UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE QUÍMICA

IDENTIFICACIÓN DE LINCRNAS COMO BIOMARCADORES DE PREDICCIÓN DE RESPUESTA A LA QUIMIOTERAPIA NEOADYUVANTE, EN PACIENTES CON CÁNCER DE MAMA LOCALMENTE AVANZADO MEDIANTE ANÁLISIS DE TRANSCRIPTOMA

TESIS

QUE PARA OBTENER EL TÍTULO DE

QUÍMICA FARMACÉUTICO BIÓLOGA

PRESENTA

LAURA MARIANA CONTRERAS ESPINOSA

CD. MX. **AÑO 2019**

JURADO ASIGNADO:

PRESIDENTE.	**Profesor: Dr. Francisco Javier Plasencia de la Parra**
VOCAL.	**Profesor: Dr. Javier Andrés Juárez Díaz**
SECRETARIO.	**Profesor: Dr. Cristian Gabriel Oliverio Arriaga Canon**
1er. SUPLENTE.	**Profesora: Dra. Verónica Garrocho Villegas**
2° SUPLENTE.	**Profesor: Alberto García Lozano**

SITIO DONDE SE DESARROLLÓ EL TEMA:

INSTITUTO NACIONAL DE CANCEROLOGÍA (INCAN)

ASESOR DEL TEMA:

Dr. Cristian Gabriel Oliverio Arriaga Canon

SUSTENTANTE:

CONTRERAS ESPINOSA LAURA MARIANA

Agradecimientos

Investigación realizada gracias al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la UNAM en el proyecto IA207017. Agradezco a la DGAPA-UNAM la beca recibida.

Abreviaturas y siglas

ºC: Grado celsius.

BMPR1B: Proteína receptora asociada a morfogénesis del hueso 1 B (*BMPR1B*, por sus siglas en inglés).

CaMa: Cáncer de Mama.

cDNA: DNA complementario.

ChIP-Sec: Secuenciación por Inmunoprecipitación de la cromatina (del inglés *Chromatin Inmunoprecipitation*).

CMLA: Cáncer de Mama Localmente Avanzado.

DNA: Ácido desoxirribonucleico.

EDTA: Ácido Etilaminotetraacético.

FDR: Tasa de falsos descubrimientos (por sus siglas en inglés *False Discovery Rate*).

H3K4me1: Monometilación de la lisina 4 en la histona 3.

H3K4me3: Trimetilación de la lisina 4 en la histona 3.

H3K9me1: Monometilación de la lisina 9 en la histona 3.

H3K9me3: Trimetilación de la lisina 9 en la histona 3.

H3K27ac: Acetilación de la lisina 27 en la histona 3.

H3K27me3: Trimetilación de la lisina 27 en la histona 3.

H3K36me3: Trimetilación de la lisina 36 en la histona 3.

HER2: Receptor 2 del factor de crecimiento epidérmico humano (del inglés *Human Epidermal Growth Factor Receptor 2 or ERBB2*).

IARC: Agencia Internacional de la Investigación en Cáncer (por sus siglas en inglés).

INCan: Instituto Nacional de Cancerología.

Kb: Kilobase.

lncRNA: RNA largo no codificante (del inglés *long non-coding RNA*).

lincRNA: RNA largo no codificante intergénico (del inglés *long intergenic non-coding RNA*).

lincRNA-ACR: RNA largo no codificante intergénico asociado a quimiorresistencia (del inglés *long intergenic non-coding RNA Associated to Chemoresitance*).

min: Minuto.

miRNA: Micro RNA.

mRNA: RNA mensajero.

ncRNA: RNA no codificante.

nt: Nucleótidos.

NTC: Control Negativo.

OMS: Organización Mundial de la Salud.

pb: Pares de bases.

PC: Potencial Codificante.

PCR: Reacción en Cadena de la Polimerasa.

PolII: RNA Polimerasa II.

QR: Quimiorresistencia.

QT: Quimioterapia.

QTNeo: Quimioterapia Neoadyuvante.

RAP: Purificación de RNA Antisentido (del inglés *RNA Antisense Purification*).

RCB: Carga Residual Tumoral (por sus siglas en inglés *Residual Cancer Burden*).

RE: Receptor de Estrógenos.

RIN: Número de Integridad del RNA (del inglés *RNA Integrity number*).

RNA: Ácido ribonucleico.

RNA-ChIP: Inmunoprecipitación de la cromatina y del RNA.

RNA-Sec: Secuenciación masiva en paralelo de RNA.

ROC: Característica Operativa del Receptor (del inglés *Receiver Operating Characteristic*).

RP: Receptor de Progesterona.

RPC: Respuesta Patológica Completa.

RPM: Revoluciones por minuto.

RT: Enzima reverso-transcriptasa.

s: Segundo.

SFB: Suero Fetal Bovino

SS: Secretaría de Salud.

TN: Triple Negativo.

TPM: Transcritos por Millón.

Resumen

El cáncer de mama es un problema de salud pública en México, y se detecta en un 70% en etapas localmente avanzadas al momento del diagnóstico. En el Instituto Nacional de Cancerología la quimioterapia neoadyuvante ha surgido como la terapia estándar para el manejo del cáncer de mama localmente avanzado; sin embargo, los estudios han demostrado que menos del 50% de las pacientes presentan respuesta patológica completa con este esquema terapéutico, por lo que es necesario buscar marcadores moleculares de predicción de respuesta a la quimioterapia neoadyuvante. Los RNA largos no codificantes intergénicos (lincRNA, por sus siglas en inglés) se definen como transcritos mayores a 200 bases que no codifican a proteínas y se ha reportado en estudios de secuenciación masiva en paralelo de RNA (RNA-Sec) que presentan perfiles de expresión anormales en cáncer de mama. A pesar de ello, no se conoce el valor predictivo de los lincRNA en las terapias para este padecimiento. En este trabajo se identificaron a los lincRNA como marcadores moleculares de predicción en la respuesta a la quimioterapia neoadyuvante, en pacientes con cáncer de mama localmente avanzado mediante el análisis del transcriptoma por RNA-Sec. Para ello, se realizó un análisis de expresión diferencial del transcriptoma con base en resultados de RNA-Sec obtenidos de 12 muestras de pacientes con adenocarcinoma mamario (etapas clínicas IIB-IIIC, subtipo molecular Luminal B), sin tratamiento oncológico previo, en un estudio de tipo casos y controles. Como resultado de este trabajo se identificaron 74 lincRNA diferencialmente expresados asociados con la resistencia a la quimioterapia neoadyuvante, de los cuales se validó experimentalmente el lincRNA-ACR por PCR en tiempo real. Los resultados indican que el lincRNA-ACR se sobreexpresa en el grupo de pacientes que presentan QR, y al evaluar su capacidad como método predictivo de respuesta a la quimioterapia neoadyuvante se determinó que el biomarcador tiene una sensibilidad del 93.7% y una especificidad del 85.7%. En conclusión, existe un perfil de expresión diferencial de lincRNA característico del grupo de pacientes QR que se asocia con la resistencia a la quimioterapia neoadyuvante. Particularmente, el lincRNA-ACR tiene una alta sensibilidad y especificidad en la predicción de respuesta a la quimioterapia neoadyuvante, lo que sugiere su uso como un marcador molecular clínico en cáncer de mama.

Abstract

Breast cancer is a public health problem in Mexico, and is detected at 70% in locally advanced stages. In the National Institute of Cancerology (INCan) neoadjuvant chemotherapy has emerged as the standard therapy for the management of locally advanced breast cancer, however studies have shown that less than 50% of patients have complete pathological response with this therapeutic scheme, so it is necessary to look for molecular markers to predict neoadjuvant chemotherapy response. Long non-coding intergenic RNAs (lincRNAs) are defined as transcripts greater than 200 bases that do not encode proteins and have been reported in parallel sequencing studies of RNA (RNA-Seq) that present abnormal expression profiles in CaMa. Despite this, the predictive value of lincRNAs in therapies for this condition is not known. The aim of this study was to identify lincRNAs as molecular markers of prediction in the response to neoadjuvant chemotherapy, in patients with locally advanced breast cancer by analyzing the transcriptome by RNA-Seq. A differential expression analysis of the transcriptome was performed based on results obtained from RNA-Seq of 12 samples from patients with breast adenocarcinoma (clinical stages IIB-IIIC, Luminal B), without previous oncological treatment, in a case-control study. We identified 74 differentially expressed lincRNAs associated with resistance to neoadjuvant chemotherapy, of which the lincRNA-ACR was validated experimentally by real-time PCR. The results indicate that the lincRNA-ACR is over-expressed in the group of QR cases, and when assessing its capacity as a predictive method of response to chemotherapy, it was determined that it is associated with a sensitivity of 93.7% and a specificity of 85.7%. As a conclussion, there is a differential expression profile of lincRNAs characteristic of the QR control group that is associated with resistance to neoadjuvant chemotherapy. Particularly, lincRNA-ACR is associated with a high sensitivity and specificity in the prediction of response to neoadjuvant chemotherapy, which suggests its use as a clinical marker in breast cancer.

TABLA DE CONTENIDO

ÌNDICE DE FIGURAS

ÌNDICE DE TABLAS

1. Introducción

1.1. Historia natural de la enfermedad

El cáncer de mama (CaMa) se define como un conjunto de padecimientos caracterizados por el crecimiento celular descontrolado en el tejido epitelial que recubre la mama (Figura 1). Al progresar es capaz de extenderse a toda la zona anatómica mamaria y fuera de ella en el proceso de metástasis[1,2]. Aunque el crecimiento tumoral es característico del tejido epitelial mamario, el desarrollo del CaMa puede iniciarse con la aparición de células malignas en cualquier zona de la mama, llevando a la formación de tejido neoplásico, que se extenderá a los ganglios linfáticos cercanos y finalmente invadirá otros órganos[2]. Este proceso de progresión se ha descrito en 5 etapas principales[3].

- **Incepción**. Aparición y proliferación de células neoplásicas, principalmente de origen epitelial.
- **Crecimiento intraepitelial**. El tumor progresa al incluir en su radio de crecimiento células estromales, linfáticas y vasculares, generando un foco tumoral.
- **Invasión inicial**. La confluencia de focos tumorales resulta en la formación de un tumor de mayor tamaño, que es capaz de propagarse dentro y fuera de la mama.
- **Difusión regional**. La propagación tumoral se extiende hasta los ganglios linfáticos regionales cercanos.
- **Difusión sistémica**. La propagación tumoral se extiende hasta órganos sistémicos en el proceso de metástasis.

Durante la etapa de crecimiento intraepitelial las células de diferentes linajes se incorporan al tumor, por lo que la progresión del cáncer puede incluir diversos fenotipos histológicos [2,3], que se clasifican en dos grupos principales: los carcinomas *in situ* y los carcinomas invasivos o infiltrantes[4]. Los carcinomas *in situ* se caracterizan por ser benignos y suelen subdividirse en ductales o lobulares dependiendo de la ubicación del foco tumoral inicial[2]. En cambio, los carcinomas invasivos o infiltrantes se componen por células capaces de diseminarse en la

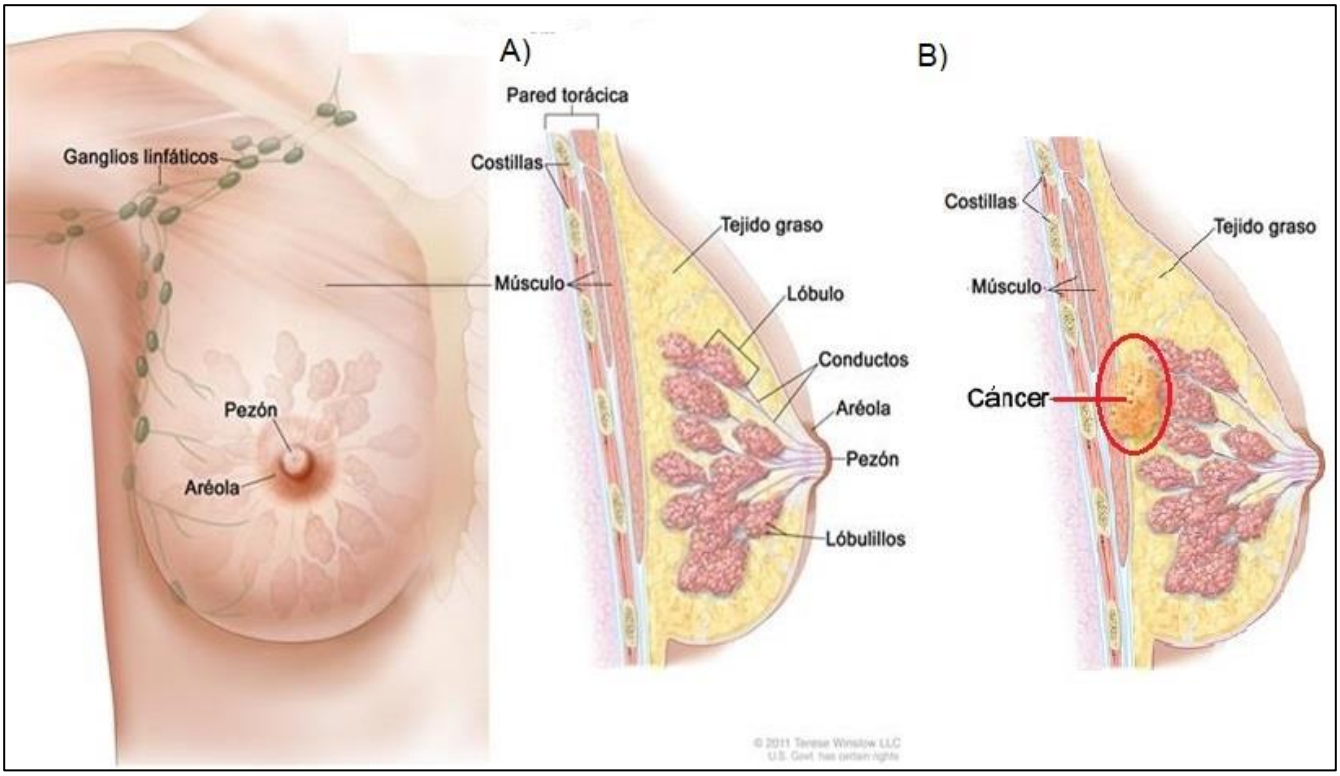


Figura 1. Anatomía de la mama. La mama humana se localiza entre la segunda y la sexta costilla, en el plano anterior respecto a las costillas. A) La unidad funcional de la mama es el lóbulo glandular, compuesto a su vez por un conjunto de lobulillos delineados por tejido epitelial y rodeada por tejido graso, cuya función es la producción láctea. Cada mama tiene alrededor de 20 lóbulos, y la leche producida por éstos se transporta por los conductos hacia el pezón. B) La mayoría de los tumores se originan en el epitelio de los conductos, aunque también pueden originarse en el epitelio lobular. Imagen modificada de Dellaire, *et al*, 2014[1].

zona mamaria e invadir otros tejidos, por lo cual es el tipo de CaMa que principalmente deriva en cáncer metastásico a nodos linfáticos, pulmones, hueso, cerebro y piel[3]. La metástasis es mortal y se considera la última etapa en la progresión carcinogénica, de ahí la importancia de la detección del CaMa en etapas tempranas de su desarrollo. Actualmente, el diagnóstico del CaMa en México ocurre en etapas avanzadas, lo que se conoce como cáncer de mama localmente avanzado (CMLA) y se relaciona con un mal pronóstico y con un aumento en el índice de mortalidad, convirtiéndolo en uno de los principales problemas de salud pública en nuestro país[5].

1.2. Epidemiología

El CaMa es una de las principales causas de ***morbilidad*** y ***mortalidad*** relacionadas a neoplasias malignas a nivel mundial, siendo la primera causa de muerte por cáncer en la población femenil (Figura 2) y la quinta causa de muerte global con un estimado de 2.088 millones de casos nuevos en 2018, lo que corresponde al 11.6% del total de diagnósticos de cáncer[6]. En 2015 la Organización Mundial de la Salud (OMS) reportó al CaMa como la octava causa de muerte en mujeres a nivel mundial, contando 571,000 de las 8.8 millones de muertes por cáncer[6].

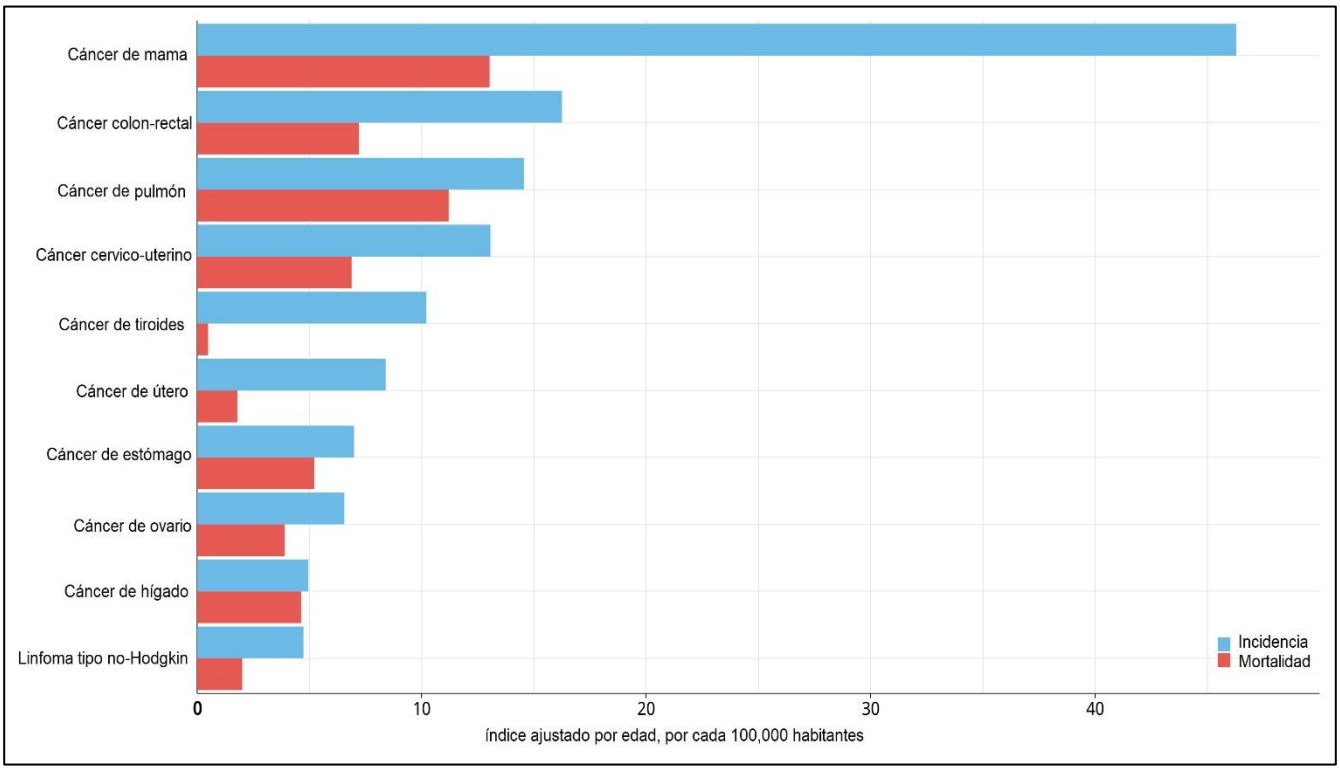


Figura 2. El CaMa es la principal causa de muerte por cáncer en el mundo. Gráfico de barras que presenta los índices de incidencia y mortalidad de los diferentes tipos de cáncer en la población femenil a nivel mundial (presentado como la cantidad de personas afectadas por cada 100,000 habitantes en el eje de las abscisas). Gráfico obtenido en Cancer Today[6].

La mayor ***incidencia*** de CaMa se reporta en las regiones del norte de América y en Europa. En Estados Unidos de Norte América se registraron 252,710 nuevos casos diagnosticados con CaMa, además de 40,610 decesos relacionados a este padecimiento en 2017[7]. Según lo reportado por la Agencia Internacional de la Investigación en Cáncer (IARC, por sus siglas en inglés) en 2012, en México el CaMa se posiciona en primer lugar de incidencia con un índice del 35.4% y una ***prevalencia*** en cáncer a 5 años con un índice del 21%[6]. Asimismo, el Instituto

Nacional de Cancerología (INCan) reportó 317 ingresos de pacientes con tumores mamarios en 2016, de los cuales 287 son clasificados como tumores malignos[8], lo que se relaciona con el informe epidemiológico de la Secretaría de Salud (SS) del 2017, que indica un estimado de 9,746 diagnósticos de tumor maligno de la mama, mientras el último reporte del 2018 registra 12,054 nuevos casos a nivel nacional[9].

En cuanto a mortalidad, la OMS indica que el 22% del total de las muertes prematuras en México son relacionadas a cáncer[7], de las cuales el 18% corresponde a decesos asociados al CaMa en la población femenil mexicana, colocándolo como la primera causa de muerte por cáncer en mujeres en nuestro país[10]. En 2008 se registró que 5 de cada 100,000 personas fallecían por causa del CaMa en el total de la población mexicana, mientras que por cada millón de mujeres se reportaron 97 decesos; posteriormente la SS informó 5,338 decesos por CaMa en 2013, y 6,693 en el reporte del 2016, lo que representó el 6% del total de defunciones en ese año[7]. Como muestra la información mencionada, el CaMa afecta a la población femenil en México en proporciones mayores conforme pasa el tiempo, lo que se relaciona con diferentes factores de riesgo que aumentan la probabilidad de desarrollar este padecimiento en la población mexicana.

1.3. Factores de riesgo

El CaMa se considera una enfermedad multifactorial, y se han descrito como principales factores de riesgo aquellos asociados a la respuesta hormonal, el estilo de vida, así como los componentes hereditarios[1]. La respuesta a los ciclos hormonales es uno de los principales mecanismos de desarrollo del CaMa, por lo que factores como la edad (> 50 años), edad en que ocurre la menarquía, embarazos después de los 35 años e interrumpidos, así como la menopausia son elementos que tienen correlación con el desarrollo de CaMa[1]. En cuanto al estilo de vida, algunos hábitos como fumar y la dieta, la exposición a dosis elevadas de radiación y la condición de obesidad aumentan el riesgo de desarrollar este padecimiento[11].
Por otro lado, la historia clínica y familiar de un paciente son indicadores de riesgo cuando hay identificados familiares con CaMa o si la paciente ha sido diagnosticada con CaMa anteriormente[4,11]. Los antecedentes familiares implican la herencia de mutaciones genéticas que se han asociado a la predisposición a CaMa, como los genes *BRCA1/2*[12,13], la proteína

p53 y el receptor 2 del factor de crecimiento epidérmico humano (*HER2/ERBB2*)[14]. Además, el estudio de mutaciones en ***genes conductores*** en el desarrollo de CaMa como *AKT1*[15] y *GATA3*[16]*,* ha permitido establecer que cambios en su secuencia como variaciones de un solo nucleótido, inserción de repetidos o amplificaciones del mismo alteran la función de su producto proteíco y aumentan la probabilidad de desarrollar cáncer, por lo que también se consideran factores de riesgo[4,17].

Finalmente, la identificación de los factores de riesgo genético se ha logrado gracias al estudio molecular del CaMa, ya que éstos genes son parte del conjunto de elementos moleculares que se encuentran involucrados en el proceso carcinogénico y permiten a las células cancerosas llevar a cabo las funciones indispensables para su subsistencia, contribuyendo con el desarrollo tumoral[17].

1.4. Patogénesis molecular del cáncer de mama

El desarrollo del CaMa ocurre a nivel celular cuando en el epitelio mamario una célula aumenta su potencial replicativo y, en consecuencia, es capaz de formar tejido mamario neoplásico[18]. Las características de las células cancerosas son la consecuencia de una transformación celular en la cual se encuentra alterado el genoma, el transcriptoma, el proteoma y el metaboloma de la célula, de modo que ésta presenta diferencias en su comportamiento biológico y que a su vez comparte con otras células cancerosas, por tanto, se conocen como *Características distintivas del cáncer* (Figura 3)[19].

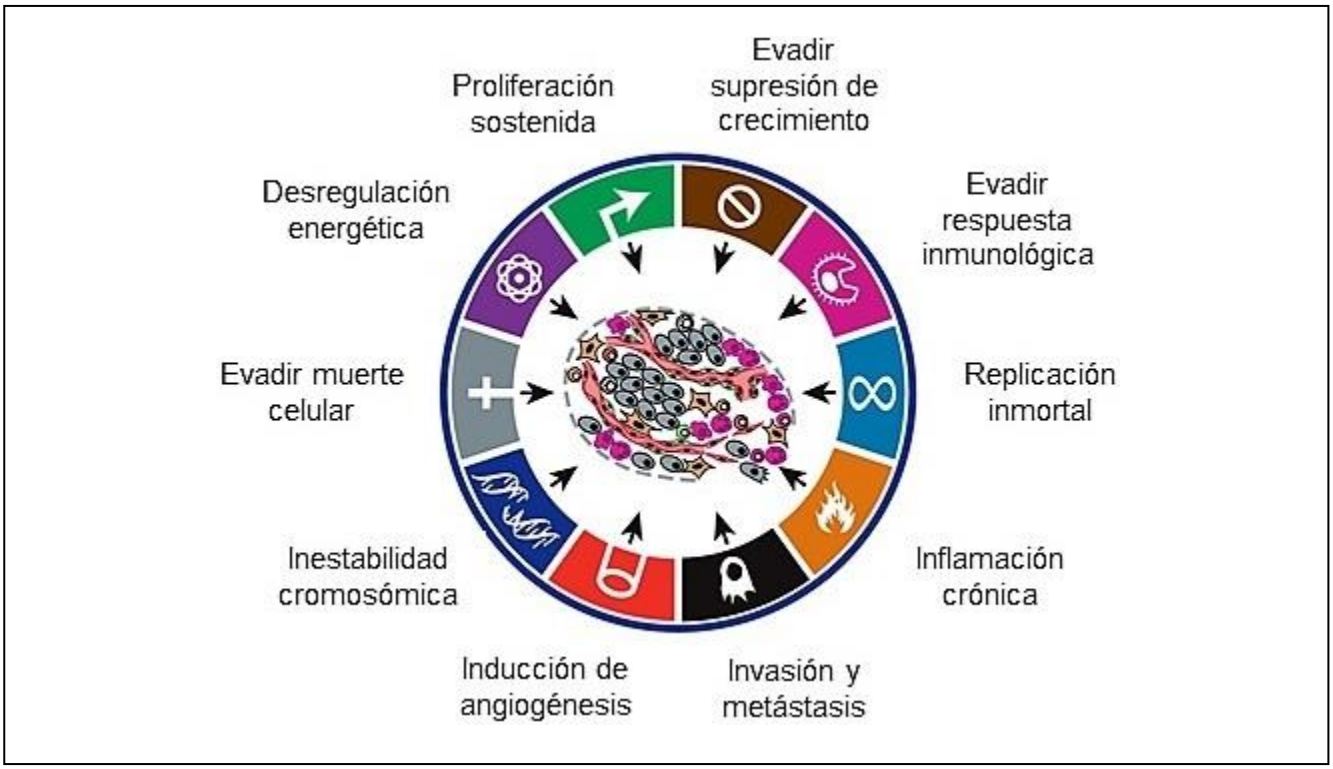


Figura 3. Características distintivas del cáncer. Las células cancerosas se caracterizan por tener anormalidades funcionales comunes, que son determinantes para su desarrollo y supervivencia, logrando así el inicio de la carcinogénesis. Estas características les permiten en general evadir la respuesta inmunológica, la muerte celular y promover la supervivencia celular. Modificada de Hanahan y Weinberg, 2011[19].

Las particularidades que tienen las células cancerosas se explican por la alteración molecular de las diferentes vías de señalización intracelular, que son esenciales para regular adecuadamente las funciones celulares. Una de las vías cuya función fisiológica se encuentra modificada en cáncer es PI3K/Akt/mTOR (Figura 4)[20], que es la principal vía reguladora de la proliferación y del crecimiento celular, porque responde a estímulos externos que promueven el desarrollo y el crecimiento de la célula, tales como la disponibilidad de nutrientes y de factores de crecimiento[21]. En cáncer, la vía PI3K/Akt/mTOR se encuentra activa de manera constitutiva, alterando además otras vías relacionadas con la señalización intracelular como Jak/STAT/Wnt, por lo que aumenta el potencial proliferativo y se promueve la progresión del ciclo celular, llevando así al desarrollo de neoplasias[19].

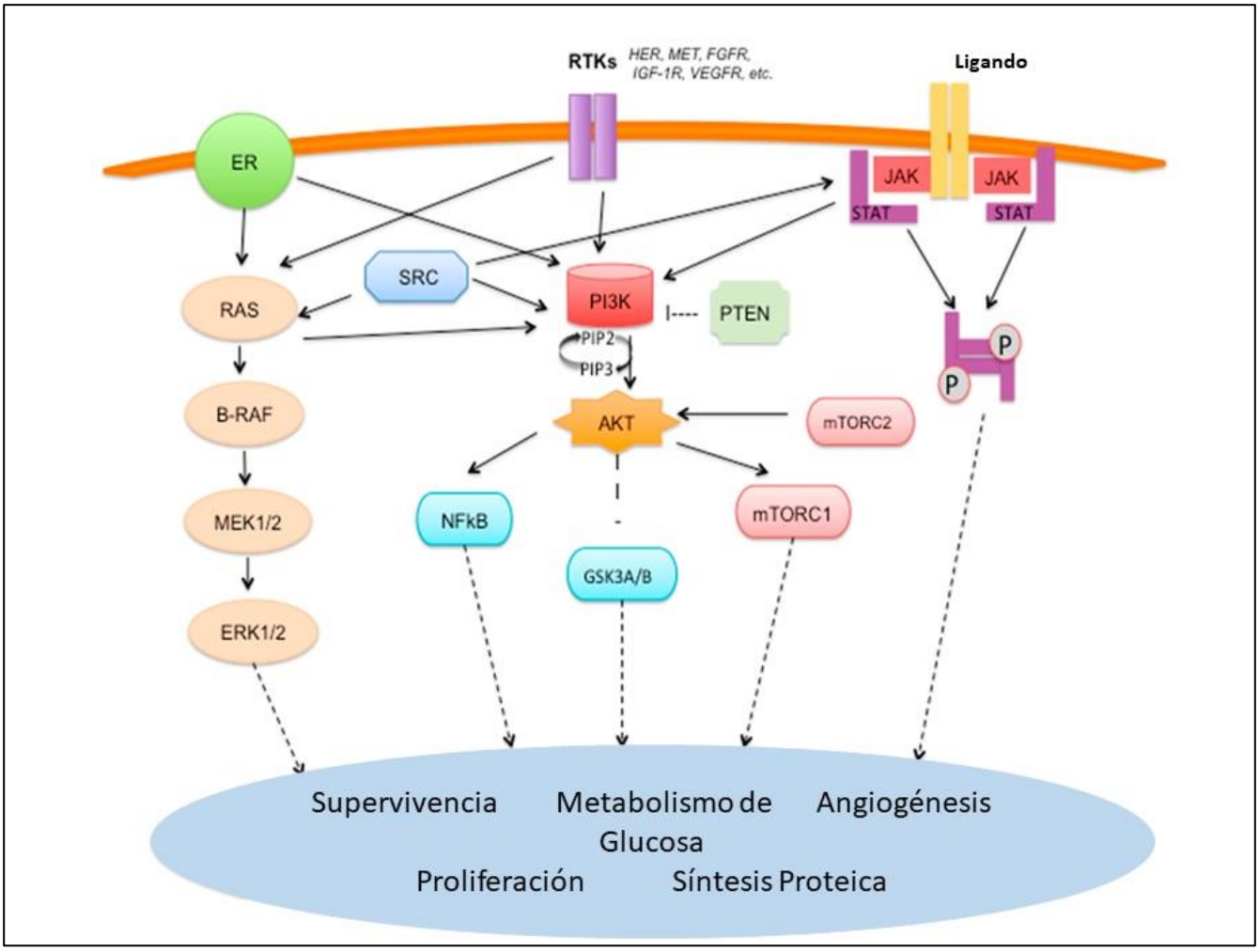


Figura 4. Principales vías de señalización intracelular que contribuyen a la patogénesis molecular del CaMa. La principal alteración de la función normal de las células mamarias comienza con el fenómeno de respuesta a factores de crecimiento o estímulos hormonales, que desencadenan la señalización intracelular regulada por elementos clave en los eventos de supervivencia, metabolismo, proliferación y síntesis proteica. Además, las células pueden incluso alterar sus funciones, como iniciar de manera anómala el proceso de angiogénesis. Modificada de Toss, *et al*, 2017[22].

Otras vías que se encuentran alteradas funcionalmente en CaMa son aquellas afectadas por la pérdida de función de la proteína p53 y las vías de reparación del DNA, lo cual tiene consecuencia directamente en la tasa de mutación, lo que produce la transformación de una célula normal a una célula cancerosa[23]. En ese sentido, la información genética dentro de una célula cancerosa acumula mutaciones hasta que presenta un fenotipo que le permite principalmente evadir la apoptosis, evitar el control del ciclo celular, promover la proliferación y la supervivencia[19]. Esto se debe a que las mutaciones afectan la función de oncogenes[23], que en condiciones fisiológicas se involucran en la regulación de estas funciones[20]. Un ejemplo de ello son aquellos genes descritos anteriormente como factores de riesgo, como *BRCA1/2*,

que participa en la reparación del daño al DNA, y en condiciones patológicas se asocia a la inestabilidad cromosómica[24]. Otro ejemplo es la amplificación de *HER2/*ERBB2, que promueve la proliferación celular y se asocia con la resistencia a tratamiento en CaMa[25]. Asimismo, cambios en la tasa de expresión de algunos otros genes también contribuyen con el desarrollo de tumores, como es el caso de los receptores de estrógeno (RE) y progesterona (RP)[25]. Se ha reportado que la disminución de la expresión de RE en estadíos tempranos se asocia a fenotipos tumorales menos agresivos, mientras un aumento en la expresión de esta misma proteína en estadíos más avanzados promueve el aumento en la tasa de proliferación celular e inestabilidad cromosómica[26].

En suma, establecer los mecanismos moleculares que llevan al desarrollo de CaMa ha permitido mejorar las técnicas de diagnóstico y apoyar tanto en el diseño de terapias, como en el desarrollo de protocolos de seguimiento del paciente oncológico, a través de la identificación de moléculas que pueden servir como marcadores clínicos, que permitan definir patrones de expresión genética que comparten los diferentes grupos de pacientes y, en consecuencia, llevar al desarrollo de la clasificación molecular.

1.5. Diagnóstico

En México, el diagnóstico clínico se basa principalmente en estudios de imagen, como la mastografía, el ultrasonido mamario y la imagen por resonancia magnética, en conjunto con la realización de una biopsia con aguja de corte (Trucut) en lesiones que pueden ser palpables o no[27]. Al confirmarse la presencia de CaMa se determina el tipo tumoral de la paciente de acuerdo a la caracterización molecular, que se realiza identificando la expresión de receptores hormonales (como RE, RP y HER2) y determinando la clasificación histológica con base en el sistema de estadificación TNM, que se basa en el tamaño tumoral (T), si hay nodos linfáticos afectados (N) y el desarrollo de metástasis (M)[28]. En nuestro país, los tumores localmente avanzados representan el 70% de las etapas clínicas al momento del diagnóstico, donde el 53% de los casos se diagnostican como CMLA, que se define como cáncer en los estadios IIB a IIIC con extensión a los ganglios linfáticos[29]. Estas características se relacionan con un mal pronóstico para las pacientes, por lo cual se promueve la detección temprana del CaMa a

través de la auto-exploración, para mejorar el manejo del paciente oncológico en caso de presentar este padecimiento.

Debido a la alta heterogeneidad biológica y clínica que presentan los tumores mamarios, la caracterización molecular de los tumores es actualmente una prueba de rutina en el diagnóstico del CaMa[4] y es la que proporciona más información acerca de la condición de la paciente, por lo que es un elemento de apoyo para las decisiones médicas en el área oncológica[27].

1.5.1 Clasificación molecular del cáncer de mama

El sistema molecular de clasificación tumoral para CaMa se basa actualmente en la expresión de proteínas y se realiza mediante la detección de los receptores RE, RP y HER2, con el uso de técnicas inmunohistoquímicas[25]. Hasta ahora, se han descrito 5 grupos en esta clasificación: *Luminal A, Luminal B, HER2+, Basal* y *Bajo en Claudina*[30] (Figura 5), los cuales han demostrado tener gran utilidad como una herramienta clínica de apoyo para conocer el pronóstico de la paciente y para la elección adecuada del tratamiento. Particularmente, en el INCan se ha reportado que de las pacientes que ingresan con diagnóstico de CMLA, aproximadamente el 55% presentan las características moleculares del subtipo luminal B, por lo que es un grupo de interés para su estudio en este instituto.

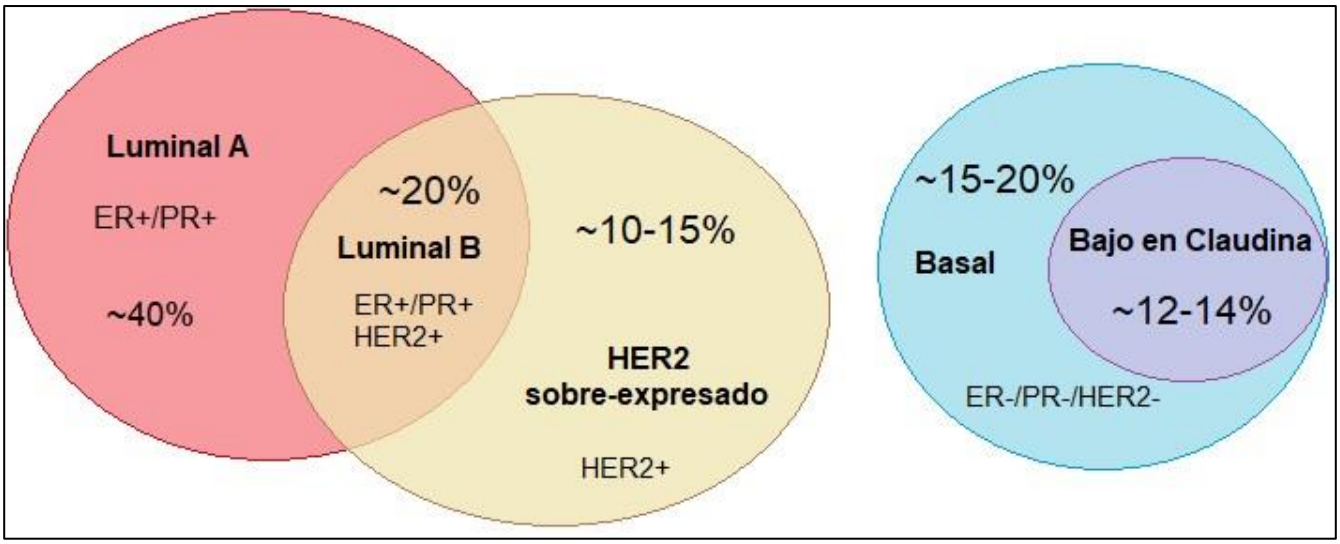


Figura 5. La clasificación molecular permite identificar el tipo tumoral en el diagnóstico de CaMa. Se han identificado 5 tipos principales de tumores en CaMa de acuerdo a los patrones de expresión génica: Los luminales A y B, los HER2 enriquecidos, los basales y los de baja expresión de claudina. Modificada de Malhotra, *et al*, 2010[4].

Debido a la heterogeneidad molecular que presentan los tumores mamarios, los estudios de análisis masivos de expresión génica basados en transcriptomas han demostrado que existen grupos de genes con mayor capacidad de distinguir entre los tipos moleculares que los marcadores por inmunohistoquímica utilizados actualmente. A partir de estos estudios se han desarrollado pruebas moleculares basadas en firmas genéticas que proporcionan información sobre el pronóstico de la paciente y predicen los beneficios de la terapia para las pacientes[31].

1.5.2. Pruebas moleculares en cáncer de mama

Las pruebas moleculares pertenecen al concepto de *Diagnóstico molecular*, que incluye todas las técnicas de *Biología molecular* empleadas en la identificación de marcadores moleculares, ya sea a nivel de DNA, RNA o proteínas, y que sean informativos sobre un estado fisiológico o patológico. En el caso del CaMa, podemos nombrar las pruebas Prosigna (PAM50)[32], Oncotype Dx y Mammaprint[33] como algunos ejemplos de pruebas moleculares basadas en genes codificantes, y que son de uso clínico[34].

Prosigna es un análisis genético que se basa en el perfil de expresión de 50 genes diferentes, entre ellos *FOXA1, FGFR4, HER2/ERBB2* y *BCL2* que ha mostrado tener utilidad en pacientes con CaMa en estadíos I y II, con expresión de los receptores hormonales RE y RP[32]. Esta prueba proporciona información sobre la recurrencia a 10 años, el subtipo molecular de la paciente y tiene valor predictivo para la respuesta a la terapia adyuvante y terapia hormonal[35]. Por otra parte, Oncotype Dx se compone de un panel de 21 genes para pronosticar recurrencia a 10 años después del diagnóstico, además de ser útil en la predicción acerca de los beneficios de la terapia adyuvante[33].

Asimismo, Mammaprint es una prueba molecular basada en microarreglos de expresión, la cual fue aprobada por la Administración de Alimentos y Medicamentos de Estados Unidos (FDA, por sus siglas en inglés) para su uso clínico. Mammaprint es una firma genética de 70 genes, en la cual se consideran genes de referencia, genes relacionados al riesgo de metástasis y genes asociados a la angiogénesis[33]. El análisis en conjunto de la expresión de estos genes proporciona información del riesgo de recurrencia y de desarrollo de metástasis a 10 años, además de tener un valor predictivo en la respuesta a la terapia adyuvante[36].

En resumen, la clasificación molecular del CaMa es un sistema que permite obtener información relevante acerca de la condición patológica de la paciente, contribuyendo así con la ***medicina de precisión***, que tiene como objetivo permitir a los médicos el manejo adecuado del paciente oncológico basado en las características del mismo[37]. Como parte de ello, el uso de marcadores moleculares como los incluidos en las firmas genéticas hace posible optimizar el manejo del paciente, además de mejorar la selección del tratamiento adecuado.

1.6. Tratamiento

La elección del tratamiento para un caso clínico de CaMa no depende de la consideración de factores aislados, sino que requiere de tener toda la información posible acerca de la situación patológica de la paciente. Existen dos modalidades generales de terapia: la adyuvante y la neoadyuvante. Su selección dependerá de la situación de la paciente, del criterio del médico encargado y de la evidencia clínica[28].

El tratamiento adyuvante es una estrategia médica en la cual se administran agentes terapéuticos después de la cirugía primaria, con el objetivo de evitar el desarrollo de micrometástasis y reducir la mortalidad asociada a CaMa. Sin embargo, esta modalidad terapéutica ofrece beneficios sólo a pacientes en etapas tempranas del desarrollo del cáncer, las cuales se diagnostican en menos del 30% en México[38].

Por otro lado, el tratamiento neoadyuvante es una modificación de la modalidad adyuvante, en la que antes de la cirugía se administra el tratamiento farmacológico, que puede ser endócrino o quimioterapia (QT), o su administración conjunta de manera sistémica, con la finalidad de aumentar la probabilidad de éxito de la cirugía al eliminar la micrometástasis y reducir el tamaño del tumor. Esto implica no sólo mejorar el control de la paciente, sino también el pronóstico, y supone una mejoría en su calidad de vida[39].

En particular, la administración de la quimioterapia neoadyuvante (QTNeo) es el estándar terapeútico en el INCan para pacientes con CMLA (Figura 6). Consiste en la administración de agentes quimioterapéuticos y, en el caso de las pacientes con subtipo molecular HER2+, se administra de manera conjunta con terapia biológica en la cual se utilizan anticuerpos

monoclonales. Entre los agentes farmacológicos se encuentra el fluorouracilo, un anti-neoplásico cuya estructura química es de una pirimidina análoga al uracilo, que se diferencia de éste por tener un grupo fluoruro en la posición C-5 en lugar de un átomo de hidrógeno, lo que altera el metabolismo de los nucléosidos y, por lo tanto, es capaz de incorporarse a las moléculas de RNA y DNA, llevando a la muerte celular[40]. Otro fármaco utilizado en este tratamiento es la adriamicina, una antraciclina que causa daño a las células cancerosas por dos vías: intercalándose en el complejo DNA-Topoisomerasa II en el proceso de reparación de daño a DNA, o generando radicales libres al ser oxidada por el metabolismo celular[41]. Finalmente, la ciclofosfamida es un agente alquilante[42] que puede causar daño mediante diferentes mecanismos como añadir un grupo alquilo a las bases nitrogenadas en el DNA y el RNA, lo que puede ocasionar la formación de enlaces intercatenarios en el DNA e interferir con la síntesis del mismo, así como en el proceso de transcripción. Además, este mismo agente puede provocar el desapareamiento de bases entre las hebras de DNA, llevando a la generación de mutaciones[43].

En el caso particular donde una paciente expresa el receptor HER2, es posible incluir en el cuadro la administración de trastuzumab, un anticuerpo monoclonal que interactúa directamente con la proteína HER2, inhibiendo la cascada de señalización que éste receptor activa, la cual contribuye con los procesos de proliferación y crecimiento celular[44]. Este esquema ha demostrado ser el más benéfico para pacientes con CMLA, ya que aumenta hasta 25% la tasa de supervivencia a 10 años en este grupo[45]. No obstante las pacientes que presentan ***respuesta patológica completa*** (RPC) con la QTNeo representan menos del 50%, mientras el resto de las pacientes desarrolla ***quimiorresistencia*** (QR)[46,47].

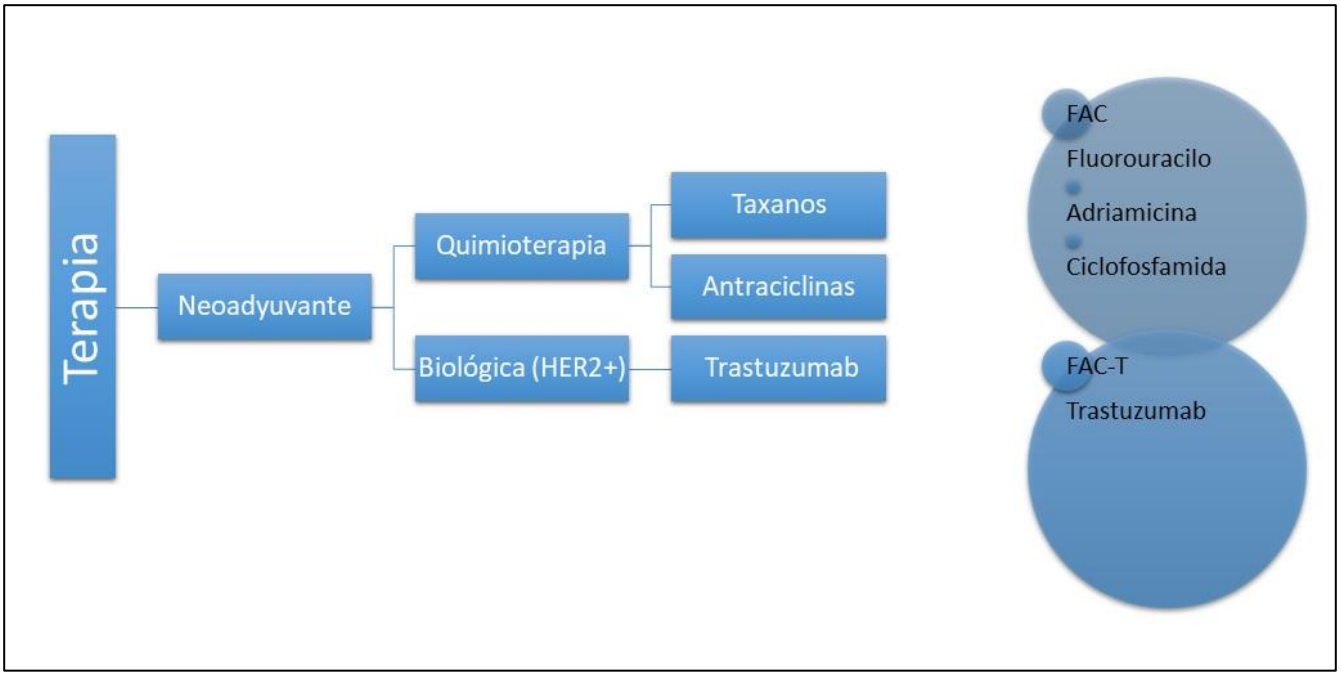


Figura 6. Cuadro de tratamiento en la quimioterapia neoadyuvante implementado en el INCan. El cuadro neoadyuvante consiste en dos etapas, la primera incluye la administración de agentes terapéuticos del cuadro de quimioterapia (taxanos y antraciclinas) y biológico (trastuzumab). La quimioterapia administrada en el INCan se conoce como cuadro FAC, y si éste se combina con terapia biológica se conoce como FAC-T. La combinación es viable si la paciente es positiva a la expresión de HER2, y resulta en un mejor pronóstico.

Debido al alto porcentaje de pacientes con CMLA resistentes a la QTNeo, es necesario definir de manera más apropiada los criterios de elección terapéutica, entre ellos los marcadores de predicción de respuesta a tratamiento, que en su mayoría han sido desarrollados para la terapia adyuvante (como Mammaprint y Oncotype Dx), por lo que es necesario desarrollar nuevos marcadores para su aplicación en la predicción de respuesta a la terapia neoadyuvante.

1.6.1 Biomarcadores moleculares de predicción en cáncer de mama

El Instituto Nacional de Salud de los Estados Unidos de América define un biomarcador como cualquier molecula biológica (metabolitos, electrolitos o macromoléculas como el RNA, DNA y proteínas) que se puedan detectar en el cuerpo, como en la sangre, en los tejidos y en los fluidos corporales mediante métodos invasivos (biopsia, muestras de sangre) o no invasivos (muestra de orina, heces, saliva), y que además son medibles con reproducibilidad[48,49]. Los biomarcadores pueden ser indicadores de procesos fisiológicos, patológicos o de respuesta a intervenciones terapéuticas, y se clasifican de acuerdo a su función como biomarcadores *de*

exposición, de susceptibilidad y *de efecto/respuesta*, por lo que se consideran signos médicos que proporcionan información objetiva del estado de salud de un paciente[50].

En cuanto a los biomarcadores utilizados en el estudio del cáncer destacan los marcadores moleculares, que presentan la ventaja de ser indicadores objetivos de parámetros clínicos como susceptibilidad, riesgo, pronóstico, monitoreo de recurrencia y riesgo de desarrollo de metástasis. Además, se ha reportado su utilidad en el proceso de diagnóstico oncológico y para la predicción de respuesta a las diferentes terapias[50]. Los biomarcadores moleculares de predicción de respuesta a terapia son indicadores del posible beneficio que puede representar el tratamiento para un paciente. En particular, este tipo de marcadores son indicadores de resistencia a la terapia y su uso ha permitido mejorar la elección de tratamiento en la práctica clínica[51].

En CaMa, el uso de biomarcadores moleculares de predicción es común en la práctica clínica. Se basa principalmente en la detección de proteínas por técnicas inmunohistoquímicas y de genes codificantes por pruebas moleculares[52], como es la firma genética Oncotype Dx, ya que han demostrado asociarse con una alta ***sensibilidad*** y ***especificidad***[53]. No obstante, la mayoría de estas moléculas aún se encuentran en la etapa de protocolos clínicos, dado que no han sido aprobados para su uso estandarizado en clínica como marcadores individuales, como ha sido el caso para *CYP2D6* y *STAT3*[54].

En las guías clínicas de Europa y de Estados Unidos se recomienda la detección de la expresión de genes como *Ki67, RE, PR, uPA, PAI-1* y *HER2/ERBB2* para predecir la respuesta a la QT adyuvante en CaMa como alternativa al uso de firmas genéticas; sin embargo, aún no existen estas recomendaciones para la QTNeo, por lo que es necesaria la identificación de biomarcadores moleculares de predicción de respuesta a esta modalidad terapéutica[52].

Estudios recientes basados en los análisis bioinformáticos realizados a partir de datos generados por RNA-Sec han demostrado que existen perfiles de expresión no sólo de genes codificantes, sino también de RNA no codificantes (ncRNA) asociados con la respuesta de pacientes a la QTNeo, en específico perfiles de expresión de RNA largos no codificantes (lncRNA)[55], por lo que se requiere investigar más acerca de la relación que existe entre los lncRNA y la respuesta a la QTNeo en CaMa.

1.7. Generalidades de los lncRNA

Las evidencias científicas han llegado a establecer que sólo una pequeña parte del genoma en humanos codifica proteínas, que es el equivalente a menos del 2% del genoma total, siendo el 98% restante DNA no codificante. En años recientes se confirmó que gran parte de las secuencias no codificantes se transcriben, dando como resultado la síntesis de distintos ncRNA entre los que se encuentran transcritos de naturaleza no codificante como los microRNA, los RNA ribosomales y los lncRNA. Los lncRNA se definen como transcritos que carecen de potencial codificante, cuya longitud es mayor a 200 bases, se componen de 1 a 3 exones y se clasifican de acuerdo a su función y localización genómica. Entre las modificaciones bioquímicas para los lncRNA se ha reportado que éstos presentan en general la adición del nucleótido modificado 7-metil guanosina (caperuza o *cap*) en la región 5', muestran la poliadenilación del extremo 3' (aunque existen lncRNA que no tienen esta modificación), sufren procesamiento post-transcripcional (*splicing*) y son generalmente sintetizados por la RNA polimerasa II (PolII)[56]. La importancia de los lncRNA radica en que éstos se asocian principalmente a la regulación de procesos que contribuyen con la homeostasis celular y su desregulación puede llevar al desarrollo de procesos patológicos, como el cáncer[57].

1.7.1. Clasificación de los lncRNA

A pesar de la gran diversidad de lncRNA identificados hasta la fecha, en la actualidad no existe una clasificación satisfactoria para este tipo de RNA debido a su amplia gama de tamaños, funciones y patrones diferenciales de expresión. Sin embargo, frecuentemente se les ha clasificado dependiendo de su localización en el genoma, por ejemplo, los podemos localizar en regiones intergénicas, en exones, en intrones o en regiones no traducidas de los genes codificantes, lo cual ha servido de apoyo en la clasificación de los lncRNA (Figura 7). En particular, los lncRNA intergénicos (lincRNA) representan el 47% del biotipo de lncRNA, y entre sus principales funciones se encuentran el remodelado de la estructura de la cromatina, la regulación de dominios génicos, así como la regulación de la traducción, entre otros. Además, éstos regulan procesos celulares como el crecimiento, la proliferación y la diferenciación, incluso actualmente se sabe que intervienen en vías de señalización intracelular en algunos

tipos de cáncer, lo cual establece la gran importancia de este tipo de transcritos de naturaleza no codificante en el proceso carcinogénico[56].

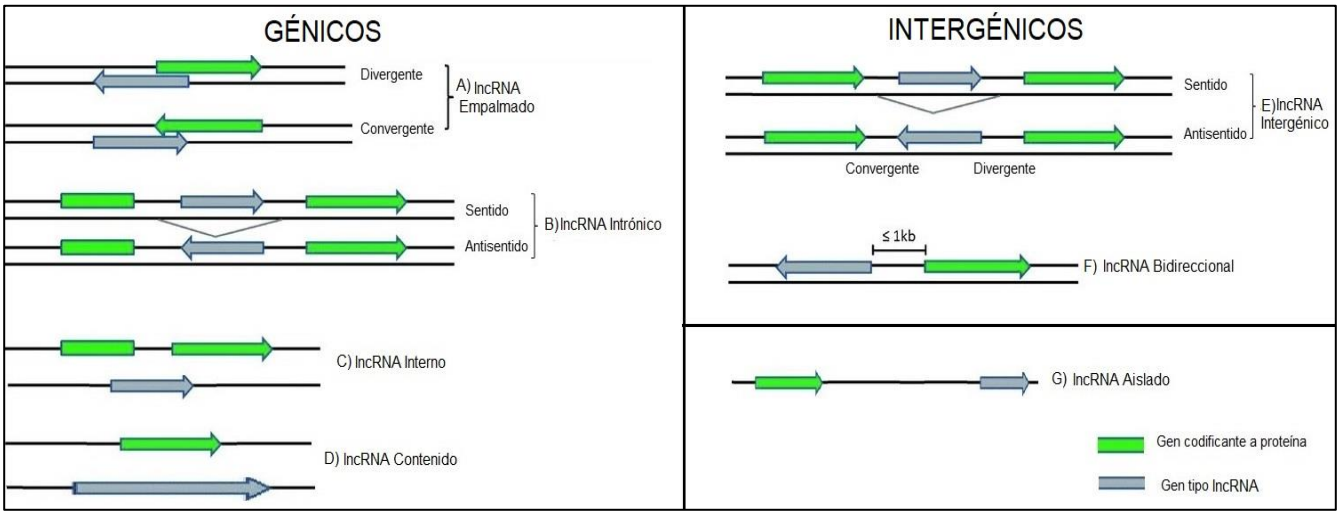


Figura 7. Clasificación de los lncRNA. De acuerdo con su localización en el genoma, los lncRNAs se clasifican como génicos (izquierda) e intergénicos (derecha). A) Los lncRNAs génicos pueden estan empalmados cuando su secuencia coincide en la posición genómica con algún gen codificante (en este caso pueden ser convergentes o divergentes). B) Los lncRNAs intrónicos se localizan dentro de los intrones de genes. C) Los lncRNAs internos se refieren a aquellos que se encuentran completamente contenidos en un gen codificante. D) Los lncRNA contenidos se refieren a cuando un gen codificante se encuentra completamente empalmado dentro de un lncRNA. E) Los lncRNAs intergénicos son los que se encuentran adyacentes a un gen codificante o cuando el lncRNA está ubicado entre dos genes. F) Los lncRNAs bidireccionales son aquellos sintetizados en sentido opuesto al gen adyacente. G) Los lncRNAs aislados son todos aquellos que no se encuentran en ninguna de las condiciones mencionadas anteriormente. Tanto los lncRNA génicos como los intergénicos pueden sintetizarse en la cadena sentido como antisentido. Modificada de Derrien, *et al*, 2012[58].

Por otro lado, se ha determinado que los lncRNA pueden regular a sus genes blanco en *cis* y en *trans*. La regulación en *cis* es cuando la expresión del lncRNA puede afectar la expresión del gen y/o genes adyacentes en el mismo cromosoma donde éste es sintetizado. Por su parte, la regulación en *trans* se refiere a cuando un lncRNA regula un gen, o grupo de genes, situados a una gran distancia dentro del mismo cromosoma e incluso en cromosomas distintos donde se está dando la síntesis del lncRNA. Para ambos tipos de regulación, ya sea en *cis* o *trans,* los mecanismos de regulación del lncRNA puede ser diversos, como el mecanismo de señuelo[59,60] (Figura 8). Finalmente, debido a que los lncRNA desempeñan funciones regulatorias, la expresión de los lncRNA está a su vez controlada por otros mecanismos celulares que permiten mantener la homeostasis celular, entre los cuales está la regulación a nivel genético mediada por factores transcripcionales y la regulación epigenética, llevada a cabo a través de las modificaciones post-traduccionales de histonas y la metilación del DNA[60].

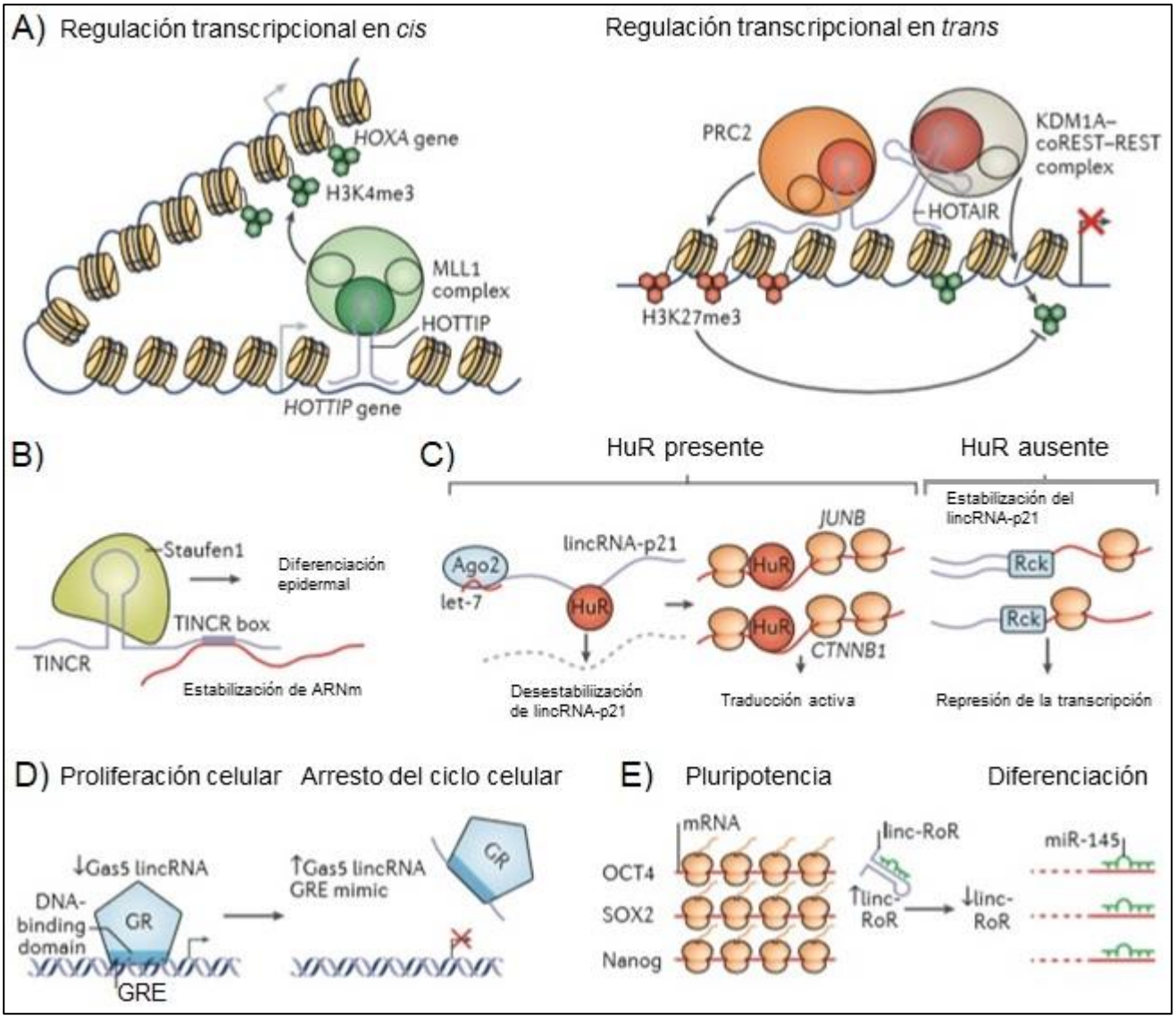


Figura 8. Principales mecanismos de acción asociados a la función de lncRNA. A) Las principales funciones de los lncRNA están asociadas a la regulación transcripcional en *cis* en donde el gen blanco se encuentra cercano al locus del lncRNA, por ejemplo la interacción de HOTTIP con la cromatina, el cual regula la transcripción génica a modo de señal en el locus de los genes homeóticos. Por otro lado, la regulación en *trans* se lleva a cabo en sitios a distancias mayores a 5 kb del locus del lncRNA, un ejemplo de ello es la actividad reguladora de HOTAIR asociada al silenciamiento transcripcional, en la que este transcrito actúa por el mecanismo de guía. Los mecanismos a través de los cuales los lncRNA llevan a cabo estas funciones son principalmente al actuar como plataformas para proteínas, por ejemplo, B) TINCR en el núcleo y C) lincRNA-p21 en el citoplasma, regulando la traducción de sus respectivos mRNA blanco. También pueden funcionar como señuelos, siendo el caso de los lncRNA D) lincRNA-Gas5 que está involucrado en la regulación de la proliferación celular, y de E) lincRNA-RoR con el miR-145, que participa en el proceso de diferenciación celular. Modificada de Ransohoff, ***et al***, 2018[56].

1.7.2. Regulación epigenética de los lncRNA

Los lncRNA se regulan por diferentes mecanismos moleculares como se mencionó anteriormente, entre estos mecanismos se encuentra la regulación epigenética a través de las modificaciones post-traduccionales de las histonas, las cuales conforman al nucleosoma y actúan en la célula al regular la estructura de la cromatina, donde dependiendo de la combinación de modificaciones post-traduccionales de las histonas se puede llevar a un estado activo o de silenciamiento génico, dependiendo de una gran gama de señales regulatorias (Figura 9)[61].

Entre las modificaciones covalentes que se presentan en las histonas se encuentran la acetilación, la metilación, la fosforilación, la ubiquitinación, la sumoilación y la desaminación[62]. Por otro lado, se ha visto que ninguna de estas modificaciones post-traduccionales de las histonas se coloca aleatoriamente, dichas modificaciones siguen un patrón específico a lo largo del promotor y del cuerpo de los genes, lo cual define si la transcripción debe llevarse a cabo o no. Las marcas post-traduccionales de histonas que regulan la transcripción tanto de genes codificantes como de lncRNA son: H3K4me1, H3K4me3, H3K27ac, H3K9ac, H3K36me3, H3K27me3 y H3K9me3[60]. En conjunto, estas marcas de cromatina pueden definir la existencia de unidades transcripcionales activas por lo que han sido de gran utilidad para el estudio de los lncRNA ya que han servido para identificarlos a lo largo del genoma.

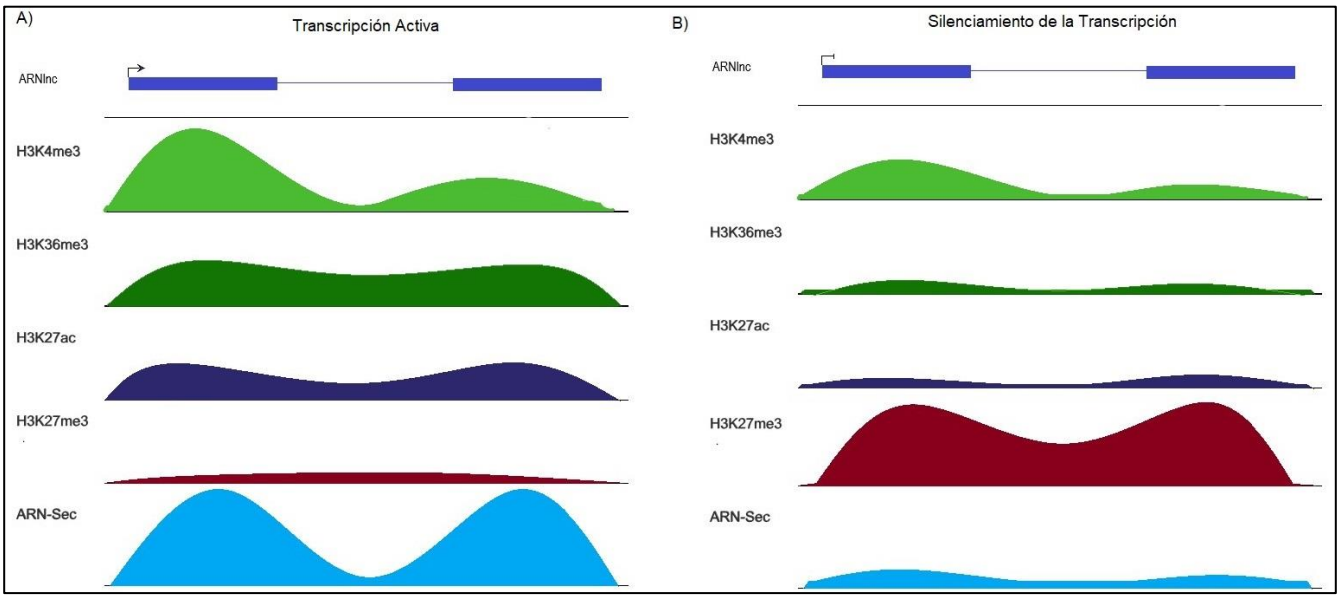


Figura 9. La expresión de los lncRNA es regulada por las marcas post-traduccionales de histonas. A) Transcripción activa. La localización de las marcas post-traduccionales de histonas en el sitio de inicio de la transcripción de los lncRNA define unidades transcripcionalmente activas, como el posicionamiento de la H3K4me3, así como las marcas H3K27ac y H3K36me3 sobre el cuerpo del gen. B) Silenciamiento de la

transcripción. El enriquecimiento de la marca H3K27me3, junto con la baja frecuencia de posicionamiento de las marcas de activación de la transcripción implican el silenciamiento transcripcional de lncRNA.

En suma, la expresión y función de los lncRNA se encuentra regulada por mecanismos tanto genéticos como epigenéticos los cuales permiten que la expresión de lncRNA sea específica en los diferentes procesos biológicos, tipos celulares y tejidos. Así, la alteración de estos mecanismos regulatorios se ha asociado con cambios en los patrones de expresión de los lncRNA en cáncer[63], por lo que estos transcritos de naturaleza no codificante son parte del estudio de este conjunto de patologías y han servido para la búsqueda de nuevos blancos terapéuticos y biomarcadores moleculares de uso clínico.

1.7.3. Los lincRNA y su asociación con el cáncer

Existe evidencia de que los patrones de expresión de los lincRNA en células cancerosas son diferentes a los de las células de tejido normal. Un ejemplo es *EPIC1*, que es un lincRNA identificado a partir de un análisis de expresión diferencial del transcriptoma entre diferentes tipos de cáncer. Los resultados de este estudio muestran que la expresión de *EPIC1* se encuentra alterada principalmente en CaMa, y está sobreexpresado en el tejido neoplásico mamario respecto al tejido normal. Adicionalmente, en el estudio se determinó que el mecanismo molecular por el cual *EPIC1* contribuye a la carcinogénesis es mediante la interacción física con el factor transcripcional MYC, y por tanto es un regulador de la transcripción, así como de la regulación de la expresión de los genes blanco de MYC[64]. Esto demuestra que los lincRNA, como *EPIC1*, están implicados en el proceso carcinogénico y corrobora la importancia de los lincRNA en el estudio del cáncer.

Además, las evidencias experimentales han determinado que los lincRNA se han asociado también con las características principales del cáncer, ya que se ha visto que cambios en los niveles de expresión de algunos lincRNA se relacionan con su desarrollo y progresión[65]. Existe evidencia experimental de que la sobreexpresión de los lincRNA *HOTAIR* y *BCAR4 c*orrelaciona con el desarrollo de metástasis, mientras que la subexpresión del lincRNA *Gas5* promueve la viabilidad de células cancerosas en la mama, al permitir la evasión de la apoptosis (Figura 10)[66,67].

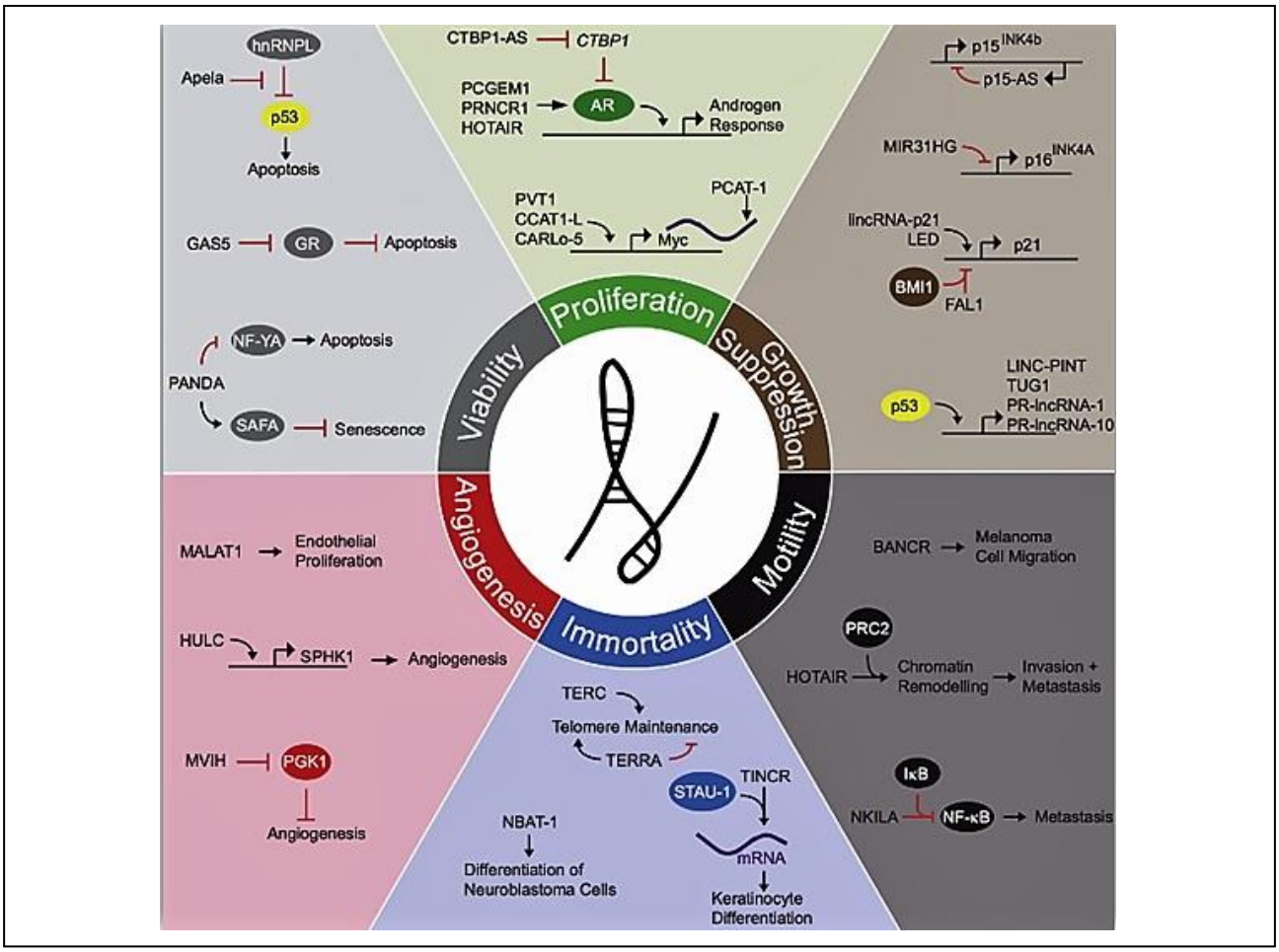


Figura 10. Los lncRNA y su asociación con las características principales del cáncer. Los estudios de correlación en la expresión de lncRNA y cáncer han demostrado su relevancia en el proceso carcinogénico, al promover procesos patológicos. Un ejemplo es HOTAIR, cuya sobreexpresión promueve la invasión y la metástasis. Por otro lado, la subexpresión de TINCR lleva a la desregulación del ciclo celular, lo que tiene como consecuencia la disminución en la tasa de diferenciación de los queratinocitos humanos. Modificada de Schmitt y Chang, 2016[65].

Asimismo, existen algunos estudios que demuestran la asociación del lincRNA con la respuesta a una terapia particular en cáncer. Por ejemplo, en CaMa se ha visto que *CCAT2* se ha asociado a la sobrevida global después de recibir QT adyuvante. De la misma manera, el lincRNA Gas*5*[67] y *linc-ROR*[68] se han relacionado con la resistencia a QT sugieriendo la importancia de los lincRNA y su asociación a la QT. Incluso, algunos lincRNA ya han sido propuestos como biomarcadores para predecir los beneficios de la QT adyuvante (*AK024118, U79277, AK000974* y *BC040204*[67]). En cuanto a su uso en otras terapias, se ha descrito que los lincRNA *TINCR*[69]*, LINC00160* y *LINC01016*[70] están asociados a la resistencia a terapia endócrina, mientras el lncRNA *ATB* se ha visto asociado a la resistencia a trastuzumab en la terapia biológica[71].

En resumen, los lincRNA son moléculas con potencial valor predictivo en las terapias para CaMa y aún no se han analizado a profundidad aquellos relacionados con la respuesta a QTNeo, siendo necesaria su investigación en el desarrollo de biomarcadores para esta terapia en CaMa. La mayoría de los lincRNA que se han relacionado con la respuesta a terapia se han identificado a través del análisis del transcriptoma con el apoyo de herramientas bioinformáticas, por lo que es una metodología útil en su estudio y su asociación con el cáncer.

1.7.4. Herramientas bioinformáticas para el estudio de lincRNA

La secuenciación masiva en paralelo de RNA, o RNA-Sec, es una tecnología que se ha perfeccionado con el desarrollo de la secuenciación de siguiente generación (del inglés *Next Generation Sequencing*), lo que permite hacer el análisis del transcriptoma, que incluye la secuenciación de transcritos de naturaleza no codificante, llevando así a la identificación de transcritos no anotados, como los lincRNA, y hace posible la asociación de sus perfiles de expresión con información clínica, lo que ha llevado a la identificación de nuevos biomarcadores para su aplicación en diagnóstico, pronóstico y predicción de respuesta a tratamiento en cáncer[17].

Un estudio por RNA-Sec consiste en la secuenciación de RNA proveniente de un especimen biológico, se prepara una biblioteca de cDNA a partir de éste y posteriormente se lleva a cabo la secuenciación (Figura 11)[72]. Dependiendo de la pregunta de investigación, la secuenciación puede realizarse por extremo sencillo (*single-end*) o por extremos pareados (*paired-end*), siendo éste último el estudio más completo, ya que nos permite identificar los transcritos sintetizados a partir de la hebra sentido o antisentido del DNA. A partir de esto se determinan la ***cobertura*** y la ***profundidad*** del experimento, que deben ser apropiadas para garantizar la confiabilidad de los resultados finales del análisis bioinformático. Por ejemplo, si se pretende descubrir nuevos transcritos, la profundidad apropiada es de 100 millones de lecturas, mientras que para un análisis de expresión diferencial sólo se necesitan de 15-25 millones de lecturas por muestra secuenciada[73].

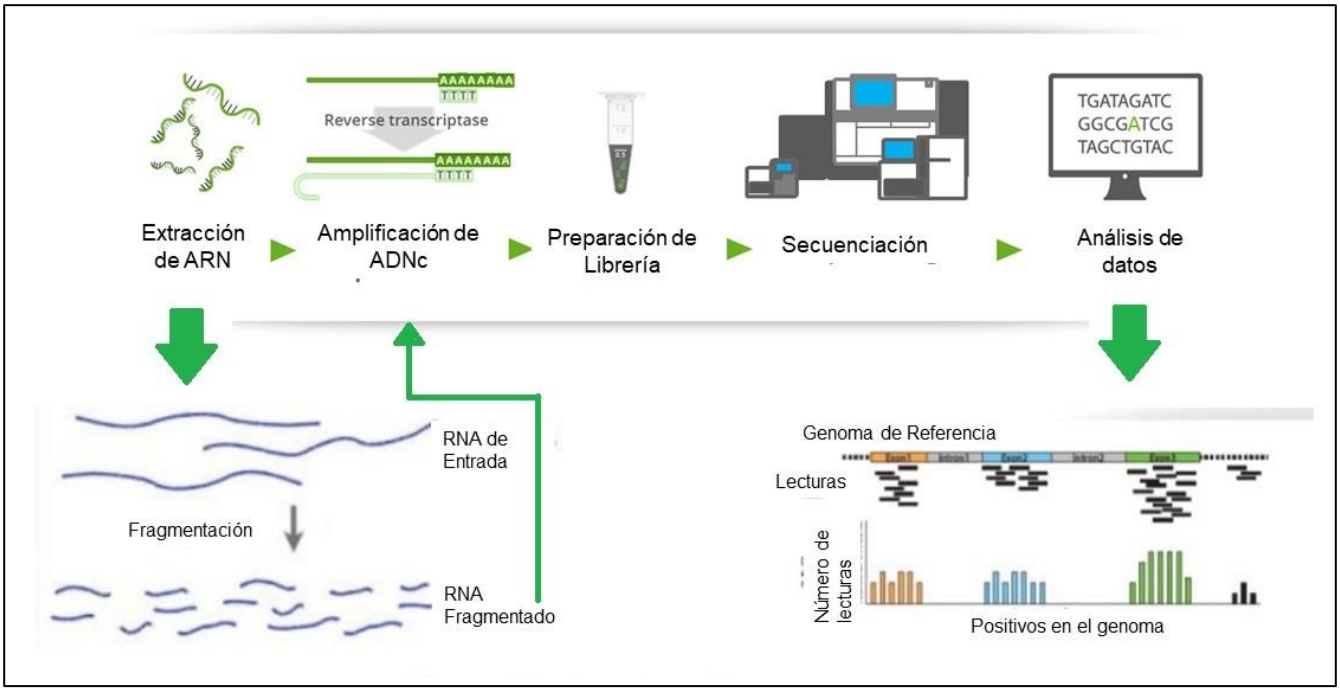


Figura 11. Descripción general de un experimento de RNA-Sec. Para un análisis de transcriptoma debe extraerse el RNA con la calidad adecuada para asegurar la confiabilidad de los resultados. Posteriormente pasará por un proceso de fragmentación (hasta obtener fragmentos de 200-500 nt) y a partir de ellos se generará la biblioteca de cDNA donde se utilizarán adaptadores para su identificación después de la secuenciación. Los resultados de secuenciación deben ser analizados para la identificación y cuantificación de los transcritos, para finalmente poder realizar análisis que determinen la relación de expresión con diferentes condiciones.

Los resultados de un experimento de RNA-Sec implican la obtención de un volumen grande de datos, ya que el archivo que el secuenciador proporciona incluye todas las lecturas de las secuencias de nucleótidos identificadas y, adicionalmente, le asigna a cada base secuenciada un puntaje de calidad, lo que se conoce como archivo FASTQ. Para poder utilizar la información contenida en esta clase de archivos se requiere del uso de herramientas bioinformáticas y de la infraestructura computacional adecuada para el procesamiento de la información de la secuenciación masiva[74]. Actualmente, existen plataformas disponibles en línea, como Galaxy[75], que contiene un paquete de herramientas bioinformáticas comunes a los flujos de trabajo en el procesamiento de datos provenientes de RNA-Sec para la generación de reportes de calidad, de alineamiento de lecturas al genoma de referencia, cuantificación de transcritos y para realizar análisis de expresión diferencial[76].

Los estudios de expresión diferencial se llevan a cabo después de la cuantificación de transcritos y se basan en un análisis estadístico para determinar cambios cuantitativos en la expresión de transcritos comparando entre diferentes condiciones, lo cual los convierte en

estudios útiles en el área clínica para la comparación de estados fisiológicos con diferentes patologías[77]. Con este tipo de abordaje, es posible asociar los resultados de la secuenciación masiva con las diferentes variables clínicas, y se ha descubierto más información que contribuye con el entendimiento acerca de la naturaleza tumoral, llevando a la identificación de nuevos blancos terapéuticos y al desarrollo de biomarcadores[78].

En conclusión, la secuenciación masiva tipo RNA-Sec es una herramienta altamente sensible y precisa que permite determinar la expresión génica global a través del estudio del transcriptoma, teniendo por ventaja la capacidad de analizar incluso los transcritos de naturaleza no codificante. También permite comparar cantidades masivas de información bajo diferentes condiciones patológicas, por lo que representa una metodología de utilidad clínica en la identificación de nuevos biomarcadores. Para el caso particular del CaMa, los biomarcadores utilizados para la predicción de respuesta a tratamiento de los pacientes oncológicos, son principalmente genes codificantes detectados a través de técnicas inmunohistoquímicas y se utilizan sólo para la terapia adyuvante, por lo que el uso de una tecnología de frontera como lo es la RNA-Sec llevará a la identificación de nuevos biomarcadores que permiten predecir la respuesta a la QTNeo, como los lincRNA, cuyo valor predictivo en esta terapia no ha sido explorado a profundidad.

2. Justificación

En este trabajo se generará nuevo conocimiento sobre posibles marcadores moleculares de predicción en pacientes con cáncer de mama localmente avanzado que estará basado en un lincRNA o una firma de expresión genética de lincRNA, lo que podría ayudar a elegir con mayor seguridad a las pacientes que se pueden beneficiar de terapias personalizadas, así como predecir qué pacientes no responderán a la quimioterapia neoadyuvante y así proponer nuevas estrategias terapéuticas que realmente beneficien a este subgrupo de pacientes. Por lo tanto, los resultados de este proyecto también beneficiarían el avance e implementación de la medicina de precisión en nuestro país.

3. Planteamiento del problema

En México, el 70% de los diagnósticos de cáncer de mama están en etapas avanzadas, lo que disminuye la probabilidad de que respondan a tratamiento. En el Instituto Nacional de Cancerología, el tratamiento estándar para estas pacientes es la quimioterapia neoadyuvante. En la actualidad, la elección de tratamiento se apoya en el uso de marcadores moleculares basados en la detección de genes codificantes y de proteínas mediante técnicas inmunohistoquímicas. Sin embargo, la predicción de respuesta para la quimioterapia neoadyuvante carece de marcadores adecuados que ayuden en la gestión adecuada de los tratamientos oncológicos en cáncer de mama localmente avanzado. La tecnología RNA-Sec permite el análisis del transcriptoma y la comparación entre diferentes condiciones patológicas, por lo que es una herramienta útil en el descubrimiento de nuevos biomarcadores de predicción de respuesta a la quimioterapia neoadyuvante en pacientes con cáncer de mama. Por otro lado, se ha demostrado que los lincRNA son biomoléculas que presentan especificidad en su expresión en CaMa, por lo que la evaluación de sus perfiles de expresión tiene un potencial valor predictivo en la respuesta a quimioterapia neoadyuvante, que aún no ha sido explorado a profundidad.

4. Pregunta de investigación

¿El perfil de expresión de lincRNA puede ser utilizado como marcador de predicción de respuesta a la quimioterapia neoadyuvante en pacientes mexicanas con cáncer de mama localmente avanzado?

5. Hipótesis

Existirá un perfil de expresión genética basado en lincRNA con valor predictivo en la resistencia a la quimioterapia neoadyuvante en pacientes mexicanas con cáncer de mama localmente avanzado.

6. Objetivo general

Identificar lincRNA como marcadores moleculares de predicción de respuesta a la quimioterapia neoadyuvante en pacientes mexicanas con cáncer de mama localmente avanzado mediante análisis del transcriptoma.

6.1 Objetivos particulares

1. Seleccionar datos de secuenciación masiva en paralelo de RNA (RNA-Sec) que cumplan con los criterios de calidad para análisis bioinformático.
2. Determinar los perfiles de expresión diferencial de los lincRNA en muestras de pacientes con cáncer de mama localmente avanzado, considerando la respuesta patológica.
3. Realizar la caracterización *in silico* de los lincRNA seleccionados.
4. Realizar la validación molecular a través de análisis de expresión de los lincRNA seleccionados en líneas celulares y muestras de pacientes mexicanas mediante RT-PCR en tiempo real.
5. Determinar la especificidad y sensibilidad de los lincRNA candidatos como biomarcadores predictivos de la respuesta a la quimioterapia neoadyuvante.

7. Metodología

7.1. Etapa A. Recolección de muestras y obtención de datos de secuenciación

7.1.1. Criterios de inclusión de las pacientes

La Unidad de Investigación Biomédica en Cáncer del INCan cuenta con un banco de tumores que incluye 350 biopsias de pacientes diagnosticadas con CaMa y que forman parte de la población de ingresos en el instituto. En el presente estudio (integrado al proyecto *Cátedras Conacyt*, número 930) se incluyeron aquellas muestras correspondientes a pacientes clasificadas con CMLA del subtipo molecular Luminal B, en los estadios IIA y IIIB, positivas y negativas a la expresión de HER2, que no presentaran metástasis ni se les hubiera administrado tratamiento al momento de la toma de la biopsia, que ocurrió en el periodo 2012-2015 (Figura 12). En total se analizaron 28 muestras de pacientes (**Apéndice A**, Tablas A.1 y A.2), que firmaron con anterioridad el consentimiento informado aprobado por el comité de ética del INCan (012/048/OMI) (CB/806). El protocolo de investigación de la presente tesis se encuentra registrado ante los comités de investigación y ética del INCan con el número (018/055/DII) (CEI/1302/18).

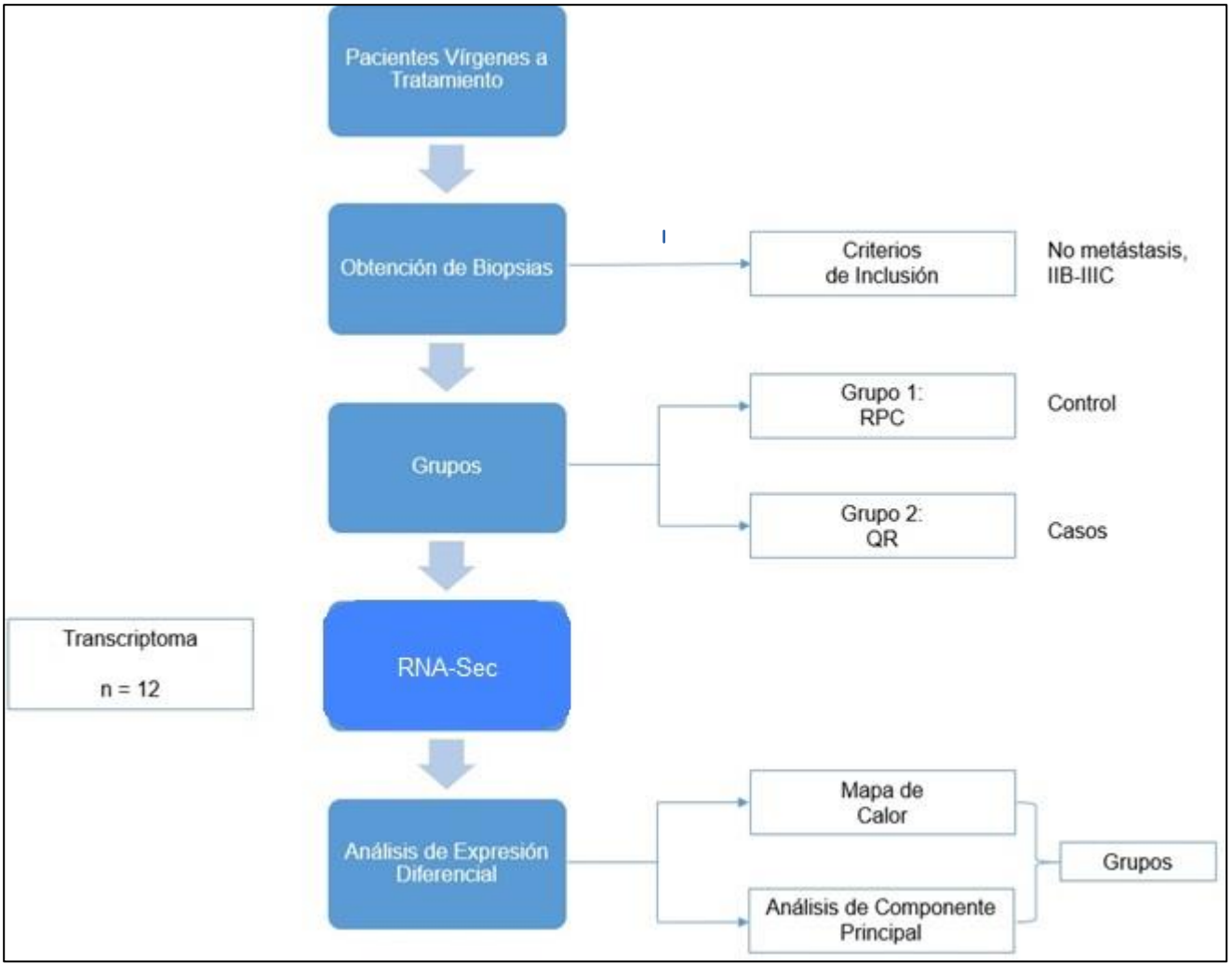


Figura 12. Flujo de trabajo para identificar los perfiles de expresión de lincRNA con valor predictivo en la respuesta a QTNeo mediante RNA-Sec. El análisis del transcriptoma fue realizado a un conjunto de pacientes del INCan que cumplieran con los criterios de inclusión establecidos en este estudio. Al seleccionarse a las candidatas, se extrajo RNA de tejido fresco de la biopsia tumoral previa a la administración de la QTNeo. Después de obtener los resultados del seguimiento clínico de respuesta al tratamiento, se agruparon las muestras de pacientes como controles si presentaron RPC a la QTNeo, y como casos si presentaron resistencia. Posteriormente, se procesó el RNA total de las muestras mediante secuenciación masiva de RNA (RNA-Sec), y los resultados se analizaron con herramientas bioinformáticas para corroborar la formación de grupos entre los casos y controles.

7.1.2. Purificación de RNA total de las muestras

Se purificó el RNA total de las biopsias de las 28 pacientes con el uso del kit AllPrep de QIAGEN (Cat. No. 80204), para la purificación simultánea de RNA, DNA y miRNA a partir de tejido fresco. La calidad y la concentración del RNA se determinó con el bioanalizador Tape Station 2200 (Agilent Technologies) y se almacenó a -80ºC, hasta su posterior uso.

7.1.3. Secuenciación masiva en paralelo de RNA (RNA-Sec)

A partir del RNA total se prepararon alícuotas de RNA de 12 de las 28 muestras, con 1.2 µg de RNA a una concentración de 100 ng/µl por muestra. El experimento de secuenciación fue de extremos pareados, con una profundidad de 20-30 millones de lecturas, utilizando la plataforma HiSeq 2500 de la compañía Illumina. Las alícuotas se entregaron al Laboratorio de Genómica de la Unidad de Red de Apoyo a la Investigación (RAI), en el Instituto Nacional de Ciencias Médicas y Nutrición "Salvador Zubirán" para su secuenciación, que estuvo a cargo del Dr. Inti Alberto de la Rosa.

7.2. Etapa B. Análisis bioinformático y validación experimental

7.2.1. Análisis bioinformático

7.2.1.1. Análisis de expresión diferencial

Para lograr los objetivos antes mencionados se inició comprobando la calidad de los datos de secuenciación. Posteriormente, se realizó el alineamiento de las lecturas con el genoma de referencia (hg38) y la cuantificación de transcritos no codificantes. Para el análisis de expresión diferencial se construyeron dos grupos de pacientes: (1) el grupo QR que contiene 7 muestras subtipo Luminal B resistentes a la QTNeo, y (2) el grupo RPC, que está compuesto de 4 muestras sensibles al mismo tratamiento y se definió como el grupo control. Del resultado de este análisis se seleccionaron 10 lincRNA sobreexpresados y 10 lincRNA subexpresados (Figura 13).

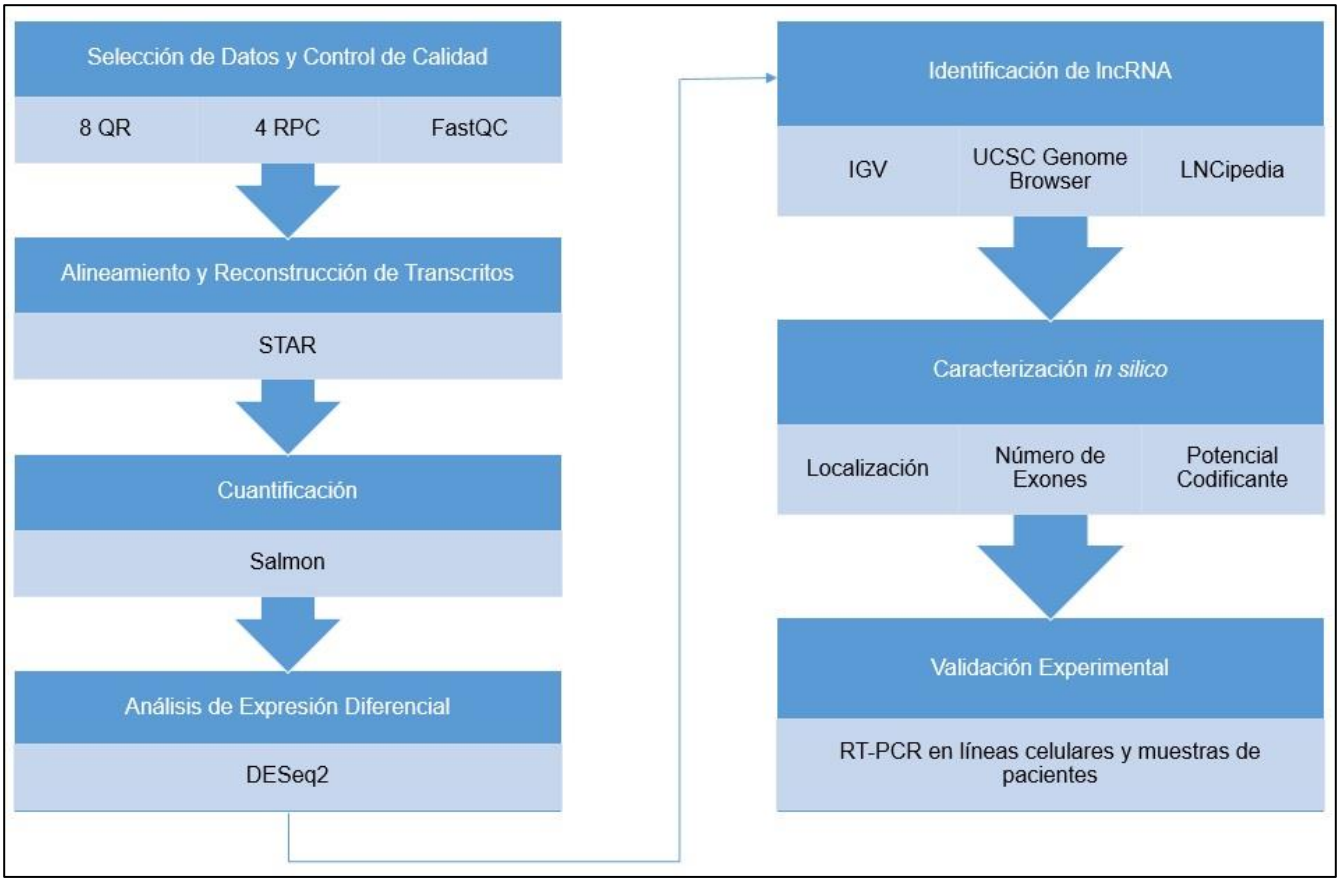


Figura 13. Estrategia bioinformática y experimental para identificar los perfiles de expresión de lincRNA en muestras de pacientes con CMLA. El flujo de trabajo consistió en obtener muestras de pacientes vírgenes a tratamiento que pertenecen al banco de tumores de CaMa del INCan. Se purificó RNA total de las mismas y se secuenciaron por RNA-Sec. Después, se analizaron los resultados de secuenciación por métodos bioinformáticos, realizando un análisis de expresión diferencial para identificar a los lincRNA diferencialmente expresados en pacientes con resistencia a QTNeo, mediante la construcción de un mapa de calor y un análisis de componente principal, para determinar la formación de grupos entre las pacientes incluidas en el estudio.

7.2.1.1.1. Análisis de calidad de los datos de secuenciación

El análisis se llevó a cabo con los archivos FastQ obtenidos de la secuenciación. Este tipo de archivos se caracterizan por contener las secuencias en formato FASTA con un carácter asociado que indica la calidad por cada base secuenciada. Además, incluye un identificador de la secuencia y la información del organismo de origen. La interpretación de los caracteres de calidad se realiza utilizando la herramienta *FastQC Read Quality Report* del servidor Galaxy, con los parámetros preestablecidos por la misma[75].

El reporte genera un análisis de 11 parámetros de calidad que pueden ser calificados como calidad aceptable, mala calidad, o inaceptable. Los parámetros se muestran en la Tabla 1.

Tabla 1. Parámetros de calidad evaluados en el reporte

Parámetros de calidad	Calificación	Valor de calidad
Cuadro de estadísticos básicos	Aceptable	Secuencias de baja calidad = 0
Calidad de la secuencia por base	Aceptable	X > 20Q
Puntaje de calidad por secuencia	Aceptable	Ausencia de regiones de mala calidad
Contenido de bases en la secuencia	Aceptable	Distribución del puntaje promedio por secuencia mayor a 20
Contenido GC por base	Aceptable	Contenido homogéneo por cada nucleótido detectado
Contenido GC por secuencia	Aceptable	Distribución normal
Contenido N por base	Aceptable	Cero
Distribución de la longitud de la secuencia	Aceptable	Distribución de longitud mayor a 120 nt
Niveles de duplicación de la secuencia	Aceptable	Menor al total de secuencias analizadas
Secuencias sobrerrepresentadas/Contenido de adaptador	Aceptable	Cero
Contenido de Kmer	Aceptable	Cero

Para que un archivo de secuenciación se considere adecuado para su uso en un análisis bioinformático, es requisito que tenga calificación *Aceptable* en el cuadro de estadísticos básicos, además de tener una calificación *Aceptable* en la calidad de la secuencia por base (**Apéndice B**).

7.2.1.1.2. Cambio de formato de plataforma de secuenciación

Una vez analizada la calidad, los archivos tipo FastQ deben ser modificados con un cambio en el formato de la plataforma de secuenciación. La información del archivo FastQ está codificada de acuerdo con el tipo de secuenciador utilizado; y se han establecido formatos estándar de codificación dependiendo de la plataforma utilizada, como es el caso de Illumina. Para nuestros archivos, el cambio de plataforma consistió en modificar la codificación del puntaje de calidad asignado a cada secuencia después de la secuenciación, con la finalidad de homogeneizar los archivos en un formato estándar que permitió su procesamiento adecuado en las diferentes herramientas de Galaxy, evitando errores de interpretación del código de secuencias. Esto se llevó a cabo con la herramienta *FastQ Groomer,* particularmene el formato de cambio fue para la plataforma Sanger & Illumina 1.8+.

7.2.1.1.3. Alineamiento y mapeo de las secuencias

La herramienta *STAR* del mismo servidor es capaz de hacer una limpieza adicional de las lecturas de los datos de secuenciación, depurando del archivo secuencias con baja calidad. Esto sirve para asegurar la calidad del proceso de alineamiento de las lecturas al genoma. Adicionalmente mapea las secuencias y realiza un conteo preliminar de los transcritos, que es de utilidad para llevar a cabo el análisis de expresión diferencial. Los resultados de cuantificación corresponden sólo a los transcritos anotados en el genoma de referencia.

El archivo de mapeo para las secuencias fue el genoma de referencia hg38 proporcionado por la herramienta (*Homo sapiens (b38)*: hg38, que es la versión más actualizada del genoma humano) y el modelo genómico de *splicing* fue el archivo *gencode.v27(GRCh38.p10)*[82]. El tipo de experimento de RNA-Sec se estableció del tipo lecturas pareadas y los parámetros restantes fueron los prestablecidos en la herramienta.

7.2.1.1.4. Cuantificación de transcritos

Para la elaboración de los gráficos de caja (del inglés *boxplot*) y el mapa de calor por agrupación jerárquica se utilizó el archivo de cuantificación de transcritos. La unidad de medida estandarizada que se utilizó fue transcritos por millón (TPM). La herramienta *Salmon* cuantifica

los transcritos presentes en la muestra secuenciada utilizando el archivo tipo FastQ de la misma, y un genoma de referencia en formato FASTA (archivo *Homo_sapiens.GRCh38.ncrna.fa.gz)*[79]*.* El análisis se llevó a cabo estableciendo que eran lecturas pareadas, indicando el primer archivo como lecturas sentido y se indicó el uso del genoma *gencode.v27(GRCh38.p10)*[80]. Dado que el archivo tipo FASTA de referencia corresponde sólo a los genes no codificantes, los resultados de cuantificación sólo consideran los RNA no codificantes presentes en la muestra.

7.2.1.1.5. Análisis de expresión diferencial

El análisis de expresión diferencial se realizó definiendo dos grupos de pacientes: (1) el grupo control (RPC) y (2) el grupo de casos (QR). Al ser un ***estudio de casos y controles anidado***, el efecto que se analiza es la sensibilidad a la QTNeo (i.e., si es sensible o no al tratamiento) y su asociación con los perfiles de expresión de lincRNA, por lo que el uso de un tejido calibrador (tejido normal mamario) no es apropiado en este análisis y no se incluyó. En la herramienta *DESeq2* los grupos se definen en niveles de factor: el nivel de factor o grupo QR incluyó los archivos de cuantificación de transcritos de las muestras codificadas para el estudio, como CM1, CM2, CM3, CM4, CM5, CM6 y CM12, de igual modo el nivel de factor o grupo RPC incluye las muestras CM7, CM9, CM10 y CM11. El orden de niveles de factor indica el orden de comparación entre los grupos. El grupo QR se identificó como el factor 1 y, en consecuencia, los resultados del análisis indican los lincRNA sobreexpresados, así como los subexpresados, en el grupo de casos.

La herramienta *DESeq2* incluye en sus resultados un reporte detallado que se compone por un archivo tipo PDF que contiene un gráfico de análisis de componente principal, un cuadro de organización jerárquica y un gráfico MA, lo cual permite la visualización gráfica de los perfiles de expresión de ncRNA en las muestras analizadas de las pacientes. Además, se genera un archivo tipo tabular con la lista de los ncRNA diferencialmente expresados en el grupo QR, que incluye la siguiente información por gen: (a) identificador de *Ensembl*; (b) promedio de conteo de lecturas; (c) tasa de cambio en la expresión (log2FC); (d) error estándar; (e) valor p; y (f) la tasa de descubrimientos falsos FDR (por sus siglas en inglés *False Discovery Rate*), que es un valor estadístico que permite conocer la tasa de descubrimientos que son incorrectos en una comparación múltiple y esto permite descartar aquellos lncRNA que no tengan cambios

estadísticamente significativos en su expresión. Debido a que el objetivo de este trabajo fue la identificación de perfiles de expresión diferencial de lincRNA, fue necesario seleccionar de los resultados el conjunto de lincRNA sobreexpresados y subexpresados, los cuales se asocian a un valor de FDR ≤ 0.05, y se generaron gráficos de análisis de componente principal y mapas de calor adicionales.

7.2.1.1.6. Elaboración de gráficos

Los gráficos de caja y mapa de calor por agrupación jerárquica se elaboraron con el paquete para la interfaz *R Studio ggplot2*, mediante el uso de las funciones *barplot, boxplot* y *heatmap*, utilizando los parámetros prestablecidos por la interfaz.

7.2.1.1.7. Selección de datos del análisis de expresión diferencial

Para la elección de lincRNA candidatos se filtraron los resultados del análisis de expresión diferencial utilizando la interfaz *R Studio*, a partir del cual se obtuvo un subconjunto de información de los lincRNA. El subconjunto se generó extrayendo el listado de identificadores de GENCODE de lncRNA y posteriormente se identificaron los lincRNA utilizando la base de datos de ncRNA de *Ensembl*.

Se seleccionaron aquellos lincRNA que tuvieran asociados valores de FDR ≤ 0.05, y que adicionalmente cumplieran con tener asociados valores de log2FC (indicador que se define como el logaritmo base 2 de la tasa de cambio de la expresión del transcrito, del inglés *Fold Change*) mayores o iguales al punto de corte, que se estableció en 1.5 para los lincRNA sobreexpresados y en -1.5 para los subexpresados. Finalmente, se eligieron los 10 lincRNA sobreexpresados y los 10 lincRNA subexpresados que fueran estadísticamente significativos.

7.2.1.2. *Caracterización in silico de los lincRNA candidatos*

Para los 10 lincRNA sobreexpresados se buscó cada una de las características de la Tabla 2

Tabla 2. Características de los lincRNA en la caracterización *in silico.*

Concepto	Parámetro
Localización genómica (con base en el genoma hg38).	Regiones intergénicas
Longitud del transcrito.	200 pb < Longitud
Número de exones.	1-3
Identificación de isoformas (si existen, identificar aquellas presentes en la muestra).	Presencia
Identificación del gen adyacente al lincRNA.	Presencia
Determinación del potencial codificante.	X ≤ 0
Identificación de las marcas post-traduccionales de histonas	Presencia/Ausencia
	H3K4me1
	H3K4me3
	H3K9me3
	H3K27ac
	H3K27me3
	H3K36me3
Buscar asociaciones de los lincRNA con funciones o patologías por *Gene Ontology.*	Presencia/Ausencia

Para los primeros 5 puntos se utilizó la información de las bases de datos *Ensembl*[79] y *LNCipedia*[81]. La determinación del potencial codificante se realizó con la herramienta *Coding Potential Calculator*[82]*,* disponible en línea, que clasifica a los transcritos de acuerdo a la Tabla 3.

Tabla 3. Clasificación de transcritos de acuerdo a su potencial codificante.

Tipo de Transcrito	Valor del Potencial Codificante (PC)
Codificante	0 < PC
No codificante	PC ≥ 0

Se analizó la secuencia de la isoforma más abundante en las muestras de pacientes analizadas por RNA-Sec de cada transcrito, que se obtuvo de la base de datos *LNCipedia*[81].

Para la caracterización de los factores transcripcionales y proteínas asociadas a la región promotora de los lincRNA candidatos, así como de las marcas de cromatina, las lecturas de RNA-Sec y el posicionamiento de la RNA PolII, se utilizaron las bases de datos *UCSC Genome Browser*[83] y *WashU Epigenome Browser*[84], en la línea celular MCF-7 y en fibroblastos mamarios humanos. Para obtener información de *WashU Epigenome Browser,* fue necesario conocer la posición de los lincRNA candidatos en el genoma hg19, ya que no está actualizado con el genoma hg38. Además, se utilizaron las herramientas *BDGP*[85] (en *Drosophila melanogaster*) y *Promoter2.0*[86] (en vertebrados), para determinar la existencia de posibles regiones promotoras en las secuencias de nucleótidos analizadas. En ambas herramientas, puntajes cercanos a 1.0 indican que la zona es una posible región promotora.

En el caso de las marcas de cromatina, no existe información en las bases de datos para todas las modificaciones post-traduccionales de histonas antes mencionadas para todos los lincRNA en la línea celular MCF-7 (Clave GSE31755, Tabla 4) y para los fibroblastos mamarios humanos (Clave GSE16368, Tabla 1). Los resultados indican en general el enriquecimiento de las marcas H3K4me1 (asociada a potenciadores de la transcripción), H3K4me3 (asociada a sitios promotores) y la marca H3K27ac (asociada a activación de la transcripción). Las demás marcas se incluyeron de acuerdo a la disponibilidad de la información en cada base de datos.

Tabla 4: Claves de acceso en GEO Datasets para la información pública consultada en WashU Epigenome Browser.

Elemento de análisis *in silico*	**Espécimen**	**Clave GEO**	**Tipo de Experimento**	**Plataforma**
H3K4me1	Fibroblastos mamarios humanos, sanos	GSM1127065	ChIP-Seq	Illumina HiSeq 2000
H3K4me3	Fibroblastos mamarios humanos, sanos	GSM1127085	ChIP-Seq	Illumina HiSeq 2000
	MCF-7	GSM945269	ChIP-Seq	Illumina Genome Analyzer
H3K36me3	Fibroblastos mamarios humanos, sanos	GSM1127068	ChIP-Seq	Illumina HiSeq 2000
	MCF-7	GSM970217	ChIP-Seq	Illumina Genome Analyzer
H3K27ac	Fibroblastos mamarios humanos, sanos	GSM1127064	ChIP-Seq	Illumina HiSeq 2000
	MCF-7	GSM945854	ChIP-Seq	Illumina Genome Analyzer
H3K9me3	MCF-7	GSM945857	ChIP-Seq	Illumina Genome Analyzer
H3K27me3	Fibroblastos mamarios humanos, sanos	GSM1127134	ChIP-Seq	Illumina HiSeq 2000
	MCF-7	GSM970218	ChIP-Seq	Illumina Genome Analyzer

RNA-Seq	Fibroblastos mamarios humanos, sanos	GSM1127100	RNA-Seq	Illumina HiSeq 2000
	MCF-7	GSM767851	RNA-Seq	Illumina Genome Analyzer IIx
Input	Fibroblastos mamarios humanos, sanos	GSM1127089	ChIP-Seq	Illumina HiSeq 2000
	MCF-7	GSM970217	ChIP-Seq	Illumina Genome Analyzer

Para completar la caracterización *in silico* se llevó a cabo la asociación a funciones o patologías de cada lincRNA utilizando la base de datos *FARNA*[87], que incluye un compendio de estudios de asociación o correlación de ncRNA con procesos celulares y diferentes patologías basado en análisis de coexpresión de genes codificantes y genes no codificantes. Para este estudio, se identificaron sólo las funciones asociadas a vías de señalización intracelular en la progresión y regulación del ciclo celular, apoptosis y supervivencia celular. La correlación con patologías se restringió a los estudios en cáncer y CaMa.

7.2.1.3. Determinación del perfil de expresión transcripcional de lincRNA en líneas celulares de CaMa

Se realizó la búsqueda de resultados de secuenciación por RNA-Sec de las líneas celulares MCF-10A, MCF-7, BT474 y MDA-MB-231 utilizando la base de datos GEO Datasets[92], para obtener la clave del proyecto (Tabla 5). Con esa información, por medio de la base de datos del Archivo Europeo de Nucleótidos (ENA, del inglés *European Nucleotide Archive)*[88] se descargaron los archivos en el servidor Galaxy, donde se procesaron con el mismo flujo de trabajo, tal y como se mencionó previamente para las muestras de pacientes. La selección de archivos se llevó a cabo considerando que cada resultado fuera la secuenciación de cultivos celulares utilizados como control y que no hubiesen sido administrados con ningún tipo de vehículo, disolvente o control de vector.

Tabla 5. Claves de acceso en GEO Datasets para los archivos de RNA-Sec de líneas celulares.

Línea Celular	Clave de GEO	Clave de SRA	Tipo	Plataforma de Secuenciación
MCF-10[a]	GSM1172882	SRX317727	Paired-end	Illumina Genome Analyzer IIx
	GSM1897320	SRX1293333		Illumina HiSeq 2000
	GSM1915044	SRX1361306		Illumina Genome Analyzer IIx
MCF-7	GSM2072527	SRX1603568		Illumina HiSeq 2000
	GSM2072571	SRX1603615		
	GSM2072572	SRX1603616		
BT474	GSM1172853	SRX317702		Illumina Genome Analyzer IIx
	GSM1466928	SRX671583		Illumina HiSeq 2000
	GSM1897280	SRX1293293		
MDA-MB-231	GSM2242132	SRX1960593		Illumina HiSeq 2000
	GSM2791584	SRX3210867		
	GSM2791576	SRX3210859		

7.2.2. Validación experimental

7.2.2.1. Purificación de RNA total de líneas celulares

Para estandarizar y validar el método de cuantificación de la expresión de los lincRNA candidatos se cultivaron las líneas celulares MCF-10A, MCF-7, BT474 y MDA-MB-231, y se obtuvo el RNA total de cada una de ellas, comprobando la calidad del mismo y determinando su concentración.

7.2.2.1.1. Cultivo de las líneas celulares

Las líneas celulares que se usaron en este trabajo son todas derivadas de tejido mamario, y se cultivaron en condiciones libres de antibiótico.

- **MCF-10A** (CRL-10317). Proveniente de tejido de glándula mamaria de *Homo sapiens* (humano), cuyo origen es un paciente con fibrosis quística, por lo que se considera una línea celular transformada. Su morfología es epitelial. Su crecimiento se llevó a cabo en medio basal de crecimiento de células mamarias epiteliales (MEBM), suplementado con suero fetal bovino (SFB) al 10%, en condiciones de atmósfera de CO_2 al 5%, a 37ºC.

- **MCF-7** (HTB-22). Proveniente del tejido de glándula mamaria de *H. sapiens*, derivada de adenocarcinoma en un sitio de metástasis. Se considera una línea celular tumorigénica, de morfología epitelial, que corresponde al subtipo molecular luminal A. Su crecimiento se llevó a cabo en medio mínimo esencial Eagle's (EMEM) suplementado con SFB al 10% e insulina recombinante (0.01 mg/mL), en condiciones de atmósfera de $C0_2$ al 5%, a 37ºC.

- **BT474** (HTB-20). Proveniente de tejido de los ductos de la glándula mamaria de *H. sapiens*, derivada de un carcinoma. Se considera una línea celular tumorigénica, de morfología epitelial, que corresponde al subtipo molecular luminal B. Su crecimiento se llevó a cabo en medio 46-X Hybri Care suplementado con SFB al 10% y bicarbonato de sodio en concentración de 1.5 g/L, en condiciones de atmósfera de CO_2 al 5%, a 37ºC.

- **MDA-MB-231** (HTB-26). Proveniente de tejido de la glándula mamaria de *H. sapiens*, derivada de adenocarcinoma en un sitio de metástasis. Se considera una línea celular tumorigénica, de morfología epitelial, que corresponde al subtipo molecular triple negativo. Su crecimiento se llevó a cabo en medio Leibovitz's L-15 suplementado con SFB al 10%, en condiciones de atmósfera libre de CO_2, a 37ºC.

7.2.2.1.2. Purificación de RNA total

De los cultivos de las líneas celulares MCF-10A, MCF-7, BT474 y MDA-MB-231, realizados en botellas de cultivo de 25 cm^2, se lisaron las células utilizando el reactivo de trabajo TRIzol Reagent (Ambion, Life Technologies, de Thermo Fisher Scientific). Se retiró el medio de cultivo de cada caja y se procedió a lavar con amortiguador fosfato salino (PBS, por sus siglas en inglés), para después añadir 1 mL de TRIzol, distribuyéndolo en toda la superficie de la caja de manera homogénea, y se almacenó la mezcla en un tubo eppendorf de 1.5 mL.

Para la extracción del RNA total se añadió a la mezcla con TRIzol 0.2 mL de cloroformo, se agitó vigorosamente y se centrifugó a 14,000 revoluciones por minuto (RPM), durante 15 minutos (min) a 4ºC, conservando la fase acuosa (o fase superior) y se adicionó 1 mL de isopropanol, que posteriormente se incubó por 20 min, y se centrifugó a 14,000 RPM por 15 min a 4ºC, conservando la pastilla (centrífuga Eppendorf).

La pastilla, que corresponde al RNA extraído, se lavó con etanol al 70% y se centrifugó a 14,000 RPM por 5 min a 4ºC, dos veces. Se retiró el sobrenadante y se dejó secar la pastilla a temperatura ambiente por 5 minutos. Finalmente, la pastilla se eluyó en 50 µL de agua con dietil pirocarbonato (DEPC) libre de RNAasas, se incubó a 50ºC durante 10 min, y se colocó en hielo (el RNA extraído se conservó a -80ºC).

7.2.2.1.3. Cuantificación de RNA

Cuantificación con NanoDrop: La cuantificación se llevó a cabo con el dispositivo NanoDrop (Acceso Lab, de Thermo Fisher Scientific), determinando primero la señal de absorbancia de una solución blanco (agua DEPC libre de RNasas) para ácidos nucleicos, especificando el procedimiento para RNA. Una vez determinado el blanco, se cuantificó 1 µL del RNA de cada línea celular, verificando los radios de absorbancia 260/280 (para determinar contaminación con DNA o proteínas) y 260/230 (que identifica contaminación con fenoles). El radio 260/280 se considera ideal cuando su valor es cercano a 2.0, mientras que el rango ideal del radio 260/230 es de 2.0 a 2.2.

7.2.2.1.4. Determinación de la intensidad de la calidad e integridad del RNA con el uso del Bioanalizador

Se cuantificó y se determinó la calidad del RNA extraído de cada línea celular mediante el uso del bioanalizador TapeStation 2200 (Agilent Technologies), preparando alícuotas a una concentración de 100 µg/µL. Para lograrlo, se añadió 1 µL de cada alícuota de RNA a 5 µL de solución amortiguadora desnaturalizante de RNA ("RNA Screen Tape Sample Buffer" y "Screen Tape" para RNA, cat. 5190-6506, de Agilent Technologies), y la mezcla se incubó a 72ºC por 3 min, para después colocarla en hielo y cargarla al bioanalizador. El resultado de este análisis nos dio información sobre la integridad del RNA (RIN) y la concentración del RNA. Se considera que el RNA tiene buena calidad cuando el valor del RIN es igual o mayor a 7.0.

7.2.2.2. *Cuantificación de los transcritos tipo lincRNA candidatos*

Se realizó la cuantificación de la expresión relativa de los lincRNA seleccionados por PCR en tiempo real. Para la validación del método de cuantificación se determinó primero la expresión en las líneas celulares mencionadas y posteriormente se realizó en las 28 muestras de pacientes. El análisis de resultados se llevó a cabo por el método delta Cq (ΔCq), y la significancia estadística se determinó por un análisis tipo t-student a una cola, no pareada.

7.2.2.2.1. Diseño de oligonucléotidos

Utilizando la secuencia de cada lincRNA candidato (proveniente de LNCipedia[81]), se diseñaron pares de oligonucleótidos para PCR en tiempo real con el uso de la herramienta *Primer-Blast*[89]. Los parámetros de diseño fueron los pre-establecidos por la herramienta.

Se seleccionaron los pares de oligonucleótidos con mayor contenido de GC (%GC > 50\%), con amplicones de longitud entre 120 a 150 bases, posicionados en los exones, y que cumplieran con lo siguiente.

- Presencia de las bases G o C en el extremo 3' de cada oligonucleótido.
- Valor de auto-complementariedad menor o igual a 3.00.
- Valor de Tm cercano a 60ºC (58ºC-62ºC).
- Longitud de cada oligonucleótido no mayor a 24 bases.

Se corroboró *in silico* la existencia de un amplicón por cada par de oligonucleótidos con la herramienta en línea *In-Silico PCR* del servidor UCSC Genome Browser[83], y se solicitó su síntesis (Tabla 6).

Tabla 6. Información de los oligonucleótidos diseñados para los experimentos de PCR en tiempo real.

Nombre	Longitud (bases)	Longitud del amplicón (bases)
Seq ID 1	19	101
Seq ID 2	18	101
Seq ID 3	20	133
Seq ID 4	20	133

7.2.2.3. Síntesis de cDNA

Para la cuantificación de transcritos tipo lincRNA por PCR en tiempo real fue necesario obtener cDNA a partir del RNA extraído de las líneas celulares y de las muestras de pacientes.

1. ***Tratamiento con DNasa I***. El uso de la enzima DNasa I se justifica para garantizar que el RNA no esté contaminado con DNA genómico. El procedimiento consistió en preparar para cada alícuota de RNA un tubo de reacción que contuviera 1 µg de RNA, 1 µL de solución amortiguadora 10X para DNasa I con MgCl2, 1 µL de enzima DNasa I (50 U/µL) y agua calidad *Biología molecular* estéril (c.b.p un volumen de reacción de 10 µL). La mezcla se incubió a 37ºC por 40 min, y posteriormente se añadió 1 µL de ácido etilaminotetraacético (EDTA) 50 mM, incubando a 65ºC por 10 min. Finalmente se dejó reposar en hielo y se conservó a -80ºC (se utilizó el kit DNase I, RNase-free, ref. EN0525, molecular biology, de Thermo Fisher Scientific).

2. ***Tratamiento con transcriptasa reversa***. Para la síntesis de cDNA se mezclaron 10 µL de RNA tratado con DNasa con: 4.2 µL de agua bi-destilada, desionizada y estéril, 2 µL

de Solución Amortiguadora PCR 10X (II), 2 µL de *Random Primers*, 0.8 µL de dNTPs 10 mM y 1 µL de enzima Reverso-transcriptasa de alta capacidad (RT) a 50 U/µL (se utilizó el kit High Capacity cDNA Reverse Transcription, ref. 4368814, applied biosystems, de Thermo Fisher Scientific). La reacción se procesó en un termociclador como se muestra en la Tabla 7. Al final se diluyó el cDNA 1:4 en agua calidad *Biología molecular*.

Tabla 7. Programa de termociclador para la síntesis de cDNA.

Paso	**1**	**2**	**3**	**4**
Temperatura (°C)	25	37	80	4
Tiempo (min)	10	120	5	∞

3. ***Evaluación de la calidad del cDNA por PCR en tiempo real***: La calidad del cDNA obtenido se analizó con un procedimiento de PCR en tiempo real, con el uso del equipo QuantStudio 3 (applied biosystems, de Thermo Fisher Scientific), preparando la siguiente reacción: 5 µL de SYBR Green/ROX qPCR *Master Mix* (ref. K0223, molecular biology, de Thermo Scientific), 2.2 µL de agua calidad *Biología molecular*, 0.3 µL de oligonucléotidos 10 µM sentido y antisentido, 2.5 µL de cDNA (en el caso del control negativo [NTC], se añadió la misma cantidad de agua). Los oligonucleótidos sentido y antisentido generan un amplicón de la secuencia del gen constitutivo (Seq. ID. 3 y 4), con lo que se corrobora la adecuada amplificación de los productos de reacción. En este experimento, se procesó un duplicado +RT, un duplicado -RT y un duplicado de NTC y se estableció el programa de reacción en el termociclador, como se muestra en la Tabla 8. Se considera que la calidad del cDNA es buena cuando el valor del Cq es mayor o igual a 14 para la amplificación del gen constitutivo.

Tabla 8. Programa de termociclador para PCR en tiempo real.

Etapa de desnaturalización		**Etapa de Amplificación**			**Etapa de disociación**		
50°C	95°C	95°C	60°C	72°C	95°C	60°C	95°C
2 min	10 min	15 s	30 s	30 s	15 s	1 min	1 s

7.2.2.4. Validación del método de cuantificación de transcritos por PCR en tiempo real

La validación del método de cuantificación se realizó con el cDNA de la línea celular MCF-7. El procedimiento se llevó a cabo con la mismas características experimentales mencionadas en el apartado **Evaluación de la calidad del cDNA por PCR en tiempo real**, y se incluyó por cuadruplicado +RT, -RT y NTC para los oligonucleótidos del gen constitutivo y para el gen problema.

7.2.2.4.1. Determinación de la eficiencia de amplificación de los oligonucleótidos

La eficiencia de amplificación de lo oligonucleótidos se llevó a cabo realizando una PCR en tiempo real con las mismas condiciones de reacción mencionadas en el apartado **Evaluación de la calidad del cDNA por PCR en tiempo real**:

- Se realizaron diluciones seriales 1:10, 1:100, 1:1000 y 1:10000 del cDNA de la línea celular MCF-7, a partir de la dilución primaria 1:4.

- Para cada dilución se preparó un cuadruplicado técnico, tanto para lo oligonucleótidos del gen constitutivo como para los de la secuencia problema.

Los resultados de Cq se relacionaron linealmente con la dilución correspondiente, y se llevó a cabo una regresión lineal para la determinación del valor de la pendiente, que se relaciona directamente con el porcentaje de eficiencia. Para ello, se utilizó la herramienta disponible en Thermo Fisher Web Tools → qPCR Efficiency Calculator[95]. Se considera que un par de oligonucleótidos es eficiente y su uso es óptimo en una PCR en tiempo real cuando la eficiencia de amplificación de los oligonucleótidos se encuentra entre el 95% y el 105%.

7.2.2.4.2. Cuantificación relativa de transcritos mediante el método ΔΔCq

La cuantificación de transcritos se llevó a cabo con el método de cuantificación relativa bajo las mismas condiciones experimentales que las indicadas en el apartado **Validación del método de cuantificación de transcritos por PCR en tiempo real**, utilizando el cDNA de las líneas celulares MCF-10A, MCF-7, BT474 y MDA-MB-231, con cuadruplicados técnicos para

cada línea celular, tanto para los oligonucleótidos del gen constitutivo como para los del gen problema, además de incluirse cuadruplicados -RT y NTC para cada condición. La interpretación de los resultados de amplificación se llevó a cabo por el método ΔCq (Ecuación 1) y por el método ΔΔCq (Ecuación 2).

Ecuación 1: $\Delta Cq = Cq_{(gen\ problema)} - Cq_{(gen\ constitutivo)}$

Ecuación 2: $\Delta\Delta Cq = \Delta Cq_{(muestra\ problema)} - \Delta Cq_{(muestra\ control)}$
$= (Cq_{(gen\ problema)} - Cq_{(gen\ constitutivo)})_{muestra\ problema} - (Cq_{(gen\ problema)} - Cq_{(gen\ constitutivo)})_{referencia}$

Ecuación 3: Cuantificación relativa $\Delta\Delta Cq = 2^{-\Delta\Delta Cq}$

Se determinó el promedio de cuantificación relativa ΔΔCq para los cuadruplicados y se calculó el error estándar asociado. La cuantificación de expresión relativa en el caso de las líneas celulares fue realizada por ambos métodos (ΔCq y ΔΔCq), y en el caso de las muestras de pacientes, al no contar con un tejido calibrador la cuantificación se limita al método ΔCq.

7.2.2.5. *Análisis estadístico*

Por tratarse de un análisis de casos y controles, la literatura disponible en estudios similares propone determinar la significancia estadística de los marcadores seleccionados entre el grupo de casos QR y el grupo control RPC mediante el análisis con t-student no pareada a un extremo, con intervalo de confianza del 95%[55]. En el caso de la expresión en líneas celulares, se llevó a cabo un análisis ANOVA, seguido de una prueba de Tukey, considerando el mismo intervalo de confianza, utilizando la herramienta *One-Way ANOVA Calculator*[90].

Para determinar la especificidad y sensibilidad del valor predictivo de los candidatos se construyó una curva ROC ajustada utilizando la herramienta en línea del Hospital Johns Hopkins[91] La construcción de la curva se realiza proporcionando la información de cuantificación por el método ΔCq y el estatus clínico de las pacientes (si presentó quimiorresistencia o no). El ajuste de la curva se realiza con un intervalo del 95% de confianza.

Adicionalmente, se realizó una prueba de asociación estadística entre los niveles de expresión de los candidatos y la condición de respuesta a la QTNeo, mediante el cálculo de la razón de momios, a un intervalo de confianza del 95%, utilizando la herramienta en línea *MedCalc's Odds Ratio Calculator*[92].

8. Resultados

8.1. Análisis bioinformático

8.1.1 Selección de datos de RNA-Sec que cumplen con los criterios de calidad para realizar el análisis bioinformático

Para demostrar la asociación de los perfiles de expresión de lincRNA con la respuesta a la QTNeo, se llevó a cabo un análisis del transcriptoma de muestras de pacientes con CMLA a través de resultados de RNA-Sec. Con base en los criterios de inclusión descritos en la sección **Metodología**, se seleccionaron 12 muestras de pacientes del banco de tumores mamarios del INCan para incluirse en el protocolo. En la Tabla 9 se describen las principales características de este conjunto de muestras, de las cuales la mayoría de ellas corresponden al subtipo molecular Luminal B (n = 11), y sólo una pertenece al subtipo Luminal A. A partir de la información del seguimiento clínico, se construyeron dos grupos de estudio dependiendo de la respuesta patológica: el grupo de casos (QR), incluye a las pacientes que presentaron resistencia a la QTNeo (n = 8) y el grupo control (RPC), que contiene a las pacientes que resultaron ser sensibles a este tratamiento (n = 4); la inclusión de las muestras con respuesta parcial en ambos grupos se basó en la etapa clínica, la expresión de HER2 y el porcentaje de expresión de Ki67 (**Apéndice A**). En general, se observó que las pacientes con resistencia a la QTNeo son negativas a la expresión del receptor HER2. Por otro lado, la calidad del RNA de estas muestras se asoció en todos los casos a un valor de RIN mayor a 7.0, por lo que cumplieron con los criterios de aceptación para el presente estudio.

Tabla 9. Características clínicas relevantes en el estudio para el análisis de expresión de lincRNAs, y parámetros de calidad de la secuenciación.

Muestra	Subtipo Molecular	Respuesta Patológica	Expresión de HER2	RIN	Total de Lecturas (Millones)
Grupo de Casos (QR)					
CM-1	Luminal B	Resistencia	-	7.2	25
CM-2	Luminal B	Resistencia	+	7.4	41
CM-3	Luminal B	Resistencia	-	8.9	26
CM-4	Luminal B	Parcial	-	7.5	41
CM-5	Luminal B	Resistencia	-	7.9	38
CM-6	Luminal B	Resistencia	-	8.2	36
CM-8	Luminal A	Resistencia	-	-	-
CM-12	Luminal B	Parcial	+	9.1	25
Grupo Control (RPC)					
CM-7	Luminal B	Parcial	-	8.9	49
CM-9	Luminal B	Completa	+	8.5	28
CM-10	Luminal B	Completa	+	8.9	20
CM-11	Luminal B	Completa	-	8.6	17

Los resultados del procedimiento de secuenciación indican que, como se planificó, la cobertura de secuenciación para las muestras en general fue mayor a 20 millones de lecturas (Tabla 9). Por otro lado, el reporte de calidad muestra 11 parámetros que evalúan los puntajes de calidad asignados por el secuenciador a cada fragmento de RNA procesado. El resumen global de calidad se muestra en el Cuadro de estadísticos básicos con información general del proceso de secuenciación (Figura 14 A), representando el primer criterio de aceptación de los resultados, considerando que el número de secuencias detectadas con mala calidad debe ser cero. En el caso de las muestras de pacientes, los Cuadros de estadísticos básicos fueron

calificados como *Aceptables.* Por lo tanto, cumplen con los principales parámetros de calidad del proceso de secuenciación.

A)

Measure	Value
Filename	CM4_S45_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	41018770
Sequences flagged as poor quality	0
Sequence length	35-126
%GC	49

B)

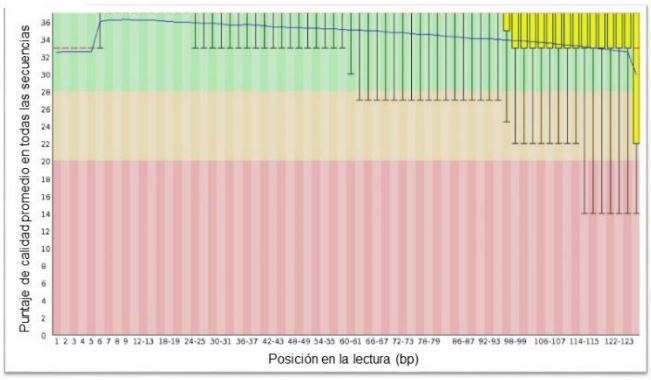


Figura 14. Parámetros básicos del reporte de calidad de los resultados de RNA-Sec. A) Se muestra el resumen principal del reporte de calidad, y contiene la información más relevante del proceso de secuenciación. Incluye datos como el identificador de la muestra, el tipo de archivo, la plataforma de secuenciación, el número total de secuencias analizadas, la cantidad de secuencias con baja calidad, el rango de longitud de los fragmentos y el porcentaje global de guaninas y citosinas (GC). B) Se muestra el gráfico de puntaje de calidad promedio de la secuenciación por cada fragmento, en el cual se distinguen 3 zonas de colores: la roja corresponde al intervalo de valores de calidad que son calificados como Inaceptables, la zona amarilla corresponde al puntaje de Mala calidad y la zona verde indica las lecturas de mejor calidad (Aceptable). Ambas figuras son representativas del reporte de calidad de todas las muestras de pacientes secuenciadas.

El siguiente parámetro determinante de la calidad es el puntaje global de calidad asignado por base en las secuencias. La calidad de las bases se analiza determinando el promedio del puntaje de calidad dependiendo de la posición en el fragmento y se considera *Aceptable* cuando el valor de puntaje de calidad es mayor o igual a 20Q, lo que se identifica como las regiones amarilla y verde (Figura 14 B). Para las 11 muestras analizadas, la calidad se califica como *Aceptable* en este parámetro, con lo que se determinó que los 12 archivos de secuenciación eran apropiados para su uso en los análisis bioinformáticos.

Los resultados de secuenciación para las 12 muestras de pacientes analizadas cumplieron con los criterios de calidad bioinformáticos, ya que en ningún caso se indicó la presencia de lecturas de mala calidad, además los puntajes de calidad por base fueron mayores a 20Q para todas las muestras de pacientes (**Apéndice B**). En conclusión, los reportes corroboraron que los resultados de RNA-Sec de las muestras analizadas cuentan con la calidad apropiada para su análisis bioinformático.

8.1.2. Determinación de los perfiles de expresión diferencial de lincRNA

Después de corroborar la calidad de los resultados de secuenciación, se realizó el análisis bioinformático de los mismos. Con el objetivo de establecer si existían lincRNA que presentaran diferencias en su expresión entre las pacientes del grupo de casos QR y el grupo control RPC, se realizó un análisis de expresión diferencial del transcriptoma. Los resultados mostraron que existe un conjunto de mRNA y ncRNA diferencialmente expresados entre ambos grupos. Sin embargo, al considerar ambos tipos de transcritos, un análisis de componente principal no permite distinguir entre el grupo de casos (QR) y el grupo control (RPC) (**Apéndice C**).

Para determinar si este efecto era producido por la expresión de los mRNA o los lncRNA, se llevaron a cabo análisis de expresión diferencial independientes para ambos tipos de transcritos, y se encontró que para el caso de los mRNA no era posible distinguir entre las pacientes del grupo de casos y las pacientes del grupo control (Figura 15 A). En cambio, el perfil de expresión de los lncRNA es capaz de definir al grupo de casos QR (Figura 15 B). En la Figura 15 se aprecia la similitud que existe entre las muestras de pacientes que presentan resistencia a la QTNeo en cuanto a la expresión de lncRNA. En este gráfico se observa que la muestra CM-8 se encuentra a una distancia mayor de todas las demás muestras del grupo de casos QR (azul), puesto que, a pesar de ser resistente a la QTNeo, pertenece al subtipo Luminal A, a diferencia de las otras que son Luminal B. Estos transcritos se han descrito como moléculas con especificidad de expresión en diferentes condiciones biológicas, y los resultados de este primer acercamiento sugieren la existencia de un conjunto de lncRNA que se relaciona con la resistencia a la QTNeo en pacientes mexicanas con CMLA subtipo Luminal, positivas y negativas a la expresión HER2. Esto sugiere que el perfil de expresión de los lncRNA es más especifíco de la resistencia a QTNeo respecto al perfil de mRNA en estas pacientes.

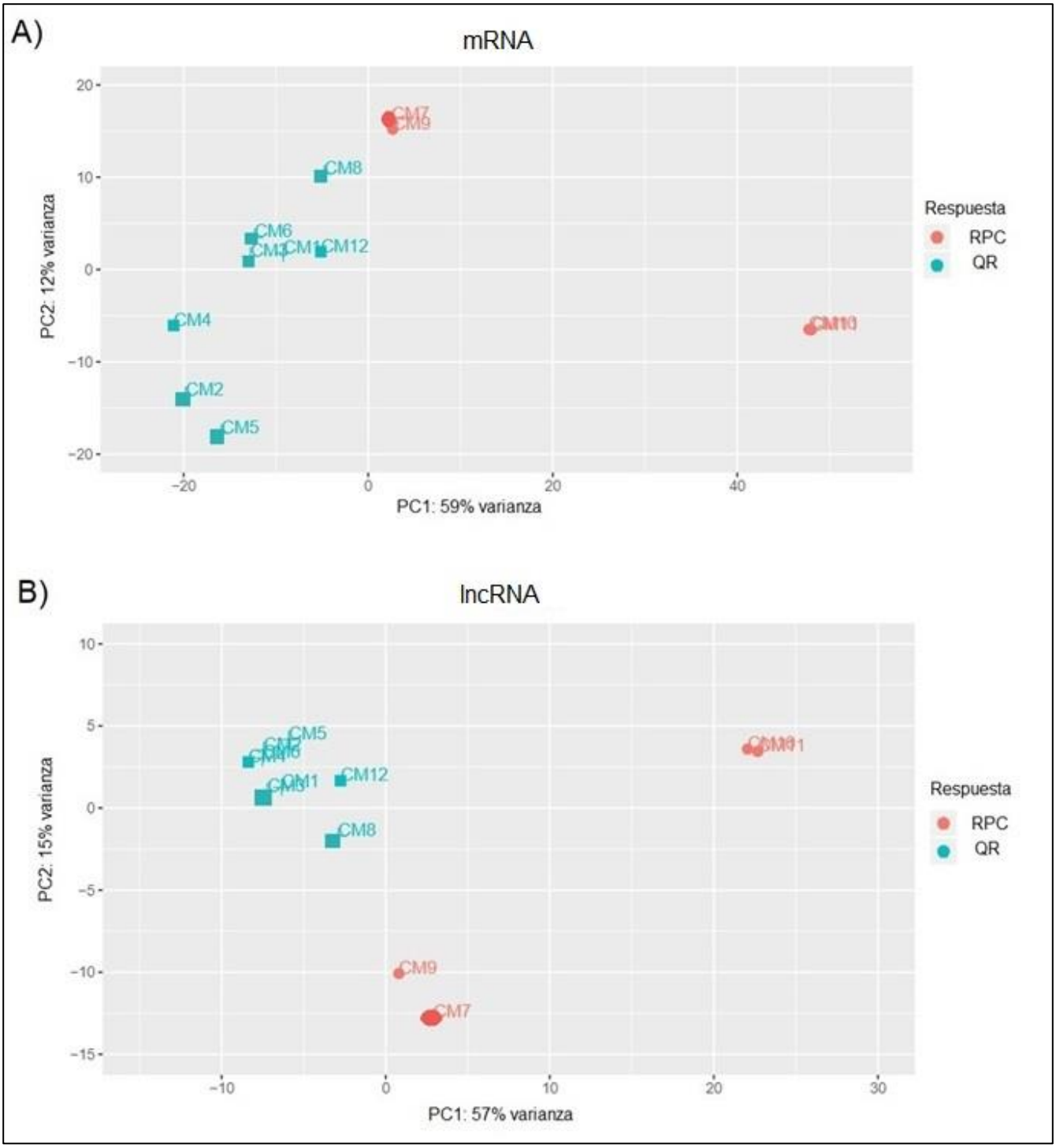


Figura 15. El perfil de expresión diferencial de los lncRNA permite la distinción entre grupos de pacientes. A) Gráfico del análisis de componente principal en el que se observa que la expresión de mRNA no permite la adecuada distinción entre las pacientes del grupo de casos QR (azul) y el grupo control RPC (rojo). B) Gráfico del análisis de componente principal en el que se observa que las muestras del grupo de casos QR presentan similitud respecto a los perfiles de expresión de lncRNA (azul), y se distinguen de aquellas que presentan la RPC (rojo); n= 12.

Debido a que el conjunto de lncRNA es capaz de definir al grupo de casos QR, se quiso conocer si el subconjunto de los lncRNA conformado por los lincRNA era suficiente para obtener el mismo resultado. Para ello, se realizó un análisis de expresión diferencial sólo considerando este subconjunto de transcritos. Se identificaron 17 lincRNA sobreexpresados y 57 lincRNA subexpresados de acuerdo al punto de corte establecido y que fueran estadísticamente significativos en el grupo de casos QR (Figura 16). Por tanto, existe un perfil de expresión de lincRNA que está diferencialmente expresado en el grupo de pacientes resistentes a la QTNeo.

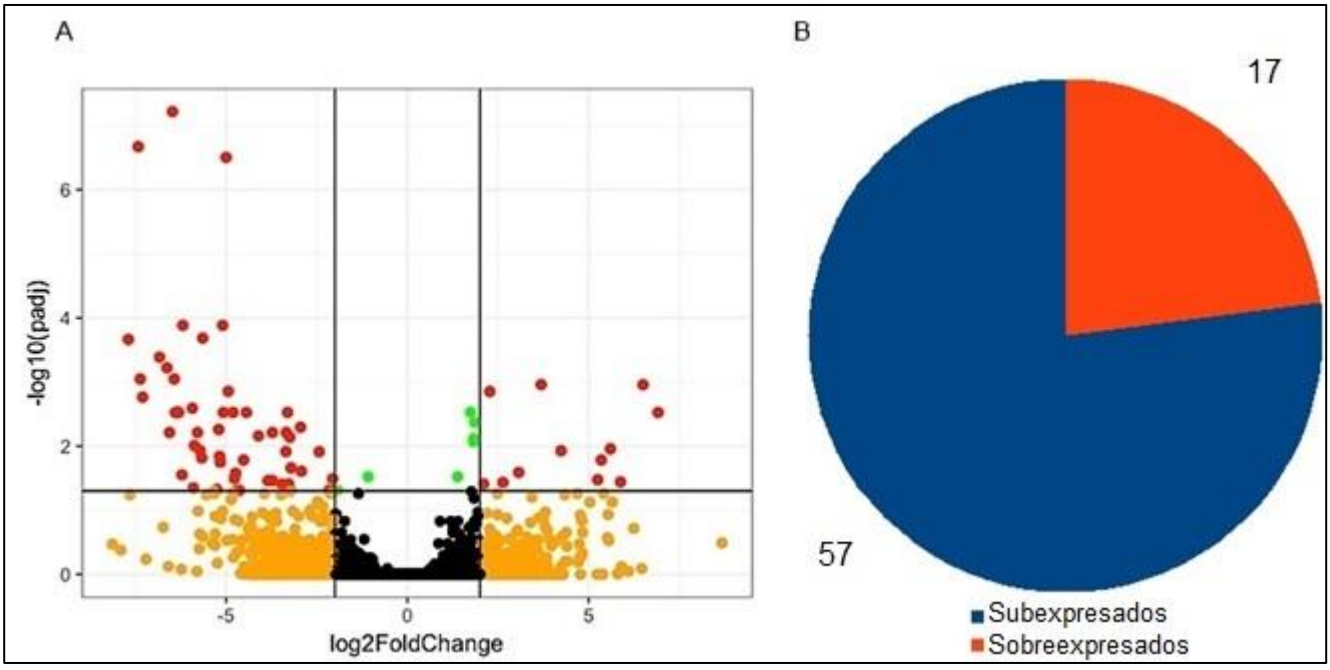


Figura 16. El análisis de expresión diferencial del transcriptoma de las muestras de pacientes identifica lincRNA sobreexpresados y subexpresados en el grupo QR.A) El gráfico de tipo volcán elaborado a partir de los resultados del análisis de expresión diferencial comprobó que existe un conjunto de lincRNA diferencialmente expresados en el grupo QR, que se ilustran en los cuadrantes superiores izquierdo y derecho del gráfico (rojo). B) Gráfico de pastel que ilustra a los 57 lincRNA subexpresados (azul) y los 17 sobreexpresados (rojo), de acuerdo a los criterios de expresión diferencial mencionados en la **Metodología**; FDR < 0.05**.**

Una vez identificado el perfil de lincRNA diferencialmente expresados en las pacientes del grupo de casos QR, se construyó un mapa de calor con el objetivo de visualizar la condición de expresión de cada lincRNA en las muestras de las pacientes (Figura 17). Las columnas del mapa de calor presentan el perfil de expresión particular de lincRNAs para cada paciente, y se observa que los perfiles de expresión para las muestras del grupo de casos QR son similares entre ellos, por lo que además, son capaces de agruparse jerárquicamente dentro del mapa

de calor, generando así dos conjuntos independientes de pacientes, que corresponden al grupo de casos QR y al grupo control RPC. El conjunto que contiene a las pacientes QR también incluye dos subconjuntos principales, uno de ellos está más cercano a las pacientes RPC y corresponde a aquellas pacientes que, a pesar de presentar respuesta parcial al tratamiento, se consideraron clínicamente sensibles al mismo, como si presentaran respuesta patológica completa (Tabla 9), por lo que los resultados de este análisis corresponden con las características clínicas antes mencionadas.

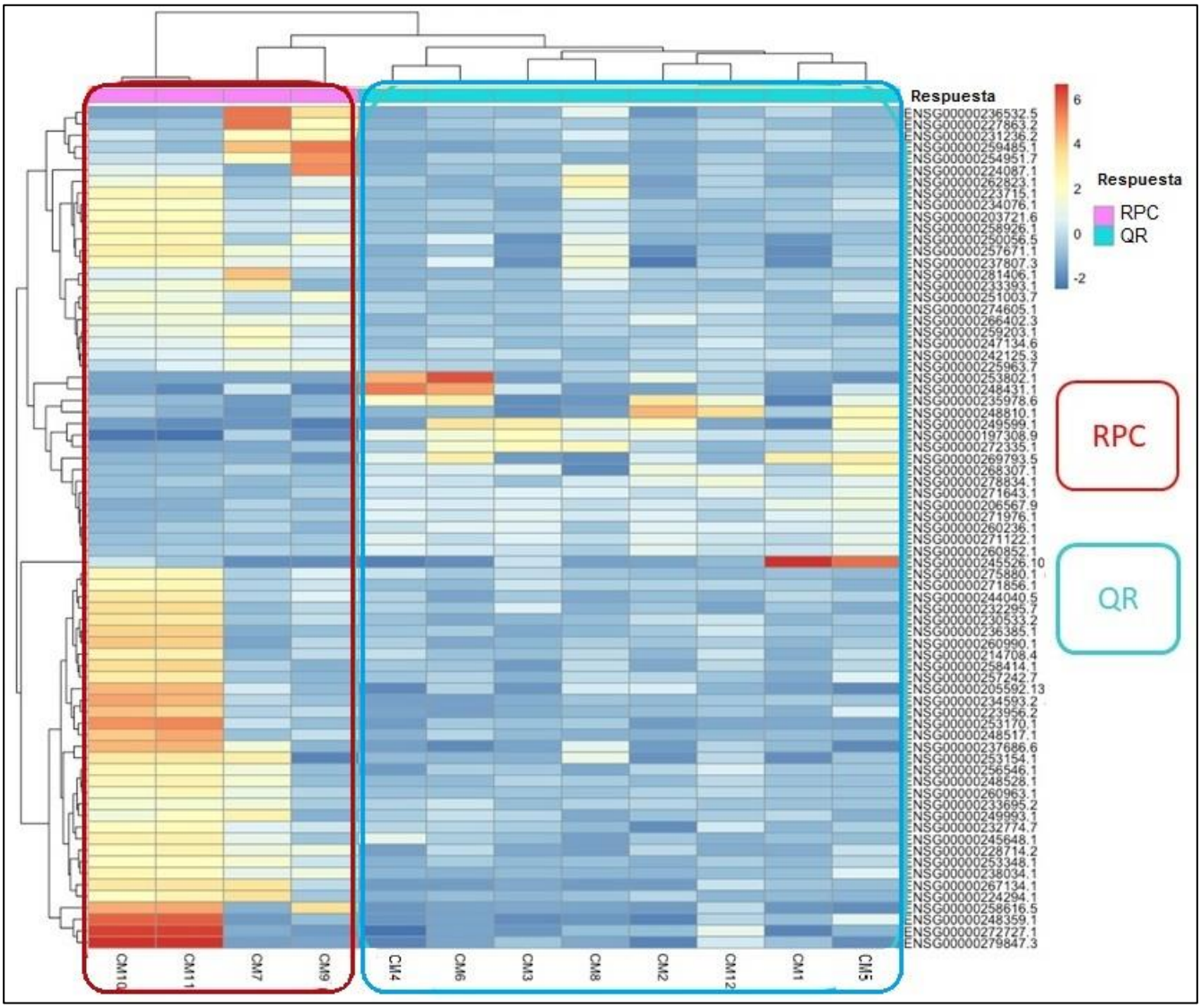


Figura 17. La agrupación jerárquica de las muestras respecto a la expresión de lincRNA permite la distinción entre los grupos QR y RPC. La construcción del mapa de calor permite visualizar el perfil de expresión individual de cada muestra de CMLA analizada (columnas), con la ayuda de un código de colores, donde el azul representa los lincRNA subexpresados y el color rojo los sobreexpresados (filas). El análisis por agrupación jerárquica en este gráfico distingue entre las muestras que presentan resistencia a la QTNeo (cuadro azul) y aquellas con RPC (cuadro rojo) (n = 12); FDR < 0.05. La escala de colores está basada en el valor numérico log (TPM) de cada lincRNA.

Con la construcción del mapa de calor fue posible identificar diferentes conjuntos de lincRNA que pueden tener asociación con la respuesta a QTNeo, por lo anterior, a nosotros nos interesó enfocarnos en aquellos lincRNAs que exclusivamente se sobreexpresan en el grupo de pacientes con resistencia a la QTNeo (Figura 17).

En suma, el análisis de los perfiles de expresión de lincRNA entre los dos grupos permitió la identificación de los lincRNA diferencialmente expresados (basados en el valor log2FC) en el grupo de casos QR. Con esa información, se seleccionaron algunos de ellos para su caracterización *in silico*, para finalmente elegir aquellos que pudieran ser validados experimentalmente.

8.1.3. Caracterización in silico de los lincRNA con expresión diferencial en el grupo QR

Con el objetivo de identificar lincRNA con potencial valor predictivo en la respuesta a QTNeo, se seleccionaron los 10 lincRNA con mayores valores de tasa de cambio, lo cual determina cuáles son los principales lincRNA sobreexpresados en las pacientes que presentan resistencia a la QTNeo, y se caracterizaron por métodos *in silico*. Aunque este estudio se dirigió principalmente a la búsqueda de lincRNA sobreexpresados en la resistencia a QTNeo, también se incluyó el análisis de lincRNA subexpresados (información no mostrada).

Con el propósito de conocer las características principales de cada lincRNA diferencialmente expresado en el grupo de casos (QR) se determinó el símbolo, la posición genómica, la longitud de la isoforma canónica, la cadena a partir de la cual se sintetiza y las funciones por ontología de genes, ya que no existen reportes científicos sobre éstas (Tabla 10). Además, se verificó la localización a menos de 5 kb de un gen codificante y el número de exones, ya que junto con los parámetros anteriores son capaces de definir a un lincRNA. Los resultados muestran que los lincRNA con sobreexpresión en el grupo de casos (QR) se sintetizan en su mayoría a partir de la cadena antisentido, su longitud se encuentra en el intervalo de 340 bases a 4.5 kb y su posición es adyacente a un gen codificante (Tabla 10). Esto es indicativo de que todos los genes no codificantes identificados cumplen con los criterios de clasificación de los lincRNA.

Tabla10.Los 10 lincRNA sobre-expresados en el grupo de casos QR y sus principales caracteristicas

Símbolo	Posición	Longitud (pb)	*log2FC	FDR	Cadena	•Funciones
BMPR1B-AS1	chr4:94,743,800-94,757,533 B	511	6.91	0.003	Antisentido	Apoptosis
AC105999.2	chr8:40,298,741-40,311,261	881	6.50	0.001	Antisentido	Muerte celular Programada
LINC00461	chr5:88,507,546-88,684,808	3558	5.88	0.036	Antisentido	Vía TGF-β, actividad transcripcional del complejo SMAD2/SMAD3:4
AC005150.1	chr4:162,740,668-162,742,048	628	5.60	0.011	Sentido	Regulación de la expresión proteica
LINC02432	chr4:141,321,129-141,326,007	2839	5.25	0.033	Antisentido	Interacción con AGO2
lincRNA-ACR	—	2214	3.69	0.001	Antisentido	—
AC093297.3	chr5:44,826,076-44,828,592	2517	3.07	0.026	Sentido	-
LINC02560	chr19:58,400,221-58,400,679	459	2.63	0.037	Sentido	Asociado a receptores de membrana tipo tirosín-cinasa
AC022007.1	chr3:10,006,418-10,011,095	1897	2.27	0.001	Antisentido	Biogénesis de organelos, mantenimiento, apoptosis, mitogénesis, respuesta inmunológica
AC112220.4	chr3:33,795,688-33,796,950	1263	1.83	0.007	Antisentido	-

*Se muestran los 10 lincRNA con mayores valores de log2FC obtenidos del análisis de expresión diferencial, lo que se interpreta como sobre-expresión en el grupo QR. Se eligieron aquellos con un valor de FDR < 0.05, para considerar su sobre-expresión significativa.•Las funciones fueron identificadas por ontología de genes, con ayuda del servidor FARNA.

Por otro lado, la búsqueda de asociación funcional mostró que la mayoría de los lincRNA sobreexpresados están involucrados en procesos que regulan la proliferación celular, el ciclo celular, la regulación transcripcional y la muerte celular programada. Entre los reguladores de la proliferación celular se encuentra el lincRNA *LINC02560*, ya que se ha relacionado con la expresión de receptores membranales tipo tirosín-cinasa, cuya principal función es la activación de vías de señalización que regulan la proliferación celular, como lo son los receptores de la familia HER. El lincRNA *LINC00461* es uno de los tres genes con mayor valor de log2FC identificado en este análisis, y se ha asociado con la regulación del ciclo celular a través de las proteínas SMAD2/3, que afectan la vía TGF-β. Asimismo, algunos de los lincRNA identificados, como *AC005150.1* y *LINC02432*, se han relacionado con la regulación transcripcional a través de su interacción con proteínas. Finalmente, la muerte celular programada es la función que identificamos como la que más se relaciona con este conjunto de lincRNA, ya que en su regulación participan *AC105999.2, AC022007.1* y *BMPR1B*-AS1, que es el lincRNA con mayor valor de log2FC, y se ha asociado con el proceso de apoptosis; este lincRNA se encuentra adyacente al gen que codifica a la proteína receptora asociada a morfogénesis del hueso 1B (*BMPR1B*, por sus siglas en inglés). La función del gen codificante se asocia al desarrollo del tejido óseo durante el proceso embrionario, y su expresión es basal en tejido mamario normal, por lo que la sobreexpresión en tejidos tumorales sugiere su papel en el desarrollo de CaMa. El análisis por ontología de genes sugiere que es posible que los lincRNA identificados estén relacionados con el proceso carcinogénico, debido a que regulan procesos que se encuentran alterados en esta condición patológica.

Otra función identificada en el análisis anterior fue la regulación de la respuesta inmunológica, a la cual se asocian *AC022007.1* y el lincRNA que denominaremos lincRNA-ACR (RNA largo no codificante asociado a quimiorresistencia). De este lincRNA, se sabe que está involucrado en la diferenciación de los linfocitos T y que puede regular la expresión del gen codificante adyacente, que representa un factor transcripcional importante en linfocitos y en tejido mamario. Esta evidencia aunada a que este lincRNA no se expresa en el tejido mamario normal, sugieren que puede estar involucrado en el proceso carcinogénico en CaMa, y regular la respuesta a QTNeo, por lo que la investigación se enfocó principalmente en la caracterización de este transcrito no codificante.

En resumen, los resultados del análisis de expresión diferencial permiten la identificación de un perfil de expresión de lincRNA compuesto por 10 candidatos que sirve para predecir la respuesta a la QTNeo, dado que su expresión define al grupo de pacientes que presentaron resistencia al tratamiento. Entre ellos, el lincRNA lincRNA-ACR se consideró el candidato más factible para su caracterización *in silico* y validación experimental en muestras de pacientes, con la finalidad de determinar su potencial predictivo en la respuesta a la QTNeo en pacientes mexicanas con CMLA.

8.1.4. Determinación de la localización genómica y las marcas de cromatina asociadas al lincRNA-ACR

La clasificación de los lncRNA está basada principalmente en la localización genómica. En particular, los lincRNAs se caracterizan por encontrarse en regiones intergénicas dentro del genoma y por no componerse de más de 3 exones. Como parte de los resultados de la sección anterior, se encontró que el gen lincRNA-ACR se localiza en el cromosoma 10 y su longitud es de aproximadamente 2 kb. Por su posición, sabemos además que el gen codificante más cercano se encuentra al menos a 1 kb de distancia, por lo que la región en que se ubica es intergénica. El lincRNA-ACR se sintetiza a partir de la cadena antisentido, se compone de 2 exones, y el transcrito tipo se caracteriza por contener ambos exones (Figura 18 A). Este lincRNA se caracteriza por tener 4 isoformas (Figura 18 B). Con todo lo anterior, se corroboró que particularmente el lincRNA-ACR es un transcrito que tiene las características principales del biotipo lincRNA.

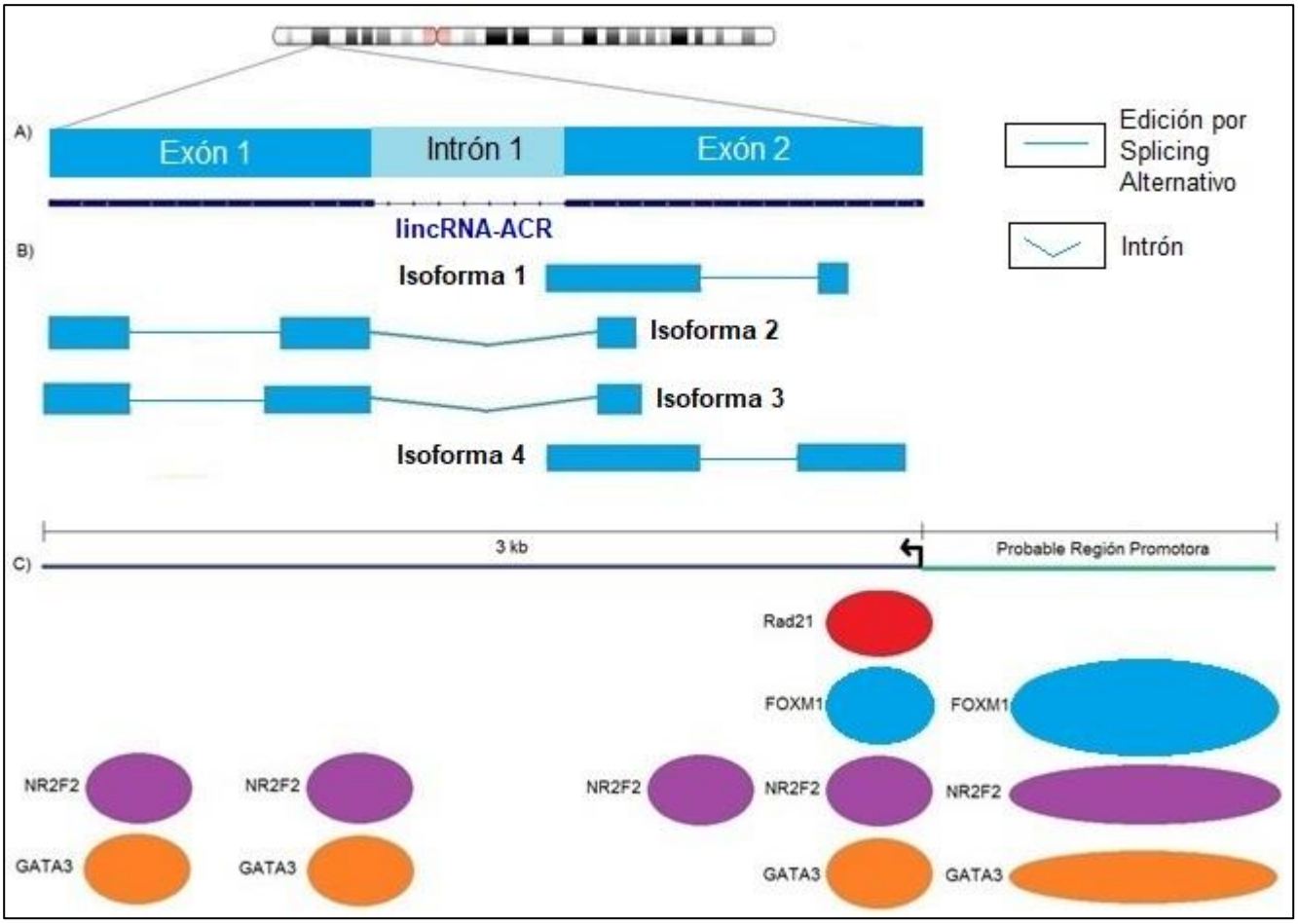


Figura 18. El lincRNA-ACR es un RNA largo de naturaleza no codificante, de tipo intergénico. A) Posición y estructura génica, el lincRNA-ACR se localiza en el cromosoma 10, y se compone por 2 exones. El transcrito tipo, que mide aprox. 2 kb, incluye estos 2 exones. B) Isoformas, el lincRNA-ACR tiene 4 isoformas, todas de menor longitud que el transcrito tipo, y se originan por el procesamiento post-transcripcional de los transcritos. C) El análisis in silico de la probable región promotora mostró la localización de factores transcripcionales como Rad21. Las proteínas FOXM1, NR2F2 y GATA3 se localizan además sobre el cuerpo del gen.

Con el objetivo de definir si la región genómica en la que se localiza el lincRNA-ACR corresponde a una unidad transcripcional y saber cuál era el estado de expresión de este gen no codificante en CaMa, se determinó mediante análisis *in silico* el enriquecimiento de las marcas de cromatina: H3K4me3, H3K27ac, H3K27me3, H3K9me3 y la H3K36me3 sobre el cuerpo del gen lincRNA-ACR con resultados de ChIP-Seq disponibles en bases de datos públicas para la línea celular MCF-7 (Figura 19). Encontramos que la modificación asociada a la elongación de la transcripción H3K27ac, se encuentra enriquecida sobre el cuerpo del gen, mientras la marca H3K4me3 asociada a promotores activos se encuentra posicionada a 500 bases del primer exón del lincRNA-ACR. El análisis *in silico* de de la secuencia promotora dio un puntaje de 0.8, y 0.7 en las dos bases de datos que se analizaron, lo cual indica que la

secuencia analizada es la posible región promotora de lincRNA-ACR Por otro lado, no se encontró el enriquecimiento de las marcas asociadas a silenciamiento transcripcional H3K9me3 y H3K27me3 sobre el cuerpo o el promotor de este gen no codificante, además, el análisis por RNA-Sec muestra expresión de este transcrito en esta línea celular. Esto sugiere que el gen lincRNA-ACR corresponde a una unidad transcripcional activa en la línea celular de CaMa MCF-7.

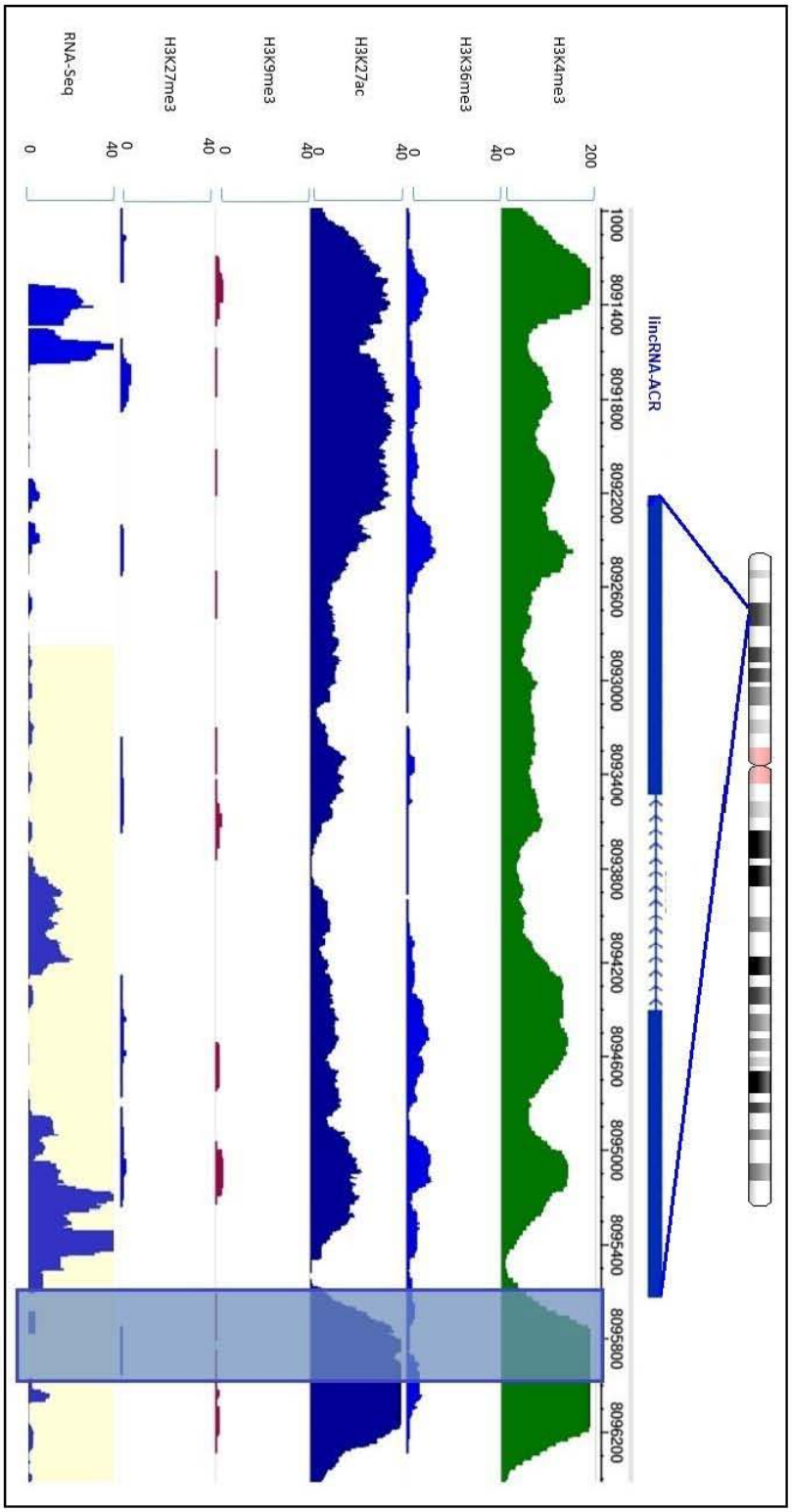


Figura 19. Caracterización *in silico* de las marcas de cromatina asociadas al lincRNA-ACR en la línea celular tumorigénica MCF-7. A) Posición genómica del lincRNA-ACR. B) Estructura del transcrito lincRNA-ACR. C) Histogramas de resultados de experimentos de ChIP-Seq para la posición genómica del lincRNA-ACR en el modelo celular neoplásico MCF-7, donde se incluyen: la modificación post-traduccional de histona H3K4me3, característica de zonas promotoras, a 500 pb río arriba del primer exón del lincRNA-ACR (región azul). Se muestra también la distribución de las modificaciones asociadas a silenciamiento transcripcional H3K27me3 y la H3K9me3, así como las modificaciones asociadas a transcripción activa como H3K36me3 y H3K27ac. Además, se muestran los resultados de RNA-Sec para este transcrito.

Una vez que determinamos que el lincRNA-ACR tiene las características de una unidad transcripcionalmente activa en un modelo celular neoplásico de CaMa, quisimos conocer el estado transcripcional de este gen en tejido normal mamario. Para ello, realizamos el mismo análisis *in silico* con la información disponible para los fibroblastos mamarios humanos sanos (Figura 20), y se determinó el enriquecimiento de las marcas de cromatina H3K4me1, H3K4me3, H3K36me3, H3K27ac y H3K27me3. Los resultados mostraron que las modificaciones post-traduccionales de histonas asociadas a la activación transcripcional H3K4me1 y H3K4me3 se encuentran poco enriquecidas en la región promotora y a lo largo del cuerpo del gen no codificante. Sin embargo, a diferencia del modelo neoplásico, en fibroblastos mamarios no se detectó el enriquecimiento de la H3K27ac ni de la H3K36me3 sobre el cuerpo del gen. En cambio, la modificación de histona asociada a la represión transcripcional H3K27me3 se encontró enriquecida sobre todo el cuerpo del gen, lo cual muestra que todo el cuerpo del gen se encuentra mayoritariamente enriquecido con marcas de cromatina asociadas a silenciamiento, incluso los datos de RNA-Sec mostraron una baja frecuencia de expresión del lincRNA-ACR. En conjunto, los resultados sugieren que en condiciones neoplásicas el lincRNA-ACR se encuentra activado y sobreexpresado de acuerdo a los datos obtenidos de ChIP-Sec y RNA-Sec, mientras que en condiciones normales del tejido mamario este lincRNA contiene marcas bivalentes de cromatina, que han sido asociadas a genes "preparados" para la transcripción (del inglés *poised*) y que sugieren una expresión basal de lincRNA-ACR en tejido normal.

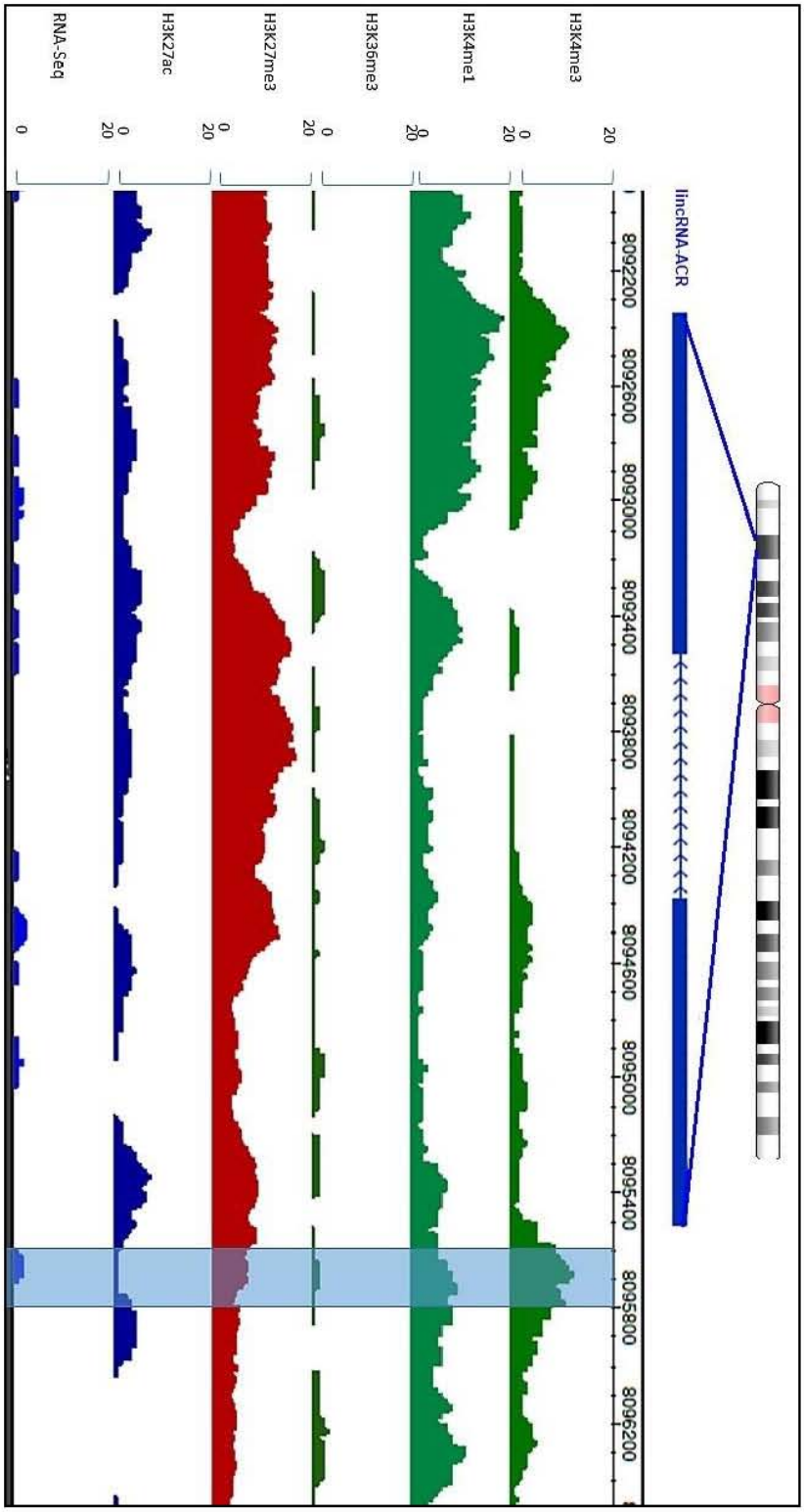


Figura 20. Caracterización in silico de las marcas de cromatina asociadas al lincRNA-ACR en tejido mamario sano. A) Posición genómica del lincRNA-ACR. B) Estructura del transcrito lincRNA-ACR. C) Histogramas de resultados de experimentos de ChIP-Seq para la posición genómica del lincRNA-ACR en fibroblastos mamarios humanos, donde se incluyen las modificaciones post-traduccionales de histonas: asociadas a promotores H3K4me1 y H3K4me3 a 500 pb río arriba del primer exón del lincRNA-ACR. Se muestran también las modificaciones asociadas a transcripción activa H3K36me3 y la H3K27ac sobre el cuerpo del gen, así como la modificación asociada a silenciamiento de la transcripción H3K27me3. Además, se muestran resultados de RNA-Sec para este transcrito.

8.1.5 Determinación del potencial codificante del lincRNA-ACR

A diferencia de los mRNA, los lncRNA se caracterizan por carecer de marcos abiertos de lectura. Una de las estrategias bioinformáticas que se ha utilizado para establecer si estos transcritos de naturaleza no codificante dan origen a algún péptido es mediante el análisis de la secuencia para buscar si hay marcos abiertos de lectura en su secuencia, lo que se conoce como potencial codificante. Con el objetivo de determinar que el lincRNA-ACR corresponde a un transcrito de naturaleza no codificante, se analizó la secuencia de la isoforma estudiada con el uso del servidor *CPC: Coding Potential Calculator.* La herramienta basa su cálculo en la evaluación del marco de lectura, su longitud, su localización entre un codón de inicio y un codón de paro, además de la homología de la secuencia del transcrito con secuencias codificantes depositadas en las bases de datos. El puntaje final que se obtiene de este proceso se interpreta de la siguiente manera: los valores positivos mayores a 1 indican secuencias que corresponden a transcritos codificantes, mientras los valores negativos menores a -1 corresponden a transcritos no codificantes, que no contienen marcos abiertos de lectura.

En nuestros análisis consideramos transcritos que corresponden a genes cuya secuencia da lugar a la síntesis de un péptido, como *BRCA1, CCND1* y *GAPDH*, utilizados a modo de controles positivos de genes codificantes, los cuales obtuvieron puntajes de potencial codificante mayores a 1, lo que corrobora que la herramienta es capaz de distinguir una secuencia codificante (Figura 21). Asimismo, en los análisis se incluyeron secuencias no codificantes de lincRNA como *HOTAIR* y *PVT1*, que previamente han sido validados como lincRNA, así como nuestro RNA candidato lincRNA-ACR. A diferencia de los genes codificantes, los valores obtenidos para estas secuencias fueron menores a -1, lo cual indica que la herramienta CPC puede identificar secuencias no codificantes. En particular, el potencial codificante de la secuencia de RNA del lincRNA-ACR obtuvo un puntaje con valor de -1.05 similar al valor obtenido de HOTAIR, uno de los lincRNAs mejor caracterizado (Figura 21) lo que indica que el lincRNA es un gen no codificante ya que carece de un marco abierto de lectura en su secuencia.

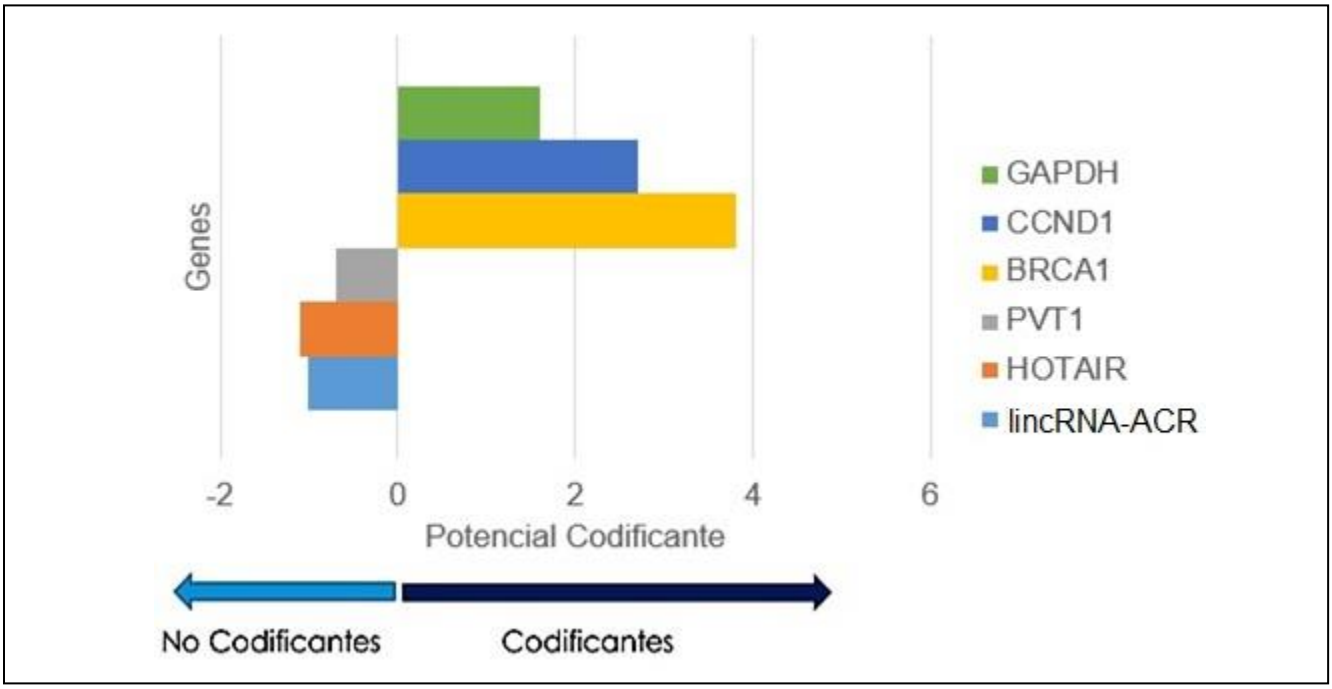


Figura 21. El lincRNA-ACR es un RNA no codificante largo. El gráfico de barras muestra el análisis del potencial codificante de la secuencia del lincRNA-ACR. Los valores positivos (cuadrante derecho) corresponden a secuencias con potencial codificante o genes codificantes, mientras los valores negativos corresponden a secuencias con potencial codificante nulo (cuadrante izquierdo) o genes no codificantes. El potencial codificante del lincRNA-ACR obtuvo un valor de -1.05, por lo que se considera un transcrito no codificante (azul). Otros lincRNA que se incluyeron en el análisis fueron HOTAIR (PC = -1.15, naranja) y PVT1 (PC = -0.7, gris) e incluyeron también transcritos con potencial codificante como controles de genes codificantes como GAPDH (PC = 1.6, verde), CCND1 (PC = 2.6, azul rey) y BRCA1 (PC = 3.9, amarillo). Los valores de potencial codificante fueron calculados con la herramienta CPC[8].

En conclusión, los análisis *in silico* de la secuencia del lincRNA-ACR indican que corresponde a un transcrito de naturaleza no codificante, que no da lugar a la síntesis de péptidos, y corrobora que pertenece al biotipo de RNA largos no codificantes.

8.1.6 Determinación de la localización celular del lincRNA-ACR

Las evidencias experimentales han establecido que los lncRNA se encuentran localizados tanto en el núcleo celular como en el citoplasma. Las últimas evidencias científicas muestran que los lncRNA tienden a encontrarse enriquecidos al interior del núcleo celular, debido a que se encuentran involucrados en la regulación epigenética de varios genes. Con el objetivo de determinar en qué compartimiento celular se encuentra acumulado lincRNA-ACR nosotros utilizamos la información de la base de datos *lncATLAS*, que recopila la información sobre la acumulación de lncRNAs en núcleo y citoplasma a partir de lo reportado en estudios de

secuenciación por RNA-Sec. Si el valor del radio es positivo, indica que el lncRNA se localiza en citoplasma, por lo que un valor negativo significa que su localización es nuclear. En la línea celular MCF-7 se observó que el lincRNA-ACR se acumula en el núcleo celular, y este fenómeno se observa también en otras líneas celulares (Figura 22), lo que sugiere que el lincRNA-ACR es un lincRNA que se acumula en el núcleo celular.

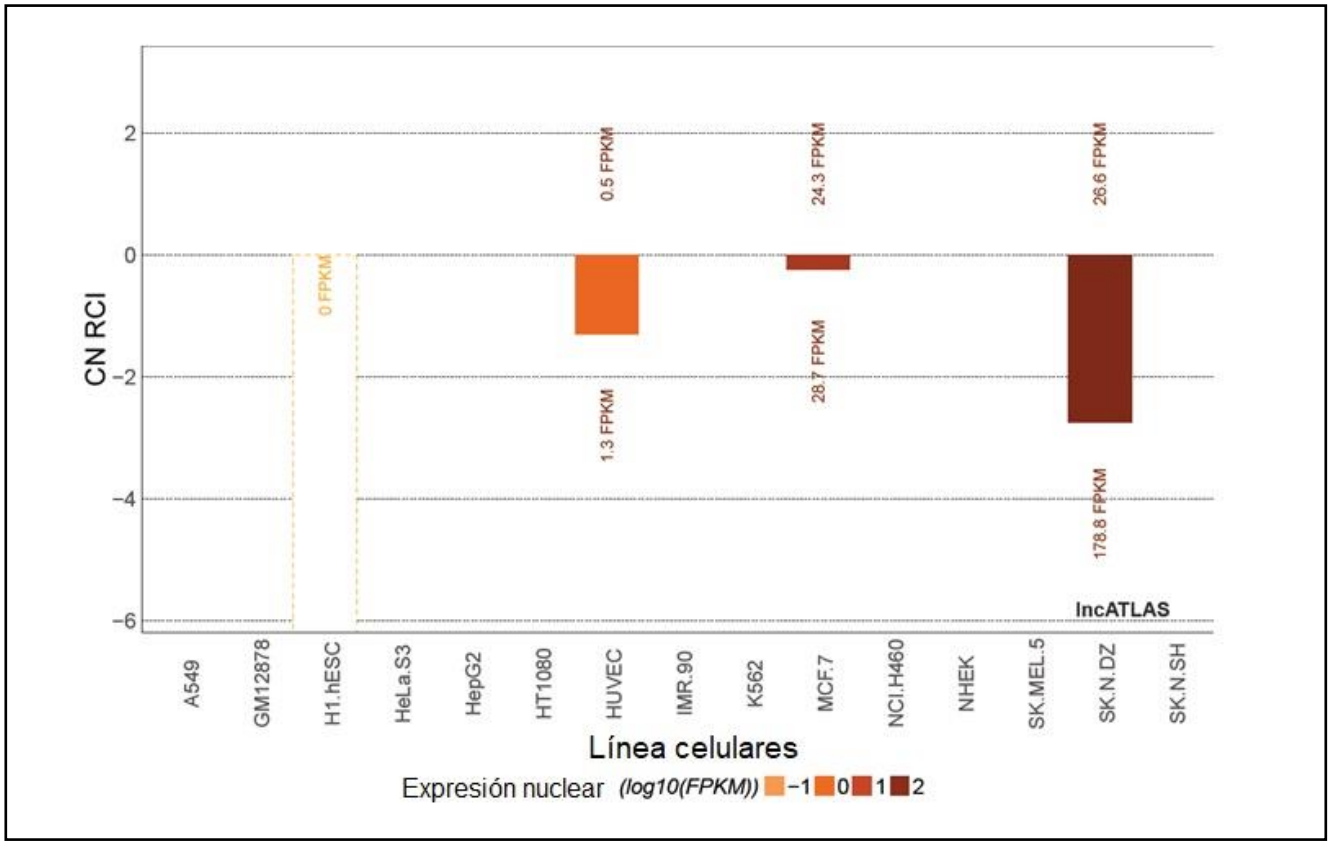


Figura 22. El lincRNA-ACR se localiza en el núcleo celular. Se observa que el radio de concentración relativa núcleo/citoplasma (CN RCI, eje de las ordenadas, escala de -6 a 2) para el lincRNA-ACR es negativo para el modelo celular MCF-7, tendencia que se observa en las líneas celulares HUVEC y SK.N.DZ, lo que indica que la localización de este transcrito es nuclear. Gráfico obtenido en lncATLAS.crg.eu

8.1.7 Determinación de la expresión de lincRNA-ACR en líneas celulares de cáncer de mama, a partir de datos de RNA-Sec

El análisis de las marcas de cromatina en la línea celular MCF-7 nos permitió distinguir que el lincRNA-ACR es un gen transcripcionalmente activo, que se expresa con mayor frecuencia en la línea celular MCF-7, respecto a la expresión en fibroblastos mamarios. Con la finalidad de evaluar la expresión del lincRNA-ACR en líneas celulares de CaMa que representan los diferentes subtipos moleculares de CaMa, se analizaron datos de RNA-Sec de las líneas

celulares MCF-10A, MCF-7, BT474 y MDA-MB-231. La primera línea celular es un modelo no tumorigénico de células mamarias transformadas, por lo que fue utilizada como un control de expresión; la línea celular MCF-7 es un modelo tumorigénico de CaMa luminal A, debido a que se caracteriza por la expresión de receptores RE, RP y no expresa HER2; BT474 es un modelo de CaMa luminal B que se caracteriza por la expresión de receptores RE, RP y HER2. Además, se analizó la línea celular MDA-MB-231, que corresponde a un modelo de CaMa triple negativo (TN), debido a que no expresa ninguno de los receptores antes mencionados.

Los archivos de secuenciación fueron obtenidos de la base de datos *GEO Datasets*, y descargados a *Galaxy* mediante el servidor ENA. Para llevar a cabo la cuantificación de transcritos, se realizó conforme a lo establecido en la sección **Metodología**. Como se observa en la Figura 23 A, la línea celular MCF-10A presenta expresión basal del lincRNA-ACR, lo que coincide con la expresión del tejido normal mamario. Por su parte, de los demás modelos celulares analizados, la línea celular MCF-7 es la que presentó la mayor expresión del lincRNA-ACR, seguida de la línea celular MDA-MB-231 y la que presentó menor expresión fue la línea celular BT474. Por lo anterior, la línea celular MCF-7 se eligió como el modelo celular para estandarizar la validación del método experimental para la cuantificación del lincRNA-ACR por PCR en tiempo real.

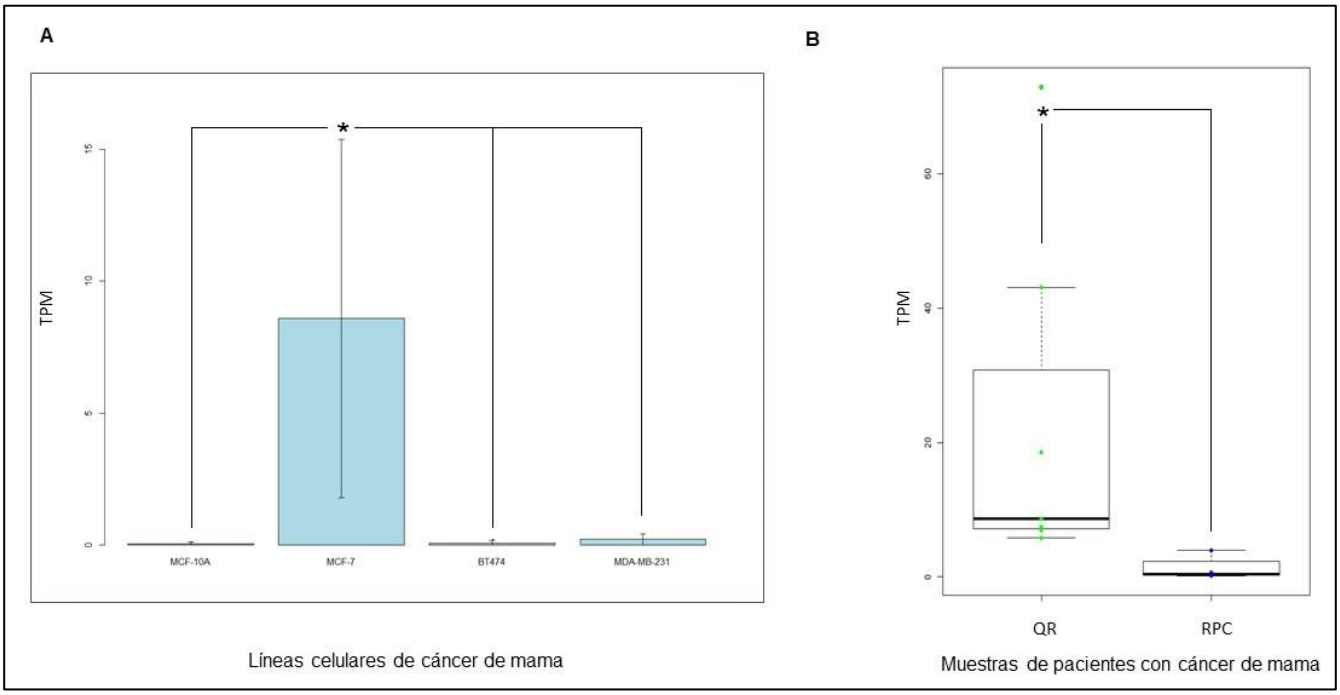


Figura 23. El lincRNA-ACR se expresa en líneas celulares tumorigénicas y en muestras de pacientes resistentes a la QTNeo. A) Análisis de expresión a partir de resultados públicos de RNA-Sec para las líneas celulares MCF-10A, MCF-7, BT474 y MDA-MB-231, donde se observa que las líneas celulares tumorigénicas

presentan mayor expresión del lincRNA que la línea celular control MCF-10A (p = 0.001). B) Gráfico de caja que ilustra la diferencia de expresión del lincRNA-ACR entre el grupo QR y el grupo RPC, siendo los pacientes del grupo QR quienes sobre-expresan este lincRNA (n=11, p = 0.004).

Para conocer los niveles de expresión en las muestras de pacientes evaluadas por RNA-Sec, se extrajo la información de cuantificación del lincRNA-ACR para cada paciente del grupo de casos (QR) y del grupo control (RPC). Lo que se observó fue que todas las pacientes del grupo QR sobreexpresan este transcrito no codificante en comparación con las pacientes del grupo control (Figura 23 B), lo que es indicativo de la posible asociación de lincRNA-ACR con la resistencia a la QTNeo.

En conclusión, el análisis a partir de los datos obtenidos de RNA-Sec para las diferentes líneas celulares de CaMa mostró que en todas las líneas celulares neoplásicas evaluadas (MCF-7, BT474 y MDA-MB-231) el lincRNA-ACR se sobreexpresa, en comparación con la línea celular de tejido mamario normal MCF-10A. Por otro lado, también se observó que esta sobreexpresión es característica de las pacientes con resistencia a la QTNeo, lo que sugiere que este transcrito de naturaleza no codificante podría estar asociado a la resistencia en la QTNeo.

8.2 Validación experimental

8.2.1 Validación experimental del lincRNA-ACR en líneas celulares y muestras de pacientes

Los resultados del análisis de la expresión del lincRNA-ACR a partir de datos de RNA-Sec en líneas celulares y muestras de pacientes confirmaron que este gen no codificante se sobreexpresa en CaMa. Para corroborar esto experimentalmente, se validó un método de cuantificación relativa del lincRNA-ACR por PCR en tiempo real en la línea celular MCF-7, y se cuantificaron además los niveles del transcrito en las mismas líneas celulares de CaMa (BT474, MDA-MB-231) y en la línea celular MCF-10A. Para ello, se extrajo RNA y se verificó que tuviera la calidad adecuada para su uso en un procedimiento de PCR en tiempo real (Figura 24). Como puede observarse, el RNA de cada una de las líneas celulares tiene un valor asociado de RIN mayor a 8, por lo que el RNA obtenido tiene la calidad e integridad apropiada

para hacer ensayos moleculares y cumple con los criterios para garantizar resultados de PCR en tiempo real.

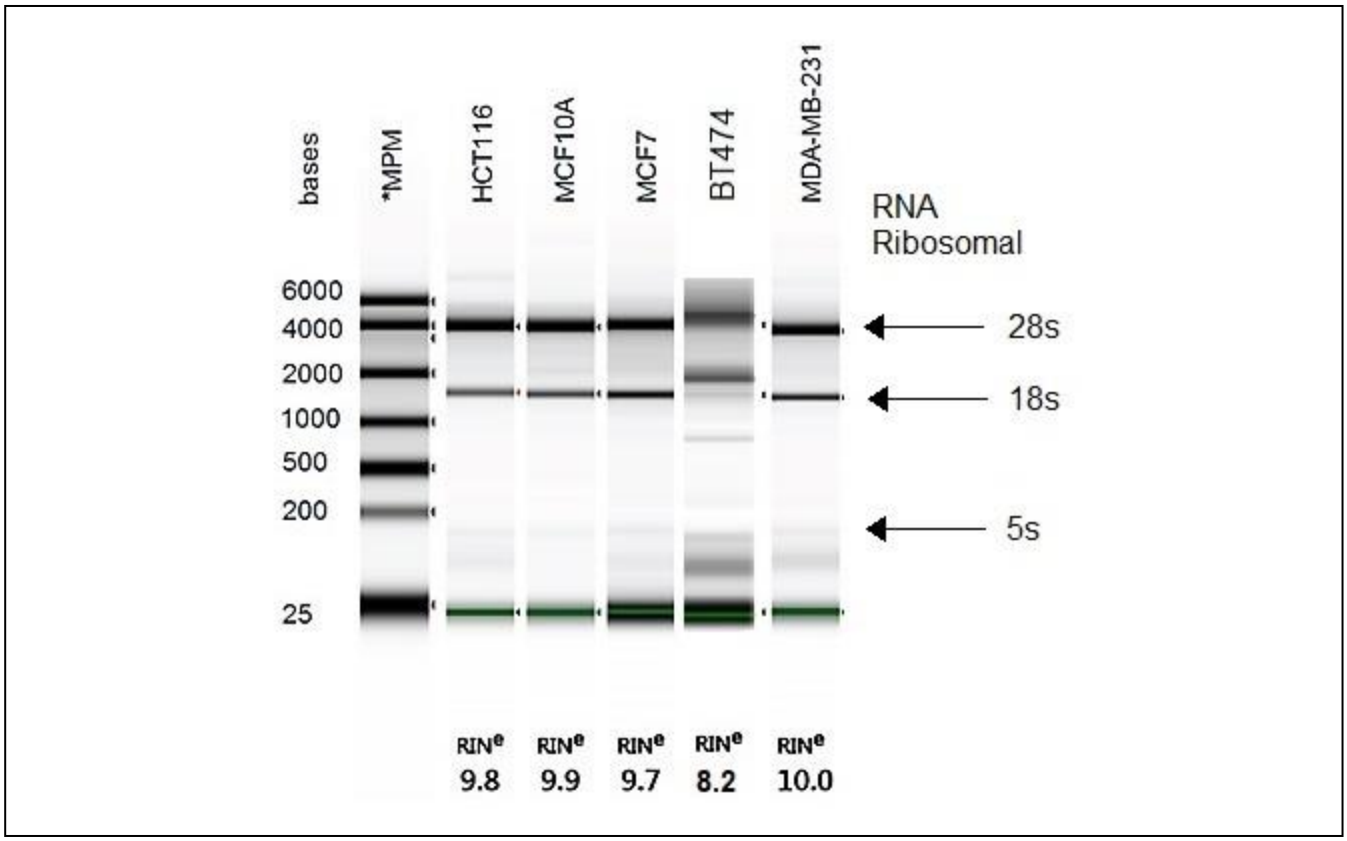


Figura 24. Calidad e integridad del RNA. Gel de electroforesis capilar que muestra la integridad del RNA (RIN), en donde se observa también la integridad de los RNA ribosomales 28s, 18s y 5s, para el RNA correspondiente a las líneas celulares MCF-10A, MCF-7, BT474, MDA-MB-231, se incluyó un control de RNA de alta calidad (HCT116).

Como parte del proceso de validación del método de cuantificación relativa por PCR en tiempo real, se procedió a evaluar la eficiencia de amplificación de los oligonucleótidos diseñados para el lincRNA-ACR y para el gen constitutivo. Los resultados mostraron que el porcentaje de eficiencia de amplificación para los oligonucleótidos del lincRNA-ACR (Seq. ID 1 y 2) y para el gen constitutivo (Seq. ID 3 y 4) se encuentra en el intervalo de 95% a 105% (Figura 25), por lo que se consideran oligonucleótidos funcionales en un procedimiento de PCR en tiempo real.

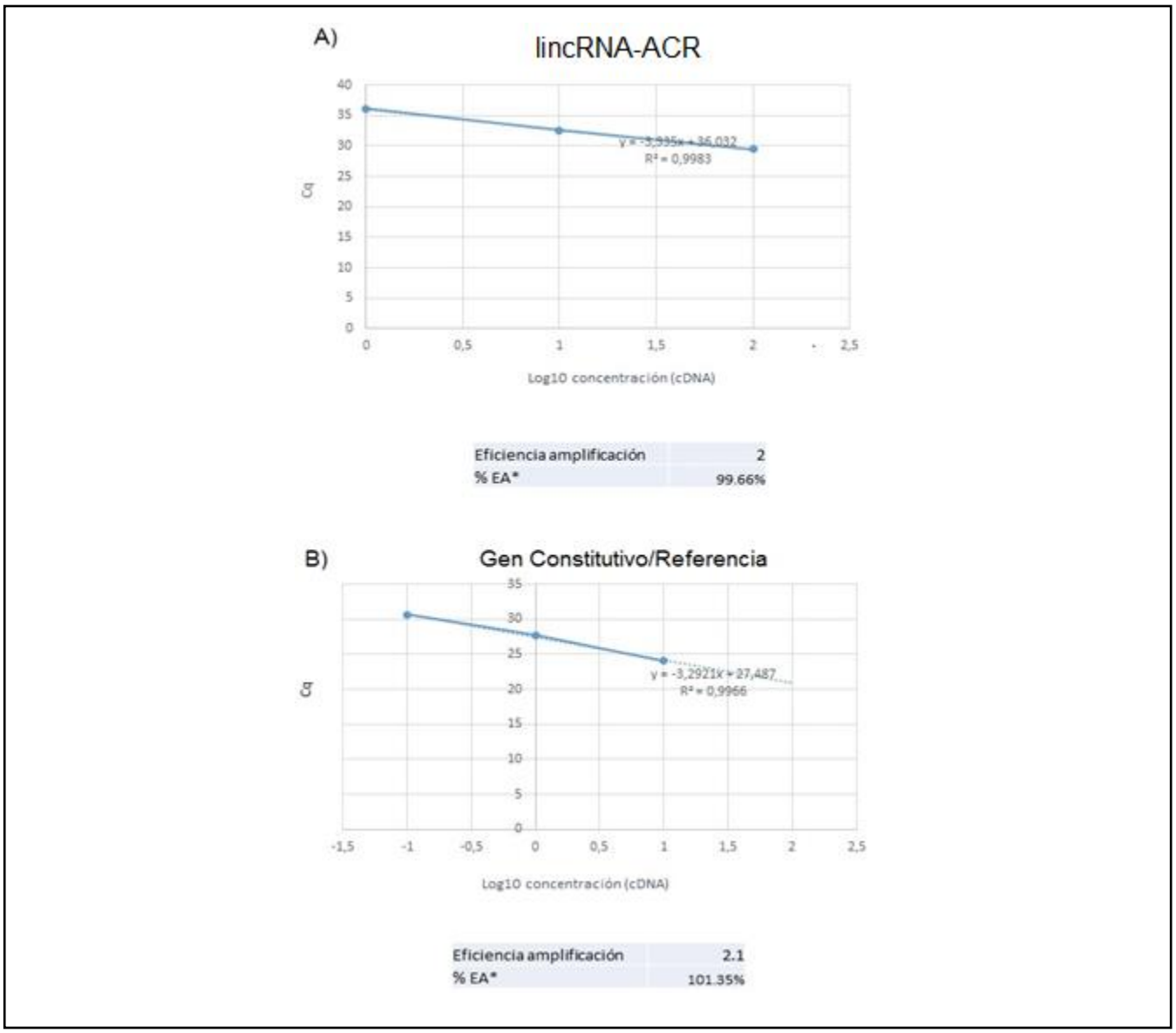


Figura 25. Eficiencia de amplificación de los oligonucleótidos de lincRNA-ACR. A) Curva de eficiencia de amplificación de los oligonucleótidos para lincRNA-ACR (Seq. ID No. 1 y 2). B) Curva de eficiencia de amplificación de los oligonucleótidos del gen constitutivo/referencia (Seq. ID. No 3 y 4). La determinación de la eficiencia de amplificación se realizó con cDNA de la línea celular MCF-7, y se obtuvo una eficiencia de 99.66% para el lincRNA-ACR y de 101.35% para los oligonucleótidos del gen referencia. El análisis se realizó en la herramienta en línea disponible en la página web de Thermo Fisher Scientific: qPCR Efficiency Calculator.

Una vez comprobada la eficiencia de amplificación de los oligonucleótidos para el lincRNA-ACR y el gen constitutivo, se procedió a cuantificar los niveles de expresión relativa del lincRNA-ACR por PCR en tiempo real, con el uso del método ΔΔCq, con el propósito de validar los resultados del análisis de los datos de RNA-Sec para las diferentes líneas celulares de CaMa, utilizando como calibrador la línea celular MCF-10A (Figura 26). Respecto a los niveles del calibrador, se encontró que la línea celular MCF-7 presenta un incremento en la expresión

del lincRNA-ACR de más de 300 veces, seguido de la sobreexpresión de la línea celular MDA-MB-231 con un incremento de 100 veces, y la línea celular BT474, que presenta los valores menores de expresión del lincRNA-ACR cuando se compara con el calibrador. Entonces, los resultados muestran que todas las líneas celulares de CaMa presentan mayor expresión del lincRNA-ACR que la línea celular no neoplásica que en este caso fue MFC-10A, por lo que la sobreexpresión del lincRNA-ACR es característica de las líneas celulares de CaMa analizadas. Además, el patrón de expresión de las diferentes líneas celulares coincidió con lo observado en el análisis *in silico*. Por lo tanto, las líneas celulares neoplásicas mamarias presentan sobreexpresión del lincRNA-ACR cuando éstas se comparan con una línea celular no neoplásica (MCF-10A).

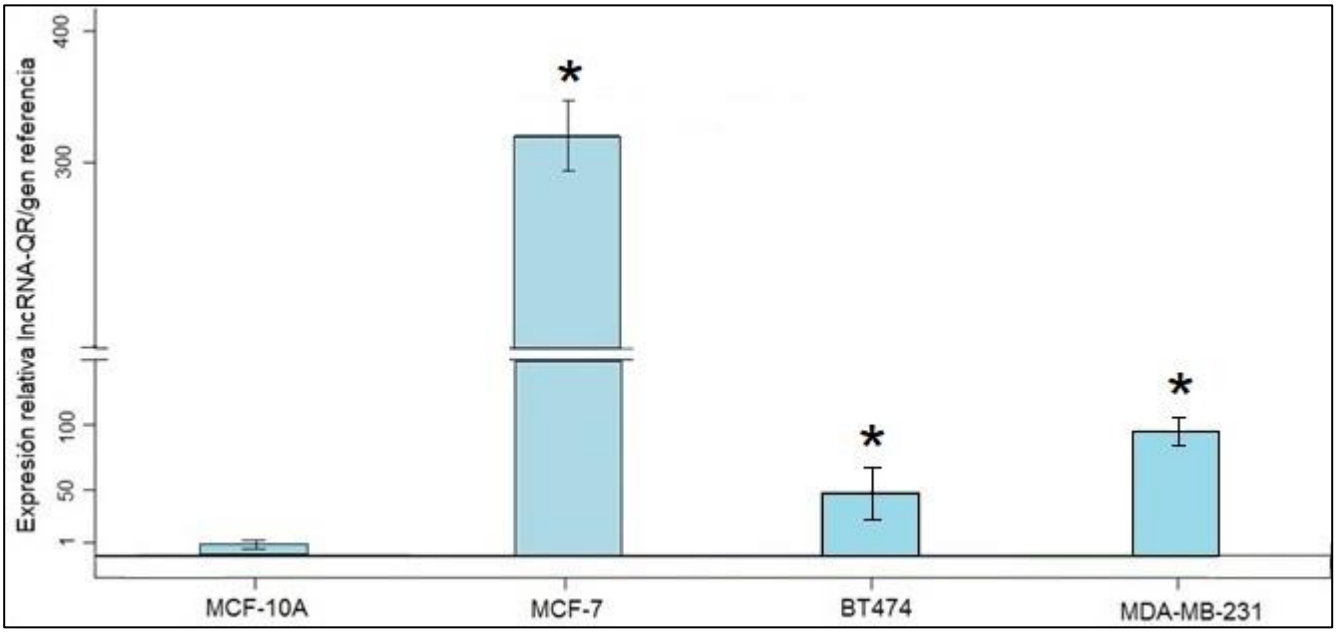


Figura 26. Expresión del lincRNA-ACR en líneas celulares de CaMa. La cuantificación relativa del lincRNA-ACR mediante un procedimiento de PCR en tiempo real mostró que la línea celular MCF-7 expresa 300 más el lincRNA, respecto al calibrador MCF-10A. Por otro lado, la línea celular MDA-MB-231 expresa al lincRNA-ACR en una frecuencia 100 veces mayor, y la línea celular BT474 lo expresa 50 veces más (ANOVA con prueba de Tuckey, p = 0.0001).

Los resultados de cuantificación del lincRNA-ACR en las líneas celulares, en conjunto con lo hallado en los análisis por RNA-Sec, nos permitió determinar su sobreexpresión en CaMa, por lo que ahora nosotros quisimos determinar el patrón de expresión de lincRNA-ACR en muestras de pacientes, mediante PCR en tiempo real. Para ello, se utilizaron las muestras de las pacientes del estudio de RNA-Sec y se incluyeron otras pacientes de una cohorte independiente, que en conjunto representan 28 pacientes. Éstas fueron divididas en dos

grupos: (1) el grupo de casos (QR) que contiene las pacientes con resistencia a la QTNeo (n= 14), y (2) el grupo control (RPC) que se compone de las pacientes sensibles a este tratamiento (n= 14). En la Figura 27 A se observan los resultados obtenidos de la cuantificación de expresión relativa por el método ΔCq para muestras de pacientes con CMLA subtipo Luminal B. Se encontró que las muestras del grupo de casos (QR) sobreexpresan el lincRNA-ACR en contraste con el grupo (RPC), donde el lincRNA-ACR se encuentra subexpresado. Cabe señalar que la sobreexpresión del grupo de casos (QR) es de 130 veces respecto al grupo control (RPC). Para ilustrar esta diferencia entre los grupos, se construyó un gráfico de caja (Figura 27 B), en el que se observa que el valor promedio de la cuantificación de expresión relativa en el grupo de casos (QR) es 100 veces mayor que la del grupo control RPC. Esto sugiere que el lincRNA-ACR se sobreexpresa en pacientes que presentan resistencia al tratamiento por lo que este transcrito no codificante podría ser un posible biomarcador de resistencia a la QTNeo en pacientes con CMLA subtipo Luminal B.

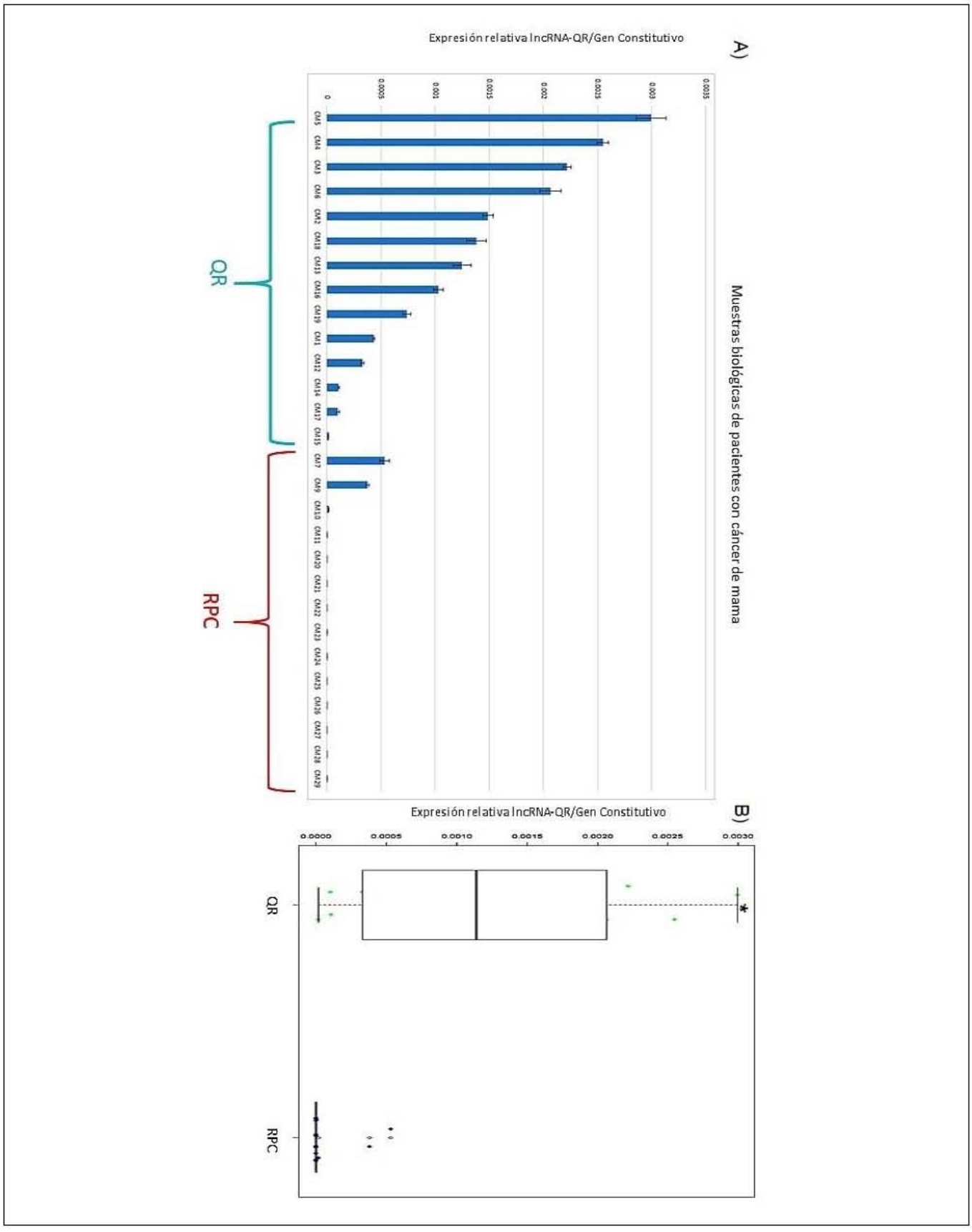


Figura 27. Validación experimental del lincRNA-ACR en pacientes que recibieron QTNeo. A) Validación de lincRNA-ACR por PCR en tiempo real. Panel izquierdo: En azul se presenta la expresión relativa del lincRNA-

ACR en las pacientes QR (casos). En rojo se presentan las pacientes con RPC (controles). QR n=14 y RPC n=14. Todas las pacientes pertenecen al subtipo molecular de CaMa Luminal B (HER2+/HER2-). B) El gráfico de caja ilustra la diferencia de expresión del lincRNA-ACR entre el grupo QR y el grupo RPC, en donde se distingue que el lincRNA-ACR es distintivo de las pacientes resistentes a QTNeo (n = 11, p = 0.0009).

Los resultados hasta ahora obtenidos muestran que el lincRNA-ACR es un transcrito de naturaleza no codificante, del biotipo lncRNA, que está sobreexpresado en pacientes diagnosticadas con CMLA, subtipo Luminal B, quienes además presentaron resistencia a la QTNeo, por lo que nos preguntamos si existía asociación estadísticamente significativa entre la sobreexpresión del lincRNA-ACR y la resistencia a la QTNeo. Para ello, llevamos a cabo un análisis de asociación mediante el cálculo de la razón de momios, y lo que se obtuvo fue un valor de la razón de probabilidades de 15:1 ($p = 0.003$), lo que puede interpretarse como que existe la probabilidad del 93.7% de presentar resistencia a la QTNeo si se detecta la sobreexpresión del lincRNA-ACR en una paciente con CMLA subtipo Luminal B. En suma, existe una asociación entre la sobreexpresión del lincRNA-ACR y la resistencia a la QTNeo en pacientes con CMLA subtipo Luminal B, lo que sugiere que puede ser utilizado como un método predictivo de respuesta a la QTNeo.

Una vez que determinamos que existía asociación significativa entre la sobreexpresión del lincRNA-ACR y la condición de resistencia a la QTNeo, nos preguntamos si éste transcrito podría ser utilizado en la práctica clínica como un biomarcador molecular para predecir la respuesta a la QTNeo en pacientes con CaMa. Con el objetivo de determinar si el lincRNA-ACR podía funcionar como biomarcador de predicción, llevamos a cabo la construcción de una curva ROC ajustada (Figura 28), que permite determinar el valor de expresión relativa en el cual se tiene la mayor especificidad y sensibilidad para la detección de pacientes resistentes a la QTNeo. A partir de este análisis, se determinó que el valor de expresión relativa del lincRNA-ACR que tiene la mayor especificidad y sensibilidad es de 0.0001, a partir del cual los valores mayores representan pacientes que presentarán resistencia a la QTNeo. Particularmente en este análisis, la curva ROC también proporciona información sobre la capacidad predictiva de un método, a través del valor del área bajo la curva (ABC), en donde valores mayores a 0.5 indican que el método es capaz de predecir la resistencia a QTNeo. Nuestros resultados muestran que el uso del lincRNA-ACR como método predictivo tiene un valor ABC = 0.96, lo que indica que la probabilidad de que el método detecte adecuadamente a una paciente resistente a la QTNeo es del 96%. Por lo tanto, los resultados obtenidos de

este análisis sugieren que el uso del lincRNA-ACR como biomarcador de predicción de respuesta a tratamiento es un método capaz de distinguir entre pacientes que presentarán resistencia a la QTNeo de aquellas que no, por lo que puede utilizarse como una herramienta clínica en la gestión de tratamientos en CaMa.

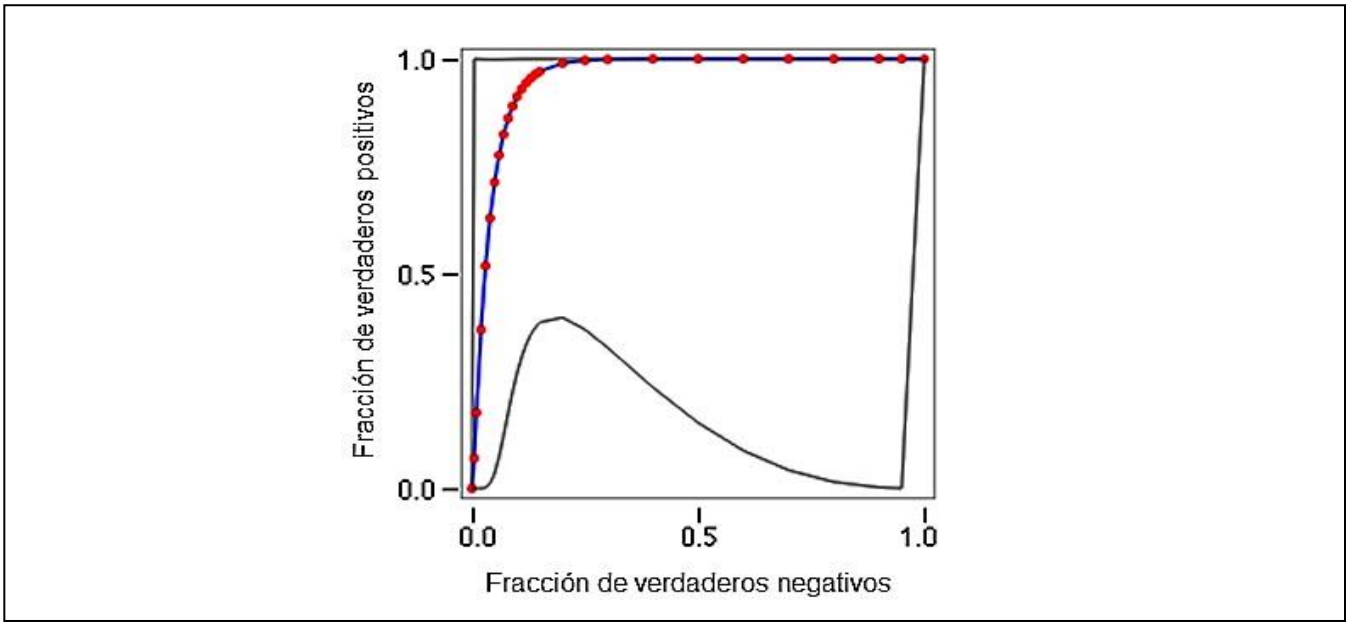


Figura 28. El lincRNA-ACR es un potencial biomarcador de predicción de la respuesta a la QTNeo. La construcción de la curva ROC ajustada permitió la identificación del punto de corte de cuantificación del lincRNA-ACR para la detección de pacientes resistentes a QTNeo, con un valor de expresión relativa de 0.0001, el cual da como resultado una especificidad del 93.7% y una sensibilidad del 85.7%, además del valor del área bajo la curva de 0.96.

En resumen, los resultados de este estudio mostraron que existe un perfil de expresión de lincRNA, identificado a través de estudios por RNA-Sec, que corresponde a pacientes con CMLA subtipo Luminal B, que presentan resistencia a la QTNeo, con lo que es posible distinguir entre la condición de resistencia y la condición de sensibilidad a este tratamiento. Derivado de este análisis, se identificó un lincRNA, al que denominamos lincRNA-ACR, que debido a sus características y a su perfil de expresión en este grupo de pacientes, podría ser candidato para su uso clínico como un biomarcador de predicción. Los resultados del análisis estadístico en la validación experimental de este lincRNA por PCR en tiempo real sugieren su uso como un biomarcador molecular de predicción de respuesta a terapia, ya que corroboramos su asociación con la condición de resistencia a la QTNeo, y se encontró que tiene una alta sensibilidad y especificidad en la predicción de respuesta a la QTNeo en el grupo de pacientes con CMLA, subtipo Luminal B.

9. Discusión

El presente trabajo es un estudio de descubrimiento de nuevos biomarcadores de predicción de respuesta a la QTNeo basado en perfiles de expresión de lincRNA, en pacientes mexicanas con CMLA, subtipo Luminal B, a través de la tecnología RNA-Sec, que es una metodología de frontera que ha mostrado ser de gran utilidad en el desarrollo de la medicina de precisión dentro del área de *Oncología molecular*. El RNA-Sec permite llevar a cabo el análisis del transcriptoma, lo cual implica evaluar la expresión de todos los transcritos que se expresan en un momento y tiempo determinado en el contexto del tumor, y dentro de la información que es posible obtener de estos estudios se encuentran los perfiles de expresión de los lincRNA, esto ha permitido conocer a mayor profundidad las características moleculares de los tumores mamarios[93]. En particular, los lincRNA son transcritos que se han asociado con funciones regulatorias en condiciones fisiológicas[57], y se sabe que están involucrados en diferentes procesos celulares relacionados con la patogénesis del CaMa[64], por lo que se han propuesto como nuevos biomarcadores moleculares de utilidad clínica en el diagnóstico, pronóstico y predicción de respuesta a tratamiento, contribuyendo así con el desarrollo de la medicina de precisión en cáncer.

Actualmente, la medicina de precisión para el manejo de pacientes en CaMa se ha apoyado del uso de firmas genéticas como Mammaprint, Prosigna y Oncotype Dx[33], que son pruebas moleculares basadas en la expresión de genes codificantes, y dependiendo de la prueba permite conocer el subtipo molecular del tumor mamario, además de proporcionar información sobre el pronóstico de la paciente (Prosigna y Mammaprint). En cuanto a la predicción de respuesta al tratamiento, se ha demostrado que las firmas genéticas como Mammaprint[36] y Oncotype Dx[94] proporcionan información sobre el beneficio que obtendrá la paciente de la quimioterapia adyuvante. Sin embargo, hay pocos estudios enfocados en la asociación de estas firmas genéticas con la respuesta a la QTNeo[95], que es actualmente uno de los esquemas de tratamiento para las pacientes con CMLA que ha presentado mejores beneficios[45] y a la fecha no existe una firma molecular de uso clínico rutinario que prediga la respuesta a la QTNeo por lo que es necesario la búsqueda de nuevos marcadores moleculares que permitan distinguir a las pacientes que se van a beneficiar de este tipo de tratamiento.

Una limitante que presenta el uso de las firmas moleculares como Prosigna y Mammaprint es que sólo se basan en el uso de genes codificantes, y esto restringe la información que puede obtenerse de la muestra analizada. Con esta premisa, en un estudio realizado por Berger y colaboradores[93], en el que se analizaron archivos de RNA-Sec de tumores mamarios disponibles en bases de datos públicas (TCGA), se demostró que tanto los perfiles de expresión de genes codificantes como de genes no codificantes de los tumores mamarios son determinantes para distinguir entre los grupos de pacientes y estas evidencias no coinciden con la agrupación basada actualmente en genes codificantes conocida como PAM50. A pesar de que PAM50 ha sido de utilidad clínica, los estudios actuales hechos por Berger y colaboradores sugieren que las firmas de expresión como Prosigna y Mammaprint limitan la información que puede obtenerse del tumor mamario, representando un obstáculo en el proceso de diagnóstico y elección de tratamiento. En este mismo estudio, nuevos análisis bioinformáticos sugieren que existe una nueva red de interacción génica basada en lncRNA asociada a los diferentes tipos de tumores mamarios[93].

En la actualidad, los lncRNA son moléculas que por sus características biológicas se han propuesto como marcadores moleculares oncológicos, ya que presentan perfiles de expresión específicos en los diferentes tipos de cáncer, en las diferentes etapas de la progresión[65] y tienen una alta estabilidad en fluidos corporales como suero, plasma y orina, que son los principales tipos de muestras que se obtienen en la clínica. Un ejemplo del uso de lncRNA como marcadores oncológicos de gran importancia y de uso rutinario en la clínica para el manejo del paciente es PCA3, que es un lincRNA que ha sido aprobado por la FDA para su uso en el diagnóstico molecular de cáncer de próstata[96]. En particular, en CaMa existen estudios que demuestran la asociación que existe entre la expresión de lincRNA y el desarrollo de esta patología. Tal es el caso de HOTAIR, que es un lincRNA cuya sobreexpresión se ha asociado a la resistencia en la terapia con tamoxifén así como al desarrollo de metástasis[97,98]. Otro ejemplo es lincRNA-ATB que es un transcrito que se encuentra sobreexpresado en pacientes con CaMa, y se ha asociado con la resistencia a trastuzumab en estudios *in vitro*[71]. Asimismo, MALAT-1 es otro lincRNA cuya sobreexpresión se ha propuesto como marcador pronóstico en CaMa, ya que se ha visto asociado al desarrollo de metástasis y a la recurrencia a 5 años[98]. De la misma manera, TINCR es otro lincRNA que se ha asociado a la resistencia a terapia endócrina y se sabe que este transcrito no codificante se encuentra sobreexpresado en CaMa asociándose al aumento del potencial proliferativo de las células mamarias[69].

Finalmente, lincROR ha sido asociado con la capacidad invasiva del tumor, mediante la regulación de miR-145 y la vía MAPK/ERK, donde su sobreexpresión se ha asociado a un mal pronóstico[68]. Todo lo anterior es parte de la evidencia que demuestra que los lincRNA son moléculas importantes en el desarrollo y la progresión del CaMa, ya que sus perfiles de expresión se han asociado con la regulación de vías de señalización intracelulares como la proliferación celular, el ciclo celular y la resistencia a tratamiento, por lo que sus perfiles de expresión pueden ser utilizados como una herramienta clínica de diagnóstico, pronóstico y predicción de respuesta a los distintos tratamientos utilizados en CaMa.

Nosotros en particular demostramos que el análisis del transcriptoma mediante RNA-Sec permite la identificación de lincRNA asociados a la respuesta a la QTNeo en pacientes mexicanas con CMLA subtipo Luminal B. A través de un estudio de casos y controles, donde nuestro grupo de casos (QR) que incluyó a las pacientes que presentaron resistencia a la QTNeo fue comparado con el grupo control de pacientes que fueron sensibles al tramiento (RPC), fue posible identificar un perfil de expresión de lincRNAs relacionado con la resistencia a la QTNeo. Dentro de nuestros hallazgos, encontramos que existe un perfil de expresión basado en lincRNA que es característico de las pacientes mexicanas con CMLA subtipo Luminal B resistentes a la QTNeo, el cual no se ha reportado en estudios anteriores. Este perfil de expresión se compone de 74 genes no codificantes del biotipo lincRNAs: 57 subexpresados y 17 sobreexpresados, los cuales se encuentran diferencialmente expresados en las pacientes con resistencia a la QTNeo, respecto al grupo de pacientes sensibles a este tratamiento. En particular, nosotros nos enfocamos en el conjunto de lincRNA sobreexpresados debido a que este tipo de transcritos son muy pocos abundantes en los distintos tipos de células, encontrando hasta un transcrito por célula como HOTTIP[99].

El análisis por ontología de genes mostró que el grupo de lincRNA sobreexpresados mostraron estar relacionados principalmente con la regulación de los procesos de proliferación, apoptosis y la respuesta inmunológica, sugiriendo que su sobreexpresión podría estar asociada con el desarrollo carcinogénico afectando vías de señalización intracelular o alterando la regulación transcripcional a través de su interacción con otros genes y/o proteínas (Tabla 10). Entre los lincRNA sobreexpresados en las pacientes que mostraron resistencia a la QTNeo, nosotros seleccionamos a lincRNA-ACR, ya que fue el lincRNA del que se encontró más información en la literatura a diferencia de los demás lincRNA, los cuales no han sido caracterizados ni

reportados hasta la fecha[100]. El transcrito lincRNA-ACR se caracterizó de manera bioinformática y su expresión se validó experimentalmente por PCR en tiempo real, sin embargo, no existen a la fecha reportes sobre su caracterización molecular, funcional, de su posible papel en la patogénesis del CaMa, ni de su asociación clínica a la resistencia a la QTNeo. Debido a todo lo anterior, nosotros decidimos enfocarnos a estudiar este lincRNA en las muestras de pacientes con CMLA.

Los resultados experimentales por PCR en tiempo real de líneas celulares corroboraron lo observado en el análisis de los datos de RNA-Sec en la cuantificación del lincRNA-ACR, por ello decidimos estandarizar el método de cuantificación en la línea celular MCF-7, que es la que presentó la mayor expresión de este transcrito, y se utilizó para cuantificar la expresión relativa del lincRNA-ACR en muestras de pacientes. Los resultados de la cuantificación por PCR en tiempo real de este transcrito sugieren que el lincRNA-ACR se sobreexpresa exclusivamente en las líneas neoplásicas de CaMa, así como en las pacientes que muestran resistencia a la QTNeo. De manera interesante, nuestros resultados coinciden con lo reportado en la clínica acerca de la respuesta a la QTNeo y la expresión de HER2, ya que las pacientes que mostraron ser resistentes a tratamiento son en su mayoría negativas a la expresión de este receptor, mientras las que muestran sensibilidad a este tratamiento son HER2+, lo cual sugiere que existe una relación entre la sobreexpresión del lincRNA-ACR y la expresión de HER2, por lo que se necesita incluir más pacientes para corroborar esta asociación y tener mayor robustez en los análisis estadísticos.

Por otro lado, para trasladar una metodología a su uso en clínica es necesario conocer la capacidad de detección de la misma. Una herramienta útil para ello es el análisis por curvas ROC. Este tipo de gráficos permiten el análisis de diferentes puntos de corte, y determina la sensibilidad y especificidad del biomarcador. El punto de corte óptimo es aquel que en el gráfico se relaciona con la mayor especificidad y sensibilidad posible para el método que se evalúa[101]. Otro aspecto importante del gráfico es el cálculo del área bajo la curva, que es indicativo de la capacidad del método para predecir un resultado, o en este caso para detectar una paciente resistente a la QTNeo, por lo que valores del área bajo la curva mayores a 0.5 indican que el método es capaz de realizar esta distinción. En este caso, nuestros resultados de la curva ROC sugieren que el uso del lincRNA-ACR como marcador de respuesta a terapia en pacientes con CMLA subtipo Luminal B cuenta con la suficiente sensibilidad y especificidad

para identificar a las pacientes resistentes a la QTNeo. Sin embargo, a pesar de haber obtenido altos valores de sensibilidad y especificidad es necesario incluir en el análisis un número mayor de muestras y analizar de manera independiente a las pacientes de acuerdo a la expresión de HER2.

Los resultados de los análisis estadísticos indicaron que existía diferencia significativa de la expresión del lincRNA-ACR en el grupo de pacientes con resistencia a la QTNeo; sin embargo, es necesario establecer si realmente existe una asociación significativa entre la expresión de este transcrito no codificante y la condición de resistencia a la QTNeo en CMLA, por lo que realizamos el cálculo de la razón de momios o probabilidades. Este tipo de análisis es utilizado comúnmente en estudios epidemiológicos para identificar factores de riesgo. Sin embargo, el diseño del cálculo también es útil para realizar análisis en estudios de casos y controles, que es el tipo de estudio que se presentó en esta tesis, por lo que con ello justificamos el uso de este análisis para determinar la asociación con la resistencia al tratamiento[102].

El cálculo de la razón de momios indica que la probabilidad de desarrollar resistencia a la QTNeo con la sobreexpresión del lincRNA-ACR es de más del 90% en pacientes con CMLA subtipo Luminal B, lo que confirma que su detección puede ser utilizada como técnica clínica de apoyo a la gestión de tratamiento. A pesar de que los resultados estadísticos sugieren el uso del lincRNA-ACR como un biomarcador molecular de predicción de respuesta a tratamiento, el estudio tiene en general una limitante en cuanto al número de muestras analizadas, ya que es reducido respecto a lo reportado en otros análisis de búsqueda de biomarcadores[55,100], lo cual significa para nosotros un sesgo estadístico. Es por ello que ya se ha planteado como una perspectiva prioritaria ampliar la cohorte de pacientes, incluyendo el reclutamiento de un mayor número de pacientes del subtipo Luminal B, tanto de muestras de tejido fresco como de muestras embebidas en bloques de parafina (como parte de un estudio retrospectivo), además de muestras de pacientes de los demás subtipos moleculares (Luminal A, TN y HER2 enriquecido).

Por otro lado, la relación que existe entre la expresión del lincRNA-ACR y la resistencia a la QTNeo podría estar relacionada con algún mecanismo molecular asociado con la resistencia farmacológica a la QTNeo. El esquema utilizado en el INCan consiste en la administración de los agentes quimioterapéuticos 5-fluorouracilo, adriamicina y ciclofosfamida, los cuales

producen daño a las células cancerosas mediante diferentes mecanismos. Todos ellos tienen en común que actúan sobre los procesos de replicación y reparación de DNA, así como la síntesis de RNA, ya que generan cambios químicos en los ácidos nucleicos. En el caso de los agentes farmacológicos administrados en la QTNeo se han descrito algunos mecanismos moleculares de resistencia, entre los que destacan aquellos en los cuales las moléculas que participan en la detección de daño son capaces de activar la maquinaria de reparación de DNA y la detección de errores en la síntesis de mRNA[40,41,103], evitando así el efecto citotóxico de la terapia. Se sabe que los lincRNA son capaces de regular la actividad génica interactuando con las proteínas y el DNA[60], por lo que la asociación del lincRNA-ACR con la resistencia a la QTNeo podría estar dada por la interacción directa del lincRNA-ACR con estas proteínas, o bien regulando la expresión de las mismas, manteniendo así activos los mecanismos de reparación del daño causado por el fluorouracilo, la adriamicina y la ciclofosfamida. En cuanto al trastuzumab, la expresión del receptor HER2 se asocia con la respuesta a la QTNeo, y la expresión del lincRNA-ACR es basal en presencia de esta proteína, por lo que no encontramos asociación con la resistencia a la terapia biológica.

Finalmente, el presente trabajo es un estudio sin precedentes de secuenciación por RNA-Sec en pacientes mexicanas con CMLA subtipo Luminal B. Existen otros trabajos en los cuales se evalúan firmas genéticas basadas en lncRNA como biomarcadores de predicción a la respuesta a QTNeo en pacientes con CaMa. Sin embargo, ninguno de ellos es específico para la población mexicana, no están desarrollados para la detección de resistencia a la QTNeo y no proporcionan información acerca de la asociación, la sensibilidad o especificidad del uso de los lincRNA reportados[55,104], por lo que este trabajo proporciona información relevante acerca de los perfiles de expresión de lincRNA en la resistencia a QTNeo en CaMa, y contribuye con el desarrollo de la medicina de precisión oncológica en México, lo que impactará de manera relevante en el manejo del paciente oncológico en CaMa en un futuro.

10. Conclusiones

De los resultados obtenidos en este trabajo, se concluye lo siguiente.

- Las pacientes con CMLA subtipo Luminal B, positivas y negativas a la expresión de HER2 que presentan resistencia a la QTNeo y las que presentan RPC tienen perfiles específicos de expresión de lincRNA, que permiten la distinción entre los grupos.

- El lincRNA-ACR, identificado en el análisis de expresión diferencial, presenta las características que definen a los lncRNA intergénicos, corresponde a una unidad transcripcionalmente activa que presenta expresión basal en el tejido normal mamario humano, y que se sobreexpresa en líneas celulares de CaMa.

- El lincRNA-ACR es detectable mediante la técnica de PCR en tiempo real a partir de RNA obtenido de biopsias, lo que permite su cuantificación para la predicción de respuesta a la QTNeo.

- El uso del lincRNA-ACR como método de predicción de respuesta a la QTNeo tiene una sensibilidad del 93.7% y una especificidad del 85.7%, lo que lo convierte a este transcrito no codificante en un potencial biomarcador molecular de predicción de respuesta a la QTNeo.

En conclusión, el presente estudio demostró que existe un conjunto de lincRNA asociados con la respuesta a la quimioterapia neoadyuvante en pacientes mexicanas con cáncer de mama localmente avanzado, entre los cuales destacó el lincRNA-ACR, cuya especificidad y sensibilidad en la detección de resistencia a la quimioterapia neoadyuvante lo muestran como un potencial biomarcador de respuesta a este tratamiento.

Actualmente el presente trabajo se encuentra en el proceso de trámite de patente bajo el título “Biomarcador Molecular para la Predicción de la Respuesta a la Quimioterapia Neoadyuvante en Pacientes con Cáncer de Mama Localmente Avanzado, mediante detección por PCR” con el número de solicitud Mx/a/2018/015065.

11. Perspectivas

El presente proyecto se dirigió específicamente a la búsqueda de biomarcadores moleculares de predicción de respuesta a la QTNeo en las pacientes con CMLA, subtipo Luminal B, positivas y negativas a la expresión de HER2. No obstante, debido a los resultados obtenidos durante la realización del proyecto y la falta de literatura científica encontrada para el lincRNA-ACR, queda como propuesta para continuar el proyecto lo siguiente.

- Ampliar la cohorte de muestras de pacientes con CMLA, subtipo Luminal B, para determinar de forma más adecuada la sensibilidad y especificidad del biomarcador lincRNA-ACR, y lograr su inclusión como método predictivo de rutina en la práctica clínica.
- Caracterizar los perfiles de expresión del lincRNA-ACR en muestras de pacientes con CMLA, incluyendo los subtipos moleculares: Luminal A, HER2 enriquecido, basal y triple negativo.
- Llevar a cabo la caracterización molecular del lincRNA-ACR en CaMa que incluye la caracterización experimental del promotor del lincRNA-ACR, mediante ensayos de luciferasa, marcas de cromatina y factores transcripcionales asociados al locus del lincRNA-ACR por ChIP-Sec, interacciones físicas de lincRNA-ACR con proteínas utilizando RNA-ChIP y RAP, así como análisis bioinformáticos que incluyen análisis de la co-expresión del lincRNA-ACR con otros genes, determinación de las redes de interacción génica del lincRNA-ACR e identificación de los blancos de regulación del lincRNA-ACR.
- Determinar experimentalmente la contribución del lincRNA-ACR con la resistencia a agentes quimioterapéuticos utilizados en los esquemas de la QTNeo.
- Realizar la caracterización de los otros lincRNA identificados en este trabajo, para continuar la búsqueda de candidatos a biomarcadores de predicción en la respuesta a QTNeo en CMLA, y generar un panel de expresión que permita distinguir con mayor certeza entre las pacientes que presentarán resistencia a la terapia de aquellas que responderán.

12. Referencias

1. Dellaire, Graham, B. M. Breast Cancer Genomics. in *Cancer Genomics* 213–232 (Academic Press, 2014).
2. Gallager, H. S. The developmental pathology of breast cancer. *Cancer* **46**, 905–907 (1980).
3. González Blanco, I., Hervás, G. & M, J. Historia natural del cáncer de mama. *Toko-Ginecol. Práctica* 264–269
4. Malhotra, G. K., Zhao, X., Band, H. & Band, V. Histological, molecular and functional subtypes of breast cancers. *Cancer Biol. Ther.* **10**, 955–960 (2010).
5. Salud, S. de. Cáncer de Mama. Introducción. *gob.mx* Available at: http://www.gob.mx/salud. (Accessed: 14th March 2018)
6. Cancer today. Available at: http://gco.iarc.fr/today/home. (Accessed: 30th March 2018)
7. WHO | Mortality Burden Disease. *WHO* Available at: http://www.who.int. (Accessed: 30th March 2018)
8. INCan 2016. Available at: http://incan-mexico.org/incan/incan.jsp?iu_p=/incan/pub/estatico/direccion/incan-numeros.xml. (Accessed: 20th May 2018)
9. Salud, S. de. Boletín Epidemiológico Sistema Nacional de Vigilancia Epidemiológica Sistema Único de Información. *gob.mx* Available at: http://www.gob.mx/salud. (Accessed: 11th February 2018)
10. WHO | Top 10 causes of death. *WHO* doi:/entity/gho/mortality_burden_disease/causes_death/top_10/en/index.html
11. Breast Cancer Treatment. *National Cancer Institute* Available at: https://www.cancer.gov. (Accessed: 31st March 2018)
12. Castilla, L. H. *et al.* Mutations in the BRCA1 gene in families with early-onset breast and ovarian cancer. *Nat. Genet.* **8**, 387–391 (1994).
13. Easton, D. F. *et al.* A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am. J. Hum. Genet.* **81**, 873–883 (2007).
14. Darb-Esfahani, S. *et al.* Role of TP53 mutations in triple negative and HER2-positive breast cancer treated with neoadjuvant anthracycline/taxane-based chemotherapy. *Oncotarget* **7**, 67686–67698 (2016).
15. Ju, X. *et al.* Akt1 governs breast cancer progression in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7438–7443 (2007).
16. Si, W. *et al.* Dysfunction of the Reciprocal Feedback Loop between GATA3- and ZEB2-Nucleated Repression Programs Contributes to Breast Cancer Metastasis. *Cancer Cell* **27**, 822–836 (2015).
17. Verigos, J. & Magklara, A. Revealing the Complexity of Breast Cancer by Next Generation Sequencing. *Cancers* **7**, 2183–2200 (2015).
18. Breast Cancer Research Program, Congressionally Directed Medical Research Programs. Available at: http://cdmrp.army.mil. (Accessed: 19th March 2018)
19. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
20. Weinberg, R. *The Biology of Cancer, Second Edition*. (Garland Science, 2013).
21. Alberts, B. *et al. Molecular Biology of the Cell*. (Garland Science, 2002).
22. Toss, A. *et al.* Molecular Biomarkers for Prediction of Targeted Therapy Response in Metastatic Breast Cancer: Trick or Treat? *Int. J. Mol. Sci.* **18**, (2017).

23. DeVita, V. T., Lawrence, T. S., Rosenberg, S. A., DePinho, R. A. & Weinberg, R. A. *DeVita, Hellman, and Rosenberg's Cancer: Principles and Practice of Oncology*. (Lippincott Williams and Wilkins, 2011).
24. Narod, S. A. & Salmena, L. BRCA1 and BRCA2 mutations and breast cancer. *Discov. Med.* **12**, 445–453 (2011).
25. Allison, K. H. Molecular pathology of breast cancer: what a pathologist needs to know. *Am. J. Clin. Pathol.* **138**, 770–780 (2012).
26. Bombonati, A. & Sgroi, D. C. The molecular pathology of breast cancer progression. *J. Pathol.* **223**, 307–317 (2011).
27. Secretaría de Salud. Guía de Referencia Rápida: Diagnóstico y Tratamiento del Cáncer de Mama en Segundo y Tercer Nivel de Atención. (2009).
28. Arce C, Bargalló, E., Villaseñor Y. Oncoguía. Cáncer de Mama. (2011).
29. del Castillo, C. del, Acevedo, J. C., Peralta, O. & Solá, A. Cáncer de Mama Localmente Avanzado y Cáncer de Mama Inflamatorio. in *II Jornada Chilena de Consenso en Cáncer de Mama* 105–109
30. Liu, Z., Zhang, X.-S. & Zhang, S. Breast tumor subgroups reveal diverse clinical prognostic power. *Sci. Rep.* **4**, 4002 (2014).
31. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10869–10874 (2001).
32. Wallden, B. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics* **8**, (2015).
33. Xin, L., Liu, Y.-H., Martin, T. A. & Jiang, W. G. The Era of Multigene Panels Comes? The Clinical Utility of Oncotype DX and MammaPrint. *World J. Oncol.* **8**, 34–40 (2017).
34. Netto, G. J., Saad, R. D. & Dysert, P. A. Diagnostic molecular pathology: current techniques and clinical applications, part I. *Proc. Bayl. Univ. Med. Cent.* **16**, 379–383 (2003).
35. PAM50 (Prosigna) | Susan G. Komen®. Available at: https://ww5.komen.org/BreastCancer/PAM50.html. (Accessed: 13th April 2018)
36. Research, A. A. for C. MammaPrint Reduces Breast Cancer Overtreatment. *Cancer Discov.* **6**, OF4–OF4 (2016).
37. Reference, G. H. What is precision medicine? *Genetics Home Reference* Available at: https://ghr.nlm.nih.gov/primer/precisionmedicine/definition. (Accessed: 17th November 2018)
38. Chew, H. K. Adjuvant therapy for breast cancer. *West. J. Med.* **174**, 284–287 (2001).
39. Choi, M. *et al.* Evaluation of Pathologic Complete Response in Breast Cancer Patients Treated with Neoadjuvant Chemotherapy: Experience in a Single Institution over a 10-Year Period. *J. Pathol. Transl. Med.* **51**, 69–78 (2017).
40. Zhang, N., Yin, Y., Xu, S.-J. & Chen, W.-S. 5-Fluorouracil: Mechanisms of Resistance and Reversal Strategies. *Molecules* **13**, 1551–1569 (2008).
41. Thorn, C. F. *et al.* Doxorubicin pathways: pharmacodynamics and adverse effects. *Pharmacogenet. Genomics* **21**, 440–446 (2011).
42. Humans, I. W. G. on the E. of C. R. to. *CYCLOPHOSPHAMIDE*. (International Agency for Research on Cancer, 2012).
43. Cyclophosphamide. Available at: https://www.drugbank.ca/drugs/DB00531. (Accessed: 17th January 2019)
44. Hudis, C. A. Trastuzumab — Mechanism of Action and Use in Clinical Practice. *N. Engl. J. Med.* **357**, 39–51 (2007).
45. Specht, J. & Gralow, J. R. Neoadjuvant chemotherapy for locally advanced breast cancer. *Semin. Radiat. Oncol.* **19**, 222–228 (2009).
46. Doval, D. C., Dutta, K., Batra, U. & Talwar, V. Neoadjuvant chemotherapy in breast cancer: review of literature. *J. Indian Med. Assoc.* **111**, 629–631 (2013).

47. Herrmann, A., Hall, A. & Zdenkowski, N. Women's Experiences with Deciding on Neoadjuvant Systemic Therapy for Operable Breast Cancer: A Qualitative Study. *Asia-Pac. J. Oncol. Nurs.* **5**, 68–76 (2018).
48. González Hernández, Á. *Principios de bioquímica clínica y patología molecular (2a. ed.)*. (Elsevier Health Sciences Spain - T, 2014).
49. Strimbu, K. & Tavel, J. A. What are Biomarkers? *Curr. Opin. HIV AIDS* **5**, 463–466 (2010).
50. Henry, N. L. & Hayes, D. F. Cancer biomarkers. *Mol. Oncol.* **6**, 140–146 (2012).
51. Amin, S. & Bathe, O. F. Response biomarkers: re-envisioning the approach to tailoring drug therapy for cancer. *BMC Cancer* **16**, (2016).
52. Duffy, M. J. *et al.* Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). *Eur. J. Cancer* **75**, 284–298 (2017).
53. Tan, W., Yang, M., Yang, H., Zhou, F. & Shen, W. Predicting the response to neoadjuvant therapy for early-stage breast cancer: tumor-, blood-, and imaging-related biomarkers. *Cancer Manag. Res.* **10**, 4333–4347 (2018).
54. Han, H. S. & Magliocco, A. M. Molecular Testing and the Pathologist's Role in Clinical Trials of Breast Cancer. *Clin. Breast Cancer* **16**, 166–179 (2016).
55. Wang, G., Chen, X., Liang, Y., Wang, W. & Shen, K. A Long Noncoding RNA Signature That Predicts Pathological Complete Remission Rate Sensitively in Neoadjuvant Treatment of Breast Cancer. *Transl. Oncol.* **10**, 988–997 (2017).
56. Ransohoff, J. D., Wei, Y. & Khavari, P. A. The functions and unique features of long intergenic non-coding RNA. *Nat. Rev. Mol. Cell Biol.* **19**, 143–157 (2018).
57. Huarte, M. The emerging role of lncRNAs in cancer. *Nat. Med.* **21**, 1253–1261 (2015).
58. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
59. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **17**, 47–62 (2016).
60. Ma, L., Bajic, V. B. & Zhang, Z. On the classification of long non-coding RNAs. *RNA Biol.* **10**, 925–933 (2013).
61. DNA Packaging: Nucleosomes and Chromatin | Learn Science at Scitable. Available at: https://www.nature.com/scitable/topicpage/dna-packaging-nucleosomes-and-chromatin-310. (Accessed: 9th August 2018)
62. Chellappan Srikumar. *Chromatin Protocols*. (Human Press, 2015).
63. Sun, M. & Kraus, W. L. From Discovery to Function: The Expanding Roles of Long Non-Coding RNAs in Physiology and Disease. *Endocr. Rev.* er00009999 (2015). doi:10.1210/er.0000-9999
64. Wang, Z. *et al.* lncRNA Epigenetic Landscape Analysis Identifies EPIC1 as an Oncogenic lncRNA that Interacts with MYC and Promotes Cell-Cycle Progression in Cancer. *Cancer Cell* **33**, 706-720.e9 (2018).
65. Schmitt, A. M. & Chang, H. Y. Long Noncoding RNAs in Cancer Pathways. *Cancer Cell* **29**, 452–463 (2016).
66. Dos Anjos Pultz, B. *et al.* Far beyond the usual biomarkers in breast cancer: a review. *J. Cancer* **5**, 559–571 (2014).
67. Malih, S., Saidijam, M. & Malih, N. A brief review on long noncoding RNAs: a new paradigm in breast cancer pathogenesis, diagnosis and therapy. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **37**, 1479–1485 (2016).
68. Peng, W., Huang, J., Yang, L., Gong, A. & Mo, Y.-Y. Linc-RoR promotes MAPK/ERK signaling and confers estrogen-independent growth of breast cancer. *Mol. Cancer* **16**, 161 (2017).

69. Xu, S., Kong, D., Chen, Q., Ping, Y. & Pang, D. Oncogenic long noncoding RNA landscape in breast cancer. *Mol. Cancer* **16**, 129 (2017).
70. Amorim, M., Salta, S., Henrique, R. & Jerónimo, C. Decoding the usefulness of non-coding RNAs as breast cancer markers. *J. Transl. Med.* **14**, 265 (2016).
71. Shi, S.-J. *et al.* LncRNA-ATB promotes trastuzumab resistance and invasion-metastasis cascade in breast cancer. *Oncotarget* **6**, 11652–11663 (2015).
72. Ventola, G. M. M. *et al.* Identification of long non-coding transcripts with feature selection: a comparative study. *BMC Bioinformatics* **18**, 187 (2017).
73. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
74. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, 951–969 (2015).
75. Galaxy. Available at: https://usegalaxy.org/. (Accessed: 10th January 2018)
76. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).
77. Differential gene expression analysis. *EMBL-EBI Train online* (2016). Available at: https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/differential-gene. (Accessed: 18th October 2018)
78. Wu, D., Rice, C. M. & Wang, X. Cancer bioinformatics: A new approach to systems clinical medicine. *BMC Bioinformatics* **13**, 71 (2012).
79. Ensembl browser 91. Available at: https://www.ensembl.org. (Accessed: 1st March 2018)
80. GENCODE - Home page. Available at: https://www.gencodegenes.org/. (Accessed: 10th January 2018)
81. LNCipedia. Available at: https://lncipedia.org.
82. CPC - Coding Potential Calculator. Available at: http://cpc.cbi.pku.edu.cn. (Accessed: 1st March 2018)
83. UCSC Genome Browser Home. Available at: https://genome.ucsc.edu/. (Accessed: 10th January 2018)
84. WashU EpiGenome Browser. Available at: http://epigenomegateway.wustl.edu/browser. (Accessed: 1st March 2018)
85. BDGP: Home. Available at: http://www.fruitfly.org/. (Accessed: 8th November 2018)
86. Promoter 2.0 Prediction Server. Available at: http://www.cbs.dtu.dk/services/Promoter/. (Accessed: 23rd April 2018)
87. FARNA: Knowledgebase of Annotated Functions of Non-coding RNA Transcripts. Available at: http://www.cbrc.kaust.edu.sa/farna. (Accessed: 1st March 2018)
88. European Nucleotide Archive < EMBL-EBI. Available at: https://www.ebi.ac.uk/ena. (Accessed: 10th January 2018)
89. Primer-Blast. Available at: https://www.ncbi.nlm.nih.gov/tools/primer-blast/.
90. One-Way ANOVA Calculator. Available at: https://www.socscistatistics.com/tests/anova/Default2.aspx. (Accessed: 6th November 2018)
91. ROC Analysis: Web-based Calculator for ROC Curves. Available at: http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html. (Accessed: 26th June 2018)
92. Schoonjans, F. MedCalc's Odds ratio calculator. *MedCalc* Available at: https://www.medcalc.org/calc/odds_ratio.php. (Accessed: 31st October 2018)
93. Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33**, 690-705.e9 (2018).
94. McVeigh, T. P. & Kerin, M. J. Clinical use of the Oncotype DX genomic test to guide treatment decisions for patients with invasive breast cancer. *Breast Cancer Targets Ther.* **9**, 393–400 (2017).

95. Ohara, A. M. *et al.* PAM50 for prediction of response to neoadjuvant chemotherapy for ER-positive breast cancer. *Breast Cancer Res. Treat.* (2018). doi:10.1007/s10549-018-5020-7
96. Wang, T. *et al.* Diagnostic significance of urinary long non-coding PCA3 RNA in prostate cancer. *Oncotarget* **8**, 58577–58586 (2017).
97. Xue, X. *et al.* LncRNA HOTAIR enhances ER signaling and confers tamoxifen resistance in breast cancer. *Oncogene* **35**, 2746–2755 (2016).
98. Cerk, S. *et al.* Current Status of Long Non-Coding RNAs in Human Breast Cancer. *Int. J. Mol. Sci.* **17**, (2016).
99. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120–124 (2011).
100. Zhang, Y. *et al.* Long intergenic non-coding RNA expression signature in human breast cancer. *Sci. Rep.* **6**, 37821 (2016).
101. Hoo, Z. H., Candlish, J. & Teare, D. What is an ROC curve? *Emerg. Med. J. EMJ* **34**, 357–359 (2017).
102. Balasubramanian, H., Ananthan, A., Rao, S. & Patole, S. Odds ratio vs risk ratio in randomized controlled trials. *Postgrad. Med.* **127**, 359–367 (2015).
103. Gamcsik, M. P., Dolan, M. E., Andersson, B. S. & Murray, D. Mechanisms of resistance to the toxicity of cyclophosphamide. *Curr. Pharm. Des.* **5**, 587–605 (1999).
104. Liu, X. *et al.* Digital gene expression profiling analysis and its application in the identification of genes associated with improved response to neoadjuvant chemotherapy in breast cancer. *World J. Surg. Oncol.* **16**, 82 (2018).
105. Chemoresistance. *Dictionary of medical definitions* (2018).
106. Pathologic Complete Response. *Dictionary of cancer terms* (2018).

Apéndice A: Descripción de pacientes y reportes de calidad por muestra de los resultados de secuenciación

A.1: Descripción de pacientes

Tabla A.1: Descripción de las pacientes incluídas en el estudio de RNA-Sec.

Variable		Media	DE
Edad,	años	47.83	7.76
Tamaño del tumor,	cm	6.92	2.68
Ki67,	(%)	45.67	21.84
IMC,	Kg/cm^2	27.49	4.14
Variable		n	%
Edad,	años		
	< 40 años	2	16.67
	40 años o más	10	83.33
Comorbilidad			
	Diabetes mellitus	2	16.67
	Hipertensión	2	16.67
Menopausia		6	50.00
IMC, Kg/cm^2			
	Normopeso (19-24.9)	3	25.00
	Sobrepeso (25-29)	6	50.00
	Obesidad (>30)	3	25.00
Estadio clínico			
	IIB	3	25.00
	IIIA	7	58.33
	IIIB	2	16.67
Tipo Histológico	Carcinoma ductal		
	infiltrante	12	100.00
Grado histológico			
	1	0	0
	2	6	50.00
	3	6	50.00
Sensibilidad a QTNeo			
	RPC	2	16.67
	Parcial	3	25.00
	Resistencia	7	58.33
Fenotipo			
	RE/RP(+)	12	100.00
	RE/RP (-)	0	0
	Luminal A	1	8.33
	Luminal B	11	91.66
	HER2+	4	33.34
	HER2-	8	66.68

DE: Desviación estándar, IMC: índice de masa corporal. HER2(+): Tumores HER2 enriquecidos

A.2: Características clínicas de las pacientes incluidas en el estudio.

Tabla A.2: Características de las pacientes incluidas en el estudio **(cohorte completa).**

ID	Subtipo Molecular	Etapa Clínica	Edad	Respuesta Patológica	Expresión de HER2	Expresión de RE	Expresión de RP	Expresión de Ki67 (%)	NIA	Inclusión en estudio de RNA-Sec	Total de Lecturas (millones)
Grupo de Casos (QR)											
CM-1	Luminal B	IIIB	52	Resistencia	-	+	+	60	7.2	Sí	25
CM-2	Luminal B	IIB	66	Resistencia	+	+	+	90	7.4	Sí	41
CM-3	Luminal B	IIIA	52	Resistencia	-	+	+	18	8.9	Sí	26
CM-4	Luminal B	IIIA	49	Resistencia	-	+	+	25	7.5	Sí	41
CM-5	Luminal B	IIIA	55	Resistencia	-	+	+	35	7.9	Sí	38
CM-6	Luminal B	IIB	42	Resistencia	-	+	+	30	8.2	Sí	36
CM-8	Luminal A	IIIA	41	Resistencia	-	+	+	10	NA	Sí	NA
CM-12	Luminal B	IIIA	39	Resistencia	+	+	+	50	9.1	Sí	25
CM-13	Luminal B	IIIA	65	Resistencia	-	+	+	20	6.9	No	-
CM-14	Luminal B	IIIB	45	Resistencia	-	+	+	80	6.5	No	-
CM-15	Luminal B	IIIA	46	Resistencia	-	+	+	40	7.0	No	-
CM-16	Luminal B	IIIB	52	Resistencia	-	+	+	40	6.8	No	-
CM-17	Luminal B	IIIA	57	Resistencia	-	+	+	20	6.8	No	-
CM-18	Luminal B	IIIA	54	Resistencia	-	+	+	30	6.8	No	-
CM-19	Luminal B	IIB	39	Resistencia	-	+	+	20	6.8	No	-
Grupo Control RPC											
CM-7	Luminal B	IIIA	47	Completa	-	+	+	80	8.9	Sí	49
CM-9	Luminal B	IIIA	68	Completa	+	+	+	50	8.5	Sí	28
CM-10	Luminal B	IIIB	48	Completa	+	-	+	50	8.9	Sí	20
CM-11	Luminal B	IIB	40	Completa	-	+	+		8.6	Sí	17
CM-20	Luminal B	IIIA	45	Completa	-	-	+	60	5.6	No	-
CM-21	Luminal B	IIIA	44	Completa	+	-	+	20	8.8	No	-
CM-22	Luminal B	IIIA	50	Completa	-	+	+	50	7.2	No	-
CM-23	Luminal B	IIIA	54	Completa	+	-	+	40	8.1	No	-
CM-24	Luminal B	IIIB	40	Completa	-	+	+	50	4.8	No	-
CM-25	Luminal B	IIB	58	Completa	-	+	+	50	6.6	No	-

CM-26	Luminal B	IIB	55	Completa	+	+	+	40	7.5	No	-
CM-27	Luminal B	IIIA	44	Completa	+	+	+	40	8.4	No	-
CM-28	Luminal B	IIIA	39	Completa	-	+	-	90	8.5	No	-
CM-29	Luminal B	IIIC	62	Completa	+	+	+	30	7.0	No	-

Apéndice B: Reportes de calidad

En esta sección se muestran los reportes de calidad completos de las muestras CM-4 y CM-10, representativas del grupo de casos QR y del grupo control RPC, respectivamente, Se incluyen los reportes de las lecturas 1 y 2, ya que el experimento de RNA-Sec fue del tipo extremos pareados (*paired-end)*.

B.1.1 Muestra CM-4 (Lectura 1)

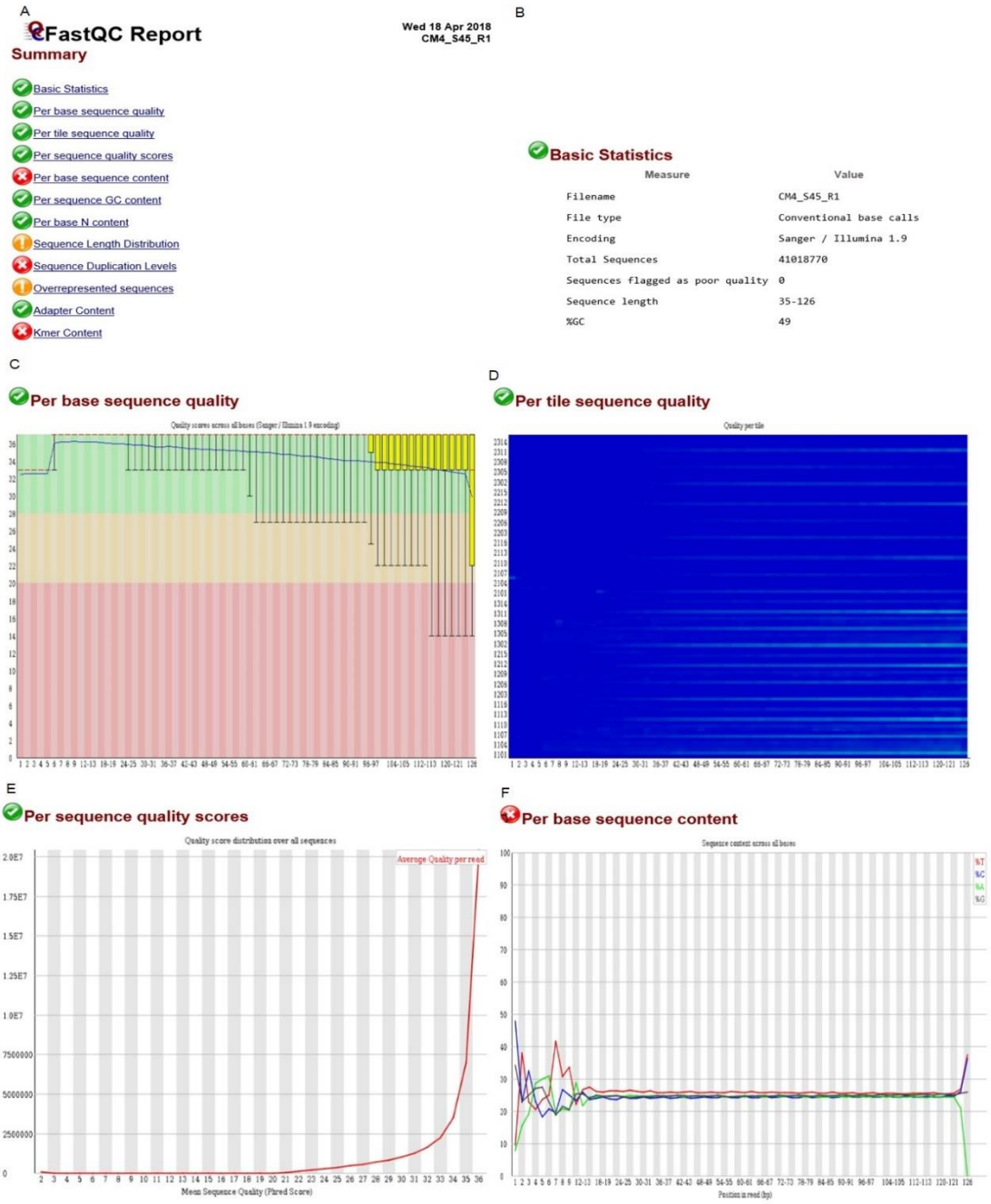


Figura B1. Reporte de calidad de la muestra CM-4. A) Parámetros del reporte de calidad. B) Cuadro de estadísticos básicos. C) Gráfico de análisis de la calidad de la secuencia por base del fragmento de RNA

secuenciado. D) Gráfico de azulejos que analiza la calidad de la secuencia por posición de la base. E) Gráfico de calidad por secuencia que muestra la cantidad de secuencias con baja calidad por posición del fragmento de RNA. F) Gráfico de contenido de base por secuencia.

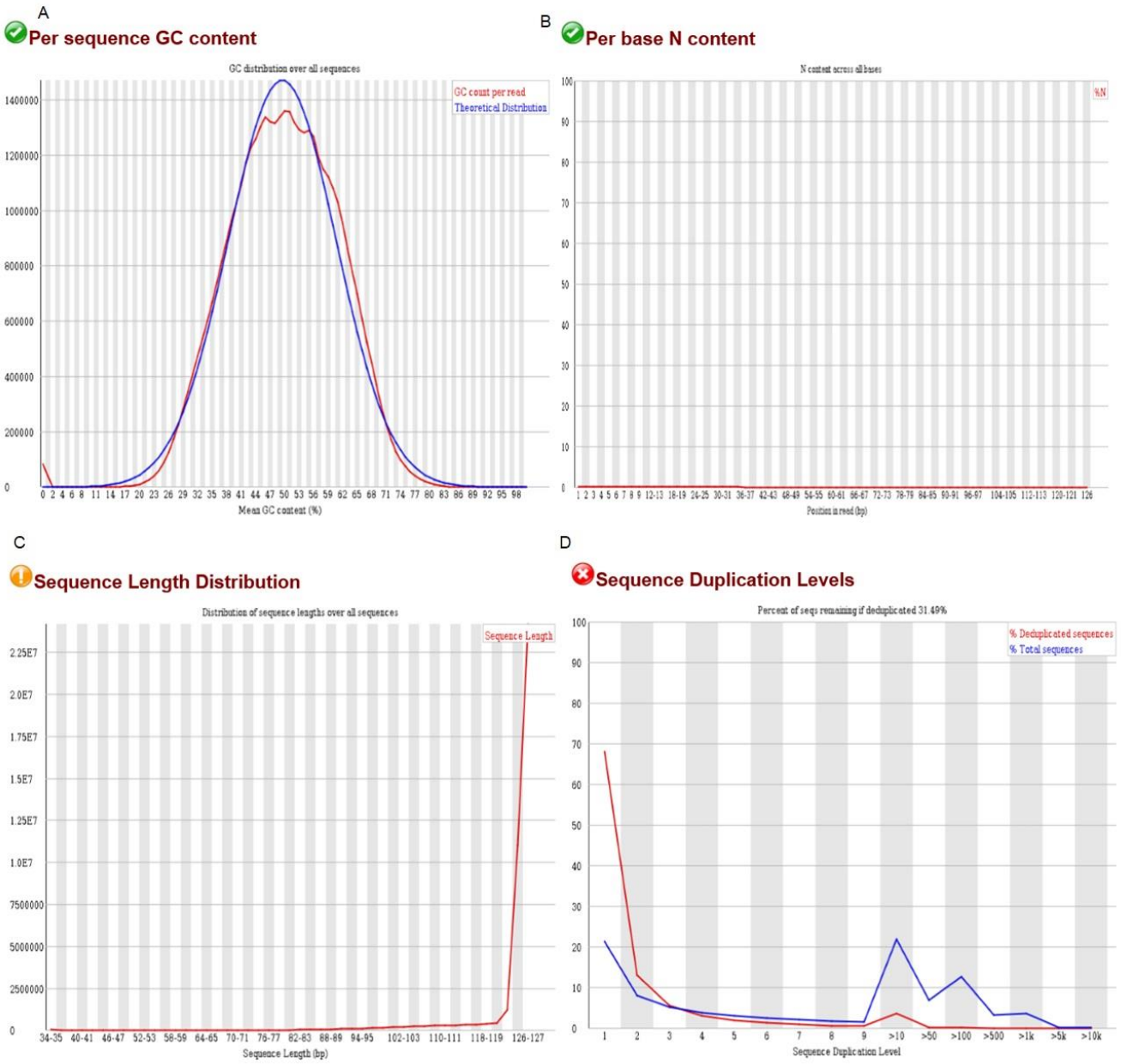


Figura B2. Reporte de calidad de la muestra CM-4. A) Contenido global de las bases G y C por posición en la secuencia, que consiste en evaluar el promedio de bases G y C detectadas en la secuenciación (rojo) y el valor estimado por el equipo de secuenciación (azul). B) contenido por base de nucleótidos no identificados, que debe ser igual a cero. C) Gráfico de distribución de longitudes de secuencia, cuyos valores deben encontrarse en el intervalo de 120-127 nt para ser “aceptable”. D) Gráfico de proporción de secuencias duplicadas, que evalúa que la cantidad de secuencias duplicadas (rojo) sea menor respecto al total de secuencias detectadas (azul).

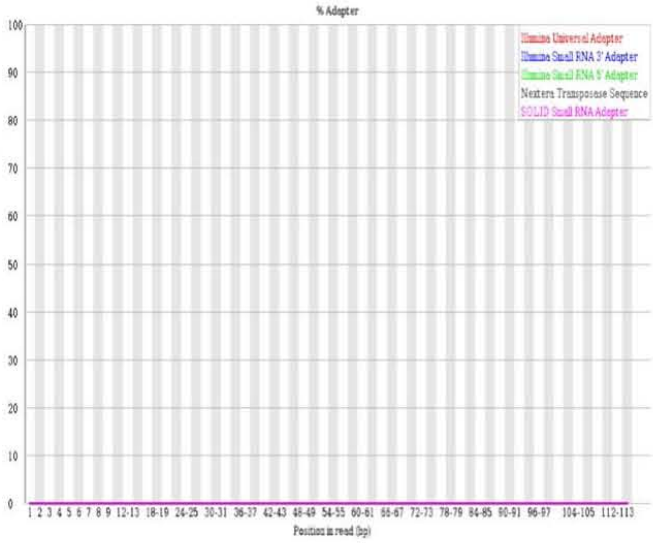
A

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	83929	0.20461120604055166	No Hit

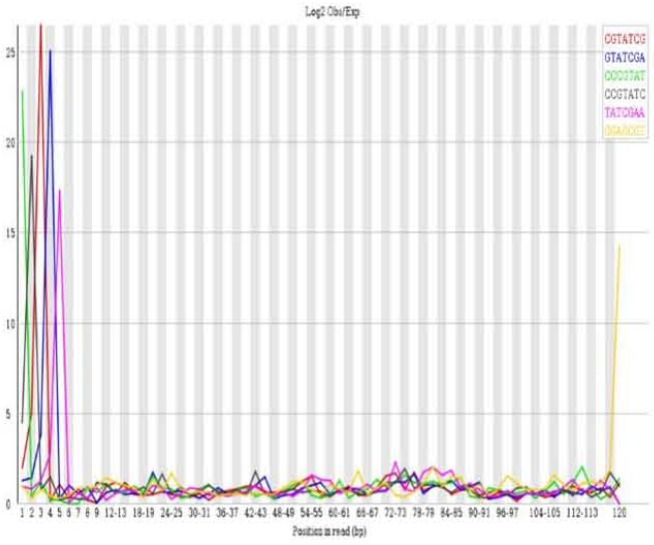
B

Adapter Content



C

Kmer Content



D

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CGTATCG	6950	0.0	26.427357	3
GTATCGA	6920	0.0	25.015312	4
CCCGTAT	9330	0.0	22.775791	1
CCGTATC	10215	0.0	19.244236	2
TATCGAA	10610	0.0	17.310852	5
GGAGCGT	17540	0.0	14.270679	120
CGATAAC	9900	0.0	14.171284	1
TACCGAC	2905	0.0	13.676897	120
ATCGAAG	16195	0.0	13.442556	6
ATAACGT	10820	0.0	13.124487	3
TAAGCGA	4105	0.0	13.008199	9
CTCGTTT	12330	0.0	12.473395	1
TTCGCGC	2940	0.0	12.3746605	7
CTCGGTT	10405	0.0	11.565336	1
CTAAGCG	4660	0.0	11.458939	8
TATCGCG	1605	0.0	11.3338	9
GTCCGAT	4920	0.0	11.095946	1
ATACGTA	5030	0.0	10.998247	112-113
CTCGTAT	6940	0.0	10.911306	1
CGCGTAA	2390	0.0	10.806903	1

Produced by FastQC (version 0.11.5)

Figura B3. Reporte de calidad de la muestra CM-4. A) Secuencias sobrerrepresentadas (que debe ser cero), B) Contenido de adaptadores (que debe ser cero), C) contenido de Kmer, que debe ser cercano a cero y D) la Tabla de porcentajes de representación de los Kmer identificados.

B.1.2 Muestra CM-4 (Lectura 2)

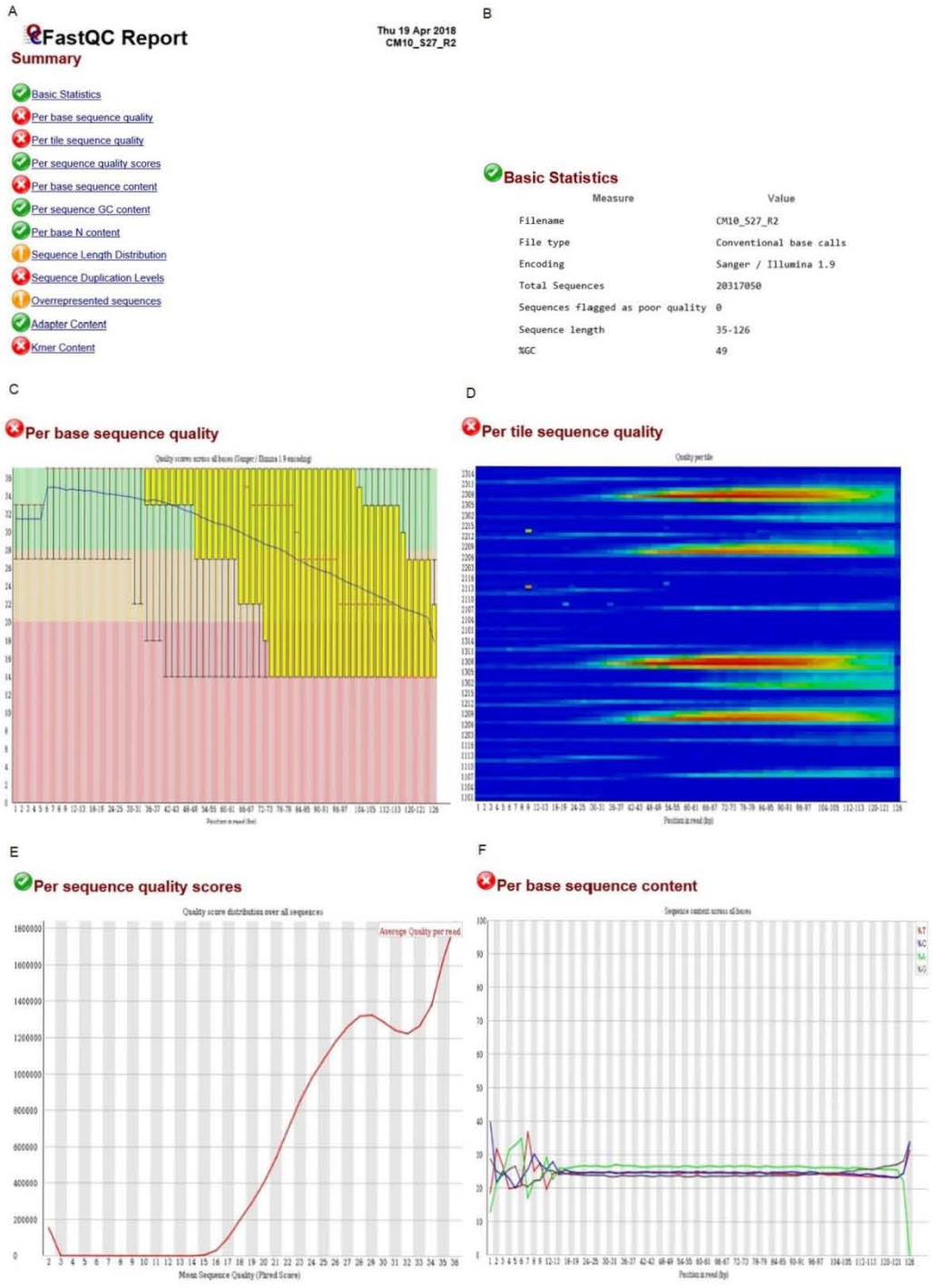


Figura B4. Reporte de calidad de la muestra CM-4. A) Parámetros del reporte de calidad. B) Cuadro de estadísticos básicos. C) Gráfico de análisis de la calidad de la secuencia por base del fragmento de RNA secuenciado. D) Gráfico de azulejos que analiza la calidad de la secuencia por posición de la base. E) Gráfico de calidad por secuencia que muestra la cantidad de secuencias con baja calidad por posición del fragmento de RNA. F) Gráfico de contenido de base por secuencia.

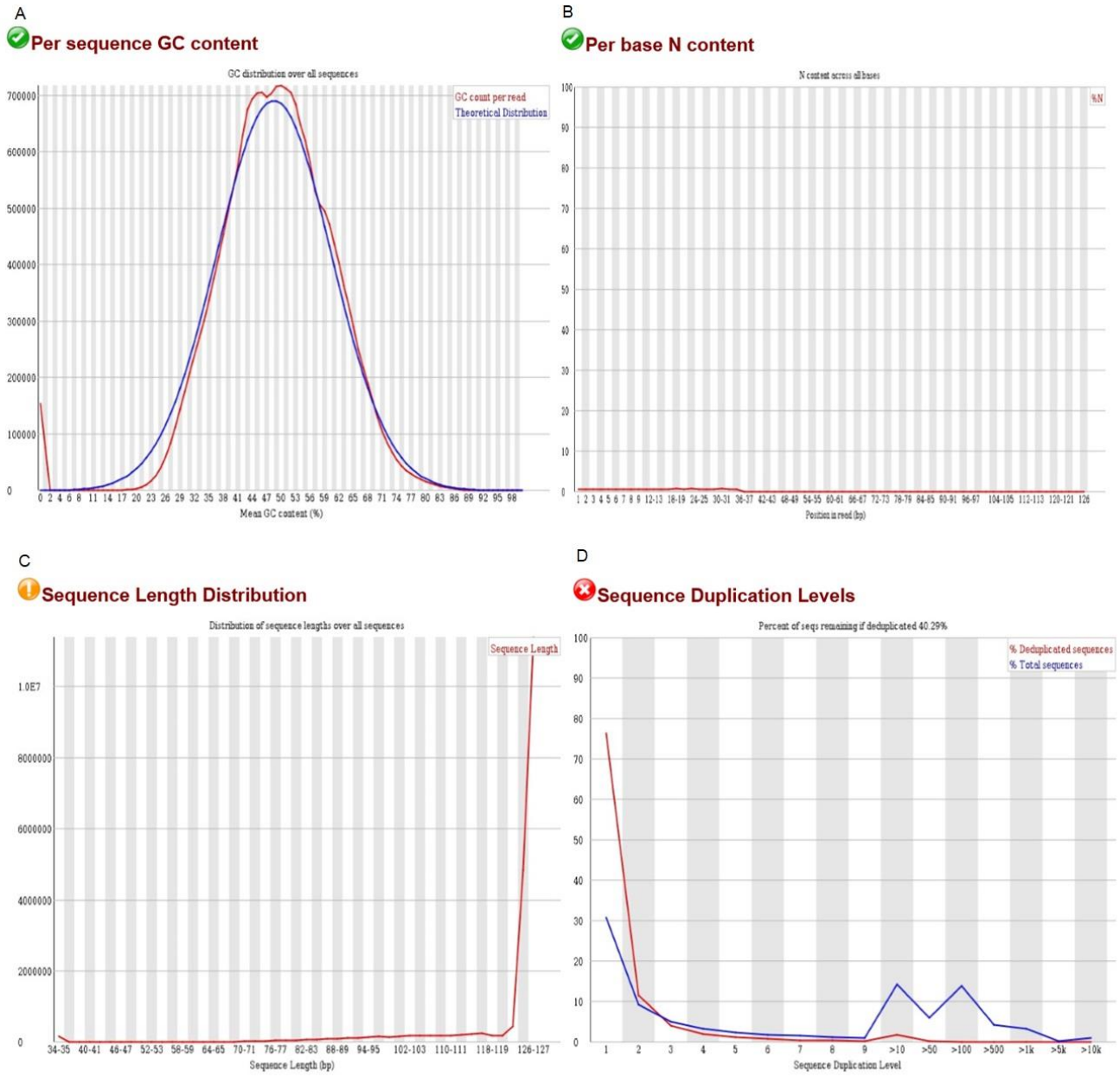


Figura B5. Reporte de calidad de la muestra CM-4. A) Contenido global de las bases G y C por posición en la secuencia, que consiste en evaluar el promedio de bases G y C detectadas en la secuenciación (rojo) y el valor estimado por el equipo de secuenciación (azul). B) contenido por base de nucleótidos no identificados, que debe ser igual a cero. C) Gráfico de distribución de longitudes de secuencia, cuyos valores deben encontrarse en el intervalo de 120-127 nt para ser Aceptable. D) Gráfico de proporción de secuencias duplicadas, que evalúa que la cantidad de secuencias duplicadas (rojo) sea menor respecto al total de secuencias detectadas (azul).

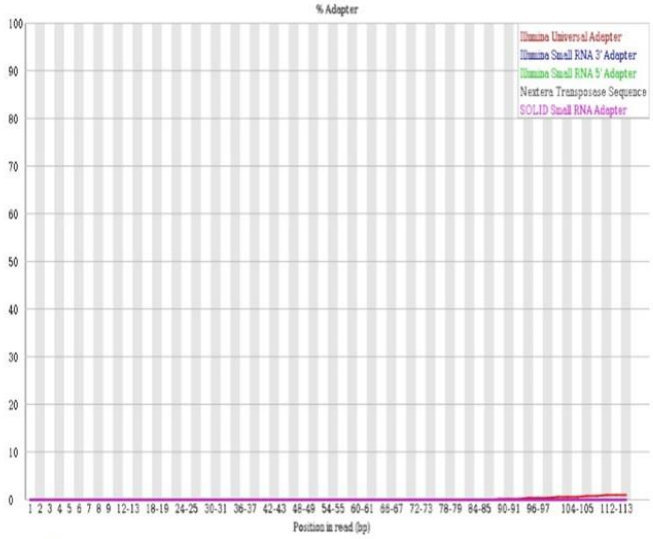
A

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	153844	0.7572162297183892	No Hit

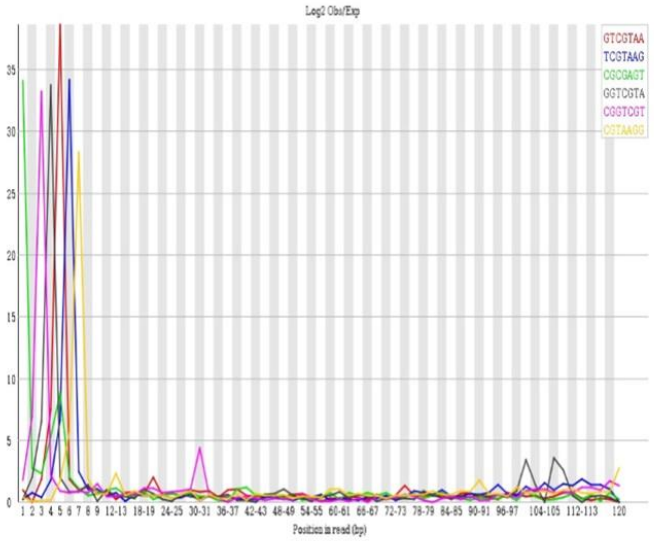
B

Adapter Content



C

Kmer Content



D

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GTCGTAA	3885	0.0	38.598145	5
TCGTAAG	4435	0.0	34.201416	6
CGCGAGT	3905	0.0	34.11403	1
GGTCGTA	4805	0.0	33.726574	4
CGGTCGT	4510	0.0	33.24736	3
CGTAAGG	5505	0.0	28.287027	7
CGCGGTC	5745	0.0	27.303665	1
GCGGTCG	5720	0.0	27.020124	2
CGACCTA	3070	0.0	22.166798	4
GCGAACG	3120	0.0	20.887394	3
TACGCAC	3190	0.0	19.887918	9
CGGCGAA	3530	0.0	19.277441	1
GCGAGTG	7375	0.0	18.92352	2
AACGCGA	2095	0.0	17.067427	3
TAAGGGC	9360	0.0	16.821865	9
CGCAAAT	12445	0.0	16.800741	28-29
GAACGCG	4050	0.0	16.37672	5
GGCGAAC	4235	0.0	15.796219	2
TAGGTCG	2465	0.0	15.675387	3
CAACGCG	2585	0.0	15.616585	2

Produced by FastQC (version 0.11.5)

Figura B6. Reporte de calidad de la muestra CM-4. A) Secuencias sobrerrepresentadas (que debe ser cero), B) Contenido de adaptadores (que debe ser cero), C) contenido de Kmer, que debe ser cercano a cero y D) la tabla de porcentajes de representación de los Kmer identificados.

B.1.3 Muestra CM-10 (Lectura 1)

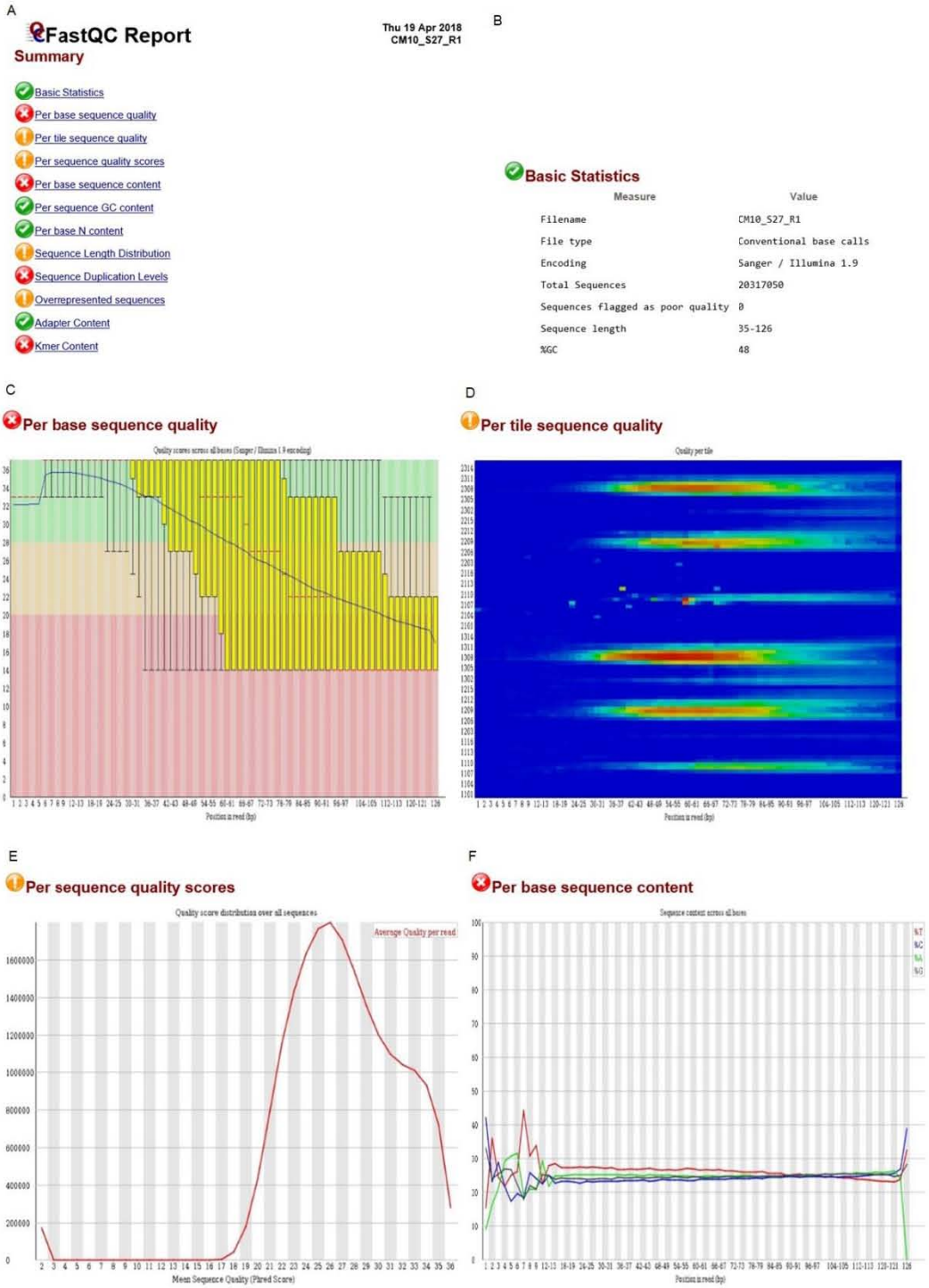


Figura B7. Reporte de calidad de la muestra CM-10. A) Parámetros del reporte de calidad. B) Cuadro de estadísticos básicos. C) Gráfico de análisis de la calidad de la secuencia por base del fragmento de RNA secuenciado. D) Gráfico de azulejos que analiza la calidad de la secuencia por posición de la base. E) Gráfico de calidad por secuencia que muestra la cantidad de secuencias con baja calidad por posición del fragmento de RNA. F) Gráfico de contenido de base por secuencia.

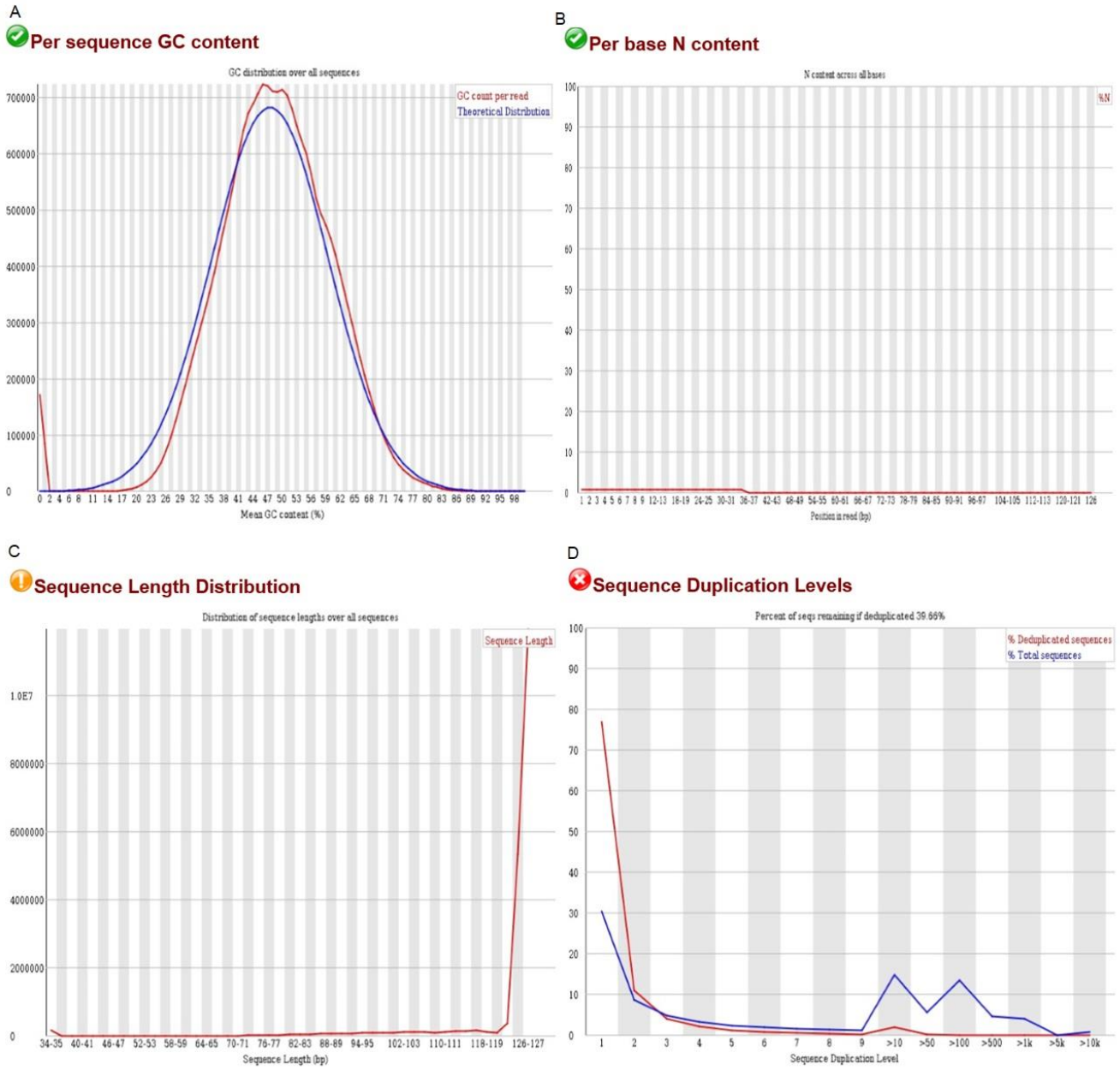


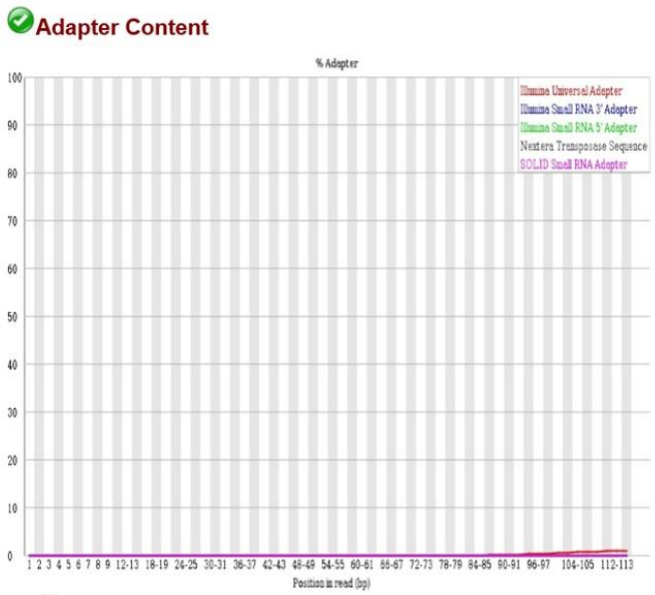
Figura B8. Reporte de calidad de la muestra CM-10. A) Contenido global de las bases G y C por posición en la secuencia, que consiste en evaluar el promedio de bases G y C detectadas en la secuenciación (rojo) y el valor estimado por el equipo de secuenciación (azul). B) contenido por base de nucleótidos no identificados, que debe ser igual a cero. C) Gráfico de distribución de longitudes de secuencia, cuyos valores deben encontrarse en el intervalo de 120-127 nt para ser Aceptable. D) Gráfico de proporción de secuencias duplicadas, que evalúa que la cantidad de secuencias duplicadas (rojo) sea menor respecto al total de secuencias detectadas (azul).

A

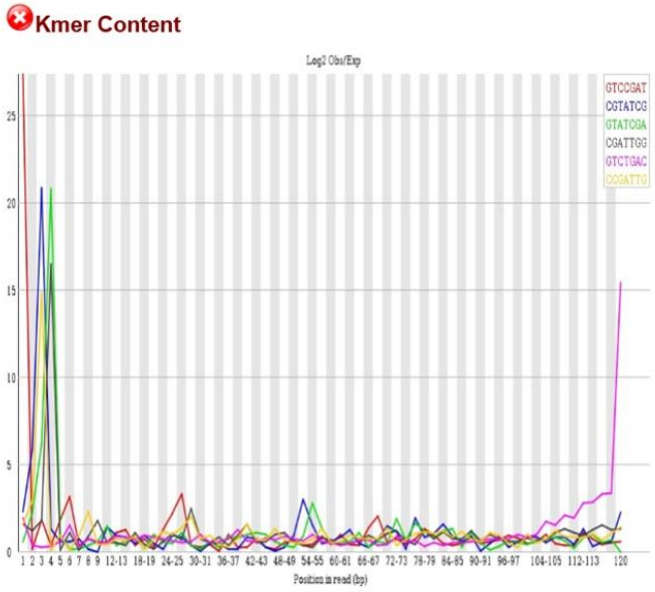
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	171727	0.8452358979280948	No Hit

B



C



D

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GTCCGAT	3135	0.0	27.31124	1
CGTATCG	2990	0.0	20.8386	3
GTATCGA	2890	0.0	20.753588	4
CGATTGG	5090	0.0	16.47398	4
GTCTGAC	10695	0.0	15.4263	120
CCGATTG	5520	0.0	14.87428	3
CCCGTAT	6075	0.0	13.998072	1
GTCGGAC	3875	0.0	13.685332	120
TCCGATT	7630	0.0	13.203148	2
CTGAACC	21190	0.0	13.115618	120
TATCGAA	7915	0.0	13.095524	5
CGAGATG	4300	0.0	13.017862	120
CCGTATC	6390	0.0	12.758017	2
AACTCCG	8645	0.0	12.154919	120
ACCGTCG	2255	0.0	11.758461	120
CGCGTAA	1200	0.0	11.163701	1
GATAACG	4535	0.0	11.0427475	2
CGATAAC	4650	0.0	10.897526	1
ATATCGT	2430	0.0	10.783507	4
ATAACGT	4575	0.0	10.564354	3

Produced by FastQC (version 0.11.5)

Figura B9. Reporte de calidad de la muestra CM-10. A) Secuencias sobrerrepresentadas (que debe ser cero), B) Contenido de adaptadores (que debe ser cero), C) contenido de Kmer, que debe ser cercano a cero y D) la tabla de porcentajes de representación de los Kmer identificados.

B.1.4 Muestra CM-10 (Lectura 2)

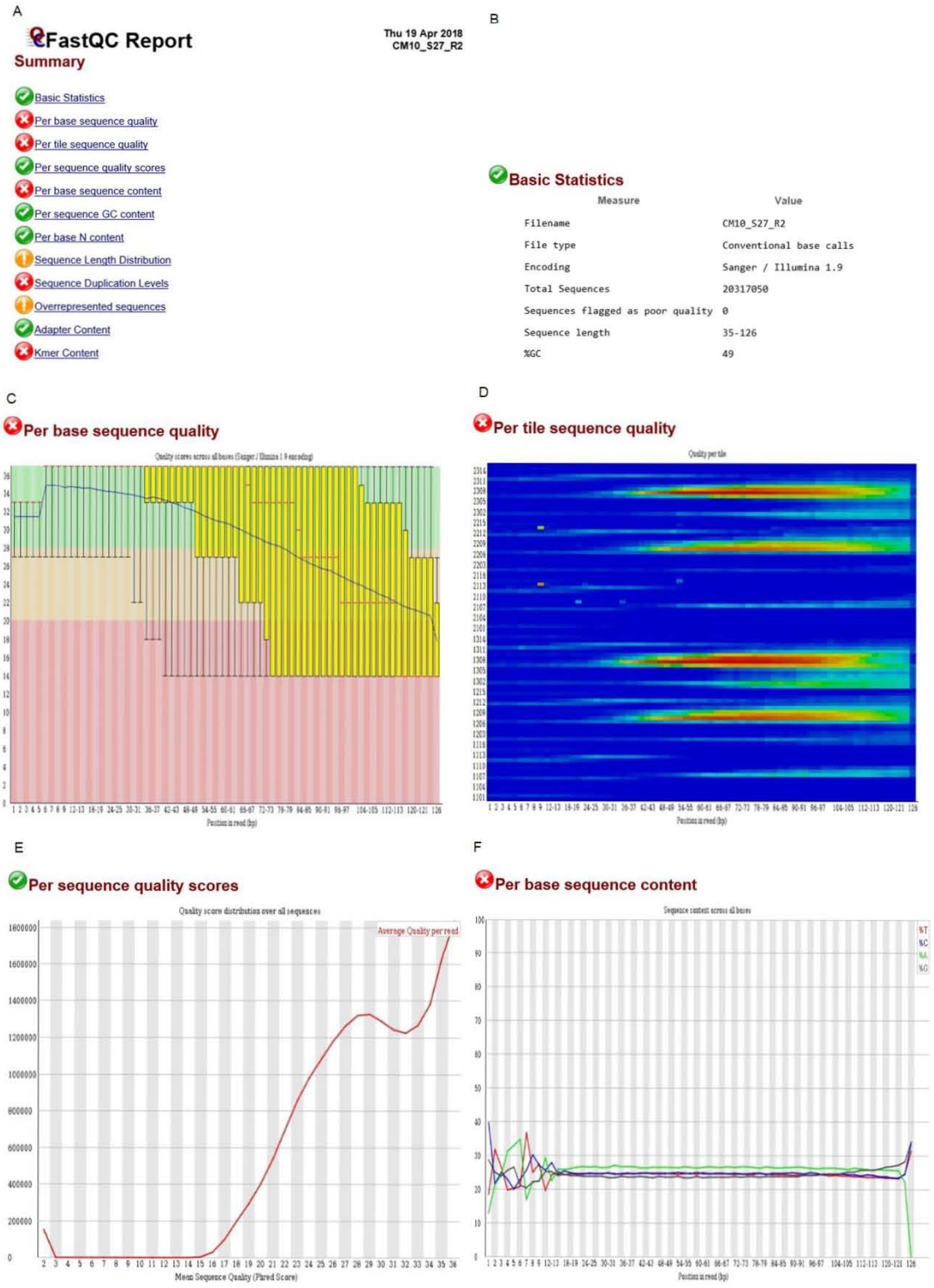


Figura B10. Reporte de calidad de la muestra CM-10. A) Parámetros del reporte de calidad. B) Cuadro de estadísticos básicos. C) Gráfico de análisis de la calidad de la secuencia por base del fragmento de RNA secuenciado. D) Gráfico de azulejos que analiza la calidad de la secuencia por posición de la base. E) Gráfico de

calidad por secuencia que muestra la cantidad de secuencias con baja calidad por posición del fragmento de RNA. F) Gráfico de contenido de base por secuencia.

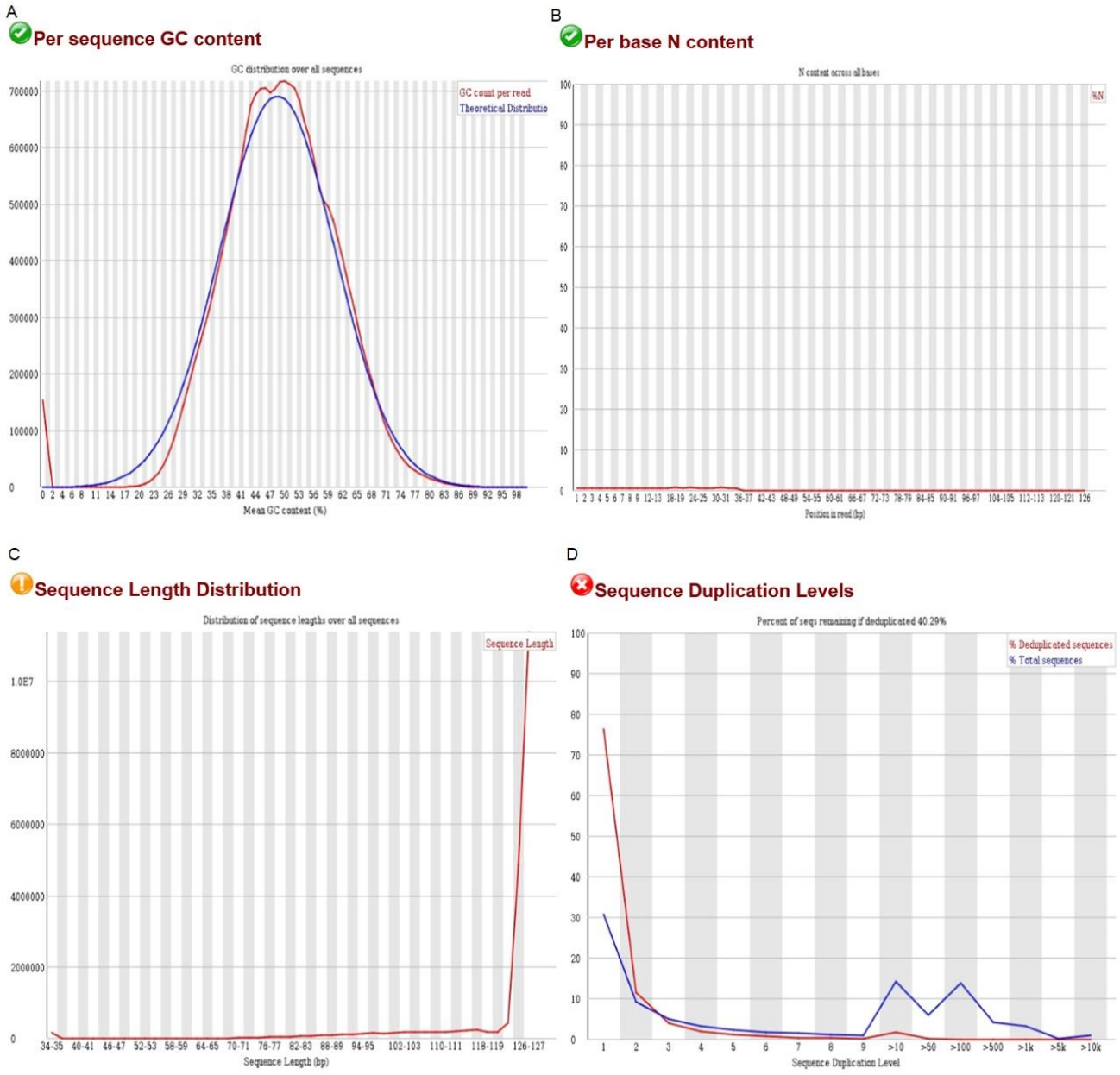


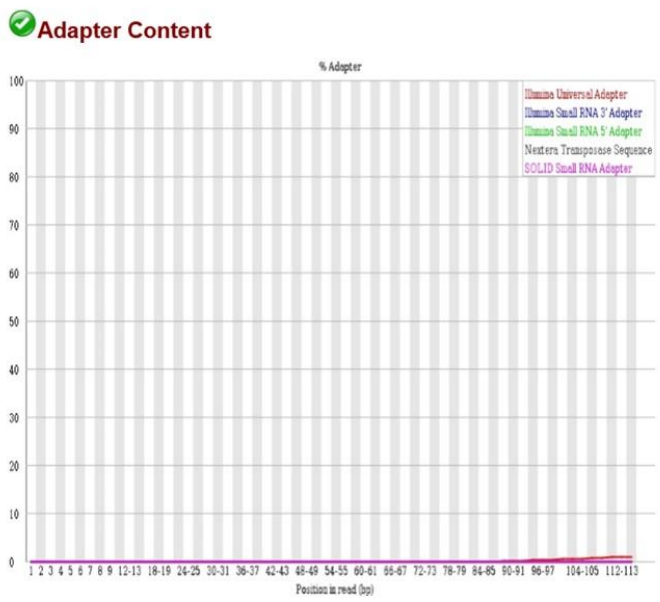
Figura B11. Reporte de calidad de la muestra CM-10. A) Contenido global de las bases G y C por posición en la secuencia, que consiste en evaluar el promedio de bases G y C detectadas en la secuenciación (rojo) y el valor estimado por el equipo de secuenciación (azul). B) contenido por base de nucleótidos no identificados, que debe ser igual a cero. C) Gráfico de distribución de longitudes de secuencia, cuyos valores deben encontrarse en el intervalo de 120-127 nt para ser Aceptable. D) Gráfico de proporción de secuencias duplicadas, que evalúa que la cantidad de secuencias duplicadas (rojo) sea menor respecto al total de secuencias detectadas (azul).

A

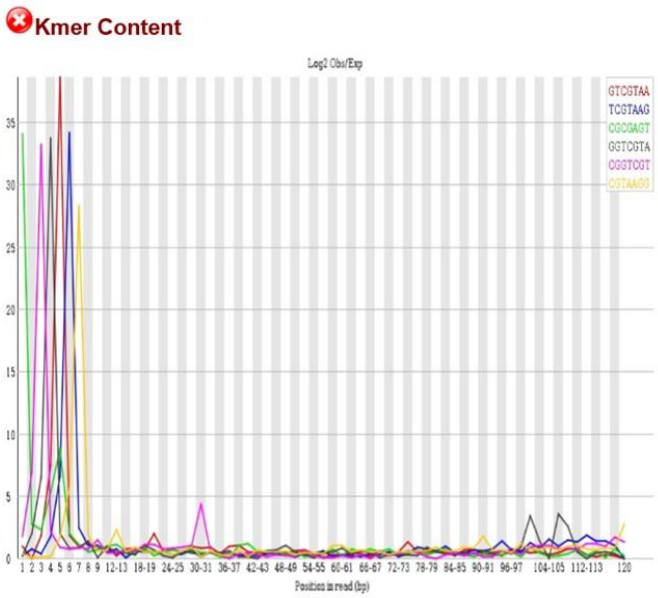
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	153844	0.7572162297183892	No Hit

B



C



D

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GTCGTAA	3885	0.0	38.598145	5
TCGTAAG	4435	0.0	34.201416	6
CGCGAGT	3905	0.0	34.11403	1
GGTCGTA	4805	0.0	33.726574	4
CGGTCGT	4510	0.0	33.24736	3
CGTAAGG	5505	0.0	28.287027	7
CGCGGTC	5745	0.0	27.303665	1
GCGGTCG	5720	0.0	27.020124	2
CGACCTA	3070	0.0	22.166798	4
GCGAACG	3120	0.0	20.887394	3
TACGCAC	3190	0.0	19.887918	9
CGGCGAA	3530	0.0	19.277441	1
GCGAGTG	7375	0.0	18.92352	2
AACGCGA	2095	0.0	17.067427	3
TAAGGGC	9360	0.0	16.821865	9
CGCAAAT	12445	0.0	16.800741	28-29
GAACGCG	4050	0.0	16.37672	5
GGCGAAC	4235	0.0	15.796219	2
TAGGTCG	2465	0.0	15.675387	3
CAACGCG	2585	0.0	15.616585	2

Produced by FastQC (version 0.11.5)

Figura B12. Reporte de calidad de la muestra CM-10. A) Secuencias sobrerrepresentadas (que debe ser cero), B) Contenido de adaptadores (que debe ser cero), C) contenido de Kmer, que debe ser cercano a cero y D) la tabla de porcentajes de representación de los Kmer identificados.

Apéndice C. Resultados del análisis de expresión diferencial de los genes codificantes entre el grupo control (QR) y el grupo de casos (RPC).

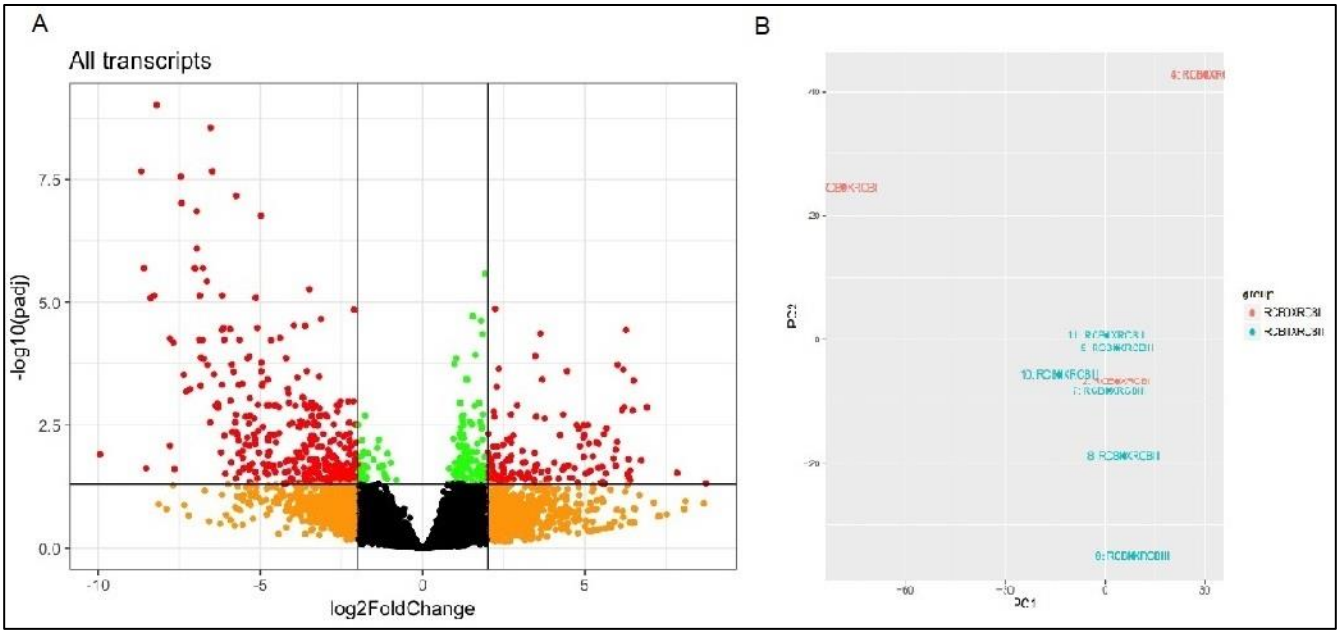


Figura C1. Resultados del análisis de expresión diferencial de mRNA y lncRNA entre el grupo control QR y el grupo de casos RPC. A) Gráfico de tipo volcán que representa los 169 genes sobreexpresados y los 231 genes subexpresados en el grupo control QR (rojo, cuadrantes derecho e izquierdo, respectivamente). B) Análisis de componente principal basado en el perfil transcripcional completo de las muestras de pacientes, que no es capaz de distinguir entre los grupos. Análisis realizados considerando un valor de FDR < 0.05.

Apéndice D: Habilidades aprendidas durante la realización del proyecto de tesis

Durante la realización de este proyecto, las habilidades adquiridas fueron:

- Análisis Bioinformático:
  - Realizar e interpretar análisis de calidad de datos de secuenciación de RNA-Sec.
  - Manejo de herramientas bioinformáticas en el servidor Galaxy y en el lenguaje de programación R para realizar alineamiento de secuencias al genoma de referencia, cuantificación de transcritos, análisis de expresión diferencial.
  - Manejo de plataformas en línea para análisis de regiones promotoras y localiación celular.
  - Manejo del lenguaje R para análisis estadístico y análisis bioinformático.
  - Diseño de oligonucleótidos.
  - Cuantificación relativa mediante PCR en tiempo real por los métodos ΔCq y ΔΔCq.
- Habilidades Experimentales
  - Cultivo celular.
  - Extracción de RNA total de líneas celulares y muestras de pacientes.
  - Cuantificación de RNA y análisis de integridad del RNA.
  - RT-PCR punto final para la síntesis de cDNA.
  - PCR en tiempo real para cuantificación relativa de expresión génica.

Apéndice E: Portada de la publicación del capítulo, en el libro *Analyzing Network Data in Biology and Medicine*, de la editorial Cambridge University Press

Analyzing Network Data in Biology and Medicine

An interdisciplinary textbook for biological, medical and computational scientists

Edited by

Nataša Pržulj
University College London

2 Epigenetic Data and Disease

Rodrigo González-Barrios, Marisol Salgado-Albarrán, Nicolás Alcaráz,
Cristian Arriaga-Canon, Lissania Guerra-Calderas, Laura Contreras-Espinoza and Ernesto Soto-Reyes

GLOSSARY

Chromatin conformation capture (5C, HiC): A technique used to profile all chromatin interactions in specific regions of the genome by the hybridization of a mixture of DNA primers to chromosome conformation capture (3C) templates followed by high-throughput sequencing.

Cis-acting: regions of non-coding DNA, which regulate transcription within the same chromosome.

CpG Island: Genomic regions with a minimum of 200 bp, with a G–C content greater than 50% and observed/expected CpG ratio above 60%.

Enhancer: A cis-acting regulatory sequence that markedly increases expression of a neighbouring gene. Enhancers are typically capable of operating over considerable distances (sometimes ~50 kb) upstream or downstream of the gene and in either orientations.

Epigenomics: Is the systematic analysis of the global state of gene expression modulated by epigenetic processes such as DNA methylation, posttranslational modifications of histones non-coding RNA and the organization of chromatin inside the nucleus.

Euchromatin: Less densely packed or open chromatin that is often associated with active transcription.

Global hypomethylation: Loss of DNA methylation across the genome that commonly occurs in cancer cells.

Heterochromatin: Tightly packed form of chromatin that lack high number of genes and is commonly constituted by repetitive sequences in the genome, which is associated with inactive transcription and serves as a structural element of the chromosome.

Local Hypermethylation: Gain of methylation that occurs at specific regulatory regions that alters the normal state of transcription in diseases like cancer.

Hi-C contact matrix: Matrix which displays all chromatin interactions found within a genomic range. Firstly, the genome is partitioned into bins of fixed size. Then, a

Apéndice F: Carátula del documento oficial con la aceptación de la solicitud de patente para la técnica de detección del biomarcador lincRNA-ACR para predecir resistencia a la QTNeo

gob mx

Instituto Mexicano de la Propiedad Industrial

Solicitud de Patente de Invención o de Registro de Modelo de Utilidad o de Registro de Diseño Industrial

Homoclave del formato
IMPI-00-009
Fecha de publicación del formato en el DOF
24 / 05 / 2018

Folio y Fecha de Recepción
INSTITUTO MEXICANO DE LA PROPIEDAD INDUSTRIAL
Dirección Divisional de Patentes
Solicitud MX/a/2018/015065
Expediente
Fecha: 5/DIC/2018 Hora: 11:51:08
Folio: MX/E/2018/090583 304888

Datos generales de la solicitud
(X) Solicitud de Patente de Invención
( ) Solicitud de Registro de Modelo de Utilidad
( ) Solicitud de Registro de Diseño Industrial, especifique.
( ) Modelo Industrial ( ) Dibujo Industrial

Datos generales del o de los solicitante(s)

Personas físicas
CURP (opcional):
Nombre(s):
Primer apellido:
Segundo apellido:
Nacionalidad:
Teléfono (lada, número, extensión):
Correo electrónico (opcional):
( ) Continúa en anexo

Personas morales
RFC (opcional): INC481125HL8
Denominación o razón social:
INSTITUTO NACIONAL DE CANCEROLOGÍA
Nacionalidad: MEXICANA
Teléfono (lada, número, extensión):
01 55 5628 0411
Correo electrónico (opcional): vmoralesp@incan.edu.mx
( ) Continúa en anexo

Domicilio del o de los solicitante(s)
Código postal: 14080
Calle: Avenida San Fernando
Número exterior: 22
Número interior:
Colonia: Sección XVI
Municipio o demarcación territorial: Tlalpan
Localidad
Entidad Federativa: Ciudad de México
Entre calles (opcional): Enoch Cancino y Tlalpan
País: Mexico
Calle posterior (opcional):

Datos generales del o de los inventor(es) o diseñador(es)
CURP (opcional):
Nombre(s): CRISTIAN GABRIEL OLIVERIO
Primer apellido: ARRIAGA
Segundo apellido: CANON
Nacionalidad: MEXICANA
Teléfono (lada, número, extensión): 044 55 6249 5845
Correo electrónico (opcional): cristiancanon@hotmail.com
(X) Continúa en anexo

MÉXICO GOBIERNO DE LA REPÚBLICA COFEMER IMPI

Contacto:
Página 1 de 4

gob mx

Instituto Mexicano de la Propiedad Industrial

Domicilio del o de los inventor(es) o diseñador(es)	
Código postal: 16050	
Calle: Portal	
Número exterior: 4	Número interior: 4
Colonia: Jardines del Sur	
Municipio o demarcación territorial: Xochimilco	Localidad:
Entidad Federativa: Ciudad de México	Entre calles (opcional)
País: México	Calle posterior (opcional)

Datos generales del o de los apoderado(s)	
CURP (opcional):	Registro General de Poderes (opcional): RGP-DCAJ-33939
Nombre(s): VICTOR RAMÓN	RFC (opcional):
Primer apellido: MORALES	Teléfono (lada, número, extensión): 01 55 5628 0411
Segundo apellido: PEÑA	Correo electrónico (opcional): vmoraleop@incan.edu.mx
	◯ Continúa en anexo

Domicilio para oír y recibir notificaciones	
Código postal: 14080	
Calle: Avenida San Fernando	
Número exterior: 2	Número interior: Puerta 1
Colonia: Barrio del Niño Jesús	
Municipio o demarcación territorial: Tlalpan	Localidad:
Entidad Federativa: Ciudad de México	Entre calles (opcional): Viaducto Tlalpan y Niño Jesús
País: México	Calle posterior (opcional)

Datos generales de los autorizados para oír y recibir notificaciones			
Nombre(s):	Primer apellido:	Segundo apellido:	CURP (opcional):
VICTOR RAMÓN	MORALES	PEÑA	◯ Continúa en anexo

Datos de la solicitud
Denominación o título de la invención, modelo de utilidad o diseño industrial: BIOMARCADOR MOLECULAR PARA LA PREDICCIÓN DE LA RESPUESTA A LA QUIMIOTERAPIA NEOADYUVANTE EN PACIENTES CON CANCER DE MAMA LOCALMENTE AVANZADO, MEDIANTE DETECCIÓN POR PCR.
Fecha de divulgación previa (DD / MM / AAAA): / /

Divisional de la solicitud	
No. Expediente en trámite:	Figura jurídica:
Fecha de presentación (DD / MM / AAAA): / /	

PCT
No. de solicitud internacional:
Fecha de presentación internacional (DD / MM / AAAA): / /

Prioridad o prioridades reclamada(s)		
País (oficina) de origen:	Fecha de presentación (DD/MM/AAAA):	Número de serie:
	/ /	
		◯ Continúa en anexo

Bajo protesta de decir verdad, manifiesto que los datos asentados en esta solicitud son ciertos.

VICTOR RAMÓN MORALES PEÑA

Nombre y firma del solicitante o su apoderado.

Apéndice G: Glosario

Cobertura: Número de veces que una base está presente en las lecturas de secuenciación.

Especificidad: Probabilidad de que la prueba o biomarcador utilizado detecte a un individuo sano, es decir, que presente un resultado negativo en el estudio.

Estudio de casos y controles anidado: Un estudio de casos y controles es un estudio epidemiológico, observacional y analítico, en el cual los sujetos son seleccionados en función de si presentan (casos) o no (controles) el efecto de interés, para después asociarlo con una característica particular. Cuando los individuos son seleccionados de una población que pertenece a una cohorte, se le denomina estudio de tipo anidado.

Genes conductores: Genes cuyas mutaciones o alteraciones en la función se relacionan directamente con la progresión del cáncer.

Incidencia: Tasa de morbilidad que indica la cantidad de personas que se diagnostican con una enfermedad por cada 100,000 habitantes.

Medicina de Precisión: Enfoque médico para el tratamiento y prevención de una enfermedad, que toma en cuenta la variabilidad individual de genes, ambiente y estilo de vida para los pacientes[37].

Morbilidad: Proporción de personas con un padecimiento en una población definida, es decir, la cantidad de personas que padecen una enfermedad por cada 100,000 habitantes.

Mortalidad: Cantidad de personas que fallecen a causa de un padecimiento por cada 100,000 habitantes.

Prevalencia: Tasa de morbilidad que indica la proporción de personas que viven con la enfermedad por cada 100,000 habitantes.

Profundidad: Número de veces que un transcrito es amplificado por el secuenciador durante el proceso de secuenciación.

Quimiorresistencia: Capacidad del tumor de evadir el daño causado por los agentes quimioterapéuticos[105].

Respuesta Patológica Completa: Ausencia de evidencia clínica que demuestre la presencia o progresión del desarrollo tumoral en el tejido mamario después de la administración terapéutica, por lo que se considera que el tratamiento es exitoso[106].

Sensibilidad: Probabilidad de que la prueba o biomarcador utilizado detecte a un individuo enfermo, es decir, que presente un resultado positivo en el estudio.