con potencial de pluripotencia, muy utilizado en estudios de desarrollo o cáncer. Una de las aplicaciones de esta herramienta es la identificación de células madre/progenitoras para determinar el origen al inferir las trayectorias de diferenciación, técnicas ampliamente utilizadas actualmente en el análisis de datos de scRNA-seq y una de las motivaciones para desarrollar el método. Se mostró la aplicación a un conjunto de datos de mama sana; se identificaron las células de origen utilizando el algoritmo y luego se utilizó una herramienta de inferencia de trayectorias (SLINGSHOT) para inferir el proceso de diferenciación del epitelio mamario. El análisis mostró una trayectoria bifurcada, lo que apoya la hipótesis de que las células madre mamarias son bipotentes. Además, este enfoque permitió identificar los genes clave variables a lo largo del proceso de diferenciación.

La herramienta ORIGINS2, por su parte, amplía la funcionalidad de la versión anterior de la técnica al calcular actividades de PPINs en general, asociadas con cualquier proceso biológico o conjunto de genes, lo cual permite la cuantificación de módulos biológicos específicos en poblaciones celulares heterogéneas. En particular, a partir del análisis de datos de scRNA-seq de cáncer de mama triple negativo (TN), ORIGINS2 permitió diferenciar poblaciones celulares con distintos procesos biológicos activos, tales como el ciclo celular y la migración. A diferencia de los métodos que se basan exclusivamente en la expresión génica, ORIGINS2 incorpora información previa de interacciones proteína-proteína. Una perspectiva futura que surge de este trabajo es que ORIGINS2 podría emplearse para mapear datos de scRNA-seq a un espacio de actividades asociadas con procesos biológicos específicos. Una ventaja de esta transformación es que reduciría la dimensionalidad de los datos, y que a diferencia de los métodos usuales de reducción de dimensiones, cada vector representaría una función biológica. Utilizando esta nueva base, se podrían aplicar herramientas estándar de análisis de datos de scRNA-seq como la reducción de dimensionalidad (en caso de que el espacio de

actividades tenga alta dimensionalidad), visualización de datos, estudio de trayectorias de diferenciación, entre otras.

En el Capítulo 5 se llevó a cabo un análisis cuantitativo de características relevantes en cáncer de mama, realizando un estudio separado por subtipo para evaluar las diferencias entre los tres subtipos más prevalentes: ER+, HER2+ y TN. Para ello, se utilizaron muestras de scRNA-seq de donantes de bases de datos públicas y se definieron diversos parámetros, que incluyeron las alteraciones en el número de copias (CNAs), la entropía, la heterogeneidad tumoral y la actividad de PPINs vinculadas a procesos biológicos de interés, como la EMT, el ciclo celular y la identidad celular luminal o basal. A diferencia de estudios previos que generalmente se enfocan en alguna de estas características en particular o realizan análisis cualitativos, esta estrategia permite una evaluación cuantitativa integrada de estos aspectos en el contexto del cáncer. Los resultados obtenidos a nivel de células individuales mostraron patrones de actividad distintos entre los subtipos de cáncer de mama. Por ejemplo, se observó que la actividad del ciclo celular y la entropía aumentan progresivamente en los subtipos ER+, HER2+ y TN, siendo este último el más agresivo y con mayor proporción de células mitóticas. Esto contribuye a la caracterización del cáncer de mama y a futuro se podría extender la aplicación a otros tipos de cáncer.

Los patrones de alteración en CNAs también difirieron entre los subtipos estudiados, con valores significativamente elevados y mayor variabilidad en HER2+ y TN en comparación con ER+, lo cual refuerza la idea de que la heterogeneidad genómica y las alteraciones cromosómicas juegan un papel clave en la progresión y el comportamiento clínico de estos subtipos. Una observación interesante fue la ausencia de un grupo celular con actividad elevada de la EMT (ACT^{EMT}), un proceso biológico que se sabe que ocurre en estos subtipos y que está asociado a la agresividad. Esto podría deberse a la baja proporción o a la ausencia de células que experimentan este proceso en las muestras estudiadas. Otra explicación alternativa puede ser que la sobreexpresión de genes asociados a la EMT como ZEB1, ZEB2 y SNAIL puede inducir cambios fenotípicos sin mostrar un aumento en la actividad de la PPIN asociada a la EMT.

En la Parte II de la tesis, utilizando como punto de partida otros trabajos de la literatura, se introdujo un modelo teórico para estudiar la interacción entre los circuitos de pluripotencia y la EMT. El módulo de EMT incluye los factores de transcripción de la familia SNAIL y ZEB, así como los miARNs miR-200 y miR-34. El módulo de pluripotencia considera los factores de Yamanaka OCT4, SOX2 y NANOG. Se analiza el comportamiento de ambos módulos por separado y luego se los conectan, en base a evidencia experimental y predicciones bioinformáticas. El modelo integrado predice que la influencia de la pluripotencia sobre la EMT puede modular el comportamiento celular al expandir la región de parámetros que permite la existencia del estado híbrido y modular la reversibilidad de la transición inversa (MET). Estos resultados son novedosos, ya que aportan una explicación a observaciones experimentales contradictorias sobre la reversibilidad de la EMT.

El modelo acoplado de la EMT y la pluripotencia desarrollado predice que el fenotipo epitelial corresponde a un estado diferenciado y no pluripotente, el fenotipo mesenquimal al más pluripotente, y el fenotipo híbrido E/M a un estado intermedio con un amplio rango de valores de NANOG, un marcador de pluripotencia. Se ha propuesto en la literatura que la ventaja metastásica de las células híbridas E/M se debe a su capacidad para migrar en clústers y evadir el sistema inmune, mientras que las células mesenquimales, al migrar individualmente, son más susceptibles a ser detectadas y eliminadas *in vivo*. En este aspecto, las predicciones del modelo sugieren que las células mesenquimales presentan la mayor pluripotencia. Una

posible explicación de esta predicción podría ser que solo una pequeña fracción de las células mesenquimales logra evadir el sistema inmune e invadir órganos distantes. En cambio, las células híbridas E/M, con niveles moderadamente elevados de NANOG, podrían tener un mayor potencial metastásico debido a su combinación de pluripotencia, capacidad migratoria, adhesión celular y evasión inmune. Es decir, es posible que las células con mayor pluripotencia no representen necesariamente el mayor riesgo de metástasis, sino que la conjunción de distintas propiedades sea la que proporcione la capacidad metastásica. En conclusión, el modelo predice que las redes de EMT y pluripotencia operan de manera conjunta desempeñando un papel clave en la progresión tumoral y la diseminación metastásica.

En conclusión, en esta tesis se realizan contribuciones en aspectos metodológicos, al desarrollar herramientas para el análisis transcriptómico de células individuales y para el modelado de redes regulatorias génicas. Asimismo, se hacen aportes en el ámbito de la biología del cáncer, al aplicar el conjunto de herramientas desarrolladas y aprendidas para el estudio de la heterogeneidad, la pluripotencia y la EMT. Un aspecto fundamental de esta tesis es que fue desarrollada íntegramente utilizando datos públicos y herramientas computacionales de código abierto y gratuitas. Esta consideración fue tenida en cuenta desde la planificación inicial de la tesis, motivada en parte por la disponibilidad limitada de recursos en Argentina y aprovechando esta ventaja que brinda el enfoque teórico y computacional. Además, la creciente tendencia en la ciencia hacia la disponibilización de datos fortalece la calidad de la investigación y amplía las oportunidades para quienes desarrollamos ciencia en contextos con recursos limitados, permitiendo contribuir al avance del conocimiento sin la necesidad de generar datos propios. En línea con este principio, y con el objetivo de fomentar la reproducibilidad y accesibilidad de los resultados, todo el código y los análisis desarrollados en esta tesis son de código abierto y acceso gratuito junto con las publicaciones derivadas de esta tesis.

Apéndices

Apéndice A: Tablas de la Parte I

A

Proceso biológico	***Gene ontology BP***	***Gene Ontology ID***	**Descripción**
Proliferación celular	*Cell population proliferation*	GO:0008283	Multiplicación o reproducción de células, que da lugar a la expansión de una población celular.
Diferenciación celular	*Cell differentiation*	GO:0030154	Proceso de desarrollo en el cual una célula relativamente no especializada adquiere características estructurales y/o funcionales especializadas que la caracterizan como una célula específica. La diferenciación incluye los procesos involucrados en el compromiso de una célula hacia un destino específico y su desarrollo subsecuente hasta alcanzar el estado maduro.
Ciclo celular mitótico	*Mitotic cell cycle process*	GO:1903047	Un proceso que es parte del ciclo celular mitótico.
Migración celular	*Cell migration*	GO:0016477	Movimiento controlado y autopropulsado de una célula desde un sitio hacia un destino guiado por señales moleculares.
Respuesta inmune	*Immune response*	GO:0006955	Cualquier proceso del sistema inmunológico que funciona en la respuesta de un organismo ante una amenaza potencial interna o invasiva.
Diferenciación de células madre	*Stem cell differentiation*	GO:0048863	Proceso en el cual una célula relativamente no especializada adquiere las características especializadas de una célula madre. Una célula madre es aquella que retiene la capacidad de dividirse y proliferar durante toda la vida para proporcionar células progenitoras que pueden diferenciarse en células especializadas.
Ciclo celular	*Cell cycle*	GO:0007049	Progresión de eventos bioquímicos y estados morfológicas que ocurren en una célula durante sucesivos eventos de replicación celular o replicación nuclear. El ciclo celular comprende la replicación y segregación del material genético, seguida de la división de la célula.
Replicación del ADN	*DNA replication*	GO:0006260	Proceso metabólico celular en el que una célula duplica una o más moléculas de ADN.

Continúa en la siguiente página

Proceso biológico	***Gene ontology BP***	***Gene Ontology ID***	**Descripción**
Reparación del ADN	*DNA repair*	GO:0006281	Proceso de restauración del ADN después de un daño. Los genomas están sujetos a daño por agentes químicos y físicos del ambiente y por radicales libres o agentes alquilantes generados endógenamente en el metabolismo. El ADN también se daña debido a errores durante su replicación. Se han descrito diversas rutas de reparación del ADN que incluyen la reversión directa, la reparación por escisión de bases, la reparación por escisión de nucleótidos, la fotorreactivación y la reparación de rupturas de doble cadena.
Respuesta inflamatoria	*Inflammatory response*	GO:0006954	Reacción defensiva inmediata (por vertebrados) ante una infección o lesión causada por agentes químicos o físicos. El proceso se caracteriza por vasodilatación local, extravasación de plasma hacia los espacios intercelulares y acumulación de glóbulos blancos y macrófagos.
Proliferación de células madre	*Stem cell proliferation*	GO:0072089	Multiplicación o reproducción de células madre, que da lugar a la expansión de una población de células madre.
Secreción hormonal	*Hormone secretion*	GO:0046879	Liberación regulada de hormonas, las cuales tienen un efecto regulador específico sobre un órgano o grupo de células en particular.
Respuesta inflamatoria aguda	*Acute inflammatory response*	GO:0002526	Inflamación que comprende una respuesta rápida, de corta duración y relativamente uniforme ante una lesión aguda o un desafío antigénico, y que se caracteriza por la acumulación de líquido, proteínas plasmáticas y leucocitos granulocíticos. Una respuesta inflamatoria aguda ocurre en cuestión de minutos u horas, y se resuelve en unos pocos días o se convierte en una respuesta inflamatoria crónica.
Respuesta inflamatoria crónica	*Chronic inflammatory response*	GO:0002544	Inflamación de larga duración (semanas o meses) en la cual la inflamación activa, la destrucción tisular y los intentos de reparación ocurren simultáneamente. Aunque puede seguir a una inflamación aguda, la inflamación crónica con frecuencia comienza como una respuesta de bajo grado, persistente y, a menudo, asintomática.
Respuesta inmune adaptativa	*Adaptive immune response*	GO:0002250	Respuesta inmunitaria mediada por células que expresan receptores específicos para antígenos y que permiten una respuesta secundaria mejorada a exposiciones subsecuentes al mismo antígeno (memoria inmunológica).

Continúa en la siguiente página

Proceso biológico	*Gene ontology BP*	*Gene Ontology ID*	Descripción
Respuesta inmune humoral	*Humoral immune response*	GO:0006959	Respuesta inmunitaria mediada a través de un fluido corporal.
Respuesta inmune innata	*Innate immune response*	GO:0045087	Respuestas defensivas mediadas por componentes codificados en la línea germinal que reconocen directamente componentes de patógenos potenciales.

Tabla A.1.: Procesos biológicos para los cuales se calcula la actividad de la PPIN asociada y el nombre original del BP (por sus siglas en inglés, *Biological Process*) en la base de datos QuickGO. ID identificador del proceso y descripción que proporciona la base de datos.

Nombre de muestra	ID GEO	# células antes	Máx. % mitocondrial	# genes mín.	# genes máx,	# células después	# genes detectados
N-0372-epi	GSM4909276	12212	20	500	5000	11162	18453
TN-0126	GSM4909281	3666	20	500	7000	2766	15821
TN-0135	GSM4909282	15870	20	500	5000	15212	18431
TN-B1-0431	GSM4909286	5581	20	500	5000	5228	17747
TN-B1-0131	GSM4909285	5880	20	500	5000	5688	17773
TN-B1-0554	GSM4909283	9593	30	500	5000	6787	19462
TN-B1-0717	GSM4909284	21130	40	500	7000	16865	18279
HER2-0138	GSM4909287	4779	25	500	7000	4160	18541
HER2-0337	GSM4909288	12288	25	500	5000	8252	18319
HER2-0321	GSM4909289	16073	25	500	7000	13057	18577
HER2-0161	GSM4909290	6561	20	500	5000	6335	18335
HER2-0285	GSM4909291	19027	20	500	5000	17992	18517
ER-0001	GSM4909277	7393	30	500	5000	5754	19377
ER-0125	GSM4909278	7289	30	500	5000	4226	17798
ER-0074	GSM4909279	6676	30	500	5000	3714	17489
ER-0042	GSM4909280	5439	20	500	3500	4529	16860
ER-0052	GSM4909280	10592	20	500	5000	8102	18364
ER-0163	GSM4909304	7334	30	500	3500	5990	18312

Tabla A.2.: Detalles del preprocesamiento de las muestras: nombre y ID de GEO de las muestras, número de células y genes antes y después del filtrado y valores de corte del contenido mitocondrial del número de genes detectados por célula.

Score	Subtipos comparados	p-valores	q-valores	Significativo
$\langle ACT^{LB-dn} \rangle$	ER+ vs. HER2+	0.02247887336612529	0.035966197385800466	*
$\langle ACT^{LB-dn} \rangle$	ER+ vs. TN	0.00305279853864276	0.015224604293820772	*
$\langle ACT^{LB-dn} \rangle$	TN vs. HER2+	0.013710830078133254	0.025312301682707546	*
$\langle ACT^{LB-up} \rangle$	ER+ vs. HER2+	0.5228166539190887	0.5455478127851361	
$\langle ACT^{LB-up} \rangle$	ER+ vs. TN	0.005074868097940257	0.015224604293820772	*
$\langle ACT^{LB-up} \rangle$	TN vs. HER2+	0.008113117265565782	0.021634979374842087	*
$\langle ACT^{LM-dn} \rangle$	ER+ vs. HER2+	0.08283742515880639	0.11694695316537373	
$\langle ACT^{LM-dn} \rangle$	ER+ vs. TN	0.00305279853864276	0.015224604293820772	*
$\langle ACT^{LM-dn} \rangle$	TN vs. HER2+	0.013710830078133254	0.025312301682707546	*
$\langle ACT^{LM-up} \rangle$	ER+ vs. HER2+	0.31530245208174557	0.37836294249809466	
$\langle ACT^{LM-up} \rangle$	ER+ vs. TN	0.005074868097940257	0.015224604293820772	*
$\langle ACT^{LM-up} \rangle$	TN vs. HER2+	0.004656257247125864	0.004656257247125864	*
$\langle CNA \rangle$	ER+ vs. HER2+	0.004656257247125864	0.015224604293820772	*
$\langle CNA \rangle$	ER+ vs. TN	0.005074868097940257	0.015224604293820772	*
$\langle CNA \rangle$	TN vs. HER2+	0.9272644735252321	0.9272644735252321	
$\langle ACT^{CC} \rangle$	ER+ vs. HER2+	0.013710830078133254	0.025312301682707546	*
$\langle ACT^{CC} \rangle$	ER+ vs. TN	0.020240570577077503	0.034698120989275716	*
$\langle ACT^{CC} \rangle$	TN vs. HER2+	0.17090352023079755	0.22787136030773006	
$\langle H \rangle$	ER+ vs. HER2+	0.004656257247125864	0.015224604293820772	*
$\langle H \rangle$	ER+ vs. TN	0.013065226764425985	0.025312301682707546	*
$\langle H \rangle$	TN vs. HER2+	0.41131379177625904	0.47007290488715314	
$\langle ACT^{EMT} \rangle$	ER+ vs. HER2+	0.23533326650958586	0.2972630734857927	
$\langle ACT^{EMT} \rangle$	ER+ vs. TN	0.04532756207797217	0.06799134311695826	
$\langle ACT^{EMT} \rangle$	TN vs. HER2+	0.522816653919088	0.5455478127851361	

Tabla A.3.: Comparación de diferentes puntajes entre subtipos tumorales. Se muestran los valores p y q, junto con la indicación de significancia estadística.

Apéndice B: Figuras suplementarias

B

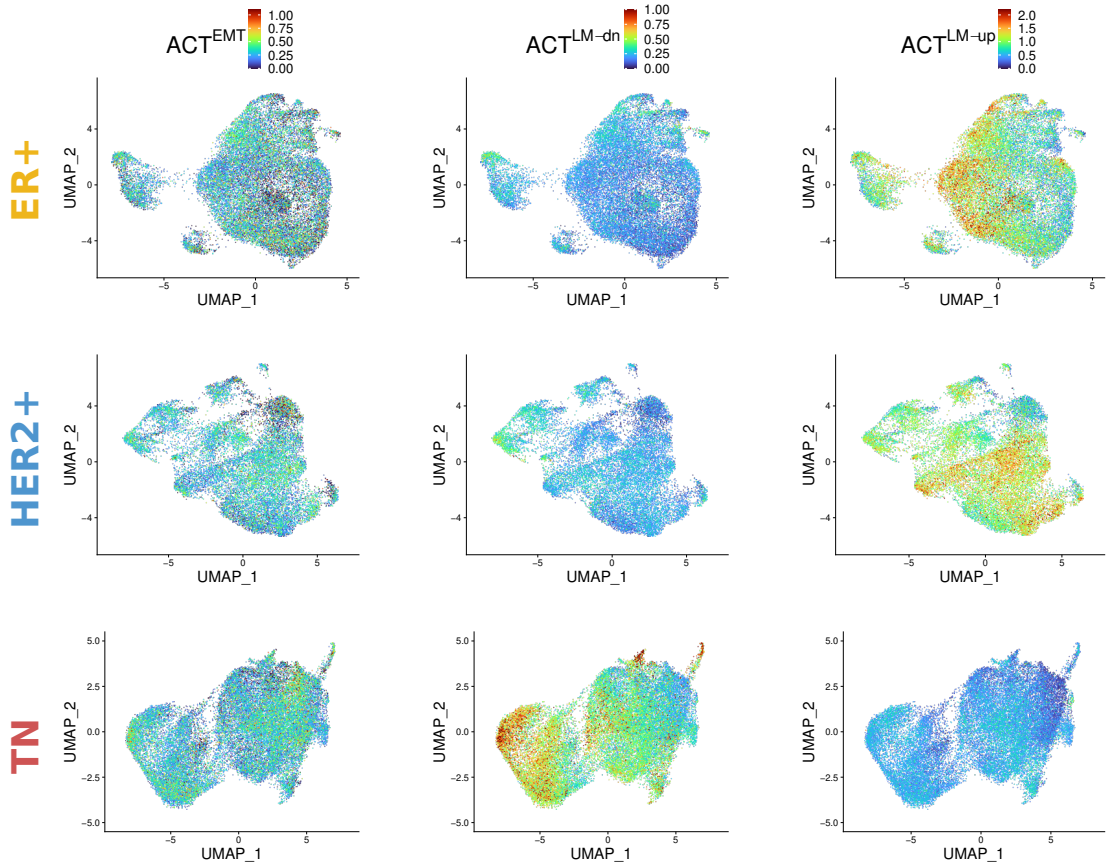

Figura B.1.: Visualización UMAP de las muestras integradas por subtipo de cáncer de mama por separado (ER+, HER2+, and TN). Código de colores de las células según los *scores* ACT^{EMT}, ACT^{LM-dn}, and $ACT^{LM\text{-}up}$

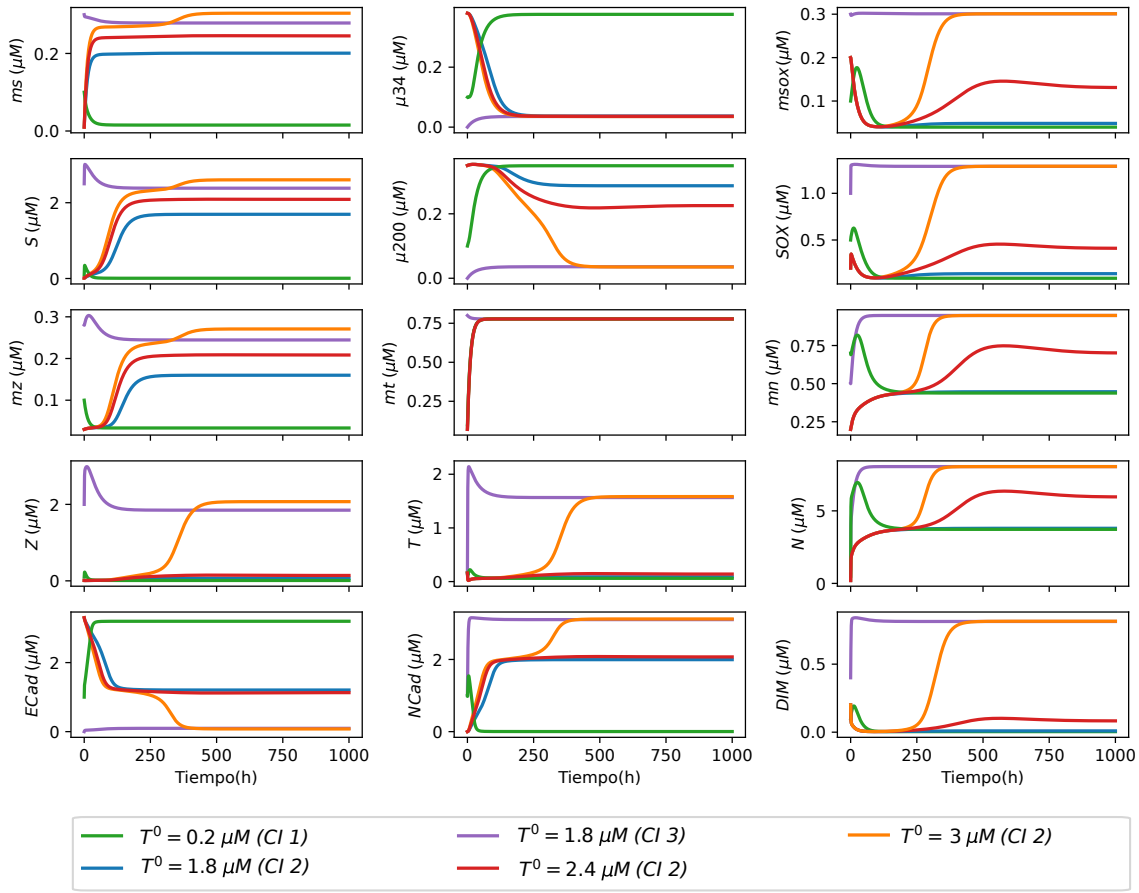

Figura B.2.: Evolución temporal de todos los elementos del circuito completo (EMT+pluripotencia) para distintas las condiciones iniciales de la Tabla D.6 y valores TGF-β exógeno T^0. Las unidades de las concentraciones se encuentran en μM.

C

Apéndice C: Formalismo regulación por factores de transcripción

En este apéndice se derivan las expresiones para modelar la regulación por factores de transcripción.

C.1. Regulación de la expresión génica por un factor de transcripción

Regulación de la transcripción por un activador

Comenzamos con el caso más simple de regulación de la expresión génica por un único factor de transcripción, en este caso un activador. Sean P el promotor del gen x y A un activador que regula la transcripción del gen x, A se une a P formando el complejo AP con tasa k_{on} y el complejo se disocia con tasa k_{off}. El diagrama de las reacciones que ocurren es de la siguiente forma:

$$A + P \rightleftarrows AP.$$

La región promotora puede encontrarse libre (P) o formando un complejo con el activador (AP). Por conservación, la concentración total de P se mantiene constante:

$$[P_T] = [P] + [AP]\,. \tag{C.1}$$

El proceso de formación del complejo AP se puede formular utilizando cinética de acción de masas. De este modo, la velocidad de formación del complejo es proporcional a la tasa de unión multiplicado por la concentración de los reactivos A y P. Por otro lado, la tasa de disociación del complejo es el producto entre la tasa de disociación y la concentración de AP. Así, se puede escribir la tasa de cambio del complejo AP:

$$\frac{d\,[AP]}{dt} = k_{on}\,[A]\,[P] - k_{off}\,[AP]\,. \tag{C.2}$$

Los eventos de unión y disociación del complejo ocurren de forma mucho más rápida en la escala temporal que la transcripción y traducción. La ec. C.2 se aproxima al estado estacionario en el cual las concentraciones no cambian, $\frac{d[AP]}{dx} = 0$ y, se encuentra un balance entre la formación y disociación del complejo:

$$k_d\,[AP] = [A]\,[P]\,, \tag{C.3}$$

donde $k_d = k_{off}/k_{on}$ es la constante de disociación. Utilizando las ecuaciones C.1 y C.3, se obtienen las probabilidades de encontrar al promotor en complejo con el activador y libre:

$$\begin{aligned} \frac{[AP]}{[P_T]} &= \frac{[A]/k_d}{1+[A]/k_d}, \\ \frac{[P]}{[P_T]} &= \frac{1}{1+[A]/k_d}. \end{aligned} \tag{C.4}$$

La primera ec. C.4 describe la probabilidad de que el promotor esté en complejo con el factor de transcripción, y es conocida como la ecuación de Michaelis-Menten en el contexto de la cinética enzimática. La constante k_d, denominada constante de Michaelis-Menten, corresponde a la concentración del activador en la cual la tasa de transcripción alcanza la mitad de su velocidad máxima. En sistemas que siguen una cinética simple de Michaelis-Menten, esta constante representa la constante de disociación del complejo AP (inversa a la afinidad entre el activador y el promotor). Valores bajos de k_d indican que el complejo AP está fuertemente unido y se disocia con poca frecuencia. La tasa de transcripción del gen x es la suma de las contribuciones cuando el promotor está en complejo con el activador y cuando está libre:

$$tasa\,de\,producción\,(activador) = g_A \frac{[AP]}{[P_T]} + g_0 \frac{[P]}{[P_T]}. \tag{C.5}$$

Utilizando las ecs. C.4 y C.5

$$tasa\,de\,producción\,(activador) = g_A \frac{[A]/k_d}{1+[A]/k_d} + g_0 \frac{1}{1+[A]/k_d}. \tag{C.6}$$

Utilizando la propiedad que $\frac{[A]/k_d}{1+[A]/k_d} + \frac{[1]/k_d}{1+[A]/k_d} = 1$ se puede reescribir la tasa de producción del ARNm de la siguiente forma:

$$tasa\,de\,producción\,(activador) = (g_A - g_0) \frac{[A]/k_d}{1+[A]/k_d} + g_0. \tag{C.7}$$

donde g_0 es la tasa basal de producción del ARNm y g_A es la máxima tasa de transcripción, que ocurre cuando A está unido al promotor. En la Figura C.1A se muestra cómo es la producción del ARNm en función de la concentración de un activador (caso $n = 1$). Cuando la concentración del activador aumenta, la función de activación también, y se activa la transcripción del gen. Esta función de regulación para activadores aumenta linealmente con $[A]$ cuando $[A] \ll k_d$ y alcanza la saturación para valores altos de $[A]$. La tasa máxima de producción de ARNm es g_A, que corresponde a la tasa de transcripción cuando el promotor está en complejo con el activador. Cuando la concentración del activador es nula, la tasa de transcripción es la basal (g_0), usualmente pequeña o nula. En la literatura se utilizan distintas convenciones para denotar los parámetros aquí llamados g_0 y g_A. Una de las convenciones más utilizadas consiste en denotar a $g_A - g_0$ mediante el parámetro β. Debido a que la transcripción basal es pequeña, frecuentemente al parámetro β se lo llama tasa de transcripción máxima. Por otro lado, se suele llamar k^0 a la tasa de transcripción basal aquí representada como g_0. Por este motivo, cuando modelemos sistemas más complejos, se utilizará dicha notación.

Regulación de la transcripción por un represor

Denotamos por R a un factor de transcripción que se une al promotor P e inhibe la transcripción. El diagrama que representa las reacciones de asociación y disociación del represor con el promotor tiene la siguiente forma:

$$R + P \rightleftarrows RP.$$

Desarrollando de manera análoga a la sección anterior para el caso de un activador, podemos obtener la tasa de producción de ARNm:

$$tasa\,de\,producción\,(represor) = g_R\,[RP] + g_0\,[P]\,. \tag{C.8}$$

Las ecs. C.4, que describen la probabilidad de que el promotor se encuentre libre o en complejo con un activador, son igualmente aplicables al caso de un represor. Además, se cumple la relación $\frac{[R]/k_d}{1+[R]/k_d} + \frac{[1]/k_d}{1+[R]/k_d} = 1$. Sustituyendo estas expresiones en la ecuación C.8, se obtiene:

$$tasa\,de\,producción\,(represor) = (g_0 - g_R)\,\frac{1}{1 + [R]\,/k_d} + g_R. \tag{C.9}$$

En la Figura C.1B se muestra cómo es la producción del ARNm en función de la concentración de un represor (caso $n = 1$). Al aumentar la concentración del inhibidor, la función de regulación disminuye, lo que provoca una reducción en la tasa de transcripción. En la ec. C.9, cuando el promotor está saturado a altas concentraciones del inhibidor, la tasa de transcripción se aproxima a g_R, comúnmente denominada fuga. Esta fuga representa una transcripción residual que ocurre incluso cuando el represor está presente en altas concentraciones. Sin embargo, en muchos casos, la fuga es despreciable y se considera que esta constante es nula. Por otro lado, cuando la concentración del represor es baja, la tasa de expresión alcanza su valor máximo (g_0), que corresponde a la tasa de transcripción del promotor libre. De manera análoga al caso de un activador, en los siguiente capítulos utilizaremos el parámetro β en lugar de $g_0 - g_R$, comúnmente referido como la tasa máxima de transcripción, y k^0 para denotar la tasa de fuga.

C.2. Cooperatividad

La mayoría de los factores de transcripción forman multímeros de varias subunidades proteicas repetidas. En este caso se puede derivar la función de regulación asumiendo que n moléculas del factor de transcripción A, en este caso un activador, se pueden unir al promotor P. Para describir este proceso, el enfoque más utilizado consiste en asumir que n moléculas de activadores (nA) se unen al promotor prácticamente en simultáneo, dando lugar al complejo A_nP. Se puede describir esta reacción como se hace en las secciones anteriores:

$$nA + P \rightleftarrows A_nP.$$

Por conservación, la concentración total del promotor P_T será igual a la suma de las concentraciones del promotor libre P y del complejo promotor-activadores A_nP:

$$[P_T] = [P] + [A_nP]\,. \tag{C.10}$$

La velocidad de formación del complejo $[A_nP]$ está dada por el producto de la concentración del promotor libre y la concentración del activador elevada a la potencia n, y la tasa de disociación del complejo será el producto de la constante de disociación :

$$\frac{d\left[A_nP\right]}{dt} = k_{on}\left[A\right]^n\left[P\right] - k_{off}\left[A_nP\right]. \tag{C.11}$$

Aproximando al estado cuasiestacionario, es decir $\frac{d[A_nP]}{dt} \approx 0$, y definiendo $k_d^n = k_{off}/k_{on}$ se obtiene la relación de balance entre la formación y disociación del complejo:

$$k_d^n\left[A_nP\right] = \left[A\right]^n\left[P\right]. \tag{C.12}$$

Utilizando las ecs. C.10 y C.12 se obtienen la probabilidad de encontrar el promotor libre y en complejo con el multímero:

$$\begin{aligned} \frac{[P]}{[P_T]} &= \frac{1}{1+\left(\left[A\right]/k_d\right)^n}, \\ \frac{[A_nP]}{[P_T]} &= \frac{\left(\left[A\right]/k_d\right)^n}{1+\left(\left[A\right]/k_d\right)^n}. \end{aligned} \tag{C.13}$$

La tasa de producción de ARNm es la suma de las probabilidades de encontrar al operador en los distintos estados (P y A_nP) multiplicado por la tasa de transcripción en cada estado:

$$tasa\,de\,producción\; = g_0\frac{[P]}{[P_T]} + g_A\frac{[A_nP]}{[P_T]}. \tag{C.14}$$

Sustituyendo las ecs. C.13 en la ec. C.14 y, utilizando la relación $\frac{1}{1+([A]/k_d)^n} + \frac{([A]/k_d)^n}{1+([A]/k_d)^n} = 1$ se tiene que la producción para un activador es:

$$tasa\,de\,producción\,(activador) = \left(g_A - g_0\right)\frac{\left(\left[A\right]/k_d\right)^n}{1+\left(\left[A\right]/k_d\right)^n} + g_0. \tag{C.15}$$

g_A representa la tasa máxima de transcripción, alcanzada cuando la concentración del activador es elevada. Por otro lado, g_0 denota la tasa de transcripción basal, correspondiente a la situación en la que la concentración del activador es nula. En muchos casos, esta constante es tan pequeña que se puede aproximar a 0. A medida que la concentración del activador aumenta, la función de activación se incrementa, estimulando la transcripción del gen.

Un desarrollo similar se puede realizar para factores de transcripción represores R:

$$tasa\,de\,producción\,(represor) = \left(g_0 - g_R\right)\frac{1}{1+\left(\left[R\right]/k_d\right)^n} + g_R. \tag{C.16}$$

En este contexto, g_0 representa la tasa de transcripción cuando el represor no está unido al operador. En ausencia de represor, la tasa total de producción de ARNm se aproxima a g_0, que corresponde a la máxima tasa de producción. Por el contrario, cuando la concentración del represor es elevada, la transcripción es reprimida, y la tasa de producción disminuye hasta aproximarse a g_R, conocida como fuga. Esta constante es generalmente pequeña y, en muchos casos, puede considerarse despreciable.

Las funciones de regulación presentadas en las Ecs. C.15 y C.16 se conocen como funciones de Hill, donde n es el coeficiente de Hill, que refleja el grado de cooperatividad en la unión del activador o represor. Cuando no hay cooperatividad ($n = 1$), la función se reduce a

la ecuación de Michaelis-Menten. Un mayor coeficiente de Hill ($n > 1$) indica una mayor cooperatividad, lo que resulta en una respuesta más pronunciada de la función de regulación, ya sea activadora o represora, como se puede ver en la Figura C.1.

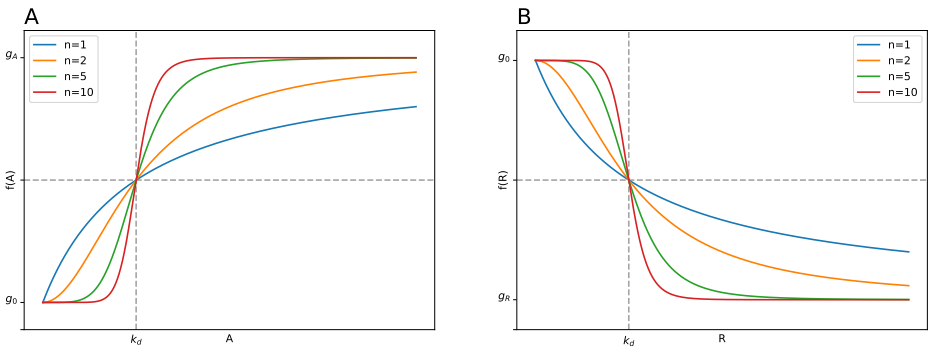

Figura C.1.: (**A**) Función de Hill para el caso de un activador para distintos valores del coeficiente de Hill. (**B**)Función de Hill para un represor para distintos valores del coeficiente de Hill. La línea de puntos gris indica el valor de la concentración del regulador (k_d) para el cual la tasa de producción del ARNm alcanza la mitad de su valor máximo.

En este planteamiento sencillo, no se consideran los estados intermedios en los cuales se encuentran unidos menos de n moléculas del factor de transcripción al promotor. Es decir, que se asume cooperatividad positiva extrema. Planteando las reacciones para el promotor uniéndose a $1, 2, ..., n$ factores de transcripción, se pueden derivar las funciones de regulación considerando los estados intermedios. La cooperatividad implica que las constantes de disociación son distintas para cada estado. Realizando un procedimiento similar para encontrar la fracción de ocupación del promotor, se llega a la ecuación de Adair [198]. Considerando que la última reacción tiene mucha mayor afinidad que los eventos anteriores, la ecuación de Adair se reduce a la función de Hill. Esta suposición es equivalente a asumir que al unirse una molécula se unen prácticamente en simultáneo n moléculas al promotor, que es la suposición que se hace al plantear las reacciones.

Cabe destacar que estos modelos son los utilizados para la regulación de proteínas. En la transcripción, para el caso de factores de transcripción que se unen como multímeros a un promotor, este análisis no aplica directamente ya que la multimerización ocurre en el citosol en vez de en el sitio de unión del promotor. Sin embargo, también se obtienen funciones de Hill con una constante distinta a la constante de disociación. A fines prácticos del modelado, se utilizan las mismas ecuaciones diferenciales con valores distintos de la constante de disociación [198].

C.3. Regulación de la transcripción por dos factores de transcripción

La transcripción de los genes es comúnmente regulada por múltiples factores de transcripción. Para ello, es necesario analizar la ocupación del promotor de un gen regulado por dos factores de transcripción. La unión de dos factores de transcripción al promotor puede ser no competitiva, los factores de transcripción tienen sitios de unión diferentes (Fig. C.2A), o competitiva, es decir, que comparten el mismo sitio de unión (Fig. C.2B).

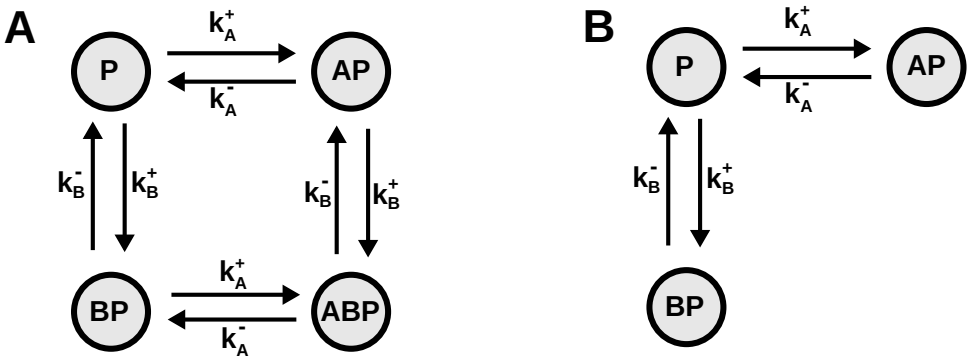

Figura C.2.: Regulación de la transcripción por dos factores de transcripción. (**A**) Caso no competitivo, A y B tienen dos sitios de unión distintos al promotor. (**B**) Caso competitivo, A y B comparten el mismo sitio de unión.

Caso no competitivo

Siguiendo un procedimiento análogo al empleado para el caso de la regulación por un único factor de transcripción, es posible derivar las fracciones de ocupación correspondientes a los cuatro estados posibles del promotor, que se esquematizan en la Figura C.2A:

$$\begin{aligned}
\frac{[P]}{[P_T]} &= \frac{1}{[1+(A/A_0)^{n_A}]}\frac{1}{[1+(B/B_0)^{n_B}]} = H^-(A)\,H^-(B)\,,\\
\frac{[AP]}{[P_T]} &= \frac{(A/A_0)^{n_A}}{[1+(A/A_0)^{n_A}]}\frac{1}{[1+(B/B_0)^{n_B}]} = H^+(A)\,H^-(B)\,,\\
\frac{[BP]}{[P_T]} &= \frac{1}{[1+(A/A_0)^{n_A}]}\frac{(B/B_0)^{n_B}}{[1+(B/B_0)^{n_B}]} = H^-(A)\,H^+(B)\,,\\
\frac{[ABP]}{[P_T]} &= \frac{(A/A_0)^{n_A}}{[1+(A/A_0)^{n_A}]}\frac{(B/B_0)^{n_B}}{[1+(B/B_0)^{n_B}]} = H^+(A)\,H^+(B)\,,
\end{aligned} \tag{C.17}$$

donde H^+ y H^- son las funciones de Hill activadoras e inhibidoras. La tasa de transcripción total del gen x tiene la siguiente forma:

$$G(A,B) = g_0\frac{[P]}{[P_T]} + g_A\frac{[AP]}{[P_T]} + g_B\frac{[BP]}{[P_T]} + g_{AB}\frac{[ABP]}{[P_T]}. \tag{C.18}$$

Uniendo las ecs. C.17 y C.18. La tasa de transcripción total de un gen regulado por dos factores de transcripción es

$$G(A,B) = g_0\,H^-(A)\;H^-(B) + g_A\,H^+(A)\;H^-(B) + g_B\,H^-(A)\;H^+(B) + g_{AB}\,H^+(A)\;H^+(B)\,. \tag{C.19}$$

Reescribiendo las tasas de producción en función de g_0, la tasa de producción en el caso de que ningún regulador está ligado al promotor, tal que $g_A = \lambda_A\,g_0,\ g_B = \lambda_B\,g_0,\ g_{AB} = \lambda_A\lambda_B\,g_0$,

se obtiene una forma alternativa de la ec. C.19 también utilizada en la literatura [204, 205, 247]:

$$G\left(A,B\right)=g_0\ H^g\left(A\right)\ H^g\left(B\right), \tag{C.20}$$

donde $H^g\left(A\right)=H^-\left(A\right)+\lambda_A H^+\left(A\right)=\frac{1+\lambda_A(A/A_0)^{n_A}}{1+(A/A_0)^{n_A}}$, que se la puede encontrar denominada como función de Hill *shifted*, aquí la llamaremos función de Hill generalizada. Si $0<\lambda_A<1$ corresponde a un represor $0\leq\lambda_A<1$, a menor valor es un represor más fuerte, siendo el represor perfecto $\lambda_A=0$. Para valores de $\lambda_A>1$ el regulador es un activador y para $\lambda_A=1$ no hay regulación. Dejaremos esta expresión para más adelante.

Por otro lado, utilizando que $1=H^+\left(X\right)+H^-\left(X\right)$ la ec. C.19 se puede reescribir de forma conveniente para analizar distintos escenarios:

$$\begin{aligned}
G_R\left(A,B\right)&=\left(g_0+g_{AB}-g_A-g_B\right)H^-\left(A\right)H^-\left(B\right)+\left(g_A-g_{AB}\right)H^-\left(A\right)+\left(g_B-g_{AB}\right)H^-\left(B\right)+g_{AB},\\
G_A\left(A,B\right)&=\left(g_0+g_{AB}-g_A-g_B\right)H^+\left(A\right)H^+\left(B\right)+\left(g_A-g_0\right)H^+\left(A\right)+\left(g_B-g_0\right)H^+\left(B\right)+g_0,\\
G_{A/R}\left(A,B\right)&=\left(g_A+g_B-g_0-g_{AB}\right)H^+\left(A\right)H^-\left(B\right)+\left(g_{AB}-g_B\right)H^+\left(A\right)+\left(g_{AB}-g_B\right)H^-\left(B\right)+g_B,
\end{aligned} \tag{C.21}$$

donde $G_R\left(A,B\right)$ y $G_A\left(A,B\right)$ son las tasas de producción de ARNm para el caso de dos represores y dos activadores. La ec. C.19 describe de manera general la producción de un gen regulado por dos factores de transcripción, considerando tanto su acción independiente como su potencial sinérgico cuando ambos están presentes. Es posible analizar distintos escenarios de regulación. Cuando los factores de transcripción actúan de manera autónoma y sin influencia mutua, este comportamiento se denomina *OR* en la literatura. Por otro lado, también se puede estudiar el caso en que ambos reguladores solo ejercen su función al actuar de forma conjunta, lo cual se conoce como *AND*.

- ***OR* (dos represores):** g_0 es la máxima tasa de producción ya que corresponde a la situación en el que ningún represor está unido. De esta forma, g_A-g_0 , g_B-g_0 y $g_{AB}-g_0$ representan la disminución de la tasa de transcripción cuando el represor A, el represor B y ambos están unidos al promotor respecto a la situación en la que el promotor se encuentra libre. Si los represores son independientes entre sí, la disminución de la transcripción cuando ambos represores están unidos al promotor debe ser igual a la suma de la disminución cuando está unido solo A o solo B: $g_{AB}-g_0=(g_A-g_0)+(g_B-g_0)$. Utilizando esta relación y la ec. C.21 para represores, se obtiene la función de regulación para dos represores independientes:

  $$G_R^{OR}\left(A,B\right)=\left(g_A-g_{AB}\right)\ H^-\left(A\right)+\left(g_B-g_{AB}\right)\ H^-\left(B\right)+g_{AB}.$$

  g_{AB} es la tasa de producción cuando ambos factores están unidos, es decir, es el término de fuga que se menciona anteriormente. Si tomamos el caso en el que la fuga es nula se tiene que la tasa de transcripción total es la suma de las funciones de regulación de ambos represores por separado:

  $$G_R^{OR}\left(A,B\right)=g_A\ H^-\left(A\right)+g_B\ H^-\left(B\right).$$

- ***OR* (dos activadores):** De manera análoga al razonamiento realizado en el caso anterior, el aumento de la máxima tasa de transcripción cuando ambos activadores están unidos respecto al caso de promotor libre es igual a la suma del aumento de las tasas de producción cuando están unidos cada uno por separado. Esto significa que $g_{AB}-g_0=(g_A-g_0)+(g_B-g_0)$. Tomando esta expresión y la ec. C.21 para activadores

la tasa de producción de ARNm para dos activadores que actúan independientemente es:

$$G_A^{OR}(A,B) = (g_A - g_0)\ H^+(A) + (g_B - g_0)\ H^+(B) + g_0.$$

La constante g_0 es la producción cuando ningún activador esta presente, es decir la producción basal. Considerando que la producción basal es despreciable, como se hace frecuentemente, se encuentra que la producción de transcriptos total es la suma de ambas funciones de activación:

$$G_A^{OR}(A,B) = g_A\ H^+(A) + g_B\ H^+(B).$$

- ***OR* (un activador y un represor):** Supongamos que A es activador y B represor. Considerando que el efecto combinado de los dos factores de transcripción es igual a la suma del efecto de ambos por separado, es decir, sin considerar sinergia, se tiene $g_{AB} - g_0 = (g_A - g_0) + (g_B - g_0)$. Utilizando la ec. C.21 para un activador y un represor se obtiene:

$$G_{A/R}(A,B) = (g_{AB} - g_B)\ H^+(A) + (g_0 - g_B)\ H^-(B) + g_B.$$

Donde g_B es la tasa de producción de ARNm cuando el represor está unido pero el activador no, representaría la producción basal por A y fuga por B. Si asumimos que esta tasa mínima de producción es nula, la función de regulación puede simplificarse como la suma de las funciones de *Hill* de cada factor de transcripción por separado:

$$G_{A/R}^{OR}(A,B) = g_{AB}\ H^+(A) + g_0\ H^-(B).$$

- ***AND* (dos represores):** En este caso la transcripción solo está activada cuando A y B no están unidos al promotor. Así $g_A = g_B = g_{AB}$ y $g_0 > g_{AB}$ y, utilizando la ec. C.21 para represores se tiene que:

$$G_R^{AND}(A,B) = (g_0 - g_{AB})\ H^-(A)\ H^-(B) + g_{AB}.$$

En este caso, la represión es parcial, ya que a concentraciones altas de los represores hay transcripción basal, conocida como fuga. Si la fuga es nula, se tiene que la producción del ARNm regulada por dos represores que actúan solo en conjunto es el producto de ambas funciones de regulación:

$$G_R^{AND}(A,B) = g_0\ H^-(A)\ H^-(B).$$

- ***AND* (dos activadores):** La producción del ARNm se activa solo cuando A y B están unidos. Utilizando que $g_A = g_B = g_0$ en la ec. C.21 para activadores se tiene que:

$$G_A^{AND}(A,B) = (g_{AB} - g_0)\ H^+(A)\ H^+(B) + g_0.$$

En la situación en que la producción basal es nula se llega a que la producción es el producto de las funciones de activación de cada activador por separado:

$$G_A^{AND}(A,B) = g_{AB}\ H^+(A)\ H^+(B).$$

- ***AND* (un activador y un represor):** Consideremos un escenario en el que A actúa como activador y B como represor. En este caso, la producción de ARNm está activada únicamente cuando A está unido y B no lo está, lo que implica que $g_{AB} = g_B = g_0$ y $g_A > g_{AB}$. Mediante la ec. C.21 para un sistema con un activador y un represor, la tasa de producción de ARNm se expresa como:

$$G_{A/R}^{AND}(A,B) = (g_A - g_B)\ H^+(A)\ H^-(B) + g_B,$$

Donde g_B es la tasa de producción de ARNm cuando el represor B está unido pero el activador A no. Si se asume la producción basal o fuga $(g_B = 0)$, la tasa de producción del transcripto se reduce al producto de las funciones de regulación de ambos factores de transcripción por separado:

$$G_{A/R}^{AND}(A,B) = g_A\ H^+(A)\ H^-(B).$$

Del análisis previo se observa que, en el caso de una regulación conjunta no sinérgica (*OR*) por dos factores de transcripción, la función de regulación es la suma de las funciones de regulación individuales de cada factor de transcripción. Por otro lado, si se asume que la transcripción solo se activa cuando los activadores están unidos y los represores no lo están de manera simultánea al promotor (*AND*), la función de regulación conjunta se describe como el producto de las funciones de regulación de cada factor de transcripción por separado.

Caso competitivo

Nuevamente, desarrollando de forma similar a lo realizado para la regulación por un solo factor de transcripción, se pueden obtener las fracciones de ocupación asociadas a los tres estados del promotor ilustrados en la Figura C.2B:

$$\begin{aligned}
\frac{[P]}{[P_T]} &= \frac{1}{1 + (A/A_0)^{n_A} + (B/B_0)^{n_B}} = H^-_{comp}(A,B), \\
\frac{[AP]}{[P_T]} &= \frac{(A/A_0)^{n_A}}{1 + (A/A_0)^{n_A} + (B/B_0)^{n_B}} = H^+_{comp}(A,B), \\
\frac{[BP]}{[P_T]} &= \frac{(B/B_0)^{n_B}}{1 + (A/A_0)^{n_A} + (B/B_0)^{n_B}} = H^-_{comp}(B,A).
\end{aligned} \tag{C.22}$$

Por conservación tenemos que:

$$H^-_{comp}(A,B) + H^+_{comp}(A,B) + H^-_{comp}(B,A) = 1. \tag{C.23}$$

La tasa de transcripción total es

$$G(A,B) = g_0 \frac{[P]}{[P_T]} + g_A \frac{[AP]}{[P_T]} + g_B \frac{[BP]}{[P_T]}. \tag{C.24}$$

Reemplazando la ec. C.22 en la ec. C.24 se tiene que la transcripción es

$$G(A,B) = g_0\, H^-_{comp}(A,B) + g_A\, H^+_{comp}(A,B) + g_B\, H^-_{comp}(B,A). \tag{C.25}$$

La producción de ARNm de la ec. C.25 se puede reescribir de la siguiente forma:

$$G\left(A,B\right)=g_0\frac{1+\lambda_A\left(A/A_0\right)^{n_A}+\lambda_B\left(B/B_0\right)^{n_B}}{1+\left(A/A_0\right)^{n_A}+\left(B/B_0\right)^{n_B}}, \tag{C.26}$$

donde $\lambda_A = g_A/g_0$ y $\lambda_B = g_B/g_0$ son las tasas de transcripción cuando el promotor está unido al factor de transcripción A y B respecto a la tasa de transcripción del promotor libre.

Se pueden derivar las tasas de producción de ARNm para diferentes casos particulares comúnmente utilizados en la bibliografía, considerando las distintas combinaciones de los tipos de regulación (activadores o represores):

- **Dos represores:** Utilizando las ecs. C.22 y C.24, y suponiendo que A y B inhiben la transcripción de igual forma ($g_A = g_B = g_{A/B}$), la expresión para la tasa de producción del ARNm regulado por dos represores que se unen al mismo sitio de unión es:

$$G\left(A,B\right)_R^{comp}=\left(g_0-g_{A/B}\right)\frac{1}{1+\left(A/A_0\right)^{n_A}+\left(B/B_0\right)^{n_B}}+g_{A/B}. \tag{C.27}$$

  Si la concentración de ambos represores es nula $G\left(A,B\right)_R^{comp}\approx g_0$, es decir la transcripción no se encuentra inhibida. Cuando la concentración de uno de los represores es elevada $G\left(A,B\right)_R^{comp}\approx g_{A/B}$, esto es la fuga. Para el caso en el que A y B sean represores ideales, $g_A = g_B = g_{A/B} = 0$, y la expresión se reduce a $G\left(A,B\right)_R^{comp}=g_0\frac{1}{1+\left(A/A_0\right)^{n_A}+\left(B/B_0\right)^{n_B}}$.

- **Dos activadores:** Siguiendo el mismo razonamiento, utilizando las ecs. C.22 y C.24, y también asumiendo que A y B aumentan la transcripción de igual forma, se puede derivar la tasa de transcripción para dos activadores que compiten por el mismo sitio de unión:

$$G\left(A,B\right)_A^{comp}=\left(g_{A/B}-g_0\right)\frac{\left(A/A_0\right)^{n_A}+\left(B/B_0\right)^{n_B}}{1+\left(A/A_0\right)^{n_A}+\left(B/B_0\right)^{n_B}}+g_0. \tag{C.28}$$

  En el límite en que la concentración de uno de los activadores es alta, la transcripción alcanza su valor máximo $G\left(A,B\right)_R^{comp}\approx g_{A/B}$. En contraste, cuando la concentración de ambos activadores es nula, la transcripción es igual la producción basal, es decir, $G\left(A,B\right)_R^{comp}\approx g_0$. Si la producción basal es despreciable, la tasa de transcripción se simplifica a: $G\left(A,B\right)_A^{comp}=g_{A/B}\frac{\left(A/A_0\right)^{n_A}+\left(B/B_0\right)^{n_B}}{1+\left(A/A_0\right)^{n_A}+\left(B/B_0\right)^{n_B}}$.

- **Un represor y un activador:** Siguiendo el mismo procedimiento que se realiza para los dos casos anteriores se tiene que la tasa de producción de ARNm para un gen que es regulado por un activador (A) y un represor (B) que compiten por el mismo sitio de unión es:

$$G\left(A,B\right)_{A/R}^{comp}=\left(g_0-g_B\right)\frac{1+\left(A/A_0\right)^{n_A}}{1+\left(A/A_0\right)^{n_A}+\left(B/B_0\right)^{n_B}}+g_B. \tag{C.29}$$

  Cuando la concentración del represor es nula o la concentración del activador es elevada respecto a la del represor la transcripción es máxima $G\left(A,B\right)_{A/R}^{comp}\approx g_0$. Por otro lado, si la concentración del represor es alta o del activador baja respecto a la del represor $G\left(A,B\right)_{A/R}^{comp}\approx g_B$. En el caso en que la fuga sea despreciable, es decir $g_B\approx 0$, la tasa de producción se reduce a $G\left(A,B\right)_{A/R}^{comp}=g_0\frac{1+\left(A/A_0\right)^{n_A}}{1+\left(A/A_0\right)^{n_A}+\left(B/B_0\right)^{n_B}}$.

En este enfoque, los factores de transcripción compiten por el mismo sitio de unión, lo cual ocurre biológicamente cuando las secuencias del promotor a las que se unen los distintos factores se superponen. La expresión general de la tasa de producción de ARNm regulado por dos factores de transcripción que compiten está dada por la ec. C.26. También se analizan casos particulares que frecuentemente se utilizan en los trabajos, descritos por las ecs. C.27, C.28 y C.29 como lo hacen Tian *et al.* en un modelo del circuito génico de la transición epitelio-mesenquimal [203] .

En la Figura C.3 a modo ilustrativo del efecto en la selección del tipo de regulación la tasa de producción de ARNm para los distintos tipos de regulación por dos factores de transcripción, considerando ambos activadores, ambos inhibidores y uno de cada tipo.

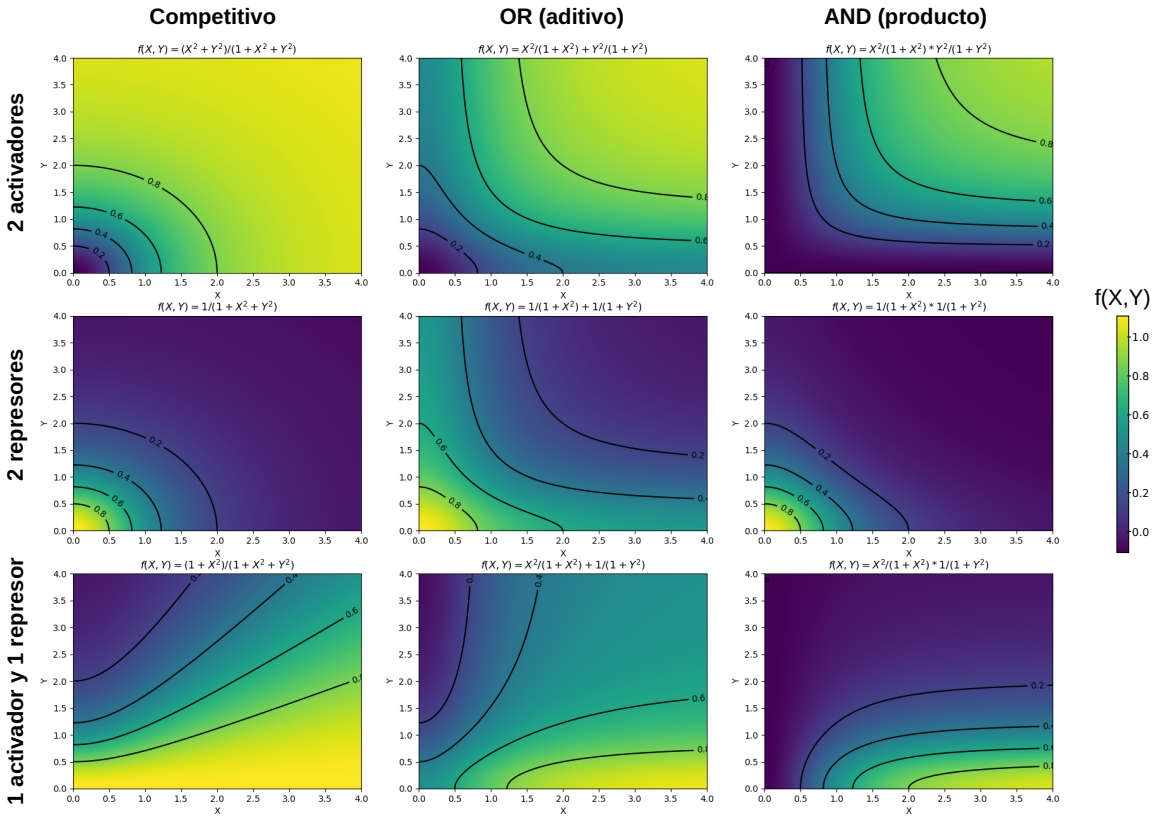

Figura C.3.: Regulación por dos factores de transcripción para distinta combinación de reguladores (activador/represor) y distintas formas de acción de la regulación conjunta. La escala está normalizada entre 0 y la máxima tasa de producción de ARNm y las variables se encuentran adimensionalizadas.

C.4. Regulación por múltiples factores de transcripción

Las expresiones derivadas en la sección anterior para la regulación por dos factores de transcripción pueden extenderse al caso de múltiples factores de transcripción. Así, la función de regulación generalizada de la ec. C.20 para N factores de transcripción es:

$$G^g = g_0 \prod_i^N H^g\left(X_i, \lambda_i\right), \tag{C.30}$$

Donde g_0 es la producción del ARNm cuando el promotor está libre, $H^g\left(X_i, \lambda_{r_i}\right)$ son las funciones de *Hill* generalizadas para los reguladores i, y el parámetro λ_i determina determina si el regulador actúa como activador ($\lambda_i > 1$) o represor ($0 \leq \lambda_i < 1$).

Si la regulación se activa exclusivamente cuando los activadores están unidos al promotor y los inhibidores no lo están (*AND*), la tasa de producción de ARNm se describe como el producto de las funciones de *Hill* de los reguladores por separado:

$$G^{AND} = \left(g_{máx} - g_{min}\right) \prod_{i}^{N} H\left(X_i\right) + g_{min}, \tag{C.31}$$

Al factor $g_{máx} - g_{min}$ se lo suele denotar con el símbolo β y se lo suele denotar como la tasa máxima de transcripción, aunque técnicamente no es la máxima ya que puede haber una tasa basal de transcripción. Respecto a la tasa de transcripción basal, aquí denotada como g_{min}, en los siguientes capítulos la denotaremos mediante el símbolo k^0.

En el caso particular de que la acción conjunta de los reguladores sea aditiva, sin interacción sinérgica, la tasa de producción se expresa como la suma de las funciones de *Hill* de cada regulador por separado:

$$G^{OR} = \left(g_{máx} - g_{min}\right) \sum_{i}^{N} H\left(X_i\right) + g_{min}, \tag{C.32}$$

donde $g_{máx}$ es la tasa de producción máxima, alcanzada cuando los activadores están unidos y los represores no, y g_{min} cuando los activadores no están unidos y los represores sí. Las funciones de *Hill*, $H\left(X_i\right)$ son las funciones de *Hill* para represores o inhibidores dependiendo de la naturaleza de la regulación del factor de transcripción X_i.

Otra posibilidad es considerar que varios factores de transcripción compiten por el mismo sitio de unión. La ec. C.26 puede generalizarse para describir la competencia entre N factores de transcripción:

$$G^{comp} = g_0 \frac{1 + \sum_N^i \lambda_i \left(X_i / X_i^0\right)^{n_X}}{1 + \sum_N^{X_i} \left(X_i / X_i^0\right)^{n_X}}, \tag{C.33}$$

donde $\lambda_i = g_i / g_0$ es la tasa de transcripción cuando el regulador i está unido al promotor respecto a la tasa de transcripción cuando el promotor se encuentra libre. X_i es la concentración del regulador, X_i^0 es la constante de equilibrio y g_0 es la tasa de transcripción del promotor libre.

Apéndice D: Tablas de la Parte II

D

En este apéndice se encuentran las tablas que contienen los valores de los parámetros y condiciones iniciales utilizados en los modelos de redes de regulación génica de la Parte II.

Parámetro	**Descripción**	**Valor**
T^0	Concentración exógena de T	$0-5\mu M$
$k^0_{\mu 200}$	Producción basal de $\mu 200$	$0{,}0002\mu M\, h^{-1}$
$\beta_{S,Z\to\mu 200}$	Tasa de producción máxima de $\mu 200$ dependiente de S y Z	$0{,}012\mu M\, h^{-1}$
$K_{S\to\mu 200}$	Constante de saturación media de $\mu 200$ dependiente de S	$5\mu M$
$K_{Z\to\mu 200}$	Constante de saturación media de $\mu 200$ dependiente de Z	$0{,}2\mu M$
$d_{\mu 200}$	Tasa de degradación de $\mu 200$	$0{,}035h^{-1}$
k^0_{mz}	Tasa de transcripción basal de mz	$0{,}0003\mu M\, h^{-1}$
β_{mz}	Tasa de transcripción máxima de mz	$0{,}06\mu M\, h^{-1}$
$K_{S\to mz}$	Constante de saturación media de mz dependiente de S	$3{,}5\mu M$
d_{mz}	Tasa de degradación de mz	$0{,}09h^{-1}$
β_Z	Tasa de traducción de Z	$17\mu M\, h^{-1}$
$K_{\mu 200\to Z}$	Constante de saturación media de Z dependiente de $\mu 200$	$0{,}06\mu M$
d_Z	Tasa de degradación de Z	$1{,}66h^{-1}$
$k^0_{\mu 34}$	Producción basal de $\mu 34$	$0{,}0012\mu M\, h^{-1}$
$\beta_{S,Z\to\mu 34}$	Tasa de producción máxima de $\mu 34$ dependiente de S y Z	$0{,}0012\mu M\, h^{-1}$
$K_{S\to\mu 34}$	Constante de saturación media de $\mu 34$ dependiente de S	$0{,}15\mu M$
$K_{Z\to\mu 34}$	Constante de saturación media de $\mu 34$ dependiente de Z	$0{,}36\mu M$
$d_{\mu 34}$	Tasa de degradación de $\mu 34$	$0{,}035h^{-1}$
k^0_{ms}	Producción basal de ms	$0{,}0006\mu M\, h^{-1}$
β_{ms}	Tasa de producción máxima de ms	$0{,}03\mu M\, h^{-1}$
$K_{T\to ms}$	Constante de saturación media de ms dependiente de T	$1{,}6\mu M$
d_{ms}	Tasa de degradación de ms	$0{,}09h^{-1}$
β_S	Tasa de traducción de S	$17\mu M\, h^{-1}$
$K_{\mu 34\to S}$	Constante de saturación media de S dependiente de $\mu 34$	$0{,}08\mu M$
d_S	Tasa de degradación de S	$1{,}66h^{-1}$
k^0_{mt}	Producción basal de mt	$0{,}07\mu M\, h^{-1}$
d_{mt}	Tasa de degradación de mt	$0{,}09h^{-1}$
β_T	Tasa de traducción de T	$3\mu M\, h^{-1}$
$K_{\mu 200\to T}$	Constante de saturación media de T dependiente de $\mu 200$	$0{,}06\mu M$
d_T	Tasa de degradación de T	$1{,}1h^{-1}$
$\beta_{S\to E}$	Tasa de producción máxima de $ECad$ dependiente de S	$1\mu M\, h^{-1}$
$K_{S\to E}$	Constante de saturación media de $ECad$ dependiente de S	$0{,}2\mu M$
$\beta_{Z\to E}$	Tasa de producción máxima de $ECad$ dependiente de Z	$0{,}6\mu M\, h^{-1}$
$K_{Z\to E}$	Constante de saturación media de $ECad$ dependiente de Z	$0{,}5\mu M$

Continúa en la siguiente página

Parámetro	Descripción	Valor
d_E	Tasa de degradación de $ECad$	$0{,}5h^{-1}$
$\beta_{S\to N}$	Tasa de producción máxima de $NCad$ dependiente de S	$1\mu M\, h^{-1}$
$K_{S\to N}$	Constante de saturación media de $NCad$ dependiente de S	$0{,}2\mu M$
$\beta_{Z\to N}$	Tasa de producción máxima de $NCad$ dependiente de Z	$0{,}6\mu M\, h^{-1}$
$K_{Z\to N}$	Constante de saturación media de $NCad$ dependiente de Z	$0{,}5\mu M$
d_N	Tasa de degradación de $NCad$	$0{,}5h^{-1}$
$n_{S\to\mu 200}, n_{Z\to\mu 200}$, $n_{S\to mz}, n_{\mu 200\to Z}$, $n_{S\to\mu 34}, n_{Z\to\mu 34}$, $n_{T\to ms}, n_{\mu 34\to S}$, $n_{\mu 200\to T}, n_{S\to E}$, $n_{Z\to E}, n_{S\to N}$, $n_{Z\to N}$	Coeficientes de Hill	2

Tabla D.1.: Parámetros utilizados en el modelo de transición epitelio-mesenquimal (ecs. 7.5 y Figura 7.5).

Variable	$CI\,1(\mu M)$	$CI\,2(\mu M)$
$[\mu 200]$	0,35	0,1
$[mz]$	0,03	0,1
$[Z]$	0,01	0,1
$[\mu 34]$	0,38	0,1
$[ms]$	0,01	0,1
$[S]$	0,01	0,1
$[mt]$	0,07	0,1
$[T]$	0,16	0,16
$[ECad]$	3,3	3
$[NCad]$	0	0

Tabla D.2.: Condiciones iniciales de las variables utilizadas en la Figura 7.5.

Parámetro	Descripción	Valor
k^0_{msox}	Producción basal de $msox$	$0,003\mu M h^{-1}$
β_{msox}	Tasa de transcripción máxima de $msox$	$0,03\mu M h^{-1}$
$K_{DIM\to msox}$	Constante de saturación media de $msox$ dependiente de DIM	$1\mu M$
$n_{DIM\to msox}$	Coeficiente de Hill de $msox$ dependiente de DIM	1
d_{msox}	Tasa de degradación de $msox$	$0,09h^{-1}$
β_{SOX}	Tasa de traducción de SOX	$10\mu M h^{-1}$
k_{on}	Tasa de dimerización de DIM	$0,5\mu M h^{-1}$
k_{off}	Tasa de de-dimerización de DIM	$0,02h^{-1}$
d_{SOX}	Tasa de dedegradación de SOX	$1h^{-1}$
d_{DIM}	Tasa de degradación de DIM	$1h^{-1}$
k^0_{mn}	Producción basal de mn	$0,003\mu M h^{-1}$
$\beta_{DIM\to mn}$	Tasa de transcripción máxima de mn dependiente de DIM	$0,03\mu M h^{-1}$
$K_{DIM\to mn}$	Constante de saturación media de mn dependiente de DIM	$1\mu M$
$n_{DIM\to mn}$	Coeficiente de Hill de mn dependiente de DIM	2
$\beta_{N\to mn}$	Tasa de transcripción máxima de mn dependiente de N	$0,06\mu M h^{-1}$
$K_{N\to mn}$	Constante de saturación media de mn dependiente de N	$3\mu M$
$n_{N\to mn}$	Coeficiente de Hill de mn dependiente de N	2
d_{mn}	Tasa de degradación de mn	$0,09h^{-1}$
β_N	Tasa de traducción de N	$17\mu M h^{-1}$
d_N	Tasa de degradación de N	$2h^{-1}$

Tabla D.3.: Parámetros utilizados en el modelo de pluripotencia (ecs. 7.7 y Figura 7.9).

Variable	$CI\,1(\mu M)$	$CI\,2(\mu M)$
$[msox]$	0,01	0,5
$[SOX]$	0,01	0,5
$[DIM]$	0,01	0,5
$[mn]$	0,01	0,5
$[N]$	0,01	0,5

Tabla D.4.: Condiciones iniciales de las variables utilizadas en el modelo de pluripotencia en la Figura 7.9.

Parámetro	Descripción	Valor
$K_{DIM\to msox}$	Constante de saturación media de $msox$ dependiente de DIM	$0,2\mu M$
$K_{DIM\to mn}$	Constante de saturación media de mn dependiente de DIM	$0,1\mu M$
$K_{\mu 200\to SOX}$	Constante de saturación media de SOX dependiente de $\mu 200$	$0,2\mu M$
$n_{\mu 200\to SOX}$	Coeficiente de Hill de SOX dependiente de $\mu 200$	2
$\beta_{SOX\to \mu 200}$	Tasa de producción máxima de $\mu 200$ dependiente de SOX	$0{,}001\mu M\,h^{-1}$
$K_{SOX\to \mu 200}$	Constante de saturación media de $\mu 200$ dependiente de SOX	$0,4\mu M$
$n_{SOX\to \mu 200}$	Coeficiente de Hill de $\mu 200$ dependiente de SOX	2

Tabla D.5.: Parámetros utilizados en el modelo acoplado.

Variable	$CI\,1(\mu M)$	$CI\,2(\mu M)$	$CI\,3(\mu M)$
$[\mu 200]$	0,1	0,35	0
$[mz]$	0,1	0,03	0,28
$[Z]$	0,1	0,01	2
$[\mu 34]$	0,1	0,38	0
$[ms]$	0,1	0,01	0,3
$[S]$	0,1	0,01	2,5
$[mt]$	0,1	0,07	0,8
$[T]$	0,16	0,16	0,16
$[ECad]$	1	3,3	0
$[NCad]$	1	0	1
$[moct]$	0,1	0,2	0,3
$[SOX]$	0,2	0,2	1
$[DIM]$	0,1	0,2	0,4
$[mn]$	0,7	0,2	0,5
$[N]$	0,5	0,2	1

Tabla D.6.: Condiciones iniciales de las variables utilizadas en el modelo acoplado en las Figuras 8.5 y B.2.

Trabajos publicados

En el marco de esta tesis doctoral se realizaron las siguientes publicaciones:

- Senra, D., Guisoni, N., & Diambra, L. (2022). *ORIGINS: A protein network-based approach to quantify cell pluripotency from scRNA-seq data*. MethodsX, 9, 101778. https://doi.org/10.1016/j.mex.2022.101778 (**Capítulo 3**).
- Senra, D., Guisoni, N., & Diambra, L. (2023). *Cell annotation using scRNA-seq data: A protein-protein interaction network approach.* MethodsX, 10, 102179. https://doi.org/10.1016/j.mex.2023.102179 (**Capítulo 4**).
- Senra, D., Guisoni, N., & Diambra, L. (2025). *Unraveling Tumor Heterogeneity: Quantitative Insights from Single-cell RNA Sequencing Analysis in Breast Cancer Subtypes*. Gene Expression. https://doi.org/10.14218/GE.2024.00071 (**Capítulo 5**).
- Senra, D., Diambra, L. & Guisoni, N.: *Mathematical Modeling of Gene Regulatory Networks in Epithelial-Mesenchymal Transition and Pluripotency*, en preparación (**Parte II**).
- Adicionalmente, durante el primer año del doctorado avancé en la investigación iniciada en la tesina de grado (*Dinámica de filopodios en células tumorales en cultivo: análisis de imágenes y desarrollo de modelos computacionales* bajo la dirección de Nara Guisoni y Luciana Bruno). Este trabajo dio lugar a la publicación del artículo: Senra, D., Páez, A., Gueron, G., Bruno, L., & Guisoni, N. (2020). *Following the footprints of variability during filopodial growth*. European Biophysics Journal, 49(7), 643–659. https://doi.org/10.1007/s00249-020-01473-6.

Bibliografía

[1]A. C. Rios, N. Y. Fu, G. J. Lindeman y J. E. Visvader, "In situ identification of bipotent stem cells in the mammary gland," *Nature*, vol. 506, n.º 7488, págs. 322-327, 2014.

[2]K. A. Gieniec y F. M. Davis, "Mammary basal cells: Stars of the show," *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, vol. 1869, n.º 1, pág. 119 159, 2022.

[3]M.-L. Asselin-Labat y col., "Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation," *Nature cell biology*, vol. 9, n.º 2, págs. 201-209, 2007.

[4]W. Song y col., "Hormones induce the formation of luminal-derived basal cells in the mammary gland," *Cell Research*, vol. 29, n.º 3, págs. 206-220, 2019.

[5]J. L. Inman, C. Robertson, J. D. Mott y M. J. Bissell, "Mammary gland development: cell fate specification, stem cells and the microenvironment," *Development*, vol. 142, n.º 6, págs. 1028-1042, 2015.

[6]N. Harbeck y col., "Breast cancer," *Nature reviews Disease primers*, vol. 5, n.º 1, pág. 66, 2019.

[7]H. Sung y col., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, n.º 3, págs. 209-249, 2021.

[8]M. Arnold y col., "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, págs. 15-23, 2022.

[9]Q. H. Nguyen y col., "Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity," *Nature communications*, vol. 9, n.º 1, págs. 1-12, 2018.

[10]A. G. Waks y E. P. Winer, "Breast cancer treatment: a review," *Jama*, vol. 321, n.º 3, págs. 288-300, 2019.

[11]E. Orrantia-Borunda, P. Anchondo-Nuñez, L. E. Acuña-Aguilar, F. O. Gómez-Valles y C. A. Ramírez-Valdespino, "Subtypes of breast cancer," *Breast Cancer [Internet]*, 2022.

[12]N. Eliyatkin, E. Yalcin, B. Zengel, S. Aktaş y E. Vardar, "Molecular classification of breast carcinoma: from traditional, old-fashioned way to a new age, and a new way," *Journal of breast health*, vol. 11, n.º 2, 2015.

[13]S. Masood, "Breast cancer subtypes: morphologic and biologic characterization," *Women's Health*, vol. 12, n.º 1, págs. 103-119, 2016.

[14]B. Tran y P. L. Bedard, "Luminal-B breast cancer and novel therapeutic targets," *Breast Cancer Research*, vol. 13, n.º 6, pág. 221, 2011.

[15]T. Reya, S. J. Morrison, M. F. Clarke e I. L. Weissman, "Stem cells, cancer, and cancer stem cells," *nature*, vol. 414, n.º 6859, págs. 105-111, 2001.

[16]L. Yang y col., "Targeting cancer stem cell pathways for cancer therapy," *Signal transduction and targeted therapy*, vol. 5, n.º 1, pág. 8, 2020.

[17]X. Chu y col., "Cancer stem cells: advances in knowledge and implications for cancer therapy," *Signal Transduction and Targeted Therapy*, vol. 9, n.º 1, pág. 170, 2024.

[18]U. Kapoor-Narula y N. Lenka, "Cancer stem cells and tumor heterogeneity: Deciphering the role in tumor progression and metastasis," *Cytokine*, vol. 157, pág. 155 968, 2022.

[19]M. Barberis, "Breast Cancer Heterogeneity," *Diagnostics*, vol. 11, n.º 9, pág. 1555, 2021.

[20]G. Turashvili y E. Brogi, "Tumor heterogeneity in breast cancer," *Frontiers in medicine*, vol. 4, pág. 227, 2017.

[21]F. Tang y col., "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature methods*, vol. 6, n.º 5, págs. 377-382, 2009.

[22]B. Hwang, J. H. Lee y D. Bang, "Single-cell RNA sequencing technologies and bioinformatics pipelines," *Experimental & molecular medicine*, vol. 50, n.º 8, págs. 1-14, 2018.

[23]B.-S. Jang, W. Han e I. A. Kim, "Tumor mutation burden, immune checkpoint crosstalk and radiosensitivity in single-cell RNA sequencing data of breast cancer," *Radiotherapy and Oncology*, vol. 142, págs. 202-209, 2020.

[24]Y. Lu y col., "Complement signals determine opposite effects of B cells in chemotherapy-induced immunity," *Cell*, vol. 180, n.º 6, págs. 1081-1097, 2020.

[25]S. Ding, X. Chen y K. Shen, "Single-cell RNA sequencing in breast cancer: Understanding tumor heterogeneity and paving roads to individualized therapy," *Cancer Communications*, vol. 40, n.º 8, págs. 329-344, 2020.

[26]T. K. Olsen y N. Baryawno, "Introduction to single-cell RNA sequencing," *Current protocols in molecular biology*, vol. 122, n.º 1, e57, 2018.

[27]R. A. Amezquita y col., "Orchestrating single-cell analysis with Bioconductor," *Nature methods*, vol. 17, n.º 2, págs. 137-145, 2020.

[28]L. Yu, Y. Cao, J. Y. Yang y P. Yang, "Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data," *Genome biology*, vol. 23, n.º 1, pág. 49, 2022.

[29]H. C. Nguyen, B. Baik, S. Yoon, T. Park y D. Nam, "Benchmarking integration of single-cell differential expression," *Nature Communications*, vol. 14, n.º 1, pág. 1570, 2023.

[30]W. Saelens, R. Cannoodt, H. Todorov e Y. Saeys, "A comparison of single-cell trajectory inference methods," *Nature biotechnology*, vol. 37, n.º 5, págs. 547-554, 2019.

[31]M. D. Luecken y col., "Benchmarking atlas-level data integration in single-cell genomics," *Nature methods*, vol. 19, n.º 1, págs. 41-50, 2022.

[32]Y. Ryu, G. H. Han, E. Jung y D. Hwang, "Integration of single-cell RNA-seq datasets: a review of computational methods," *Molecules and cells*, vol. 46, n.º 2, págs. 106-119, 2023.

[33]M. Setty, V. Kiseliovas, J. Levine, A. Gayoso, L. Mazutis y D. Pe'er, "Characterization of cell fate probabilities in single-cell data with Palantir," *Nature biotechnology*, vol. 37, n.º 4, págs. 451-460, 2019.

[34]H. Chen y col., "Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM," *Nature communications*, vol. 10, n.º 1, págs. 1-14, 2019.

[35]S. V. Stassen, G. G. Yip, K. K. Wong, J. W. Ho y K. K. Tsia, "Generalized and scalable trajectory inference in single-cell omics data with VIA," *Nature communications*, vol. 12, n.º 1, págs. 1-18, 2021.

[36]F. A. Wolf y col., "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells," *Genome biology*, vol. 20, n.º 1, págs. 1-9, 2019.

[37]J. Cao y col., "The single-cell transcriptional landscape of mammalian organogenesis," *Nature*, vol. 566, n.º 7745, págs. 496-502, 2019.

[38]C. R. Banerji y col., "Cellular network entropy as the energy potential in Waddington's differentiation landscape," *Scientific reports*, vol. 3, n.º 1, págs. 1-7, 2013.

[39]A. E. Teschendorff, P. Sollich y R. Kuehn, "Signalling entropy: A novel network-theoretical framework for systems analysis and interpretation of functional omic data," *Methods*, vol. 67, n.º 3, págs. 282-293, 2014.

[40]A. E. Teschendorff y T. Enver, "Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome," *Nature communications*, vol. 8, n.º 1, págs. 1-15, 2017.

[41]W. Chen, S. J. Morabito, K. Kessenbrock, T. Enver, K. B. Meyer y A. E. Teschendorff, "Single-cell landscape in mammary epithelium reveals bipotent-like cells associated with breast cancer risk and outcome," *Communications biology*, vol. 2, n.º 1, págs. 1-13, 2019.

[42]D. Grün y col., "De novo prediction of stem cell identity using single-cell transcriptome data," *Cell stem cell*, vol. 19, n.º 2, págs. 266-277, 2016.

[43]M. Guo, E. L. Bao, M. Wagner, J. A. Whitsett e Y. Xu, "SLICE: determining cell differentiation and lineage based on single cell entropy," *Nucleic acids research*, vol. 45, n.º 7, e54-e54, 2017.

[44]N. P. Palmer, P. R. Schmid, B. Berger e I. S. Kohane, "A gene expression profile of stem cell pluripotentiality and differentiation is conserved across diverse solid and hematopoietic cancers," *Genome biology*, vol. 13, n.º 8, págs. 1-13, 2012.

[45]G. S. Gulati y col., "Single-cell transcriptional diversity is a hallmark of developmental potential," *Science*, vol. 367, n.º 6476, págs. 405-411, 2020.

[46]A.-L. Barabasi y Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature reviews genetics*, vol. 5, n.º 2, págs. 101-113, 2004.

[47]T. G. O. Consortium, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Research*, vol. 47, n.º D1, págs. D330-D338, 2018, ISSN: 0305-1048. DOI: 10.1093/nar/gky1055. dirección: https://doi.org/10.1093/nar/gky1055.

[48]*Gene Ontology Data Archive*, https://release.geneontology.org/.

[49]I. Rodchenkov y col., "Pathway Commons 2019 Update: integration, analysis and exploration of pathway data," *Nucleic Acids Research*, vol. 48, n.º D1, págs. D489-D497, 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gkz946. dirección: https://doi.org/10.1093/nar/gkz946.

[50]Y. Wang y col., "Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine," *Journal of Experimental Medicine*, vol. 217, n.º 2, 2020.

[51]D. Pellin y col., "A comprehensive single cell transcriptional landscape of human hematopoietic progenitors," *Nature communications*, vol. 10, n.º 1, págs. 1-15, 2019.

[52]A. Wang y col., "Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes," *Elife*, vol. 9, e62522, 2020.

[53]S. Umar, "Intestinal stem cells," *Current gastroenterology reports*, vol. 12, n.º 5, págs. 340-348, 2010.

[54]S. Cui y P.-Y. Chang, "Current understanding concerning intestinal stem cells," *World journal of gastroenterology*, vol. 22, n.º 31, pág. 7099, 2016.

[55]N. Barker, "Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration," *Nature reviews Molecular cell biology*, vol. 15, n.º 1, págs. 19-33, 2014.

[56]A. Birbrair y P. S. Frenette, "Niche heterogeneity in the bone marrow," *Annals of the new York Academy of Sciences*, vol. 1370, n.º 1, págs. 82-96, 2016.

[57]S. J. Szilvassy, "The biology of hematopoietic stem cells," *Archives of medical research*, vol. 34, n.º 6, págs. 446-460, 2003.

[58]M. Zhao y L. Li, "Regulation of hematopoietic stem cells in the niche," *Science China Life Sciences*, vol. 58, n.º 12, págs. 1209-1215, 2015.

[59]C. Baumgartner y col., "An ERK-dependent feedback mechanism prevents hematopoietic stem cell exhaustion," *Cell stem cell*, vol. 22, n.º 6, págs. 879-892, 2018.

[60]C. Dussiau y col., "Hematopoietic differentiation is characterized by a transient peak of entropy at a single-cell level," *BMC biology*, vol. 20, n.º 1, págs. 1-15, 2022.

[61]K. Wiesner, J. Teles, M. Hartnor y C. Peterson, "Haematopoietic stem cells: entropic landscapes of differentiation," *Interface focus*, vol. 8, n.º 6, pág. 20 180 040, 2018.

[62]A. Ciechanowicz, "Stem cells in lungs," *Stem Cells*, págs. 261-274, 2019.

[63]S. D. Reynolds y A. M. Malkinson, "Clara cell: progenitor for the bronchiolar epithelium," *The international journal of biochemistry & cell biology*, vol. 42, n.º 1, págs. 1-4, 2010.

[64]E. L. Rawlins y col., "The role of Scgb1a1+ Clara cells in the long-term maintenance and repair of lung airway, but not alveolar, epithelium," *Cell stem cell*, vol. 4, n.º 6, págs. 525-534, 2009.

[65]D. J. Weiss y col., "Stem cells and cell therapies in lung biology and lung diseases," *Proceedings of the American thoracic society*, vol. 8, n.º 3, págs. 223-272, 2011.

[66]R. G. Crystal, "Airway basal cells. The "smoking gun" of chronic obstructive pulmonary disease," *American journal of respiratory and critical care medicine*, vol. 190, n.º 12, págs. 1355-1362, 2014.

[67]K. Street y col., "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC genomics*, vol. 19, n.º 1, págs. 1-16, 2018.

[68]K. Van den Berge y col., "Trajectory-based differential expression analysis for single-cell sequencing data," *Nature communications*, vol. 11, n.º 1, págs. 1-13, 2020.

[69]M. D. Luecken y F. J. Theis, "Current best practices in single-cell RNA-seq analysis: a tutorial," *Molecular systems biology*, vol. 15, n.º 6, e8746, 2019.

[70]T. Stuart y col., "Comprehensive integration of single-cell data," *Cell*, vol. 177, n.º 7, págs. 1888-1902, 2019.

[71]M. K. Jaakkola, F. Seyednasrollah, A. Mehmood y L. L. Elo, "Comparison of methods to detect differentially expressed genes between single-cell populations," *Briefings in bioinformatics*, vol. 18, n.º 5, págs. 735-743, 2017.

[72]L. Diambra, A. M. Alonso, S. Sookoian y C. J. Pirola, "Single cell gene expression profiling of nasal ciliated cells reveals distinctive biological processes related to epigenetic mechanisms in patients with severe COVID-19," *Computers in Biology and Medicine*, vol. 148, pág. 105 895, 2022.

[73]D. Senra, N. Guisoni y L. Diambra, "ORIGINS: A protein network-based approach to quantify cell pluripotency from scRNA-seq data," *MethodsX*, vol. 9, pág. 101 778, 2022.

[74]D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'donovan y R. Apweiler, "QuickGO: a web-based tool for Gene Ontology searching," *Bioinformatics*, vol. 25, n.º 22, págs. 3045-3046, 2009.

[75]B. Pal y col., "A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast," *The EMBO journal*, vol. 40, n.º 11, e107333, 2021.

[76]I. Tirosh y col., "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq," *Science*, vol. 352, n.º 6282, págs. 189-196, 2016.

[77]S. Aibar y col., "SCENIC: single-cell regulatory network inference and clustering," *Nature methods*, vol. 14, n.º 11, págs. 1083-1086, 2017.

[78]S. Z. Wu y col., "A single-cell and spatially resolved atlas of human breast cancers," *Nature genetics*, vol. 53, n.º 9, págs. 1334-1347, 2021.

[79]M. Karaayvaz y col., "Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq," *Nature communications*, vol. 9, n.º 1, pág. 3588, 2018.

[80]W. Chung y col., "Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer," *Nature communications*, vol. 8, n.º 1, pág. 15 081, 2017.

[81]L. Ren y col., "Single cell RNA sequencing for breast cancer: present and future," *Cell Death Discovery*, vol. 7, n.º 1, pág. 104, 2021.

[82]A. Dave y col., "The Breast Cancer Single-Cell Atlas: Defining cellular heterogeneity within model cell lines and primary tumors to inform disease subtype, stemness, and treatment options," *Cellular Oncology*, vol. 46, n.º 3, págs. 603-628, 2023.

[83]A. Marusyk, M. Janiszewska y K. Polyak, "Intratumor heterogeneity: the rosetta stone of therapy resistance," *Cancer cell*, vol. 37, n.º 4, págs. 471-484, 2020.

[84]J. Liu, H. Dang y X. W. Wang, "The significance of intertumor and intratumor heterogeneity in liver cancer," *Experimental & molecular medicine*, vol. 50, n.º 1, e416-e416, 2018.

[85]S. Ramón y Cajal y col., "Clinical implications of intratumor heterogeneity: challenges and opportunities," *Journal of Molecular Medicine*, vol. 98, págs. 161-177, 2020.

[86]J. Xu y col., "Delving into the heterogeneity of different breast cancer subtypes and the prognostic models utilizing scRNA-Seq and Bulk RNA-Seq," *International Journal of Molecular Sciences*, vol. 23, n.º 17, pág. 9936, 2022.

[87]L. Ma y col., "Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer," *Cancer cell*, vol. 36, n.º 4, págs. 418-430, 2019.

[88]L. Harbers, F. Agostini, M. Nicos, D. Poddighe, M. Bienko y N. Crosetto, "Somatic copy number alterations in human cancers: an analysis of publicly available data from the cancer genome atlas," *Frontiers in oncology*, vol. 11, pág. 700 568, 2021.

[89]Y. Zhang y col., "Copy number alterations that predict metastatic capability of human breast cancer," *Cancer research*, vol. 69, n.º 9, págs. 3795-3801, 2009.

[90]W. Han y col., "DNA copy number alterations and expression of relevant genes in triple-negative breast cancer," *Genes, Chromosomes and Cancer*, vol. 47, n.º 6, págs. 490-499, 2008.

[91]J. Peng y col., "Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma," *Cell research*, vol. 29, n.º 9, págs. 725-738, 2019.

[92]C. Neftel y col., "An integrative model of cellular states, plasticity, and genetics for glioblastoma," *Cell*, vol. 178, n.º 4, págs. 835-849, 2019.

[93]O. Gandrillon y col., "Entropy as a measure of variability and stemness in single-cell transcriptomics," *Current Opinion in Systems Biology*, vol. 27, pág. 100 348, 2021.

[94]C. R. Banerji, S. Severini, C. Caldas y A. E. Teschendorff, "Intra-tumour signalling entropy determines clinical outcome in breast and lung cancer," *PLoS computational biology*, vol. 11, n.º 3, e1004115, 2015.

[95]D. DeTomaso, M. G. Jones, M. Subramaniam, T. Ashuach, C. J. Ye y N. Yosef, "Functional interpretation of single cell similarity maps," *Nature communications*, vol. 10, n.º 1, pág. 4376, 2019.

[96]M. G. Davey, S. O. Hynes, M. J. Kerin, N. Miller y A. J. Lowery, "Ki-67 as a prognostic biomarker in invasive breast cancer," *Cancers*, vol. 13, n.º 17, pág. 4455, 2021.

[97]T. O. Nielsen y col., "Assessment of Ki67 in breast cancer: updated recommendations from the international Ki67 in breast cancer working group," *JNCI: Journal of the National Cancer Institute*, vol. 113, n.º 7, págs. 808-819, 2021.

[98]J. Felipe Lima, S. Nofech-Mozes, J. Bayani y J. M. Bartlett, "EMT in breast carcinoma—a review," *Journal of clinical medicine*, vol. 5, n.º 7, pág. 65, 2016.

[99]Y. Wang y B. P. Zhou, "Epithelial-mesenchymal transition in breast cancer progression and metastasis," *Chinese journal of cancer*, vol. 30, n.º 9, pág. 603, 2011.

[100]Y. Chen, B. Pal, G. J. Lindeman, J. E. Visvader y G. K. Smyth, "R code and downstream analysis objects for the scRNA-seq atlas of normal and tumorigenic human breast tissue," *Scientific Data*, vol. 9, n.º 1, pág. 96, 2022.

[101]A. P. Patel y col., "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, n.º 6190, págs. 1396-1401, 2014.

[102]A. S. Venteicher y col., "Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq," *Science*, vol. 355, n.º 6332, eaai8478, 2017.

[103]Y. Hao y col., "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, n.º 13, págs. 3573-3587, 2021.

[104]R. Gao y col., "Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes," *Nature biotechnology*, vol. 39, n.º 5, págs. 599-608, 2021.

[105]A. Serin Harmanci, A. O. Harmanci y X. Zhou, "CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data," *Nature communications*, vol. 11, n.º 1, pág. 89, 2020.

[106]Y. H. Choi y J. K. Kim, "Dissecting cellular heterogeneity using single-cell RNA sequencing," *Molecules and cells*, vol. 42, n.º 3, pág. 189, 2019.

[107]X. Duan, W. Wang, M. Tang, F. Gao y X. Lin, "Dissecting Cellular Heterogeneity Based on Network Denoising of scRNA-seq Using Local Scaling Self-Diffusion," *Frontiers in Genetics*, vol. 12, pág. 811 043, 2022.

[108]C. Hafemeister y R. Satija, "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression," *Genome biology*, vol. 20, n.º 1, pág. 296, 2019.

[109]C. Mayer y col., "Developmental diversification of cortical inhibitory interneurons," *Nature*, vol. 555, n.º 7697, págs. 457-462, 2018.

[110]Gene set of GO:0007049 from QuickGO, https://www.ebi.ac.uk/QuickGO/term/GO:0007049, Accessed: 2023-11-28, 2023.

[111]Gene set of GO:0010718 from from MSigDB, https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/GOBP_POSITIVE_REGULATION_OF_EPITHELIAL_TO_MESENCHYMAL_TRANSITION.html, Accessed: 2023-11-28, 2023.

[112]E. Charafe-Jauffret y col., "Gene expression profiling of breast cell lines identifies potential new basal markers," *Oncogene*, vol. 25, n.º 15, págs. 2273-2284, 2006.

[113]Gene set of CHARAFE BREAST CANCER LUMINAL VS BASAL DN from MSigDB, https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN.html, Accessed: 2023-11-28, 2006.

[114]Gene set of CHARAFE BREAST CANCER LUMINAL VS BASAL UP from MSigDB, https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP.html, Accessed: 2023-11-28, 2006.

[115]Gene set of CHARAFE BREAST CANCER LUMINAL VS MESENCHYMAL DN from MSigDB, https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_DN.html, Accessed: 2023-11-28, 2006.

[116]A. Subramanian y col., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, n.º 43, págs. 15 545-15 550, 2005.

[117]A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo y J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, n.º 12, págs. 1739-1740, 2011.

[118]S. Kannan, M. Farid, B. L. Lin, M. Miyamoto y C. Kwon, "Transcriptomic entropy benchmarks stem cell-derived cardiomyocyte maturation against endogenous tissue at single cell level," *PLOS Computational Biology*, vol. 17, n.º 9, págs. 1-21, sep. de 2021.

[119]B. D. MacArthur e I. R. Lemischka, "Statistical mechanics of pluripotency," *Cell*, vol. 154, n.º 3, págs. 484-489, 2013.

[120]M. Krzywinski y N. Altman, "Points of significance: comparing samples–part II," *nature methods*, vol. 11, n.º 4, pág. 355, 2014.

[121]Y. Chen y G. Smyth, *Data, R code and output Seurat Objects for single cell RNA-seq analysis of human breast tissues*, Available at https://figshare.com/articles/dataset/Data_R_code_and_output_Seurat_Objects_for_single_cell_RNA-seq_analysis_of_human_breast_tissues/17058077, "2" de "2022". DOI: "10.6084/m9.figshare.17058077.v1".

[122]D. Senra, N. Guisoni y L. Diambra, "Unraveling Tumor Heterogeneity: Quantitative Insights from Single-cell RNA Sequencing Analysis in Breast Cancer Subtypes," *Gene Expression*, n.º 000, págs. 0–0, 2025.

[123]B. Orsetti y col., "Genetic profiling of chromosome 1 in breast cancer: mapping of regions of gains and losses and identification of candidate genes on 1q," *British journal of cancer*, vol. 95, n.º 10, págs. 1439-1447, 2006.

[124]T. Baslan y col., "Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing," *elife*, vol. 9, e51480, 2020.

[125]L. Zhang, N. Feizi, C. Chi y P. Hu, "Association analysis of somatic copy number alteration burden with breast cancer survival," *Frontiers in Genetics*, vol. 9, pág. 421, 2018.

[126]M. K. Ibragimova, M. M. Tsyganov, A. M. Pevzner y N. V. Litviakov, "Transcriptome of breast tumors with different amplification status of the long arm of chromosome 8," *Anticancer research*, vol. 41, n.º 1, págs. 187-195, 2021.

[127]R. M. Rodrigues-Peres y col., "Copy number alterations associated with clinical features in an underrepresented population with breast cancer," *Molecular Genetics & Genomic Medicine*, vol. 7, n.º 7, e00750, 2019.

[128]X. Dai y col., "Breast cancer intrinsic subtype classification, clinical use and future trends," *American journal of cancer research*, vol. 5, n.º 10, pág. 2929, 2015.

[129]X. Dai, H. Cheng, Z. Bai y J. Li, "Breast cancer cell line classification and its relevance with breast tumor subtyping," *Journal of Cancer*, vol. 8, n.º 16, pág. 3131, 2017.

[130]Y. Herdiana, N. Wathoni, S. Shamsuddin y M. Muchtaridi, "α-Mangostin nanoparticles cytotoxicity and cell death modalities in breast cancer cell lines," *Molecules*, vol. 26, n.º 17, pág. 5119, 2021.

[131]H. Schwarzenbach y P. B. Gahan, "Predictive value of exosomes and their cargo in drug response/resistance of breast cancer patients," *Cancer Drug Resistance*, vol. 3, n.º 1, pág. 63, 2020.

[132]M. Luo y col., "Breast cancer stem cells: current advances and clinical implications," *Mammary Stem Cells: Methods and Protocols*, págs. 1-49, 2015.

[133]J. S. Crabtree y L. Miele, "Breast cancer stem cells," *Biomedicines*, vol. 6, n.º 3, pág. 77, 2018.

[134]X. Zhang, K. Powell y L. Li, "Breast cancer stem cells: biomarkers, identification and isolation methods, regulating mechanisms, cellular origin, and beyond," *Cancers*, vol. 12, n.º 12, pág. 3765, 2020.

[135]S. A. Narod, "Breast cancer in young women," *Nature reviews Clinical oncology*, vol. 9, n.º 8, págs. 460-470, 2012.

[136]N. Klauber-DeMore, "Tumor biology of breast cancer in young women," *Breast disease*, vol. 23, n.º 1, págs. 9-15, 2006.

[137]A. Bharat, R. L. Aft, F. Gao y J. A. Margenthaler, "Patient and tumor characteristics associated with increased mortality in young women ($\leq$ 40 years) with breast cancer," *Journal of surgical oncology*, vol. 100, n.º 3, págs. 248-251, 2009.

[138]A. C. Voogd y col., "Differences in risk factors for local and distant recurrence after breast-conserving therapy or mastectomy for stage I and II breast cancer: pooled results of two large European randomized trials," *Journal of clinical oncology*, vol. 19, n.º 6, págs. 1688-1697, 2001.

[139]Y. Wu, M. Sarkissyan y J. V. Vadgama, "Epithelial-mesenchymal transition and breast cancer," *Journal of clinical medicine*, vol. 5, n.º 2, pág. 13, 2016.

[140]D. Hanahan y R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, n.º 5, págs. 646-674, 2011.

[141]E. Hay, "Organization and fine structure of epithelium and mesenchyme in the developing chick embryo," en *Epithelial-Mesenchymal Interactions; 18th Hahnemann Symposium, 1968*, Williams & Wilkins, 1968.

[142]E. D. Hay, "An overview of epithelio-mesenchymal transformation," *Cells Tissues Organs*, vol. 154, n.º 1, págs. 8-20, 1995.

[143]J. Yang y col., "Guidelines and definitions for research on epithelial–mesenchymal transition," *Nature reviews Molecular cell biology*, vol. 21, n.º 6, págs. 341-352, 2020.

[144]R. Kalluri, R. A. Weinberg y col., "The basics of epithelial-mesenchymal transition," *The Journal of clinical investigation*, vol. 119, n.º 6, págs. 1420-1428, 2009.

[145]A. Dongre y R. A. Weinberg, "New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer," *Nature reviews Molecular cell biology*, vol. 20, n.º 2, págs. 69-84, 2019.

[146]S. Gerstberger, Q. Jiang y K. Ganesh, "Metastasis," *Cell*, vol. 186, n.º 8, págs. 1564-1579, 2023.

[147]J. Fares, M. Y. Fares, H. H. Khachfe, H. A. Salhab e Y. Fares, "Molecular principles of metastasis: a hallmark of cancer revisited," *Signal transduction and targeted therapy*, vol. 5, n.º 1, pág. 28, 2020.

[148]E. W. Thompson y D. F. Newgreen, "Carcinoma invasion and metastasis: a role for epithelial-mesenchymal transition?" *Cancer research*, vol. 65, n.º 14, págs. 5991-5995, 2005.

[149]D. Ribatti, R. Tamma y T. Annese, "Epithelial-mesenchymal transition in cancer: a historical overview," *Translational oncology*, vol. 13, n.º 6, pág. 100 773, 2020.

[150]L. Bornes, G. Belthier y J. van Rheenen, "Epithelial-to-mesenchymal transition in the light of plasticity and hybrid E/M states," *Journal of clinical medicine*, vol. 10, n.º 11, pág. 2403, 2021.

[151]J. Xu, S. Lamouille y R. Derynck, "TGF-β-induced epithelial to mesenchymal transition," *Cell research*, vol. 19, n.º 2, págs. 156-172, 2009.

[152]A. Z. Ayob y T. S. Ramasamy, "Cancer stem cells as key drivers of tumour progression," *Journal of biomedical science*, vol. 25, págs. 1-18, 2018.

[153]F. Rossi, H. Noren, R. Jove, V. Beljanski y K.-H. Grinnemo, "Differences and similarities between cancer and somatic stem cells: therapeutic implications," *Stem cell research & therapy*, vol. 11, n.º 1, pág. 489, 2020.

[154]L. Walcher y col., "Cancer stem cells—origins and biomarkers: perspectives for targeted personalized therapies," *Frontiers in immunology*, vol. 11, pág. 1280, 2020.

[155]G.-Q. Ling, D.-B. Chen, B.-Q. Wang y L.-S. Zhang, "Expression of the pluripotency markers Oct3/4, Nanog and Sox2 in human breast cancer cell lines," *Oncology letters*, vol. 4, n.º 6, págs. 1264-1268, 2012.

[156]L. You, X. Guo e Y. Huang, "Correlation of cancer stem-cell markers OCT4, SOX2, and NANOG with clinicopathological features and prognosis in operative patients with rectal cancer," *Yonsei medical journal*, vol. 59, n.º 1, págs. 35-42, 2018.

[157]G. Basati, H. Mohammadpour y A. Emami Razavi, "Association of high expression levels of SOX2, NANOG, and OCT4 in gastric cancer tumor tissues with progression and poor prognosis," *Journal of gastrointestinal cancer*, vol. 51, págs. 41-47, 2020.

[158]T.-Y. Fu y col., "Association of OCT 4, SOX 2, and NANOG expression with oral squamous cell carcinoma progression," *Journal of Oral Pathology & Medicine*, vol. 45, n.º 2, págs. 89-95, 2016.

[159]B. Bussolati, S. Bruno, C. Grange, U. Ferrando y G. Camussi, "Identification of a tumor-initiating stem cell population in human renal carcinomas," *The FASEB Journal*, vol. 22, n.º 10, págs. 3696-3705, 2008.

[160]S.-H. Chiou y col., "Coexpression of Oct4 and Nanog enhances malignancy in lung adenocarcinoma by inducing cancer stem cell–like properties and epithelial–mesenchymal transdifferentiation," *Cancer research*, vol. 70, n.º 24, págs. 10 433-10 444, 2010.

[161]A. Amsterdam, C. Raanan, L. Schreiber, O. Freyhan, L. Schechtman y D. Givol, "Localization of the stem cell markers LGR5 and Nanog in the normal and the cancerous human ovary and their inter-relationship," *Acta Histochemica*, vol. 115, n.º 4, págs. 330-338, 2013.

[162]Y. Pan, J. Jiao, C. Zhou y col., "Nanog is highly expressed in ovarian serous cystadenocarcinoma and correlated with clinical stage and pathological grade," *Pathobiology*, vol. 77, n.º 6, págs. 283-288, 2010.

[163]U. Ezeh, P. Turek, R. Reijo y A. Clark, "Human embryonic stem cell genes OCT4, NANOG, STELLAR, and GDF3 are expressed in both seminoma and breast carcinoma," *Cancer*, vol. 104, n.º 10, págs. 2255-2265, 2005.

[164]T. Lin, Y. Ding y J. Li, "Overexpression of Nanog protein is associated with poor prognosis in gastric adenocarcinoma," *Medical Oncology*, vol. 29, n.º 2, págs. 878-885, 2012.

[165]J. Wen, J. Park, K. Park y col., "Oct4 and Nanog expression is associated with early stages of pancreatic carcinogenesis," *Pancreas*, vol. 39, n.º 5, págs. 622-626, 2010.

[166]S. Chiou, C. Yu, C. Huang y col., "Positive correlations of Oct-4 and Nanog in oral cancer stem-like cells and high-grade oral squamous cell carcinoma," *Clinical Cancer Research*, vol. 14, n.º 13, págs. 4085-4095, 2008.

[167]Y. Guo, S. Liu, P. Wang y col., "Expression profile of embryonic stem cell-associated genes Oct4, Sox2 and Nanog in human gliomas," *Histopathology*, vol. 59, n.º 4, págs. 763-775, 2011.

[168]A. Gillis, H. Stoop, K. Biermann y col., "Expression and interdependencies of pluripotency factors LIN28, OCT3/4, NANOG and SOX2 in human testicular germ cells and tumours of the testis," *International Journal of Andrology*, vol. 34, n.º 4 Pt 2, e160-e174, 2011.

[169]X. Zhou y col., "Expression of the stem cell marker, Nanog, in human endometrial adenocarcinoma," *International journal of gynecological pathology*, vol. 30, n.º 3, págs. 262-270, 2011.

[170]C. R. Jeter, T. Yang, J. Wang, H.-P. Chao y D. G. Tang, "Concise review: NANOG in cancer stem cells and tumor development: an update and outstanding questions," *Stem cells*, vol. 33, n.º 8, págs. 2381-2390, 2015.

[171]M. Lee, E. J. Nam, S. W. Kim, S. Kim, J. H. Kim e Y. T. Kim, "Prognostic impact of the cancer stem cell–related marker NANOG in ovarian serous carcinoma," *International Journal of Gynecologic Cancer*, vol. 22, n.º 9, 2012.

[172]H.-M. Meng y col., "Over-expression of Nanog predicts tumor progression and poor prognosis in colorectal cancer," *Cancer biology & therapy*, vol. 9, n.º 4, págs. 295-302, 2010.

[173]T. Nagata y col., "Prognostic significance of NANOG and KLF4 for breast cancer," *Breast cancer*, vol. 21, págs. 96-101, 2014.

[174]X. Lu, S. J. Mazur, T. Lin, E. Appella e Y. Xu, "The pluripotency factor nanog promotes breast cancer tumorigenesis and metastasis," *Oncogene*, vol. 33, n.º 20, págs. 2655-2664, 2014.

[175]E. E. Ibrahim y col., "Embryonic NANOG activity defines colorectal cancer stem cells and modulates through AP1-and TCF-dependent mechanisms," *Stem cells*, vol. 30, n.º 10, págs. 2076-2087, 2012.

[176]J. Zhang y col., "NANOG modulates stemness in human colorectal cancer," *Oncogene*, vol. 32, n.º 37, págs. 4397-4405, 2013.

[177]T. K. W. Lee, A. Castilho, V. C. H. Cheung, K. H. Tang, S. Ma e I. O. L. Ng, "CD24+ liver tumor-initiating cells drive self-renewal and tumor initiation through STAT3-mediated NANOG regulation," *Cell stem cell*, vol. 9, n.º 1, págs. 50-63, 2011.

[178]C. R. Jeter y col., "NANOG promotes cancer stem cell characteristics and prostate cancer resistance to androgen deprivation," *Oncogene*, vol. 30, n.º 36, págs. 3833-3845, 2011.

[179]M. Zbinden, A. Duquet, A. Lorente-Trigos, S.-N. Ngwabyt, I. Borges y A. Ruiz i Altaba, "NANOG regulates glioma stem cells and is essential in vivo acting in a cross-functional network with GLI1 and p53," *The EMBO journal*, vol. 29, n.º 15, págs. 2659-2674, 2010.

[180]C.-S. Niu, D.-X. Li, Y.-H. Liu, X.-M. Fu, S.-F. Tang y J. Li, "Expression of NANOG in human gliomas and its relationship with undifferentiated glioma cells," *Oncology reports*, vol. 26, n.º 3, págs. 593-601, 2011.

[181]H. Niwa, J.-i. Miyazaki y A. G. Smith, "Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells," *Nature genetics*, vol. 24, n.º 4, págs. 372-376, 2000.

[182]V. Karwacki-Neisius y col., "Reduced Oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by Oct4 and Nanog," *Cell stem cell*, vol. 12, n.º 5, págs. 531-545, 2013.

[183]T. W. Theunissen, A. L. Van Oosten, G. Castelo-Branco, J. Hall, A. Smith y J. C. Silva, "Nanog overcomes reprogramming barriers and induces pluripotency in minimal conditions," *Current Biology*, vol. 21, n.º 1, págs. 65-71, 2011.

[184]J. Balzeau, M. R. Menezes, S. Cao y J. P. Hagan, "The LIN28/let-7 pathway in cancer," *Frontiers in genetics*, vol. 8, pág. 31, 2017.

[185]S. Roy, R. R. Sunkara, M. Y. Parmar, S. Shaikh y S. K. Waghmare, "EMT imparts cancer stemness and plasticity: new perspectives and therapeutic potential," *Frontiers in Bioscience-Landmark*, vol. 26, n.º 2, págs. 238-265, 2020.

[186]I. Fabregat, A. Malfettone y J. Soukupova, "New insights into the crossroads between EMT and stemness in the context of cancer," *Journal of clinical medicine*, vol. 5, n.º 3, pág. 37, 2016.

[187]G. Babaei, S. G.-G. Aziz y N. Z. Z. Jaghi, "EMT, cancer stem cells and autophagy; The three main axes of metastasis," *Biomedicine & Pharmacotherapy*, vol. 133, pág. 110 909, 2021.

[188]A.-P. Morel, M. Lièvre, C. Thomas, G. Hinkal, S. Ansieau y A. Puisieux, "Generation of breast cancer stem cells through epithelial-mesenchymal transition," *PloS one*, vol. 3, n.º 8, e2888, 2008.

[189]A. D. Rhim y col., "EMT and dissemination precede pancreatic tumor formation," *Cell*, vol. 148, n.º 1, págs. 349-361, 2012.

[190]J. E. Choi y col., "Expression of epithelial-mesenchymal transition and cancer stem cell markers in colorectal adenocarcinoma: Clinicopathological significance," *Oncology reports*, vol. 38, n.º 3, págs. 1695-1705, 2017.

[191]G. V. Vijay y col., "GSK3β regulates epithelial-mesenchymal transition and cancer stem cell properties in triple-negative breast cancer," *Breast Cancer Research*, vol. 21, págs. 1-14, 2019.

[192]W. Zhou y col., "Snail contributes to the maintenance of stem cell-like phenotype cells in human pancreatic cancer," *PloS one*, vol. 9, n.º 1, e87409, 2014.

[193]A. Singh y J. Settleman, "EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer," *Oncogene*, vol. 29, n.º 34, págs. 4741-4751, 2010.

[194]Z. Cai, Y. Cao, Y. Luo, H. Hu y H. Ling, "Signalling mechanism (s) of epithelial–mesenchymal transition and cancer stem cells in tumour therapeutic resistance," *Clinica Chimica Acta*, vol. 483, págs. 156-163, 2018.

[195]M. Singla, A. Kumar, A. Bal, S. Sarkar y S. Bhattacharyya, "Epithelial to mesenchymal transition induces stem cell like phenotype in renal cell carcinoma cells," *Cancer cell international*, vol. 18, págs. 1-13, 2018.

[196]M. Luo, M. Brooks y M. S. Wicha, "Epithelial-mesenchymal plasticity of breast cancer stem cells: implications for metastasis and therapeutic resistance," *Current pharmaceutical design*, vol. 21, n.º 10, págs. 1301-1310, 2015.

[197]G. Cooper, "Regulation of transcription in eukaryotes," *The cell: A molecular approach*, págs. 374-378, 2000.

[198]B. P. Ingalls, *Mathematical modeling in systems biology: an introduction*. MIT press, 2013.

[199]J. O'Brien, H. Hayder, Y. Zayed y C. Peng, "Overview of microRNA biogenesis, mechanisms of actions, and circulation," *Frontiers in endocrinology*, vol. 9, pág. 402, 2018.

[200]U. Alon, *An introduction to systems biology: design principles of biological circuits*. Chapman y Hall/CRC, 2019.

[201]R. C. Lee, R. L. Feinbaum y V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," *cell*, vol. 75, n.º 5, págs. 843-854, 1993.

[202]R. C. Friedman, K. K.-H. Farh, C. B. Burge y D. P. Bartel, "Most mammalian mRNAs are conserved targets of microRNAs," *Genome research*, vol. 19, n.º 1, págs. 92-105, 2009.

[203]X.-J. Tian, H. Zhang y J. Xing, "Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition," *Biophysical journal*, vol. 105, n.º 4, págs. 1079-1089, 2013.

[204]M. Lu, M. K. Jolly, R. Gomoto, B. Huang, J. Onuchic y E. Ben-Jacob, "Tristability in cancer-associated microRNA-TF chimera toggle switch," *The journal of physical chemistry B*, vol. 117, n.º 42, págs. 13 164-13 174, 2013.

[205]M. Lu, M. K. Jolly, H. Levine, J. N. Onuchic y E. Ben-Jacob, "MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination," *Proceedings of the National Academy of Sciences*, vol. 110, n.º 45, págs. 18 144-18 149, 2013.

[206]G. A. Burger, E. H. Danen y J. B. Beltman, "Deciphering epithelial–mesenchymal transition regulatory networks in cancer through computational approaches," *Frontiers in oncology*, vol. 7, pág. 162, 2017.

[207]J. G. T. Zañudo y col., "Towards control of cellular decision-making networks in the epithelial-to-mesenchymal transition," *Physical biology*, vol. 16, n.º 3, pág. 031 002, 2019.

[208]J. Zhang y col., "TGF-β–induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops," *Science signaling*, vol. 7, n.º 345, ra91-ra91, 2014.

[209]H. Ungefroren, "Autocrine TGF-β in cancer: review of the literature and caveats in experimental analysis," *International journal of molecular sciences*, vol. 22, n.º 2, pág. 977, 2021.

[210]H. Siemens y col., "miR-34 and SNAIL form a double-negative feedback loop to regulate epithelial-mesenchymal transitions," *Cell cycle*, vol. 10, n.º 24, págs. 4256-4271, 2011.

[211]N. H. Kim y col., "A p53/miRNA-34 axis regulates Snail1-dependent cancer cell epithelial–mesenchymal transition," *Journal of Cell Biology*, vol. 195, n.º 3, págs. 417-433, 2011.

[212]U. Wellner y col., "The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs," *Nature cell biology*, vol. 11, n.º 12, págs. 1487-1495, 2009.

[213]C. P. Bracken y col., "A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition," *Cancer research*, vol. 68, n.º 19, págs. 7846-7854, 2008.

[214]U. Burk y col., "A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells," *EMBO reports*, vol. 9, n.º 6, págs. 582-589, 2008.

[215]P. A. Gregory y col., "An autocrine TGF-β/ZEB/miR-200 signaling network regulates establishment and maintenance of epithelial-mesenchymal transition," *Molecular biology of the cell*, vol. 22, n.º 10, págs. 1686-1698, 2011.

[216]J. P. Thiery, H. Acloque, R. Y. Huang y M. A. Nieto, "Epithelial-mesenchymal transitions in development and disease," *cell*, vol. 139, n.º 5, págs. 871-890, 2009.

[217]S. Peiro y col., "Snail1 transcriptional repressor binds to its own promoter and controls its expression," *Nucleic acids research*, vol. 34, n.º 7, págs. 2077-2084, 2006.

[218]L. Dunipace, J. M. McGehee, J. Irizarry y A. Stathopoulos, "The proximal enhancer of the snail gene mediates negative autoregulatory feedback in Drosophila melanogaster," *Genetics*, iyaf058, 2025.

[219]L. Marucci, "Nanog dynamics in mouse embryonic stem cells: results from systems biology approaches," *Stem cells international*, vol. 2017, n.º 1, pág. 7 160 419, 2017.

[220]H. Fatma y H. R. Siddique, "Cancer cell plasticity, stem cell factors, and therapy resistance: how are they linked?" *Cancer and Metastasis Reviews*, vol. 43, n.º 1, págs. 423-440, 2024.

[221]M.-L. Wang, S.-H. Chiou y C.-W. Wu, "Targeting cancer stem cells: emerging role of Nanog transcription factor," *OncoTargets and therapy*, págs. 1207-1220, 2013.

[222]Z. Zhang e Y. Zhang, "Transcriptional regulation of cancer stem cell: regulatory factors elucidation and cancer treatment strategies," *Journal of Experimental & Clinical Cancer Research*, vol. 43, n.º 1, pág. 99, 2024.

[223]V. Chickarmane, C. Troein, U. A. Nuber, H. M. Sauro y C. Peterson, "Transcriptional dynamics of the embryonic stem cell switch," *PLoS computational biology*, vol. 2, n.º 9, e123, 2006.

[224]M. A. Shea y G. K. Ackers, "The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation," *Journal of molecular biology*, vol. 181, n.º 2, págs. 211-230, 1985.

[225]V. Chickarmane y C. Peterson, "A computational model for understanding stem cell, trophectoderm and endoderm lineage determination," *PLoS one*, vol. 3, n.º 10, e3478, 2008.

[226]T. Kalmar y col., "Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells," *PLoS biology*, vol. 7, n.º 7, e1000149, 2009.

[227]I. Glauche, M. Herberg e I. Roeder, "Nanog variability and pluripotency regulation of embryonic stem cells-insights from a mathematical model analysis," *PloS one*, vol. 5, n.º 6, e11238, 2010.

[228]M. Herberg, I. Glauche, T. Zerjatke, M. Winzi, F. Buchholz e I. Roeder, "Dissecting mechanisms of mouse embryonic stem cells heterogeneity through a model-based analysis of transcription factor dynamics," *Journal of The Royal Society Interface*, vol. 13, n.º 117, pág. 20 160 167, 2016.

[229]M. Herberg, T. Kalkan, I. Glauche, A. Smith e I. Roeder, "A model-based analysis of culture-dependent phenotypes of mESCs," *PloS one*, vol. 9, n.º 3, e92496, 2014.

[230]S. Godwin, D. Ward, E. Pedone, M. Homer, A. G. Fletcher y L. Marucci, "An extended model for culture-dependent heterogenous gene expression and proliferation dynamics in mouse embryonic stem cells," *NPJ systems biology and applications*, vol. 3, n.º 1, pág. 19, 2017.

[231]R. H. Clewley, W. Sherwood, M. LaMar y J. Guckenheimer, "PyDSTool, a software environment for dynamical systems modeling," *URL http://pydstool. sourceforge. net*, 2007.

[232]M. K. Jolly, B. Huang, M. Lu, S. A. Mani, H. Levine y E. Ben-Jacob, "Towards elucidating the connection between epithelial–mesenchymal transitions and stemness," *Journal of The Royal Society Interface*, vol. 11, n.º 101, pág. 20 140 962, 2014.

[233]M. K. Jolly y col., "Coupling the modules of EMT and stemness: A tunable 'stemness window'model," *Oncotarget*, vol. 6, n.º 28, pág. 25 161, 2015.

[234]F. Bocci, H. Levine, J. N. Onuchic y M. K. Jolly, "Deciphering the dynamics of epithelial-mesenchymal transition and cancer stem cells in tumor progression," *Current Stem Cell Reports*, vol. 5, págs. 11-21, 2019.

[235]S. A. Mani y col., "The epithelial-mesenchymal transition generates cells with properties of stem cells," *Cell*, vol. 133, n.º 4, págs. 704-715, 2008.

[236]O. H. Ocaña y col., "Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1," *Cancer cell*, vol. 22, n.º 6, págs. 709-724, 2012.

[237]P. Samavarchi-Tehrani y col., "Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming," *Cell stem cell*, vol. 7, n.º 1, págs. 64-77, 2010.

[238]A. Grosse-Wilde y col., "Stemness of the hybrid epithelial/mesenchymal state in breast cancer and its association with poor survival," *PloS one*, vol. 10, n.º 5, e0126522, 2015.

[239]R. Strauss y col., "Analysis of epithelial and mesenchymal markers in ovarian cancer reveals phenotypic heterogeneity and plasticity," *PloS one*, vol. 6, n.º 1, e16186, 2011.

[240]Y.-X. Lu y col., "Regulation of colorectal carcinoma stemness, growth, and metastasis by an miR-200c-Sox2–negative feedback loop mechanism," *Clinical cancer research*, vol. 20, n.º 10, págs. 2631-2642, 2014.

[241]X. Hua y col., "The inhibitory effect of compound ChlA-F on human bladder cancer cell invasion can be attributed to its blockage of SOX2 protein," *Cell Death & Differentiation*, vol. 27, n.º 2, págs. 632-645, 2020.

[242]C. Peng y col., "A unilateral negative feedback loop between miR-200 microRNAs and Sox2/E2F3 controls neural progenitor cell-cycle exit and differentiation," *Journal of Neuroscience*, vol. 32, n.º 38, págs. 13 292-13 308, 2012.

[243]Y. Chen y X. Wang, "miRDB: an online database for prediction of functional microRNA targets," *Nucleic acids research*, vol. 48, n.º D1, págs. D127-D131, 2020.

[244]V. Agarwal, G. W. Bell, J.-W. Nam y D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *elife*, vol. 4, e05005, 2015.

[245]G. Wang y col., "Critical regulation of miR-200/ZEB2 pathway in Oct4/Sox2-induced mesenchymal-to-epithelial transition and induced pluripotent stem cell generation," *Proceedings of the National Academy of Sciences*, vol. 110, n.º 8, págs. 2858-2863, 2013.

[246]Q. Quan y col., "Cancer stem-like cells with hybrid epithelial/mesenchymal phenotype leading the collective invasion," *Cancer science*, vol. 111, n.º 2, págs. 467-476, 2020.

[247]W. Shi y col., "Hill Function-based Model of Transcriptional Response: Impact of Nonspecific Binding and RNAP Interactions," *arXiv preprint arXiv:2403.01702*, 2024.