UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE MEDICINA

BIOMEDICINA

CARACTERIZACIÓN FUNCIONAL DEL LINCRNA GATA3-AS1 EN LÍNEAS CELULARES DE CÁNCER DE MAMA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRA EN CIENCIAS BIOLÓGICAS

PRESENTA:

Q. F. B. CONTRERAS ESPINOSA LAURA MARIANA

TUTOR PRINCIPAL DE TESIS: DR. CRISTIAN GABRIEL OLIVERIO ARRIAGA CANON

INSTITUTO NACIONAL DE CANCEROLOGÍA.

COMITÉ TUTOR: DRA. LORENA AGUILAR ARNAL

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM.

COMITÉ TUTOR: DR. RODRIGO GONZÁLEZ BARRIOS DE LA PARRA

INSTITUTO NACIONAL DE CANCEROLOGÍA.

CIUDAD UNIVERSITARIA, CD. MX., NOVIEMBRE, 2021

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE MEDICINA

BIOMEDICINA

CARACTERIZACIÓN FUNCIONAL DEL LINCRNA GATA3-AS1 EN LÍNEAS CELULARES DE CÁNCER DE MAMA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRA EN CIENCIAS BIOLÓGICAS

PRESENTA:

Q. F. B. CONTRERAS ESPINOSA LAURA MARIANA

TUTOR PRINCIPAL DE TESIS: DR. CRISTIAN GABRIEL OLIVERIO ARRIAGA CANON

INSTITUTO NACIONAL DE CANCEROLOGÍA.

COMITÉ TUTOR: DRA. LORENA AGUILAR ARNAL

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM.

COMITÉ TUTOR: DR. RODRIGO GONZÁLEZ BARRIOS DE LA PARRA

INSTITUTO NACIONAL DE CANCEROLOGÍA.

CIUDAD UNIVERSITARIA, CD. MX. 2021

COORDINACIÓN DEL POSGRADO EN CIENCIAS BIOLÓGICAS

ENTIDAD FACULTAD DE MEDICINA

OFICIO CPCB/1035/2021

ASUNTO: Oficio de Jurado

M. en C. Ivonne Ramírez Wence
Directora General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que en la reunión ordinaria del Subcomité de Biología Evolutiva, Ecología, Manejo Integral de Ecosistemas y Sistemática del Posgrado en Ciencias Biológicas, celebrada el día **6 de septiembre de 2021** se aprobó el siguiente jurado para el examen de grado de **MAESTRA EN CIENCIAS BIOLÓGICAS** en el campo de conocimiento de **Biomedicina** de la estudiante **CONTRERAS ESPINOSA LAURA MARIANA** con número de cuenta **311170911** con la tesis titulada **"CARACTERIZACIÓN FUNCIONAL DEL LINCRNA GATA3-AS1 EN LÍNEAS CELULARES DE CÁNCER DE MAMA",** realizada bajo la dirección del **DR. CRISTIAN GABRIEL OLIVERIO ARRIAGA CANON,** quedando integrado de la siguiente manera:

Presidente:	DR. LUIS ALONSO HERRERA MONTALVO
Vocal:	DR. FELIPE VACA PANIAGUA
Vocal:	DRA. KARLA FABIOLA MEZA SOSA
Vocal:	DRA. RUTH RUIZ ESPARZA GARRIDO
Secretario:	DRA. LORENA AGUILAR ARNAL

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Ciudad Universitaria, Cd. Mx., a 09 de noviembre de 2021

COORDINADOR DEL PROGRAMA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

DR. ADOLFO GERARDO NAVARRO SIGÜENZA

AGRADECIMIENTOS

Extiendo mi agradecimiento al Programa de Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México. Agradezco los recursos destinados a la realización de este Proyecto de Investigación.

Asimismo, quiero resaltar que esta investigación fue realizada gracias al apoyo proporcionado por el Consejo Nacional de Ciencia y Tecnología (CONACyT, CVU: 1003211). Agradezco al CONACyT la beca recibida.

Finalmente, agradezco el apoyo y dirección del tutor principal de este proyecto, el Dr. Cristian Gabriel Oliverio Arriaga Canon, así como a los miembros del comité tutor, la Dra. Lorena Aguilar Arnal y el Dr. Rodrigo González Barrios de la Parra por el apoyo proporcionado durante el desarrollo de este proyecto de investigación.

AGRADECIMIENTOS

Agradezco al Instituto Nacional de Cancerología (INCan), particularmente al Laboratorio de Carcinogénesis, que es dirigido por el Dr. Luis Alonso Herrera Montalvo, por los recursos destinados a la realización de este Proyecto de Investigación.

DEDICATORIA

Este trabajo está dedicado a mi familia, que es integrada por mis padres, mis hermanas, mis tíos y mi abuelita. Ha sido muy importante para mí su apoyo y comprensión durante mis estudios de posgrado. También dedico este trabajo a mis amigos. La realización de este trabajo no hubiera sido posible sin la diversión y el apoyo que mutuamente nos hemos proporcionado.

Incluyo también mi agradecimiento nuevamente al Laboratorio de Carcinogénesis del INCan, en especial a la M. Clementina Castro, que siempre nos ha apoyado, guiado y aconsejado. El éxito de este trabajo también fue posible debido al apoyo de mis compañeros de laboratorio, a quienes agradezco su atención y su ayuda.

Quiero incluir también en esta dedicatoria a Rubén, porque seguimos creciendo juntos, a nuestro ritmo, acompañándonos.

Además, durante la pandemia por SARS-CoV-2 mi cuarto y la sala de mi hogar se convirtieron en mis principales cubículos de trabajo. Dentro de estos cubículos, hubo dos compañeros de "trabajo" a los que quiero agradecer, que son mis pequeños cachorros eternos Turín y Pinta. Especialmente a Pinta, porque ella estuvo a mi lado durante la realización de los análisis bioinformáticos e *in silico* que incluye este trabajo, así como en todo el proceso de escritura de esta tesis (incluso en los tutorales). Su presencia me hizo más amena la cuarentena por la pandemia.

Finalmente, quiero dedicar este trabajo a mi abuelo, Rafael Espinosa Guerrero. Sé que le hubiera asombrado mucho lo que las computadoras pueden hacer actualmente, y también sé que su interés en mi trabajo hubiera motivado mis días, como lo hacía cuando aún se contaba entre los presentes. Con esto, se han acumulado las cosas de las que teníamos que hablar desde que ya no tenemos las reuniones dominicales frente al televisor.

"Cuando creíamos que teníamos todas las respuestas, de pronto cambiaron todas las preguntas".

Mario Benedetti, poeta uruguayo (1920-2009).

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS Y SÍMBOLOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS Y SÍMBOLOS

°C: Grados Celsius

µL: Microlitro

ASO: Oligonucleótido Antisentido (por sus siglas en inglés *Antisense Oligonucleotides*)

ATAC-Sec: Ensayo para la identificación de la cromatina accesible a la transposasa usando secuenciación (del inglés *Assay for Transposase-Accessible Chromatin*)

C: Citosina

c. b. p.: Cuanto baste para

CAGE: Análisis de la expresión de genes con modificación Cap 5' (del inglés *Cap analysis gene expression*)

CCLE: Enciclopedia de líneas celulares de cáncer (del inglés *Cancer Cell Line Encyclopedia*)

cDNA: DNA Copia

ChIP-Sec: Secuenciación de la inmunoprecipitación de la cromatina (del inglés *Chromatin Immunoprecipitation*)

CO_2: Dióxido de Carbono

DEPC: Dietilpirocarbonato

DNA: Ácido desoxirribonucleico (por sus siglas en inglés *Desoxiribonucleic Acid*)

DTT: Dithiothreitol

EDTA: Ácido etilaminotetraacético

ENCODE: Enciclopedia de Elementos de DNA (del inglés *Encyclopedia of DNA Elements*)

F: Sentido (del inglés *Forward*)

FARNA: Anotación de funciones de transcritos de RNA no codificantes humanos (del inglés *Database of Function Annotation of human non-coding RNA transcripts*)

FDR: Tasa de descubrimientos falsos (del inglés *False Discovery Rate*)

FPKM: Fragmentos de lecturas por kilobase de transcrito por cada millón de lecturas mapeadas (del inglés *fragments per kilobase of transcript per million reads mapped*)

g: Gramo

G: Guanina

GEO: Tomo Colectivo de Expresión de Genes (del inglés *Gene Expression Omnibus*)

GTEx: Proyecto de Expresión Genotipo-Tejido (del inglés *Genotype-Tissue Expression Project*)

HCl: Ácido clorhídrico

HEPES: Ácido 4-(2-hidroxietil)piperazin-1-iletanosulfónico

HER2: Receptor 2 del factor de crecimiento epidérmico humano (del inglés *human epidermal growth factor receptor 2*)

HMS LINCS: Biblioteca de Firmas Celulares Integradas Basadas en Redes de la Escuela de Medicina de la Universidad de Harvard (del inglés *Harvard Medical School Library of Integrated Network-based Cellular Signatures*)

KEGG: Enciclopedia de Genes y Genomas de Kyoto (del inglés *Kyoto Encyclopedia of Genes and Genomes*)

L: Litro

lncRNA: RNA largo no codificante

lincRNA: RNA largo no codificante intergénico

M: Molar

$MgCl_2$: Cloruro de Magnesio

min: Minutos

miRNA: microRNA

mg: Miligramo

mL: Mililitro

mRNA: RNA mensajero

NaCl: Cloruro de Sodio

ncRNA: RNA no codificante

NES: puntaje normalizado de enriquecimiento (del inglés *normalized enrichment score*)

NTC: Control negativo

nM: Nanomolar

P/V: Partes por volúmen

pb: pares de bases.

PBS: Solución amortiguadora de fosfatosPCR: Reacción en cadena de la polimerasa (del inglés *Polymerase Chain Reaction*)R: Antisentido (del inglés *Reverse*)

RNA: Ácido Ribonucleico (por sus siglas en inglés *Ribonucleic Acid*)

RNA-PET: Secuenciación masiva en paralelo de RNA de extremo pareado marcado (del inglés *paired end tag*)

RNA-Sec: Secuenciación masiva en paralelo de RNA

RPM: Revoluciones por minuto

RPKM: Lecturas por kilobase de transcrito por cada millón de lecturas mapeadas (del inglés *reads per kilobase of transcript per million reads mapped*)

RT-PCR: Reverso Transcripción por PCR

RT-qPCR: PCR cuantitativa (tiempo real) integrada con reverso transcripción

SFB: Suero Fetal Bovino

STRING: Herramienta de búsqueda para Recabar la Información de interacción de Genes/Proteínas (del inglés *Search Tool for the Retrieval of Interacting Genes/Proteins*)

TANRIC: Atlas de RNA no codificantes en Cáncer (del inglés *The Atlas of ncRNA in Cancer*)

TCGA: Programa del Atlas del Genoma del Cáncer (del inglés *The Cancer Genome Atlas*)

TPM: Transcritos por millón (del inglés *Transcripts per million*)

TSS: Sitio de inicio de la transcripción (del inglés *Transcription Start Site*)

U: Unidades de enzima

1.0 RESUMEN

Los RNA largos no codificantes intergénicos (lincRNAs) divergentes son genes no codificantes que se caracterizan por localizarse adyacentes a otro gen y por su expresión tejido-específica. Estos RNAs no codificantes pueden regular la expresión de sus genes adyacentes mediante un mecanismo molecular de regulación en *cis*, ya que regulan el locus del cual se transcribieron de manera alelo-específica. Actualmente, se han identificado lincRNAs divergentes que se asocian con el desarrollo de cáncer de mama, puesto que su perfil de expresión define las características clínicas de los tumores mamarios. No obstante, existen pocos trabajos en la literatura científica que caractericen funcional y molecularmente estos RNAs no codificantes, y que demuestren que éstos llevan a cabo su función mediante un mecanismo de regulación en *cis* sobre sus genes adyacentes. En nuestro trabajo previo, identificamos a *GATA3-AS1*, un lincRNA con expresión específica en pacientes con cáncer de mama que se asocia con la respuesta a la quimioterapia neoadyuvante y que se transcribe de manera divergente con respecto a su gen codificante adyacente *GATA3*. Sin embargo, a la fecha no existen reportes acerca de su función ni de la regulación sobre la expresión de su gen adyacente *GATA3* en cáncer de mama. En este trabajo, demostramos que *GATA3-AS1* cumple con el primer postulado de la regulación en *cis* ya que su expresión y la de *GATA3* tienen una correlación positiva (correlación de Pearson = 0.8, valor *p* =0.001). Por otro lado, el silenciamiento de *GATA3-AS1* usando oligonucleótidos antisentido en la línea celular MCF-7, demostró que este lincRNA también cumple con el segundo postulado de la regulación en *cis*, ya que la disminución en los niveles de expresión de *GATA3-AS1* se relacionan con la disminución de la expresión del mRNA de su gen adyacente *GATA3* (n=6, valor $p < 0.001$). En conclusión, nuestros resultados sugieren que *GATA3-AS1* regula la expresión de su gen adyacente *GATA3* mediante un mecanismo de regulación en *cis* de manera positiva en la línea celular MCF-7, el cual no había sido descrito previamente en la literatura.

2.0 ABSTRACT

Divergent intergenic long non-coding RNAs are non-coding genes characterized by being located adjacent to another gene(s) and by being expressed in a tissue-specific manner. These non-coding RNAs can regulate the expression of their adjacent gene(s) through a *cis*-acting regulatory mechanism, since they regulate the locus from which they were transcribed in an allele-specific way. Currently, divergent intergenic long non-coding RNAs have been identified as being associated with breast cancer development, since their expression profile define the clinical characteristics of breast tumors. However, there are few works in the scientific literature that functionally and molecularly characterize these non-coding RNAs, and that demonstrate these RNAs carry out their functions through a *cis*-regulatory mechanism on their adjacent genes. In our previous work, we identified *GATA3-AS1*, a long non-coding RNA with specific expression in breast cancer patients which has been associated with the response to neoadjuvant chemotherapy. Moreover, *GATA3-AS1* is divergently transcribed to its adjacent coding gene *GATA3*. Until now, there are no reports about its function or the regulation of the expression of its adjacent gene *GATA3* in breast cancer. Our results demonstrate that *GATA3-AS1* complies with the first postulate of *cis* regulation since *GATA3-AS1* and *GATA3* expression are positively correlated (Pearson's correlation = 0.8, *p*-value = 0.001). On the other hand, *GATA3-AS1* knock down in the MCF-7 cell line showed that it also complies with the second postulate of *cis*-regulation, since the decrease in *GATA3-AS1* expression level is related to the decrease observed for its adjacent gene *GATA3* (n = 6, *p*-value <0.001). In conclusion, our results show that *GATA3-AS1* regulates the expression of its adjacent gene *GATA3* through a *cis*-regulatory mechanism in a positive way in the MCF-7 cell line, which had not been previously described.

3.0 INTRODUCCIÓN

3.1 Generalidades de los RNAs largos no codificantes (lncRNAs)

El genoma humano está compuesto por dos tipos de secuencias génicas: las codificantes, que dan lugar a proteínas, y las no codificantes, que se caracterizan por no generar proteínas como producto final[1]. Por otro lado, la evidencia científica sugiere que la transcripción del genoma humano es extensiva en todo el material genético[1], por lo que no se limita sólo a los genes codificantes. Se ha reportado que el 77% de los genes que se transcriben corresponde a RNAs no codificantes (ncRNAs) [3], que incluyen varios biotipos de RNA, como los RNA largos no codificantes (lncRNAs), los cuales constituyen aproximadamente el 10% del genoma no codificante[4]. Se ha reportado que los lncRNAs se asocian a funciones biológicas regulatorias, como la regulación transcripcional y postranscripcional de sus genes blanco, así como al desarrollo de diferentes padecimientos como el cáncer[5,6], por lo que es necesario profundizar la investigación acerca de la biología de los lncRNAs.

Evidencias experimentales han establecido que los lncRNAs se definen como transcritos mayores a 200 pares de bases que no presentan marcos abiertos de lectura, lo que significa que no dan origen a una proteína o un polipéptido[6]. Estos transcritos de naturaleza no-codificante se clasifican principalmente por su localización en el genoma como génicos o intergénicos[8,9]. Los lncRNAs génicos se definen como aquellos cuya secuencia genética empalma con uno o más nucleótidos de un gen codificante[8], mientras que los lncRNAs intergénicos (lincRNAs) son aquellos que se localizan a menos de 50 Kb de distancia de su gen más cercano, el cual se define como el gen adyacente, y sus secuencias no se empalman con ningún nucleótido[10]. Además, los lincRNAs se caracterizan por la longitud de sus transcritos, que en promedio es de 1 Kb, y contienen entre 1-3 exones, en comparación con los RNA mensajeros (mRNAs), cuyos transcritos tienen una la longitud promedio de 2.9 Kb y contienen alrededor de 10 exones[10]. Actualmente, se ha reportado que los lincRNAs son poco abundantes, incluso son

órdenes de magnitud menos abundantes en comparación con los mRNAs[10], llegando a tener hasta un sólo transcrito o molécula de lincRNA por célula[11], lo cual es relevante ya que existe evidencia científica que indica que sus bajos niveles de expresión están relacionados con su especificidad de expresión en tejidos humanos, así como con la regulación de sus genes blanco[2].

Particularmente, los lincRNAs que se sintetizan en sentido opuesto a la dirección de la transcirpción de su gen adyacente se definen como lincRNAs divergentes[9], y se ha reportado que representan aproximadamente el 1.7% del genoma humano[8]. Además, su cercanía con los genes adyacentes a ellos, ha sido relevante para el estudio de sus funciones intracelulares, ya que se ha reportado que regulan la expresión de sus genes adyacentes. Por ejemplo, el lincRNA *NBR2* modula la transcripción de su gen adyacente *BRCA1* en el desarrollo tumoral en cáncer de mama. Por lo tanto, el estudio de las características y las funciones biológicas de los lincRNAs divergentes es relevante para el entendimiento del desarrollo de padecimientos crónico-degenerativos como el cáncer[12].

3.1.1 Características de los lincRNAs divergentes

Los lincRNAs divergentes se caracterizan porque su longitud promedio es de 1 Kb, además de que sus transcritos se componen por dos exones[2] y su localización celular es principalmente en el núcleo celular[13]. Además, su transcripción es de tipo bidireccional, ya que la síntesis de sus transcritos se lleva a cabo en sentido opuesto a la transcripción de su gen adyacente[9]. Cuando la distancia entre los sitios de inicio de la transcripción (TSS) del lincRNA divergente y del gen adyacente es menor a 1 Kb, la región genómica donde se encuentran los promotores de ambos genes se define como un promotor compartido[2]. Sin embargo, el promotor del lincRNA divergente y el promotor de su gen adyacente deben ser independientes para que su transcripción se considere divergente[16].

Asimismo, se ha reportado que los lincRNAs divergentes se encuentran conservados en secuencias cortas o motivos específicos en los mamíferos, particularmente si su gen adyacente es codificante[15], lo cual ha llevado a la conclusión de que la cercanía con el gen codificante favorece la conservación del lincRNA divergente con el objetivo de preservar su función regulatoria sobre el gen adyacente. Por ejemplo, *MYMLR* es un lincRNA conservado que activa la transcripción de su gen adyacente *MYC*, y se ha demostrado que el abatimiento de la expresión de este lincRNA divergente tiene como consecuencia la disminución en la expresión de *MYC*, así como de sus genes blanco, lo que correlaciona con la disminución de la proliferación en la línea celular de adenocarcinoma de pulmón A549[15]. Por lo tanto, la evidencia científica sugiere que la conservación de los lincRNAs divergentes como *MYMLR*[16,17] está relacionada con las funciones intracelulares en las que participan mediante la regulación de la expresión de sus genes adyacentes.

Entonces, la función regulatoria que los lincRNAs divergentes ejercen sobre su gen adyacente está modulada por diferentes mecanismos moleculares que regulan la transcripción de estos lincRNAs, lo que permite el mantenimiento de sus niveles de expresión para regular las funciones intracelulares que éstos llevan a cabo.

3.1.2 Mecanismos de regulación de la expresión de los lincRNAs divergentes

La regulación de la expresión de los lincRNAs divergentes está modulada por diferentes mecanismos moleculares, entre los que se encuentran algunos factores epigenéticos, como las modificaciones post-traduccionales de histonas[18] y la regulación del promotor mediada por factores transcripcionales[14]. En conjunto, la regulación de la transcripción de los lincRNAs divergentes tiene como consecuencia el mantenimiento adecuado de las funciones que éstos desempeñan en las vías de señalización intracelulares, y se relaciona a su vez con las bajas tasas de expresión de los mismos y la especificidad de su expresión en diferentes tejidos humanos[13].

3.1.2.1 Regulación de la expresión de los lincRNAs divergentes mediada por modificaciones post-traduccionales de histonas

Las modificaciones post-traduccionales de histonas que principalmente se encuentran enriquecidas en las regiones genómicas donde se localizan los lincRNAs divergentes son aquellas que se asocian con la activación de la transcripción, como la trimetilación de la lisina 4 en la histona 3 (H3K4me3), la trimetilación de la lisina 36 en la histona 3 (H3K36me3) y la monometilación de la lisina 4 en la histona 3 (H3K4me1), que también regula la activación transcripcional de genes codificantes[16]. Particularmente, se ha reportado que el enriquecimiento de la H3K4me1 lleva a cabo una función dual de regulación en los loci donde se localizan lincRNAs divergentes, debido a que además de activar su transcripción, también modula la expresión de los genes adyacentes, principalmente si éstos son codificantes, ya que el enriquecimiento de la H3K4me1 se asocia con la presencia de elementos potenciadores de la transcripción o *enhancers*[18,20] en el promotor compartido. Incluso, se ha observado que la interacción del lincRNA divergente con el elemento *enhancer* promueve la expresión de sus genes codificantes adyacentes al potenciar su transcripción[20] (Figura 1).

Otra modificación post-traduccional de histonas importante en la regulación transcripcional de los lincRNAs divergentes es la trimetilación de la lisina 9 en la histona 3 (H3K9me3), que se ha identificado como una marca de silenciamiento epigenético en genes codificantes. De manera contraria, se ha establecido que en los loci a partir de los cuales se sintetizan lincRNAs divergentes, la acumulación de la H3K9me3 se encuentra asociada con la transcripción activa de los mismos[18] y con la especificidad de la expresión en diferentes tejidos humanos[9]. Por otra parte, las modificaciones post-traduccionales en la lisina 27 de la histona 3 (H3K27) también participan en la regulación de la expresión de los lincRNAs divergentes. Por ejemplo, Luo y colaboradores demostraron que la acetilación de la H3K27 (H3K27ac) se encuentra enriquecida en los TSSs de lincRNAs divergentes que se expresan activamente en células pluripotentes, mientras que la trimetilación de la lisina-27 (H3K27me3) se encuentra en niveles

menores sobre el TSS y en el cuerpo del gen al compararla con la H3K27ac, por lo que la H3K27me3 podría tener una función bivalente en la regulación transcripcional que mantiene la especificidad de expresión de estos lincRNAs[12] (Figura 1). En síntesis, las modificaciones post-traduccionales de histonas son parte del mecanismo molecular que regula la expresión de los lincRNAs divergentes, principalmente al modular la activación de su transcripción.

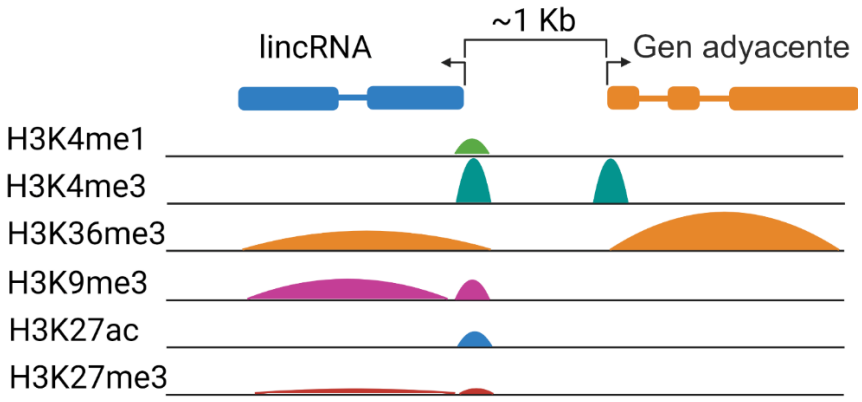


Figura 1. Los loci de los que provienen los lincRNAs divergentes se caracterizan por la presencia de modificaciones post-traduccionales de histonas relacionadas con la activación de la transcripción. La modificación post-traduccional de histona H3K4me3 está enriquecida en los promotores de los lincRNAs divergentes y de los genes codificantes adyacentes (turquesa), mientras que la H3K36me3 se encuentra enriquecida en el cuerpo de los genes (naranja), y está asociada a la activación transcripcional. Sin embargo, la H3K36me3 se encuentra en niveles menores en los cuerpos de los genes de los lincRNAs con respecto a los genes codificantes, lo que explica que los niveles de expresión de los lincRNAs sean menores comparados con los de los mRNAs[12]. Asimismo, la H3K27ac (azul) y la H3K4me1 (verde) están enriquecidas en los promotores de los lincRNAs divergentes y se asocian a la activación de la transcripción de los lincRNAs[18,20–22]. Finalmente, la H3K9me3 (morado) y la H3K27me3 (rojo)

también se localizan en el promotor y en el cuerpo de los genes de los lincRNAs divergentes[10,21,22]. Figura creada con BioRender.com

En resumen, las modificaciones post-traduccionales de histonas que se encuentran enriquecidas en los loci a partir de los cuales se transcriben los lincRNAs divergentes, se relacionan con la activación de la transcripción (H3K4me3 y la H3K36me3) y con a la especificidad de expresión en tejidos (H3K9me3 y la H3K27me3). Por otro lado, el enriquecimiento de la H3K27ac y la H3K4me1 en el promotor, regulan la expresión tanto del lincRNA como de su gen adyacente al asociarse con la actividad de elementos *enhancer*, por lo cual, regulan la transcripción de lincRNAs divergentes. Sin embargo, existen otros mecanismos moleculares que regulan la expresión de los lincRNAs divergentes, como los mediados por factores transcripcionales, que también están relacionados con las vías de señalización intracelulares en las que estos lincRNAs participan.

3.1.2.2 Regulación de la expresión de lincRNAs divergentes a través de su región promotora

La regulación de la expresión de los lincRNAs divergentes se lleva a cabo mediante mecanismos moleculares que involucran elementos epigenéticos, como las modificaciones post-traduccionales de histonas, así como elementos genéticos, que regulan directamente la región promotora del lincRNA, y está mediado principalmente por la interacción con factores transcripcionales[6,9].

Se ha reportado que para los promotores compartidos de lincRNAs divergentes, la regulación mediada por factores transcripcionales es diferente que para los genes codificantes. Por un lado, la región promotora independiente de los genes codificantes presenta motivos de unión para factores transcripcionales que no están empalmados entre ellos, por lo que su transcripción ocurre en mayor frecuencia, y se conocen como

"promotores fuertes"[14] (Figura 2). Por el contrario, en los promotores de los lincRNAs divergentes se ha observado que los motivos de unión a factores transcripcionales se encuentran empalmados entre ellos dentro de la secuencia promotora, lo que se asocia con la regulación de la expresión tejido-específica de los lincRNAs divergentes. La evidencia científica sugiere que la arquitectura del promotor de los lincRNAs divergentes se encuentra conservada en mamíferos a nivel de secuencia, al igual que los promotores de los genes codificantes, lo que significa que ha sido un mecanismo de regulación evolutivamente útil que preserva la función de los genes ubicados en loci divergentes[7,14].

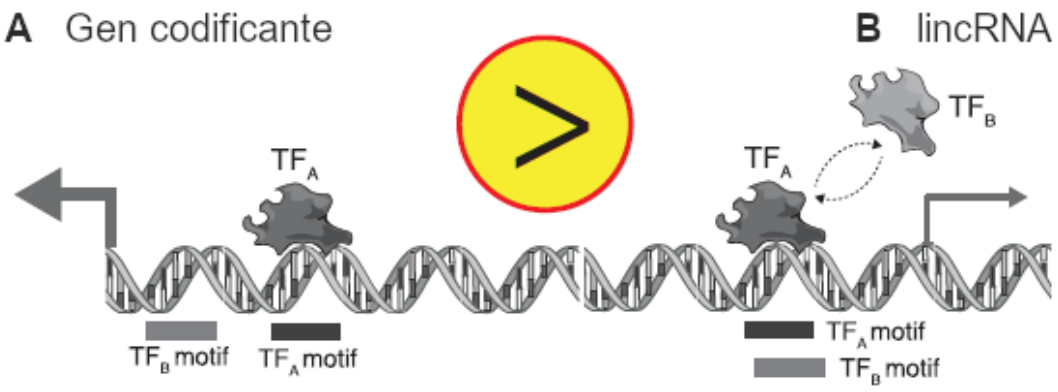


Figura 2. Los promotores de los lincRNAs divergentes tienen motivos de unión para factores transcripcionales que regulan la activación de su transcripción. A) En los promotores compartidos de lincRNAs divergentes y genes codificantes adyacentes, los factores transcripcionales (TF_A y TF_B) interactúan con más estabilidad con la región promotora del gen codificante adyacente, por lo que la interacción con la RNA Pol II es más eficiente y promueve la transcripción del gen codificante. B) En los promotores de los lincRNAs divergentes, la estabilidad de la interacción entre los factores transcripcionales y sus motivos de unión es menor debido al empalme de las secuencias de los motivos, ya que compiten por la interacción con la región promotora, por lo que la estabilidad de la interacción con la RNA Pol II es menor y los lincRNAs divergentes se expresan en niveles menores respecto a los genes codificantes. Modificada de Mattioli, *et al,* 2019[14].

Un ejemplo de regulación de la transcripción de lincRNAs divergentes por factores transcripcionales es la regulación mediada por MYC*,* que regula en general a los promotores de los lncRNAs. El gen *MYC* pertenece a los oncogenes asociados a la proliferación celular en varios tipos de cáncer[23], y la pérdida de su función provoca la disminución en la expresión de lncRNAs[10], entre los que se encuentran algunos lincRNAs divergentes[7]. La regulación que ejerce MYC sobre los lincRNAs divergentes ocurre de manera simultánea con la regulación de la estructura de la cromatina mediada por DICER1, que promueve el enriquecimiento de la H3K4me3 y la H3K36me3 en los loci de los lincRNAs[10], por lo que la regulación conjunta de MYC y Dicer1 se relaciona con la activación de la transcripción de lincRNAs divergentes. Por otro lado, se han identificado factores transcripcionales que regulan a los promotores de los lincRNAs divergentes de manera tejido-específica, como es el caso del receptor de estrógeno (ER)[25,26]. La alteración de la función del ER produce la desregulación de los niveles de expresión de lincRNAs divergentes[26], principalmente para aquellos que están relacionados con las vías de señalización intracelulares que regulan la proliferación y el ciclo celular, demostrando la importancia de la regulación que ejercen los factores transcripcionales sobre la transcripción de los lincRNAs divergentes[28].

En la actualidad, existen pocos reportes científicos acerca de la regulación de los promotores de lincRNAs divergentes mediada por factores transcripcionales. Por lo anterior, se ha hecho uso de herramientas bioinformáticas que permitan predecir la presencia de motivos de unión a factores transcripcionales en las secuencias promotoras de lincRNAs divergentes, como por ejemplo la herramienta *FIMO*[28]. Los resultados obtenidos a partir del análisis hecho con *FIMO* han sido validados mediante experimentos de inmunoprecipitación de cromatina (*ChIP,* por sus siglas en inglés), por ejemplo, en la línea celular HepG2 de cáncer hepático, en la que se identificó la presencia del motivo de unión para el factor transcripcional NRF1 en la región promotora del lncRNA *DLEU1* mediante el uso de *FIMO*, lo cual fue validado posteriormente mediante un ensayo experimental de *ChIP*[14]. Por lo tanto, el uso de las herramientas

bioinformáticas ha contribuido importantemente al entendimiento de los mecanismos moleculares que regulan la expresión de lincRNAs divergentes.

En suma, la regulación de la expresión de lincRNAs divergentes hacia sus regiones promotoras está mediada tanto por elementos epigenéticos que incluyen las modificaciones post-traduccionales de histonas así como elementos genéticos que incluyen la unión específica de factores transcripcionales como MYC. Igualmente, la secuencia promotora de los lincRNAs divergentes ha conservado a través de la evolución una estructura basada en secuencias empalmadas de motivos de unión a factores transcripcionales que contribuye con su expresión tejido-específica y con la modulación de las funciones que los lincRNAs divergentes desempeñan en las diferentes vías de señalización intracelulares. No obstante, a la fecha existe limitada evidencia científica que permita establecer con claridad las funciones regulatorias que los lincRNAs divergentes llevan a cabo, por lo que es necesario profundizar la investigación respecto a este tema.

3.1.3 Funciones de los lincRNAs divergentes

Se ha reportado que los lincRNAs divergentes se encuentran acumulados y enriquecidos en la fracción de la cromatina[29], por lo cual este tipo de transcritos de naturaleza no-codificante pueden interactuar con factores transcripcionales y con complejos remodeladores de la cromatina en los promotores de sus genes adyacentes. Por lo anterior, aunado a su cercanía con algunos genes, a los lincRNAs divergentes se les ha asociado con funciones que pueden regular la expresión de estos genes adyacentes a través del mantenimiento de la estructura de la cromatina, para que sea permisiva para la activación de la transcripción[13], lo cual regula de manera directa la transcripción de dichos genes.

3.1.3.1 Mecanismos moleculares de acción de los lincRNAs divergentes

Los lincRNAs divergentes pueden llevar a cabo sus funciones mediante la interacción con otras biomoléculas como las proteínas, el DNA, los mRNAs y los miRNAs, con las cuales puede interactuar incluso de manera simultánea, lo que permite que los lincRNAs divergentes lleven a cabo sus mecanismos moleculares de regulación génica mediante la interacción con los remodeladores de la cromatina, así como con factores transcripcionales para regular la transcripción de sus genes adyacentes[10,18].

Se han propuesto tres modelos por los cuales los lincRNAs divergentes pueden llevar a cabo sus mecanismos de regulación de la expresión de sus genes adyacentes, que son el modelo de regulación de guía (del inglés *guide*), el de plataforma (del inglés *scaffold*) y el de potenciador (del inglés *enhancer*)[10]. El modelo de guía permite la formación del complejo ribonucleoproteíco y en particular el lincRNA tiene la capacidad de dirigir al complejo a sus secuencias específicas en el genoma. Por otro lado, en el modelo de plataforma, los lincRNAs divergentes tienen la capacidad de interactuar con un conjunto de proteínas que se unen al lincRNA para formar un complejo ribonucleoproteico que sirve como una plataforma de ensamblaje para llevar a cabo la activación de un gen en específico. Finalmente, mediante el modelo de potenciador o *enhancer*, se lleva a cabo el acercamiento e interacción a larga distancia con la formación de asas de cromatina, que permiten al lincRNA, junto con las proteínas asociadas, interactuar con los genes a los cuales regulan y de esta manera potenciar su expresión (Figura 3)[30]. En conjunto, estos mecanismos moleculares permiten a los lincRNAs divergentes modular la transcripción de sus genes blanco, así como ejercer su función en las diferentes vías de señalización intracelulares.

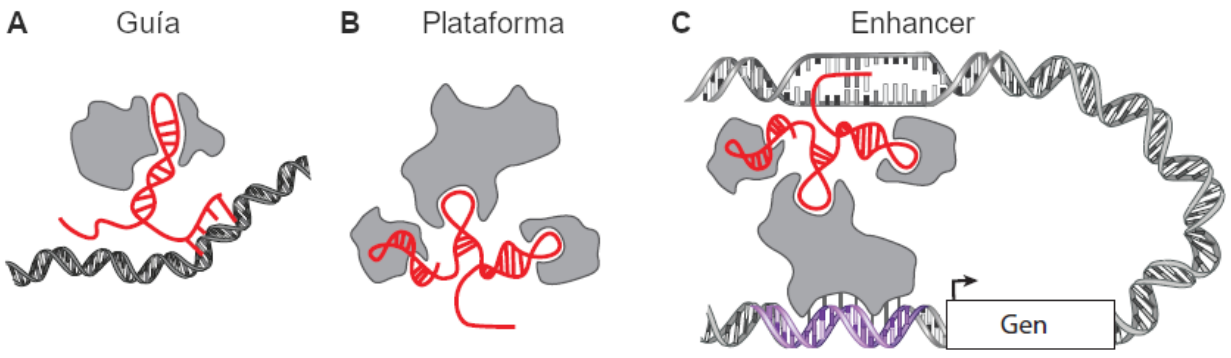


Figura 3. Los lincRNAs divergentes pueden llevar a cabo sus funciones de regulación génica mediante diferentes mecanismos moleculares de acción. A) Los lincRNAs divergentes pueden regular la expresión de sus genes adyacentes como guías de proteínas, al interactuar simultáneamente con las proteínas y el DNA, lo que estabiliza la interacción de la proteína con su gen blanco. B) La función de plataforma de los lincRNAs estabiliza su interacción de diferentes proteínas al mismo tiempo, lo que permite la formación de complejos proteínicos. C) Los lincRNAs funcionan como *enhacers* al interactuar con el promotor de su gen adyacente y al mismo tiempo con un elemento *enhancer* que potencia su expresión, lo cual es un mecanismo relevante para los lincRNAs divergentes y para la regulación de sus genes adyacentes. Modificada de los artículos de Rinn y Chang, 2012[29], y Gao, *et al*, 2020[32].

Asimismo, se ha reportado que los lincRNAs divergentes pueden regular la expresión de sus genes adyacentes mediante estos mecanismos moleculares debido a la cercanía que existe entre ellos, lo que se ha definido como regulación en *cis*[13], que es un mecanismo por el cual los lincRNAs divergentes regulan la expresión y la función de sus genes adyacentes[12,19].

3.1.3.2 El mecanismo de regulación en cis y sus funciones biológicas

El mecanismo de regulación en *cis* se define como la actividad regulatoria de un lncRNA sobre la expresión de su gen adyacente[19], que se caracteriza por ser alelo-específica, lo

que significa que su efecto regulatorio ocurre en el mismo locus a partir del cual se sintetiza el lincRNA[18], ya que, debido a la cercanía con su gen adyacente, una sola molécula de lincRNA es suficiente para llevar a cabo su acción regulatoria[13]. Además, la evidencia científica sugiere que el mecanismo de regulación en *cis* es una función de los lincRNAs divergentes que se ha conservado en conjunto con los elementos regulatorios en *cis* que se encuentran en sus regiones promotoras, como los elementos *enhancer*[2,33], por lo que su conservación podría estar relacionada con la importancia de las funciones biológicas que los lincRNAs divergentes llevan a cabo en las vías de señalización intracelulares que regulan.

Para identificar a un lncRNA que regula la expresión de su gen adyacente mediante el mecanismo molecular de regulación en *cis*, Guttman y Rinn propusieron tres postulados experimentales basados en la interacción del lncRNA con su gen adyacente:

- Primer Postulado: la expresión del gen codificante adyacente correlaciona con la expresión del lncRNA en todas las condiciones tanto en su contexto natural, así como en el contexto experimental.
- Segundo postulado: la pérdida de la expresión y de la función del lncRNA tendrá un efecto en la expresión de su gen codificante adyacente.
- Tercer postulado: el lncRNA modulará la expresión del gen codificante adyacente exclusivamente en el alelo a partir del cual ambos fueron sintetizados[18].

Por lo tanto, se considera que un lincRNA divergente regula en *cis* la expresión de su gen adyacente si cumple con los tres postulados (Figura 4), de lo contrario, la regulación de la expresión de su gen adyacente se lleva a cabo mediante la regulación en *trans*.

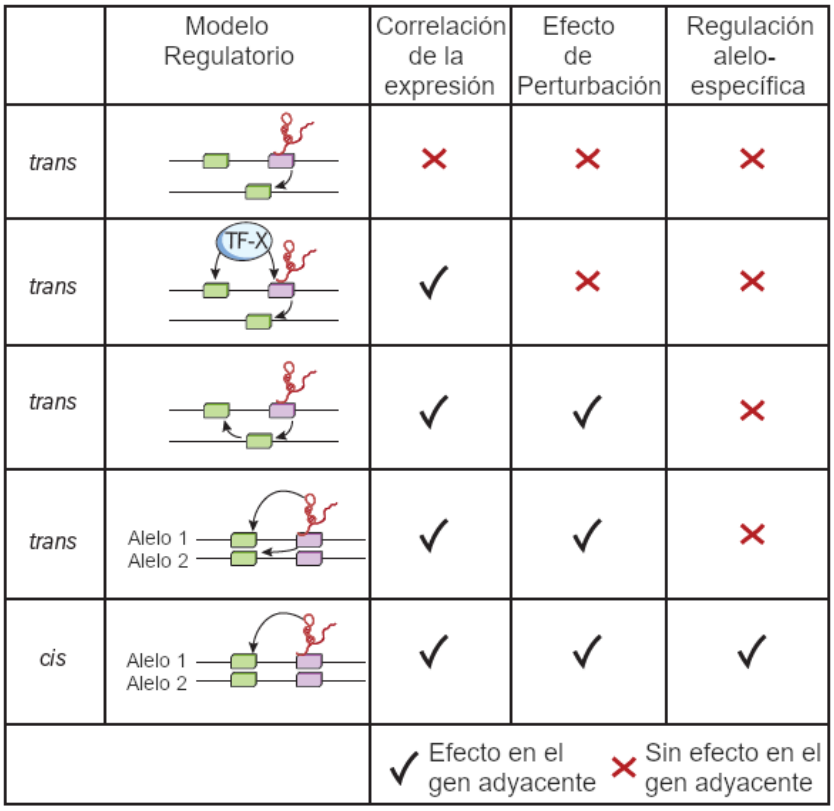


Figura 4. Postulados del mecanismo de regulación en *cis*. Esquema que representa las condiciones en las que se considera que un lncRNA regula a su gen adyacente mediante el mecanismo molecular en *cis*. Como se observa en la figura, al cumplir con las tres condiciones enumeradas en los postulados (correlación en la expresión, efecto de perturbación y regulación alelo-específica) se puede afirmar que un lncRNA regula a su gen adyacente mediante el mecanismo molecular en *cis*. De lo contrario, el mecanismo de regulación se considera en *trans*[19]. Modificada de Guttman y Rinn, 2012[18].

Además, la regulación en *cis* mediada por un lincRNA divergente puede tener el efecto de silenciamiento o de activación de la expresión del gen adyacente, al regular de manera directa el reclutamiento de factores transcripcionales[19], o de manera indirecta mediante la interacción con remodeladores de la cromatina[18]. Por ejemplo, la evidencia científica sugiere que el lincRNA divergente *MYMLR* regula la activación de la transcripción de su gen adyacente *MYC* en *cis* al reclutar a PCBP2 a la región promotora de *MYC*, que

adicionalmente interactúa con un elemento *enhancer* para potenciar su expresión, lo cual promueve la proliferación celular en la línea celular de cáncer de pulmón A549[15].

Por otro lado, los lincRNAs divergentes regulan la expresión de sus genes adyacentes de manera indirecta al participar en el mantenimiento y remodelado de la cromatina mediante la interacción con complejos remodeladores de la cromatina, como el complejo MLL1 (lisina metiltransferasa 2A), que contribuyen con el posicionamiento de modificaciones post-traduccionales de histonas, como la H34me3, o con la formación de asas de cromatina, como se ha visto en el locus de *HOTTIP* y los genes *HOXA*[7,10,19]. Por ejemplo, se ha descrito que los lincRNAs divergentes que regulan en *cis* la expresión de su gen adyacente se encuentran enriquecidos en cromatina activa, e interactúan con proteínas como WDR5, que forma parte de los complejos remodeladores epigenéticos en los que se encuentran la metiltransferasa que se conoce como MLL1, que se encarga de depositar la H3K4me3 y de la H3K4me2 en los promotores de sus genes adyacentes[29]. Adicionalmente, se ha reportado que estos lincRNAs, interactúan con las cohesinas y la proteína CTCF, las cuales en conjunto regulan la formación de asas de cromatina y re-localizan a los promotores de sus genes adyacentes para que interactúen con elementos de regulación distales, como los *enhancers*[34]. Entonces, a través de estos mecanismos, los lincRNAs divergentes pueden activar la transcripción de sus genes adyacentes, por lo que se considera que la regulación positiva de la transcripción es una de las principales funciones biológicas que llevan a cabo este tipo de transcritos de naturaleza no-codificante.

Actualmente, la evidencia científica sugiere que los lincRNAs divergentes participan en la regulación de vías de señalización intracelulares involucradas en la diferenciación celular, al activar la transcripción de sus genes adyacentes en *cis*[35]. Por ejemplo, se ha reportado que los lincRNAs divergentes regulan la diferenciación de células pluripotentes[12,18,35], la diferenciación de células troncales a células del endodermo[36], la diferenciación de células musculares[37] y la diferenciación de células neuronales[38]. Asimismo, los lincRNAs divergentes se han asociado también a procesos biológicos como la organogénesis y el desarrollo embrionario[12], como es el caso del lincRNA

Foxd3as, que regula la activación de la expresión de su gen adyacente *FOXD3* durante el desarrollo embrionario en un modelo biológico de ratón[39]. No obstante, aún son pocos los lincRNAs divergentes que han sido experimentalmente caracterizados de manera molecular, por lo que es necesario profundizar la investigación sobre sus funciones biológicas.

En resumen, los lincRNAs divergentes que regulan la expresión de sus genes adyacentes en *cis* han sido poco caracterizados experimentalmente. Sin embargo, se ha determinado que los estos lincRNAs regulan la expresión de sus genes adyacentes mediante diferentes mecanismos moleculares, como los de plataforma, guía y potenciador, que además se han asociado a la regulación positiva de la transcripción de sus genes adyacentes. Esta función es particularmente importante para tratar de entender cómo los cambios en los perfiles de expresión de los lincRNAs divergentes podrían tener un efecto en la regulación de las vías de señalización intracelulares involucradas en el desarrollo de padecimientos como el cáncer.

3.2 Los lincRNAs divergentes en cáncer

El cáncer es una enfermedad multifactorial que se caracteriza por el crecimiento y la proliferación no regulada de las células, llevando a la formación de tejido neoplásico y al desarrollo tumoral[40]. Dentro de los factores que contribuyen con el desarrollo del cáncer, se ha reportado a los lincRNAs divergentes, ya que durante el desarrollo de células tumorales sus perfiles de expresión[41,42] y sus funciones regulatorias sobre las vías de señalización intracelulares en las que participan se encuentran alteradas, lo que contribuye a la carcinogénesis[42,43].

Asimismo, los lincRNAs divergentes se han relacionado con el desarrollo tumoral al asociarse con la alteración en la regulación de la transcripción de sus genes codificantes adyacentes[41]. Un ejemplo es *MYMLR*, que se ha asociado con el desarrollo de cáncer de pulmón al promover la sobrexpresión de su gen adyacente *MYC* al activar su

transcripción, lo cual regula positivamente la proliferación y el ciclo celular[15]. Otro ejemplo es el lincRNA *SLC16A1-AS1*, que regula positivamente la expresión de su gen codificante adyacente *SLC16A1/MCT1* al reclutar al factor transcripcional E2F1 a su promotor y promoviendo la reprogramación metabólica, que es una de las características del cáncer de vejiga[44]. Debido a la cercanía de los loci de *MYMLR* y *SLC16A1-AS1* con los loci de sus respectivos genes codificantes, se ha sugerido que estos lincRNAs divergentes pueden regularlos en *cis*, pero no se ha demostrado experimentalmente hasta ahora[15,19,44]. Adicionalmente, se ha reportado que los perfiles de expresión de los lincRNAs divergentes pueden definir molecularmente patologías como el cáncer[45], por lo que se han considerado como biomarcadores moleculares en la práctica clínica[46,47], esto debido a que su expresión es tejido-específica[48], y también a que presentan expresión diferencial entre los distintos tipos de cáncer[42]. Por lo anterior, los lincRNAs han demostrado ser de gran utilidad no sólo en el estudio del cáncer, sino también en la práctica clínica como biomarcadores diagnósticos[49].

A pesar de que ya existen reportes en la literatura acerca de las funciones de algunos lncRNAs en cáncer de mama[53], y su utilidad como biomarcadores moleculares en el pronóstico[50] y la respuesta a tratamiento[51], hay pocos trabajos de investigación científica dedicados a caracterizar las funciones de los lincRNAs divergentes en las células mamarias neoplásicas, que demuestren que modulan a sus genes adyacentes en *cis* en el desarrollo de cáncer de mama[52–54].

El cáncer de mama es la neoplasia maligna más común en mujeres a nivel mundial, siendo la principal causa de muerte por cáncer en la población femenina[55]. De acuerdo con su perfil molecular, el cáncer de mama puede ser clasificado en subtipos moleculares basados en la expresión de genes particulares, principalmente de los receptores hormonales de los tumores mamarios, lo cual ha permitido definir seis subtipos moleculares intrínsecos: luminal A, luminal B, enriquecido en el receptor 2 del factor de crecimiento epidérmico humano (*HER2*), basal, los normales y los bajos en claudina[56,57]. Debido a su especificidad de expresión, los lncRNAs se han reportado como genes no codificantes que se expresan diferencialmente en el tejido mamario neoplásico los

cuales, en conjunto con otros genes codificantes y no codificantes, participan en las vías de señalización intracelulares que regulan la carcinogénesis en la glándula mamaria, por lo que los lincRNAs divergentes pueden tener utilidad en la práctica clínica[46,51,52].

Finalmente, para caracterizar las funciones biológicas de los lincRNAs divergentes en cáncer de mama y los mecanismos moleculares mediante los cuales llevan a cabo la regulación de las vías de señalización intracelulares en las que participan, en este estudio se han utilizado diversas líneas celulares de cáncer de mama como modelos de estudio[60] (Tabla suplementaria 1, Apéndice A), debido a que éstas provienen de cultivos primarios de tumores mamarios de pacientes con diferentes fenotipos moleculares[61,62], por lo que han sido de utilidad en la investigación básica y clínica al replicar las condiciones intracelulares que definen a los diferentes tumores mamarios[63].

En conclusión, se ha reportado en la literatura científica que los lincRNAs divergentes contribuyen con el desarrollo tumoral. Sin embargo, las funciones que estos lincRNAs tienen en el desarrollo del cáncer de mama aún no han sido caracterizadas, por lo cual es necesario profundizar la investigación al respecto mediante el uso de modelos biológicos útiles para el estudio del cáncer de mama, como lo son las líneas celulares.

4.0 ANTECEDENTES

En nuestro trabajo previo identificamos al lincRNA *GATA3-AS1*, que es un lincRNA divergente que se sobreexpresa en líneas celulares neoplásicas de cáncer de mama (Figura 5). El lincRNA *GATA3-AS1* regula positivamente la expresión de su gen adyacente *GATA3* en el proceso de diferenciación de los linfocitos Th2. Gibbons y colaboradores sugieren que la regulación se lleva a cabo a través de la formación de una triple hélice entre *GATA3-AS1* y la secuencia de DNA del primer intrón del gen *GATA3*, lo que permite el reclutamiento de enzimas remodeladoras de la cromatina, como el complejo multipeptídico MLL, que se encarga de depositar las modificaciones post-traduccionales de histonas asociadas a transcripción activa en el dominio génico, como la H3K27ac y la H3K4me3[64], facilitando la transcripción del gen *GATA3*, el cual regula el proceso de diferenciación de los linfocitos Th2. Debido a la relación que existe entre la expresión de *GATA3* y *GATA3-AS1* en linfocitos Th2, en este mismo trabajo se sugiere que *GATA3-AS1* podría regular en *cis* a su gen adyacente *GATA3*[65].

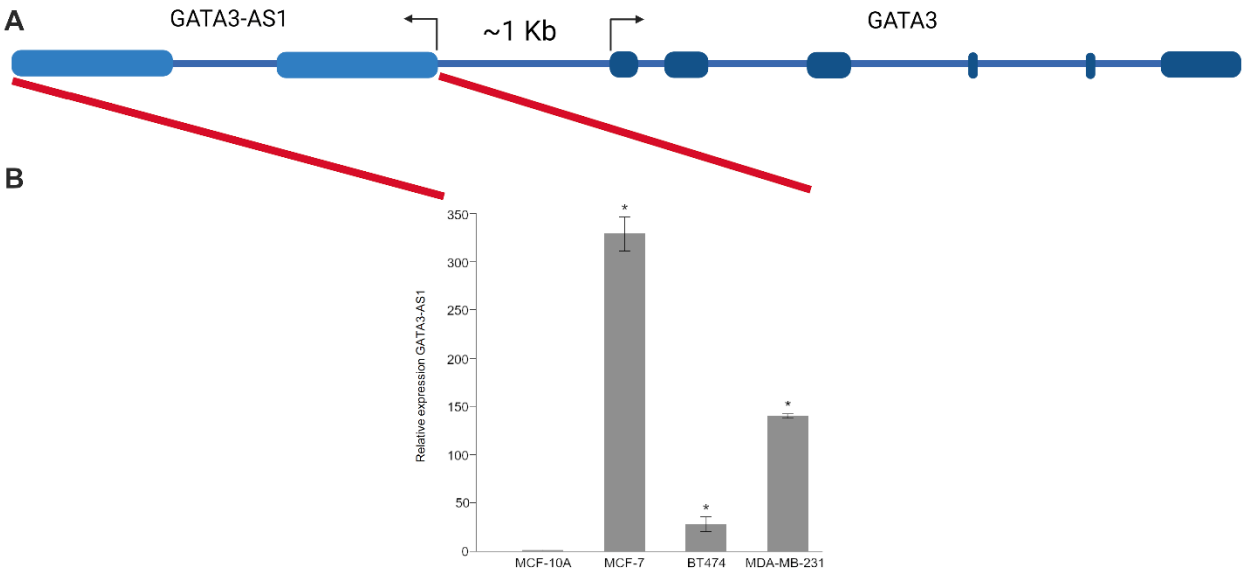


Figura 5. El lincRNA *GATA3-AS1* se encuentra en un locus de transcripción divergente junto con su gen adyacente *GATA3*. A) El lincRNA *GATA3-AS1* (azul claro) se encuentra

localizado en el brazo corto del cromosoma 10. Su locus está aproximadamente a una kilobase del gen que codifica a la proteína *GATA3* (azul oscuro), que es un factor transcripcional. B) El lincRNA *GATA3-AS1* se sobreexpresa en las líneas celulares de cáncer de mama MCF-7, BT474 y MDA-MB-231, al comparar su expresión con la línea celular transformada MCF-10A. Imagen creada con BioRender.com, modificada de Laura Contreras-Espinosa, *et al*, 2021[66].

Finalmente, en el artículo del grupo de Zhang y colaboradores[71], así como en los antecedentes directos de nuestro laboratorio, se ha reportado que *GATA3-AS1* se sobreexpresa de manera específica en pacientes con cáncer de mama subtipo luminal y que además son resistentes a la quimioterapia neoadyuvante, por lo cual se ha propuesto el uso de este lincRNA como un biomarcador de predicción de respuesta a terapia (Figura 6). No obstante, a la fecha no hay estudios científicos que expliquen este perfil de expresión de *GATA3-AS1* en cáncer de mama, su asociación con este padecimiento ni la función que cumple en el tejido mamario sano y cómo se ve afectada en las condiciones de neoplasia. En consecuencia, tampoco hay reportes que describan los procesos celulares en los que participa *GATA3-AS1* en el tejido mamario, los mecanismos a través de los cuales podría cumplir su función en las células del epitelio mamario, cuál es el mecanismo molecular con el que *GATA3-AS1* contribuye con el fenotipo neoplásico, ni si existe asociación entre la función de *GATA3* y la expresión de *GATA3-AS1* en cáncer de mama.

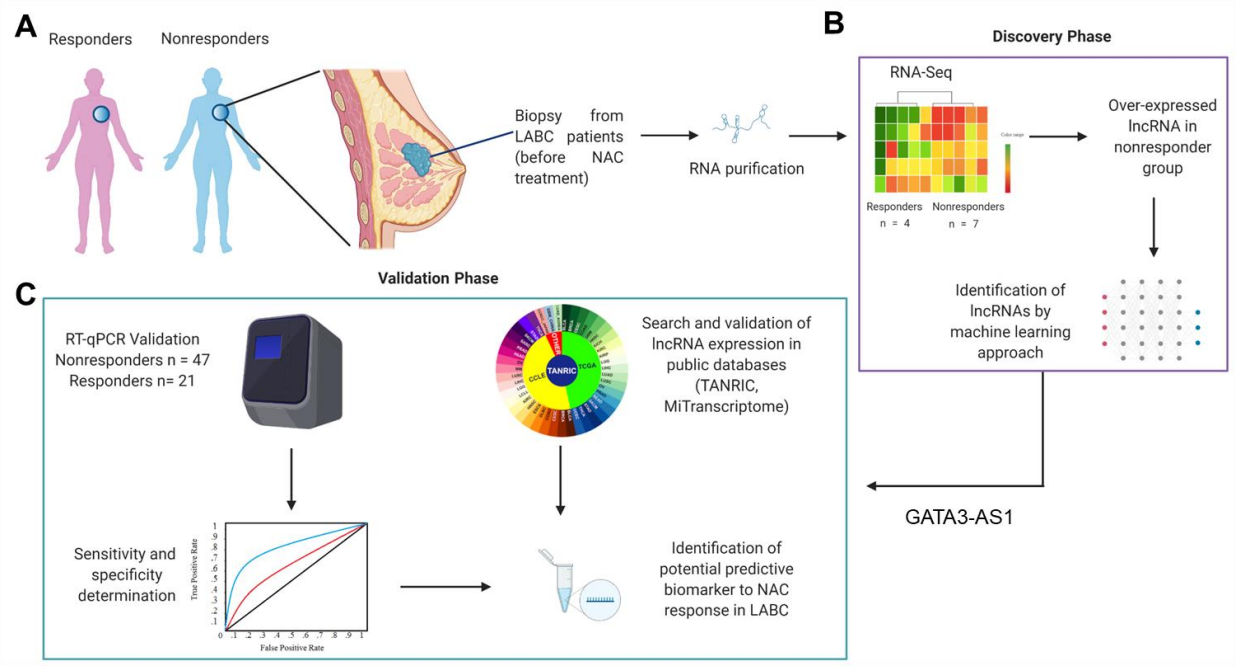


Figura 6. El lincRNA *GATA3-AS1* es un biomarcador molecular de predicción de respuesta a la quimioterapia neoadyuvante en pacientes con cáncer de mama localmente avanzado del subtipo Luminal B. A) El estudio se dividió en dos fases: La fase de descubrimiento y la fase de validación. Para el estudio, se utilizaron muestras de tumores primarios de pacientes con cáncer de mama que respondieron a tratamiento y de pacientes resistentes a la quimioterapia neoadyuvante (*Responder* y *Nonresponder*, respectivamente), cuyas biopsias fueron tomadas previas al tratamiento. B) De las pacientes de este estudio, 11 de ellas participaron en la fase de descubrimiento, por lo cual a partir de sus muestras de tumor primario se extrajo RNA, se construyeron librerías poli-A y posteriormente fueron secuenciadas por RNA-Sec (extremos pareados). Los resultados de secuenciación se analizaron con métodos bioinformáticos implementando el algoritmo de bosques aleatorios de aprendizaje de máquina (de inglés *machine learning*), con lo cual se identificaron los genes diferencialmente expresados, particularmente los lncRNAs que se sobreexpresan en las pacientes resistentes a tratamiento, entre los cuales se identificó al lincRNA *GATA3-AS1* como un potencial biomarcador de respuesta a la quimioterapia neoadyuvante. C) El lincRNA *GATA3-AS1* fue seleccionado para la fase de validación por RT-qPCR, así como para su validación en las bases de datos públicas *TANRIC (https:// https://www.tanric.org/) y TCGA (https:// https://portal.gdc.cancer.gov/). LABC: Locally*

Advanced Breast Cancer, NAC: Neoadjuvant Chemotherapy. Laura Contreras-Espinosa, *et al*, 2021[71].

Por lo tanto, en este proyecto se busca determinar si el lincRNA *GATA3-AS1* modula la expresión de su gen adyacente *GATA3* en *cis*, a través de la evaluación del efecto biológico de abatir los niveles de expresión de *GATA3-AS1* sobre la expresión de *GATA3* y sus genes blanco en líneas celulares de cáncer de mama.

5.0 PLANTEAMIENTO DEL PROBLEMA

Antecedentes preliminares del proyecto de investigación en nuestro laboratorio determinaron que el lincRNA *GATA3-AS1* está diferencialmente expresado en las líneas celulares de cáncer de mama MCF-7, BT-474 y MDA-MB-231 con respecto a la línea celular transformada de epitelio normal mamario MCF-10A. En la literatura científica se ha sugerido que *GATA3-AS1* regula la expresión de su gen adyacente *GATA3* mediante el mecanismo molecular de regulación en *cis*. Sin embargo, no hay evidencia experimental de esto en cáncer de mama, por lo que este trabajo busca determinar el mecanismo regulatorio mediante el cual *GATA3-AS1* regula la expresión de su gen adyacente *GATA3* en líneas celulares de cáncer mama.

6.0 PREGUNTA DE INVESTIGACIÓN

¿El lincRNA *GATA3-AS1* regula la activación de la expresión de su gen adyacente *GATA3* por un mecanismo molecular en *cis* en líneas celulares de cáncer de mama?

7.0 HIPÓTESIS

Si el lincRNA *GATA3-AS1* regula la activación de su gen adyacente *GATA3* por un mecanismo molecular de activación en *cis*, entonces al abatir los niveles de expresión de *GATA3-AS1* disminuirán los niveles de expresión de *GATA3*.

8.0 OBJETIVOS

8.1 General

- Determinar el mecanismo de regulación en *cis* mediante el cual, el lincRNA *GATA3-AS1* activa la expresión de su gen adyacente *GATA3*, en líneas celulares de cáncer de mama.

8.2 Particulares

1. Realizar la caracterización *in silico* de los loci genómicos del lincRNA *GATA3-AS1* y de *GATA3* en líneas celulares de cáncer de mama.
2. Abatir los niveles de expresión del lincRNA *GATA3-AS1* para determinar el efecto sobre la expresión de *GATA3*.
3. Determinar el efecto de abatir los niveles de expresión del lincRNA *GATA3-AS1* sobre la regulación en *cis* de la expresión de *GATA3*.
4. Determinar el efecto de abatir los niveles de expresión del lincRNA *GATA3-AS1* sobre la morfología de la línea celular MCF-7.

9.0 ESTRATEGIA EXPERIMENTAL

9.0.1 ABORDAJE

La determinación del mecanismo de regulación el lincRNA divergente *GATA3-AS1* lleva a cabo sobre su gente adyacente *GATA3* para activar su transcripción, se realizará en dos etapas: el análisis bioinformático y la validación experimental.

La fase de análisis bioinformático consistirá en la recopilación de información de bases de datos públicas acerca de la secuencia, la localización genómica, la localización celular y la conservación evolutiva de *GATA3-AS1* y de *GATA3*. Asimismo, se buscarán datos de acceso público de RNA-Seq para determinar la expresión diferencial de ambos transcritos en tejidos neoplásicos y sanos, particularmente en el tejido neoplásico mamario y de líneas celulares de cáncer de mama, que se procesará para determinar mediante análisis estadísticos la relación que existe entre la expresión de *GATA3-AS1* y *GATA3*. De igual manera, se analizará la información de la regulación de la expresión a nivel del promotor de *GATA3-AS1* mediante la búsqueda de datos públicos de inmunoprecipitación de la cromatina para factores transcripcionales y modificaciones post-traduccionales de histonas, así como el análisis de la secuencia promotora para identificar motivos de unión a factores transcripcionales. Finalmente, se realizará la búsqueda de información acerca las funciones biológicas que están asociadas a *GATA3-AS1* y a *GATA3*, y se validarán mediante el análisis de expresión diferencial que se hará a partir de datos del transcriptoma de la línea celular MCF-7 mediante el análisis de enriquecimiento de conjuntos de genes (*GSEA*, por sus siglas en inglés).

La fase de validación experimental se llevará a cabo evaluando la relación de la expresión de *GATA3-AS1* y *GATA3* en líneas celulares de cáncer de mama MCF-10 (línea celular transformada de fenotipo no cancerosos), MCF-7 (subtipo luminal A), BT-474 (subtipo luminal B) y MDA-MB-231 (subtipo triple negativo), mediante RT-qPCR. Por

otro lado, se validará también la localización celular de *GATA3-AS1* mediante un experimento de fraccionamiento celular. Finalmente, para validar la relación entre la expresión de *GATA3-AS1* y *GATA3* se llevará a cabo un experimento de abatimiento de la expresión de *GATA3-AS1* mediante el uso de oligonucleótidos antisentido que serán transfectados en la línea celular MCF-7, y se analizarán los cambios en la morfología del cultivo celular mediante microscopía de campo claro (Figura 7).

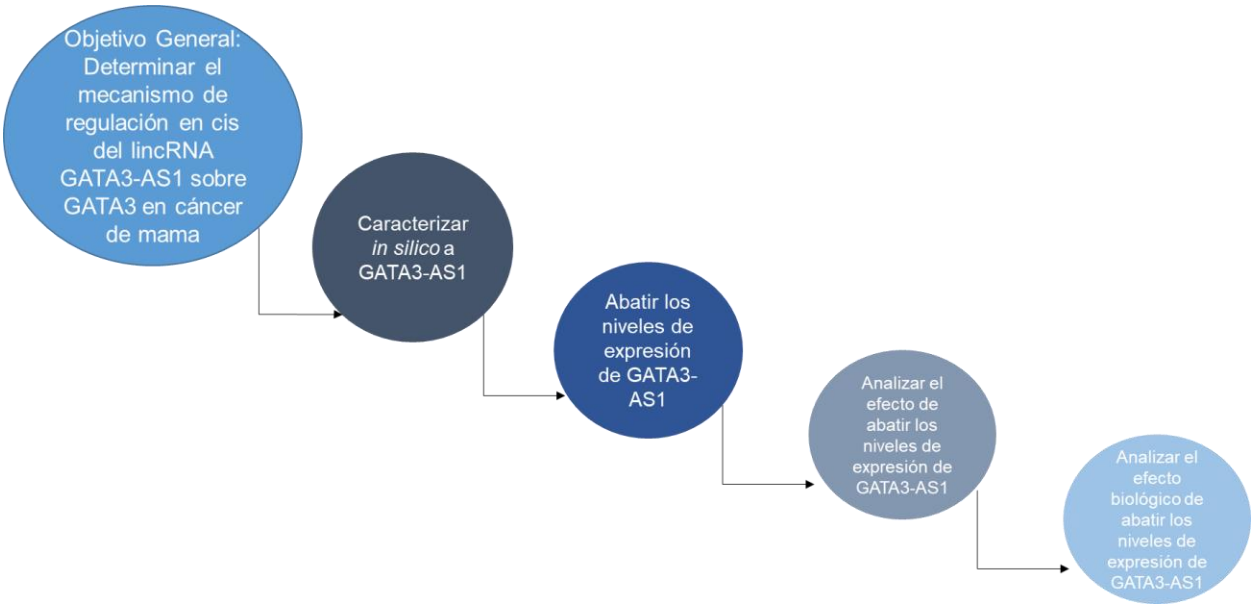


Figura 7. Flujo de trabajo general que se seguirá en la estrategia experimental para determinar el mecanismo molecular de regulación mediante el que *GATA3-AS1* activa la transcripción de *GATA3* en líneas celulares de cáncer de mama.

10.1 METODOLOGÍA

10.2 Análisis bioinformático

10.2.1 Caracterización in silico de los genes GATA3-AS1 y GATA3

La caracterización *in silico* del locus *GATA3-AS1*/*GATA3* se realizó identificando la localización genómica del mismo mediante la plataforma *Ensembl*[68],en la especie *Homo sapiens* (humano) en la versión del genoma hg38. Adicionalmente, se obtuvo la secuencia de aproximadamente 1000 pares de bases (pb) río arriba del TSS de *GATA3-AS1*.

10.2.2 Análisis de conservación de las secuencias genómicas de GATA3-AS1 y GATA3

El análisis de conservación de las regiones genómicas correspondientes al lincRNA *GATA3-AS1* y a su gen adyacente *GATA3* se realizó de manera independiente para cada gen con la herramienta *Comparative Genomics* (*Ensembl* release 104, genoma de *Homo sapiens, v. hg38*), con la opción de *alineamiento de texto* de la base de datos *Ensembl*[68]. Para el análisis, se introdujo como archivo de entrada la secuencia génica en formato FASTA considerando el cuerpo del gen y 1000 pb río arriba del sitio de inicio del primer exón.

- *GATA3-AS1*: se analizó la región chr10:8050383-8054492, cadena antisentido.
- *GATA3*: se analizó la región chr10:8052363-8075399, cadena sentido.

El análisis de alineamiento se realizó con el uso de los 90 genomas de mamíferos euterios, mediante el algoritmo *EPO* (*Enredo, Pecan, Ortheus*) extendido.

10.2.3 Determinación de la localización celular del lincRNA GATA3-AS1

Se realizó la búsqueda de Información acerca de la localización celular de *GATA3-AS1* en la línea celular MCF-7 en la base de datos *lncATLAS*[69,70]. Para validar los resultados obtenidos, se comparó con información contenida en la misma base de datos para los genes no codificantes *H19* (de localización citoplásmica) y *MALAT1* (de localización nuclear), utilizándolos como controles en el análisis bioinformático.

10.2.4 Análisis de expresión de GATA3-AS1 y GATA3 a partir de datos públicos de secuenciación

Para validar la expresión de *GATA3-AS1* en tejidos humanos (sanos y tumorales) se consultó la base de datos pública *MiTranscriptome* (v. beta)[45], que es un compendio de la expresión de lincRNAs en tejidos humanos enriquecida en datos de secuenciación masiva en paralelo de RNA (RNA-Sec) provenientes del *The Cancer Genome Atlas* (*TCGA,* release 29.0), del Centro de Patología Traslacional de Michigan y de *Encyclopedia of DNA Elements* (*ENCODE,* v.120.0), cuyos datos están normalizados como fragmentos de lecturas por kilobase de transcrito por cada millón de lecturas mapeadas (FPKM). Por otro lado, la expresión del gen adyacente *GATA3* se consultó en la base de datos *GEPIA2* (v.1)[78,79], que es un compendio de la expresión de mRNAs en tejidos humanos enriquecida de datos de RNA-Sec pertenecientes a proyectos del TCGA, cuyos datos están normalizados como transcritos por millón (TPM).

Asimismo, se obtuvieron las tablas de conteo en formato .csv, provenientes de los datos de secuenciación por RNA-Sec de las cincuenta líneas celulares de cáncer de mama provenientes de la base de datos de la *Cancer Cell Line Encyclopedia* (*CCLE*)[73], a través del portal del *Genotype-Tissue Expression* (*GTEx,* v8)[74,75]. Estos datos también se encuentran normalizados como TPM. A partir de las tablas de conteo, se extrajo la

Información de la expresión del lincRNA *GATA3-AS1* y de su gen adyacente *GATA3* para cada línea celular para la construcción de gráficos de barras de expresión.

Adicionalmente, se obtuvieron las tablas de conteo provenientes de los datos de secuenciación por RNA-Sec de las treinta y cinco líneas celulares de cáncer de mama provenientes de la base de datos de la *Harvard Medical School Library of Integrated Network-based Cellular Signatures* (*HMS LINCS*)[76,77]. Estos datos se encuentran normalizados en lecturas por kilobase de transcrito por cada millón de lecturas mapeadas (RPKM). A partir de las tablas de conteo se extrajo la Información de la expresión del lincRNA *GATA3-AS1* y de su gen adyacente *GATA3* para cada línea celular para la construcción de gráficos de barras. Finalmente, se obtuvieron datos de secuenciación por RNA-Sec en la base de datos del *Gene Expression Omnibus* (*GEO*) correspondientes a las líneas celulares MCF-10A, MCF-7, BT-474 y MDA-MB-231, cuya Información se encuentra en la Tabla 1.

Tabla 1: Claves de acceso en la base de datos GEO para los archivos de RNA-Sec de líneas celulares de cáncer de mama.

Línea Celular	**Clave de *GEO***	**Clave de SRA**	**Tipo**	**Plataforma de Secuenciación**
MCF-10A	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		

Los archivos de secuenciación *.fastq* fueron seleccionados de acuerdo con los siguientes criterios: Que provinieran de un experimento de secuenciación de extremos pareados (del inglés *paired-end*) y que además, los cultivos de los que provinieran no hubiesen sido administrados con vehículos como DMSO, disolventes o que hubiesen sido parte de experimentos de transfección.

Los archivos fueron analizados utilizando la Plataforma Galaxy[78,79], mediante el siguiente flujo de trabajo:

- Se realizó el análisis de calidad con la herramienta *FastQC Read Quality Reports* (Versión Galaxy 0.72+galaxy1): Se consideró para su uso que la calidad de los archivos debía ser Aceptable (Valor Q > 20).
- Se realizó el alineamiento y el mapeo con la herramienta *STAR* (Versión Galaxy 2.7.8a).
- Se realizó el conteo de transcritos con la herramienta *Salmon* (Versión Galaxy 1.3.0+galaxy1): Los conteos se normalizaron como TPM.

Finalmente, se extrajo la Información de conteo de los genes *GATA3-AS1* y *GATA3* para la construcción de un gráfico de cajas y se determinó la significancia estadística de las diferencias en la expresión del lincRNA *GATA3-AS1* y *GATA3* mediante un análisis de varianzas (ANOVA) seguido de una prueba de Tukey (valor $p < 0.05$).

10.2.5 Análisis de correlación de los niveles de expresión de GATA3-AS1 y GATA3 a partir de datos públicos de secuenciación

El análisis de correlación de la expresión del lincRNA *GATA3-AS1* y su gen adyacente *GATA3* se realizó con la Información extraída de los resultados de secuenciación de las bases de datos *CCLE* y *HMS LINCS* mencionadas en el apartado ***Análisis de expresión de GATA3-AS1 y GATA3 a partir de datos públicos de secuenciación.*** De este análisis se construyó un gráfico de puntos (*Dot plot*). Adicionalmente, se consultó la base de datos pública *The Atlas of ncRNA in Cancer* (*TANRIC*)[80,81] para obtener la información de correlación de la expresión del lincRNA *GATA3-AS1* y de su gen adyacente *GATA3* a partir de los datos de secuenciación por RNA-Sec provenientes de pacientes con cáncer de mama que se encuentran depositados en los proyectos del *TCGA*. Finalmente, el análisis de correlación se llevó a cabo calculando el coeficiente de correlación de

Pearson, considerando el valor del coeficiente de correlación > 0.5 y el valor *p* <0.05, ya que permite validar la significancia estadística del resultado de correlación en la expresión de genes.

10.2.6 Caracterización funcional in silico del lincRNA GATA3-AS1 y de GATA3

Las funciones a las que se encuentra asociado el lincRNA *GATA3-AS1* fueron consultadas en la base de datos de *Function Annotation of non-coding RNA* (*FARNA*)[82,83] y filtradas considerando aquellas con valor de tasa de falsos descubrimientos (FDR) menor a 0.05. Con la información recopilada se construyó un gráfico de puntos. Por otro lado, las funciones asociadas a la proteína GATA3 fueron consultadas en la *Search Tool for the Retrieval of Interacting Genes/Proteins* (*STRING*, v.11.0)[84]. Se seleccionaron sólo aquellas que provinieran de evidencia experimental, experimentos de coexpresión y de predicciones bioinformáticas, y cuya asociación tuviera significancia estadística cumpliendo con los siguientes parámetros: Fuerza de interacción (del inglés *strength*) mayor a 1.5 y valor FDR < 0.05.

10.2.7 Caracterización in silico de la región promotora del lincRNA GATA3-AS1

10.2.7.1 Construcción de los mapas de cromatina para el locus GATA3-AS1/GATA3

Los mapas de cromatina del locus *GATA3-AS1*/*GATA3* para las líneas celulares MCF-10A y MCF-7 se construyeron con la información disponible en la base de datos *WashU Epigenome Browse*r (release 2018-2021, v53.3.1)[85]. Los mapas de cromatina incluyen la información de los experimentos:

- MCF-10A:
    - Secuenciación de inmunoprecipitación de cromatina (ChIP-Sec): MYC, RNAPOL2A (P-Ser5), RNAPOL2A (P-Ser2), H3K4me3, H3K4me1, H3K4me2, H3K27ac, H3K9ac, H3K36me3, H3K9me2, H3K9me3, H3K27me3
    - DNasa I
    - Ensayo para la identificación de cromatina accesible a transposasa usando secuenciación (ATAC-Sec), para la identificación de regiones de cromatina abierta o accesible para la transcripción.
- MCF-7:
    - ChIP-Sec: MYC, *GATA3*, SUZ12, ESR1, RNAPOL2A (P-Ser2), H3K4me3, H3K4me1, H3K4me2, H3K27ac, H3K9ac, H3K36me3, H3K9me2, H3K9me3, H3K27me3
    - DNasa I
    - ATAC-Sec
    - Análisis de la expresión de genes con modificación Cap 5’ (CAGE) en la cadena sentido (+) y antisentido (-).
    - RNA-Sec en la cadena sentido (+) y antisentido (-).
    - Secuenciación masiva en paralelo de RNA de extremo pareado marcado (RNA-PET) en la cadena sentido (+) y antisentido (-).

Los experimentos de ChIP-Sec, se analizaron con la opción *Peak Finder* (v53.3.1), que permite determinar si la señal de enriquecimiento observada en una región genómica dada es significativa.

Por otro lado, se utilizó la herramienta *Toolkit for Cistrome Data Browser* (v. 2014) disponible en la Plataforma CistromeDB[86,87], que contiene información de experimentos de ChIP-Sec para identificar los factores transcripcionales y las modificaciones post-traduccionales de histonas que podrían regular la expresión del locus *GATA3-AS1*/*GATA3*. Para ello, se seleccionó la región chr10:8053465-8054672, que comprende aproximadamente 1200 pb entre el sitio de inicio del primer exón de *GATA3-AS1* y el sitio de inicio del primer exón de *GATA3*. Posteriormente, se buscó la asociación entre la presencia de factores transcripcionales y de modificaciones post-traduccionales de histonas cuyo potencial regulatorio predicho fuese mayor o igual a 0.2 para la línea celular MCF-7. El potencial regulatorio se calcula utilizando los valores numéricos de la expresión del gen de interés, del pico de enriquecimiento y en el caso particular de los factores transcripcionales, del enriquecimiento del motivo de unión en la región genómica de interés. En conjunto, la tasa de cambio final obtenida de este análisis permite predecir la activación o la represión de la transcripción del gen de interés y la probabilidad de que sea regulado por los factores transcripcionales o modificaciones post-traduccionales de histonas que se analizan[88].

10.2.7.2 Análisis de los motivos de reconocimiento de proteínas de unión a DNA en la región promotora del lincRNA GATA3-AS1

Para la identificación y predicción de motivos de unión a factores transcripcionales en la región promotora de *GATA3-AS1*, se obtuvo la secuencia de DNA que comprende 1000 pb río arriba del TSS de *GATA3-AS1* (chr10: 8054492-8055492), a partir de la información disponible en la base de datos *Ensembl*[68]. El análisis de motivos de unión

se realizó con la paquetería informática *MEME Suite* (versión 5.3.3)[28,89], como se indica a continuación:

- Se realizó la búsqueda de secuencias correspondientes a motivos de unión con la herramienta *MEME* (versión 5.3.3)
- Se validó la presencia de los motivos de unión y frecuencia de cada motivo seleccionado con la herramienta *FIMO* (versión 5.3.3)
- Se escaneó la localización de los motivos de unión con mayor frecuencia de alineamiento y significativamente enriquecidos en la región promotora con la herramienta *MAST* (versión 5.3.3)
- Se comparó la identidad de cada motivo de unión identificado contra la base de datos HUMAN (*Homo Sapiens*) DNA-HOCOMOCO Human (v11 core) con la herramienta *Tomtom* (versión 5.3.3) con la finalidad de identificar los probables factores transcripcionales o las proteínas de unión a DNA que reconocen los motivos identificados. La herramienta también permite obtener los gráficos correspondientes a los *LOGOS* de cada motivo de unión identificado. Un *LOGO* es la representación gráfica de los nucleótidos que componen al motivo de unión.

Para la construcción del esquema de la región promotora, se consideraron sólo los factores transcripcionales o proteínas de unión a DNA cuyo motivo de unión consenso tuviera la mayor cantidad de bases alineadas con el motivo de unión identificado en la región promotora, cuyo valor q fuese menor o igual a 0.1 y su valor *p* menor a 0.001 (intervalo de confianza al 95%).

Los factores transcripcionales y/o proteínas de unión a DNA seleccionados fueron utilizados en un análisis de redes para identificar las vías de señalización en las que participan utilizando la herramienta de análisis funcional de la base de datos *STRING*[84], de la cual se obtuvo la información del análisis de las vías de señalización de la base de datos de la *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*)[90]. Para validar la

información recabada en *STRING*, se realizó el mismo análisis con la herramienta *GeneMania*[91,92]. Con ello se construyeron gráficos de puntos (*Dot plot*).

10.2.8 Análisis de las interacciones del lincRNA GATA3-AS1 con proteínas, RNA y DNA.

Para identificar los motivos de reconocimiento de proteínas de unión a RNA presentes en la secuencia de nucleótidos de *GATA3-AS1*, se obtuvo la secuencia de cDNA del lincRNA *GATA3-AS1* (isoforma canónica *GATA3-AS1*-201) de la base de datos *Ensembl*[68], de la cual se obtuvo la secuencia complementaria y se convirtió a código de RNA. Posteriormente, se analizó la secuencia en la herramienta *MEME-Suite*[28,89] tal como se indica en el apartado ***Análisis de los motivos de reconocimiento de proteínas de unión a DNA en la región promotora del lincRNA GATA3-AS1*** con las siguientes especificaciones: Para la herramienta *Tomtom* (versión 5.3.3) se utilizó la base de datos RNA-Ray2013 *Homo Sapiens*.

Asimismo, se realizó la predicción de interacciones del lincRNA *GATA3-AS1* con la herramienta *catRAPID omics* (en la modalidad de interacciones entre un transcrito dado y el proteoma) de la base de datos *catRAPID*[97,98]. Se seleccionaron sólo aquellas cuyos valores de poder discriminatorio (la probabilidad de distinguir correctamente una interacción verdadera positiva de una falsa positiva) y de fuerza de interacción (la probabilidad de que el motivo de unión analizado presente en la secuencia tenga identidad comparado con el motivo reconocimiento de la proteína de unión a RNA) fueran mayores a 90, ya que al ser más cercano a 100, la probabilidad de la interacción es mayor. Adicionalmente, se realizó la búsqueda de interacciones asociadas a *GATA3-AS1* con la base de datos *RNAInter*[95,96], cuyo puntaje de interacción estuviera en el rango de valores 0.5-1.0, debido a que valores menores a 0.5 indican interacciones débiles con poca probabilidad de ocurrencia. Asimismo, se obtuvo un gráfico de círculo con las 50 interacciones estadísticamente significativas que proporciona la plataforma. Finalmente,

se seleccionaron sólo las interacciones que estuvieran presentes en los resultados de ambos análisis, considerando los puntos de corte descritos anteriormente para cada herramienta utilizada.

Por otro lado, se analizaron las posibles interacciones con miRNAs usando la herramienta DIANA (v3)[101], con lo que se obtuvo una lista de miRNAs predichos para interactuar con el lincRNA *GATA3-AS1*. Con esa lista, se analizó la interacción del lincRNA *GATA3-AS1* con otras biomoléculas, incluidos miRNAs, del lincRNA *GATA3-AS1* con la herramienta miRNet[98]. La misma herramienta proporciona el análisis de vías de señalización enriquecidas, y se filtraron para obtener sólo aquellas con valor *p* menor a 0.05.

10.2.9 Análisis de las vías de señalización enriquecidas en la línea celular MCF-7 en las que participan GATA3-AS1 y GATA3

Se llevó a cabo el análisis de expresión diferencial del transcriptoma completo (que incluye genes codificantes y no codificantes) con la herramienta *DESeq2* (versión 1.28.1) en *R/RStudio* (versión 4.0.5), utilizando las tablas de conteo obtenidas de los seis archivos de secuenciación de las líneas celulares MCF-10A y MCF-7 descritas en el apartado ***Análisis de expresión de GATA3-AS1 y GATA3 a partir de datos públicos de secuenciación*** (Tabla 2) y se utilizó como control a la línea celular MCF-10A. De este análisis se construyó un gráfico de volcán, el cual permite identificar los genes diferencialmente expresados que son estadísticamente significativos en la línea celular MCF-7.

Los resultados del análisis de expresión diferencial fueron utilizados para realizar el análisis de enriquecimiento de vías de señalización y procesos celulares con la herramienta *fgsea* (versión 1.14.0) y se consideraron sólo aquellos con valor *p* ajustado

menor o igual a 0.05, con valores de puntaje normalizado de enriquecimiento (*NES*) mayores a 1.0 y que contuvieran en el conjunto de genes analizado a los genes *GATA3* y *ESR*. Una vez aplicados estos filtros, se identificaron adicionalmente los procesos en los que también estuviera contenido *GATA3-AS1*. Con la información obtenida se construyó un gráfico de puntos (*Dot plot*) que representa el enriquecimiento de cada proceso biológico respecto al tamaño de los puntos y su color, en la línea celular MCF-7.

Finalmente, los resultados del análisis de expresión diferencial se filtraron para identificar aquellos genes diferencialmente expresados en la línea celular MCF-7 cuyos valores de tasa de cambio (*log2 Fold Change*) fuesen menores a -1.5 y mayores a 1.5, y que además su valor de *p* ajustada fuera menor a 0.05. Con esa información se construyó un mapa de calor y un gráfico de volcán que muestran los genes diferencialmente expresados con significancia estadística en la línea celular MCF-7, en ambos casos

10.3 Análisis Experimental

10.3.1 Cultivo de las líneas celulares de mama.

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 (línea celular transformada). Su morfología es epitelial. Su crecimiento se llevó a cabo en medio Eagle con modificación de Dulbecco con mezcla de nutrientes F12 Ham (DMEM F12), suplementado con suero fetal bovino (SFB) al 5%, en condiciones de atmósfera de CO_2 al 5%, a 37ºC[99].

- **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 Eagle's modificado de Dulbecco (DMEM) suplementado con SFB al 10% e insulina recombinante (0.01 mg/mL), en condiciones de atmósfera de $C0_2$ al 5%, a 37ºC[100].

- **BT-474** (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[101].

- **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 DMEM GlucoMAX suplementado con SFB al 10%, en condiciones de atmósfera de CO_2 al 5%, a 37ºC[102,103].

10.3.2 Purificación de RNA Total de líneas celulares

10.3.2.1 Extracción de RNA

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 *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 diluyó en 50 µL de agua con dietil pirocarbonato (DEPC) libre de RNasas, se incubó a 50ºC durante 10 min, y se colocó en hielo (el RNA extraído se conservó a -80ºC)[104].

10.3.2.2 Cuantificación (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[105].

10.3.3 RT-qPCR

10.3.3.1 Diseño de oligonucleótidos

Utilizando la secuencia de los lncRNAs *GATA3-AS1*, *MALAT1*, *XIST1.1*, *NEAT1.2*, *H19.1* (proveniente de *LNCipedia*[106,107]) y los genes codificantes *GATA3* y *RPS28* (obtenidas de *Ensembl*[68]), se diseñaron pares de oligonucleótidos para la PCR en tiempo real (RT-qPCR) con el uso de la herramienta *Primer-Blast*[108]. Los parámetros de diseño fueron los preestablecidos por la herramienta.

Se seleccionaron los pares de oligonucleótidos con mayor contenido de GC (%GC > 50\%), con amplicones de longitud entre 100 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 autocomplementariedad 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*[109], y se solicitó su síntesis[110] (Tabla 2).

Tabla 2. Información de los oligonucleótidos diseñados para los experimentos de PCR en tiempo real

Nombre	Secuencia (5'-3')	Longitud (bases)	Longitud del amplicón (bases)
GATA3-AS1 F	CGCAGACAGAAAAGAAGCCG	19	101
GATA3-AS1 R	GCTGGAATGGGAAGGGACTT	18	101
GATA3 F	GTGCTCGGAGGGTTTCTTGT	20	144
GATA3 R	TGCACGCTGGTAGCTCATAC	20	144
RPS28 F	CGATCCATCATCCGCAATG	20	133
RPS28 R	AGCCAAGCTCAGCGCAAC	20	133
XIST1.1 F	CTCTTCATTGTTCCTATCTGCC	22	105
XIST1.1 R	CTGACTTCCTTCAGTGTGTTC	21	105
NEAT1.2 F	CACAACGCAGATTGATGCC	19	100
NEAT 1.2 R	TCCGAGAAACGCACAAGAAG	20	100
H19.1 F	GCATGCTCCAGAGGGAATCG	20	127
H19.1 R	GCTTCAACTGATTCCGTGGC	20	127
MALAT1-104 F	GGATTCCAGGAAGGAGCGAG	20	104
MALAT1-104 R	AGGATCCTCTACGCACAACG	20	104

Para la cuantificación de transcritos tipo lncRNA y mRNA por PCR en tiempo real fue necesario obtener cDNA a partir del RNA extraído de las líneas celulares de cáncer de mama.

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 $MgCl_2$, 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 incubó 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 bidestilada, 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 3. Al final se diluyó el cDNA 1:4 en agua calidad *Biología molecular*.

Tabla 3. 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, usando el 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 oligonucleótidos 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 de 40 ciclos (40x), como se muestra en la Tabla 4. 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 4. Programa de termociclador para PCR en tiempo real

Etapa de desnaturalización		**Etapa de Amplificación (40x)**			**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

10.3.3.2 *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 las mismas características experimentales mencionadas en el apartado **Evaluación de la calidad del cDNA por PCR en tiempo real**, y se incluyó por cuadruplicado técnico +RT, -RT y NTC para los oligonucleótidos del gen constitutivo y para el gen no codificante problema.

10.3.3.3 *Determinación de la eficiencia de amplificación de los oligonucleótidos*

La eficiencia de amplificación de los oligonucleótidos se llevó a cabo realizando una PCR en tiempo real bajo 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 1:10,000 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 los 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*[111]. 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 90% y el 105%[112,113].

10.3.3.4 *Cuantificación relativa del lncRNA y el mRNA*

La cuantificación de los transcritos de los lncRNAs *GATA3-AS1*, *MALAT1*, *XIST1.1*, *NEAT1.2*, *H19.1* y de los mRNAs de *GATA3* y *RPS28,* 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 triplicados biológicos y cuadruplicados técnicos para cada línea celular,

tanto para los oligonucleótidos del gen constitutivo como para los del gen no codificante 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)[113,114].

Ecuación 1: $\Delta Cq = 0.05\ ^{(Cq_{(gen\ problema)} - Cq_{(gen\ constitutivo)})}$

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 del experimento de fraccionamiento celular, al no contar con una línea calibradora la cuantificación se limita al método ΔCq (normalizando con el RNA total)[115].

10.3.4 Fraccionamiento celular

Para la separación de los diferentes compartimentos celulares, se realizó un protocolo de separación por centrifugación, para lo cual se cultivaron 4 cajas de cultivo de 75 cm^2 de cada línea celular[116]. El procedimiento experimental consistió en:

- **Preparación de soluciones**: Durante el protocolo se utilizaron las siguientes soluciones, utilizando en todos los casos agua grado molecular (tratada con DEPC):

    - **Solución amortiguadora de lisis citoplásmico**: Solución compuesta por NP40 al 0.15%, TrisHCl 10 mM (M, pH ~7), NaCl 150 mM y agua cuanto baste para (c. b. p.) aforar.
    - **Solución amortiguadora de sucrosa**: Solución compuesta por TrisHCl 10 mM, NaCl 150 mM, sucrosa en polvo al 24% partes por volumen (P/V), y agua c. b. p.
    - **Solución PBS/EDTA**: Solución compuesta por PBS 1X y EDTA 10 mM.
    - **Solución amortiguadora de glicerol**: Solución compuesta por TrisHCl 20 mM, NaCl, 75 mM, EDTA 0.5 mM, glicerol al 50%, DTT 85 mM (se añade en el momento de uso) y agua c. b. p.
    - **Solución amortiguadora de lisis nuclear**: Solución compuesta por HEPES 20 mM, $MgCl_2$ 7.5 mM, EDTA 0.2 mM, NaCl 0.3 mM, urea 1 M, NP40 1% y DTT 1 mM (se añade en el momento de uso).

- **Procedimiento de fraccionamiento celular**

    - Las cajas de cultivo fueron tratadas con tripsina para la separación de las células. Posteriormente, se centrifugaron a 1200 RPM por 5 min y se disolvió la pastilla celular en 5 mL de PBS frío. Esta suspensión celular se centrifugó a 2000 RPM por 5 min.
    - Se retiró el PBS, se disolvió la pastilla celular en 200 µL de solución amortiguadora de lisis citoplásmico y se incubó en hielo por 5 min.
    - Se añadieron 500 µL de solución amortiguadora de sucrosa, lentamente hasta ver la formación de 2 fases, y se centrifugó a 14000 RPM por 10 min a 4°C. Se retiró el sobrenadante y se colocó en un tubo nuevo con 300 µL de TRIzol (RNA citoplásmico).
    - La pastilla de núcleos se lavó con 100 µL de solución PBS/EDTA y se disolvió en 200 µL de solución amortiguadora de glicerol. Al dispersar las células, se añadieron 200 µL de solución de lisis nuclear y se agitó con vórtex por 10 segundos. Posteriormente, se incubó la suspensión en hielo por 2 min y se centrifugó a 14000 RPM por 2 min a 4°C. El sobrenadante se transfirió a un tubo nuevo con 300 µL de TRIzol (RNA en nucleoplasma).
    - La pastilla restante se disolvió en 100 µL de PBS frío y se sonicó con el equipo COVARIS, con 3 ciclos programados de sonicación por con descansos de 30 segundos, a 50 W del poder de sonicación, en un tiempo total de 180 segundos. A la suspensión final se le añadió 1 mL de TRIzol (RNA asociado a cromatina).
    - Las suspensiones con TRIzol se almacenaron a -80°C, para posteriormente extraer el RNA como se describe en la sección *Purificación de RNA Total de líneas celulares.*

- El análisis de expresión por qPCR se realizó como se indica en el apartado ***Cuantificación relativa de transcritos***, de acuerdo con los siguientes supuestos experimentales:

    - Control de acumulación en cromatina: XIST1.1
    - Controles de acumulación nuclear: NEAT1.2, MALAT1
    - Control de acumulación en citoplasma: H19.1
    - Gen endógeno: RPS28

10.3.5 Transfección de ASOs en la línea celular MCF-7

10.3.5.1 Diseño de ASOs

Para el diseño de los oligonucleótidos antisentido (ASOs) dirigidos contra *GATA3-AS1*, se consultó la secuencia del transcrito *GATA3-AS1*-201 en *Ensembl*[68] y se seleccionó sólo la secuencia correspondiente a los exones presentes en el transcrito maduro. A partir de ello, se utilizaron las herramientas *RNAFold*[117] y *RNAstructure*[118] para determinar la probable estructura secundaria del lincRNA *GATA3-AS1*. En dichas estructuras, se buscaron regiones de al menos 20 bases de longitud, enriquecidas en GC, que no tuvieran bases repetidas más de 4 veces y que además correspondieran con estructuras no complementarias dentro de la estructura secundaria (es decir, la presencia de tallo-asa o *loops*). Aquellas que cumplieran con esos criterios se seleccionaron para obtener la secuencia complementaria en la herramienta *Reverse Complement*[119] y poder analizar las propiedades termodinámicas de la secuencia complementaria en la herramienta *OligoAnalyzer* de IDT[120]. Las secuencias con menor probabilidad de dimerización, heterodimerización, autocomplementariedad y con temperaturas de fusión entre 58-60°C se probaron en la herramienta *In Silico PCR*[109] para determinar que fueran dirigidas únicamente al transcrito del lincRNA *GATA3-AS1*[121,122]. Las secuencias que cumplieron con estas características se enlistan en la tabla 5.

Tabla 5: Secuencia de ASOs utilizados para la supresión de la expresión de *GATA3-AS1*

ASO	***Secuencia***
ASOA	5’ mA*mG*mU* mG*mG*A* T*T*T* G*G*A* G*T*C* mU*mU*mC* mU*mU 3’
ASOB	5’ mG*mG*mC* mU*mU*C* T*T*C* T*A*T* C*T*C* mG*mA*mA* mU*mU 3’
ASO-LacZ	5’ mG*mC*mU*mU*mC*A*T*C*C*A*C*C*A*C*A*mU*mA*mC*mA*mG 3’
ASO-MALAT1-AS2	5’ mA*mT*mG*mG*mA*G*G*T*A*T*G*A*C*A*T*mA*mT*mA*mA*mT 3’

Finalmente, los ASOs se diluyeron en solución salina amortiguadora de fosfatos de Dulbecco (D-PBS) a una concentración de 25 µM.

10.3.5.2 Preparación de la solución D-PBS

Se preparó la solución D-PBS de acuerdo con lo requerido en la tabla 6:

Tabla 6: Componentes para la preparación de D-PBS

Componente	Concentración
$CaCl_2$	25 µM
KCl	25 µM
$MgCl_2$	25 µM
NaCl	25 µM
Na_2HPO4	25 µM

Se diluyeron los componentes en agua grado Biología molecular y la solución final se esterilizó en autoclave.

10.3.5.3 *Cultivo celular y transfección de ASOs*

10.3.5.3.1 Construcción de la curva de concentración para la estandarización de la transfección de ASOs en la línea celular MCF-7

Se cultivó la línea celular MCF-7 en una caja de cultivo de 75 cm^2 cuya superficie fue tratada con exposición a gas plasma, pase 16, hasta tener una confluencia del 80%. Se contabilizaron las células y se sembraron alrededor de 96,000 células por pozo en cajas de cultivo de 6 pozos. En total se cultivaron 8 cajas con las siguientes condiciones:

- ASO A
    - 50 nM (n = 3)
    - 100 nM (n = 3)
    - 150 nM (n = 3)
    - 200 nM (n = 3)

- ASO B
    - 50 nM (n = 3)
    - 100 nM (n = 3)
    - 150 nM (n = 3)
    - 200 nM (n = 3)

- ASO A + B
    - 50 nM (n = 3)
    - 100 nM (n = 3)
    - 150 nM (n = 3)
    - 200 nM (n = 3)

- ASO LacZ
    - 50 nM (n = 2)
    - 100 nM (n = 2)
    - 150 nM (n = 2)
    - 200 nM (n = 2)

- ASO Malat1
    - 100 nM (n = 3)

Al alcanzar la confluencia celular el 50-70%, se prepararon para el experimento de transfección, que consistió en añadir OptiMEM (previamente atemperado a 37°C por 1 hora) y el ASO correspondiente en tubo de centrifugación de 15 mL (Tubos A), se incubó por 5 min a temperatura ambiente y se añadió lipofectamina RNAiMAX diluida en Optimem (98 µl de lipofectamina en 7.25 mL de OptiMEM. Tubo B). Después de incubar 15 min a temperatura ambiente, se añadió OptiMEM hasta obtener un volumen suficiente para colocar 1 mL por pozo. Posteriormente, se retiró el medio de cultivo de los pozos, se adicionó por goteo 1 mL de la mezcla del tubo A + tubo B, y se regresó a incubación a 37°C por 6 horas. Pasado ese tiempo, se adicionó a cada pozo 1 mL de medio de cultivo suplementado DMEM y se incubó a 37°C por 18 horas. Pasado ese tiempo, se repitió el experimento y a las 48 hrs posteriores a la primera transfección, se añadió TRIzol para la purificación de RNA (apartado ***Purificación de RNA Total de líneas celulares***). Debido a la baja concentración obtenida de RNA en cada pozo (concentración de RNA < 50 ng/ul), se realizó la mezcla final del RNA obtenido de cada condición (pool) con la finalidad de lograr la detección de los genes a cuantificar por RT-qPCR.

10.3.5.3.2 Transfección de ASOs a 50 nM

El experimento con la concentración estandarizada de ASOs determinada mediante la construcción de la curva de estandarización (50 nM) se realizó por triplicado biológico de acuerdo con las siguientes condiciones:

- ASO-A 50 nM *GATA3-AS1*
    - 3 cajas Petri P60
    - 3 cajas Petri P100
- ASO-MALAT1 100 nM
    - 3 cajas Petri P60
- ASO-LacZ 50 nM
    - 3 cajas Petri P60
    - 3 cajas Petri P100

Las condiciones de cultivo se encuentran en el apartado ***Construcción de la curva de concentración para la estandarización de la transfección de ASOs en la línea celular MCF-7***. Por otro lado, el análisis de expresión por RT-qPCR de RPS28, *GATA3-AS1*, *GATA3* y MALAT1 se realizó como se indica en el apartado ***Cuantificación relativa de transcritos.***

11.0 RESULTADOS

11.1 Caracterización in silico del lincRNA GATA3-AS1 y del gen codificante GATA3 en líneas celulares de cáncer de mama

Con el objetivo de caracterizar el locus genómico en el que se encuentran el lincRNA *GATA3-AS1* y su gen codificante adyacente *GATA3* (locus *GATA3-AS1*/*GATA3*), se buscó información de ambos genes en la base de datos *Ensembl*[68], a partir de la cual se identificó que *GATA3-AS1* se localiza en el cromosoma 10, su longitud es de aproximadamente 3.2 kb y se transcribe a partir de la cadena antisentido. Por otro lado, el TSS de su gen adyacente *GATA3*, que codifica para un factor transcripcional, se encuentra a una distancia aproximada de 1200 pb del TSS de *GATA3-AS1*, y se transcribe a partir de la cadena sentido, por lo que *GATA3-AS1/GATA3* se considera un locus de transcripción divergente (Figura 8A), que adicionalmente se encuentra conservado en 38 especies de mamíferos, de las cuales 7 especies de primates presentan la mayor cobertura de la secuencia conservada (Figura 8B-C).

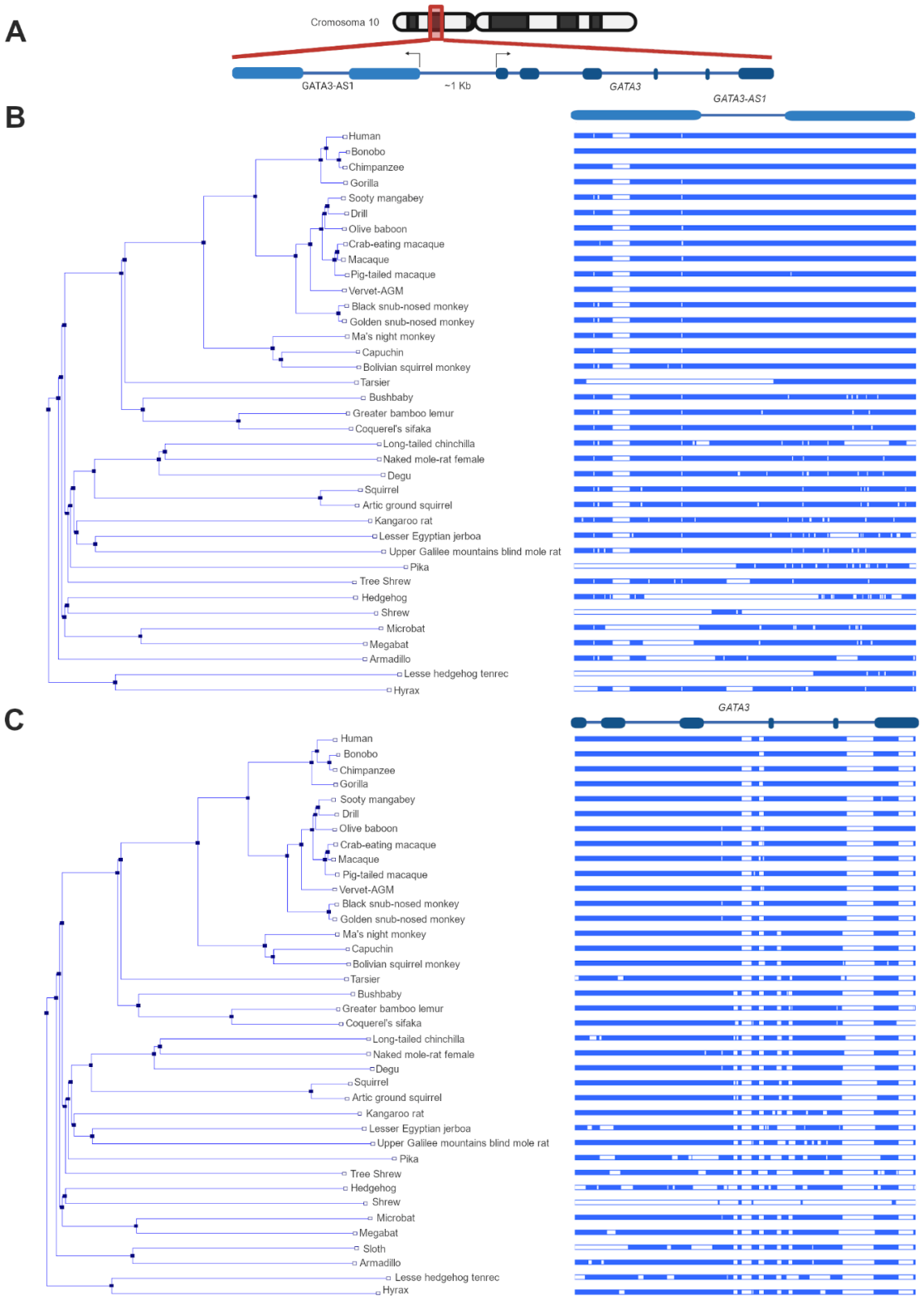


Figura 8. La transcripción del lincRNA *GATA3-AS1* es divergente. A) El locus en el que se encuentran *GATA3-AS1* (azul) y *GATA3* (marino) se localiza en la banda p14 del cromosoma 10

humano y el dominio génico mide aproximadamente 27 kb de longitud. B) Árbol filogenético del análisis de conservación evolutiva de la secuencia genética de *GATA3-AS1* en los mamíferos euterios (panel izquierdo) y su correspondiente esquema de conservación de nucleótidos (barra vertical, panel derecho). Los bloques azules representan los nucleótidos que presentan similitud con la secuencia consenso de *GATA3-AS1*. C) Árbol filogenético del análisis de conservación evolutiva de la secuencia genética de *GATA3* en los mamíferos euterios (panel izquierdo) y su correspondiente esquema de conservación de nucleótidos (barra vertical, panel derecho). Los bloques azules representan los nucleótidos que presentan similitud con la secuencia consenso de *GATA3*.

Debido a que el lincRNA *GATA3-AS1* se encuentra aproximadamente a 1 Kb de su gen adyacente *GATA3* y que su transcripción es bidireccional, es posible que la expresión de ambos genes esté relacionada como se ha demostrado para otros pares de genes como el lncRNA *EVX1-AS1* y su gen adyacente *EVX1*[12]. Para corroborar la relación que existe entre la localización genómica de *GATA3-AS1* y *GATA3*, se realizó la búsqueda en bases de datos públicas para determinar si existía información acerca de la expresión de los genes contenidos en el locus *GATA3-AS1*/*GATA3* en diferentes tejidos humanos sanos y neoplásicos. La información disponible en las bases de datos públicas indica que *GATA3-AS1* y *GATA3* se expresan diferencialmente en los tejidos neoplásicos, y particularmente se sobreexpresan en los tumores mamarios. Sin embargo, los resultados sugieren que la expresión de *GATA3-AS1* presenta mayor especificidad de expresión en cáncer de mama respecto a su gen adyacente *GATA3* (Figuras 9A-B). Por otro lado, la expresión de *GATA3-AS1* y *GATA3* presentan una correlación positiva en muestras de pacientes con cáncer de mama de acuerdo con la información disponible en la base de datos de acceso público de *TANRIC*[81] (Figura 9C). Por lo tanto, los resultados sugieren que la expresión de ambos genes está relacionada entre sí y con el fenotipo neoplásico del tejido mamario.

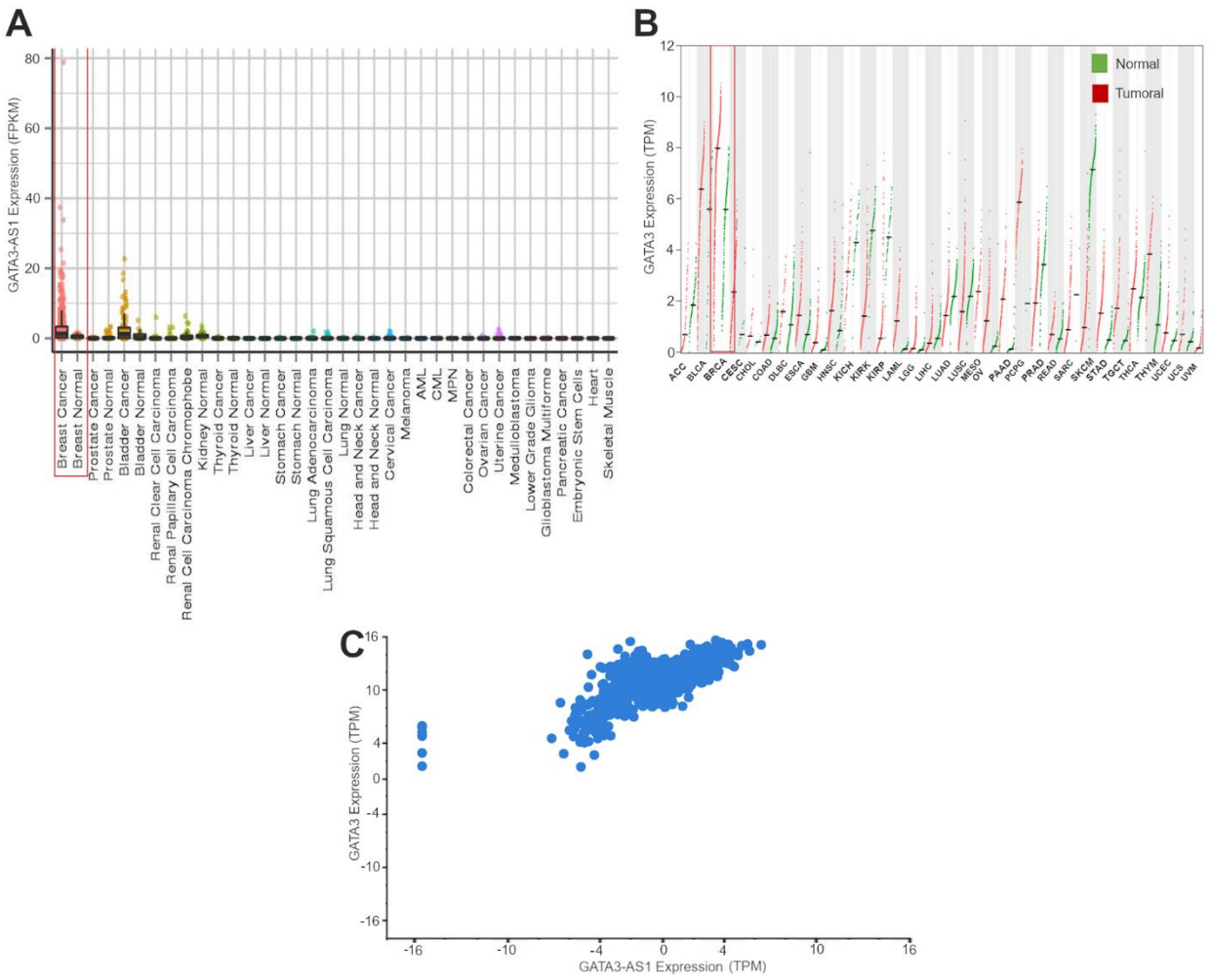


Figura 9. *GATA3-AS1* y *GATA3* se expresan diferencialmente en cáncer de mama. A) *GATA3-AS1* se sobreexpresa en cáncer de mama respecto al tejido mamario sano (rectángulo rojo) y a otros tipos de cáncer, de acuerdo con la información disponible en la base de datos *MiTranscriptome*[45] (n=6503). B) El gen *GATA3* se sobreexpresa en cáncer de mama respecto al tejido mamario sano (rectángulo rojo), de acuerdo con la información disponible en la base de datos *GEPIA*[72] (n=9135). El significado de los demás acrónimos (todos de sus siglas en inglés) se enlista a continuación: Carcinoma adrenocortical (ACC), Carcinoma Urotelial de Vejiga (BLCA), Carcinoma cervical de células escamosas y adenocarcinoma endocervical (CESC), Colangiocarcinoma (CHOL), Adenocarcinoma de colon (COAD), Linfoma difuso de células B largas (DLBC), Carcinoma esofágico (ESCA), Glioblastoma multiforme (GBM), Carcinoma de células escamosas de cabeza y cuello (HNSC), Cromófobo de riñón (KICH), Carcinoma renal de células claras (KIRC), Carcinoma renal de células papilares (KIRP), Leucemia mieloide aguda

(LAML), Glioma de bajo grado (LGG), Carcinoma hepatocelular (LIHC), Adenocarcinoma pulmonar (LUAD), Carcinoma de células escamosas de pulmón (LUSC), Mesotelioma (MESO), Cistadenocarcinoma seroso de ovario (OV), Adenocarcinoma pancreático (PAAD), Feocromocitoma y Paraganglioma (PCPG), Adenocarcinoma prostático (PRAD), Adenocarcinoma rectal (READ), Sarcoma (SARC), Melanoma cutáneo (SKCM), Adenocarcinoma de estómago (STAD), Tumor testicular de células germinales (TGCT), Carcinoma tiroideo (THCA), Timoma (THYM), Carcinoma endometrial del cuerpo uterino (UCEC), Carcinosarcoma uterino (UCS), Melanoma uveal (UVM). C) Gráfico de correlación de la expresión de *GATA3-AS1* y *GATA3* en la cohorte de cáncer de mama del *TCGA* depositada en la base de datos *TANRIC*[81] (correlación de Pearson = 0.78, valor *p* = 5.77e-17, n= 942).

Para corroborar la sobreexpresión de los genes *GATA3-AS1* y *GATA3* en cáncer de mama, se realizó el análisis de expresión de *GATA3-AS1* y *GATA3* a partir de datos de RNA-Sec de líneas celulares de cáncer de mama de la base de datos pública *HMS LINCS*[76,77]. Los resultados muestran que tanto *GATA3-AS1* como *GATA3* se expresan en las líneas celulares de cáncer de mama, destacando las mayores frecuencias de expresión de ambos genes en las líneas celulares cancerosas del fenotipo luminal MCF-7 y T-47D (Figura 10), por lo que su expresión podría estar relacionada con el fenotipo luminal.

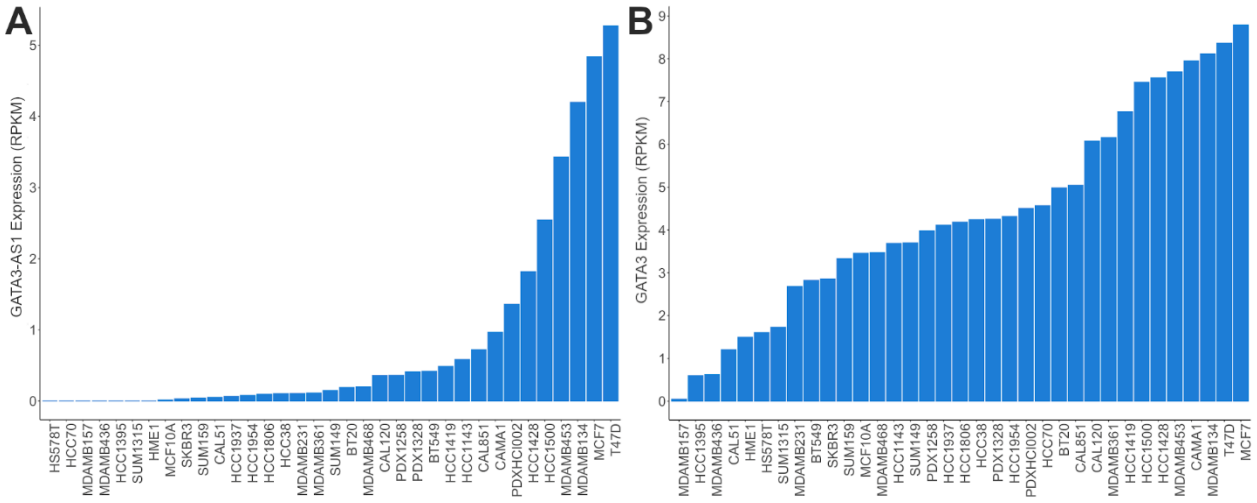


Figura 10. Expresión de *GATA3-AS1* y *GATA3* por RNA-Sec en líneas celulares de cáncer de mama. A) Gráficos de barras que muestran la expresión de *GATA3-AS1* y B) de su gen adyacente *GATA3* en las treinta y cinco líneas celulares de cáncer de mama depositadas en la base de datos *HMS LINCS*[76,77].

Para confirmar que la expresión de *GATA3-AS1* y de *GATA3* está correlacionada de forma positiva en cáncer de mama, se llevó a cabo un análisis de correlación con la información de la expresión de ambos genes obtenida de *HMS LINCS*. Como se observa en la Figura 11A, la expresión de *GATA3-AS1* y del mRNA de *GATA3* presentan una correlación positiva, destacando la línea celular MCF-7, lo cual fue corroborado en un análisis independiente de datos de RNA-Sec de las líneas celulares MCF-10A, MCF-7, BT-474 y MDA-MB-231 obtenidos de la base de datos *GEO Datasets* (Figura 11B). Asimismo, se validaron los resultados obtenidos al analizar la información de la expresión de *GATA3-AS1* y de *GATA3* en las líneas celulares de cáncer de mama de la base de datos *CCLE* (Apéndice A). En conjunto, los resultados de este análisis confirman que existe una relación positiva en la expresión de *GATA3-AS1* y *GATA3* en líneas celulares de cáncer de mama y en pacientes con cáncer de mama, tal como lo indica el primer postulado del mecanismo molecular de regulación en *cis*, que establece que la expresión del lncRNA y su gen adyacente deben estar relacionadas entre sí. Adicionalmente, la

correlación positiva en la expresión de ambos genes sugiere que la expresión de *GATA3* podría ser regulada de manera positiva por *GATA3-AS1*.

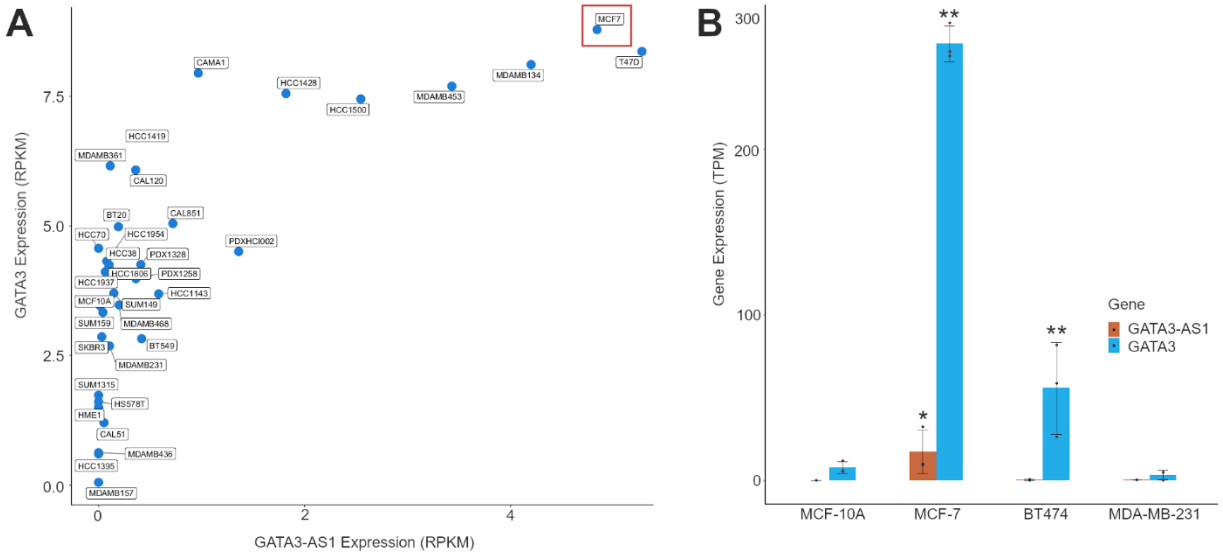


Figura 11. La expresión de *GATA3-AS1* correlaciona con la expresión de su gen adyacente *GATA3* en líneas celulares de cáncer de mama. A) Gráfico de correlación de la expresión de *GATA3-AS1* y *GATA3* en las líneas celulares de cáncer de mama depositadas en la base de datos *HMS LINCS*[76,77]. Particularmente, se observa la correlación positiva en la línea celular MCF-7 (cuadro rojo) (correlación de Pearson = 0.74, valor *p* = 3.2 e-07). B) Gráfico de barras que muestra la expresión de *GATA3* y *GATA3-AS1* en las líneas celulares MCF-10A, MCF-7, BT474 y MDA-MB-231 proveniente de *GEO Datasets,* en el cual se observa que la línea celular neoplásica MCF-7 sobreexpresa tanto a *GATA3-AS1* como a *GATA3* respecto a la línea celular transformada MCF-10A (ANOVA seguido de la prueba Tukey, **valor *p* < 0.01, * valor *p* < 0.05). Asimismo, se observa que en todas las líneas celulares analizadas la expresión de *GATA3-AS1* correlaciona con la expresión de su gen adyacente *GATA3* (correlación de Pearson = 0.8, valor *p* =0.001).

En resumen, el análisis de correlación de la expresión de *GATA3-AS1* y *GATA3* a partir de datos de secuenciación de RNA muestran que existe correlación positiva entre la

expresión de ambos genes en cáncer de mama, así como en las líneas celulares de cáncer de mama, por lo que cumple con el primer postulado de la regulación en *cis*, en el que se menciona que la expresión del lncRNA, en este caso *GATA3-AS1,* debe estar relacionada con la expresión de su gen adyacente *GATA3*.

Para determinar la probable relación funcional entre la expresión de *GATA3-AS1* y *GATA3*, se investigaron las vías de señalización intracelulares a las cuales se encuentra asociada la expresión de ambos genes por medio del uso de las plataformas *FARNA*[82,83] y *STRING*[84]. Como se muestra en la Figura 12A, el análisis de coexpresión de *GATA3-AS1* y los factores transcripcionales en cáncer de mama indica que *GATA3-AS1* está asociado a vías de señalización intracelulares involucradas en la proliferación de células endoteliales y la regulación negativa de la apoptosis, características principales del cáncer, por lo que el *GATA3-AS1* podría ser un regulador de los procesos neoplásicos en cáncer de mama. Por otro lado, *GATA3* está relacionado a vías de señalización intracelulares para el desarrollo de la glándula mamaria y también a la regulación negativa de la apoptosis (Figura 12B). Por lo tanto, la expresión de ambos genes podría estar relacionada a vías de señalización intracelulares relacionadas con el fenotipo neoplásico en cáncer de mama, como la apoptosis.

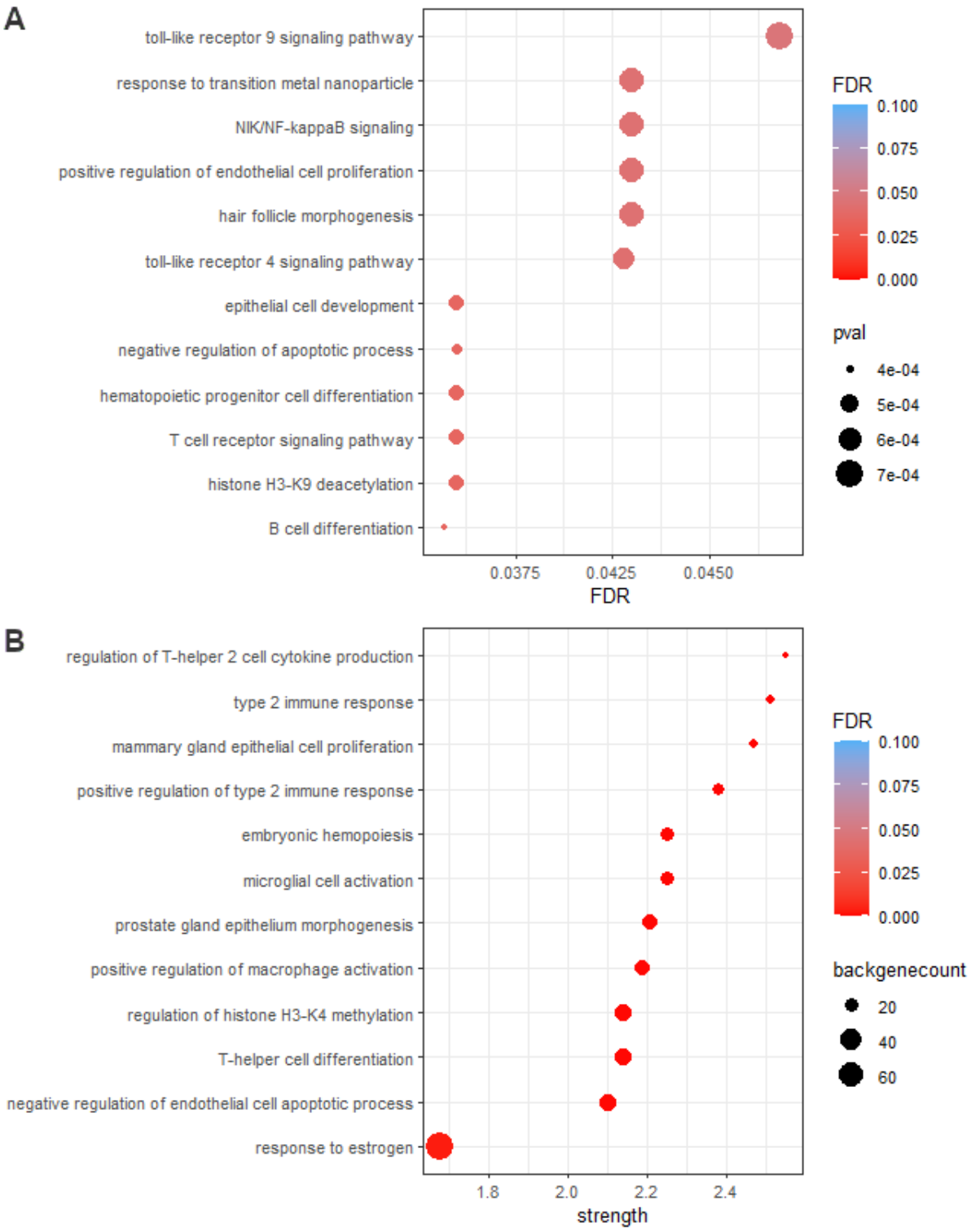


Figura 12. El lincRNA *GATA3-AS1* se asocia a vías de señalización intracelulares relacionadas al desarrollo de cáncer de mama. A) De acuerdo con la predicción de funciones

por correlación de la herramienta FARNA[83], la expresión del lincRNA *GATA3-AS1* se relaciona con las vías de señalización intracelulares enriquecidas en neoplasias mamarias. Asimismo, de acuerdo con el análisis de ontología de genes de la base de datos STRING[84], la expresión del factor transcripcional *GATA3* se relaciona con vías de señalización como la proliferación de células endoteliales en la glándula mamaria (B). En cuadros rojos se resaltan las vías de señalización a las que se relacionan de manera independiente tanto *GATA3-AS1* como *GATA3*. En cuadros azules, se resaltan las vías de señalización que son comunes a *GATA3-AS1* y a *GATA3*.

Debido a que *GATA3-AS1* y *GATA3* participan en vías de señalización intracelulares comunes en el desarrollo de tejido mamario neoplásico, la expresión de ambos genes podría estar regulada por factores transcripcionales que también se relacionan con estas vías de señalización. Para determinar si la expresión de estos genes está regulada por factores transcripcionales en común o por factores epigenéticos como las modificaciones post-traduccionales de histonas, se construyó un mapa de cromatina del locus *GATA3-AS1*/*GATA3* con la información de enriquecimiento de factores transcripcionales y de modificaciones post-traduccionales de histonas para la línea celular MCF-7. Este análisis identificó el enriquecimiento del factor transcripcional MYC en la región promotora y de la RNA Pol 2A fosforilada en la serina 2 (RNAPOL2A-Ser2), lo cual sugiere la presencia de la maquinaria transcripcional en el promotor. Además, el enriquecimiento de las modificaciones post-traduccionales de histonas asociadas con la transcripción activa de genes H3K4me1, H3K4me3 y H3K9ac en la región promotora, que coinciden con las regiones de apertura de la cromatina identificadas mediante el análisis por DNasaI y ATAC-Sec sugieren que el promotor compartido por estos genes se encuentra activo. Asimismo, se identificó que la H3K27ac y la H3K36me3 están enriquecidas en el cuerpo de ambos genes (Figura 13), por lo cual, los resultados de este análisis en conjunto sugieren que ambos genes son unidades transcripcionalmente activas en la línea celular MCF-7.

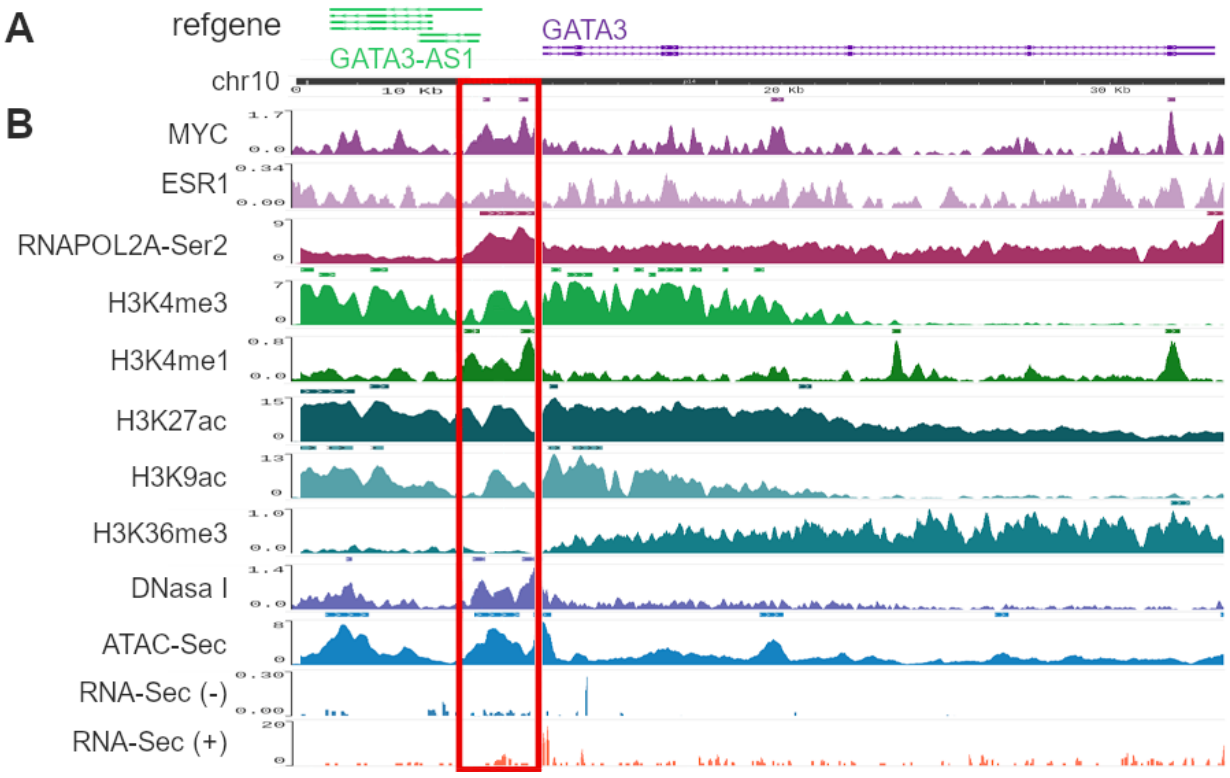


Figura 13. Mapa de cromatina del locus *GATA3-AS1*/*GATA3* en la línea celular MCF-7. Mapa de cromatina que muestra A) la localización genómica del locus *GATA3-AS1*/*GATA3*, la estructura de los transcritos de *GATA3-AS1* y *GATA3*, así como B) los histogramas de resultados de experimentos de ChIP-Sec en la línea celular neoplásica MCF-7, que incluyen ChIP-Sec para: MYC, ESR1, RNAPOL2A (P-Ser2), H3K4me3, H3K4me1, H3K27ac, H3K9ac y H3K36me3, además de análisis de DNasa I, ATAC-Sec y RNA-Sec en las cadenas sentido (+) y antisentido (-). La región promotora de 1 Kb se representa en el cuadro rojo. Los picos de enriquecimiento estadísticamente significativo se indican con cuadros superiores en cada histograma de frecuencias.

Para confirmar este resultado, un análisis adicional para la línea celular MCF-7 mostró que de este locus se obtienen transcritos individuales con la modificación Cap-5' y que en esta línea celular se expresan ambos genes (validado por RNA-Sec) (Apéndice B, Figura suplementaria 3). Igualmente, la comparación entre los mapas de cromatina de las líneas celulares MCF-7 y MCF-10A mostró que el enriquecimiento de las modificaciones post-traduccionales de histonas relacionadas con la activación

transcripcional para este locus es mayor en la línea celular MCF-7 comparado con la línea celular MCF-10A, que no es neoplásica (Apéndice B, Figura suplementaria 4), lo que sugiere que la regulación de su expresión podría estar siendo modulada por factores epigenéticos y factores transcripcionales.

Para determinar si los factores transcripcionales y las modificaciones post-traduccionales de histonas enriquecidas en esa región genómica regulan la expresión de *GATA3-AS1* y *GATA3*, se realizó la búsqueda de información experimental de datos de ChIP-Sec para la línea celular MCF-7 en la región de 1 Kb comprendida entre el TSS de *GATA3-AS1* y el TSS de *GATA3* en la plataforma de *CistromeDB*[86]. El análisis del potencial regulatorio sugiere que el enriquecimiento de las modificaciones post-traduccionales de histonas como la H3K4me1, la H3K4me3 y la H3K27ac, que están relacionadas con la activación de la transcripción, podrían regular la expresión de *GATA3-AS1* y *GATA3*, así como el enriquecimiento de los factores transcripcionales WDR5 y MYC, que se asocian a la regulación epigenética de la transcripción de los genes. Finalmente, el enriquecimiento de los receptores de andrógenos (AR) y de estrógenos (ESR1), que participan en la regulación de las vías de señalización intracelulares de respuesta hormonal en tejido mamario, también podrían regular la expresión de *GATA3-AS1* y *GATA3* en la línea celular MCF-7 (Figura 14). En resumen, la expresión del locus *GATA3-AS1*/*GATA3* podría estar regulada por la presencia de modificaciones post-traduccionales de histonas de activación transcripcional, como la H3K4me3 y por los factores de transcripción asociados a respuesta hormonal, como ESR1 en cáncer de mama.

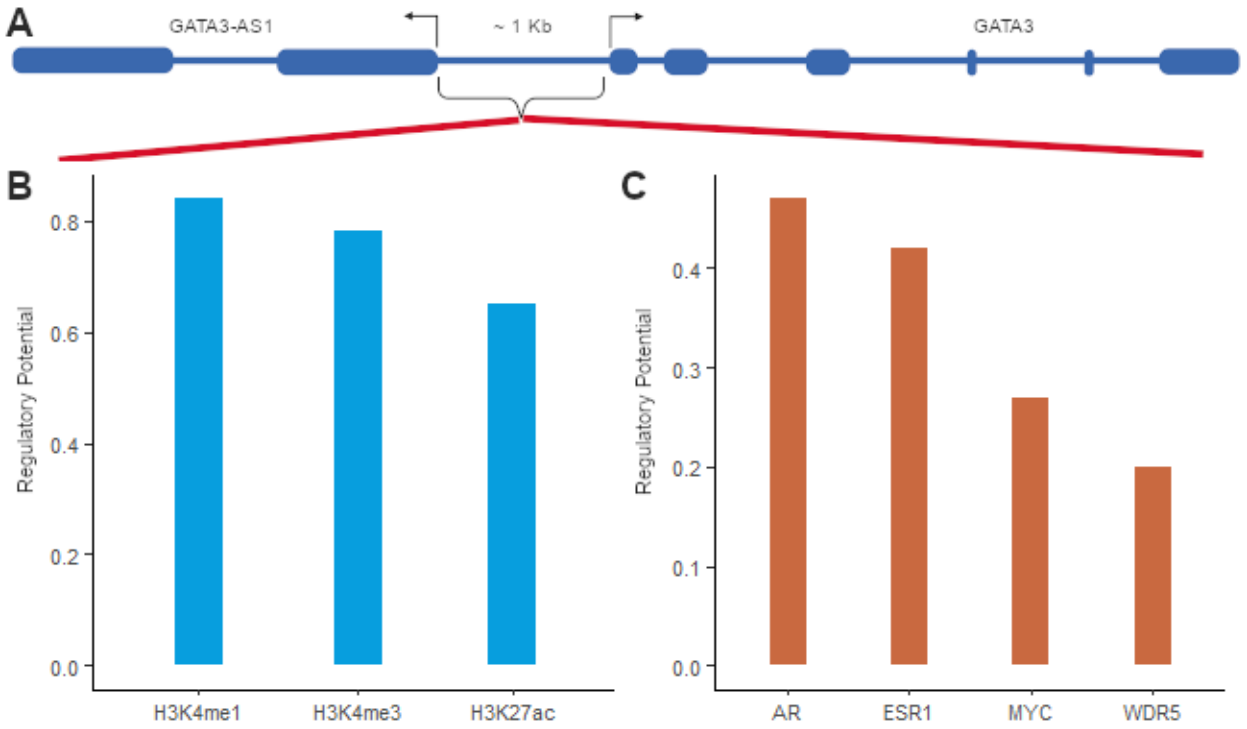


Figura 14. *GATA3-AS1* y *GATA3* son regulados por mecanismos de respuesta hormonal y factores epigenéticos. A) Esquema de la secuencia de 1 Kb de longitud entre los TSSs de *GATA3-AS1* y *GATA3* que fue analizada con la herramienta *Toolkit* de la plataforma *Cistrome DB*[86,87]. B) Gráfico de barras que representa el potencial regulatorio (eje de las ordenadas) de las modificaciones post-traduccionales de histonas H3K4me3, H3K4me1 y la H3K27ac sobre la expresión de los genes en el locus *GATA3-AS1/GATA3* en la línea celular MCF-7. C) Gráfico de barras que representa el potencial regulatorio (eje de las ordenadas) de los factores transcripcionales MYC, AR, ESR1 y de la proteína WDR5 sobre la expresión de los genes en el locus *GATA3-AS1/GATA3* en la línea celular MCF-7.

En conjunto, la caracterización bioinformática del locus *GATA3-AS1*/*GATA3* sugiere que están relacionados entre ellos en cuanto a su expresión, ya que su expresión correlaciona de manera positiva, y en cuanto a sus funciones, debido a que están asociados con vías de señalización intracelulares involucradas en el desarrollo de cáncer de mama, como la regulación negativa de la apoptosis. De igual forma, los genes *GATA3-*

AS1 y *GATA3* son unidades transcripcionalmente activas, y su expresión en cáncer de mama podría ser regulada por factores epigenéticos, como la modificación post-traduccional H3K4me3 y por factores transcripcionales como ESR1, que interactúan particularmente con sus regiones promotoras, por lo cual es importante determinar la regulación del promotor de *GATA3-AS1* sobre la expresión de este gen.

11.2 Caracterización in silico del lincRNA GATA3-AS1

11.2.1 Caracterización in silico del promotor de GATA3-AS1

Para identificar a los factores transcripcionales que interactúan con el promotor de *GATA3-AS1*, se llevó a cabo un análisis de motivos de unión de factores transcripcionales y de proteínas de unión a DNA por medio de las herramientas de la paquetería *MEME SUITE*[89]. El análisis mostró que la secuencia promotora de *GATA3-AS1* tiene motivos de unión a factores transcripcionales relacionados a la respuesta hormonal mediada por el receptor de estrógenos, como ESRRG, ESR2 y para el receptor de andrógenos AR (Figura 15). Además, se encontró la presencia de un motivo de unión para el factor transcripcional LHX3, que está relacionado con el desarrollo neuronal, que es un proceso celular que se ha identificado como enriquecido en cáncer de mama[123] (Apéndice C).

Para validar la presencia de los estos factores transcripcionales en el promotor de *GATA3-AS1*, se analizaron datos públicos de ChIP-Sec de la línea celular MCF-7. Sin embargo, no se encontró ningún experimento en esta línea celular para los factores transcripcionales ESRRG, ESR2, AR o LHX3. A pesar de esto, la información disponible permitió identificar el enriquecimiento del factor transcripcional ESR1 en la región genómica donde se identificó el motivo de reconocimiento para la proteína ESR2 (Figura 15B). Como se puede observar en la Figura 15C, los motivos de reconocimiento de DNA de ambas proteínas son similares en cuanto a su secuencia, por lo que es probable que

exista ese elemento de respuesta a estrógeno en esta región promotora. Por lo tanto, la expresión de *GATA3-AS1* podría estar regulada por factores de respuesta hormonal, como el receptor de estrógenos, que participan en el desarrollo de tejido neoplásico mamario.

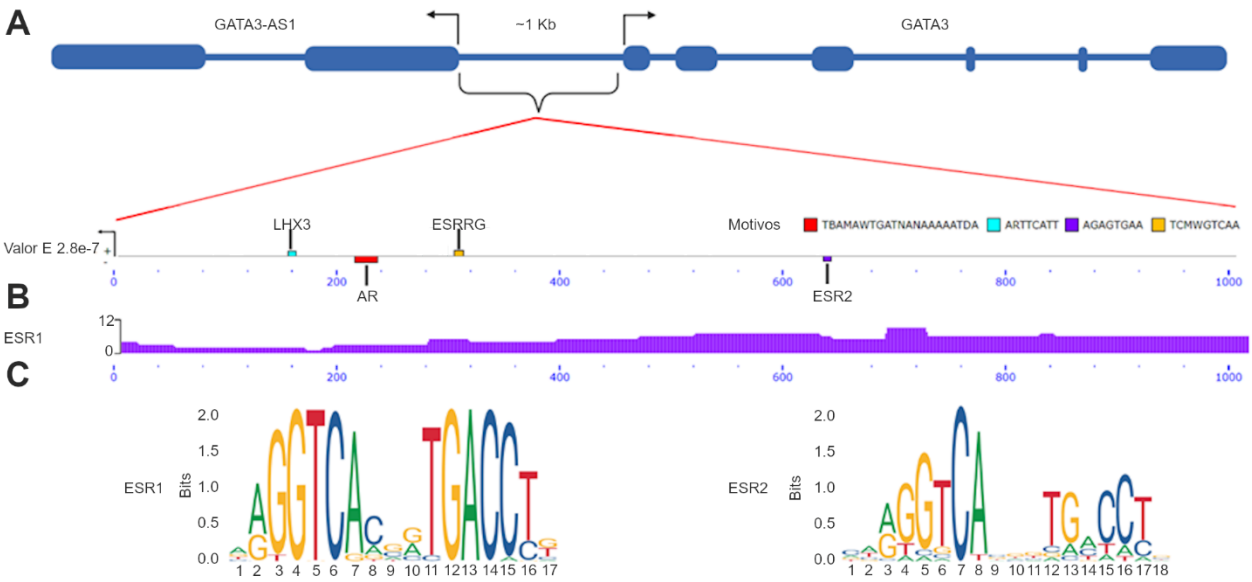


Figura 15. La secuencia promotora de *GATA3-AS1* presenta motivos de unión praa factores transcripcionales relacionados con proliferación endotelial. A) Esquema de la secuencia de la región promotora analizada a 1,024 pb río arriba del TSS de *GATA3*-AS1 (probable sitio de inicio de la transcripción). En colores se presentan los distintos motivos identificados en la cadena sentido (+) y la antisentido (-). A la izquierda se encuentra el valor *e* asociado a la identificación de motivos de unión. B) El histograma de enriquecimiento de la proteína ESR1 por ChIP-Sec muestra la presencia de este factor transcripcional sobre la región promotora de *GATA3-AS1* en la línea celular MCF-7. C) Se muestran los logos correspondientes a los motivos de reconocimiento de DNA de los factores transcripcionales ESR1 (izquierda) y de ESR2 (derecha).

Para determinar si los factores transcripcionales con motivos de unión dentro de la secuencia promotora de *GATA3-AS1* se asocian con procesos de desarrollo neoplásico

en tejido mamario se realizó un análisis funcional de los factores transcripcionales ESSRG, ESR2, ESR1, AR y LHX3 con la herramienta *STRING*[84], y se encontró que éstos participan en vías de señalización intracelulares como la respuesta a hormonas esteroides y el desarrollo de la médula de la espina dorsal, que están enriquecidas en cáncer de mama (Apéndice C, Figuras suplementarias 5 y 6), lo que sugiere que *GATA3-AS1* podría estar relacionado con el desarrollo de cáncer de mama.

11.2.2 Caracterización in silico de las funciones biológicas de GATA3-AS1

Para determinar las funciones biológicas de *GATA3-AS1*, se realizó un análisis de motivos de unión a proteínas dentro de la secuencia de RNA de la isoforma canónica *GATA3-AS1-201*, con el objetivo de identificar a las proteínas que podrían interactuar físicamente con *GATA3-AS1*. Los resultados indican que *GATA3-AS1* tiene en su secuencia de RNA motivos de unión para CPEB4, HuR, PTBP1, RBM4 y LIN28A, que son proteínas de unión a RNA (Figura 16). Estas proteínas están relacionadas con vías de señalización intracelulares involucradas en la diferenciación, el metabolismo celular y la regulación de la expresión de genes, que están enriquecidas en el desarrollo de cáncer de mama, lo cual sugiere que *GATA3-AS1* también podría estar asociado al desarrollo de cáncer de mama mediante su interacción con estas proteínas y la regulación de las vías de señalización intracelulares en las que participan.

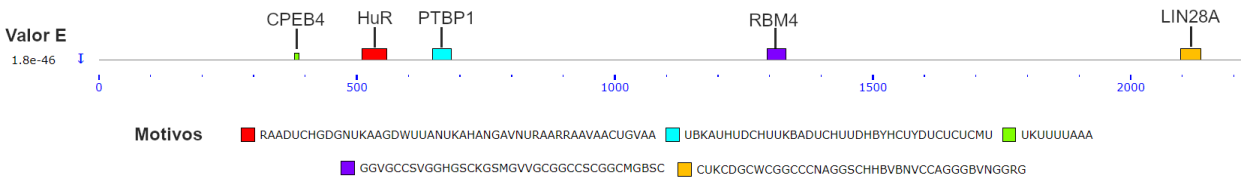


Figura 16. *GATA3-AS1* contiene motivos de unión a proteínas relacionadas con la diferenciación y el metabolismo celular. Esquema de la secuencia de RNA de la isoforma *GATA3-AS1-201* en la cual se muestran mediante bloques de colores los motivos de unión a

proteínas identificados en el análisis *in silico*. Los diferentes motivos de unión identificados están representados por los colores que se indican en la leyenda. ‘

Asimismo, se realizó un análisis de predicción de interacciones RNA-Proteína y RNA-RNA de *GATA3-AS1* con la herramienta *catRapid*[94] y se comparó con la información contenida en la plataforma *RNAInter*[95], que incluye validaciones experimentales de las interacciones. Los análisis coincidieron en que el lincRNA *GATA3-AS1* podría interactuar con la proteína CTCF, el receptor de estrógenos ESR2 y la proteína LEF1 (Figura 17, Apéndice D), que están relacionadas con el crecimiento celular, la respuesta a hormonas esteroideas y a cáncer de mama, lo que coincide con las funciones que se han descrito en los análisis anteriores.

A

Proteína	Poder Discriminatorio	Fuerza de interacción	Puntaje de interacción	Función	Fuente
CTCF	96	99	0.5831	Regulación transcripcional	ChIP-Seq (evidencia baja)
ESR2	97	100	0.5117	Regulación positiva de la transcripción	ChIP-Seq (evidencia baja)
LEF1	81	99	0.5117	Activador de la transcripción en presencia de EP300	ChIP-Seq (evidencia baja)

B

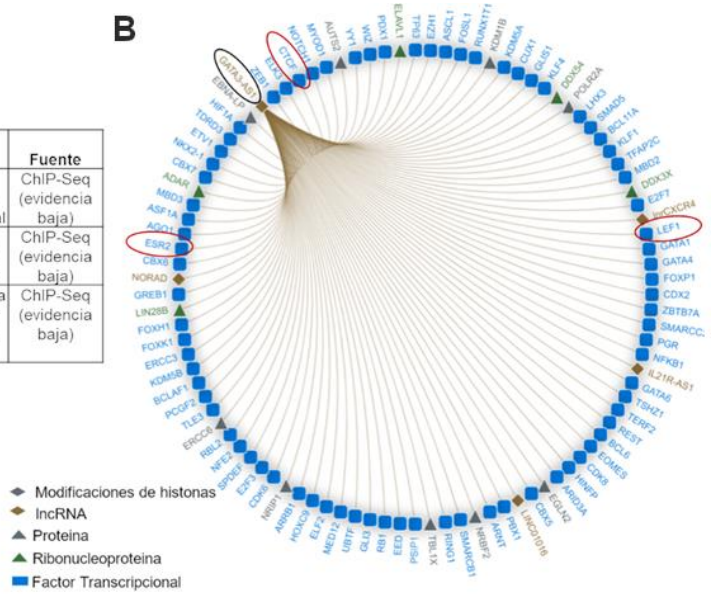


Figura 17. *GATA3-AS1* podría interactuar con factores transcripcionales que activan la transcripción. A) El análisis de predicción de interacciones de *GATA3-AS1* realizado con la herramienta *catRAPID*[97,98] coincide en al menos 3 factores transcripcionales con la información experimental contenida en la base de datos *RNAInter*[95], y que están asociados principalmente a la regulación positiva de la transcripción. En el gráfico de círculo (B) obtenido de la Plataforma

RNAInter se muestran las cincuenta interacciones principales (líneas marrones) de las predichas para *GATA3-AS1* (círculo negro) con modificaciones post-traduccionales de histonas (rombo gris), con otros lncRNAs (rombo marrón), con proteínas (triángulo gris), con complejos ribonucleoproteínicos (triángulo verde) y con factores transcripcionales (rectángulo azul). En círculos rojos se resaltan los tres factores transcripcionales que coincidieron con el análisis predictivo de *catRAPID*.

Posteriormente, se realizó un análisis de interacción de *GATA3-AS1* con proteínas y ncRNAs usando la plataforma *miRNet*[98]. Los resultados sugieren que *GATA3-AS1* participa en una red de interacción mediada por los miRNA miR-30-5p, let-7e-5p y miR-26a-5p, así como los factores transcripcionales GATA3, TP53 y ESR2 (Apéndice D, Figura Suplementaria 7), que se han relacionado con el desarrollo de cáncer de mama al regular vías de señalización como la proliferación celular, la apoptosis y la respuesta intracelular a hormonas esteroideas.

Para validar la sobreexpresión de *GATA3-AS1* y de *GATA3* en líneas celulares de cáncer de mama, se realizó un análisis de expresión diferencial del transcriptoma completo de la línea celular MCF-7. Como resultado de este análisis, se construyó un mapa de calor con los genes diferencialmente expresados en la línea celular MCF-7 respecto a la línea transformada MCF-10A, que muestra que existe un perfil de expresión diferencial de genes codificantes y no codificantes que define el fenotipo luminal neoplásico de la línea celular MCF-7, el cual incluye la sobreexpresión de *GATA3-AS1* y de *GATA3,* así como otros genes relacionados con el desarrollo de cáncer de mama, como *FOXA1* (Figura 18. Apéndice E), lo que corrobora la relación entre la expresión de ambos genes y su relación con el fenotipo neoplásico en la línea celular de cáncer de mama evaluada.

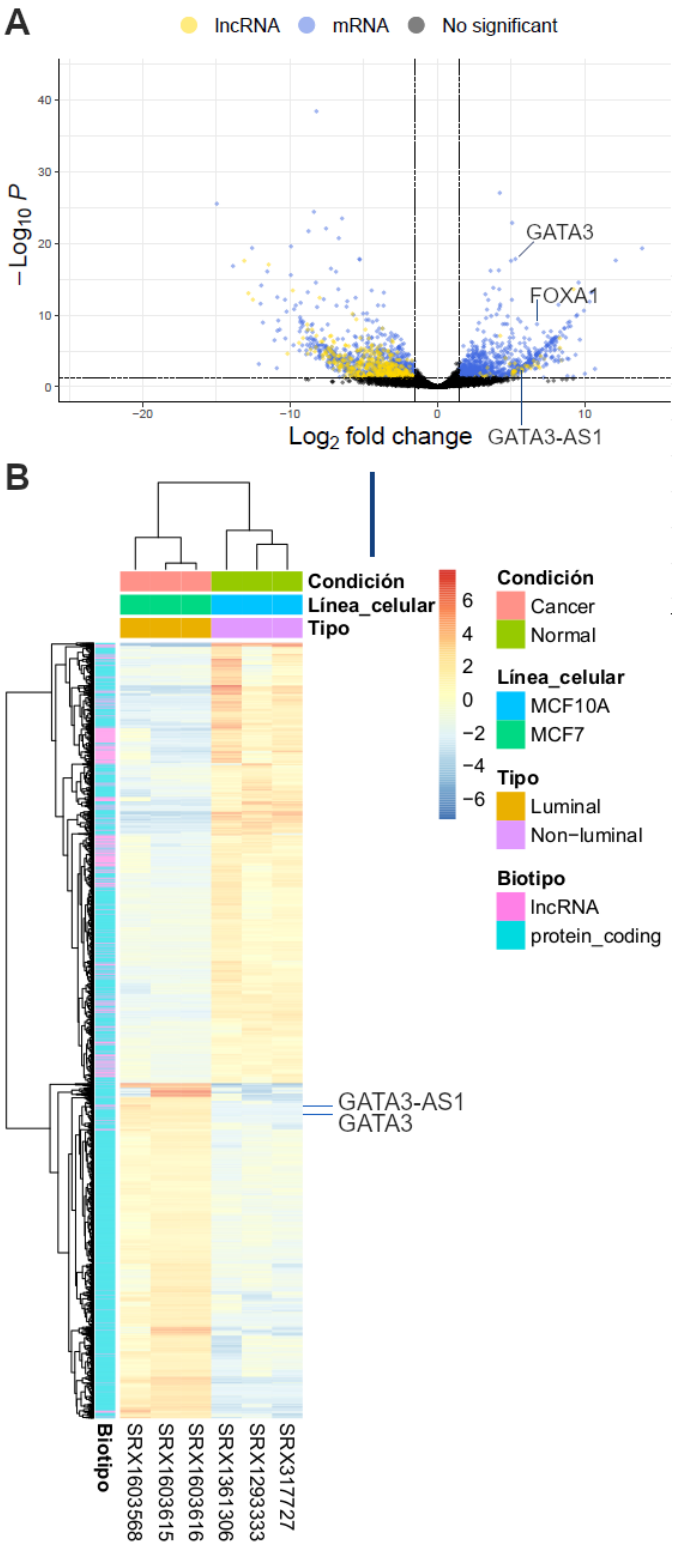


Figura 18. La sobreexpresión de *GATA3-AS1* y *GATA3* se relaciona con el fenotipo neoplásico en la línea celular MCF-7. A) El gráfico de volcán muestra los 2,707 genes

diferencialmente expresados estadísticamente significativos (panel superior izquierdo y derecho, valor *p* ajustado < 0.05) en la línea celular MCF-7, de los cuales 1,538 están subexpresados (mRNA= 1007 en azul, lncRNA= 531 en amarillo, panel superior izquierdo), mientras que 1,169 están sobreexpresados (mRNA = 1121 en azul, lncRNA = 48 en amarillo, panel superior derecho). De la misma manera, se resalta que *GATA3-AS1* y *GATA3* se sobreexpresan en la línea celular neoplásica MCF-7 (cuadrante superior derecho). En puntos negros se muestran los genes que no se encuentran diferencialmente expresados y que tampoco son estadísticamente significativos en este análisis (valor *p* ajustado > 0.05). B) El mapa de calor muestra los perfiles de expresión de los 2,707 genes diferencialmente expresados en la línea celular MCF-7 (azul) respecto a la línea celular transformada MCF-10A (rosa). Cada celda del gráfico muestra la expresión de un gen individual, así como cada columna muestre el perfil de expresión de genes de cada muestre analizada. Asimismo, se puede observar que *GATA3-AS1* y *GATA3* se sobreexpresan en la línea celular neoplásica MCF-7. La intensidad del color en cada celda indica si se encuentra sobre o subexpresado de acuerdo con lo indicado en la leyenda del gráfico.

Asimismo, con el objetivo de validar la participación de *GATA3-AS1* y *GATA3* en las vías de señalización intracelulares que se asocian al desarrollo de cáncer de mama, se realizó un análisis de enriquecimiento de conjuntos de genes en la línea celular MCF-7. Dentro de los resultados obtenidos, se observó que en ocho conjuntos de genes se encuentran *GATA3-AS1, GATA3* y *ESR1*, los cuales están relacionadas con el fenotipo luminal de cáncer de mama, la resistencia a terapia endócrina y a la vía de señalización de *ESR1* (Figura 19), lo que demuestra que *GATA3-AS1* y *GATA3* se relacionan funcionalmente en el contexto de cáncer de mama, validando los resultados obtenidos de los análisis *in silico* de la caracterización funcional de *GATA3-AS1*. Estas relaciones funcionales ya habían sido previamente reportadas para *GATA3* y *ESR*[124,125], mas no para *GATA3-AS1*, lo que contribuye al entendimiento de la función de este lincRNA en las neoplasias mamarias.

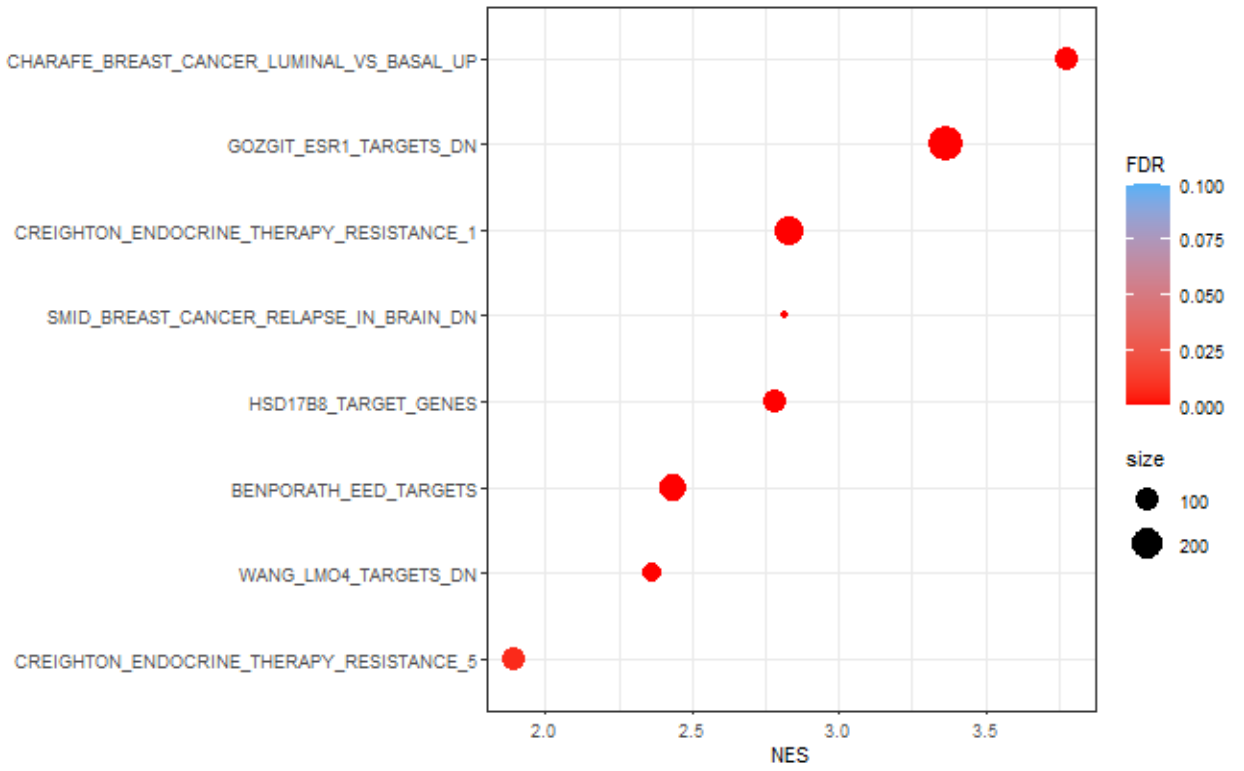


Figura 19. *GATA3-AS1* y *GATA3* participan en vías de señalización intracelulares involucradas en el desarrollo de cáncer de mama. El análisis de enriquecimiento de conjuntos de genes en la línea celular MCF-7 (respecto a la línea control MCF-10A) muestra las cincuenta vías de señalización intracelulares en las que coinciden *GATA3* y *ESR1*, de las cuales al menos en siete coinciden además con el lincRNA *GATA3-AS1* (cuadros rojos), entre los que se encuentra la recurrencia a cerebro (cuadro azul). NES: *Normalized enrichment score.*

En resumen, los resultados sugieren que *GATA3-AS1* está regulado en su región promotora por factores transcripcionales que participan en la respuesta hormonal y en el desarrollo de cáncer de mama. Asimismo, *GATA3-AS1* podría estar participando activamente en las vías de señalización intracelulares como la respuesta a hormonas esteroideas y la regulación de la apoptosis, al interactuar con proteínas como GATA3, ESR1, así como con los miRNAs miR-30-5p y let-7e-5p que también regulan estas vías de señalización en el desarrollo de cáncer de mama, lo cual fue validado con un análisis de expresión diferencial con datos de secuenciación por RNA-Sec en la línea celular MCF-7, en el que además se corroboró que *GATA3-AS1* y *GATA3* participan de manera

conjunta en las vías de señalización intracelulares relacionadas con la respuesta a estrógenos y la resistencia a terapia endócrina, aunque no ha sido descrito en la literatura científica cuál es el mecanismo molecular mediante el cual se relacionan estos dos genes. Debido a la cercanía de ambos genes y a la correlación positiva que presenta la expresión de sus transcritos, es probable que *GATA3-AS1* regule la expresión de su gen adyacente *GATA3* en *cis* tal como ocurre en linfocitos Th2[64]. Sin embargo, esto no ha sido descrito en cáncer de mama ni existe evidencia científica de ello, por lo que es necesario determinar experimentalmente si este mecanismo molecular de regulación es el que relaciona a *GATA3-AS1* y a su gen adyacente *GATA3*.

11.3 Caracterización de la expresión de GATA3-AS1 y GATA3 en la línea celular MCF-7

Con el objetivo de determinar si *GATA3-AS1* regula a su gen adyacente *GATA3* en la línea celular MCF-7, se validó la expresión de ambos genes por RT-qPCR y se corroboró que tanto *GATA3-AS1* como *GATA3* se expresan en las líneas celulares MCF-10A, MCF-7, BT-474 y MDA-MB-231, siendo consistente con los resultados de RNA-Sec que la línea celular MCF-7 sobreexpresa a *GATA3-AS1* y a *GATA3* (Figura 20A), y la correlación entre la expresión de ambos genes es positiva (Figura 20B), por lo que se eligió esta línea celular como modelo biológico de estudio para los experimentos subsecuentes.

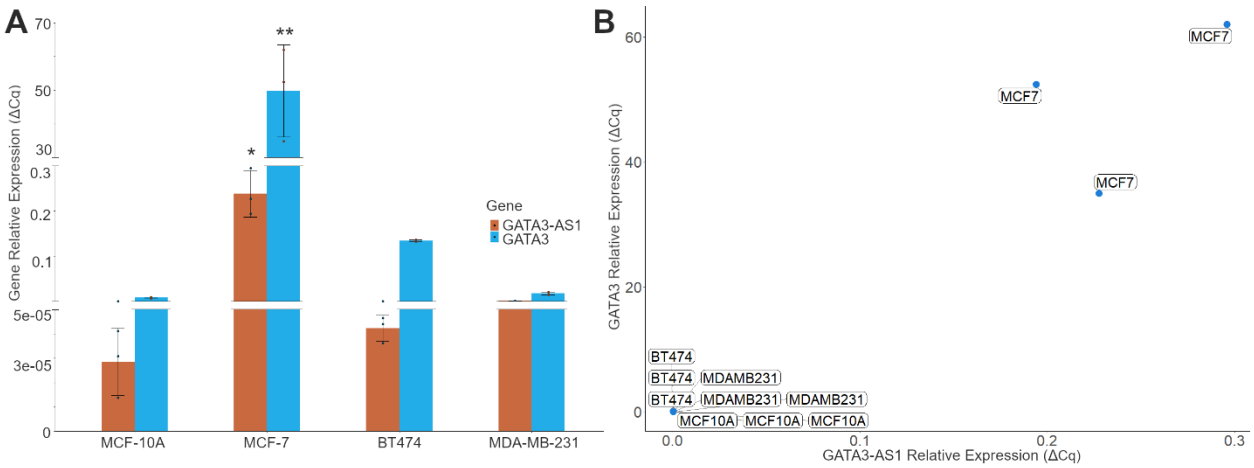


Figura 20. La línea celular MCF-7 sobreexpresa a *GATA3-AS1* y a *GATA3*. A) Gráfico de barras de la expresión relativa de *GATA3-AS1* (azul) y de *GATA3* (naranja) en líneas celulares de cáncer de mama por RT-qPCR, mediante el método de análisis por ΔCq (ANOVA seguida de la prueba Tukey, **valor $p < 0.01$, *valor $p < 0.05$). B) Gráfico de correlación de la expresión relativa de *GATA3-AS1* y *GATA3* en líneas celulares de cáncer de mama por RT-qPCR (ΔCq) (coeficiente de correlación de Spearman = 0.7, valor p =0.01). Expresión relativa al gen constitutivo *RPS28*.

Por otro lado, para caracterizar experimentalmente un lincRNA como *GATA3-AS1* es importante determinar su localización celular, debido a que ésta se relaciona con las funciones que los lincRNAs llevan a cabo, como por ejemplo la regulación de la transcripción. Mediante un experimento de fraccionamiento celular se determinó que *GATA3-AS1* se encuentra acumulado en el núcleo, particularmente asociado a cromatina, como *XIST1* y *NEAT1*, que se utilizaron como controles positivos de localización en cromatina. Asimismo, se determinó que el control positivo de localización nuclear, *MALAT1*, se encuentra enriquecido en la fracción nuclear, mientras que los controles positivos *H19* y *RPS28* están en el citoplasma (Figura 21A), lo cual fue validado con la información disponible en la base de datos lncATLAS[70] (Figura 21B). En resumen, los resultados indican que *GATA3-AS1* es un lincRNA de localización nuclear, que se

encuentra acumulado principalmente en la cromatina, lo cual podría estar relacionado con la función molecular que lleva a cabo.

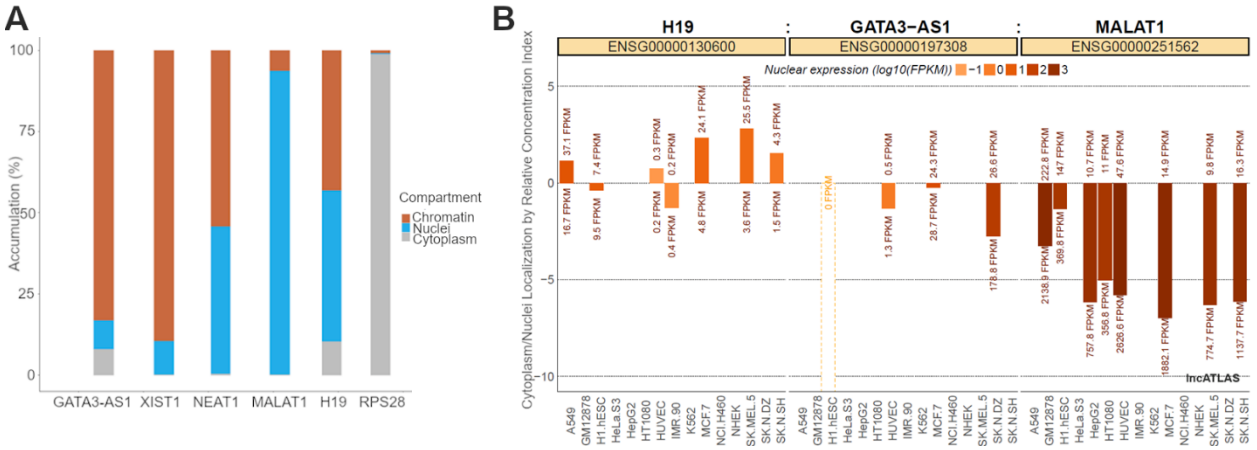


Figura 21. El lincRNA *GATA3-AS1* se encuentra localizado en la cromatina en la línea celular MCF-7. A) Gráfico de barras que muestra la expresión de *GATA3-AS1* y de los genes control en los diferentes compartimentos celulares en células MCF-7 por RT-qPCR (ΔCq). Control de cromatina: *XIST*; controles de núcleo: *NEAT1* y *MALAT1*; controles de citoplasma: *H19* y *RPS28*. Expresión relativa al gen constitutivo *RPS28*. B) Gráfico de barras del análisis del cociente citoplasma/núcleo proveniente de la base de datos *lncATLAS*[70] muestra que la localización de *GATA3-AS1* es principalmente nuclear en la línea celular MCF-7. Para validar los resultados del análisis, se incluyeron como controles a *H19* (de localización citoplasmática) y a *MALAT1* (de localización nuclear).

11.4 Abatimiento de la expresión de GATA3-AS1 mediante ASOs en la línea celular MCF-7

La acumulación en cromatina de *GATA3-AS1* sugiere que la función de este lincRNA podría estar relacionada con la regulación de la transcripción de genes. Adicionalmente, la correlación que existe entre la expresión de *GATA3-AS1* y *GATA3*, sugiere que esta función regulatoria se lleva a cabo sobre la expresión de su gen adyacente *GATA3*. Para

corroborarlo, se diseñó un experimento de abatimiento de la expresión o knock down (KD) de *GATA3-AS1* con el uso de oligonucleótidos antisentido (ASOs) dirigidos hacia *GATA3-AS1*, ya que se trata de un lincRNA nuclear[121], para evaluar si existe algún efecto sobre la expresión del gen codificante adyacente *GATA3*. Para ello, se diseñaron los ASOs sobre la secuencia del transcrito maduro de *GATA3-AS1* como se describe en ***Metodología***. Se seleccionaron 2 secuencias de ASOs (denominados ASO-A y ASO-B) para llevar a cabo el experimento de abatimiento de la expresión de *GATA3-AS1* (Figura 22).

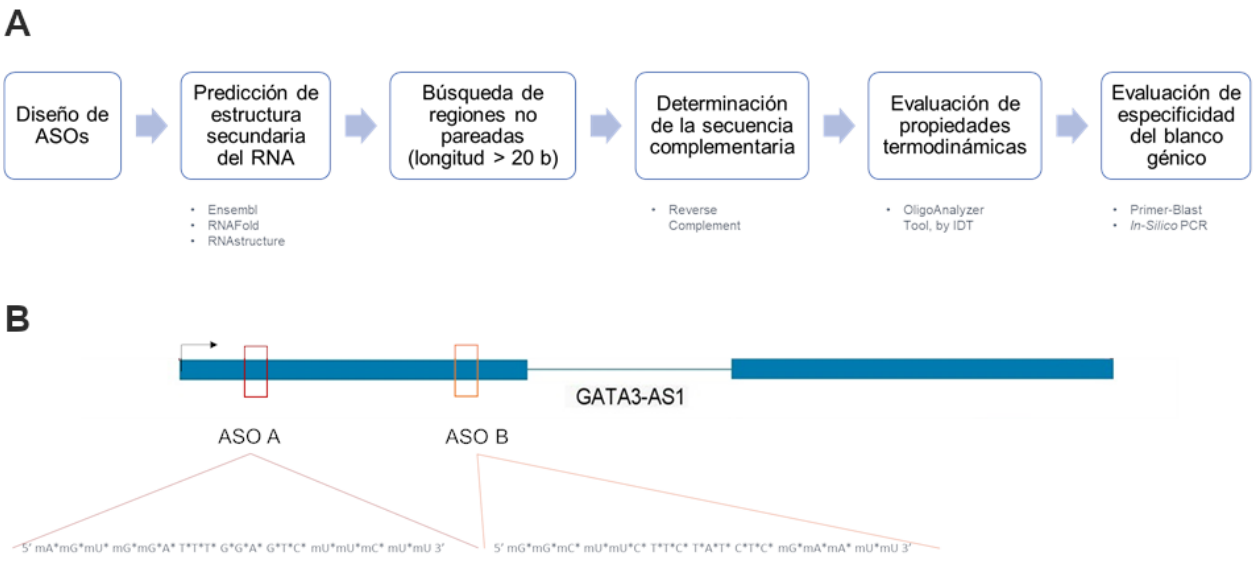


Figura 22. Esquema de diseño de los ASOs utilizados para el experimento de abatimiento de la expresión de *GATA3-AS1* en la línea celular MCF-7. A) Flujo de trabajo para el diseño de ASOs, que consistió en la predicción de la estructura secundaria de *GATA3-AS1* mediante el uso de herramientas bioinformáticas como *RNAFold* y *RNAstructure*, con la cual posteriormente se localizaron estructuras tallo-asa de longitud mayor a 20 bases que servirían como templado para el diseño de los ASOs, lo anterior fue basado en la secuencia complementaria de estas estructuras. Para cada secuencia complementaria, se evaluaron las propiedades termodinámicas con la herramienta *OligoAnalyzer* de *IDT* (https://www.idtdna.com/pages/tools/oligoanalyzer), con el objetivo de seleccionar aquellas con mayor estabilidad y menor probabilidad de dimerización durante las condiciones de la

transfección. Finalmente, se seleccionaron las secuencias más estables y se comprobó su especificidad de complementariedad con la secuencia de RNA de *GATA3-AS1* con las herramientas *Primer-Blast* e *In silico PCR* (ver sección **Metodología**). B) Esquema de la secuencia de RNA de *GATA3-AS1* en el que se muestran las dos secuencias de ASOs que se seleccionaron para el experimento de transfección, el ASOA (rojo) y el ASOB (naranja), que son complementarios al primer exón del transcrito *GATA3-AS1-201*.

Con el objetivo de establecer las condiciones experimentales para el KD de *GATA3-AS1* mediante la transfección con ASOs, se realizó la estandarización del experimento con los ASOs A y B, transfectando cada ASO independientemente o la mezcla de ambos en la línea celular MCF-7 a diferentes concentraciones (50, 100, 150 y 200 nM), utilizando como control de transfección un ASO para el gen *LacZ* (Apéndice F, Figura Suplementaria 9). Además, se realizó simultáneamente la transfección de un ASO estandarizado previamente en el laboratorio para realizar un KD de *MALAT1* en la línea celular MCF-7 (Apéndice F, Figura Suplementaria 10), con la finalidad de validar que los cambios en la expresión del transcriptoma de MCF-7 son consecuencia de la transfección de los ASOA y ASOB (ver sección ***Metodología***). A partir de los resultados obtenidos, se determinó que el ASOA a la concentración de 50 nM es la condición experimental óptima para el abatimiento de la expresión de *GATA3-AS1* en la línea celular MCF-7, por lo que se reprodujo el experimento de transfección en estas condiciones con 6 réplicas biológicas, lo que demostró que los niveles de expresión del mRNA de *GATA3* disminuyen con el abatimiento de la expresión de *GATA3-AS1* (Figura 23).

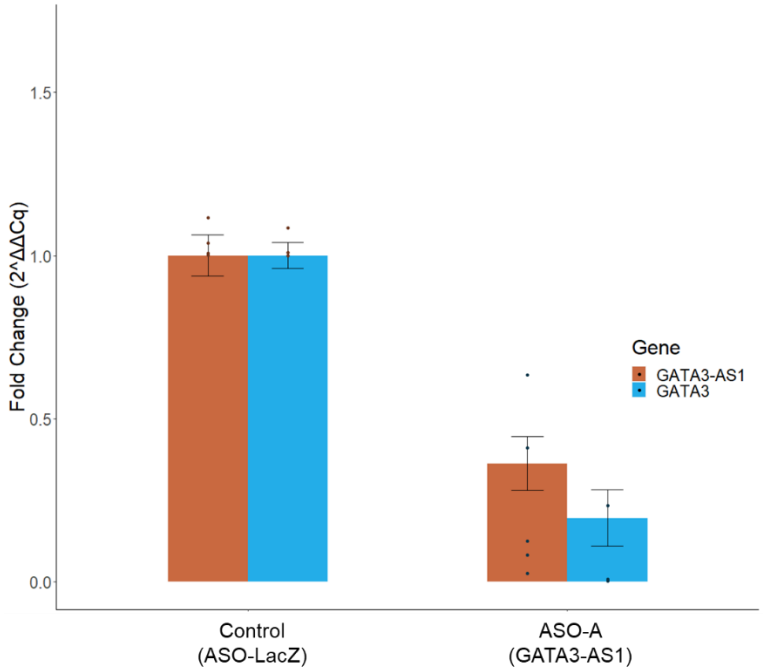


Figura 23. La expresión del gen *GATA3* disminuye con el abatimiento de la expresión del lincRNA *GATA3-AS1* en la línea celular MCF-7. Gráfico de barras que muestra la disminución en la expresión del lincRNA *GATA3-AS1* y su gen adyacente *GATA3* con la transfección del ASOA (50nM) respecto al control transfectado con el ASO-LacZ (n=6, valor $p < 0.001$).

Como se observa en la Figura 23, la expresión de *GATA3-AS1* y de *GATA3* se reduce respecto a la condición control, lo cual indica que esta disminución en la expresión es consecuencia de la transfección con el ASOA y no por el propio proceso de transfección, ya que las células transfectadas con el ASO para *LacZ*, que es un gen que no se encuentra en el genoma humano, presentan niveles mayores de la expresión de ambos genes. Por otro lado, la disminución de la expresión de *GATA3-AS1* y de *GATA3* en las células transfectadas con el ASOA corrobora la correlación positiva que existe entre la expresión de *GATA3-AS1* y *GATA3* (Figuras 9 y 11), ya que al disminuir los niveles de expresión de *GATA3-AS1* también disminuye la expresión la expresión del mRNA de *GATA3*. Este resultado valida que el lincRNA *GATA3-AS1* cumple con el segundo postulado de la regulación en *cis*, que establece que la pérdida de la función del lncRNA tendrá un efecto en la expresión de su gen adyacente. Asimismo, la expresión de *GATA3-*

AS1 se reduce en un 60%, mientras la expresión de *GATA3* se reduce hasta en un 80% durante el KD de *GATA3-AS1*, lo cual sugiere que la activación de la expresión de *GATA3* es dependiente del transcrito de *GATA3-AS1*, ya que la expresión de *GATA3* disminuye hasta 20% más que el propio lincRNA, que es el blanco del ASOA. Estos resultados, aunados a la correlación positiva de la expresión de ambos genes, sugieren que *GATA3-AS1* podría tener la función de activador en la transcripción de su gen adyacente en *cis*.

Finalmente, se analizó cualitativamente el efecto de la transfección del ASOA en el fenotipo de la línea celular MCF-7 mediante el uso de un microscopio de campo claro, para determinar si el abatimiento en la expresión de este lincRNA tenía alguna consecuencia en la morfología celular. Como se observa en la Figura 24 (paneles superiores) las células transfectadas con el ASOA disminuyeron su confluencia aproximadamente al 30% respecto a la condición control (ASOLacZ). Por otra parte, el análisis cualitativo de la morfología celular de la línea celular MCF-7 (Figura 24, páneles inferiores) muestra en el fenotipo silvestre células de morfología epitelial (plana, poligonal, con cilios cortos), agrupadas en monocapa y con contactos entre ellas. Dichas características no se mantienen en la condición de transfección con el ASOA, ya que se observan células en la periferia de la agrupación celular con forma poligonal alargada de tres lados y cilios largos que las conectan entre ellas, la cual es una morfología similar a la de los fibroblastos mamarios, que componen un tipo de tejido distinto al del epitelio mamario. En resumen, el abatimiento de la expresión del lincRNA *GATA3-AS1* induce cambios en la morfología epitelial de la línea celular MCF-7, lo cual sugiere que este lincRNA podría estar relacionado con el fenotipo neoplásico de esta línea celular.

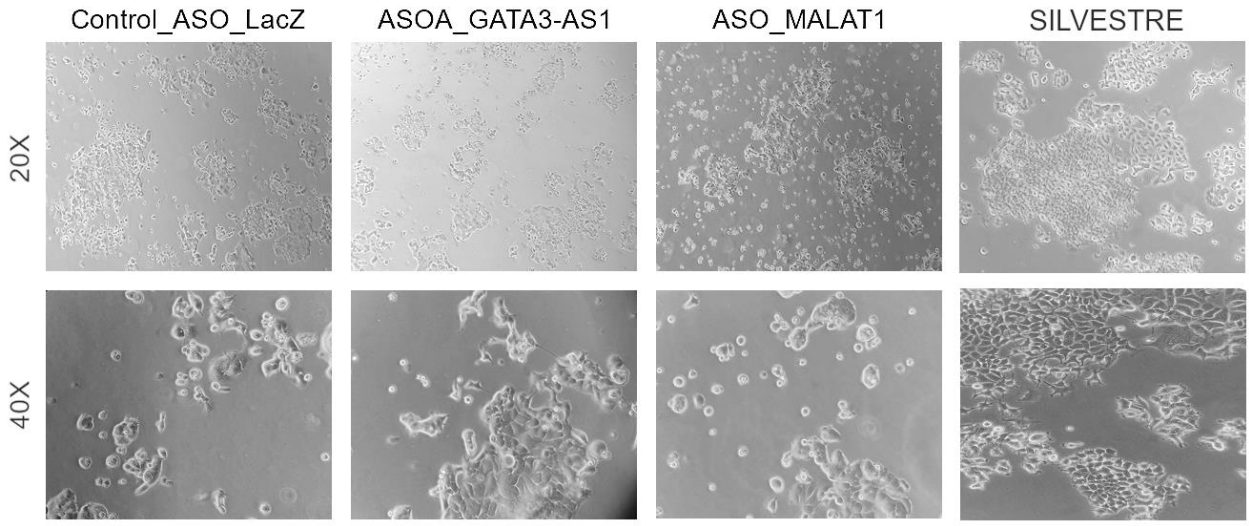


Figura 24. Cambios en la morfología de las células MCF-7 después de la transfección con el ASOA contra *GATA3-AS1*. Los páneles muestran imágenes obtenidas de visualizar una población de células MCF-7 con un microscopio de campo claro, 48 hrs posteriores a la transfección con el ASO-LacZ (panel superior izquierdo, control negativo de la transfección), con el ASOA contra *GATA3-AS1* (panel superior central) o con el ASO-MALAT1-AS2 (panel superior derecho, control positivo de la transfección) usando el objetivo 20X. En los páneles inferiores, se observan los cambios morfológicos en las células después de la transfección con el ASOA (panel inferior central) y con el ASO-MALAT1-AS2 (panel inferior derecho) al compararlos con las células de la condición control ASO-LacZ (panel inferior izquierdo) usando el objetivo 40X. El fenotipo silvestre de la línea celular MCF-7 se muestra en los páneles del extremo derecho (condición inicial del cultivo celular).

En conclusión, los resultados de este trabajo demuestran que el locus integrado por el lincRNA *GATA3-AS1* y su gen codificante adyacente *GATA3* son unidades transcripcionalmente activas, de transcripción divergente y cuya expresión está correlacionada positivamente en cáncer de mama y particularmente en la línea celular MCF-7. La expresión de ambos genes podría ser regulada por factores epigenéticos asociados con la transcripción activa y por factores transcripcionales relacionados con vías de señalización intracelulares de respuesta hormonal, como *ESR1*, de desarrollo del tejido mamario y de desarrollo de tejido neuronal, que en estudios recientes también se han relacionado con el desarrollo del cáncer de mama. Además, el lincRNA *GATA3-*

AS1 y su gen adyacente *GATA3* participan de manera conjunta en estas mismas vías de señalización intracelulares, en las cuales el lincRNA *GATA3-AS1* podría participar de manera directa mediante interacciones físicas con otras biomoléculas que también regulen estas vías de señalización intracelulares. Finalmente, se demostró experimentalmente que el lincRNA *GATA3-AS1* regula la expresión de su gen adyacente *GATA3*, ya que la expresión de ambos genes presenta correlación positiva, cumpliendo con el primer postulado de la regulación en *cis*, así como la expresión de *GATA3* disminuye al abatir los niveles de expresión de *GATA3-AS1*, por lo que cumple con el segundo postulado de la regulación en *cis*. En resumen, los resultados de este trabajo demuestran que el lincRNA *GATA3-AS1* regula la expresión de su gen adyacente *GATA3* en *cis*, lo cual sugiere que regula también las vías de señalización intracelulares involucradas en el desarrollo y mantenimiento de las células tumorales en cáncer de mama.

12.0 DISCUSIÓN

Los lncRNAs son genes no codificantes cuyos productos funcionales son transcritos de naturaleza no-codificante que se encuentran en el genoma en proporción hasta tres veces más que los genes codificantes que dan origen a proteínas[3]. En particular, los lncRNAs se han conservado a lo largo de la evolución de los mamíferos, especialmente entre los primates, por lo cual la preservación de este tipo de secuencias no codificantes puede estar relacionada con la importancia de las funciones intracelulares que desempeñan[2], como es el caso de los lincRNAs divergentes, que representan aproximadamente el 1.7% del total de secuencias genéticas contenidas en el genoma humano[8], y se caracterizan principalmente por su transcripción bidireccional[10].

La evidencia científica sugiere que estos lincRNAs divergentes contribuyen con la regulación de las funciones de sus genes adyacentes, especialmente si éstos son codificantes, lo cual está relacionado con el alto grado de conservación de estos loci[12], como es el caso del locus *GATA3-AS1*/*GATA3*, que se encuentra conservado en mamíferos y en al menos 7 especies de primates. Además, se ha reportado que en los promotores de los loci similares al locus *GATA3-AS1*/*GATA3* se han encontrado elementos de regulación en *cis* que están conservados en mamíferos, como los promotores y los elementos *enhacer*[126], como es el caso de los loci *EVX1-AS1/EVX1*[12] y *MYMLR/MYC*[15], por lo que se ha sugerido que este mecanismo de regulación se ha preservado durante la evolución debido a su importancia biológica [1,13]. Específicamente, el locus compuesto por los genes *GATA3-AS1* y *GATA3* se conserva en especies que desarrollan glándulas mamarias, por lo que la función de ambos genes podría estar relacionada con el desarrollo y mantenimiento del tejido mamario, como ha sido reportado para otros lincRNAs divergentes conservados cuya función principal es la diferenciación celular, por ejemplo *EVX1-AS1*[12].

A pesar de que *GATA3-AS1* y *GATA3* se sobreexpresan en cáncer de mama, los resultados de este trabajo sugieren que *GATA3-AS1* es un gen con expresión tejido- y cáncer-específica, ya que mediante análisis estadísticos se ha asociado a carcinoma hepatocelular[127], a la diferenciación de linfocitos Th[64] y al desarrollo de tumores mamarios[127], lo cual puede estar asociado con el enriquecimiento de la H3K4me2 y la presencia de la H3K9me3 en el promotor de *GATA3-AS1*, ya que su presencia ha sido asociada a la expresión tejido-específica de varios lincRNAs[20]. Por el contrario, *GATA3* se sobreexpresa en casi todas las neoplasias reportadas en la base de datos TCGA, por lo que no muestra una expresión tejido-específica, es decir, los mRNAs como *GATA3* tienden a tener una expresión menos tejido-específica y etapa específica, su expresión es más generalizada en los distintos tipos de neoplasias[41,43,59]. Además, existen trabajos de investigación que asocian la expresión de *GATA3* a la diferenciación de distintos tejidos humanos sanos, por lo que no es un gen con expresión tejido-específica[129]. Estos resultados sugieren que la regulación de *GATA3-AS1* sobre la expresión de *GATA3* en cáncer de mama contribuye a la modulación de su función específica para esta neoplasia, como se ha reportado para el locus *MYMLR/MYC*, que regula la proliferación celular en tumores pulmonares[15], y para el locus *SLC16A1-AS1*/*SLC16A1/MCT1*, que participan en la reprogramación metabólica de células tumorales de vejiga[44], lo que demuestra que los lincRNAs divergentes participan en el desarrollo tumoral mediante la regulación de sus genes adyacentes.

Asimismo, los resultados de este trabajo sugieren que la función regulatoria de *GATA3-AS1* sobre *GATA3* podría ser modulada a su vez por la respuesta hormonal a estrógenos, ya que la expresión de *GATA3-AS1* y *GATA3* presenta una correlación positiva principalmente en las líneas celulares MCF-7 y T47D, que expresan el receptor de estrógenos. Igualmente, el análisis mediante *FIMO* permitió identificar al menos dos motivos de unión a proteínas relacionadas con la respuesta a estrógenos en la región promotora de *GATA3-AS1*, aunque no se encontró información basada en experimentos de ChIP-Sec en las bases de datos públicas *Cistrome* o *WashU Epigenome Browser* que pudieran confirmar su enriquecimiento en la región promotora. Sin embargo, se encontró

información de experimentos de ChIP-Sec que confirman la presencia de ESR1 en el promotor de *GATA3-AS1,* para el cual no se identificó ningún motivo de unión, a pesar de que presenta similitud en su secuencia con el motivo de unión ESR2 (Figura 15C). Esto puede deberse a la estructura de la secuencia del promotor propia de los lincRNAs, ya que se ha reportado que contienen motivos de unión a factores transcripcionales empalmados entre ellos[14], lo cual podría ser la causa de que la herramienta *FIMO* no distinga adecuadamente entre los motivos de unión de ESR1 y ESR2, así como que la presencia de ESR1 no se asocie a un pico de enriquecimiento en la región promotora, como se esperaría para la región promotora de un gen codificante[130]. Finalmente, aunque la región promotora muestra el enriquecimiento del factor transcripcional MYC, tampoco se identificó ningún motivo de unión a MYC o alguna secuencia relacionada, por lo que su presencia en el promotor puede estar mediada por la interacción con una proteína adaptadora, lo cual podría comprobarse con experimentos de ChIP y espectrometría de masas.

De igual manera, el análisis funcional de los factores transcripcionales con motivos de reconocimiento en la región promotora del lincRNA *GATA3-AS1*, así como el análisis de expresión diferencial en la línea celular MCF-7 confirman la relación que existe entre la expresión de *GATA3-AS1* y *ESR1*. No obstante, la relación funcional entre ellos debe ser validada mediante la inhibición de la función del receptor de estrógenos o abatiendo su expresión en la línea celular MCF-7. Por otro lado, se identificó el enriquecimiento de vías de señalización intracelulares relacionadas con el desarrollo de células neuronales, como la especificación de células neuronales en la que participa el gen *LHX3,* que regula la especialización de células motoras neuronales en el sistema nervioso[131], y del cual se ha reportado previamente su asociación con cáncer de mama[132]. Esto último ha sido un área de investigación novedosa en cáncer de mama, ya que se ha demostrado que las vías de desarrollo neuronal, particularmente la vía de especificación de células neuronales, están enriquecidas en los tumores mamarios, así como en las líneas celulares de cáncer de mama del subtipo luminal, y los genes que participan en estas vías de señalización están relacionados principalmente con la proliferación celular[123], la

supervivencia global[125] y la resistencia a tratamiento[66] en cáncer de mama. Sin embargo, no existe evidencia experimental que justifique la relación entre *GATA3-AS1* y las vías de señalización asociadas al desarrollo neuronal, por lo que es necesario desarrollar más investigación al respecto para la caracterización funcional de *GATA3-AS1.*

Asimismo, debido a la cercanía entre los sitios de inicio de la transcripción de *GATA3-AS1* y *GATA3*, así como su transcripción bidireccional y las características de la región promotora, se ha propuesto que *GATA3-AS1* modula en *cis* la expresión de su gen adyacente *GATA3* en linfocitos Th2[64]. Sin embargo, esta hipótesis no ha sido comprobada experimentalmente bajo los postulados de la regulación en *cis* mediada por lncRNAs propuestos por Guttman y Rinn[18], así como tampoco se ha demostrado la relación funcional que existe entre ambos genes en cáncer de mama. Por lo tanto, el objetivo principal de este trabajo consistió en demostrar que la regulación de *GATA3-AS1* sobre su gen adyacente *GATA3* es llevada a cabo en *cis* en cáncer de mama, particularmente en la línea celular MCF-7.

Nuestros resultados experimentales confirman que *GATA3-AS1* cumple con el primer postulado de la regulación en *cis,* que establece que la expresión del lncRNA se relaciona con la expresión de su gen adyacente*,* debido a que la expresión de ambos genes en líneas celulares de cáncer de mama presenta una correlación positiva, así como también cumple con el segundo postulado de la regulación en *cis,* que establece que la pérdida de la función del lncRNA afecta la expresión de su gen adyacente, ya que el abatimiento de la expresión de *GATA3-AS1* tiene como consecuencia la disminución de la expresión del mRNA de su gen adyacente *GATA3*. Adicionalmente, se determinó que *GATA3-AS1* se encuentra principalmente enriquecido en la cromatina, lo cual también sugiere que la actividad regulatoria de este lincRNA es en *cis*, ya que Sun y colaboradores demostraron que los lincRNAs enriquecidos en cromatina regulan la expresión de sus genes adyacentes principalmente mediante este mecanismo[20]. Sin embargo, para demostrar experimentalmente que la regulación que ejerce *GATA3-AS1* sobre *GATA3* es mediante el mecanismo de regulación en *cis*, se debe probar que esta regulación ocurre también de manera alelo-específica[19], tal como establece el tercer postulado de la regulación en

cis[18], por lo que sería necesario llevar a cabo un experimento de edición genética mediante *CRISPR*[133] para eliminar la región genómica del locus de *GATA3-AS1* sólo en uno de los alelos y determinar si la transcripción de *GATA3* ocurre en ambos alelos o sólo en el alelo de fenotipo silvestre.

En resumen, los resultados sugieren que *GATA3-AS1* puede funcionar como un activador de la transcripción de su gen adyacente en *cis*, como se ha observado para otros loci como *MYMLR/MYC*[15] y *SLC16A1-AS1*/*SLC16A1/MCT1*[44]. Sin embargo, es necesario confirmar la interacción del transcrito de *GATA3-AS1* con la región promotora de *GATA3* mediante un experimento de precipitación de la cromatina mediante la purificación de RNA (*ChIRP*, por sus siglas en inglés), además de realizar un experimento de sobreexpresión de *GATA3-AS1* mediante la transfección de un vector de expresión para confirmar que *GATA3-AS1* regula positivamente la expresión de su gen adyacente *GATA3*.

La relación que existe entre la expresión de *GATA3-AS1* y *GATA3* sugiere que el mecanismo de activación también podría llevarse a cabo mediante la interacción con un elemento *enhancer*, esto debido también a la cercanía entre los TSSs de ambos genes y al enriquecimiento de la H3K4me1 y la H3K27ac en la región promotora, pues éstas se asocian con la presencia de elementos *enhancer*[133]. Sin embargo, es necesario corroborar estas predicciones de manera con experimentos de captura conformacional de cromatina, para determinar si existe interacción con algún elemento *enhancer*, así como analizar el enriquecimiento del complejo transcripcional Mediator asociado a *enhancers* en la región promotora[135], y si *GATA3-AS1* interactúa con éste mediante un experimento de inmunoprecipitación de RNA. Asimismo, los niveles de expresión de *GATA3* son hasta dos órdenes de magnitud mayores respecto a *GATA3-AS1* en la línea celular MCF-7, lo que sugiere que se necesitan pocos transcritos de *GATA3-AS1* para activar la expresión de *GATA3*, mientras que el KD de *GATA3-AS1* reduce la expresión de *GATA3* hasta en un 80%, lo cual podría ser indicio de que la activación de la transcripción de *GATA3* depende principalmente de *GATA3-AS1* en células MCF-7. No

obstante, es necesario corroborar la estequiometría en este mecanismo de activación de la transcripción mediante la hibridación *in situ* de *GATA3-AS1*.

Por otro lado, las células transfectadas con el ASOA contra *GATA3-AS1* pierden las características epiteliales que definen al cultivo de MCF-7 y adquieren similitudes con las células de morfología fibroblastoide, como lo son MCF-10A y MDA-MB-231. Esto es relevante ya que las líneas celulares provenientes de fibroblastos mamarios se caracterizan por no expresar *GATA3-AS1* ni *GATA3* (Figuras 10, 11 y 20), por lo que estos resultados sugieren que *GATA3-AS1* podría asociarse con la diferenciación de las células del epitelio mamario como ya ha sido demostrado para *GATA3*, cuya pérdida de función influye en la diferenciación de las células neoplásicas mamarias[136]. En conjunto con los análisis de asociación de funciones (Figura 12A), los resultados sugieren que *GATA3-AS1* podría regular vías de señalización intracelulares relacionadas con el desarrollo y el mantenimiento de células neoplásicas mamarias. No obstante, es necesario caracterizar experimentalmente la función de *GATA3-AS1* en células epiteliales del tejido mamario para confirmarlo.

Entonces, los resultados de este trabajo sugieren que el lincRNA *GATA3-AS1* regula positivamente la expresión de su gen adyacente codificante *GATA3* en *cis*. La relación que existe entre la expresión de ambos genes en cáncer de mama y en las líneas celulares de cáncer de mama sugiere que la regulación de *GATA3-AS1* sobre *GATA3* es positiva, ya que sus loci están enriquecidos con modificaciones post-traduccionales de histonas asociadas con la activación transcripcional. Esto, en conjunto con la presencia de motivos de unión a factores transcripcionales como MYC, ER, AR y la presencia de la RNA Pol II en el promotor compartido de *GATA3-AS1*/*GATA3* sugiere que el mecanismo molecular de regulación sobre el promotor de *GATA3* podría ser llevado a cabo por el reclutamiento de complejos activadores de la transcripción mediado por el lincRNA *GATA3* (Figura 25). Sin embargo, es necesaria más evidencia experimental que confirme esto.

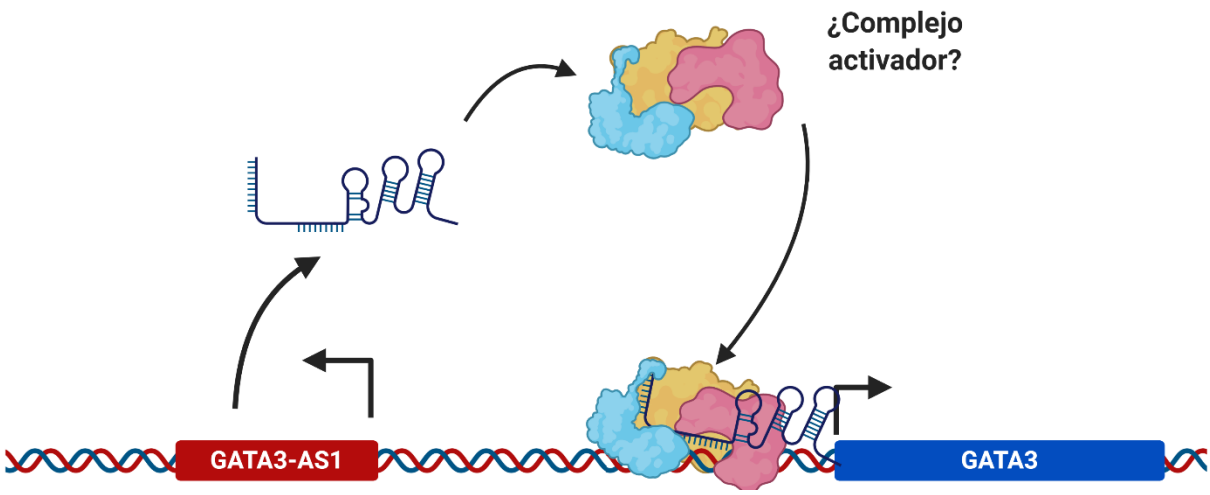


Figura 25. Mecanismo de regulación en *cis* del lincRNA *GATA3-AS1* sobre su gen adyacente *GATA3* en la línea celular MCF-7. En este trabajo, se propone que el lincRNA *GATA3-AS1* regula la expresión de su gen adyacente *GATA3* en *cis.* La evidencia sugiere que el efecto regulatorio activa la transcripción de *GATA3*, posiblemente reclutando complejos activadores, factores transcripcionales o elementos enhancer, que son los mecanismos de regulación propios de los lincRNAs.

En conclusión, los resultados anteriores sugieren que *GATA3-AS1* es un lincRNA divergente conservado, de expresión tejido- y cáncer-específica en las neoplasias mamarias, que a su vez está relacionado con la respuesta a estímulos hormonales que regulan la diferenciación y la proliferación de las células epiteliales mamarias. De igual forma, la evidencia experimental sugiere que regula la expresión de su gen codificante adyacente *GATA3* en *cis* en la línea celular MCF-7, lo que podría estar relacionado con la función que *GATA3-AS1* y *GATA3* llevan a cabo en el desarrollo de cáncer de mama. Finalmente, la determinación de la relación que existe entre el lincRNA *GATA3-AS1* y su gen adyacente *GATA3*, al igual que la caracterización del mecanismo molecular que regula su expresión en las neoplasias mamarias[56,137,138], contribuyen a un mejor entendimiento de padecimientos como el cáncer de mama y de las funciones biológicas

de los lincRNAs divergentes, que actualmente son transcritos que no han sido completamente caracterizados.

13.0 CONCLUSIONES

- El lincRNA *GATA3-AS1* cumple con el primer postulado de la regulación en *cis*, ya que su expresión y la de su gen adyacente *GATA3* presentan correlación positiva en líneas celulares de cáncer de mama y en pacientes con cáncer de mama.
- El lincRNA *GATA3-AS1* cumple con el segundo postulado de la regulación en *cis,* debido a que el abatimiento de su expresión tiene como consecuencia la disminución en la expresión de mRNA de su gen adyacente *GATA3*.
- La expresión de *GATA3-AS1* y *GATA3* podría ser regulada por factores epigenéticos que activan su transcripción, como las modificaciones post-traduccionales de histonas H3K4me3 y la H3K36me3.
- La expresión de *GATA3-AS1* y *GATA3* podría ser regulada por factores transcripcionales de respuesta hormonal como el receptor de estrógenos ESR1.
- *GATA3-AS1* y *GATA3* participan en vías de señalización intracelulares asociadas con *ESR1* en la línea celular MCF-7, por lo tanto, puedan tener una función reguladora en neoplasias mamarias.

14.0 PERSPECTIVAS

- Confirmar la disminución en la expresión de la proteína *GATA3* como consecuencia del abatimiento de la expresión del lincRNA *GATA3-AS1* mediante un experimento de Western blot.
- Llevar a cabo la caracterización del promotor del lincRNA *GATA3-AS1* mediante experimentos como ensayos de luciferasa y ChiP para confirmar la unión de los factores transcripcionales como ER y AR a su región promotora.
- Realizar el experimento de abatimiento de la expresión del receptor de estrógenos en la línea celular MCF-7 para determinar la asociación del lincRNA *GATA3-AS1* con la respuesta a estímulos hormonales por estrógenos.
- Diseñar el experimento de RNA-Sec para determinar los efectos biológicos del abatimiento del lincRNA *GATA3-AS1*, principalmente sobre la expresión de los genes blanco de su gen adyacente *GATA3*.
- Determinar si el lincRNA *GATA3-AS1* modula la expresión de su gen adyacente *GATA3* en *cis* de manera alelo específica a través de un experimento de CRISPR-Cas (tercer postulado de la regulación en *cis*).

15.0 REFERENCIAS BIBLIOGRÁFICAS

1. Evans, J.R., Feng, F.Y., and Chinnaiyan, A.M. (2016). The bright side of dark matter: lncRNAs in cancer. J Clin Invest *126*, 2775–2782.
2. Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P., and Ulitsky, I. (2015). Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. Cell Reports *11*, 1110–1122.
3. Hangauer, M.J., Vaughn, I.W., and McManus, M.T. (2013). Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. PLOS Genetics *9*, e1003569.
4. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.
5. Schmitt, A.M., and Chang, H.Y. (2016). Long Noncoding RNAs in Cancer Pathways. Cancer Cell *29*, 452–463.
6. Huarte, M. (2015). The emerging role of lncRNAs in cancer. Nat. Med. *21*, 1253–1261.
7. Quinn, J.J., and Chang, H.Y. (2016). Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. *17*, 47–62.
8. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.
9. Ma, L., Bajic, V.B., and Zhang, Z. (2013). On the classification of long non-coding RNAs. RNA Biol *10*, 925–933.
10. Ransohoff, J.D., Wei, Y., and Khavari, P.A. (2018). The functions and unique features of long intergenic non-coding RNA. Nat. Rev. Mol. Cell Biol. *19*, 143–157.
11. Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

12. Luo, S., Lu, J.Y., Liu, L., Yin, Y., Chen, C., Han, X., Wu, B., Xu, R., Liu, W., Yan, P., et al. (2016). Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. Cell Stem Cell *18*, 637–652.
13. Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26–46.
14. Mattioli, K., Volders, P.-J., Gerhardinger, C., Lee, J.C., Maass, P.G., Melé, M., and Rinn, J.L. (2019). High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. Genome Res. *29*, 344–355.
15. Kajino, T., Shimamura, T., Gong, S., Yanagisawa, K., Ida, L., Nakatochi, M., Griesing, S., Shimada, Y., Kano, K., Suzuki, M., et al. (2019). Divergent lncRNA MYMLR regulates MYC by eliciting DNA looping and promoter-enhancer interaction. The EMBO Journal *38*, e98441.
16. Galupa, R., Nora, E.P., Worsley-Hunt, R., Picard, C., Gard, C., Bemmel, J.G. van, Servant, N., Zhan, Y., Marjou, F.E., Johanneau, C., et al. (2020). A Conserved Noncoding Locus Regulates Random Monoallelic Xist Expression across a Topological Boundary. Molecular Cell *77*, 352-367.e8.
17. Daneshvar, K., Pondick, J.V., Kim, B.-M., Zhou, C., York, S.R., Macklin, J.A., Abualteen, A., Tan, B., Sigova, A.A., Marcho, C., et al. (2016). DIGIT Is a Conserved Long Noncoding RNA that Regulates GSC Expression to Control Definitive Endoderm Differentiation of Embryonic Stem Cells. Cell Reports *17*, 353–365.
18. Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.
19. Gil, N., and Ulitsky, I. (2020). Regulation of gene expression by cis -acting long non-coding RNAs. Nat Rev Genet *21*, 102–117.
20. Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M., and Lander, E.S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452–455.
21. Sun, X., Wang, Z., Hall, J.M., Perez-Cervantes, C., Ruthenburg, A.J., Moskowitz, I.P., Gribskov, M., and Yang, X.H. (2020). Chromatin-enriched RNAs mark active and repressive cis-regulation: An analysis of nuclear RNA-seq. PLOS Computational Biology *16*, e1007119.

22. Igolkina, A.A., Zinkevich, A., Karandasheva, K.O., Popov, A.A., Selifanova, M.V., Nikolaeva, D., Tkachev, V., Penzar, D., Nikitin, D.M., and Buzdin, A. (2019). H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 Histone Tags Suggest Distinct Regulatory Evolution of Open and Condensed Chromatin Landmarks. Cells *8*.
23. Capizzi, M., Strappazzon, F., Cianfanelli, V., Papaleo, E., and Cecconi, F. (2017). MIR7-3HG, a MYC-dependent modulator of cell proliferation, inhibits autophagy by a regulatory loop involving AMBRA1. Autophagy *13*, 554–566.
24. Achinger-Kawecka, J., Valdes-Mora, F., Luu, P.-L., Giles, K.A., Caldon, C.E., Qu, W., Nair, S., Soto, S., Locke, W.J., Yeo-Teh, N.S., et al. (2020). Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. Nature Communications *11*, 320.
25. Sun, M., Gadad, S.S., Kim, D.-S., and Kraus, W.L. (2015). Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. Mol Cell *59*, 698–711.
26. Lee, D., Yang, B., Sedano, M., Choudhari, R., and Gadad, S.S. (2020). SUN-733 Analysis of Divergent Long Noncoding RNAs in Estrogen-Regulated Transcription. Journal of the Endocrine Society *4*.
27. Zhang, Y., Wang, D.-L., Yan, H.-Y., Liao, J.-Y., He, J.-H., Hu, K.-S., Deng, W.-X., Wang, Y.-J., Xing, H.-T., Koeffler, H.P., et al. (2017). Genome-wide study of ER-regulated lncRNAs shows AP000439.3 may function as a key regulator of cell cycle in breast cancer. Oncology Reports *38*, 3227–3237.
28. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208.
29. Subhash, S., Mishra, K., Akhade, V.S., Kanduri, M., Mondal, T., and Kanduri, C. (2018). H3K4me2 and WDR5 enriched chromatin interacting long non-coding RNAs maintain transcriptionally competent chromatin at divergent transcriptional units. Nucleic Acids Research *46*, 9384–9400.

30. Bhat, S.A., Ahmad, S.M., Mumtaz, P.T., Malik, A.A., Dar, M.A., Urwat, U., Shah, R.A., and Ganai, N.A. (2016). Long non-coding RNAs: Mechanism of action and functional utility. Non-coding RNA Research *1*, 43–50.
31. Rinn, J.L., and Chang, H.Y. (2012). Genome Regulation by Long Noncoding RNAs. Annual Review of Biochemistry *81*, 145–166.
32. Gao, N., Li, Y., Li, J., Gao, Z., Yang, Z., Li, Y., Liu, H., and Fan, T. (2020). Long Non-Coding RNAs: The Regulatory Mechanisms, Research Strategies, and Future Directions in Cancers. Frontiers in Oncology *10*, 2903.
33. Chen, J., Shishkin, A.A., Zhu, X., Kadri, S., Maza, I., Guttman, M., Hanna, J.H., Regev, A., and Garber, M. (2016). Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biology *17*, 19.
34. Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol *22*, 96–118.
35. Toiber, D., Leprivier, G., and Rotblat, B. (2017). Long noncoding RNA: noncoding and not coded. cddiscovery *3*, 16104.
36. Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. PNAS *110*, 2876–2881.
37. Cipriano, A., Macino, M., Buonaiuto, G., Santini, T., Biferali, B., Peruzzi, G., Colantoni, A., Mozzetta, C., and Ballarino, M. (2021). Epigenetic regulation of Wnt7b expression by the cis-acting long noncoding RNA Lnc-Rewind in muscle stem cells. eLife *10*, e54782.
38. Wang, Y., Chen, S., Li, W., Jiang, R., and Wang, Y. (2020). Associating divergent lncRNAs with target genes by integrating genome sequence, gene expression and chromatin accessibility data. NAR Genomics and Bioinformatics *2*.
39. Han, X., Luo, S., Peng, G., Lu, J.Y., Cui, G., Liu, L., Yan, P., Yin, Y., Liu, W., Wang, R., et al. (2018). Mouse knockout models reveal largely dispensable but context-dependent functions of lncRNAs during development. Journal of Molecular Cell Biology *10*, 175–178.

40. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.
41. Carlevaro-Fita, J., Lanzós, A., Feuerbach, L., Hong, C., Mas-Ponte, D., Pedersen, J.S., and Johnson, R. (2020). Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. Commun Biol *3*, 1–16.
42. Yan, X., Hu, Z., Feng, Y., Hu, X., Yuan, J., Zhao, S.D., Zhang, Y., Yang, L., Shan, W., He, Q., et al. (2015). Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. Cancer Cell *28*, 529–540.
43. Wang, Z., Yang, B., Zhang, M., Guo, W., Wu, Z., Wang, Y., Jia, L., Li, S., Cancer Genome Atlas Research Network, Xie, W., et al. (2018). 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.
44. Logotheti, S., Marquardt, S., Gupta, S.K., Richter, C., Edelhäuser, B.A.H., Engelmann, D., Brenmoehl, J., Söhnchen, C., Murr, N., Alpers, M., et al. (2020). LncRNA-SLC16A1-AS1 induces metabolic reprogramming during Bladder Cancer progression as target and co-activator of E2F1. Theranostics *10*, 9620–9643.
45. Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*, 199–208.
46. Pan, Y., Li, C., Chen, J., Zhang, K., Chu, X., Wang, R., and Chen, L. (2016). The Emerging Roles of Long Noncoding RNA ROR (lincRNA-ROR) and its Possible Mechanisms in Human Cancers. Cell. Physiol. Biochem. *40*, 219–229.
47. Lian, Y., Yan, C., Xu, H., Yang, J., Yu, Y., Zhou, J., Shi, Y., Ren, J., Ji, G., and Wang, K. (2018). A Novel lncRNA, LINC00460, Affects Cell Proliferation and Apoptosis by Regulating KLF2 and CUL4A Expression in Colorectal Cancer. Molecular Therapy - Nucleic Acids *12*, 684–697.
48. Wu, W., Wagner, E.K., Hao, Y., Rao, X., Dai, H., Han, J., Chen, J., Storniolo, A.M.V., Liu, Y., and He, C. (2016). Tissue-specific Co-expression of Long Non-coding and Coding RNAs Associated with Breast Cancer. Sci Rep *6*, 32731.

49. Fan, C.-N., Ma, L., and Liu, N. (2018). Systematic analysis of lncRNA-miRNA-mRNA competing endogenous RNA network identifies four-lncRNA signature as a prognostic biomarker for breast cancer. J Transl Med *16*, 264.
50. Zhang, S., Wang, J., Ghoshal, T., Wilkins, D., Mo, Y.-Y., Chen, Y., and Zhou, Y. (2018). lncRNA Gene Signatures for Prediction of Breast Cancer Intrinsic Subtypes and Prognosis. Genes (Basel) *9*.
51. Zeng, Y., Wang, G., Zhou, C.-F., Zhang, H.-B., Sun, H., Zhang, W., Zhou, H.-H., Liu, R., and Zhu, Y.-S. (2019). LncRNA Profile Study Reveals a Three-LncRNA Signature Associated With the Pathological Complete Response Following Neoadjuvant Chemotherapy in Breast Cancer. Front Pharmacol *10*, 574.
52. Rossi, T., Pistoni, M., Sancisi, V., Gobbi, G., Torricelli, F., Donati, B., Ribisi, S., Gugnoni, M., and Ciarrocchi, A. (2020). RAIN Is a Novel Enhancer-Associated lncRNA That Controls RUNX2 Expression and Promotes Breast and Thyroid Cancer. Mol Cancer Res *18*, 140–152.
53. Tariq, A., Hao, Q., Sun, Q., Singh, D.K., Jadaliha, M., Zhang, Y., Chetlangia, N., Ma, J., Holton, S.E., Bhargava, R., et al. (2020). LncRNA-mediated regulation of SOX9 expression in basal subtype breast cancer cells. RNA *26*, 175–185.
54. Shin, T.-J., Lee, K.-H., Cho, H.-M., and Cho, J.-Y. (2019). Concise approach for screening long non-coding RNAs functionally linked to human breast cancer associated genes. Experimental and Molecular Pathology *108*, 89–96.
55. Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., and Cardoso, F. (2019). Breast cancer. Nature Reviews Disease Primers *5*, 1–31.
56. Bhat-Nakshatri, P., Gao, H., Sheng, L., McGuire, P.C., Xuei, X., Wan, J., Liu, Y., Althouse, S.K., Colter, A., Sandusky, G., et al. (2021). A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. CR Med *2*.
57. Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.
58. Berger, A.C., Korkut, A., Kanchi, R.S., Hegde, A.M., Lenoir, W., Liu, W., Liu, Y., Fan, H., Shen, H., Ravikumar, V., et al. (2018). A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell *33*, 690-705.e9.

59. Niknafs, Y.S., Han, S., Ma, T., Speers, C., Zhang, C., Wilder-Romans, K., Iyer, M.K., Pitchiaya, S., Malik, R., Hosono, Y., et al. (2016). The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. Nat Commun *7*, 12791.
60. Dai, X., Cheng, H., Bai, Z., and Li, J. (2017). Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. J Cancer *8*, 3131–3141.
61. Leung, E.Y., Askarian-Amiri, M.E., Singleton, D.C., Ferraro-Peyret, C., Joseph, W.R., Finlay, G.J., Broom, R.J., Kakadia, P.M., Bohlander, S.K., Marshall, E., et al. (2018). Derivation of Breast Cancer Cell Lines Under Physiological (5%) Oxygen Concentrations. Front. Oncol. *8*.
62. Holliday, D.L., and Speirs, V. (2011). Choosing the right cell line for breast cancer research. Breast Cancer Res. *13*, 215.
63. Mirabelli, P., Coppola, L., and Salvatore, M. (2019). Cancer Cell Lines Are Useful Model Systems for Medical Research. Cancers (Basel) *11*, 1098.
64. Gibbons, H.R., Shaginurova, G., Kim, L.C., Chapman, N., Spurlock, C.F.I., and Aune, T.M. (2018). Divergent lncRNA GATA3-AS1 Regulates GATA3 Transcription in T-Helper 2 Cells. Front. Immunol. *9*.
65. Zhang, H., Nestor, C.E., Zhao, S., Lentini, A., Bohle, B., Benson, M., and Wang, H. (2013). Profiling of human CD4+ T-cell subsets identifies the TH2-specific noncoding RNA GATA3-AS1. J. Allergy Clin. Immunol. *132*, 1005–1008.
66. Contreras-Espinosa, L., Alcaraz, N., De La Rosa-Velázquez, I.A., Díaz-Chávez, J., Cabrera-Galeana, P., Vega, R.R., Reynoso-Noveron, N., Maldonado-Martínez, H.A., González-Barrios, R., Montiel-Manríquez, R., et al. (2021). Transcriptome analysis identifies GATA3-AS1 as a long noncoding RNA associated with resistance to neoadjuvant chemotherapy in locally advanced breast cancer patients. J Mol Diagn, S1525-1578(21)00238–5.
67. Zhang, Y., Wagner, E.K., Guo, X., May, I., Cai, Q., Zheng, W., He, C., and Long, J. (2016). Long intergenic non-coding RNA expression signature in human breast cancer. Scientific Reports *6*, 37821.
68. Ensembl genome browser 100 https://www.ensembl.org/index.html.

69. Mas-Ponte, D., Carlevaro-Fita, J., Palumbo, E., Hermoso Pulido, T., Guigo, R., and Johnson, R. (2017). LncATLAS database for subcellular localization of long noncoding RNAs. RNA *23*, 1080–1087.
70. lncATLAS http://lncatlas.crg.eu.
71. Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res *45*, W98–W102.
72. GEPIA 2 http://gepia2.cancer-pku.cn/#index.
73. Broad Institute Cancer Cell Line Encyclopedia (CCLE) https://portals.broadinstitute.org/ccle.
74. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics *45*, 580–585.
75. GTEx Portal https://www.gtexportal.org/home/.
76. Breast Cancer Profiling Project, Gene Expression 1: Baseline mRNA sequencing on 35 breast cell lines - Dataset - HMS LINCS Database - HMS LINCS Project http://lincs.hms.harvard.edu/db/datasets/20348/main.
77. Niepel, M., Hafner, M., Pace, E.A., Chung, M., Chai, D.H., Zhou, L., Schoeberl, B., and Sorger, P.K. (2013). Profiles of Basal and stimulated receptor signaling networks predict drug response in breast cancer lines. Sci Signal *6*, ra84.
78. Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res *46*, W537–W544.
79. Galaxy https://usegalaxy.org.
80. Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., Weinstein, J.N., and Liang, H. (2015). TANRIC: An Interactive Open Platform to Explore the Function of lncRNAs in Cancer. Cancer Res. *75*, 3728–3737.
81. TANRIC: Home https://ibl.mdanderson.org/tanric/_design/basic/main.html.

82. Alam, T., Uludag, M., Essack, M., Salhi, A., Ashoor, H., Hanks, J.B., Kapfer, C., Mineta, K., Gojobori, T., and Bajic, V.B. (2017). FARNA: knowledgebase of inferred functions of non-coding RNA transcripts. Nucleic Acids Res. *45*, 2838–2848.
83. FARNA: Knowledgebase of Annotated Functions of Non-coding RNA Transcripts http://www.cbrc.kaust.edu.sa/farna.
84. STRING: functional protein association networks https://string-db.org/.
85. WashU EpiGenome Browser http://epigenomegateway.wustl.edu/browser.
86. Cistrome Project http://cistrome.org/.
87. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.-H., Brown, M., Zhang, X., Meyer, C.A., et al. (2019). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Research *47*, D729–D735.
88. Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., Tang, Q., Meyer, C.A., Zhang, Y., and Liu, X.S. (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat Protoc *8*, 2502–2515.
89. Introduction - MEME Suite https://meme-suite.org/meme/index.html.
90. KEGG Kyoto Encyclopedia of Genes and Genomes http://www.genome.jp/kegg.
91. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol *9*, S4.
92. GeneMANIA https://genemania.org/.
93. Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., and Tartaglia, G.G. (2013). catRAPID omics: a web server for large-scale prediction of protein–RNA interactions. Bioinformatics *29*, 2928–2930.
94. catRAPID http://service.tartaglialab.com/page/catrapid_group.
95. RNAInter: RNA Interactome Database https://www.rna-society.org/rnainter/.
96. Lin, Y., Liu, T., Cui, T., Wang, Z., Zhang, Y., Tan, P., Huang, Y., Yu, J., and Wang, D. (2020). RNAInter in 2020: RNA interactome repository with increased coverage and annotation. Nucleic Acids Res *48*, D189–D197.
97. Paraskevopoulou, M.D., Vlachos, I.S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., Zagganas, K., Tsanakas, P., Floros, E., Dalamagas, T., et al. (2016).

DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. Nucleic Acids Res *44*, D231–D238.

98. Chang, L., Zhou, G., Soufan, O., and Xia, J. (2020). miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Research *48*, W244–W251.
99. MCF 10A | ATCC https://www.atcc.org/products/crl-10317.
100. MCF7 | ATCC https://www.atcc.org/products/htb-22.
101. BT-474 | ATCC https://www.atcc.org/products/htb-20.
102. MDA-MB-231 | ATCC https://www.atcc.org/products/htb-26.
103. Physical Sciences in Oncology: Cell Line Protocol.
104. Rio, D.C., Ares, M., Hannon, G.J., and Nilsen, T.W. (2010). Purification of RNA using TRIzol (TRI reagent). Cold Spring Harb Protoc *2010*, pdb.prot5439.
105. Desjardins, P., and Conklin, D. (2010). NanoDrop Microvolume Quantitation of Nucleic Acids. J Vis Exp.
106. Volders, P.J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J., and Mestdagh, P. (2015). An update on LNCipedia: a database for annotated human lncRNA sequences. Nucleic Acids Res. *43*, 4363–4364.
107. LNCipedia https://lncipedia.org.
108. Primer-Blast https://www.ncbi.nlm.nih.gov/tools/primer-blast/.
109. UCSC In-Silico PCR https://genome.ucsc.edu.
110. Bustin, S., and Huggett, J. (2017). qPCR primer design revisited. Biomol Detect Quantif *14*, 19–28.
111. Life Technologies - US https://www.thermofisher.com/us/en/home.html.
112. Primer design and amplification efficiencies are crucial for reliability of quantitative PCR studies of caffeine biosynthetic N -methyltransferases in coffee | SpringerLink https://link.springer.com/article/10.1007/s13205-018-1487-5.
113. Taylor, S.C., Nadeau, K., Abbasi, M., Lachance, C., Nguyen, M., and Fenrich, J. (2019). The Ultimate qPCR Experiment: Producing Publication Quality, Reproducible Data the First Time. Trends in Biotechnology *37*, 761–774.

114. Rao, X., Huang, X., Zhou, Z., and Lin, X. (2013). An improvement of the 2ˆ(–delta delta CT) method for quantitative real-time polymerase chain reaction data analysis. Biostat Bioinforma Biomath *3*, 71–85.

115. Tian, L., Chou, H.-L., Zhang, L., and Okita, T.W. (2019). Targeted Endoplasmic Reticulum Localization of Storage Protein mRNAs Requires the RNA-Binding Protein RBP-L. Plant Physiology *179*, 1111–1131.

116. Mayer, A., and Churchman, L.S. (2017). A detailed protocol for subcellular RNA sequencing (subRNA-seq). Curr Protoc Mol Biol *120*, 4.29.1-4.29.18.

117. RNAfold web server http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi.

118. Welcome to the Predict a Secondary Structure Web Server https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html.

119. Reverse Complement https://www.bioinformatics.org/sms/rev_comp.html.

120. OligoAnalyzer Tool - primer analysis | IDT Integrated DNA Technologies. https://www.idtdna.com/pages/tools/oligoanalyzer.

121. DeVos, S.L., and Miller, T.M. (2013). Antisense oligonucleotides: treating neurodegeneration at the level of RNA. Neurotherapeutics *10*, 486–497.

122. Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., et al. (2010). The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. Molecular Cell *39*, 925–938.

123. Tan, R., Li, H., Huang, Z., Zhou, Y., Tao, M., Gao, X., and Xu, Y. (2020). Neural Functions Play Different Roles in Triple Negative Breast Cancer (TNBC) and non-TNBC. Scientific Reports *10*, 3065.

124. Lee, J.Y., Park, Y.J., Oh, N., Kwack, K.B., and Park, K.-S. (2017). A transcriptional complex composed of ER(α), GATA3, FOXA1 and ELL3 regulates IL-20 expression in breast cancer cells. Oncotarget *8*, 42752–42760.

125. Theodorou, V., Stark, R., Menon, S., and Carroll, J.S. (2013). GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. Genome Res. *23*, 12–22.

126. Wittkopp, P.J., and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet *13*, 59–69.

127. Luo, X., Zhou, N., Wang, L., Zeng, Q., and Tang, H. (2019). Long Noncoding RNA GATA3-AS1 Promotes Cell Proliferation and Metastasis in Hepatocellular Carcinoma by Suppression of PTEN, CDKN1A, and TP53. Can J Gastroenterol Hepatol *2019*, 1389653.

128. Zhang, M., Wang, N., Song, P., Fu, Y., Ren, Y., Li, Z., and Wang, J. (2020). LncRNA GATA3-AS1 facilitates tumour progression and immune escape in triple-negative breast cancer through destabilization of GATA3 but stabilization of PD-L1. Cell Prolif *53*, e12855.

129. Chou, J., Provot, S., and Werb, Z. (2010). GATA3 in Development and Cancer Differentiation: Cells GATA Have It! J Cell Physiol *222*, 42–49.

130. Rinn, J.L., and Chang, H.Y. (2020). Long Noncoding RNAs: Molecular Modalities to Organismal Functions. Annu Rev Biochem *89*, 283–308.

131. Thaler, J.P., Lee, S.-K., Jurata, L.W., Gill, G.N., and Pfaff, S.L. (2002). LIM Factor Lhx3 Contributes to the Specification of Motor Neuron and Interneuron Identity through Cell-Type-Specific Protein-Protein Interactions. Cell *110*, 237–249.

132. Dietrich, D., Lesche, R., Tetzner, R., Krispin, M., Dietrich, J., Haedicke, W., Schuster, M., and Kristiansen, G. (2009). Analysis of DNA Methylation of Multiple Genes in Microdissected Cells From Formalin-fixed and Paraffin-embedded Tissues. J Histochem Cytochem. *57*, 477–489.

133. Awwad, D.A. (2019). Beyond classic editing: innovative CRISPR approaches for functional studies of long non-coding RNA. Biol Methods Protoc *4*, bpz017.

134. Meng, H., and Bartholomew, B. (2018). Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II. Journal of Biological Chemistry *293*, 13786–13794.

135. Miao, Y., Ajami, N.E., Huang, T.-S., Lin, F.-M., Lou, C.-H., Wang, Y.-T., Li, S., Kang, J., Munkacsi, H., Maurya, M.R., et al. (2018). Enhancer-associated long non-coding RNA LEENE regulates endothelial nitric oxide synthase and endothelial function. Nat Commun *9*, 292.

136. Takaku, M., Grimm, S.A., De Kumar, B., Bennett, B.D., and Wade, P.A. (2020). Cancer-specific mutation of GATA3 disrupts the transcriptional regulatory network governed by Estrogen Receptor alpha, FOXA1 and GATA3. Nucleic Acids Research *48*, 4756–4768.

137. Amorim, M., Salta, S., Henrique, R., and Jerónimo, C. (2016). Decoding the usefulness of non-coding RNAs as breast cancer markers. Journal of Translational Medicine *14*, 265.

138. DeMichele, A., Yee, D., and Esserman, L. (2017). Mechanisms of Resistance to Neoadjuvant Chemotherapy in Breast Cancer. New England Journal of Medicine *377*, 2287–2289.

139. HMC18 DepMap Cell Line Summary https://depmap.org/portal/cell_line/ACH-000721?tab=mutation.

140. Zhang, X., Li, M., Momcilovic, O., Beardsley, A., Camarda, R., and Goga, A. (2016). Functional annotation of cancer driver genes in breast cancer patient-derived xenografts to identify a novel target for PARP inhibitors. JCO *34*, e23192–e23192.

141. Home - GEO DataSets - NCBI https://www.ncbi.nlm.nih.gov/gds.

142. Hafner, M., Mills, C.E., Subramanian, K., Chen, C., Chung, M., Boswell, S.A., Everley, R.A., Liu, C., Walmsley, C.S., Juric, D., et al. (2019). Multiomics Profiling Establishes the Polypharmacology of FDA-Approved CDK4/6 Inhibitors and the Potential for Differential Clinical Activity. Cell Chemical Biology *26*, 1067-1080.e8.

16.0 APÉNDICES

16.1 APÉNDICE A: ANÁLISIS DE EXPRESIÓN DEL LINCRNA *GATA3-AS1* Y *GATA3* EN LAS LÍNEAS CELULARES DE LA BASE DE DATOS *CCLE*

Tabla suplementaria 1: Clasificación por fenotipo de las principales líneas celulares de cáncer de mama utilizadas en la investigación.

Fenotipo	Línea celular	Referencia
Tipo-normal (transformada)	MCF-10A, HME1, HS281T, HS343T, HS606T, HS739T, HS742T	[99,139]
Luminal A	MCF-7, T47D, HCC1428, EFM19, CAMA1, HCC1500, MDA-MB-415, MDA-MB-134	[60,62]
Luminal B (HER2+)	BT474, ZR7530, MDA-MB-361, EFM192A, MDA-MB-361	[60,62]
HER2 enriquecido	AU565, SK-BR-3, HCC1954, HCC1569, HCC202, MDA-MB-453, HCC1419, HCC2218	[60]
Basal (Triple negativo)	BT20, HCC38, HCC1395, MDA-MB-231, MDA-MB-157, SUM149, MDA-MB-468, CAL120, CAL51, CAL851, CAL148, DU4475, HCC1143, HCC1187, HCC1599, HCC1806, HCC1937, HCC2157, HCC70, HDQ-P1, MDA-MB,436, MFM223, SKBR7, SUM159, HMC18, JIMT1, PDX1328, PDXHCI002	[60,140,141]
Bajo en claudina	BT549, SUM1315, Hs578T, PDX1258	[62,142]

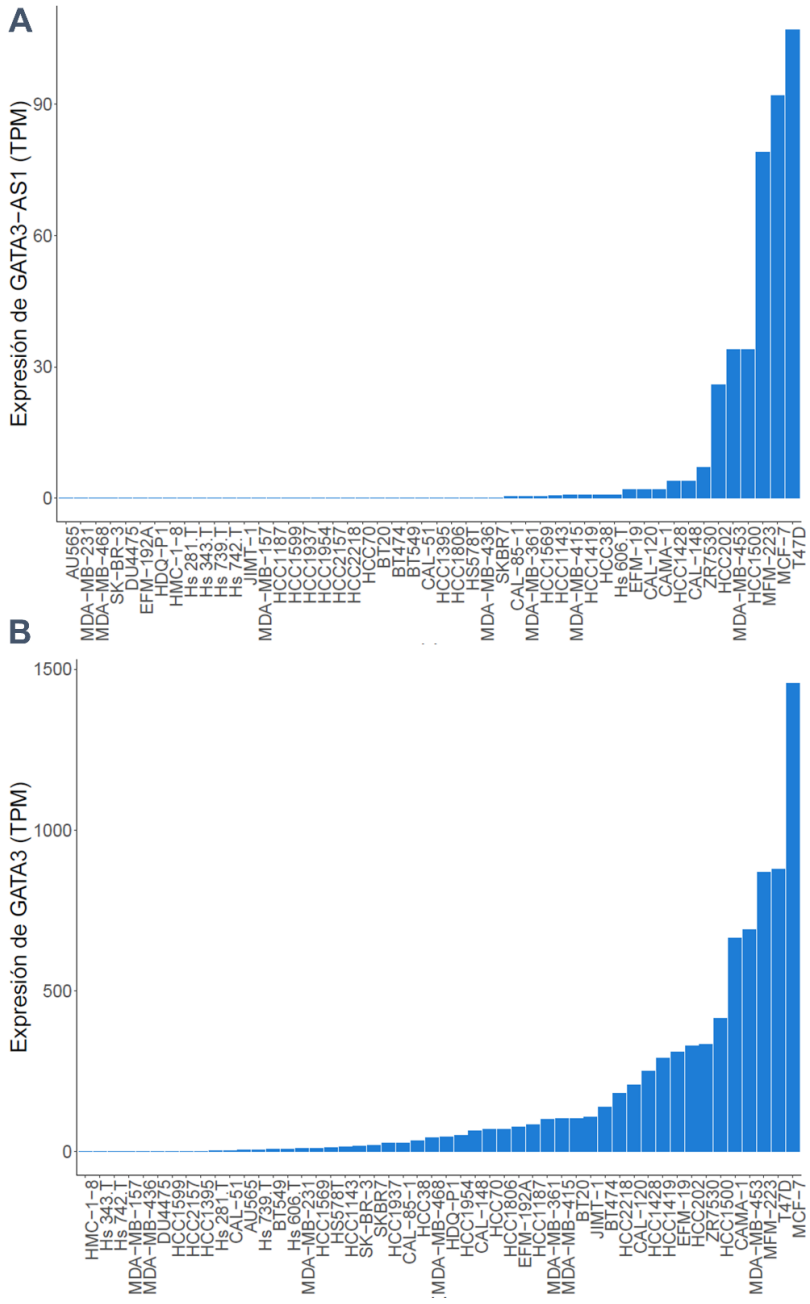


Figura suplementaria 1. Expresión del lincRNA *GATA3-AS1* y *GATA3* por RNA-Seq en líneas celulares de cáncer de mama. Gráficos de barras que muestran la expresión del lincRNA *GATA3-AS1* (A) y su gen adyacente *GATA3* (B) en las cincuenta líneas celulares de cáncer de mama depositadas en la base de datos *CCLE*[73].

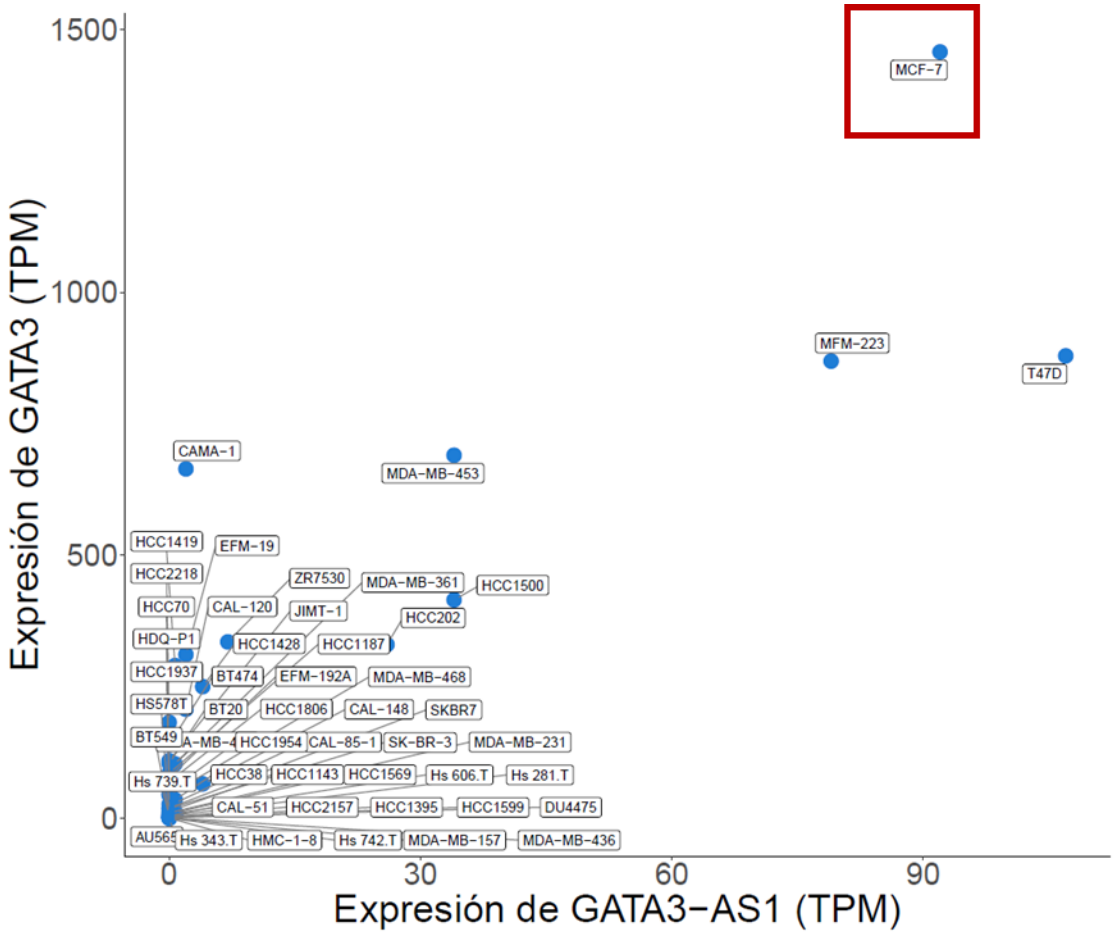


Figura suplementaria 2. La expresión del lincRNA *GATA3-AS1* correlaciona positivamente con la expresión de su gen adyacente *GATA3* en líneas celulares de cáncer de mama. A) Gráfico de correlación entre la expresión de *GATA3-AS1* y *GATA3* en las líneas celulares de cáncer de mama depositadas en *CCLE*[73]. Particularmente, se observa la correlación positiva en la línea celular MCF-7 (cuadro rojo) (correlación de Pearson =0.87, valor *p* = 2.011e-08).

16.2 APÉNDICE B: CONSTRUCCIÓN DEL MAPA DE CROMATINA PARA LA LÍNEAS CELULARES MCF-7 y MCF-10A

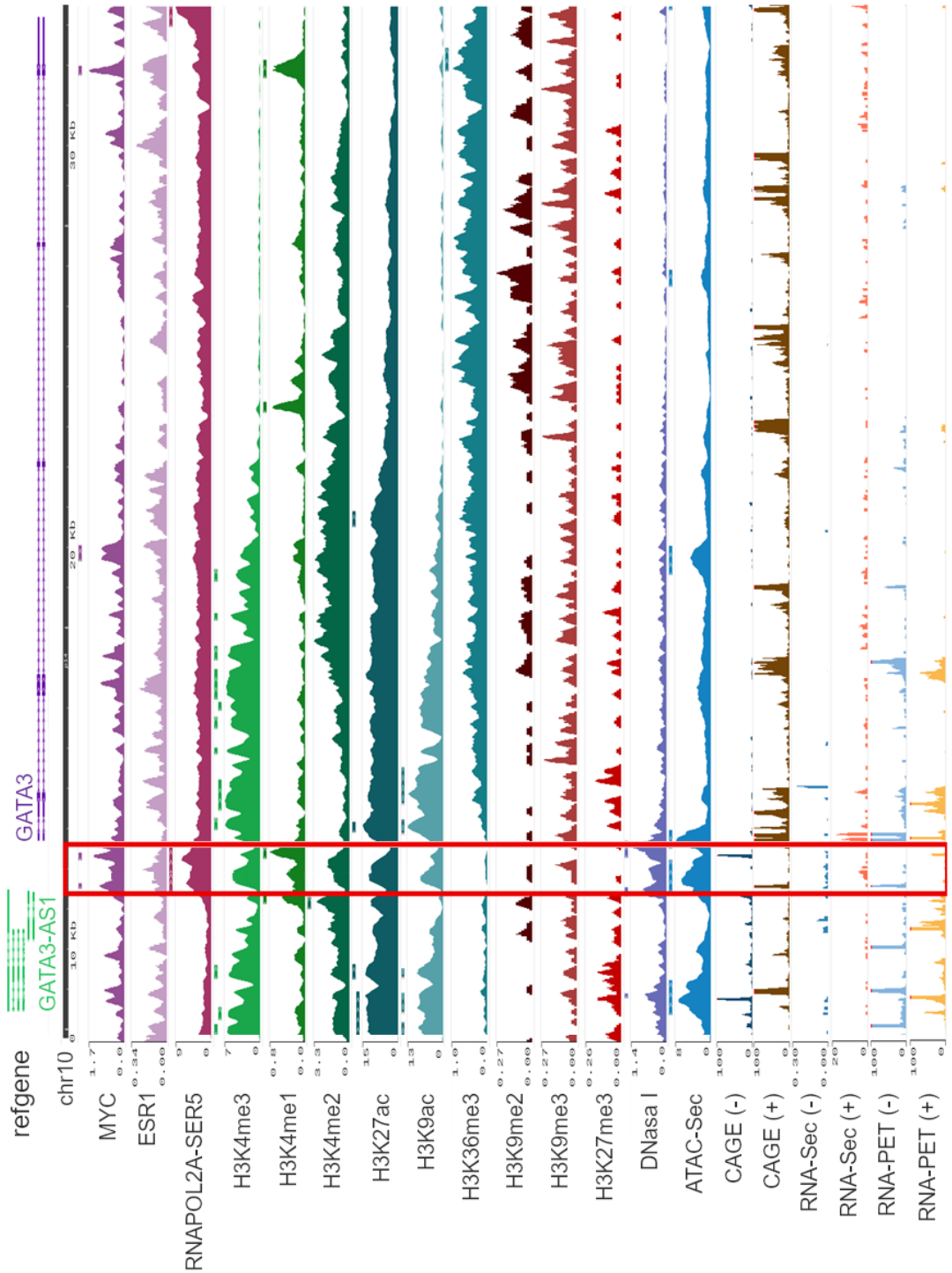


Figura suplementaria 3. Caracterización *in silico* de las marcas de cromatina asociadas al locus *GATA3-AS1*/*GATA3* en la línea celular MCF-7. Mapa de cromatina

que muestre la posición genómica del locus *GATA3-AS1*/*GATA3*, la estructura de los transcritos del lincRNA *GATA3-AS1* y *GATA3* (panel superior) y los histogramas de resultados de experimentos de ChIP-Seq en el modelo celular neoplásico MCF-7, donde se incluyen: ChIP-Sec: MYC, *GATA3*, SUZ12, ESR1, RNAPOL2A (P-Ser2), H3K4me3, H3K4me1, H3K4me2, H3K27ac, H3K9ac, H3K36me3, H3K9me2, H3K9me3, H3K27me3, DNasa I, ATAC-Sec, Análisis de la expresión de genes con modificación Cap 5' (CAGE) en la cadena sentido (+) y antisentido (-), RNA-Sec en la cadena sentido (+) y antisentido (-), Secuenciación masiva en paralelo de RNA de extremo pareado marcado (RNA-PET) en la cadena sentido (+) y antisentido (-).

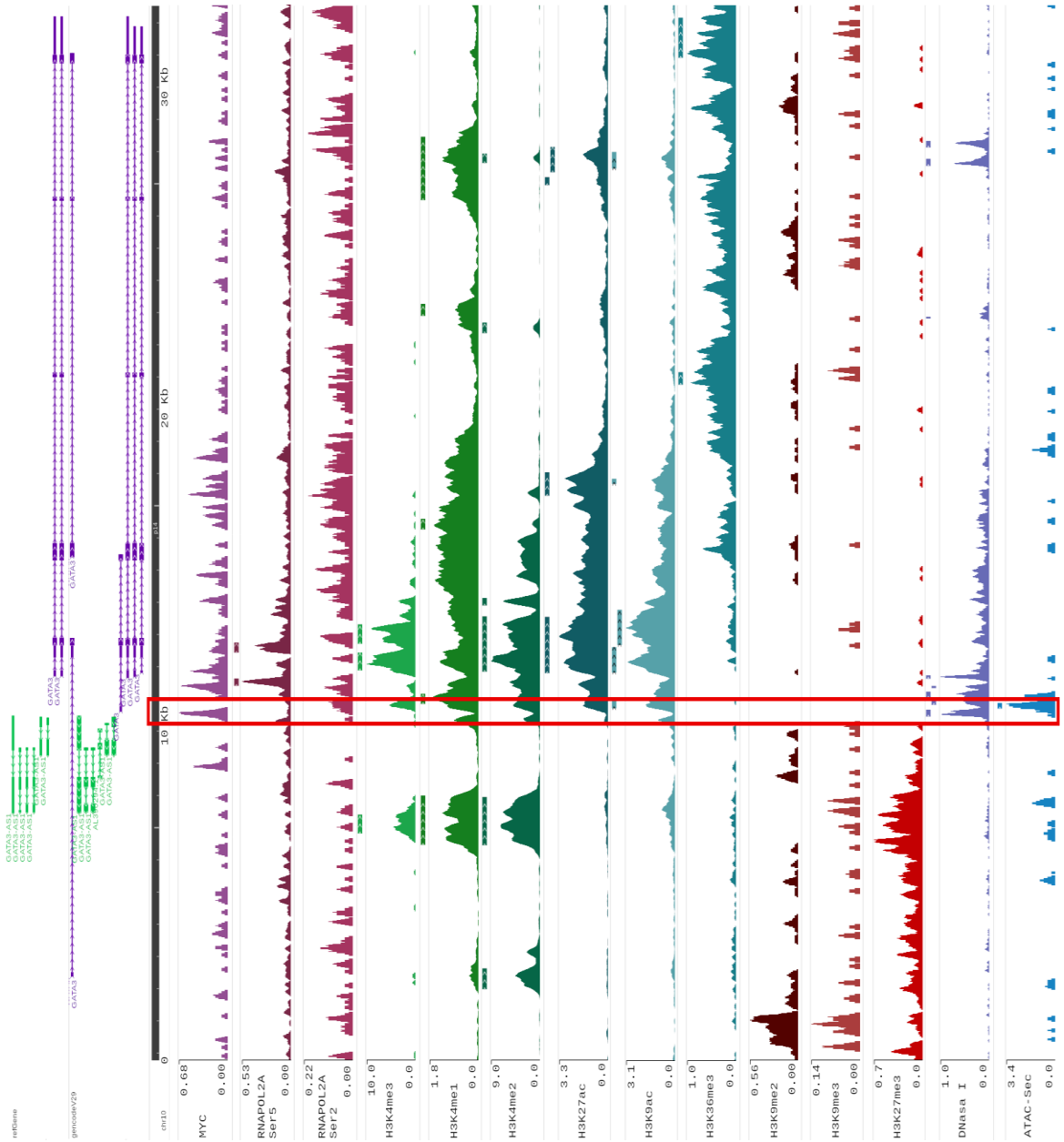
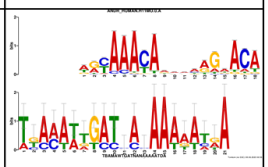
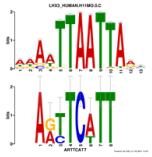
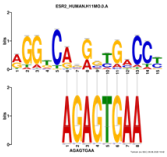


Figura suplementaria 4. Caracterización *in silico* de las marcas de cromatina asociadas al locus *GATA3-AS1*/*GATA3* en la línea celular transformada MCF-10A. Mapa de cromatina que muestre la posición genómica del locus *GATA3-AS1*/*GATA3*, la estructura de los transcritos del lincRNA *GATA3-AS1* y *GATA3* (panel superior) y los histogramas de resultados de experimentos de ChIP-Seq en el modelo celular neoplásico MCF-7, donde se incluyen: ChIP-Sec: MYC, , RNAPOL2A (P-Ser2 y P-Ser5), H3K4me3, H3K4me1, H3K4me2, H3K27ac, H3K9ac, H3K36me3, H3K9me2, H3K9me3, H3K27me3, DNasa I, ATAC-Sec.

16.3 APÉNDICE C: ANÁLISIS FUNCIONAL DE LOS FACTORES TRANSCRIPCIONALES CON MOTIVOS DE RECONOCIMIENTO EN LA REGIÓN PROMOTORA DEL LINCRNA *GATA3-AS1*

Tabla suplementaria 2: Evaluación de motivos de unión a factores transcripcionales en la región promotora del lincRNA *GATA3-AS1*

Motivo de DNA	Factor transcripcional	*P value*	LOGO
TBAMAWTGATNANAAAAATDA AGAACA	AR	1.38e-02	
ARTTCATT	LHX3	2.05e-03	
AGAGTGAA	ESR2	1.05e-02	

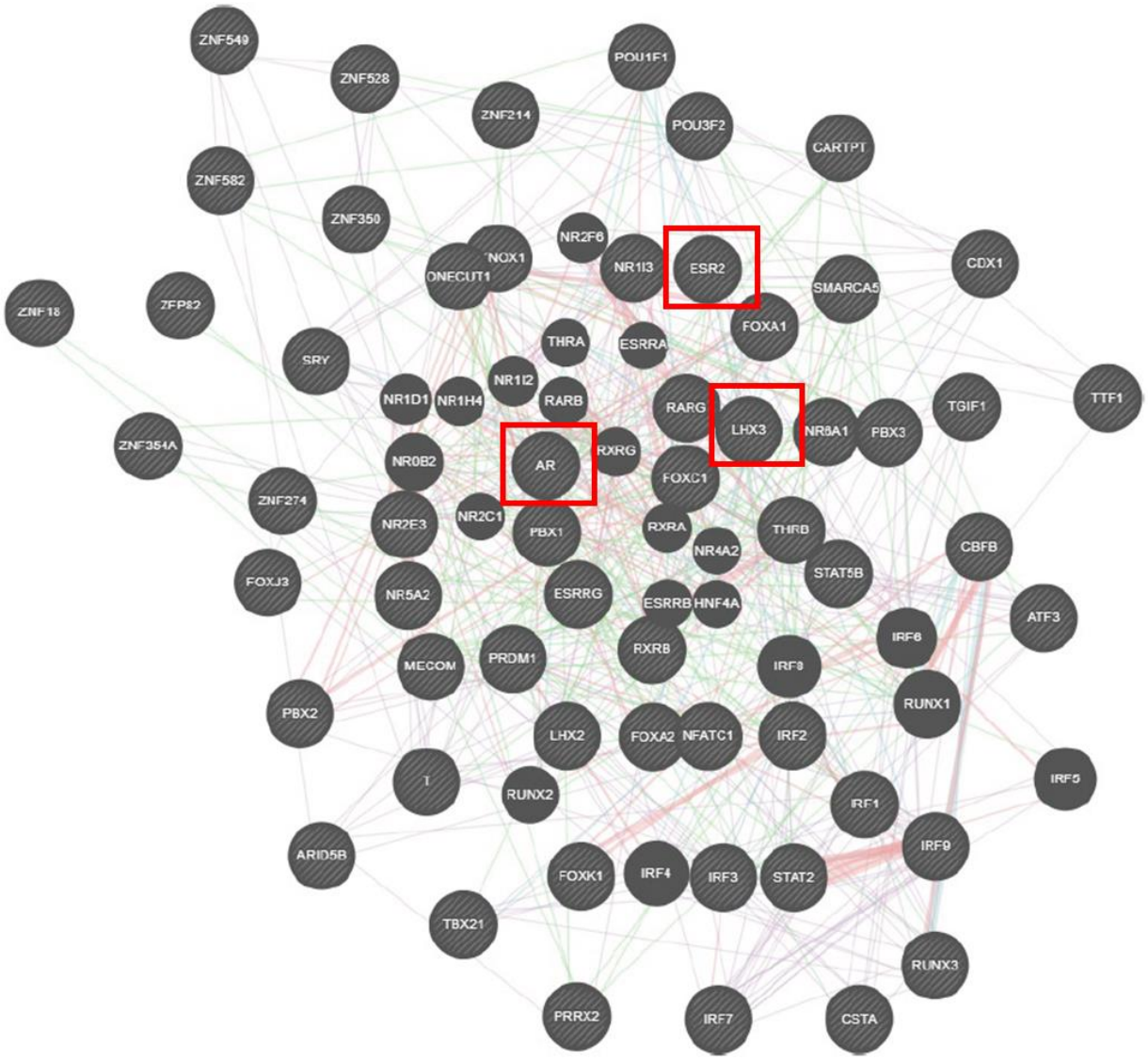


Figura suplementaria 5. Análisis de redes de los factores transcripcionales con motivos de unión a la región promotora del lincRNA *GATA3-AS1*. El análisis de redes realizado con la plataforma *GeneMANIA*[91,92] muestra las asociaciones funcionales con validación experimental de los factores transcripcionales identificados con la paquetería *MEME-SUITE*. Los nodos (gris) se conectan con líneas cuya intensidad de color (rojo) aumenta conforme al puntaje de interacción. En cuadros rojos se resaltan los factores transcripcionales AR, ESR2 y LHX3.

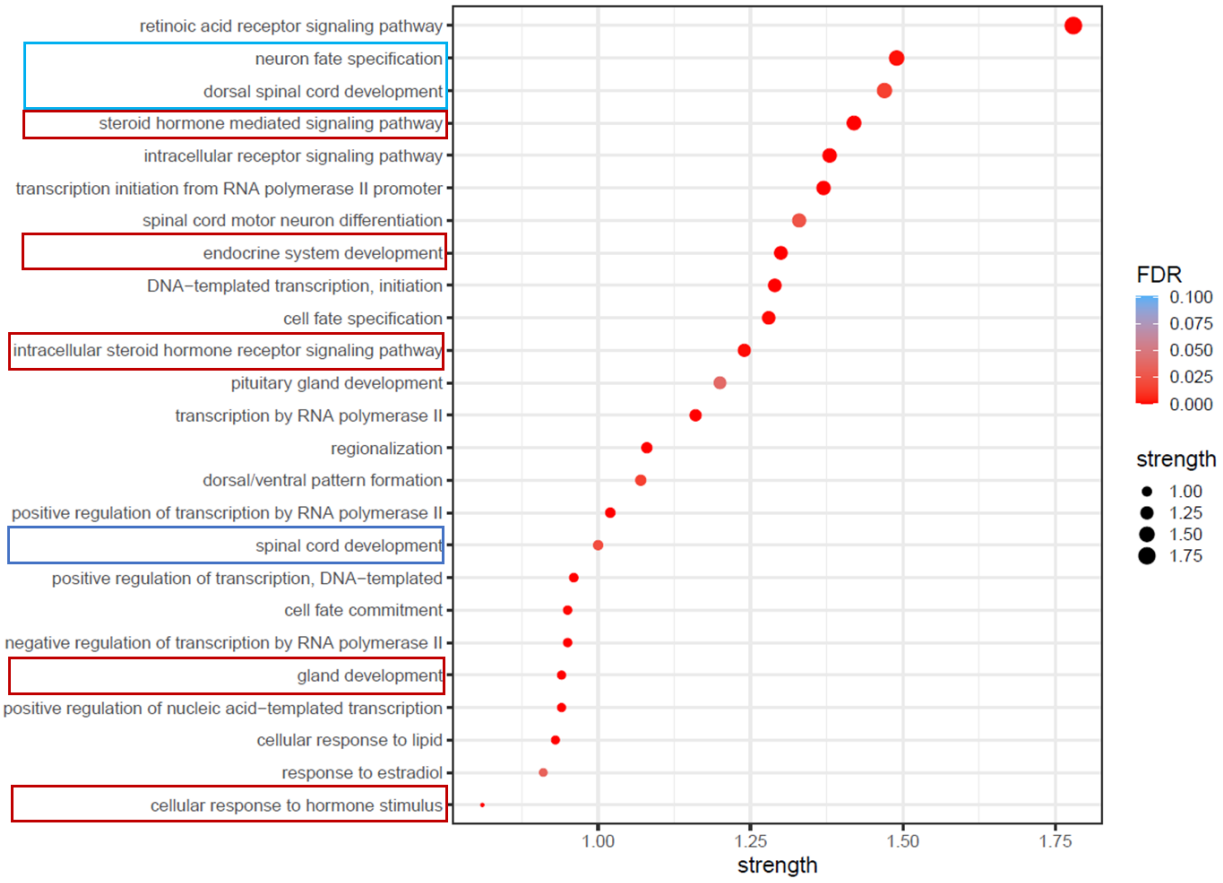


Figura suplementaria 6. Los factores transcripcionales con motivos de unión a la región promotora del lincRNA GATA3-AS1 se relacionan al desarrollo de cáncer de mama. De acuerdo con el análisis de ontología de genes realizado con la Plataforma *STRING*[84], los factores transcripcionales que podrían estar regulando la expresión del lincRNA *GATA3-AS1* se relacionan con procesos biológicos de respuesta hormonal (cuadros rojos) y relacionados con el desarrollo del sistema nerviosos (cuadros azules) que se encuentran enriquecidos en cáncer de mama.

16.4 APÉNDICE D: ANÁLISIS DE INTERACCIONES DEL LINCRNA *GATA3-AS1* CON PROTEÍNAS Y NCRNAs

Tabla suplementaria 3: Evaluación de motivos de unión a proteínas dentro del transcrito del lincRNA *GATA3-AS1*-201

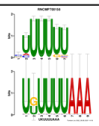
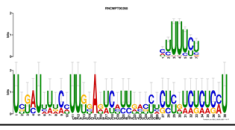
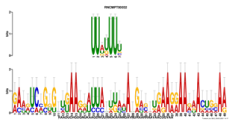
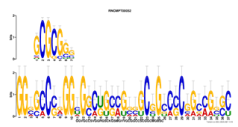
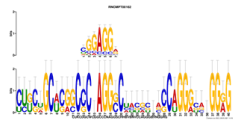
Motivo de unión a RNA	Factor transcripcional	*P value*	LOGO
UKUUUUAAA	CPEB4	0.00601204	
RAADUCHGDGNUKAAGDWUUANUK AHANGAVNURAARRAAVAACUGVAA	HuR	0.00940416	
UBKAUHUDCHUUKBADUCHUUDHBY HCUYDUCUCUCMU	PTBP1	0.00391652	
GGVGCCSVGGHGSCKGSMGVVGCGGC CSCGGCMGBSC	RBM4	0.00780476	
CUKCDGCWCGGCCCNAGGSCHHBVBN VCCAGGGBVNGGRG	LIN28A	0.00958196	

Tabla suplementaria 4: Análisis de predicción de interacciones del lincRNA *GATA3-AS1* con proteínas y factores transcripcionales mediante la herramienta *catRapid*

Proteína	Poder Discriminatorio	Fuerza de interacción	Puntaje de interacción	Función
TERF2	0.54	0.99	0.5831	Ciclo celular
MBD3	0.35	0.98	0.5831	Recluta desacetilasas de histonas y metiltransferasas
CTCF	0.96	0.99	0.5831	Regulación transcripcional
DDX3X	0.2	0.59	0.5117	Helicasa dependiente de RNA. Regulador transcripcional
LEF1	0.81	0.99	0.5117	Activador de la transcripción en presencia de EP300
ESR2	0.97	1	0.5117	Regulación positiva de la transcripción
CDX2	0.35	0.97	0.5117	Unión a DNA metilado

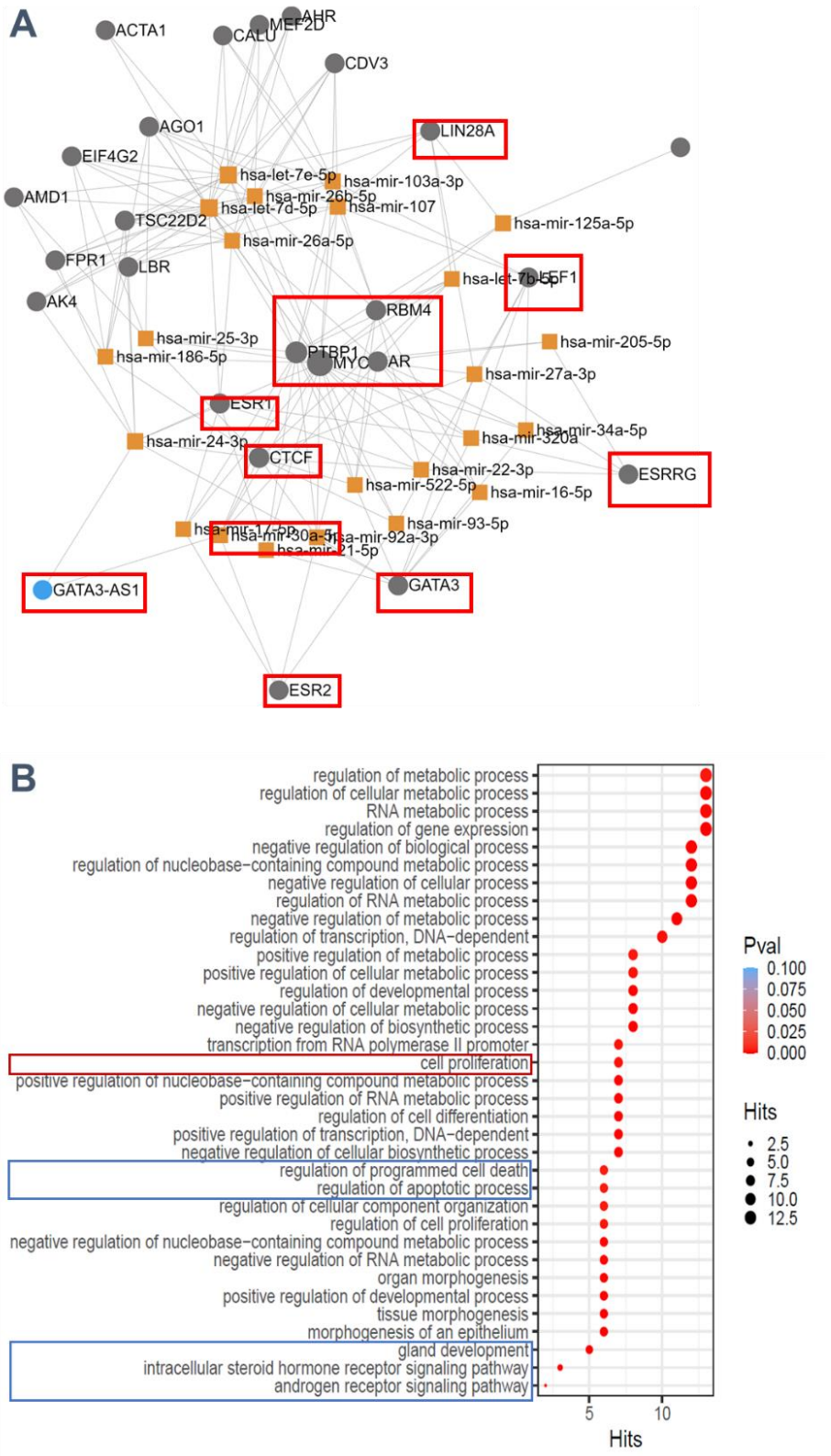


Figura suplementaria 7. Análisis funcional de redes de las proteínas y ncRNAs que interactúan con el lincRNA *GATA3-AS1*. A) El análisis de redes de la predicción de las

interacciones del lncRNA *GATA3-AS1* (nodo azul) con los factores transcripcionales y proteínas (nodos grises), además de los miRNAs (nodos naranjas). Las interacciones fueron seleccionados su valor *p* era menor a 0.05 en la Plataforma *mirNet*[98]. En cuadros rojos se resaltan los factores transcripcionales relevantes. B) El análisis de enriquecimiento de vías de señalización intracelulares tomando en cuenta sólo los genes resaltados en rojo en la figura A, muestra que en conjunto participan en vías de señalización relacionadas con cáncer de mama, como la proliferación celular (cuadro rojo) así como las vías de señalización intracelular de respuesta hormonal (cuadros azules).

16.5 APÉNDICE E: ANÁLISIS DE EXPRESIÓN DIFERENCIAL DEL TRANSCRIPTOMA COMPLETO EN LA LÍNEA CELULAR MCF-7

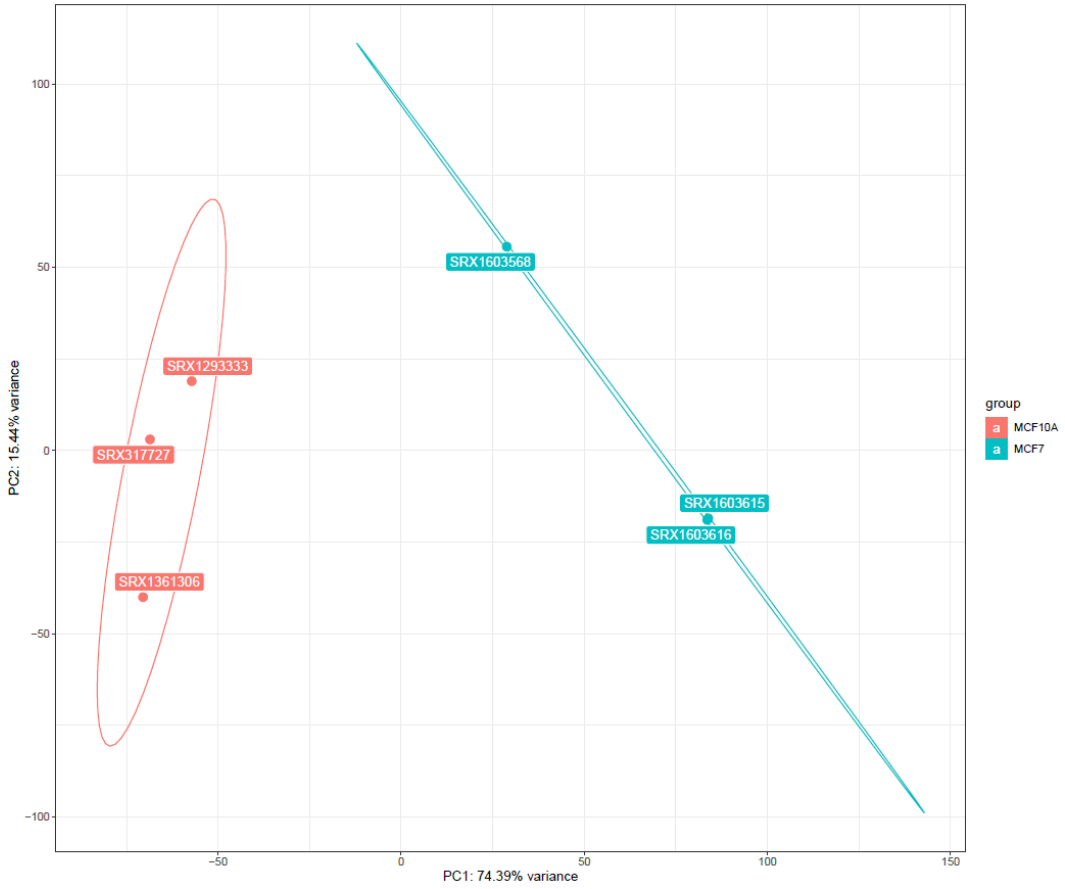


Figura suplementaria 8. Análisis de componentes principales de las réplicas biológicas de los archivos de secuenciación de RNA-Sec de las líneas celulares de cáncer de mama utilizadas en el análisis de expresión diferencial. El análisis de componentes principales muestra que el triplicado biológico de la línea celular MCF-7 (rojo) forma un grupo independiente y diferenciado del triplicado biológico de la línea celular control MCF-10A (azul), de acuerdo con el análisis del transcriptoma completo.

16.6 APÉNDICE F: ESTANDARIZACIÓN DEL EXPERIMENTO DE TRANSFECCIÓN DE ASOs PARA EL ABATIMIENTO DE LA EXPRESIÓN DEL LINCRNA *GATA3-AS1* EN LA LÍNEA CELULAR MCF-7

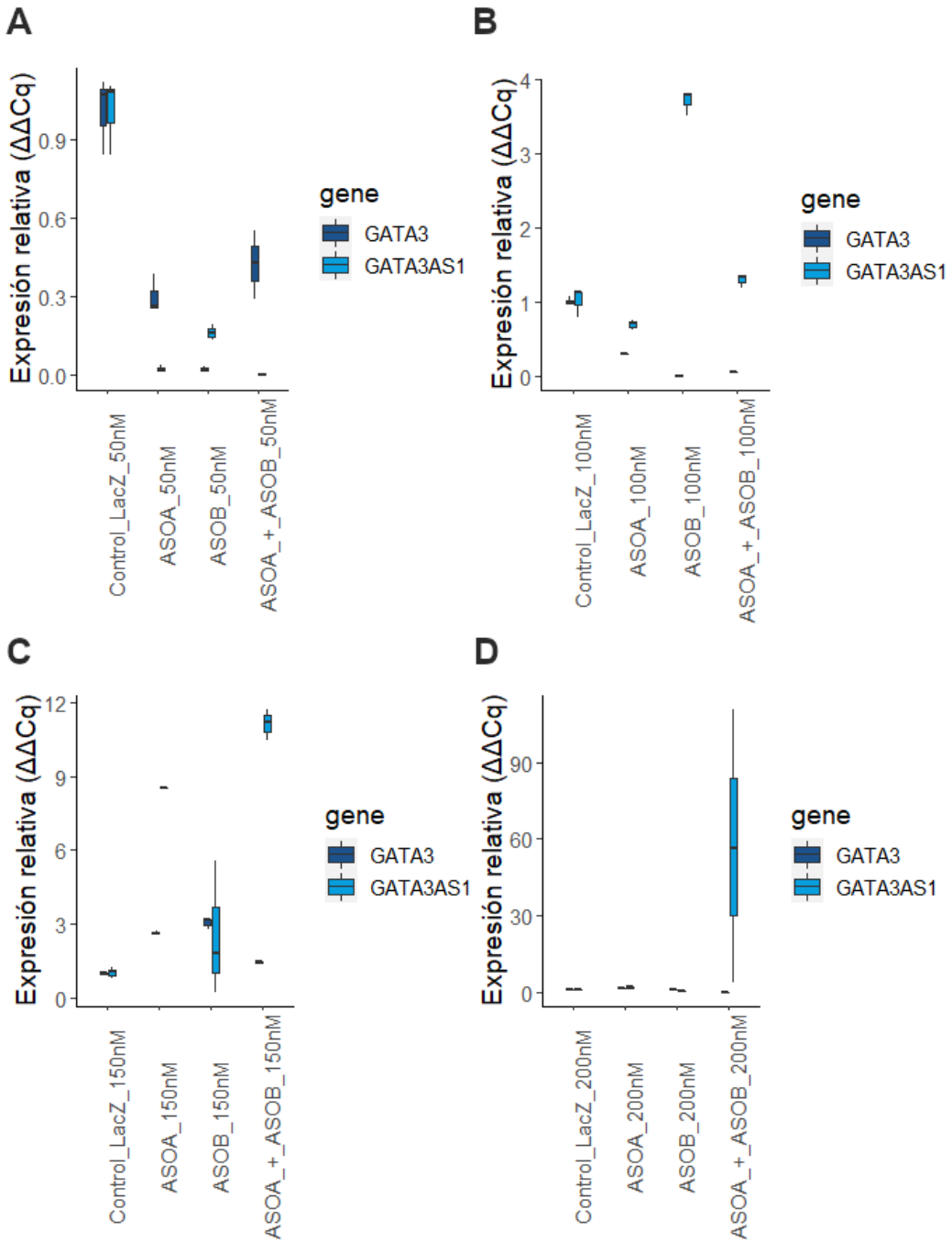


Figura suplementaria 9. Abatimiento de la expresión de *GATA3-AS1* con ASOs. El análisis por RT-qPCR (ΔΔCq) de la expresión relativa de *GATA3-AS1* y *GATA3* respecto

a *RPS28* en la condición control (LacZ), la condición de transfección con el ASO dirigido a *GATA3-AS1* A, B y la combinación A+B para las siguientes concentraciones: A) 50 nM, B) 100 nM, C) 150 nM y D) 200 nM.

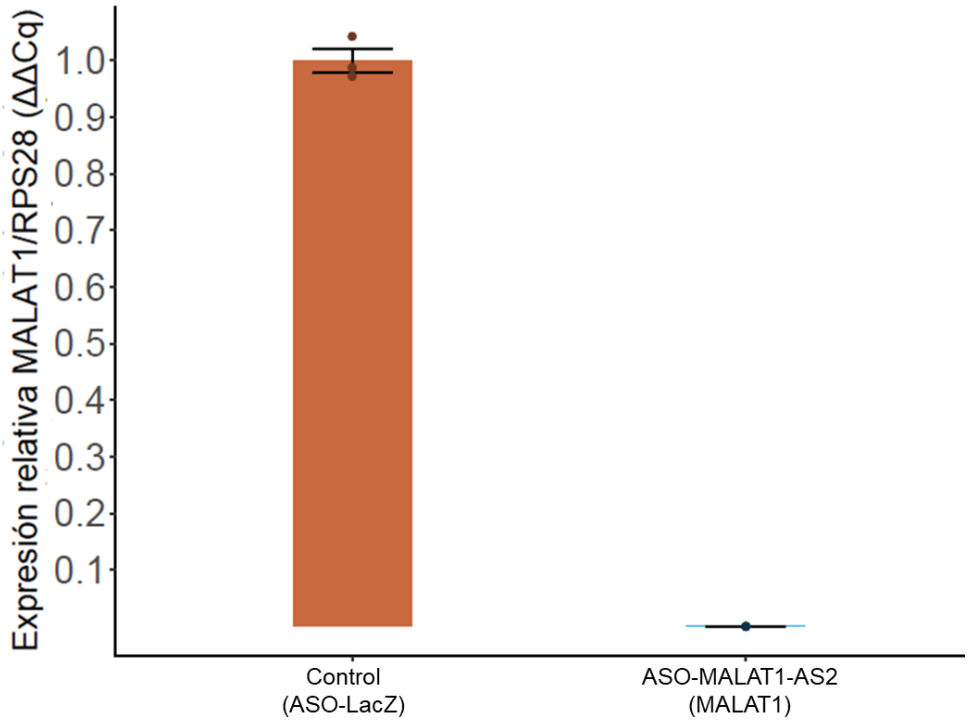


Figura suplementaria 10. Abatimiento de la expresión del gen *MALAT1* con ASOs. El análisis por RT-qPCR (ΔΔCq) de la expresión relativa de *MALAT1* respecto a *RPS28* en la condición control (LacZ) y la condición de transfección con el ASO dirigido a MALAT1 (n=3, triplicado técnico, valor *p* = 0.0002).

16.7 APÉNDICE G: PUBLICACIONES DURANTE LA REALIZACIÓN DE LA MAESTRÍA

The Journal of Molecular Diagnostics, Vol. 23, No. 10, October 2021

ELSEVIER

the Journal of Molecular Diagnostics
jmdjournal.org

Transcriptome Analysis Identifies *GATA3-AS1* as a Long Noncoding RNA Associated with Resistance to Neoadjuvant Chemotherapy in Locally Advanced Breast Cancer Patients

Laura Contreras-Espinosa,* Nicolás Alcaraz,† Inti A. De La Rosa-Velázquez,‡ José Díaz-Chávez,* Paula Cabrera-Galeana,§ Rosa Rebollar-Vega,‡ Nancy Reynoso-Noverón,¶ Héctor A. Maldonado-Martínez,‖ Rodrigo González-Barrios,* Rogelio Montiel-Manríquez,* Diana Bautista-Sánchez,* Clementina Castro-Hernández,* Rosa M. Alvarez-Gomez,** Francisco Jiménez-Trejo,†† Miguel Tapia-Rodríguez,‡‡ José A. García-Gordillo,§ Augusto Pérez-Rosas,* Enrique Bargallo-Rocha,§ Cristian Arriaga-Canon,* and Luis A. Herrera*§§

From the Unidad de Investigación Biomédica en Cáncer, Instituto Nacional de Cancerología—Instituto de Investigaciones Biomédicas, National Autonomous University of Mexico (UNAM), Tlalpan, Mexico City, México; The Bioinformatics Centre,† Department of Biology, University of Copenhagen, Copenhagen, Denmark; the Genomics Laboratory,‡ Red de Apoyo a la Investigación, Universidad Nacional Autónoma de México, Tlalpan, Mexico City, México; the Departamento de Tumores Mamarios,§ Instituto Nacional de Cancerología, Tlalpan, Mexico City, México; the Centro de Investigación en Prevención,¶ the Patología Molecular e Inmunología,‖ and the Clínica de Cáncer Hereditario,** Instituto Nacional de Cancerología, Tlalpan, Mexico City, México; the Instituto Nacional de Pediatría,†† Coyoacán, Mexico City, México; the Instituto de Investigaciones Biomédicas,‡‡ Universidad Nacional Autónoma de México, Tercer Circuito Exterior S/N, Ciudad Universitaria, Mexico City, México; and the Instituto Nacional de Medicina Genómica,§§ Tlalpan, Mexico City, México*

Accepted for publication July 7, 2021.

Address correspondence to Luis A. Herrera, Ph.D., Instituto Nacional de Medicina Genómica. Periférico Sur 4809, Arenal Tepepan, C.P. 14610. Tlalpan, CDMX, México, or Cristian Arriaga-Canon, Ph.D., Instituto Nacional de Cancerología. Av, San Fernando No. 22, Col. Sección XVI, Tlalpan, C.P. 14080, CDMX, México. E-mail: carriagac@incan.edu.mx or lherrera@inmegen.gob.mx.

Breast cancer is one of the leading causes of mortality in women worldwide, and neoadjuvant chemotherapy has emerged as an option for the management of locally advanced breast cancer. Extensive efforts have been made to identify new molecular markers to predict the response to neoadjuvant chemotherapy. Transcripts that do not encode proteins, termed long noncoding RNAs (lncRNAs), have been shown to display abnormal expression profiles in different types of cancer, but their role as biomarkers in response to neoadjuvant chemotherapy has not been extensively studied. Herein, lncRNA expression was profiled using RNA sequencing in biopsies from patients who subsequently showed either response or no response to treatment. *GATA3-AS1* was overexpressed in the nonresponder group and was the most stable feature when performing selection in multiple random forest models. *GATA3-AS1* was experimentally validated by quantitative RT-PCR in an extended group of 68 patients. Expression analysis confirmed that *GATA3-AS1* is overexpressed primarily in patients who were nonresponsive to neoadjuvant chemotherapy, with a sensitivity of 92.9% and a specificity of 75.0%. The statistical model was based on luminal B-like patients and adjusted by menopausal status and phenotype (odds ratio, 37.49; 95% CI, 6.74—208.42; $P = 0.001$); *GATA3-AS1* was established as an independent predictor of response. Thus, lncRNA *GATA3-AS1* is proposed as a potential predictive biomarker of nonresponse to neoadjuvant chemotherapy. *(J Mol Diagn 2021, 23: 1306—1323; https://doi.org/10.1016/j.jmoldx.2021.07.014)*

Supported by the National Cancer Institute of Mexico and Consejo Nacional de Ciencia y Tecnología (CONACYT) grant A3-S-46689.

L.C.-E. is a master student in the Programa de Posgrado en Ciencias Biológicas, UNAM, and received a fellowship from CONACYT with Currículum Vitae Único (CVU)-1003211.

L.C.-E. and N.A. contributed equally to this work.

Disclosures: The authors have submitted a patent at the Mexican Institute of Industrial Property; it is in process with registration key IMPI-00-009, registration number Mx/E/2018/090593, and file Mx/a/2018/015065.

Current address of I.A.D.L.R.-V., Next Generation Sequencing Core Facility, Helmholtz Zentrum Muenchen, Ingolstaedter Landstr, Neuherberg, Germany.

https://doi.org/10.1016/j.jmoldx.2021.07.014

Breast cancer (BC) is one of the main causes of death in women worldwide, with >600,000 deaths annually (*https://gco.iarc.fr/today/home*, last accessed February 25, 2021). Furthermore, as the leading cause of cancer in women, it constitutes a public health burden.[1] In particular, patients with locally advanced breast cancer (LABC) represent a heterogeneous group with variable local recurrence and global survival.[2–4] These patients have a significant risk for local recurrence and metastatic progression, in addition to presenting low rates of global survival compared with patients with BC at early stages (stage I to IIA). Hence, a comprehensive approach to LABC patients to achieve local and distant control of disease has become a challenge, as well as monitoring disease progression and treatment efficacy.[5]

On the other hand, pathologic complete response (pCR) is one of the most important parameters to consider in patient prognosis. However, different studies have shown that <50% of patients with LABC achieve pCR after neoadjuvant chemotherapy (NAC).[6] NAC was initially used in the context of LABC because it has several advantages, such as making inoperable tumors surgically resectable (stage T4, N2, or N3) and increasing the rates of breast-conserving surgery.[6] Although some biomarkers have proved useful for improving treatment efficiency, most are still in the clinical testing stage and have not yet been approved for standardized use.[7–9] Expression status of estrogen receptor (ER), progesterone receptor, and *ERBB2* [human epidermal growth factor receptor 2 (HER2)] by immunohistochemical evaluation is currently the gold standard for determining management and response to treatment.[10] In addition, gene expression panels, such as Oncotype Dx and Mammaprint, are examples of sets of biomarkers used to provide further clinical support by predicting response to chemotherapy.[11] Nevertheless, these types of tests have mostly been restricted to adjuvant chemotherapy, whereas only a few biomarkers, such as Ki-67,[12–14] have been proposed for response to NAC.

Recently, in addition to gene expression of coding genes, it has been proposed that noncoding transcripts, such as long noncoding RNAs (lncRNAs), may also serve as molecular markers for BC diagnosis and prognosis.[15–17] These transcripts are defined as having >200 bases and lacking open reading frames, making them unable to be translated into proteins.[18]

Therefore, transcriptome analysis by *ab initio* assembly established the potential importance of lncRNAs in cancer, suggesting that this kind of noncoding transcript could be useful in cancer pathogenesis and biomarker development.[19] The lncRNA *HOTAIR* has been proposed as a potential prognostic biomarker, and its overexpression was associated with metastasis-free survival and overall survival (OS), suggesting that this noncoding transcript is a powerful predictor of metastasis and death.[16] Another example is the ER-regulated lncRNA *DSCAM-AS1*, which is overexpressed in ER-positive tumors and was shown to be of clinical relevance as a good predictor of tumor progression and tamoxifen resistance.[20] In another comprehensive analysis of RNA sequencing (RNA-Seq) data from The Cancer Genome Atlas, Berger et al[15] constructed gene correlation networks and detected significant gene-lncRNA interactions in breast cancer between coding genes (*ESR1* and *DKC1*) and lncRNAs (*NEAT1*, *TUG1*, and *TERC*). Despite their potential, to date, few studies have investigated or reported lncRNAs as biomarkers of response to systemic therapies in BC[21] and specifically to NAC in BC.[22–24] Therefore, focusing on lncRNAs may aid in identifying novel and more accurate biomarkers for predicting the response to systemic neoadjuvant therapy in BC.

In this study, using RNA-Seq profiling and machine learning, first, a group of lncRNAs was identified; these lncRNAs are differentially expressed in NAC-resistant LABC Mexican patients compared with NAC-sensitive patients (nonresponders and responders, respectively) and were also predictive and stable features in random forest model. In particular, *GATA3-AS1* was identified as a divergent lncRNA that acts as a predictive biomarker of the response to NAC. Expression profiling by RT-qPCR on a larger cohort confirmed that *GATA3-AS1* is overexpressed only in nonresponder patients ($n = 68$). In addition, univariate and multivariate analyses established that *GATA3-AS1* distinguishes between nonresponder and responder patients with a sensitivity of 92.9%, a specificity of 75.0%, and an area under the curve (AUC) of approximately 0.90. Finally, *GATA3-AS1* is suggested as a novel biomarker for predicting NAC response in patients with LABC and provides the first evidence of this lncRNA as a prediction biomarker for NAC response in breast cancer patients with positive hormonal receptors that correspond to luminal B-like HER2-positive and HER2-negative phenotypes.

Materials and Methods

Breast Sample Collection

Eleven RNA samples, obtained from biopsies of female patients diagnosed with locally advanced mammary adenocarcinoma (stage IIB to IIIC) belonging to the Mexican National Cancer Institute population who were candidates for the administration of neoadjuvant chemotherapy, were sequenced by RNA-Seq. A validation cohort was collected from another 68 biopsies diagnosed with primary breast tumors, all of whom were patients at the Breast Tumor Division in the Mexican National Cancer Institute between January 2012 and December 2015. Samples were collected from tissue of the initial biopsy (taken with a thick needle) before treatment began (Figure 1). All patients were previously confirmed with primary breast carcinoma and locally advanced disease (clinical stage IIA to IIIC) by histologic studies. Patient selection was performed with the support of the breast tumor department at the Mexican National Cancer Institute, which includes oncologists and pathologists.

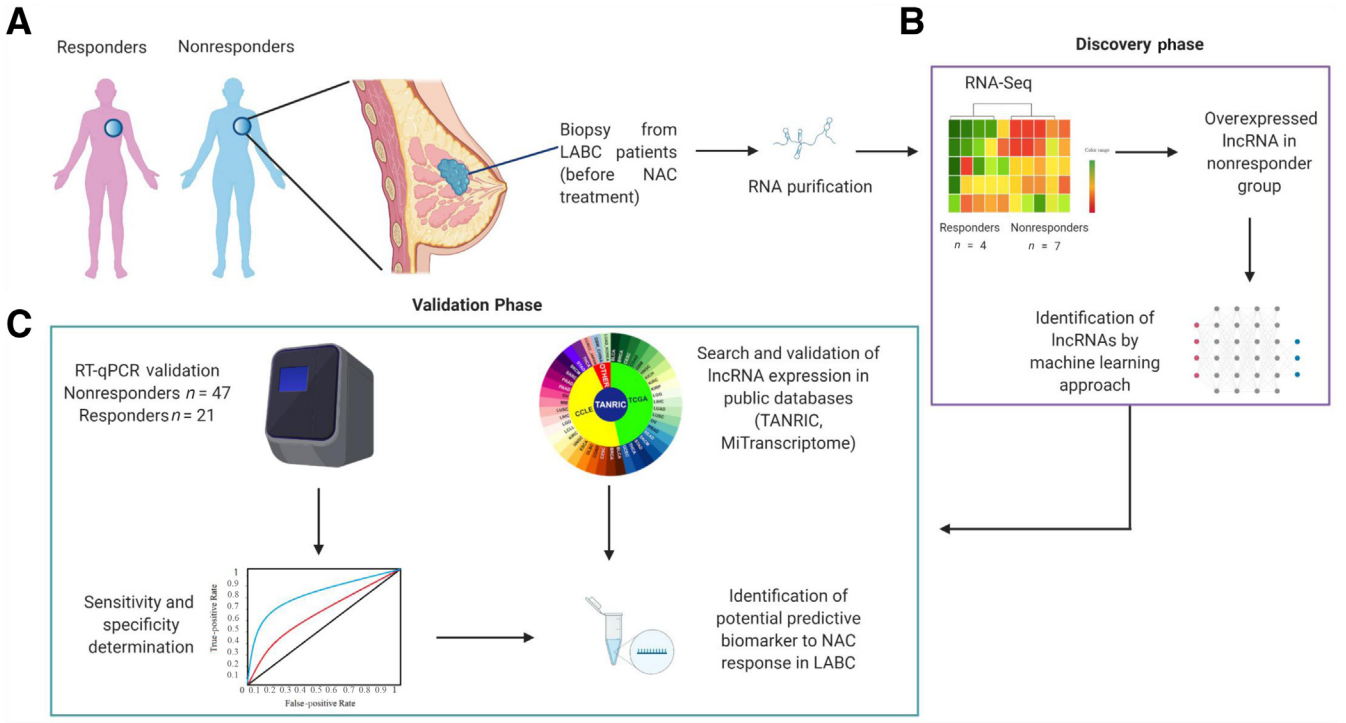


Figure 1 Experimental research design. **A:** The study was divided in two phases, discovery and validation phases, with snap-frozen pretreatment core needle biopsies obtained from primary breast cancer patients who responded to chemotherapy and from those who did not respond to treatment (responder and nonresponder patients, respectively). **B:** From them, 11 patients participated in the discovery phase, and the RNA from their samples was taken to construct a poly-A library for a paired-end RNA sequencing (*Materials and Methods*). After bioinformatic analysis of sequencing data by implementing random forest algorithm of machine learning approach, the differentially expressed genes, especially overexpressed long noncoding RNAs (lncRNAs), were identified and selected for potential prediction biomarkers. **C:** Among them, lncRNAs were identified and selected for validation phase by RT-qPCR analysis and were also validated in public databases TANRIC (*https://www.tanric.org*, last accessed February 25, 2021) and The Cancer Genome Atlas (*https://portal.gdc.cancer.gov*, last accessed February 25, 2021). Image was generated with BioRender.com. LABC, locally advanced breast cancer; NAC, neoadjuvant chemotherapy; RNA-Seq, RNA sequencing.

All patients included in this study received systemic NAC according to the recommendations of the National Comprehensive Cancer Network guidelines. Chemotherapy regimens were based on anthracyclines and taxanes in sequential scheduling, as described below: four taxane cycles every 21 days (paclitaxel, 80 mg/m^2, on days 1, 8, and 14; or docetaxel, 100 mg/m^2, on day 1) followed by four doses of fluorouracil-Adriamycin-cyclophosphamide every 21 days (fluorouracil, 500 mg/m^2, Adriamycin, 50 mg/m^2, and cyclophosphamide, 500 mg/m^2, on day 1). Subsequently, all patients underwent local control by mastectomy, breast conservation surgery mastectomy, or breast conservation surgery. At the end of the neoadjuvant regimen, treatment response was evaluated in surgical specimens obtained by the oncopathologist. A pCR was classified according to the most accepted definition: the absence of infiltrating components in the breast and lymph nodes (ypT0/is or ypN0). Luminal B-like phenotype was assigned according to the expression of estrogen and progesterone receptors (>1%), in absence or presence of HER2 overexpression. If HER2 was not overexpressed, it was considered for subtyping the expression of Ki-67 (≥20%).[25] In this study, patients were classified as responders if they presented pCR and nonresponders if they did not show pCR. Informed consent was obtained, and this study was approved by the ethical and research committee of the Mexican National Cancer Institute (018/055/DII CEI/1302/18).

External RNA-Seq Data Collection of Breast Cancer Cell Lines

RNA-Seq results of the breast cancer cell lines were obtained from the data set contained in the Breast Cancer Profiling Project, Gene Expression 1: *baseline mRNA sequencing on 35 cell lines*, which forms part of the Library of Integrated Network-Based Cellular Signatures, which includes 33 breast cancer cell lines and 2 transformed noncancerous breast cell lines (*http://lincs.hms.harvard.edu/db/datasets*, last accessed May 14, 2020).

RNA-Seq Data Analysis

The quality of the sequencing files was determined with reports generated using FastQC[26,27] version 0.11.9. Filtering of low-quality reads and adapter removal were performed with trimmomatic version 0.39.[28] Reads were mapped to the human reference genome assembly version hg38 using

STAR aligner version 2.7.1a.[29] Gene expression quantification was performed with featureCounts from the rsubread package version 1.34.7 on the STAR bam files. Gene expression quantification was also performed by aligning to the transcriptome with Salmon version 0.14.1.[30] All gene quantification was performed on gencode version 31 annotations. Tumor purity was inferred using the ESTIMATE algorithm[31] version 2.0.

Differential expression analysis was performed with DEseq2[32] version 1.24 and was subsequently filtered to analyze only the subset of lncRNAs. The cutoff points were a log2 fold change (FC) value of 1.5 for overexpressed lncRNAs and −1.5 for underexpressed lncRNAs (false discovery rate < 0.05). To corroborate grouping of the patient tumor samples, principal component analysis was performed on the basis of the expression profile of the whole transcriptome, where a batch effect for ER status was detected and added as a covariate in DEseq2 design (∼ER_status + response).

Random forest models were trained on the log-transformed normalized counts (rlog function with blind = FALSE parameter) to predict patient response, using the package randomForest[33] version 4.6.14 in a leave-two-out cross-validation scheme, where two samples (one responsive and one nonresponsive) were always left out. The mean decrease in accuracy was stored for each gene and each fold. To assess the stability of the feature importance, the quartile coefficient of dispersion of the mean decrease in accuracy over all folds was computed for each gene: $(Q_3 - Q_1)/(Q_3 + Q_1)$, where Q_3 and Q_1 are the third and first quartiles, respectively, of their mean decrease in accuracy.

ConsensusPathDB[34] was performed using clusterProfiler.[35] The Bioconductor package on all genes identified differential expression (DE) in at least one of the two quantification methods using all expressed genes as background. Gene set enrichment analysis[36] was performed on all expressed genes ordered by their DE −log10 *P* value from DEseq2. Pathway enrichment of lncRNAs was performed using the LncPath (LncRNAs2Pathways[37]) package version 1.1.

RNA Isolation and Quantitative Real-Time PCR Assays

MCF-10A, MCF-7, BT474, and MDA-MB-231 cell cultures were performed following respective ATCC (Manassas, VA) culture protocols. Total RNA was isolated from cultured cells using TRIzol (Thermo Fisher Scientific, Waltham, MA), according to the manufacturer's instructions. For patient samples, RNA was isolated using the AllPrep kit (Qiagen, Germantown, MD; number 80204); RNA concentration and quality analysis (RNA integrity number value) were performed by a Tape Station 2200 bioanalyzer (Agilent Technologies, Santa Clara, CA). Then, 1 μg RNA was treated with DNase I, RNase free (molecular biology; Thermo Fisher Scientific; reference EN0525). cDNA was synthesized from 1 μg total RNA using a High Capacity cDNA Reverse Transcription (Applied Biosystems, Thermo Fisher Scientific; reference 4368814). Finally, real-time PCR was performed with SYBR Green/ROX qPCR Master Mix (molecular biology; Thermo Fisher Scientific; reference K0223) on a QuantStudio 3 Real-Time PCR System (Applied Biosystems, Thermo Fisher Scientific). Relative gene expression was determined by fold change calculation (ΔΔCt) for cell lines and ΔCt for patient samples, normalized to *RPS28* as a housekeeping gene. The primers used are listed in Table 1.

RNA-Seq

To ensure good quality of the samples for sequencing, Nanodrop, Qubit 2.0 (Life Technologies, Carlsbad, CA), and the Agilent 2100 Bioanalyzer (Agilent Technologies) were used to detect the purity, concentration, and integrity of RNA samples, respectively. All samples had RNA integrity numbers >8.0. A total of 1 μg RNA was used to generate sequencing libraries using the TruSeq Stranded mRNA library prep kit from Illumina, Inc. (San Diego, CA), according to the manufacturer's instructions. After construction of the libraries, their concentrations and insert sizes (approximately 260 bp) were detected using Qubit 2.0 and the Agilent 2100 Bioanalyzer, respectively. The library was then sequenced using an Illumina HiSeq 2500 sequencer with paired-end 2 × 125 cycles using Illumina TruSeq version 4 sequencing by synthesis (SBS) chemistry and following the manufacturer's instructions. The depth of sequencing was >25 million reads using a HiSeq2500 sequencer by Illumina. RNA-Seq data are available from National Center for Biotechnology Information Gene Expression Omnibus (*https://www.ncbi.nlm.nih.gov/geo*, accession number GSE159448).

Immunohistochemistry

Primary antibody for GATA3 protein was obtained from BioSB (mouse monoclonal; clone EP368; catalog number BSB 3333; BioSB, Santa Barbara, CA). The reaction was performed on a Benchmark Ultra automated immunohistochemistry device (Ventana Medical Systems, Tucson, AZ). Briefly, the slides were deparaffinized at 72°C for 4 minutes on xylene-free dewaxing reagent (EZprep; Ventana Medical Systems). Antigen retrieval followed, for 8 minutes at 95°C in alkaline solution (CC1; Ventana Medical Systems). Primary antibody was incubated for 16 minutes at 36°C, diluted 1:50 (DaVinci Green diluent; catalog number PD900; Biocare Medical, Pacheco, CA). Finally, enhanced polymer-based detection with diaminobenzidine chromogen was employed (Optiview DAB Kit; Ventana Medical Systems). In between all steps, thorough washings were made (Reaction Buffer; Ventana Medical Systems). The slides were then manually counterstained with Harris hematoxylin and mounted with nonaqueous medium.

Statistical Analysis

Descriptive statistics of the main demographic and clinical variables were performed, presenting the median and

Table 1 Primers Used to Detect GATA3-AS1 Expression by RT-qPCR

Sequence 5′-3′	Bases, *n*	Amplicon, bp	Symbol
5′-CGCAGACAGAAAAGAAGCCG-3′	20	133	GATA3-AS1 F
5′-GCTGGAATGGGAAGGGACTT-3′	20		GATA3-AS1 R
5′-CGATCCATCATCCGCAATG-3′	19	101	RPS28 F
5′-AGCCAAGCTCAGCGCAAC-3′	18		RPS28 R

F, forward; GATA3-AS1, GATA3 antisense RNA 1; RT-qPCR, quantitative RT-PCR; R, reverse; RPS28, ribosomal protein S28.

interquartile range of continuous variables and the proportion of qualitative variables. Differences in the variables collected in the study groups (response versus no response) were identified according to the type of variable with the *U*-test or the χ^2 test. Identification of the effect of overexpression or down-regulation of *GATA3-AS1* on clinical response was adjusted to the main clinical variables and was performed by statistical analysis of 66% with the sample available. A $P < 0.05$ was considered statistically significant. Analysis of variance, followed by the Tukey test, was performed to determine significant differences in *GATA3-AS1* expression among different cell lines, in RNA-Seq results analysis, and in quantitative real-time PCR relative expression analysis. For clinical data, *GATA3-AS1* expression differences between responder and nonresponder groups were determined by *t*-test (two tailed, nonpaired, 95% confidence value), assuming variance homogeneity (determined by the Levene test, $P = 0.06$). This was verified by the Fisher test and χ^2 test ($P < 0.05$). In addition, because of sample size, normal distribution was determined by Lilliefors, Kolmogorov-Smirnov, and Shapiro-Wilk tests, indicating that the sample did not show normal distribution ($P > 0.05$). For that reason, a nonparametric *U*-test/Wilcoxon test was implemented, showing differences between medians ($P < 0.05$).

For The Atlas of Noncoding RNAs in Cancer (TANRIC, *https://www.tanric.org*, last accessed February 25, 2021) survival plot analysis, Cox regression was implemented to determine survival time, and the log-rank test was used to compare survival distribution. Significance was defined as $P < 0.05$. To determine the specificity and sensitivity of *GATA3-AS1* overexpression in neoadjuvant resistance prediction, receiver operating characteristic curve analysis was performed. For this analysis, the STATA software version 14 (StataCorp, College Station, TX) was used.

Results

Clinicopathologic Characteristics of Locally Advanced Breast Cancer Patients

This study was primarily focused on identifying biomarkers to predict the response to NAC in patients with LABC within luminal B-like phenotype. The main features that define each patient are the molecular subtype, age, clinical stage, and response to treatment, among others (Table 2). Expression of hormonal receptors corresponds to histologic classification of the molecular luminal B-like HER2-positive and HER2-negative phenotype (*Materials and Methods*).

Transcriptome Profiling of LncRNAs in Breast Cancer Patients Is Associated with Nonresponder Patients

With the objective of detecting lncRNAs that could potentially serve as molecular biomarkers of predicted response to NAC, transcriptome profiling was performed in a poly A−enriched population of RNAs using RNA-Seq. The discovery phase included snap-frozen pretreatment core needle biopsies from primary breast cancer patients who responded to NAC ($n = 4$) and from those who did not respond to the system treatment ($n = 7$), referred to as responder and nonresponder patients, respectively. Principal component analysis considering the complete expression profile, including mRNAs and lncRNAs, showed no distinction between groups; however, grouping based on ER status was observed and adjusted for these in subsequent analyses (Supplemental Figure S1). Next, DE analysis was performed comparing nonresponder patients with responder patients and identified 69 lncRNAs that were underexpressed and 10 lncRNAs that were overexpressed, according to the established cutoff point ($|\log2FC| > 1.5$ and false discovery rate ≤ 0.05) (Figure 2A). Unsupervised hierarchical clustering of patients shows that these DE lncRNAs place responders and nonresponders into two well-defined groups (Figure 2B).

Given that most of the lncRNAs identified in this study are synthesized from the antisense strand with limited information about their biological functions or molecular characterization, to investigate their association with different cell types, FARNA server was used (*https://www.cbrc.kaust.edu.sa/farna*, last accessed February 25, 2021), which infers the function of noncoding RNAs based on the function of their coexpressed genes in multiple data sets (Supplemental Table S1). Results demonstrated that most lncRNAs overexpressed in the nonresponder group participate in processes, such as chromatin remodeling, miRNA interactions, and cancer progression. In contrast, most underexpressed lncRNAs were related to apoptosis and interacting competing endogenous RNA networks. Derived from the fact that most lncRNAs found in this study had not been previously reported or functionally characterized, the study was focused on *GATA3-AS1* (FC, 3.02), first because this lncRNA is the only lncRNA that was overexpressed in all nonresponder

Table 2 Clinicopathologic Characteristics of the Breast Cancer Patients Analyzed

Characteristics	Nonresponders ($N = 47$)	Responders ($N = 21$)	P value
Age, P50 (P25-P75), years	49 (43—54)	54 (45.5—50.5)	0.118
Menopause			
No	34 (72.34)	9 (42.86)	0.020*
Yes	13 (27.66)	12 (57.14)	
Clinical stage			
II	3 (6.38)	3 (15)	
III	43 (91.49)	17 (85)	0.437
IV	1 (2.13)	0 (0)	
Histology			
IDC	44 (93.61)	21 (100)	
ILC	3 (6.38)	0 (0)	0.444
Grade			
Low	0 (0)	0 (0)	
Intermediate	5 (10.64)	3 (14.29)	
High	42 (89.36)	18 (85.71)	0.662
Phenotype			
Luminal B like/HER2$^-$	35 (74.47)	10 (47.62)	
Luminal B like/HER2$^+$	12 (25.53)	11 (52.38)	0.031*
GATA3-AS1 expression			
No	3 (6.4)	14 (66.7)	
Yes	44 (93.6)	7 (33.3)	<0.001*
Status			
Alive	38 (80.85)	20 (95.24)	
Dead	9 (19.15)	1 (4.76)	0.122

Data are given as *n* (%), unless otherwise indicated.
*Statistically significant.
HER2, human epidermal growth factor receptor 2; IDC, infiltrating ductal carcinoma; ILC, infiltrating lobular carcinoma; P, percentile.

patients (Figure 2B), and second, it was the only lncRNA that had been previously well characterized in lymphocytes.[38] In addition, this lncRNA is near GATA3, an important gene in breast cancer.[39] Interestingly, in RNA-Seq differential expression analyses between responder and nonresponder patients, GATA3 is not included in the significantly overexpressed genes in nonresponder patients (Supplemental Table S2).

Overexpressed and Underexpressed LncRNAs, as well as Machine Learning Analysis, Define Patients Who Are Nonresponders to NAC

RNA sequencing is a robust tool for measuring all transcripts, especially when used for identifying differentially expressed genes or noncoding RNA genes, such as lncRNAs, between sample groups.[40] Depending on the pipelines used in RNA-Seq analysis, an incorrect estimate of transcript abundance can be obtained, indicating that differences between pipelines contribute to overall uncertainty in estimates of transcript abundance.[41] Given this premise, the decision to use two different pipelines to analyze RNA sequencing data was taken. Salmon + DESeq2 and STAR + FeaturesCounts + DESeq2 were used to identify differential expression of lncRNAs. With the Salmon pipeline, 70 underexpressed lncRNAs were obtained compared with 64 lncRNAs with the STAR pipeline, and 40 lncRNAs appeared in both methods (Figure 3A and Supplemental Table S3). For overexpressed lncRNAs, when the Salmon pipeline was applied, 10 lncRNAs were found compared with 6 when the STAR pipeline was used, with an overlap of 4 lncRNAs between the two methods (Figure 3B and Supplemental Table S4). Interestingly, one of the overexpressed lncRNAs that coincided in the two pipelines was *GATA3-AS1*, the same lncRNA overexpressed in all nonresponder patients.

Furthermore, to identify lncRNAs associated with response to NAC, random forest models were built on subsets of patient samples by leaving two patients out (*Material and Methods*), and the mean decrease in accuracy of each variable was recorded for each model (Figure 3C). The top 50 genes sorted by median mean decrease in accuracy were selected, where only 6 of these were also found to be differentially expressed. This highlights the importance of considering nonlinear predictive models for biomarker discovery in complex and heterogeneous diseases, such as cancer. To assess the stability of the feature importance in the random forest models, the quartile coefficient of dispersion was calculated (Figure 3C). It was observed that *GATA3-AS1* exhibited the most stable importance across the different models. Given its high importance scores and stability across the random forest models and its detection as DE, *GATA3-AS1* was selected as a top candidate for further validation.

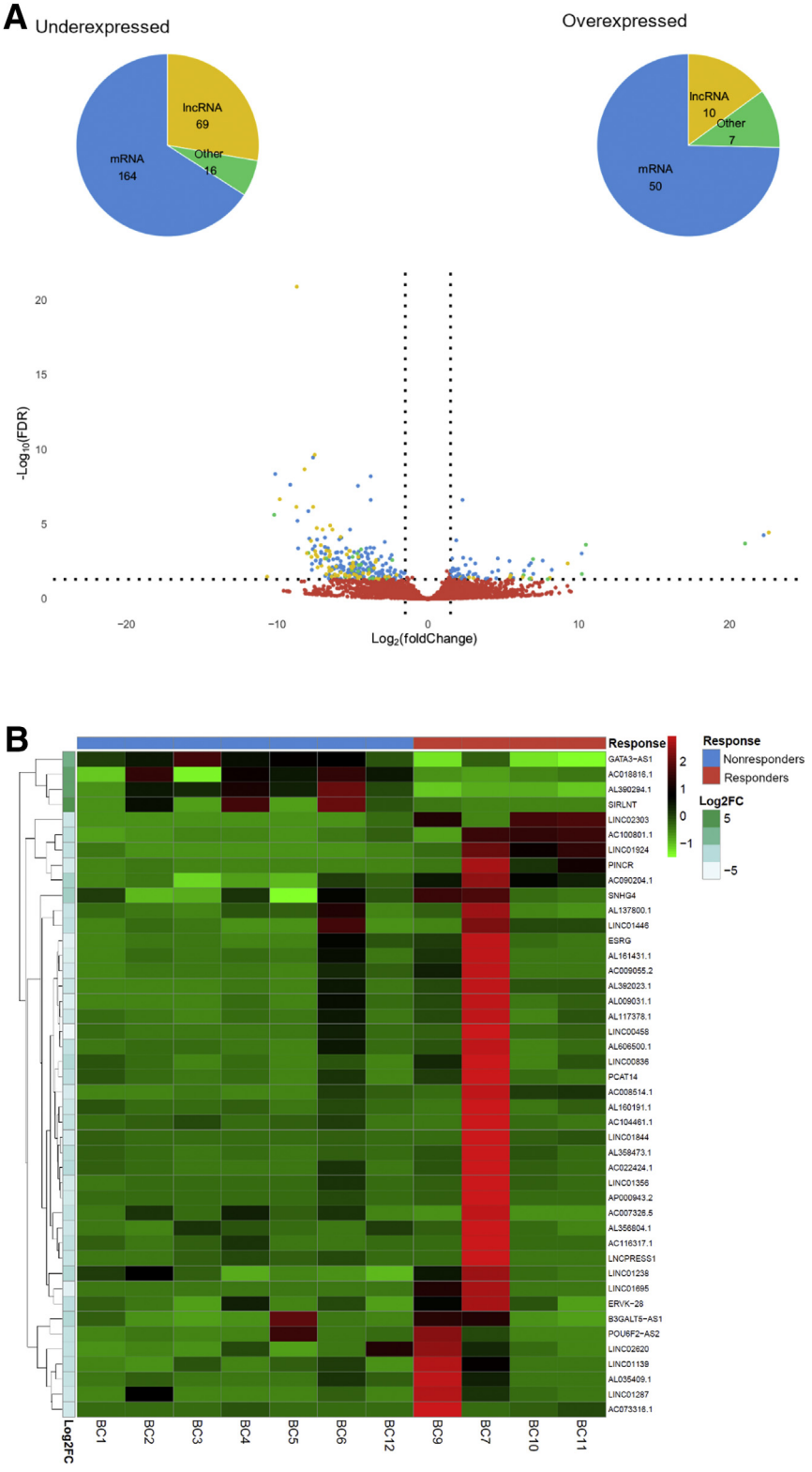


Figure 2 Transcriptome profiling of long noncoding RNAs (lncRNAs) in breast cancer (BC) patients is associated with nonresponder patients. **A:** Pie charts showing the proportion of underexpressed (**left panel**) and overexpressed (**right panel**) RNA biotypes in the nonresponder group. mRNA, lncRNA, and other RNA biotypes are indicated in blue, yellow, and green, respectively. Below is a volcano plot of the identified differentially expressed lncRNAs in nonresponder patients. Blue, yellow, and green dots correspond to mRNA, lncRNA, and other RNAs, respectively [false discovery rate (FDR) < 0.05, log2 fold change > 1.5 for up-regulation and < -1.5 for down-regulation]. Red dots correspond to RNA with no significant changes in the nonresponder group. **B:** Heat map of hierarchical clustering analysis of the top 44 differentially expressed lncRNAs between responder and nonresponder patients. Rows and columns represent differentially expressed lncRNAs and tissue samples, respectively. The color scale represents expression levels. Red and green colors represent up-regulated and down-regulated lncRNAs, respectively (FDR < 0.05). $n = 11$ (**B**). FC, fold change.

GATA3-AS1 Is Overexpressed in Breast Cancer Cell Lines and in Patients Resistant to NAC

To corroborate the expression profile of *GATA3-AS1*, RNA-Seq data from breast cancer patients and breast cancer cell lines were used. For this purpose, expression levels of *GATA3-AS1* were evaluated from RNA-Seq data in patients derived from the discovery phase. It was found that in samples from nonresponder patients, *GATA3-AS1* was overexpressed compared with responder patients, showing basal expression

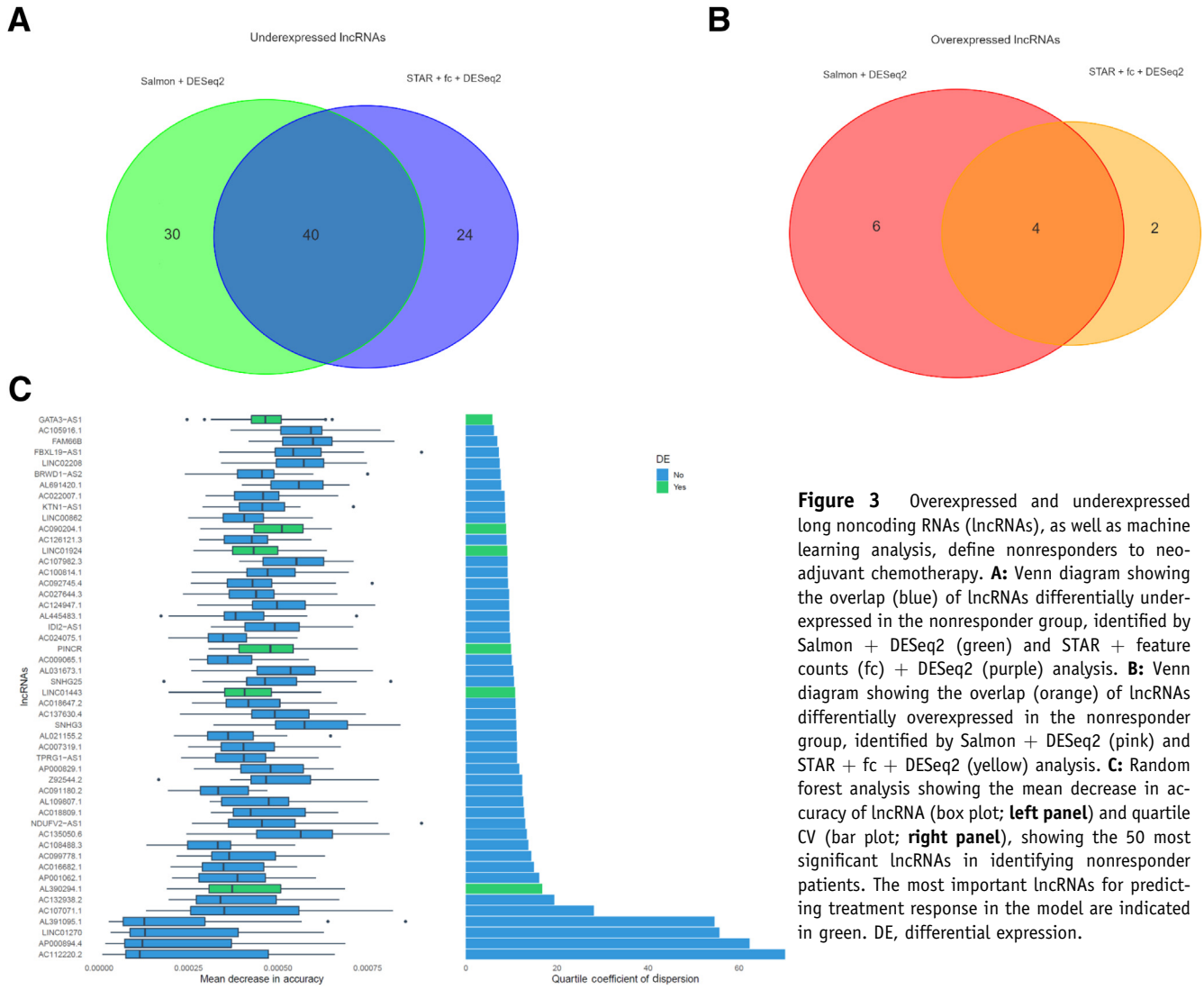


Figure 3 Overexpressed and underexpressed long noncoding RNAs (lncRNAs), as well as machine learning analysis, define nonresponders to neoadjuvant chemotherapy. **A:** Venn diagram showing the overlap (blue) of lncRNAs differentially underexpressed in the nonresponder group, identified by Salmon + DESeq2 (green) and STAR + feature counts (fc) + DESeq2 (purple) analysis. **B:** Venn diagram showing the overlap (orange) of lncRNAs differentially overexpressed in the nonresponder group, identified by Salmon + DESeq2 (pink) and STAR + fc + DESeq2 (yellow) analysis. **C:** Random forest analysis showing the mean decrease in accuracy of lncRNA (box plot; **left panel**) and quartile CV (bar plot; **right panel**), showing the 50 most significant lncRNAs in identifying nonresponder patients. The most important lncRNAs for predicting treatment response in the model are indicated in green. DE, differential expression.

when visualized by histograms (Figure 4A). Even in the expression analysis from RNA-Seq data quantified by transcript per million, *GATA3-AS1* was overexpressed only in nonresponder patients (Figure 4B), suggesting its importance as a prediction biomarker of response to NAC in LABC patients with a luminal B-like phenotype.

In addition, to evaluate oncogenic potential and tissue specificity of *GATA3-AS1*, expression validation in breast cancer cell lines was needed. RNA-Seq data from the Library of Integrated Network-Based Cellular Signatures (*http://lincs.hms.harvard.edu/db/datasets*, last accessed May 14, 2020) were used; this library includes 33 breast cancer cell lines and 2 transformed noncancerous breast cell lines (see *Materials and Methods*). RNA-Seq histograms established that several breast cancer cell lines overexpressing *GATA3-AS1* and MCF10A, a nontumorigenic epithelial cell line, exhibited basal expression levels of *GATA3-AS1* (Figure 4C). Interestingly, in bar plots where *GATA3-AS1* expression levels were measured by reads per kilobase of transcript per million reads mapped (RPKM) (Figure 4D), neoplastic cell lines, such as MDA-MB-157, MDA-MB-436, HCC1395, and CAL51, among others, showed basal expression levels, such as the noncancerous breast cell lines MCF10A and HME1. However, in the other 27 breast cancer cell lines analyzed, *GATA3-AS1* was overexpressed; in particular, MCF7 and T47D breast cancer cell lines showed the highest expression levels of this lncRNA, suggesting that GATA3-AS1 is a tissue- and stage-specific overexpressed lncRNA. Regarding *GATA3*, the adjacent coding gene, its expression was similar between responder and nonresponder patients (Supplemental Figure S2, A and B). Also, *GATA3* was widely expressed among the analyzed breast cancer cell lines (Supplemental Figure S2, C and D). Similarly, *GATA3* expression was widespread in the different cancer types analyzed (Supplemental Figure S3). On the contrary, analysis of RNA-Seq data from MiTranscriptome (*http://mitranscriptome.org*, last accessed February 25, 2021) demonstrated that *GATA3-AS1* is only overexpressed in breast and bladder cancer tissues, establishing a highly cancer-specific expression pattern and highlighting its importance in breast cancer (Supplemental Figure S4). In addition, this lncRNA is overexpressed in positive hormonal receptor phenotypes

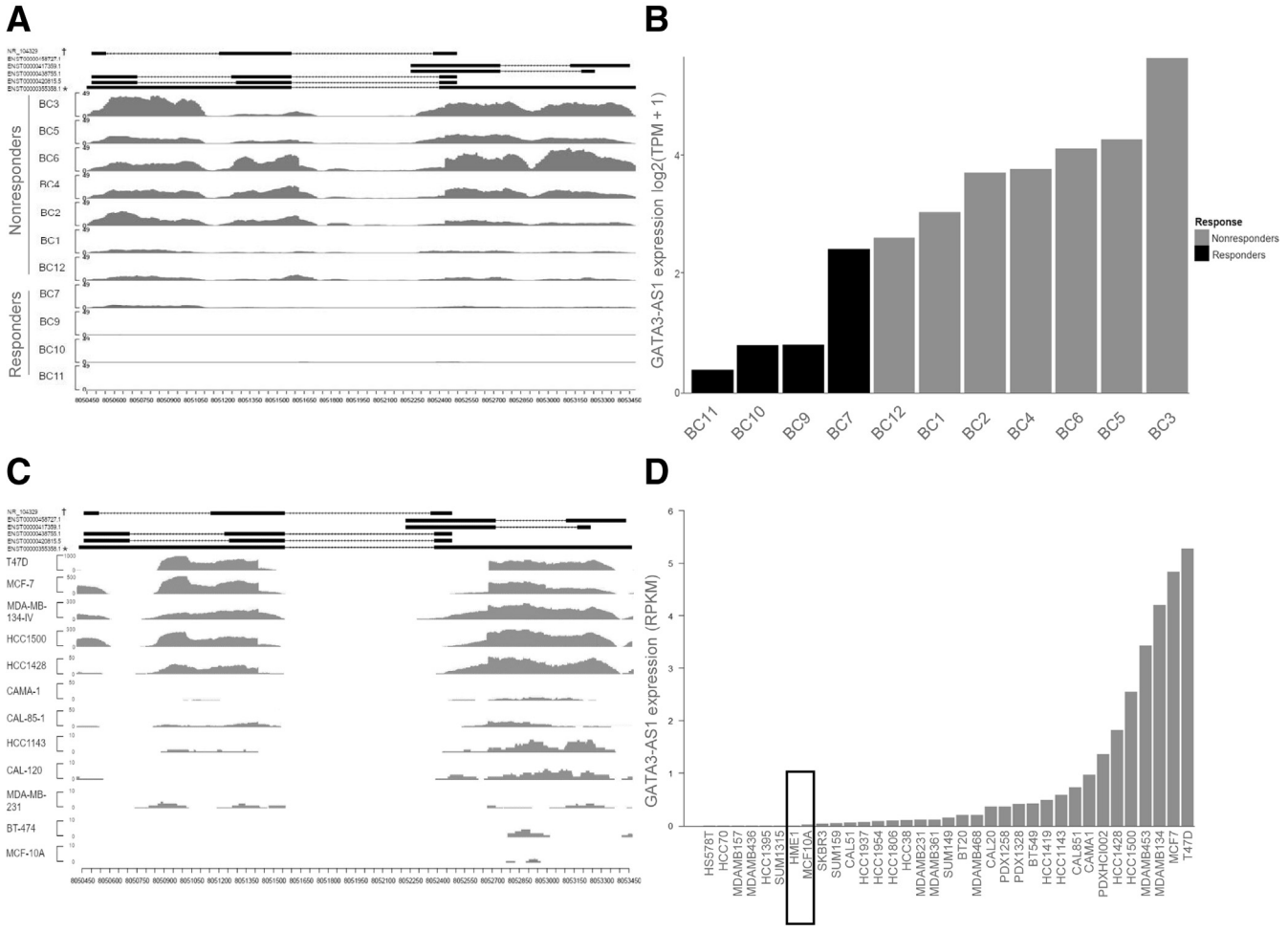


Figure 4 *GATA3-AS1* is overexpressed in breast cancer (BC) cell lines and in patients with resistance to neoadjuvant chemotherapy. **A:** The histograms show RNA-sequencing (RNA-Seq) mapped lectures in the GATA3-AS1 genomic locus using the genome hg38. Canonical GATA3-AS1 transcript and its six isoforms are shown [Ensemble identifier (ID), *https://www.ensembl.org/index.html*, last accessed February 25, 2021; **top panel**]. Nonresponder patients are represented in the **top panel**, whereas responder patients are represented on the **bottom panel**. **B:** Bar plot showing transcript per million (TPM) normalized expression of GATA3-AS1 in nonresponder (gray) and responder (black) patients. **C:** Histograms show RNA-Seq mapped lectures in the GATA3-AS1 genomic locus using the genome hg19. The canonical GATA3-AS1 transcript and its six isoforms are shown in the **top panel** (Ensemble ID). Breast cancer cell lines and transformed noncancerous breast cell line data were obtained from the Cancer Cell Line Encyclopedia (*https://portals.broadinstitute.org/ccle*, last accessed February 25, 2021). **D:** Bar plot showing reads per kilobase of transcript per million reads mapped (RPKM) normalized expression of GATA3-AS1 in 33 breast cancer cell lines (gray) and 2 transformed noncancerous breast cell lines (**black boxed area**). Data were obtained from the Library of Integrated Network-Based Cellular Signatures database (*http://lincs.hms.harvard.edu/db/datasets*, last accessed May 14, 2020). *Canonical GATA3-AS1 transcript. †Unreported isoform of GATA3-AS1 in Ensembl database; the ID was reported according to University of California, Santa Cruz, RefSeq annotation. $n = 11$ (**A**).

from RNA-Seq data obtained from TANRIC (Supplemental Figure S5), suggesting a cancer-specific expression pattern in positive hormone receptor phenotypes. Interestingly, *GATA3* was overexpressed in hormone-positive subtypes (Supplemental Figure S6) in a similar manner to the expression pattern seen for *GATA3-AS1*.

Higher Relative Expression Levels of *GATA3-AS1* Are Associated with Response to NAC

Once overexpression of *GATA3-AS1* was determined in breast cancer cell lines and nonresponder patients through RNA-Seq analysis, expression levels of *GATA3-AS1* were examined by RT-qPCR in human breast cancer cell lines and in samples from an independent cohort of patients. For breast cancer cell lines, results showed that the MCF-7 cell line exhibited increased expression levels of *GATA3-AS1* by >300-fold, followed by 100-fold overexpression in the MDA-MB-231 and BT474 cell lines, which presented 40-fold lower expression values of *GATA3-AS1* than the normal MCF10A cell line (Figure 5A). Hence, results revealed that all breast cancer cell lines analyzed in this study displayed higher expression of *GATA3-AS1* when their levels of expression were compared with nonneoplastic cell lines, such as MFC-10A.

Once expression levels of *GATA3-AS1* in human breast cancer cell lines were validated, its expression levels were analyzed by RT-qPCR in a validation cohort of 68 patients that included nonresponders and responders to NAC treatment within the luminal B-like phenotype. It was found that

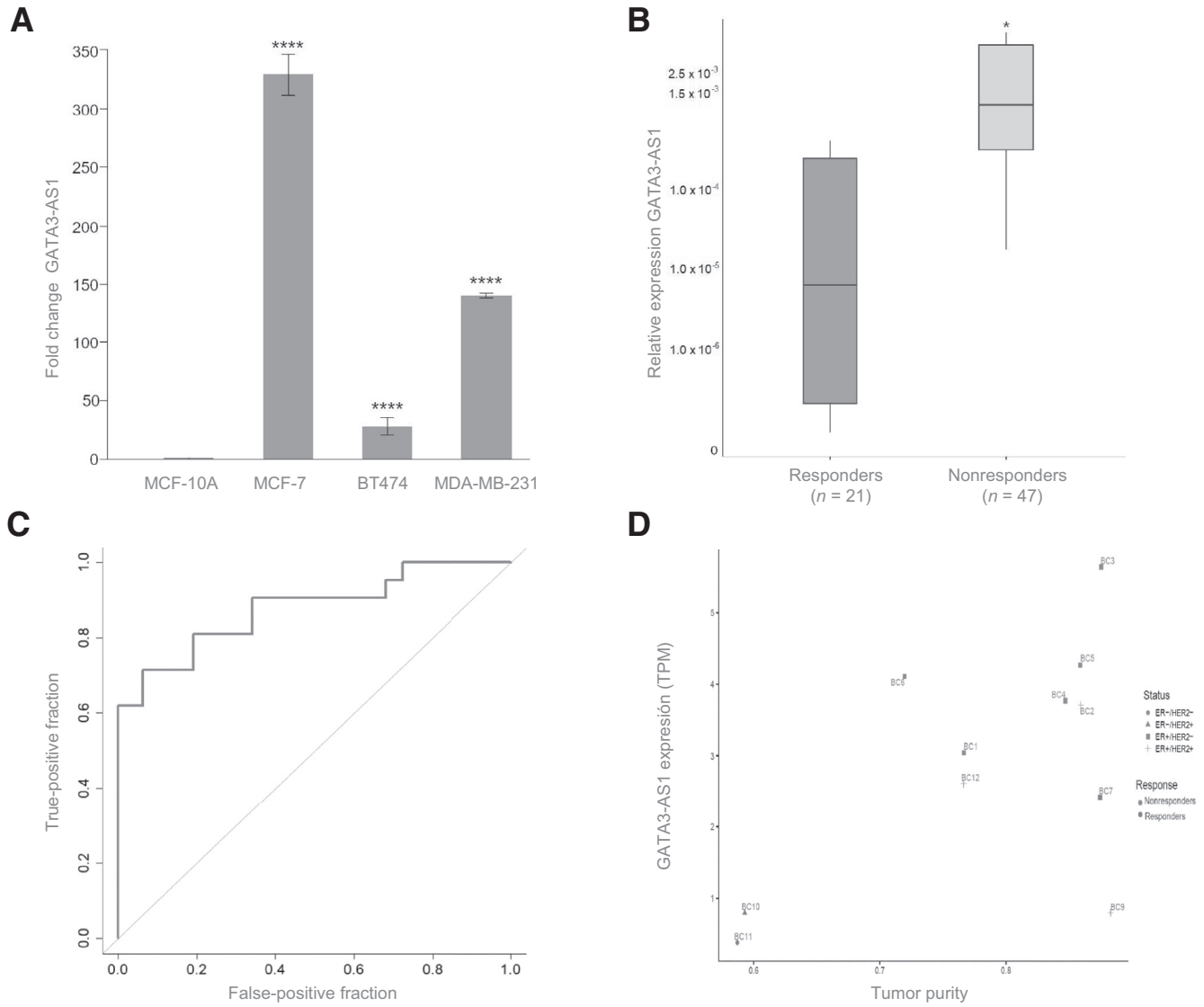


Figure 5 Higher relative expression levels of *GATA3-AS1* are associated with response to neoadjuvant chemotherapy. **A:** Bar plot of RT-qPCR quantification by fold change calculation (ΔΔCt), showing overexpression of GATA3-AS1 in breast cancer cell lines MCF-7, BT474, and MDA-MB-231, in which MCF-7 shows the highest GATA3-AS1 overexpression (analysis of variance with the Tukey test). **B:** Box plot comparing GATA3-AS1 relative expression between responder and nonresponder patients; a nonparametric *U*-test/Wilcoxon test was implemented, showing differences between medians. **C:** Receiver operating characteristic curve analysis indicates that GATA3-AS1 overexpression predicts nonresponse in neoadjuvant chemotherapy patients (cutoff = 0.000064 relative expression is normalized with *RPS28* as housekeeping gene) and is associated with a high sensitivity (92.9%) and specificity (75.0%), with $P = 0.0001$ and area under the curve = 0.876. **D:** Tumor purity analysis of the 11 patients from RNA-sequencing analysis (gray dots correspond to the nonresponders, and black dots correspond to responders). $n = 3$ (**A**); $n = 68$ (**B** and **C**). ****$P < 0.0001$ versus MCF-10A; $^{\dagger}P < 0.05$ versus nonresponders. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2.

nonresponder patients overexpressed *GATA3-AS1*, whereas responder patients exhibited underexpression of *GATA3-AS1* (Figure 5B). Subsequently, specificity and sensitivity evaluations were performed, constructing an adjusted receiver operating characteristic curve. Figure 5C shows receiver operating characteristic curve analysis to evaluate the predictive capacity of *GATA3-AS1* expression by RT-qPCR between nonresponder and responder patients. The analysis demonstrates that *GATA3-AS1* has a sensitivity of 92.9% and a specificity of 75% with an AUC of approximately 0.90, which indicates that the use of *GATA3-AS1* in the prediction of response to NAC distinguishes between patients who will not respond to NAC from those who will show response to system treatment (nonresponders versus responders, respectively). To determine whether there is a relationship between *GATA3-AS1* expression and tumor purity, RNA-Seq data analysis was used. As shown in Figure 5D, there was no relation between *GATA3-AS1* and tumor purity, suggesting that overexpression of *GATA3-AS1* and tumor purity is not associated with NAC response.

In addition, it was determined that *GATA3-AS1* and its adjacent coding gene *GATA3* were co-expressed in the cohort

Table 3 Bivariate and Multivariate Analysis to Identify Clinical Variables Related to *GATA3-AS1* Expression in Nonresponder Patients Treated with NAC

	Bivariate analysis				Multivariate analysis		
Variable	Response (21/68;33%)	OR	95% CI	*P* value	OR	95% CI	*P* value
GATA3-AS1 expression							
Yes ($n = 51$)	7 (13.73)	29.33	6.67–128.88	<0.001*	37.49	6.74–208.42	<0.001*
No ($n = 17$)	14 (82.35)	—	—	—	—	—	—
Age Dx							
	—	1.022	0.974–1.072	0.362	—	—	—
Menopause							
Yes ($n = 25$)	12 (48.00)	3.487	1.19–10.21	0.023*	0.209	0.04–0.96	0.045*
No ($n = 43$)	9 (20.93)	—	—	—	—	—	—
Phenotype							
Luminal B like ($n = 45$)	10 (22.22)	—	—	—	—	—	—
Luminal B like/HER2$^+$ ($n = 23$)	11 (47.83)	3.208	1.091–9.43	0.034*	0.30	0.74–14.74	0.117

*Statistically significant.

—, no data available; Dx, diagnostic; HER2, human epidermal growth factor receptor 2; NAC, neoadjuvant chemotherapy; OR, odds ratio.

of Mexican patients (Spearman = 0.63) (Supplemental Figure S7A) as well as in the public database TANRIC (Spearman = 0.80) (Supplemental Figure S7B). Interestingly, *GATA3* showed overexpression in breast cancer cell lines (Supplemental Figure S8A), but this coding gene was not significantly overexpressed in nonresponder patients compared with responder ones (Supplemental Figure S8B). Furthermore, *GATA3* showed low sensitivity of 52.4%, with a specificity of 73.9% and an AUC of 0.60 (Supplemental Figure S8C). Besides, no relationship was found between *GATA3* expression and tumor purity (Supplemental Figure S8D). Finally, GATA3 protein expression by immunohistochemistry was not shown to distinguish responders from nonresponders (Supplemental Figure S9), suggesting that in LABC within the luminal B-like phenotypes GATA3 is not a good molecular marker for predicting the response to NAC contrary to *GATA3-AS1*. Taken together, these results suggest that *GATA3-AS1* is overexpressed only in nonresponder patients, indicating that *GATA3-AS1* represents a potential biomarker of NAC resistance in patients with LABC within the luminal B-like phenotypes who do not respond to systemic treatment.

GATA3-AS1 Is a Predictive Molecular Biomarker of Response to NAC in Breast Cancer Patients

To determine the role of *GATA3-AS1* in prognosis, a Kaplan-Meier curve was generated. No relationship was detected for *GATA3-AS1* overexpression in OS when the cohort of 68 Mexican patients was used, suggesting that this lncRNA is not a prognostic factor in LABC patients within the luminal B-like phenotype (Supplemental Figure S10). Even when the relationship of *GATA3-AS1* with OS from TANRIC RNA-Seq luminal B phenotypes (Supplemental Figure S11A) and in all molecular subtypes (Supplemental Figure S11B) was validated, no relationship with OS was observed, confirming that *GATA3-AS1* is not related to prognosis. Therefore, these results suggest that *GATA3-AS1* is not a prognostic biomarker in the cohorts analyzed.

Furthermore, multivariate logistic regression showed that *GATA3-AS1* overexpression was an independent predictor of response adjusted by menopausal status and phenotype, proving to be statistically significant in this model with 37.49-fold (95% CI, 6.74–208.42) more probability in nonresponders with *GATA3-AS1* compared with responder patients who did not express *GATA3-AS1* (Table 3). Finally, from all analyses and results obtained in this study, *GATA3-AS1* is proposed as a novel divergent lncRNA that may serve as a potential predictive molecular biomarker of response to NAC, which could be included in clinical practice to manage Mexican LABC patients with a luminal B-like phenotype.

Functional Analysis of Long Noncoding RNAs in Nonresponder Patients to NAC

To identify potential affected pathways in nonresponder patients, lncRNA Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis was performed with LncRNAs2Pathways using the top 79 differentially expressed lncRNAs between nonresponder and responder groups from RNA-Seq data analyzed. It was found that the term taste transduction was significantly enriched for underexpressed lncRNAs ($P < 0.05$), whereas olfactory transduction and renin-angiotensin system were significantly enriched for overexpressed lncRNAs ($P < 0.05$) (Figure 6) and found that the most significantly enriched terms were taste transduction, Parkinson disease, Alzheimer disease, oxidative phosphorylation, and regulation of autophagy (Figure 6). These results indicate that lncRNAs may influence these pathways and biological processes associated with NAC response; however, more analysis is needed to confirm this.

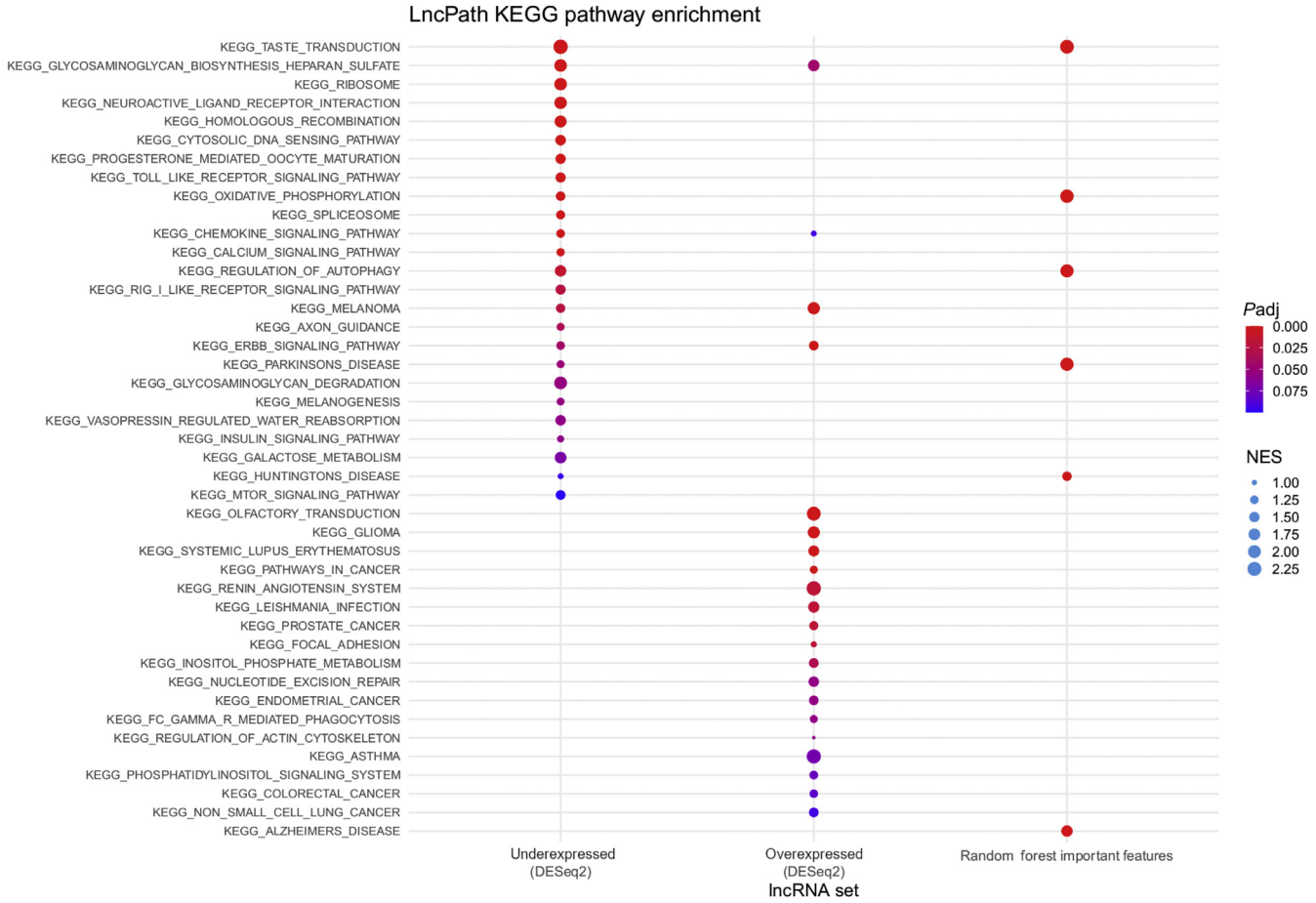


Figure 6 Functional analysis of long noncoding RNAs (lncRNAs) in nonresponder patients to neoadjuvant chemotherapy. Significantly enriched pathways for lncRNAs differentially expressed in the nonresponder group. Functional enrichment analysis was performed for underexpressed lncRNAs (**left side**), overexpressed lncRNAs (**middle**), and most significant lncRNAs (**right side**). The rows show the set that was analyzed, and columns show the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The more significant adjusted *P* value is indicated by the intensity of the red color, and the enrichment level of pathways is indicated by the normalized enrichment score (NES) score based on the size of the dot, as indicated.

Functional and Pathway Analysis of mRNAs in Patients Nonresponsive to NAC

Pathway enrichment analyses were performed using clusterProfiler on a collection of pathways from multiple sources and were collected by ConsensusPathDB[34] to determine functional enrichment analysis of the differentially expressed mRNAs in RNA-Seq data. The top 20 functionally enriched biological processes obtained from clusterProfiler analysis under Gene Ontology terms are indicated in a bar chart of the 214 differentially expressed mRNAs. The most significantly enriched biological processes were mesodermal commitment pathway, transmission across chemical synapsis, and neuronal system ($P < 0.05$) (Figure 7A). The interaction networks between enriched biological processes were analyzed, yielding an interaction network among the biological processes related to mesodermal commitment pathway and cell differentiation, as well as neuronal system and transmission across chemical synapsis, among others (Figure 7B). To identify genes involved in each functionally enriched biological process, a heat map of enriched mRNAs was constructed (Figure 7C). Finally, the network of gene pathways with significantly altered expression (FC > 1.5 and $P < 0.05$) was delineated using clusterProfiler (Figure 7D). These results suggest that mRNAs may influence these pathways and biological processes associated with NAC response.

Discussion

In general, lncRNAs are differentially expressed among human tissues.[19] These expression profiles are related to the different roles of lncRNAs in the cellular physiology of tissues, and alterations in the transcriptional rate of lncRNAs could lead to the development of pathologies, such as cancer.[42] Transcriptomic analysis of lncRNAs in different human neoplastic tissues demonstrated that these transcripts are differentially expressed among human cancers.[43] In particular, RNA-Seq and microarray studies have shown that breast cancer is characterized by specific lncRNA expression profiles among molecular subtypes

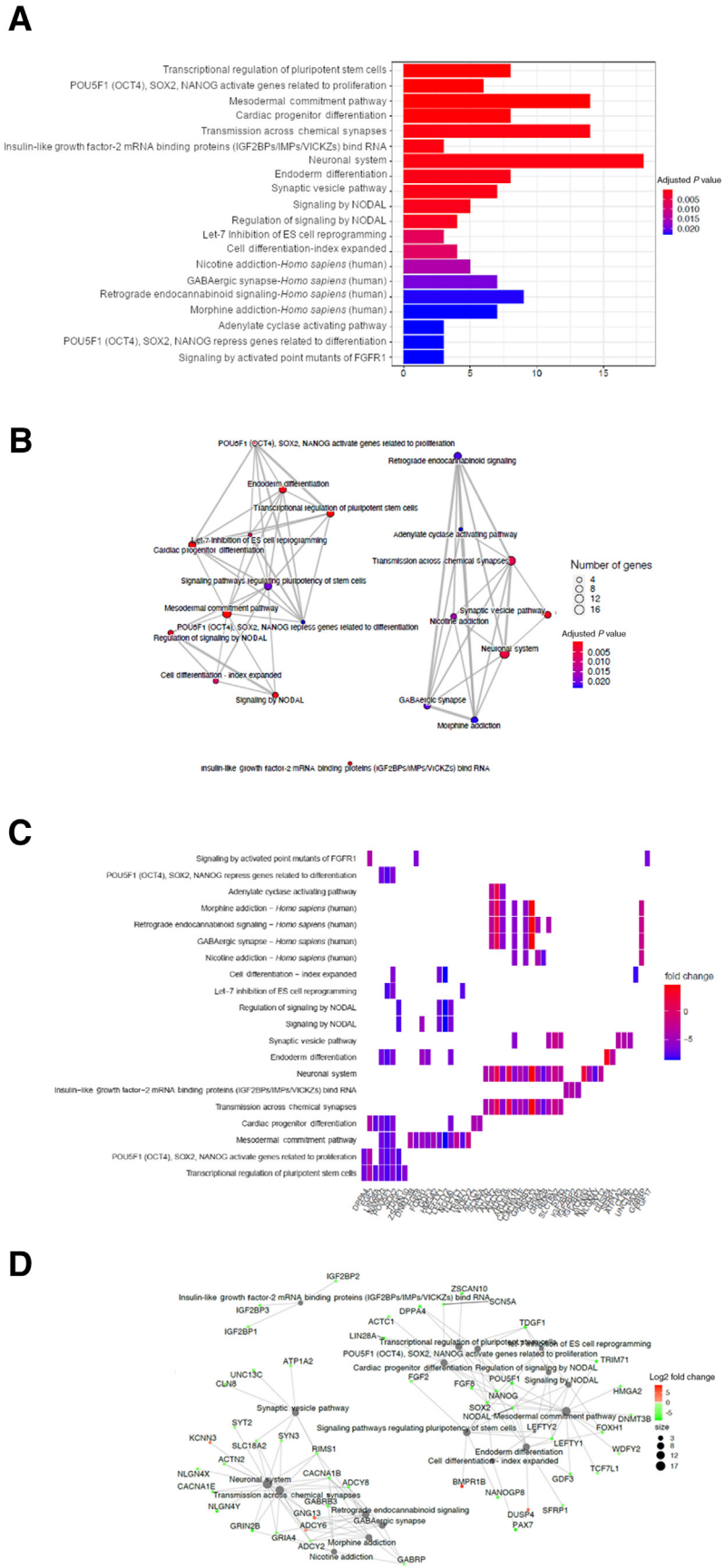


Figure 7 Functional and pathway analysis of mRNAs in nonresponder patients to neoadjuvant chemotherapy. **A:** Top 20 enriched biological processes of differentially expressed mRNAs in the nonresponder group. The significance of the *P* value is indicated by the intensity of the red color. **B:** Interaction networks of the top 20 enriched biological processes. Nodes are indicated with dots, and relationships between nodes are indicated by **gray lines**. **C:** Enrichment map for differentially expressed mRNAs in the nonresponder group. Columns show the mRNA symbol, and rows show the enriched term for biological processes. The enrichment fold change is indicated by the intensity of the color scales. **D:** Network of gene pathways showing differentially expressed mRNAs in the nonresponder group. Gene names are indicated in red and green dots, whereas gene category is indicated in gray dots. Nodes are indicated with dots, and relationships between nodes are indicated by **gray lines**. The number of genes of every node is indicated by the size of the dot, and the color of dots indicates the fold change (maximum fold change in red, and minimum fold change in green).

(luminal, HER2 enrichment, and basal)[15] that have been related to prognostic variables, such as OS, recurrence, progression, metastasis risk, treatment efficacy, and resistance to treatment.[44,45] Because of their association with treatment resistance, several lncRNAs have been identified for potential predictive use in endocrine therapy, such as *HOTAIR*,[46] *lncRNA-ATB* for antibody administration,[47] and *LINC00472* for adjuvant chemotherapy.[48] However, little is known about the association of lncRNAs and response to neoadjuvant chemotherapy, which is the standard treatment for breast cancer patients with locally advanced disease.[49,50] Several studies based on microarray assays have shown that lncRNAs are related to pathologic complete response in breast cancer patients in all molecular subtypes.[24,51] Currently, a few studies on transcriptome analysis by RNA-Seq in NAC resistance have been performed[52–54] to better understand the molecular biology of resistance to NAC in LABC patients.

In this study, a subset of long noncoding genes differentially expressed in LABC within luminal-like B phenotype patients who did not respond to NAC treatment were identified. From this subset, only the most characterized lncRNAs were selected, and identified the divergent lncRNA *GATA3-AS1*, which had been previously reported to be overexpressed in breast cancer patients.[51] In addition, *GATA3-AS1*, *RP11-279F6*, and *AC017048* showed specific and high expression levels in ER-positive (ER^+) compared with ER-negative (ER^-) cancers and normal breast tissue samples.[51] *GATA3-AS1* belongs to a class of divergent lncRNAs; it is located on chromosome 10 and is approximately 2 Kb. The transcription start site of this noncoding gene is approximately 1 Kb from the first exon of the adjacent *GATA3* gene and is synthesized from the antisense strand and consists of two exons.[38] Experimental validation by RT-qPCR of *GATA3-AS1* expression in breast cancer cell lines MCF-7, MDA-MB-231, and BT474 demonstrated an association between *GATA3-AS1* overexpression and neoplastic disease in mammary cells. This was further corroborated in samples from LABC patients in a discovery phase and an independent breast cancer patient cohort (validation phase), where it was observed that overexpression of *GATA3-AS1* is associated with patients who do not respond to systemic treatment, which is indicative of the relationship between *GATA3-AS1* overexpression and resistance to NAC. Moreover, multivariate logistic regression analysis was performed to demonstrate this association, showing that *GATA3-AS1* overexpression is an independent predictor of response and proving that *GATA3-AS1* is statistically significant in this model, with 37.49% (95% CI, 6.74%–208.42%). Moreover, high sensitivity and specificity (92.9% and 75%, respectively) with an AUC value of approximately 0.90 suggested that *GATA3-AS1* is a potential biomarker for predicting NAC response in clinical practice to improve therapy for LABC within luminal B-like phenotype patients. Several preclinical studies have identified lncRNAs that proposed as response biomarkers despite the reduced use of patient samples, such as *UCA1*,[55] *TP53COR1* (or lincRNA-p21),[56] *GAS5*,[57] and *HOTAIR*,[46] but clinical application of each lncRNA analyzed is relevant to each type of treatment in the cancer patients in whom they were evaluated. The lncRNA *GATA3-AS1*, despite having been validated in a small number of patients, might be an important biomarker of response to chemotherapy similar to the previously mentioned lncRNAs. Further studies will be needed to accurately evaluate the potential use of this lncRNA as a clinical predictive biomarker in the luminal B-like phenotype in a large cohort of patients.

Furthermore, patients in the luminal B group are of clinical interest because luminal B is a specific subtype that contains clinically aggressive ER-positive breast cancers in which patients present an intermediate prognosis with a high variety of responses to different treatments.[58] In particular, it was determined that *GATA3-AS1* is overexpressed in luminal B-like patients who do not respond to neoadjuvant chemotherapy. A clear association was observed between *GATA3-AS1* overexpression and nonresponders; however, no association was found between patient outcomes when OS was assessed against *GATA3-AS1* overexpression. Interestingly, the expression profile of *GATA3-AS1* is similar to that of lncRNAs, such as *DSCAM-AS1*, where its expression is not directly associated with prognosis but rather with response to treatment or progression of the disease.[20] Moreover, both lncRNAs exhibited higher expression in luminal B patients analyzed. It is also necessary to include patients diagnosed with luminal A breast cancer subtypes to elucidate whether it is possible to extend the predictive value of *GATA3-AS1* in this group and to generalize its use as a molecular biomarker. LncRNAs are known useful factors in treatment selection, independently of molecular cancer subtype[15]; however, this will likely require additional large cohorts to be analyzed that include both hormone-positive subtypes.

Conversely, the coding gene *GATA3* has a wide expression pattern in patients and breast cancer cell lines, as observed by RNA-Seq analysis. As previous studies have demonstrated, mRNAs tend to have less tissue- and stage-specific expression[42,59,60] in contrast to lncRNAs, which tend to have more tissue-specific[19,61,62] and stage-specific expression in disease, which is one of the main reasons lncRNAs have been proposed as molecular biomarkers in cancer.[16,63,64] Alternatively, results by receiver operating characteristic curve showed that *GATA3* had sensitivity of 52.4% and specificity of 73.9% with an AUC of 0.60 in RT-qPCR expression analyses; the results suggest that *GATA3* mRNA expression has low ability to distinguish responders from nonresponders to NAC due to its low sensitivity. Currently, GATA3 is a molecular marker in breast cancer that has had a controversial clinical role.[39] GATA3 protein expression has been associated with a favorable prognosis and increased survival in patients with invasive breast carcinoma; however, GATA3 has not been shown to be a

reliable prognostic factor regardless of ER status.[65–67] Likewise, clinical data regarding the role of GATA3 in treatment response prediction have also been controversial.[68,69] There is evidence suggesting that GATA3 could be used as a biomarker for predicting response to NAC. It was observed that absence of GATA3 is an independent pathologic complete response predictor to neoadjuvant chemotherapy through multivariate analysis, suggesting that GATA3 might be clinically useful as a predictor of poor response to chemotherapy.[70] However, these results did not evaluate the sensibility and specificity of GATA3; thus, more studies are necessary to determine the implications of GATA3 as a predictor biomarker to NAC response. To date, there is controversy over the use of GATA3 as a predictive biomarker for NAC[69–71] and, because of the low sensitivity found in the present study, the coding gene *GATA3* cannot be proposed as a biomarker for NAC resistance, even though it is co-expressed with *GATA3-AS1*. The results obtained in this study indicate that there is a moderate correlation between *GATA3-AS1* and *GATA3* mRNA expression. However, in the literature, it has been suggested that in breast cancer, *GATA3-AS1* does not regulate *GATA3* mRNA expression; instead, *GATA3-AS1* regulates GATA3 protein accumulation level through a degradation mechanism.[72] Therefore, GATA3 protein levels, measured by immunohistochemistry, cannot distinguish patients who will be resistant to neoadjuvant chemotherapy, but the possibility that GATA3 protein accumulation levels could be regulated by *GATA3-AS1*, as has been demonstrated in triple-negative breast cancer cell lines, cannot be ruled out.[72] However, this is beyond the scope of this study, and further research is needed. Thus, in locally advanced breast cancer luminal B-like phenotype, regulation of *GATA3* mRNA could not depend on *GATA3-AS1* despite there being a positive correlation in the expression levels of both transcripts.

On the other hand, there is scientific evidence that altered gene expression in breast cancer tumors is involved in neuronal-related pathways[73] and processes.[74–76] These results showed that tumors of nonresponders to NAC overexpressed lncRNA genes that are related to Kyoto Encyclopedia of Genes and Genomes pathways, such as taste transduction and Parkinson disease. This is in accordance with other reports in which lncRNA expression in breast cancer cells is related to the dysregulation of neuronal-related pathways, such as lncRNA *IRAIN* and its targets involved in cholinergic synapses[77] and the lncRNA-mRNA coexpression network, which is related to taste transduction in docetaxel-resistant breast cancer cell lines.[78] In addition, this study demonstrated that the expression profile of mRNA in breast cancer tumors is involved in the neuronal system signaling, as has been described for other processes, such as cranial nerve and neural crest development in breast cancer patients.[76,79] Together, these data indicate that genes involved in neuronal processes are regulated by lncRNAs in breast cells and, in cancer conditions, might acquire oncogenic potential, contributing to breast cancer progression, resistance to treatment, or metastasis development, as shown for lncRNA *BORG*, which is associated with brain metastasis.[80]

Finally, the divergent lncRNA *GATA3-AS1* is a noncoding transcript described in T lymphocytes with an important role in T-cell differentiation,[38] and it has been associated with respiratory pathologies, such as asthma and rhinitis.[81] In this study, it was found that *GATA3-AS1* is also expressed in mammary cells, but until now, the function of *GATA3-AS1* in breast tissue has not been described, as seen in hepatocellular carcinoma.[82] Further research is needed to establish the function and molecular mechanisms of *GATA3-AS1* in mammary tissue and how it associates with breast neoplastic disease and NAC resistance.

In conclusion, the presence of an lncRNA profile capable of defining patients nonresponsive to NAC in LABC luminal B-like patients by RNA-Seq analysis was demonstrated. Particularly, the divergent lncRNA *GATA3-AS1* showed high specificity and sensitivity associated with its predictive value in nonresponders to NAC treatment, making it the first molecular biomarker with a potential use in clinical practice in the prediction of NAC treatment response in breast cancer. Further investigation is needed to discover whether its predictive value is applicable to other molecular subtypes and to uncover the molecular mechanisms of this lncRNA in NAC resistance.

Acknowledgment

We thank the National Cancer Institute of Mexico for support.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2021.07.014*.

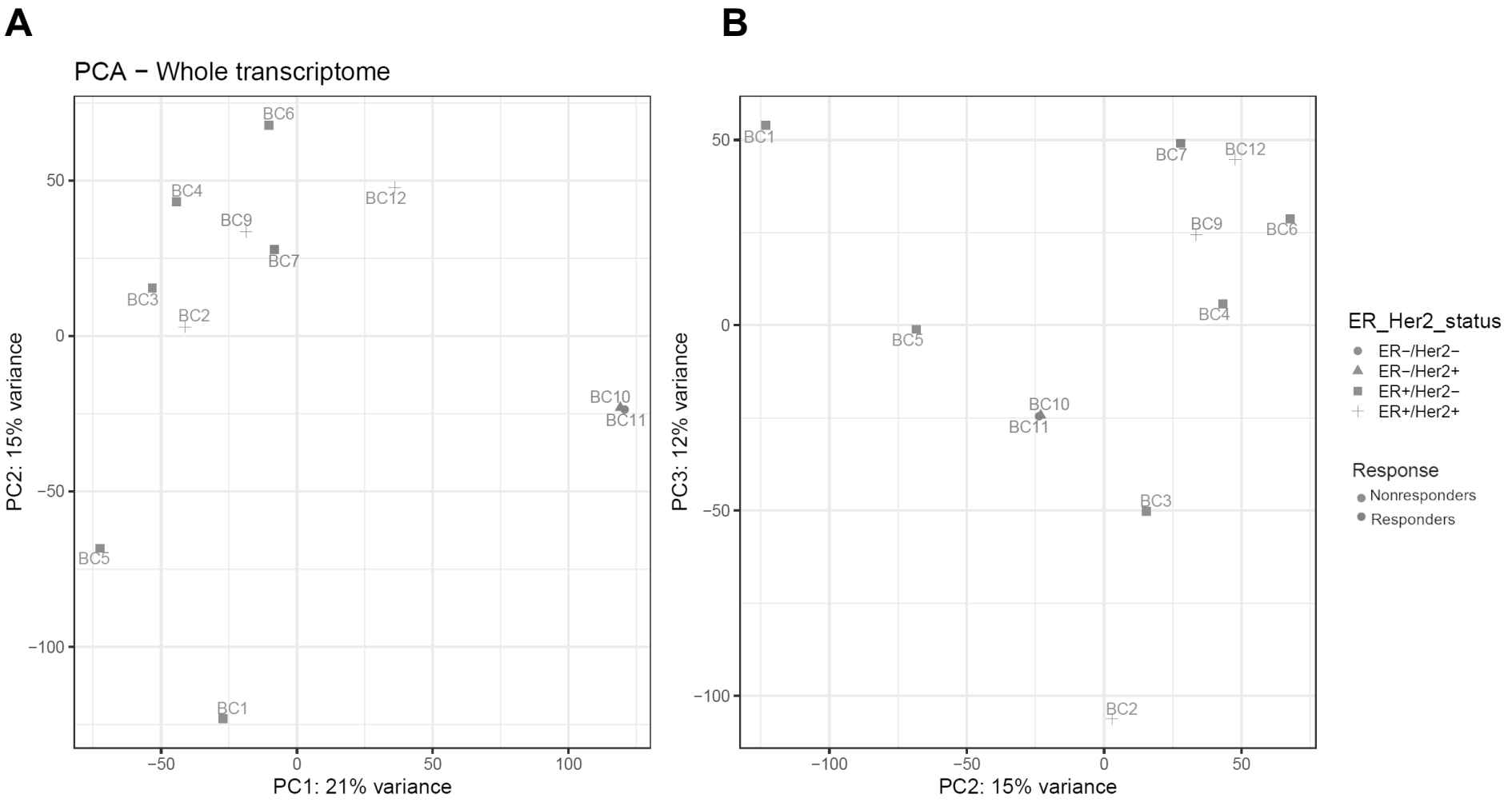
References

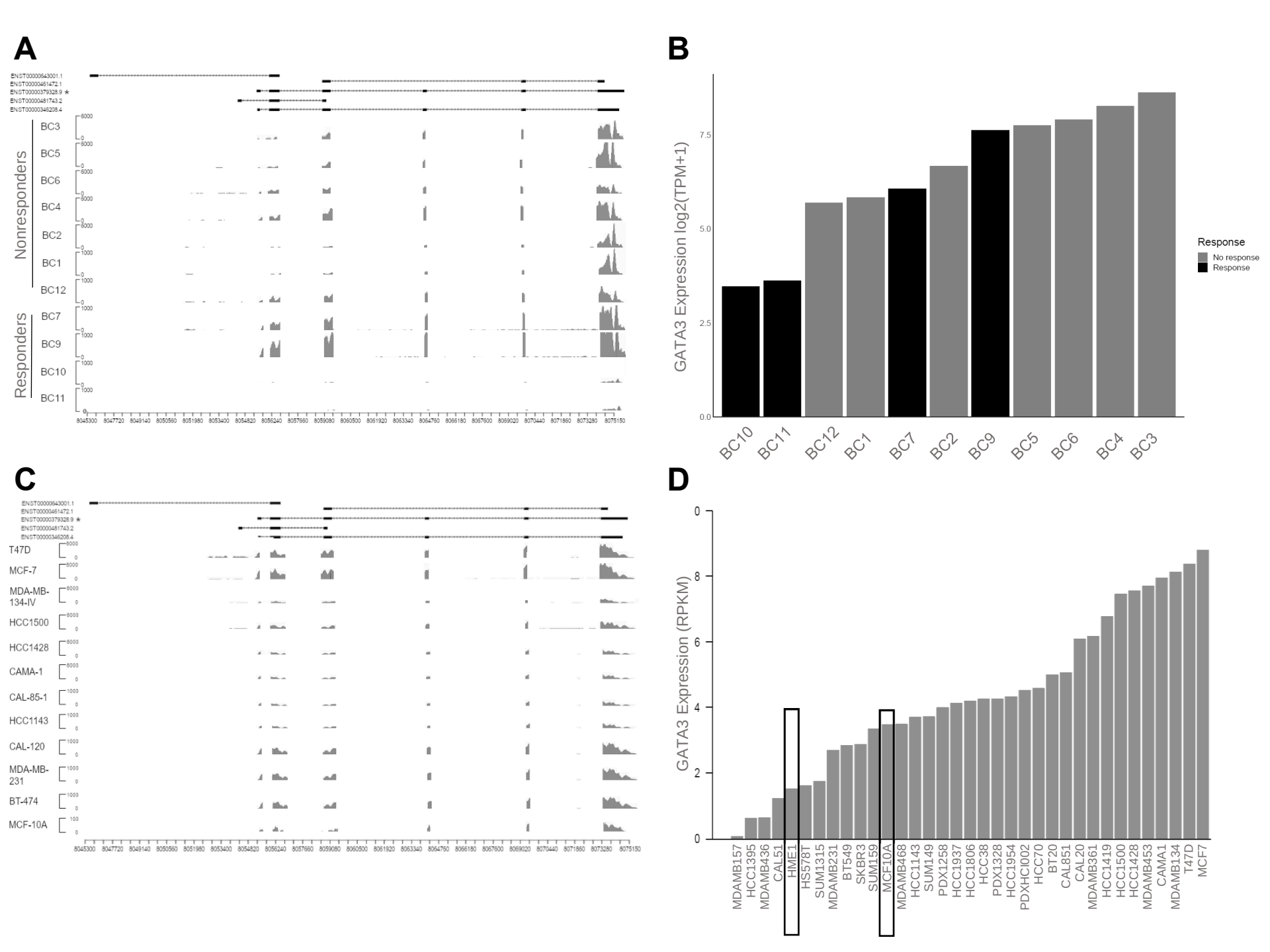
1. Miller JW, Plescia M, Ekwueme DU: Public health national approach to reducing breast and cervical cancer disparities. Cancer 2014, 120: 2537–2539
2. Raphael J, Paramsothy T, Li N, Lee J, Gandhi S: A single-institution experience of salvage therapy for patients with early and locally advanced breast cancer who progress during neoadjuvant chemotherapy. Breast Cancer Res Treat 2017, 163:11–19
3. Gardin G, Rosso R, Campora E, Repetto L, Naso C, Canavese G, Catturich A, Corvò R, Guenzi M, Pronzato P: Locally advanced non-metastatic breast cancer: analysis of prognostic factors in 125 patients homogeneously treated with a combined modality approach. Eur J Cancer 1995, 31A:1428–1433
4. Pe M, Dorme L, Coens C, Basch E, Calvert M, Campbell A, Cleeland C, Cocks K, Collette L, Dirven L, Dueck AC, Devlin N, Flechtner H-H, Gotay C, Griebsch I, Groenvold M, King M, Koller M, Malone DC, Martinelli F, Mitchell SA, Musoro JZ, Oliver K, Piault-Louis E, Piccart M, Pimentel FL, Quinten C, Reijneveld JC, Sloan J, Velikova G, Bottomley A;

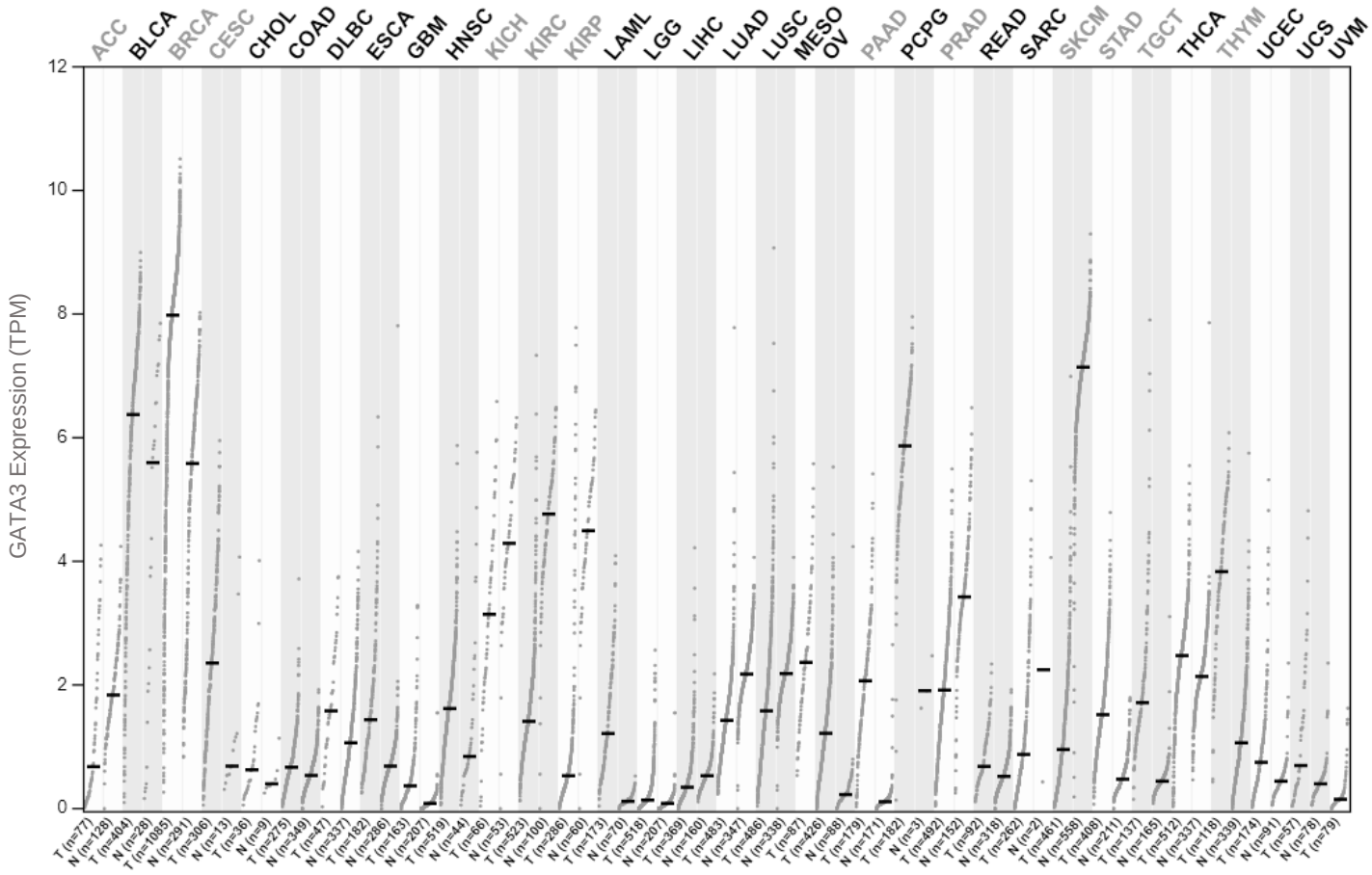
Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data Consortium (SISAQOL): Statistical analysis of patient-reported outcome data in randomised controlled trials of locally advanced and metastatic breast cancer: a systematic review. Lancet Oncol 2018, 19:e459–e469
5. Tryfonidis K, Senkus E, Cardoso MJ, Cardoso F: Management of locally advanced breast cancer-perspectives and future directions. Nat Rev Clin Oncol 2015, 12:147–162
6. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, et al: Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet 2014, 384: 164–172
7. Duffy MJ, Harbeck N, Nap M, Molina R, Nicolini A, Senkus E, Cardoso F: Clinical use of biomarkers in breast cancer: updated guidelines from the European Group on Tumor Markers (EGTM). Eur J Cancer 2017, 75:284–298
8. Han HS, Magliocco AM: Molecular testing and the pathologist's role in clinical trials of breast cancer. Clin Breast Cancer 2016, 16: 166–179
9. Pantel K, Alix-Panabières C: Liquid biopsy and minimal residual disease - latest advances and implications for cure. Nat Rev Clin Oncol 2019, 16:409–424
10. Giuliano AE, Connolly JL, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, Weaver DL, Winchester DJ, Hortobagyi GN: Breast cancer-major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin 2017, 67:290–303
11. Kwa M, Makris A, Esteva FJ: Clinical utility of gene-expression signatures in early stage breast cancer. Nat Rev Clin Oncol 2017, 14:595–610
12. Cabrera-Galeana P, Muñoz-Montaño W, Lara-Medina F, Alvarado-Miranda A, Pérez-Sánchez V, Villarreal-Garza C, Quintero RM, Porras-Reyes F, Bargallo-Rocha E, Del Carmen I, Mohar A, Arrieta O: Ki67 changes identify worse outcomes in residual breast cancer tumors after neoadjuvant chemotherapy. Oncologist 2018, 23: 670–678
13. Ellis MJ, Suman VJ, Hoog J, Goncalves R, Sanati S, Creighton CJ, DeSchryver K, Crouch E, Brink A, Watson M, Luo J, Tao Y, Barnes M, Dowsett M, Budd GT, Winer E, Silverman P, Esserman L, Carey L, Ma CX, Unzeitig G, Pluard T, Whitworth P, Babiera G, Guenther JM, Dayao Z, Ota D, Leitch M, Olson JA, Allred DC, Hunt K: Ki67 proliferation index as a tool for chemotherapy decisions during and after neoadjuvant aromatase inhibitor treatment of breast cancer: results from the American College of Surgeons Oncology Group Z1031 trial (alliance). J Clin Oncol 2017, 35:1061–1069
14. Jain P, Doval DC, Batra U, Goyal P, Bothra SJ, Agarwal C, Choudhary DK, Yadav A, Koyalla VPB, Sharma M, Dash P, Talwar V: Ki-67 labeling index as a predictor of response to neo-adjuvant chemotherapy in breast cancer. Jpn J Clin Oncol 2019, 49: 329–338
15. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al: A comprehensive pan-cancer molecular study of gynecologic and breast cancers. Cancer Cell 2018, 33:690–705.e9
16. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai M-C, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY: Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010, 464:1071–1076
17. Godinho MFE, Sieuwerts AM, Look MP, Meijer D, Foekens JA, Dorssers LCJ, van Agthoven T: Relevance of BCAR4 in tamoxifen resistance and tumour aggressiveness of human breast cancer. Br J Cancer 2010, 103:1284–1291
18. Quinn JJ, Chang HY: Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet 2016, 17:47–62
19. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, Poliakov A, Cao X, Dhanasekaran SM, Wu Y-M, Robinson DR, Beer DG, Feng FY, Iyer HK, Chinnaiyan AM: The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 2015, 47:199–208
20. Niknafs YS, Han S, Ma T, Speers C, Zhang C, Wilder-Romans K, Iyer MK, Pitchiaya S, Malik R, Hosono Y, Prensner JR, Poliakov A, Singhal U, Xiao L, Kregel S, Siebenaler RF, Zhao SG, Uhl M, Gawronski A, Hayes DF, Pierce LJ, Cao X, Collins C, Backofen R, Sahinalp CS, Rae JM, Chinnaiyan AM, Feng FY: The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. Nat Commun 2016, 7:12791
21. Campos-Parra AD, López-Urrutia E, Orozco Moreno LT, López-Camarillo C, Meza-Menchaca T, Figueroa González G, Bustamante Montes LP, Pérez-Plasencia C: Long non-coding RNAs as new master regulators of resistance to systemic treatments in breast cancer. Int J Mol Sci 2018, 19:2711
22. 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 2017, 10:988–997
23. Wang Q, Li C, Tang P, Ji R, Chen S, Wen J: A minimal lncRNA-mRNA signature predicts sensitivity to neoadjuvant chemotherapy in triple-negative breast cancer. Cell Physiol Biochem 2018, 48: 2539–2548
24. Zeng Y, Wang G, Zhou C-F, Zhang H-B, Sun H, Zhang W, Zhou H-H, Liu R, Zhu Y-S: LncRNA profile study reveals a three-LncRNA signature associated with the pathological complete response following neoadjuvant chemotherapy in breast cancer. Front Pharmacol 2019, 10:574
25. Coates AS, Winer EP, Goldhirsch A, Gelber RD, Gnant M, Piccart-Gebhart M, et al: Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the primary therapy of early breast cancer 2015. Ann Oncol 2015, 26: 1533–1546
26. Poluri RTK, Beauparlant CJ, Droit A, Audet-Walsh É: RNA sequencing data of human prostate cancer cells treated with androgens. Data Brief 2019, 25:104372
27. Kroll KW, Mokaram NE, Pelletier AR, Frankhouser DE, Westphal MS, Stump PA, Stump CL, Bundschuh R, Blachly JS, Yan P: Quality control for RNA-Seq (QuaCRS): an integrated quality control pipeline. Cancer Inform 2014, 13:7–14
28. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014, 30: 2114–2120
29. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29:15–21
30. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nat Methods 2017, 14:417–419
31. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, Mills GB, Verhaak RGW: Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013, 4:2612
32. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014, 15:550
33. Bienkowska JR, Dalgin GS, Batliwalla F, Allaire N, Roubenoff R, Gregersen PK, Carulli JP: Convergent random forest predictor: methodology for predicting drug response from genome-scale data applied to anti-TNF response. Genomics 2009, 94:423–432
34. Kamburov A, Wierling C, Lehrach H, Herwig R: ConsensusPathDB—a database for integrating human functional interaction networks. Nucleic Acids Res 2009, 37:D623–D628
35. Yu G, Wang L-G, Han Y, He Q-Y: clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012, 16: 284–287

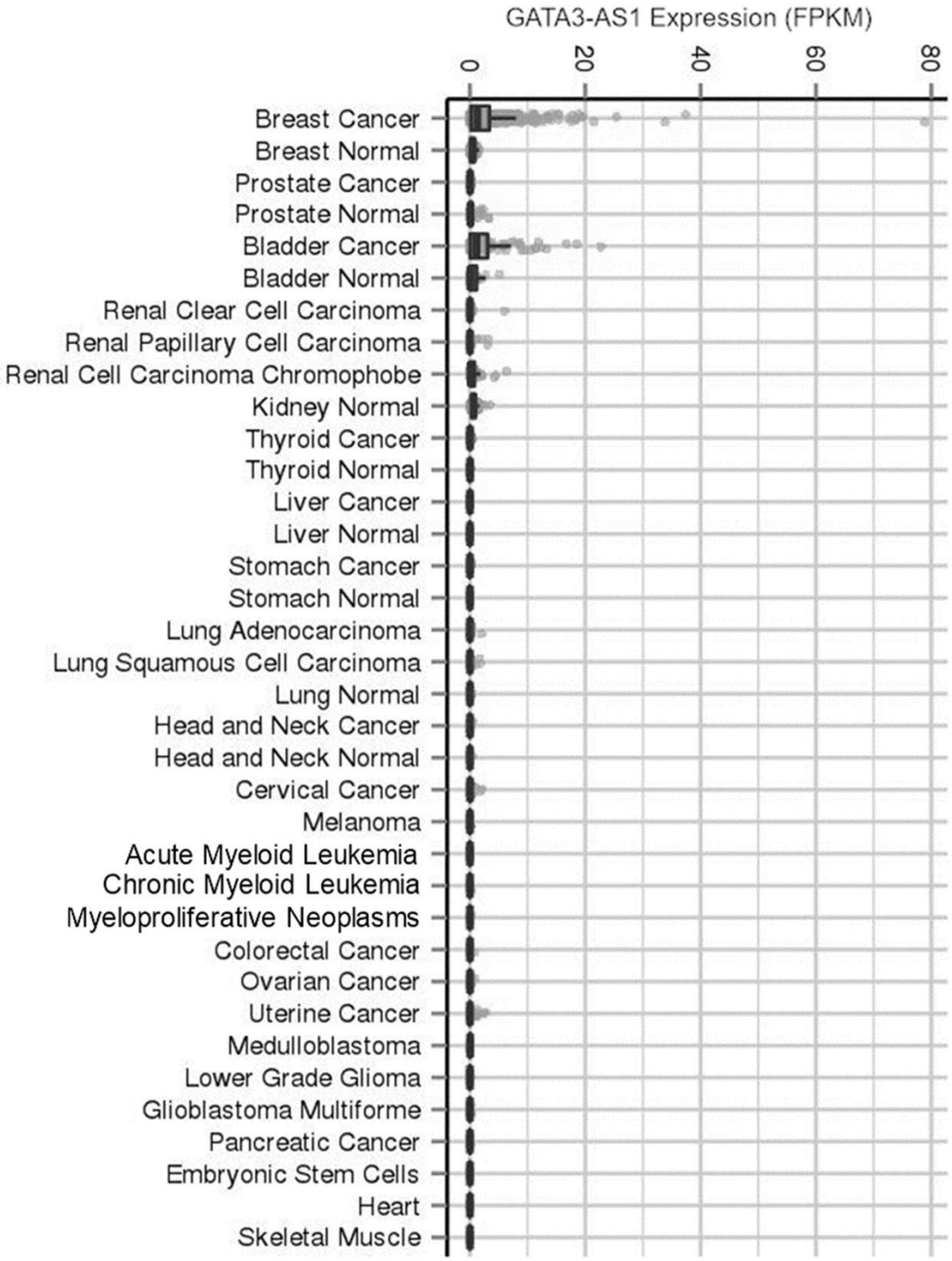
36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005, 102:15545−15550
37. Han J, Liu S, Sun Z, Zhang Y, Zhang F, Zhang C, Shang D, Yang H, Su F, Xu Y, Li C, Ren H, Li X: LncRNAs2Pathways: identifying the pathways influenced by a set of lncRNAs of interest based on a global network propagation method. Sci Rep 2017, 7: 46566
38. Gibbons HR, Shaginurova G, Kim LC, Chapman N, Spurlock CFI, Aune TM: Divergent lncRNA GATA3-AS1 regulates GATA3 transcription in T-helper 2 cells. Front Immunol 2018, 9:2512
39. Du F, Yuan P, Wang T, Zhao J, Zhao Z, Luo Y, Xu B: The significance and therapeutic potential of GATA3 expression and mutation in breast cancer: a systematic review. Med Res Rev 2015, 35:1300−1315
40. Assefa AT, De Paepe K, Everaert C, Mestdagh P, Thas O, Vandesompele J: Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. Genome Biol 2018, 19:96
41. Arora S, Pattwell SS, Holland EC, Bolouri H: Variability in estimated gene expression among commonly used RNA-seq pipelines. Sci Rep 2020, 10:2734
42. Yan X, Hu Z, Feng Y, Hu X, Yuan J, Zhao SD, Zhang Y, Yang L, Shan W, He Q, Fan L, Kandalaft LE, Tanyi JL, Li C, Yuan C-X, Zhang D, Yuan H, Hua K, Lu Y, Katsaros D, Huang Q, Montone K, Fan Y, Coukos G, Boyd J, Sood AK, Rebbeck T, Mills GB, Dang CV, Zhang L: Comprehensive genomic characterization of long non-coding RNAs across human cancers. Cancer Cell 2015, 28:529−540
43. Wang Z, Yang B, Zhang M, Guo W, Wu Z, Wang Y, Jia L, Li S; Cancer Genome Atlas Research Network, Xie W, Yang D: lncRNA epigenetic landscape analysis identifies EPIC1 as an oncogenic lncRNA that interacts with MYC and promotes cell-cycle progression in cancer. Cancer Cell 2018, 33:706−720.e9
44. Chandra Gupta S, Nandan Tripathi Y: Potential of long non-coding RNAs in cancer patients: from biomarkers to therapeutic targets. Int J Cancer 2017, 140:1955−1967
45. Chen Q, Wei C, Wang Z, Sun M: Long non-coding RNAs in anticancer drug resistance. Oncotarget 2016, 8:1925−1936
46. Xue X, Yang YA, Zhang A, Fong K-W, Kim J, Song B, Li S, Zhao JC, Yu J: LncRNA HOTAIR enhances ER signaling and confers tamoxifen resistance in breast cancer. Oncogene 2016, 35: 2746−2755
47. Shi S-J, Wang L-J, Yu B, Li Y-H, Jin Y, Bai X-Z: LncRNA-ATB promotes trastuzumab resistance and invasion-metastasis cascade in breast cancer. Oncotarget 2015, 6:11652−11663
48. Shen Y, Katsaros D, Loo LWM, Hernandez BY, Chong C, Canuto EM, Biglia N, Lu L, Risch H, Chu W-M, Yu H: Prognostic and predictive values of long non-coding RNA LINC00472 in breast cancer. Oncotarget 2015, 6:8579−8592
49. Specht J, Gralow JR: Neoadjuvant chemotherapy for locally advanced breast cancer. Semin Radiat Oncol 2009, 19:222−228
50. Klein J, Tran W, Watkins E, Vesprini D, Wright FC, Look Hong NJ, Ghandi S, Kiss A, Czarnota GJ: Locally advanced breast cancer treated with neoadjuvant chemotherapy and adjuvant radiotherapy: a retrospective cohort analysis. BMC Cancer 2019, 19:306
51. Zhang Y, Wagner EK, Guo X, May I, Cai Q, Zheng W, He C, Long J: Long intergenic non-coding RNA expression signature in human breast cancer. Sci Rep 2016, 6:37821
52. Lesurf R, Griffith OL, Griffith M, Hundal J, Trani L, Watson MA, Aft R, Ellis MJ, Ota D, Suman VJ, Meric-Bernstam F, Leitch AM, Boughey JC, Unzeitig G, Buzdar AU, Hunt KK, Mardis ER: Genomic characterization of HER2-positive breast cancer and response to neoadjuvant trastuzumab and chemotherapy-results from the ACOSOG Z1041 (alliance) trial. Ann Oncol 2017, 28:1070−1077
53. Tanioka M, Fan C, Parker JS, Hoadley KA, Hu Z, Li Y, Hyslop TM, Pitcher BN, Soloway MG, Spears PA, Henry LN, Tolaney S, Dang CT, Krop IE, Harris LN, Berry DA, Mardis ER, Winer EP, Hudis CA, Carey LA, Perou CM: Integrated analysis of RNA and DNA from the phase III trial CALGB 40601 identifies predictors of response to trastuzumab-based neoadjuvant chemotherapy in HER2-positive breast cancer. Clin Cancer Res 2018, 24:5292−5304
54. Fumagalli D, Venet D, Ignatiadis M, Azim HA, Maetens M, Rothé F, Salgado R, Bradbury I, Pusztai L, Harbeck N, Gomez H, Chang T-W, Coccia-Portugal MA, Di Cosimo S, de Azambuja E, de la Peña L, Nuciforo P, Brase JC, Huober J, Baselga J, Piccart M, Loi S, Sotiriou C: RNA sequencing to predict response to neoadjuvant anti-HER2 therapy. JAMA Oncol 2017, 3:227−234
55. Fan Y, Shen B, Tan M, Mu X, Qin Y, Zhang F, Liu Y: Long non-coding RNA UCA1 increases chemoresistance of bladder cancer cells by regulating Wnt signaling. FEBS J 2014, 281:1750−1758
56. Luo J, Wang K, Yeh S, Sun Y, Liang L, Xiao Y, Xu W, Niu Y, Cheng L, Maity SN, Jiang R, Chang C: LncRNA-p21 alters the antiandrogen enzalutamide-induced prostate cancer neuroendocrine differentiation via modulating the EZH2/STAT3 signaling. Nat Commun 2019, 10:2571
57. Chen Z, Pan T, Jiang D, Jin L, Geng Y, Feng X, Shen A, Zhang L: The lncRNA-GAS5/miR-221-3p/DKK2 axis modulates ABCB1-mediated adriamycin resistance of breast cancer via the Wnt/β-catenin signaling pathway. Mol Ther Nucleic Acids 2020, 19:1434−1448
58. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, Ruddy K, Tsang J, Cardoso F: Breast cancer. Nat Rev Dis Primers 2019, 5:1−31
59. Napoli M, Li X, Ackerman HD, Deshpande AA, Barannikov I, Pisegna MA, Bedrosian I, Mitsch J, Quinlan P, Thompson A, Rajapakshe K, Coarfa C, Gunaratne PH, Marchion DC, Magliocco AM, Tsai KY, Flores ER: Pan-cancer analysis reveals TAp63-regulated oncogenic lncRNAs that promote cancer progression through AKT activation. Nat Commun 2020, 11:5156
60. Deva Magendhra Rao AK, Patel K, Korivi Jyothiraj S, Meenakumari B, Sundersingh S, Sridevi V, Rajkumar T, Pandey A, Chatterjee A, Gowda H, Mani S: Identification of lncRNAs associated with early-stage breast cancer and their prognostic implications. Mol Oncol 2019, 13:1342−1355
61. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011, 25:1915−1927
62. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R: The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012, 22:1775−1789
63. Schmitt AM, Chang HY: Long noncoding RNAs in cancer pathways. Cancer Cell 2016, 29:452−463
64. Bhan A, Soleimani M, Mandal SS: Long noncoding RNA and cancer: a new paradigm. Cancer Res 2017, 77:3965−3981
65. Yildirim E, Bektas S, Gundogar O, Findik D, Alcicek S, Erdogan KO, Yildiz M: The relationship of GATA3 and Ki-67 with histopathological prognostic parameters, locoregional recurrence and disease-free survival in invasive ductal carcinoma of the breast. Anticancer Res 2020, 40:5649−5657
66. Pei X-H, Bai F, Smith MD, Usary J, Fan C, Pai S-Y, Ho I-C, Perou CM, Xiong Y: CDK inhibitor p18INK4c is a downstream target of GATA3 and restrains mammary luminal progenitor cell proliferation and tumorigenesis. Cancer Cell 2009, 15:389−401
67. Afzaljavan F, Sadr AS, Savas S, Pasdar A: GATA3 somatic mutations are associated with clinicopathological features and

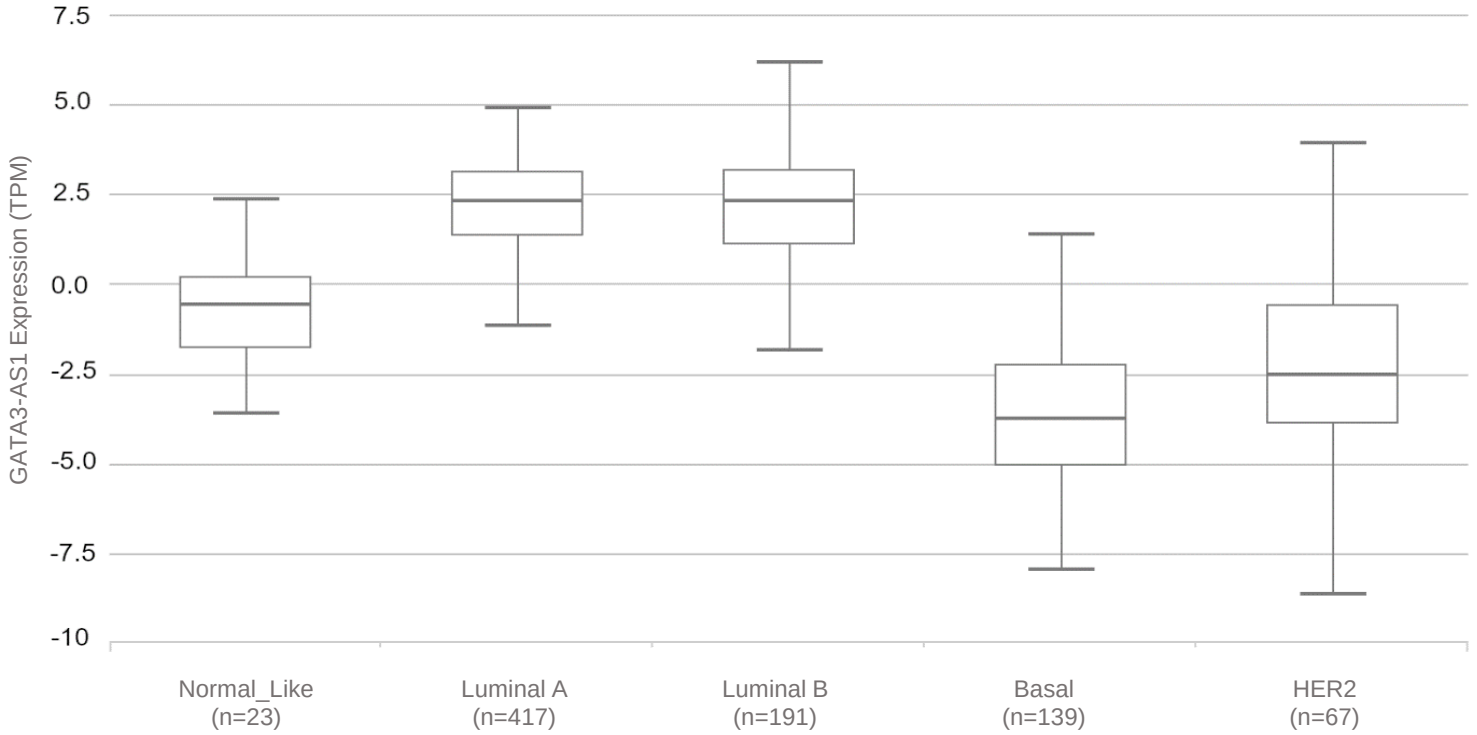
expression profile in TCGA breast cancer patients. Sci Rep 2021, 11:1679
68. Asch-Kendrick R, Cimino-Mathews A: The role of GATA3 in breast carcinomas: a review. Hum Pathol 2016, 48:37−47
69. Wasserman JK, Williams PA, Islam S, Robertson SJ: GATA-3 expression is not associated with complete pathological response in triple negative breast cancer patients treated with neoadjuvant chemotherapy. Pathol Res Pract 2016, 212:539−544
70. Tominaga N, Naoi Y, Shimazu K, Nakayama T, Maruyama N, Shimomura A, Kim SJ, Tamaki Y, Noguchi S: Clinicopathological analysis of GATA3-positive breast cancers with special reference to response to neoadjuvant chemotherapy. Ann Oncol 2012, 23: 3051−3057
71. Van Bockstal MR, Noel F, Guiot Y, Duhoux FP, Mazzeo F, Van Marcke C, Fellah L, Ledoux B, Berlière M, Galant C: Predictive markers for pathological complete response after neo-adjuvant chemotherapy in triple-negative breast cancer. Ann Diagn Pathol 2020, 49:151634
72. Zhang M, Wang N, Song P, Fu Y, Ren Y, Li Z, Wang J: LncRNA GATA3-AS1 facilitates tumour progression and immune escape in triple-negative breast cancer through destabilization of GATA3 but stabilization of PD-L1. Cell Prolif 2020, 53:e12855
73. Tan R, Li H, Huang Z, Zhou Y, Tao M, Gao X, Xu Y: Neural functions play different roles in triple negative breast cancer (TNBC) and non-TNBC. Sci Rep 2020, 10:3065
74. Diermeier SD, Chang K-C, Freier SM, Song J, El Demerdash O, Krasnitz A, Rigo F, Bennett CF, Spector DL: Mammary tumor-associated RNAs impact tumor cell proliferation, invasion, and migration. Cell Rep 2016, 17:261−274
75. Li Y, Zhang Y, Li S, Lu J, Chen J, Wang Y, Li Y, Xu J, Li X: Genome-wide DNA methylome analysis reveals epigenetically dysregulated non-coding RNAs in human breast cancer. Sci Rep 2015, 5:8790
76. Dravis C, Chung C-Y, Lytle NK, Herrera-Valdez J, Luna G, Trejo CL, Reya T, Wahl GM: Epigenetic and transcriptomic profiling of mammary gland development and tumor models disclose regulators of cell state plasticity. Cancer Cell 2018, 34: 466−482.e6
77. Pian L, Wen X, Kang L, Li Z, Nie Y, Du Z, Yu D, Zhou L, Jia L, Chen N, Li D, Zhang S, Li W, Hoffman AR, Sun J, Cui J, Hu J-F: Targeting the IGF1R pathway in breast cancer using antisense lncRNA-mediated promoter cis competition. Mol Ther Nucleic Acids 2018, 12:105−117
78. Huang P, Li F, Li L, You Y, Luo S, Dong Z, Gao Q, Wu S, Brünner N, Stenvang J: lncRNA profile study reveals the mRNAs and lncRNAs associated with docetaxel resistance in breast cancer cells. Sci Rep 2018, 8:17970
79. Sun J, Chen X, Wang Z, Guo M, Shi H, Wang X, Cheng L, Zhou M: A potential prognostic long non-coding RNA signature to predict metastasis-free survival of breast cancer patients. Sci Rep 2015, 5: 16553
80. Gooding AJ, Zhang B, Jahanbani FK, Gilmore HL, Chang JC, Valadkhan S, Schiemann WP: The lncRNA BORG drives breast cancer metastasis and disease recurrence. Sci Rep 2017, 7:12698
81. Zhang H, Nestor CE, Zhao S, Lentini A, Bohle B, Benson M, Wang H: Profiling of human CD4+ T-cell subsets identifies the TH2-specific noncoding RNA GATA3-AS1. J Allergy Clin Immunol 2013, 132:1005−1008
82. Luo X, Zhou N, Wang L, Zeng Q, Tang H: Long noncoding RNA GATA3-AS1 promotes cell proliferation and metastasis in hepato-cellular carcinoma by suppression of PTEN, CDKN1A, and TP53. Can J Gastroenterol Hepatol 2019, 2019:1389653

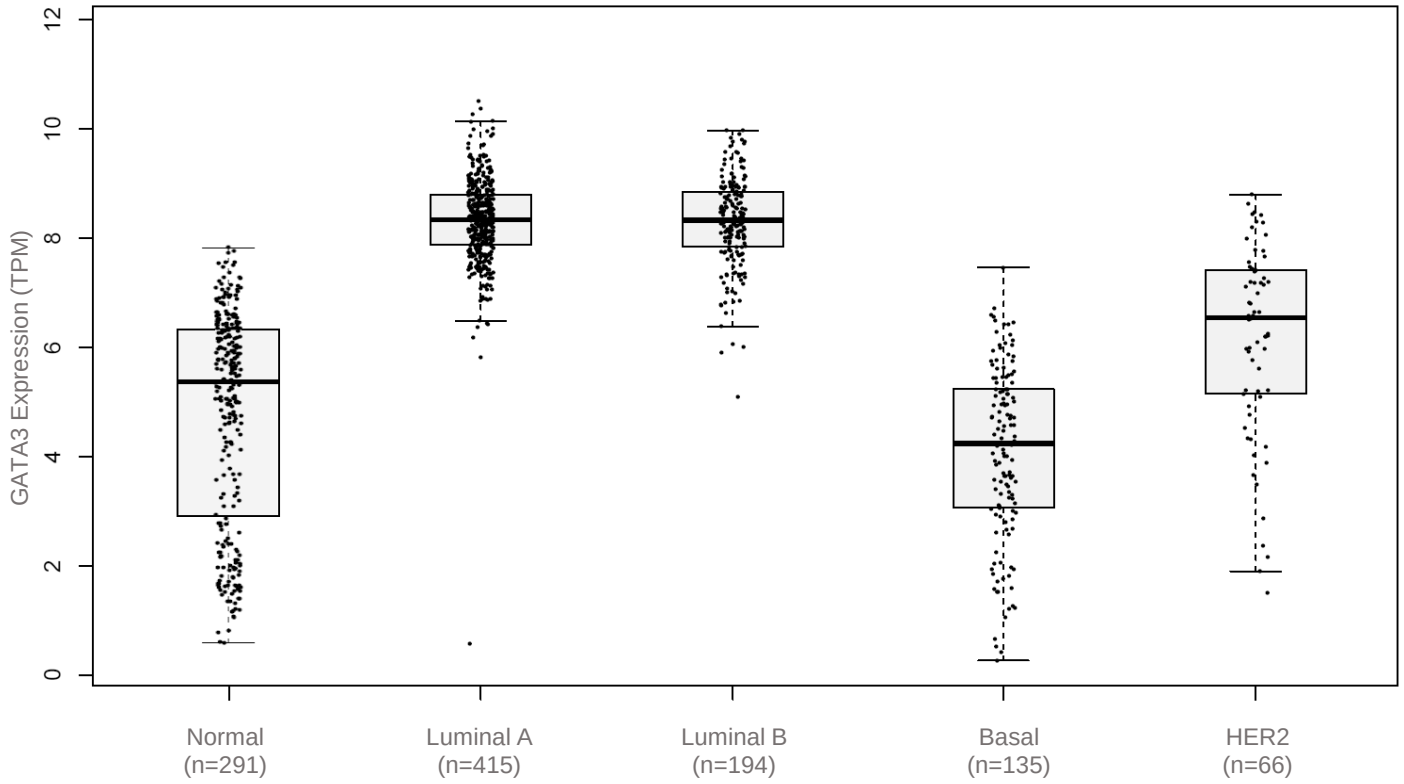

A
PCA − Whole transcriptome
PC2: 15% variance
PC1: 21% variance
BC6
BC4
BC12
BC9
BC7
BC3
BC2
BC10
BC11
BC5
BC1
B
PC3: 12% variance
PC2: 15% variance
BC1
BC7
BC12
BC9
BC6
BC4
BC5
BC10
BC11
BC3
BC2
ER_Her2_status
ER−/Her2−
ER−/Her2+
ER+/Her2−
ER+/Her2+
Response
Nonresponders
Responders


A
ENST00000643001.1
ENST00000461472.1
ENST00000379328.9
ENST00000481743.2
ENST00000346208.4
Nonresponders
BC3
BC5
BC6
BC4
BC2
BC1
BC12
Responders
BC7
BC9
BC10
BC11
8045300 8047720 8049140 8050560 8051980 8053400 8054820 8056240 8057660 8059080 8060500 8061920 8063340 8064760 8066180 8067600 8069020 8070440 8071860 8073280 8075150
B
GATA3 Expression log2(TPM+1)
0.0
2.5
5.0
7.5
Response
No response
Response
BC10
BC11
BC12
BC1
BC7
BC2
BC9
BC5
BC6
BC4
BC3
C
ENST00000643001.1
ENST00000461472.1
ENST00000379328.9
ENST00000481743.2
ENST00000346208.4
T47D
MCF-7
MDA-MB-134-IV
HCC1500
HCC1428
CAMA-1
CAL-85-1
HCC1143
CAL-120
MDA-MB-231
BT-474
MCF-10A
8045300 8047720 8049140 8050560 8051980 8053400 8054820 8056240 8057660 8059080 8060500 8061920 8063340 8064760 8066180 8067600 8069020 8070440 8071860 8073280 8075150
D
GATA3 Expression (RPKM)
0
2
4
6
8
MDAMB157
HCC1395
MDAMB436
CAL51
HME1
HS578T
SUM1315
MDAMB231
BT549
SKBR3
SUM159
MCF10A
MDAMB468
HCC1143
SUM149
PDX1258
HCC1937
HCC1806
HCC38
PDX1328
HCC1954
PDXHCI002
HCC70
BT20
CAL851
CAL20
MDAMB361
HCC1419
HCC1500
HCC1428
MDAMB453
CAMA1
MDAMB134
T47D
MCF7

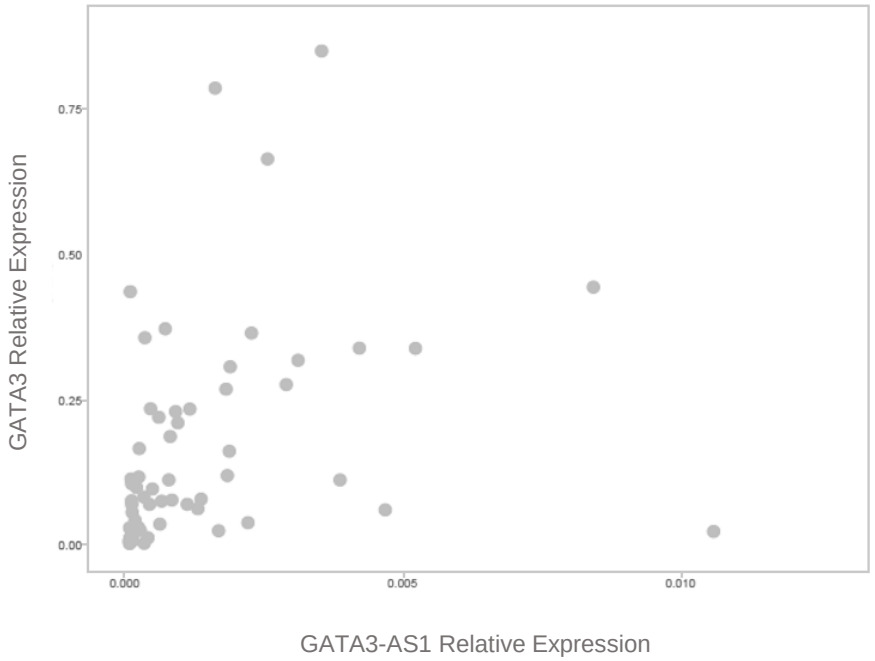

GATA3 Expression (TPM)
0
2
4
6
8
10
12
ACC
BLCA
BRCA
CESC
CHOL
COAD
DLBC
ESCA
GBM
HNSC
KICH
KIRC
KIRP
LAML
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
READ
SARC
SKCM
STAD
TGCT
THCA
THYM
UCEC
UCS
UVM


GATA3-AS1 Expression (FPKM)
0
20
40
60
80
Breast Cancer
Breast Normal
Prostate Cancer
Prostate Normal
Bladder Cancer
Bladder Normal
Renal Clear Cell Carcinoma
Renal Papillary Cell Carcinoma
Renal Cell Carcinoma Chromophobe
Kidney Normal
Thyroid Cancer
Thyroid Normal
Liver Cancer
Liver Normal
Stomach Cancer
Stomach Normal
Lung Adenocarcinoma
Lung Squamous Cell Carcinoma
Lung Normal
Head and Neck Cancer
Head and Neck Normal
Cervical Cancer
Melanoma
Acute Myeloid Leukemia
Chronic Myeloid Leukemia
Myeloproliferative Neoplasms
Colorectal Cancer
Ovarian Cancer
Uterine Cancer
Medulloblastoma
Lower Grade Glioma
Glioblastoma Multiforme
Pancreatic Cancer
Embryonic Stem Cells
Heart
Skeletal Muscle

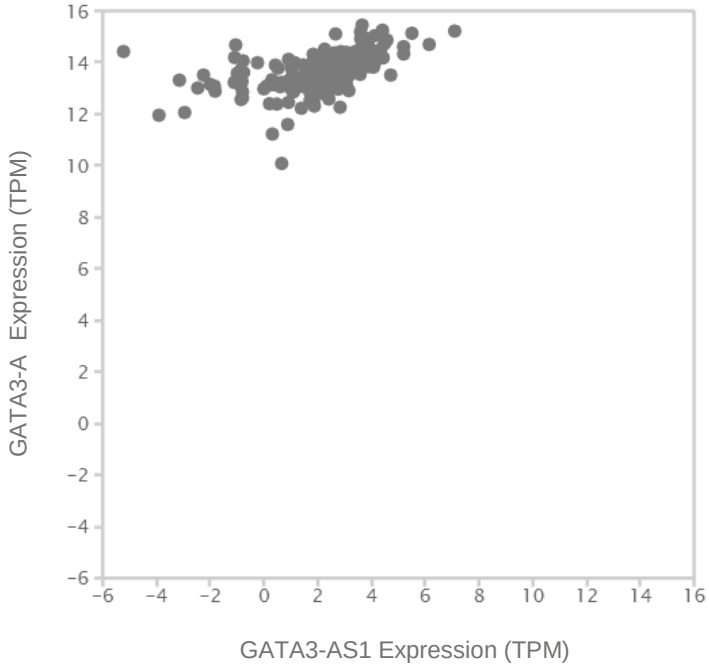

7.5
5.0
2.5
0.0
-2.5
-5.0
-7.5
-10
GATA3-AS1 Expression (TPM)
Normal_Like (n=23)
Luminal A (n=417)
Luminal B (n=191)
Basal (n=139)
HER2 (n=67)

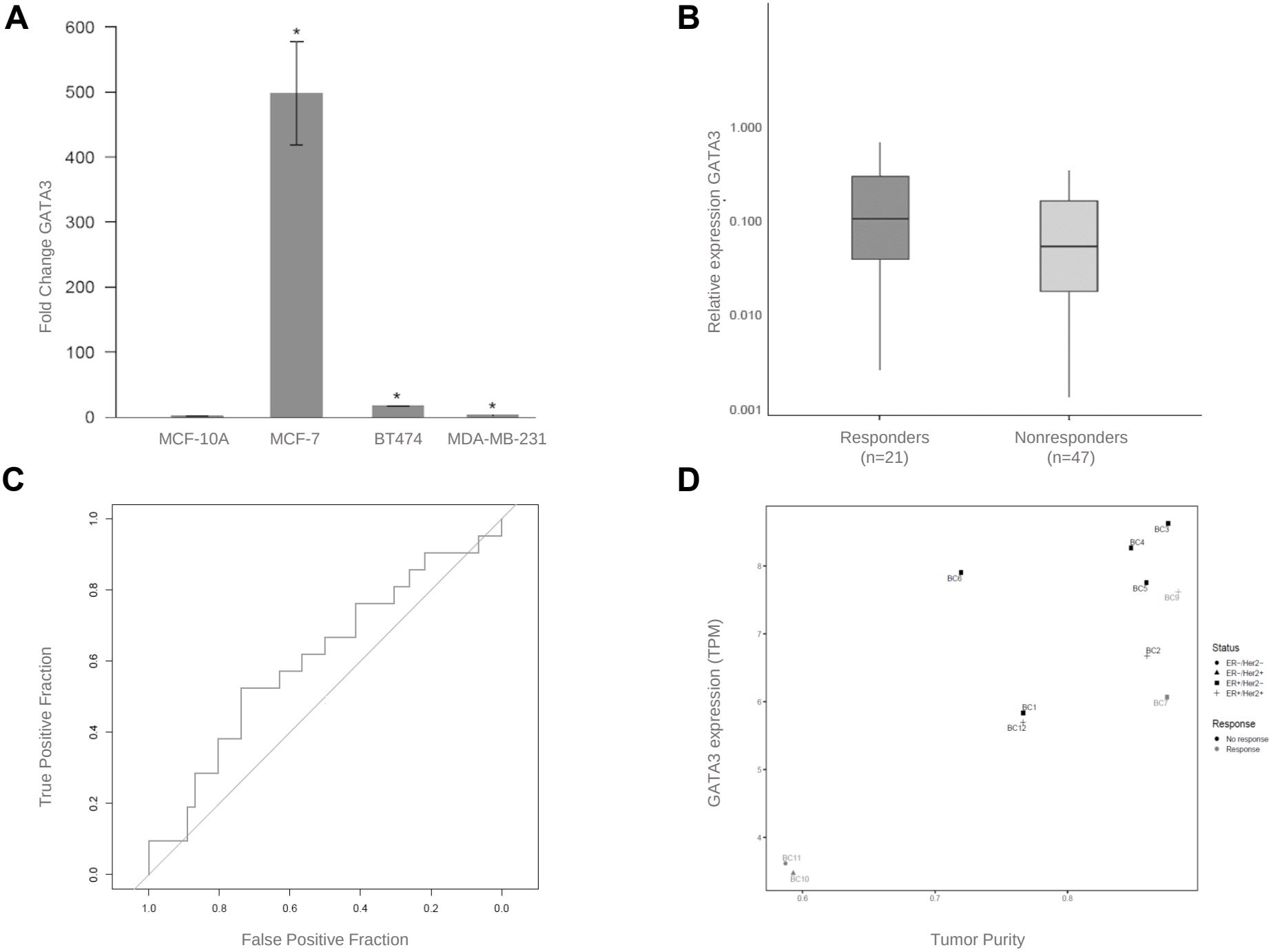

GATA3 Expression (TPM)
0
2
4
6
8
10
12
Normal
(n=291)
Luminal A
(n=415)
Luminal B
(n=194)
Basal
(n=135)
HER2
(n=66)

A


GATA3 Relative Expression
0.75
0.50
0.25
0.00
0.000
0.005
0.010
GATA3-AS1 Relative Expression

B

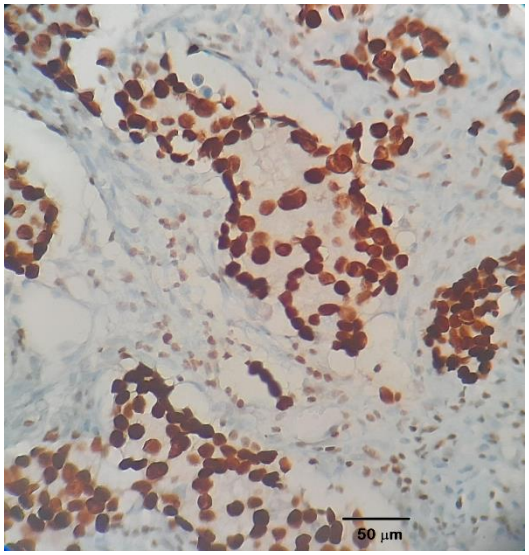

GATA3-A Expression (TPM)
16
14
12
10
8
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
8
10
12
14
16
GATA3-AS1 Expression (TPM)


A
Fold Change GATA3
600
500
400
300
200
100
0
MCF-10A
MCF-7
BT474
MDA-MB-231
B
Relative expression GATA3
1.000
0.100
0.010
0.001
Responders
(n=21)
Nonresponders
(n=47)
C
True Positive Fraction
False Positive Fraction
1.0
0.8
0.6
0.4
0.2
0.0
D
GATA3 expression (TPM)
Tumor Purity
BC3
BC4
BC6
BC5
BC9
BC2
BC1
BC12
BC7
BC11
BC10
Status
ER−/Her2−
ER−/Her2+
ER+/Her2−
ER+/Her2+
Response
No response
Response

A

Responder Patient

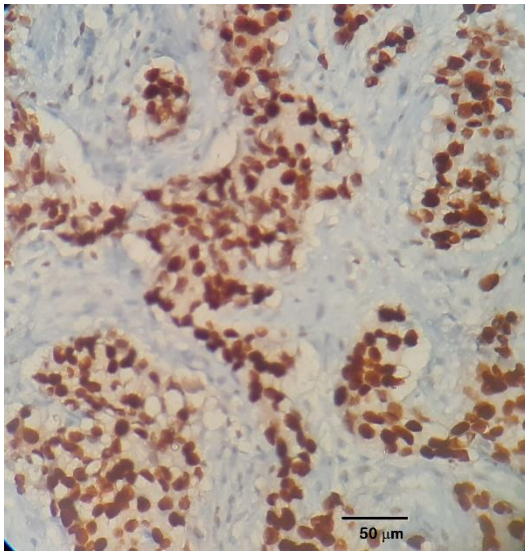
GATA3



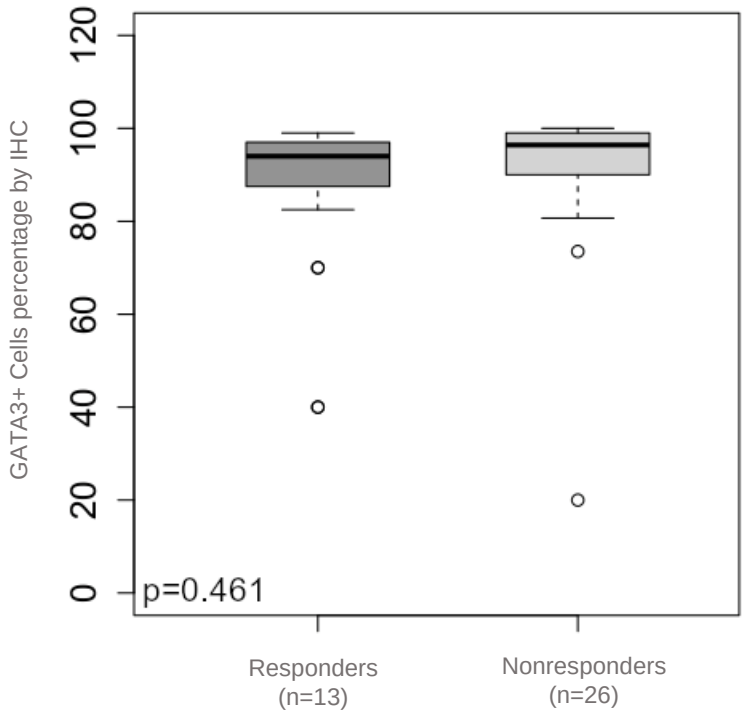
B

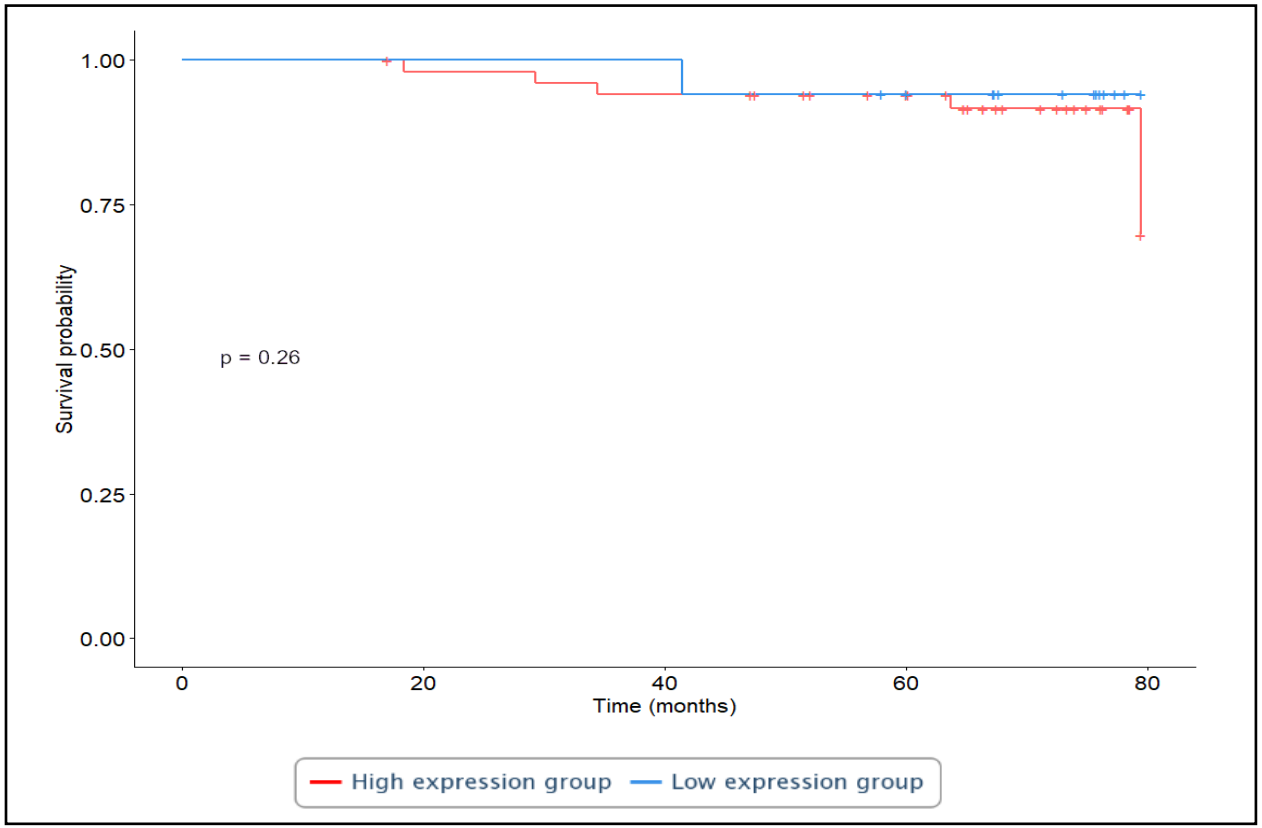
Nonresponder Patient

GATA3

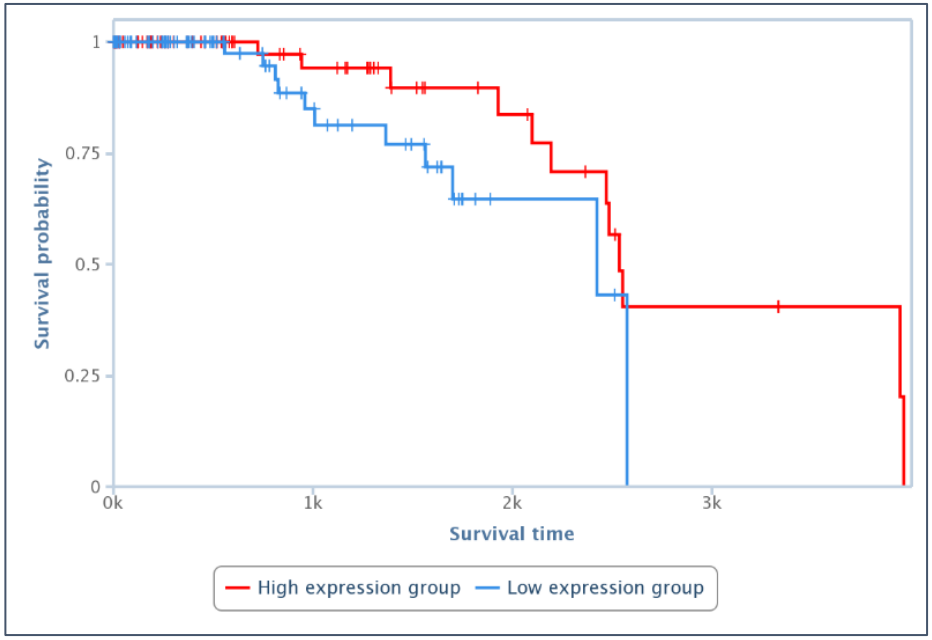


C

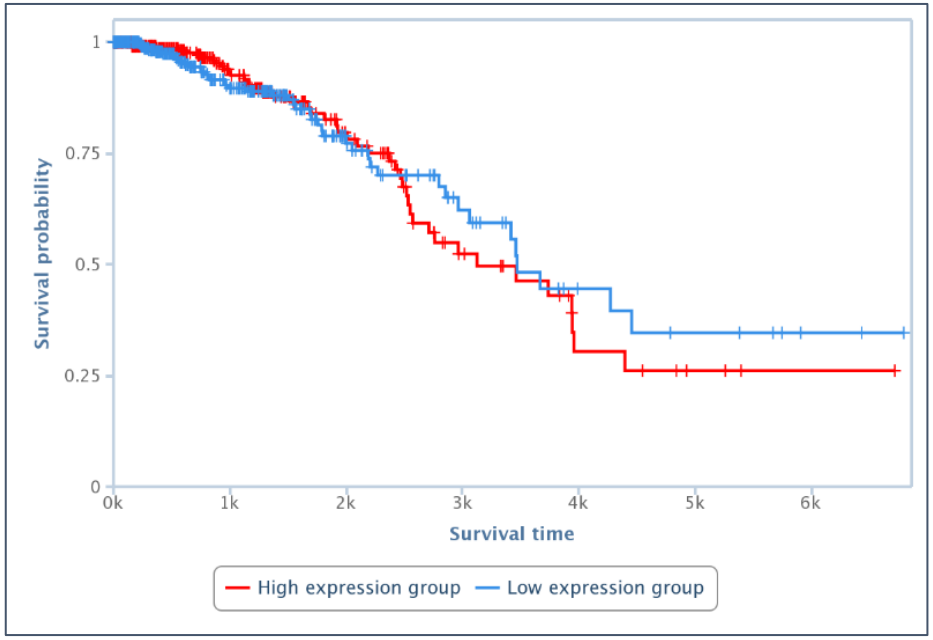



1.00
0.75
0.50
0.25
0.00
Survival probability
p = 0.26
0
20
40
60
80
Time (months)
High expression group
Low expression group

A


1
0.75
0.5
0.25
0
Survival probability
0k
1k
2k
3k
Survival time
High expression group
Low expression group

B


1
0.75
0.5
0.25
0
Survival probability
0k
1k
2k
3k
4k
5k
6k
Survival time
High expression group
Low expression group

Supplemental Table 1. Top 20 differentially expressed lncRNAs in response to neoadjuvant chemotherapy.

Symbol	**Localization***	**Length* (bases)**	**log2FC****	**Adjusted *p value******	**Strand**	**Functions▪**
Overexpressed						
*DBET*****	Chr4: 190,064,502-190,067,864	3363	22.56	3.68e-05	Forward	Chromatin remodeling via interaction with ASH1L
AC127459.1****	Chr16: 23,061,406-23,064,173	2768	9.27	4.37e-03	Reverse	Interacting in ceRNA network disturbed by rs3103067 SNP

SIRLNT	Chr8: 40,298,697-40,353,133	924	8.09	4.17e-02	Reverse	Tumor promotion by regulating miR-4766-5p/SIRT1 axis
AF111169.1****	Chr14: 76,778,952-76,782,249	721	6.83	4.37e-02	Forward	Predicted interaction with 8 miRNAs, included has-miR-188-5p
LINC02432^{+}	Chr4: 141,302,910-141,357,096	3104	6.29	2.25e-02	Reverse	NA

*TRPM2-AS*****	Chr21: 44,414,588-44,425,272	2056	6.28	3.61e-02	Reverse	Promote gastric cancer progression through sponge miR-612
AL390294.1	Chr10: 8,051,541-8,053,084	402	5.48	2.71e-02	Forward	Predicted interaction with has-miR-9-5p
AC018816.1	Chr3: 4,814,294-4,887,293	509	5.44	3.61e-02	Reverse	NA

GATA3-AS1	Chr10: 8,050,450-8,053,484	2214	3.02	4.15e-02	Reverse	Regulation of *GATA3* protein coding gene expression during T cell differentiation. Negative regulation of cell proliferation. Related to renal cell carcinoma.

AP002807.1****	Chr11: 68,050,740-68,053,762	791	2.72	3.73e-02	Forward	Predicted interaction with 3 miRNAs, included has-miR-107
Underexpressed						
LINC00458	Chr13: 54,115,783-54,142,319	1060	-9.82	2.24e-07	Reverse	Regulates endodermal lineage specification via SMAD2/3. Overexpressed in HER2-positive breast cancer. Androgen receptor signaling.

ESRG	Chr3: 54,632,122-54,639,857	3140	-8.72	7.37e-07	Reverse	Related to embryonal carcinoma
LINC01695	Chr21: 28,116,094-28,228,667	4486	-8.69	1.31e-21	Reverse	Interacting in ceRNA network disturbed by rs401431 SNP
AC008514.1	Chr5: 170,747,047-170,788,650	602	-7.44	1.11e-03	Reverse	NA
AC009055.2	Chr16: 65,190,973-65,234,914	1337	-7.37	1.89e-05	Reverse	Interacting in ceRNA network disturbed by rs4785881 SNP

LINC01844	Chr5: 142,716,229-142,761,035	2855	-7.33	1.55e-03	Forward	Related to Hirschsprung disease
PINCR	ChrX: 43,176,994-43,226,598	2228	-7.06	6.23e-03	Forward	Regulates p53 gene target, such as *BTG2, RRM2B* and *GPX1*.
LINC01356	Chr1: 112,820,170-112,850,643	1819	-6.98	2.36e-05	Reverse	Negative regulation of apoptosis. Interacting in ceRNA network disturbed by 1p13.2 CNV

AP000943.2	Chr11: 94,545,330-94,740,355	954	-6.85	2.03e-03	Reverse	Predicted interaction with has-miR-145-5p
AC073316.1	Chr7: 3,140,234-3,174,654	988	-6.74	2.30e-02	Reverse	NA

*Localization and length data were obtained from Ensembl (https://m.ensembl.org/index.html, last accessed February 25, 2021).

**log2FC = Log2 Fold Change

***Adjusted *P*-value= FDR

****Identified only with Salmon + DESeq2 Analysis

[+]Identified only with STAR + feature counts (fc) + DESeq2 Analysis.

▪ Function data were obtained from FARNA database (http://cbrc.kaust.edu.sa/farna, last accessed February 25, 2021).

Supplemental Table 2. Top differentially expressed mRNAs in response to neoadjuvant chemotherapy identify by Salmon + DESeq2 and STAR + fc + DESeq2 pipelines.

Symbol	Localization*	Length* (bases)	log2FC** Salmon + DESeq2	Adjusted P value***	log2FC** STAR + fc + DESeq2	Adjusted P value***	Strand	Functions*
Overexpressed								
PRM2	chr16:11,275,639-11,276,480	680	NA	NA	22.334	0.000	Reverse	Histone subtitute in spermatogenesis
NKX2-2	chr20:21,511,010-21,514,064	2123	NA	NA	21.017	0.000	Reverse	Transcriptional activator
ZSCAN1	chr19: 58,034,025-58,054,631	2095	8.212	0.012	7.279	0.017	Forward	Transcriptional regulation
BMPR1B	chr4: 94,757,955-95,158,453	5580	7.611	0.003	7.099	0.005	Forward	Transmembrane serine/threonine kinase receptor involved in embryogenesis

SCGN	chr6: 25,652,201-25,701,783	704	6.981	0.022	6.512	0.038	Forward	Involved in calcium flux and cell proliferation
MKX	chr10: 27,672,874-27,746,060	3846	6.376	0.013	5.957	0.017	Reverse	Cell adhesion
RUNDC3A	chr17: 44,308,413-44,318,670	1821	6.708	0.006	5.679	0.021	Forward	Associated with malignant pineal area germ cell neoplasia
SIX2	chr2: 45,005,182-45,009,645	2206	5.662	0.042	5.538	0.045	Reverse	Eye development
CYP4X1	chr1: 46,961,364-47,055,432	2256	5.405	0.003	5.459	0.003	Forward	Drug metabolism
VWA5B2	chr3: 184,229,593-184,242,329	3486	6.822	0.004	5.416	0.012	Forward	Unknown function
PTGER2	chr14:52,314,305-52,328,598	2458	NA	NA	5.408	0.046	Forward	Muscular relaxing
UNC5C	chr4:95,162,504-95,549,210	9642	NA	NA	5.362	0.023	Reverse	Axon guidance
CFAP74	chr1: 1,921,951-2,003,837	3476	2.642	0.026	4.917	0.023	Reverse	Cilium movement
ZNF835	chr19: 56,661,980-	3448	3.839	0.046	4.664	0.018	Reverse	Transcriptional regulation

	56,671,783							
DOK7	chr4: 3,463,306-3,501,482	552	4.658	0.013	4.535	0.014	Forward	Neuromuscular synaptogenesis
SIM2	chr21: 36,699,115-36,750,219	4461	5.510	0.023	4.466	0.006	Forward	Regulator of neurogenesis
HOXB5	chr17 48,591,257-48,593,96	2041	4.535	0.043	4.439	0.049	Reverse	Sequence-specific transcription factor involved in developmental regulatory system
DUSP4	chr8: 29,333,062-29,350,684	5553	3.979	0.046	3.946	0.042	Reverse	Mitogenic signal transduction regulator
LRRC6	chr8: 132,570,416-132,685,039	1672	3.136	0.004	3.175	0.004	Reverse	Cilia Motility
NAT8L	chr4: 2,059,327-2,069,089	6056	3.153	0.022	3.120	0.020	Forward	Lipogenesis regulation
NELL2	chr12: 44,508,275-44,921,848	2943	3.02	0.024	2.997	0.034	Reverse	Neuron survival
SPEF1	chr20: 3,777,504-3,781,448	1580	2.748	0.016	2.586	0.027	Reverse	Microtubule-associated protein

RORC	chr1: 151,806,071-151,832,451	2996	2.451	0.046	2.379	0.033	Reverse	Regulator of cellular differentiation and immunity
ZMYND10	chr3: 50,341,110-50,345,732	1774	2.822	0.002	2.374	0.045	Reverse	Transcription regulation and motility
NDUFAF3	chr3: 49,020,459-49,023,495	705	2.332	0.002	2.266	0.002	Forward	Essential factor of mitochondrial NADH:ubiquinone oxidoreductase complex
HSPA4L	chr4:127,781,796-127,840,733	10608	NA	NA	2.266	0.045	Forward	Chaperone activity
ABHD14B	chr3: 51,968,510-51,983,409	2056	2.015	0.01	2.207	0.005	Reverse	Transcription activator
TEX264	chr3: 51,662,693-51,704,323	1340	2.292	2.46E-07	2.123	0.000	Forward	Involved in membrane fission
SDSL	chr12: 113,422,265-113,438,277	1397	2.087	0.023	2.025	0.042	Forward	Serine dehydratase and threonine dehydratase activity
GLT8D1	chr3:52,694,486-52,706,032	1856	1.978	0.046	2.009	0.022	Reverse	Transferase activity
SPCS1	chr3:52,704,955-	1082	2.023	0.031	1.993	0.027	Forward	Processing of

	52,711,148							signal peptides
ATP6AP1L	chr5:82,278,447-82,386,977	2503	NA	NA	1.988	0.044	Forward	Involved in breast cancer
ORMDL3	chr17:39,921,041-39,927,601	2109	1.893936	0.023358	1.980	0.011	Reverse	Negative regulator of sphingolipid synthesis
ESM1	chr5:54,977,867-55,022,671	1326	NA	NA	1.960	0.021	Reverse	Involved in angiogenesis
ABHD14A	chr3:51,971,426-51,981,199	1066	NA	NA	1.912	0.048	Forward	Granule neuron development
NPRL2	chr3:50,347,330-50,350,826	1542	1.676	0.001	1.880	0.000	Reverse	Amino-acid sensing
SLC25A20	chr3:48,856,926-48,898,904	1778	1.669	0.050	1.857	0.031	Reverse	Mitocondrial transport
ZNF329	chr19:58,126,248-58,155,169	3290	1.938	0.004	1.843	0.011	Reverse	Transcriptional regulation
CCDC12	chr3:46,916,310-46,982,083	1017	1.764	0.029	1.778	0.013	Reverse	Protein binding
ZNF442	chr19:12,345,944-12,372,636	6274	NA	NA	1.736	0.013	Reverse	Transcriptional regulation
LZTFL1	chr3:45,823,316-45,916,042	4026	NA	NA	1.733	0.028	Reverse	Tumor suppressor
C1orf56	chr1:151,047,751-151,051,986	2085	NA	NA	1.695	0.006	Forward	Cellular proliferation

								control
TUSC2	chr3:50,320,027-50,328,251	1669	1.671	0.008	1.673	0.010	Reverse	Tumor suppressor
ARL6IP5	chr3:69,084,937-69,106,092	2134	1.811	0.000	1.653	0.009	Forward	Regulates intracellular concentrations of taurine and glutamate
BCDIN3D	chr12:49,836,043-49,843,106	3226	NA	NA	1.642	0.020	Reverse	miRNA processing
TMEM115	chr3:50,354,750-50,359,521	2107	1.563	0.008	1.620	0.012	Reverse	Protein transport and glycosylation in Golgi
CYB561D2	chr3:50,350,845-50,368,197	1225	1.826	0.006	1.604	0.002	Forward	Metabolism
SELENOK	chr3:53,884,417-53,891,962	340	1.669	0.012	1.550	0.016	Reverse	Stress response
GADD45B	chr19:2,476,122-2,478,259	1371	1.575	0.032	1.539	0.031	Forward	Growth and apoptosis
PMVK	chr1:154,924,740-154,942,658	1002	1.520	0.000	1.537	0.000	Reverse	Metabolism
Underexpressed								
SMIM34B	chr21: 7,784,482-7,793,954	1050	-10.125	0.000	-8.994	0.000	Reverse	Unknown function

CST4	chr20: 23,685,640-23,689,040	751	-10.125	0.000	-8.994	0.000	Reverse	Proteinase inhibitor
CLDN6	chr16: 3,014,712-3,020,071	1739	-9.129	0.000	-8.956	0.000	Reverse	Involved in tight junction
S100G	chrX: 16,649,787-16,654,674	455	-8.606	0.000	-8.333	0.000	Forward	Mineral absorption
LEFTY2	chr1: 225,936,598-225,941,492	2019	-8.648	0.000	-8.296	0.000	Reverse	Endometrial bleeding
POU5F1	chr6: 31,164,337-31,180,731	1409	-7.927	0.000	-7.954	0.000	Reverse	Early embryogenesis and embryonic stem cell pluripotency
SCN1A	chr2: 165,984,641-166,149,214	8533	-7.347	0.000	-7.891	0.000	Reverse	Neurotransmitters released
DPPA4	chr3: 109,326,141-109,339,635	2817	-6.834	0.000	-7.665	0.041	Reverse	Maintenance of active epigenetic status of target genes
TDGF1	chr3: 46,574,534-46,582,457	1956	-7.862	0.000	-7.644	0.000	Forward	Nodal signaling
CNMD	chr13: 52,703,264-	1444	-7.684	0.005	-7.513	0.002	Reverse	Growth regulation

	52,739,820							
TRIM71	chr3: 32,817,997-32,897,824	8704	-7.641	0.000	-7.413	0.001	Forward	Embryonic cell proliferation and mainteinance
LIN28A	chr1: 26,410,778-26,429,728	3458	-7.500	0.000	-7.410	0.000	Forward	Regulation of mRNAs and miRNAs involved in pluripotency and metabolism
LEFTY1	chr1: 225,886,282-225,911,382	1626	-6.817	0.001	-7.221	0.000	Reverse	Regulator of Nodal signaling
NTS	chr12: 85,874,295-85,882,992	1239	-7.070	0.001	-6.970	0.001	Forward	Regulation of fat metabolism
PTPRZ1	chr7: 121,873,089-122,062,036	8103	-5.784	0.006	-6.955	0.000	Forward	Central nervous system development
VRTN	chr14: 74,303,069-74,360,008	3408	-6.420	0.013	-6.714	0.004	Forward	Negative regulation of transcription
MYL7	chr7: 44,138,864-44,141,392	612	-6.059	0.006	-6.649	0.001	Reverse	Muscle contraction
PLA2G3	chr22: 31,134,807-31,140,508	2600	-6.638	0.000	-6.614	0.000	Reverse	Ciliogenesis

IGF2BP1	chr17: 48,997,385-49,056,145	8796	-4.339	0.036	-6.604	0.005	Forward	Regulation of gene expression
PAGE2	chrX: 55,089,018-55,092,842	456	-6.228	0.010	-6.338	0.009	Forward	Unknown function
NANOG	chr12: 7,787,794-7,799,146	5182	-6.829	0.000	-6.329	0.014	Forward	Embryonic stem cell proliferation and self-renewal
FGF17	chr8: 22,039,708-22,048,809	1320	-6.579	0.003	-6.263	0.004	Forward	Embryonic development

*Localization, length and function data were obtained from GeneCards (https://www.genecards.org/, last accessed April 08, 2021).

**log2FC = Log2 Fold Change.

***Adjusted *P*-value= FDR.

NA = Not identified with this pipeline

Supplemental Table 3. List of lncRNAs underexpressed in nonresponders patients, obtained by two different bioinformatic pipelines.

Symbol*	Localization**	Strand
Salmon + DESeq2		
LINC01405	chr12: 110,934,590-110,959,093	Forward
AC012368.1	chr2: 64,143,239-64,252,859	Forward
lnc-WDR4	chr21:42,827,613-42,838,647	Reverse
LINC00882	chr3:106449775-107240671	Reverse
AC022140.1	chr5:25404733-25445925	Reverse
AC022101.1	chr5:122311740-122423078	Forward
AC104257.1	chr8:131308545-131317632	Forward
AC087354.1	chr8:139460062-139463016	Reverse
LINC00678	chr11:27617626-27634627	Reverse
AP000446.1	chr11:83286120-83423516	Forward
LINC02253	chr15:97215812-97432094	Forward
LINC01443	chr18:14946267-14974215	Forward
AL513318.2	chr9:87008440-87042419	Reverse
AP000357.2	chr22:24691122-24696003	Reverse
AP006261.1	chr18:14404551-14430923	Reverse
AC018865.1	chr2:130082092-130082220	Reverse
AL353751.1	chr10:89283636-89292125	Forward
LINC00707	chr10:6779549-6879450	Forward
AC008517.1	chr5:92823935-92844992	Forward
MIR302CHG	chr4:112646476-112650051	Reverse
ERVH−1	chr4: 23,723,262-23,733,579	Reverse
AC079296.1	chr11:9004093-9067776	Forward
AC064802.1	chr8:114282067-114295839	Forward
AC093496.1	chr3:14272373-14303845	Reverse
AC016044.1	chr15:52800168-52805972	Reverse
LINC02582	chr18:73324941-73349879	Forward
DNAH17−AS1	chr17:78484882-78503056	Forward
AL512624.2	chr14:19,420,975-19,425,017	Forward
AL109809.5	chr20:1,729,038-1,817,765	Reverse
AC064869.1	chr2:107698737-107746941	Reverse
B3GALT5−AS1	chr21:39597147-39612910	Reverse
LINC01139	chr1:238,476,542-238,486,060	Reverse

AC073316.1	chr7:3140234-3174654	Reverse
LINC02620	chr10:104,474,939-104,480,274	Reverse
AL035409.1	chr1:77081984-77086402	Forward
LNCPRESS1	chr7:101,299,578-101,301,346	Forward
LINC01287	chr7:153,355,365-153,413,985	Reverse
LINC01695	chr21:28116094-28228667	Reverse
LINC01844	chr5:142716229-142761035	Forward
AC090204.1	chr8:32927913-33045445	Forward
AC100801.1	chr8:85833377-85951083	Forward
AP000943.2	chr11:94545330-94740355	Reverse
LINC02303	chr14:45706250-45715952	Reverse
AL356804.1	chr14:70255629-70343388	Forward
ESRG	chr3:54632122-54639857	Reverse
ERVK−28	chr19:27,638,483-27,646,483	Reverse
PCAT14	chr22:23536881-23547797	Forward
SNHG4	chr5:139274102-139284899	Forward
AC007326.5	chr22:18936411-18947741	Forward
AL137800.1	chr1:183613537-183619335	Forward
LINC01446	chr7:53655508-53811952	Reverse
LINC01356	chr1:112820170-112850643	Reverse
PINCR	chrX:43176994-43226598	Forward
AL358473.1	chr1:201023949-201028792	Forward
AC104461.1	chr1:200333193-200478669	Forward
POU6F2−AS2	chr7:38980370-39013551	Reverse
LINC00458	chr113:54115783-54142319	Reverse
AL117378.1	chr6:131901963-131920565	Forward
LINC01238	chr2:241970683-241977276	Forward
AC022424.1	chr5:5142138-5176214	Reverse
AC008514.1	chr5:170747047-170788650	Reverse
AL160191.1	chr14:70187123-70230187	Forward
AL392023.1	chr14:38190983-38202923	Forward
AC009055.2	chr16:65190973-65234914	Reverse
LINC01924	chr18:64041555-64423601	Forward
AL161431.1	chr13:109269634-109278512	Forward
LINC00836	chr10:25651712-25732935	Forward
AL009031.1	chr6:16259101-16264553	Reverse
AC116317.1	chr4:6292369-6308636	Reverse
AL606500.1	chr1:154671593-154678345	Forward
STAR + fc + DESeq2		
C1orf220	chr1:178,542,752-178,548,889	Forward

LINC01194	chr5:12,574,830-12,804,363	Forward
LINC01108	chr6:14,280,127-14,285,454	Reverse
HDAC2−AS2	chr6:113,969,701-114,471,705	Forward
AC002383.1	chr7:89443946-89496918	Forward
AC010998.3	chr10:120984966-120985596	Reverse
AC136475.3	chr11:287305-288987	Forward
AC092490.1	chr12:8788253-8795789	Forward
KRT7−AS	chr12:52,245,048-52,247,448	Reverse
AL354821.1	chr13:55535697-55583524	Reverse
LINC00648	chr14:47,764,954-47,795,302	Reverse
LINC−ROR	chr18:57,054,558-57,072,119	Reverse
KCNMB2−AS1	chr3:178526505-178937352	Reverse
SEMA6A−AS1	chr5:116,447,547-116,508,276	Forward
AL591030.1	chr6:76561328-76562590	Forward
AC005062.1	chr7:19918981-20140453	Reverse
PROSER2−AS1	chr10:11849608-11894700	Reverse
AL731571.1	chr10:124996064-125001491	Forward
AC044810.2	chr11:7754393-7905955	Reverse
AC084816.1	chr12:22699859-23188807	Forward
AC027288.1	chr12:79540203-79550535	Forward
AL392023.2	chr14:38034287-38194281	Reverse
AL512358.2	chr14:48262021-48265609	Reverse
AP001605.1	chr21:27358885-27448579	Reverse
B3GALT5−AS1	chr21:39597147-39612910	Reverse
LINC01139	chr1:238,476,542-238,486,060	Reverse
AC073316.1	chr7:3140234-3174654	Reverse
LINC02620	chr10:104,474,939-104,480,274	Reverse
AL035409.1	chr1:77081984-77086402	Forward
LNCPRESS1	chr7:101,299,578-101,301,346	Forward
LINC01287	chr7:153,355,365-153,413,985	Reverse
LINC01695	chr21:28116094-28228667	Reverse
LINC01844	chr5:142716229-142761035	Forward
AC090204.1	chr8:32927913-33045445	Forward
AC100801.1	chr8:85833377-85951083	Forward
AP000943.2	chr11:94545330-94740355	Reverse
LINC02303	chr14:45706250-45715952	Reverse
AL356804.1	chr14:70255629-70343388	Forward
ESRG	chr3:54632122-54639857	Reverse
ERVK−28	chr19:27,638,483-27,646,483	Reverse
PCAT14	chr22:23536881-23547797	Forward

SNHG4	chr5:139274102-139284899	Forward
AC007326.5	chr22:18936411-18947741	Forward
AL137800.1	chr1:183613537-183619335	Forward
LINC01446	chr7:53655508-53811952	Reverse
LINC01356	chr1:112820170-112850643	Reverse
PINCR	chrX:43176994-43226598	Forward
AL358473.1	chr1:201023949-201028792	Forward
AC104461.1	chr1:200333193-200478669	Forward
POU6F2−AS2	chr7:38980370-39013551	Reverse
LINC00458	chr113:54115783-54142319	Reverse
AL117378.1	chr6:131901963-131920565	Forward
LINC01238	chr2:241970683-241977276	Forward
AC022424.1	chr5:5142138-5176214	Reverse
AC008514.1	chr5:170747047-170788650	Reverse
AL160191.1	chr14:70187123-70230187	Forward
AL392023.1	chr14:38190983-38202923	Forward
AC009055.2	chr16:65190973-65234914	Reverse
LINC01924	chr18:64041555-64423601	Forward
AL161431.1	chr13:109269634-109278512	Forward
LINC00836	chr10:25651712-25732935	Forward
AL009031.1	chr6:16259101-16264553	Reverse
AC116317.1	chr4:6292369-6308636	Reverse
AL606500.1	chr1:154671593-154678345	Forward
Both pipelines		
B3GALT5−AS1	chr21:39597147-39612910	Reverse
LINC01139	chr1:238,476,542-238,486,060	Reverse
AC073316.1	chr7:3140234-3174654	Reverse
LINC02620	chr10:104,474,939-104,480,274	Reverse
AL035409.1	chr1:77081984-77086402	Forward
LNCPRESS1	chr7:101,299,578-101,301,346	Forward
LINC01287	chr7:153,355,365-153,413,985	Reverse
LINC01695	chr21:28116094-28228667	Reverse
LINC01844	chr5:142716229-142761035	Forward
AC090204.1	chr8:32927913-33045445	Forward
AC100801.1	chr8:85833377-85951083	Forward
AP000943.2	chr11:94545330-94740355	Reverse
LINC02303	chr14:45706250-45715952	Reverse
AL356804.1	chr14:70255629-70343388	Forward
ESRG	chr3:54632122-54639857	Reverse
ERVK−28	chr19:27,638,483-27,646,483	Reverse

PCAT14	chr22:23536881-23547797	Forward
SNHG4	chr5:139274102-139284899	Forward
AC007326.5	chr22:18936411-18947741	Forward
AL137800.1	chr1:183613537-183619335	Forward
LINC01446	chr7:53655508-53811952	Reverse
LINC01356	chr1:112820170-112850643	Reverse
PINCR	chrX:43176994-43226598	Forward
AL358473.1	chr1:201023949-201028792	Forward
AC104461.1	chr1:200333193-200478669	Forward
POU6F2−AS2	chr7:38980370-39013551	Reverse
LINC00458	chr113:54115783-54142319	Reverse
AL117378.1	chr6:131901963-131920565	Forward
LINC01238	chr2:241970683-241977276	Forward
AC022424.1	chr5:5142138-5176214	Reverse
AC008514.1	chr5:170747047-170788650	Reverse
AL160191.1	chr14:70187123-70230187	Forward
AL392023.1	chr14:38190983-38202923	Forward
AC009055.2	chr16:65190973-65234914	Reverse
LINC01924	chr18:64041555-64423601	Forward
AL161431.1	chr13:109269634-109278512	Forward
LINC00836	chr10:25651712-25732935	Forward
AL009031.1	chr6:16259101-16264553	Reverse
AC116317.1	chr4:6292369-6308636	Reverse
AL606500.1	chr1:154671593-154678345	Forward

*Symbols were directly consulted and assigned according to HGNC nomenclature rules (https://www.genenames.org/ Last access April 4th, 2021)

** Localization of genes were taken of https://www.ensembl.org/index.html (Last access April 4th, 2021)

Supplemental Table 4. List of lncRNAs overexpressed in nonresponders patients, obtained by two different bioinformatic pipelines.

Symbol*	Localization**	Strand
Salmon + DESeq2		
GATA3−AS1	**chr10:8050450-8053484**	**Reverse**
AL390294.1	chr10:8051541-8053084	Forward
AC018816.1	chr13:4814294-4887293	Reverse
SIRLNT	chr8:40298697-40353133	Reverse
TRPM2−AS	chr21:44414588-44425272	Reverse
AP002807.1	chr11:68050740-68053762	Forward
AF111169.1	chr14:76778952-76782249	Forward
AC125257.1	chr17:41848518-41851447	Reverse
AC127459.1	chr16:23061406-23064173	Reverse
DBET	chr4:190064502-190067864	Forward
STAR + fc + DEseq2		
GATA3−AS1	**chr10:8050450-8053484**	**Reverse**
AL390294.1	chr10:8051541-8053084	Forward
AC018816.1	chr13:4814294-4887293	Reverse
SIRLNT	chr8:40298697-40353133	Reverse
LINC02432	chr4:141302910-141357096	Reverse
AC006115.2	chr19:56669941-56823395	Forward
Both pipelines		
GATA3−AS1	**chr10:8050450-8053484**	**Reverse**
AL390294.1	chr10:8051541-8053084	Forward
AC018816.1	chr13:4814294-4887293	Reverse
SIRLNT	chr8:40298697-40353133	Reverse

*Symbols were directly consulted and assigned according to HGNC nomenclature rules (https://www.genenames.org/ Last access April 4th, 2021).

** Localization of genes were taken of https://www.ensembl.org/index.html (Last access April 4th, 2021).

International Journal of Infectious Diseases 105 (2021) 83–90

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Saliva is a reliable and accessible source for the detection of SARS-CoV-2

Luis A. Herrera[a,b,*], Alfredo Hidalgo-Miranda[c,*], Nancy Reynoso-Noverón[b], Abelardo A. Meneses-García[d], Alfredo Mendoza-Vargas[e], Juan P. Reyes-Grajeda[e], Felipe Vadillo-Ortega[f], Alberto Cedro-Tanda[a], Fernando Peñaloza[b], Emmanuel Frías-Jimenez[a], Cristian Arriaga-Canon[b], Rosaura Ruiz[g], Ofelia Angulo[g], Imelda López-Villaseñor[h], Carlos Amador-Bedolla[i], Diana Vilar-Compte[d], Patricia Cornejo[d], Mireya Cisneros-Villanueva[c], Eduardo Hurtado-Cordova[c], Mariana Cendejas-Orozco[c], José S. Hernández-Morales[e], Bernardo Moreno[b], Irwin A. Hernández-Cruz[b], César A. Herrera[b], Francisco García[b], Miguel A. González-Woge[b], Paulina Munguía-Garza[b], Fernando Luna-Maldonado[b], Antonia Sánchez-Vizcarra[b], Vincent G. Osnaya[b], Nelly Medina-Molotla[e], Yair Alfaro-Mora[b], Rodrigo E. Cáceres-Gutiérrez[b], Laura Tolentino-García[b], Patricia Rosas-Escobar[e], Sergio A. Román-González[a], Marco A. Escobar-Arrazola[b], Julio C. Canseco-Méndez[e], Diana R. Ortiz-Soriano[a], Julieta Domínguez-Ortiz[b], Ana D. González-Barrera[a], Diana I. Aparicio-Bautista[a], Armando Cruz-Rangel[a], Ana Paula Alarcón-Zendejas[b], Laura Contreras-Espinosa[b], Rodrigo González[b], Lissania Guerra-Calderas[b], Marco A. Meraz-Rodríguez[b], Michel Montalvo-Casimiro[b], Rogelio Montiel-Manríquez[b], Karla Torres-Arciga[b], Daniela Venegas[b], Vasti Juárez-González[b], Xiadani Guajardo-Barreto[b], Verónica Monroy-Martínez[h], Daniel Guillén[h], Jacquelina Fernández[h], Juliana Herrera[h], Renato León-Rodriguez[h], Israel Canela-Pérez[h], Blanca H. Ruíz-Ordaz[h], Rafael Valdez-Vazquez[j], Jennifer Bertin-Montoya[j], María Niembro-Ortega[j], Liudmila Villegas-Acosta[j], Daniela López-Castillo[j], Andrea Soriano-Ríos[j], Michael Gastelum-Ramos[j], Tonatiuh Zamora-Barandas[j], Jorge Morales-Baez[j], María García-Rodríguez[k], Mariano García-Martínez[k], Erik Nieto-Patlán[k], Maricarmen Quirasco-Baruch[l], Irma López-Martínez[m], Ernesto Ramírez-Gonzalez[m], Hiram Olivera-Díaz[m], Noe Escobar-Escamilla[m]

[a] *Instituto Nacional de Medicina Genómica, INMEGEN, Mexico City, Mexico*
[b] *Unidad de Investigación Biomédica en Cáncer, Instituto Nacional de Cancerología-Instituto de Investigaciones Biomédicas, UNAM, Mexico City, Mexico*
[c] *Laboratorio de Genómica del Cáncer, Instituto Nacional de Medicina Genómica, INMEGEN, Mexico City, Mexico*
[d] *Instituto Nacional de Cancerología, INCAN, Mexico City, Mexico*
[e] *Unidad de Secuenciación, Instituto Nacional de Medicina Genómica, INMEGEN, Mexico City, Mexico*
[f] *Unidad de Vinculación Científica Facultad de Medicina-INMEGEN, Mexico City, Mexico*
[g] *Secretaria de Educación, Ciencia, Tecnología e Innovación, Mexico City, Mexico*
[h] *Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Mexico City, Mexico*
[i] *Departamento de Física y Química Teórica, Facultad de Química, Universidad Nacional Autónoma de México, Mexico City, Mexico*
[j] *Unidad Temporal COVID-19, Mexico City, Mexico*
[k] *Unidad de Investigación Preclínica, Facultad de Química, Mexico City, Mexico*
[l] *Departamento de Alimentos y Biotecnología, Facultad de Química, Universidad Nacional Autónoma de México, Mexico City, Mexico*
[m] *Instituto de Diagnóstico y Referencia Epidemiológicos, InDRE, Mexico City, Mexico*

* Corresponding authors at: National Institute of Genomic Medicine, Periferico Sur 4809, Arenal Tepepan, 14610 Mexico City, Mexico.
E-mail addresses: lherrera@inmegen.gob.mx (L.A. Herrera), ahidalgo@inmegen.gob.mx (A. Hidalgo-Miranda).

https://doi.org/10.1016/j.ijid.2021.02.009

ARTICLE INFO

Article history:
Received 22 October 2020
Received in revised form 2 February 2021
Accepted 4 February 2021

Keywords:
COVID-19
SARS-CoV-2
Diagnostic test
Saliva testing
Pooling strategy

ABSTRACT

Objectives: The aim of this study was to investigate the feasibility of saliva sampling as a non-invasive and safer tool to detect severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and to compare its reproducibility and sensitivity with nasopharyngeal swab samples (NPS). The use of sample pools was also investigated.

Methods: A total of 2107 paired samples were collected from asymptomatic healthcare and office workers in Mexico City. Sixty of these samples were also analyzed in two other independent laboratories for concordance analysis. Sample processing and analysis of virus genetic material were performed according to standard protocols described elsewhere. A pooling analysis was performed by analyzing the saliva pool and the individual pool components.

Results: The concordance between NPS and saliva results was 95.2% (kappa 0.727, $p = 0.0001$) and 97.9% without considering inconclusive results (kappa 0.852, $p = 0.0001$). Saliva had a lower number of inconclusive results than NPS (0.9% vs 1.9%). Furthermore, saliva showed a significantly higher concentration of both total RNA and viral copies than NPS. Comparison of our results with those of the other two laboratories showed 100% and 97% concordance. Saliva samples are stable without the use of any preservative, and a positive SARS-CoV-2 sample can be detected 5, 10, and 15 days after collection when the sample is stored at 4 °C.

Conclusions: The study results indicate that saliva is as effective as NPS for the identification of SARS-CoV-2-infected asymptomatic patients. Sample pooling facilitates the analysis of a larger number of samples, with the benefit of cost reduction.

Introduction

The rapid spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) worldwide has generated considerable demand for medical supplies for use in fighting the pandemic. Among other problems, this has resulted in a shortage of nasopharyngeal swabs (NPS) and tests for the detection of SARS-CoV-2. Scarce consumables and invasive sample collection, which can expose medical personnel to biohazards, are obstacles to effective mass screening of the population to identify infected individuals. Mass screening is essential to identify and isolate infected individuals during reopening. Additionally, fast massive effective screening is essential in the event of a coronavirus disease 2019 (COVID-19) resurgence and for the safe return to productive activities, an approach that has been implemented by several governments around the globe. Although this situation has been addressed using different innovative approaches, such as three-dimensional printing of NPS (Callahan et al., 2020), additional solutions for sample collection that are easier and less invasive, with minimal risk to health professionals, together with strategies aiming to maximize the number of samples analyzed, must be explored.

The gold standard test for the diagnosis of SARS-CoV-2 infection involves sample collection via NPS, followed by viral RNA extraction and detection by real-time polymerase chain reaction (RT-qPCR). Recent reports have indicated that saliva is a viable option for testing with several potential advantages over NPS, including that it is a less invasive procedure, making it more viable for repeated testing. Furthermore, saliva can be self-collected by the patient with minimal guidance and intervention from healthcare personnel (Azzi et al., 2020). SARS-CoV-2 can be detected in more than 95% of saliva samples, and the virus can be cultured from saliva samples (To et al., 2020b). Detection of the virus in saliva has also been used to monitor viral load dynamics over time, indicating that the highest viral load in saliva presents during the first week after symptom onset and then declines over time (To et al., 2020a). Recently, the Food and Drug Administration (FDA) in the United States approved the first diagnostic test with the option for saliva sampling for SARS-CoV-2 detection (U.S. Food and Drug Administration, 2020a). Another study found that the home-based collection method of saliva, supervised by a clinician, performed similarly to or even better than NPS for the detection of infection (Noah et al., 2020). These findings were confirmed by recent studies, which found that saliva is more sensitive for SARS-CoV-2 detection than NPS in patients with COVID-19 (Wyllie et al., 2020). In another report, 229 paired samples from 95 patients also showed a high concordance and no significant temporal variation in viral load between the two sample types (Cheuk et al., 2020).

The combined advantages offered by saliva sampling and sample pooling result in an inexpensive diagnostic procedure suitable for assaying large numbers of samples, as has been required during the current pandemic (Abdalhamid et al., 2020; Yelin et al., 2020). Sample pooling has proven its efficacy in different applications, including retrospective testing (Hogan et al., 2020) and, more importantly, in large-scale screening of asymptomatic populations (Ben-Ami et al., 2020; Lohse et al., 2020; U.S. Food and Drug Administration, 2020b). There is work showing that pooling saliva samples for the detection of SARS-CoV-2 provides a mechanism to support testing for a greater number of individuals with substantial cost savings, especially at lower prevalence levels (Pasomsub et al., 2020a, 2020b; Watkins et al., 2020). Mirimus Clinical Labs in their SalivaClear test already use the pooling strategy to monitor and detect infections in groups of symptomatic and asymptomatic individuals (SalivaClear by Mirimus Clinical, 2020).

This study was performed to compare the reproducibility, accuracy, and feasibility of saliva sampling using NPS followed by RT-qPCR for the detection of SARS-CoV-2 in paired samples from asymptomatic clinical and laboratory personnel working in two Mexico National Institutes of Health laboratories and from asymptomatic office workers ($N = 2107$ individuals). This study presents evidence that saliva sample pooling is a reliable and inexpensive method that allows for the screening of a large number of samples.

Materials and methods

Participants

A cross-sectional study design was used to collect samples from personnel engaged in clinical and laboratory activities at Mexico's

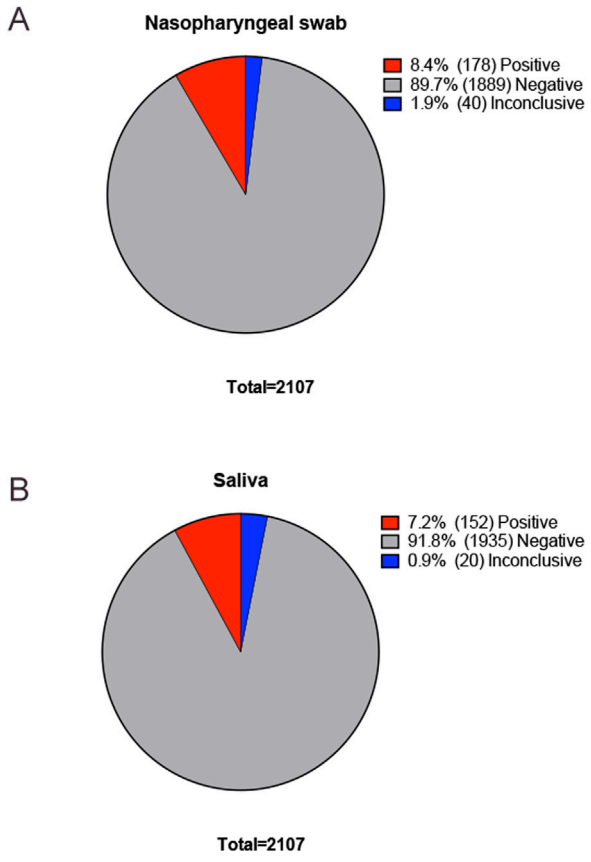


Figure 1. Frequencies and percentages of positive, negative, and inconclusive samples in 2107 paired nasopharyngeal swab and saliva samples: (A) nasopharyngeal swabs; (B) saliva samples.

National Cancer Institute and National Institute of Genomic Medicine. Consecutive asymptomatic subjects were sampled after signing an informed consent form. The study was approved by the ethics and research committees of both institutes (CEI/1479/20 and CEI 2020/21). Paired saliva and NPS samples were collected from 2107 asymptomatic healthcare and office workers to compare the two sample sources for SARS-CoV-2 detection. Additionally, saliva samples were collected from 3983 asymptomatic office workers, 2126 asymptomatic healthcare personnel, and 846 symptomatic office workers to detect SARS-CoV-2.

Sample collection

NPS were collected by a trained clinician with a flexible nylon swab that was inserted through the patient's nostrils to reach the posterior nasopharynx. It was left in place for several seconds and slowly removed while rotating. The swab was then placed in 3 mL of sterile viral transport medium. Swabs from both nostrils were deposited in a single viral transport tube. Saliva samples were self-collected by the individuals without any stimulation and without rinsing the mouth before sample collection. Five milliliters of saliva was collected in a 50-ml sterile conical centrifuge tube without preservatives. Sample collection was done within the same facilities where the viral diagnosis laboratory is located. They were also collected from nearby hospitals. As a result, the swabs and the saliva samples were processed for viral RNA extraction within 5 h after collection.

SARS-CoV-2 RNA extraction and detection

Total nucleic acid was extracted from 300 μL of viral transport medium from the NPS or 300 μL of whole saliva using the MagMAX Viral/Pathogen Nucleic Acid Isolation Kit (Thermo Fisher Scientific) and eluted into 75 μL of elution buffer. For SARS-CoV-2 RNA detection, 5 μL of RNA template was tested using the US CDC real-time RT-qPCR primer/probe sets for 2019-nCoV_N1 and 2019-nCoV_N2 and human RNase P (RP) as an extraction control. Samples were classified as positive for SARS-CoV-2 when both the N1 and N2 primer/probe sets were detected with a Ct value of less than 40 (Centers for Disease Control and Prevention, 2020). If only one of these genes was detected, the sample was labeled inconclusive. All tests were run on Thermo Fisher ABI QuantStudio 5 or QuantStudio 7 real-time thermal cyclers.

Validation of saliva performance in independent laboratories and different extraction and detection methods

For validation purposes, 60 samples that were analyzed in our laboratory were also processed in two independent authorized laboratories (30 samples in each laboratory: Instituto de Investigaciones Biomédicas and Facultad de Química, Universidad Nacional Autónoma de México) using two additional RNA extraction methods and detection systems. The additional extraction methods consisted of spin-column-based RNA extraction (Total RNA Purification Kit, Jena Biosciences) and the use of a quick extraction solution from Lucigen. The two additional methods for SARS-CoV-2 detection were conducted using the GoTaq Probe 1-

Table 1
Detection of SARS-CoV-2 by RT-qPCR between nasopharyngeal swab and saliva samples: (A) positive, negative, and inconclusive samples; (B) only positive and negative samples.

A		Nasopharyngeal swab			
		Positive	Negative	Inconclusive	Total
Saliva	Positive	139 (6.6%)	10 (0.5%)	3 (0.1%)	152 (7.2%)
	Negative	34 (1.6%)	1867 (88.6%)	34 (1.6%)	1935 (91.8%)
	Inconclusive	5 (0.2%)	12 (0.6%)	3 (0.1%)	20 (0.9%)
	Total	178 (8.4%)	1889 (89.7%)	40 (1.9%)	2107 (100%)

B		Nasopharyngeal swab		
		Positive	Negative	Total
Saliva	Positive	139 (6.8%)	10 (0.5%)	149 (7.3%)
	Negative	34 (1.7%)	1867 (91.1%)	1901 (92.7%)
	Total	173 (8.4%)	1877 (91.6%)	2050 (100%)

Prevalence positive test = 8.44% (95% CI 7.27–9.73%), sensitivity = 80.35% (95% CI 73.63–85.99%), specificity = 99.47% (95% CI 99.02–99.74%), positive predictive value = 93.29% (95% CI 88.18–96.28%), negative predictive value = 98.21% (95% CI 97.60–98.67%), positive likelihood ratio = 150.81 (95% CI 80.92–281.06), negative likelihood ratio = 0.20 (95% CI 0.15–0.27).

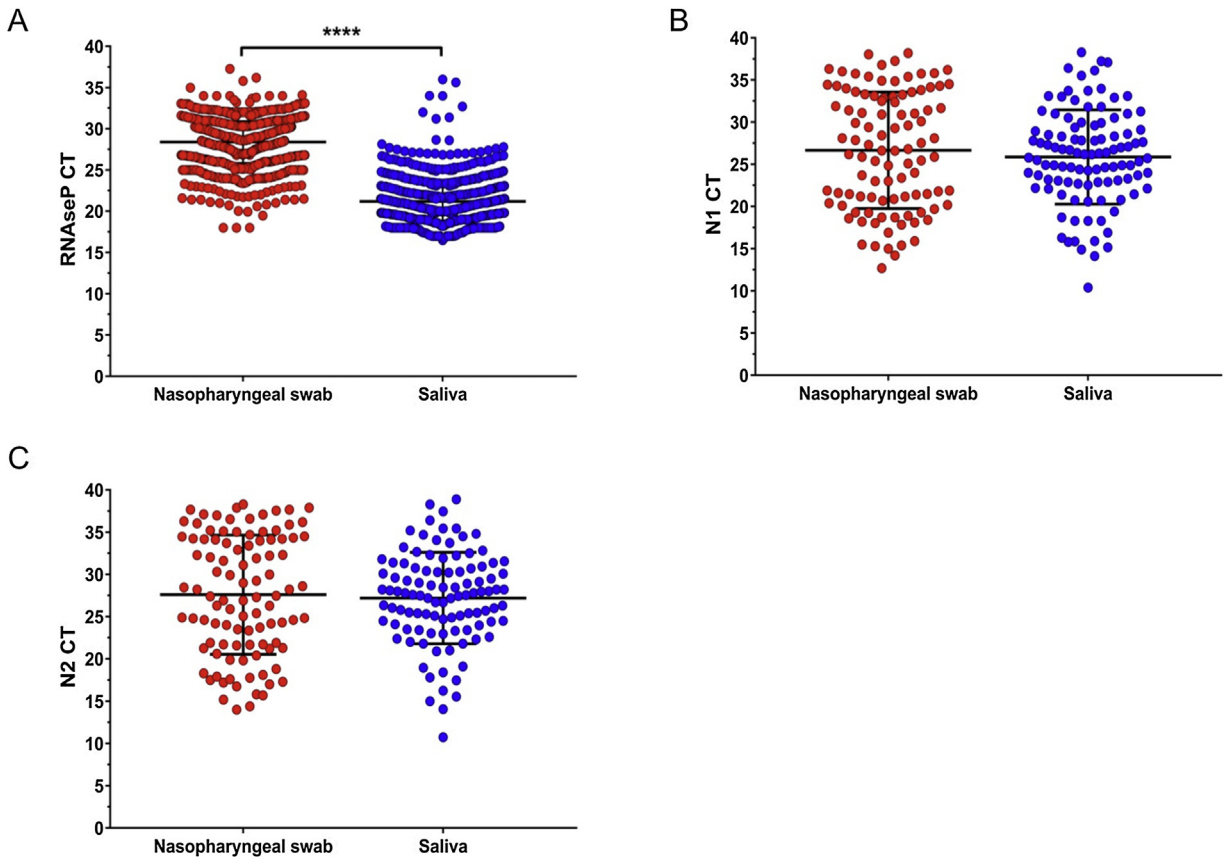


Figure 2. Cycle threshold values (Ct) in nasopharyngeal swab versus saliva. (A) RNAse P gene in all samples; (B) N1 in SARS-CoV-2-positive samples; (C) N2 in SARS-CoV-2-positive samples. RNAse P had a significantly higher concentration of total RNA in saliva compared to nasopharyngeal swab; $p < 0.00001$ (*t*-test).

Step RT-qPCR System from Promega on a 7500 ABI system and the Star Q One-step RT-qPCR from Genes2Life.

Viral copy number analysis

Copies of the SARS-CoV-2 virus were quantified using a standard curve with serial dilutions (10-fold) using the 2019-nCoV_N and Hs_RPP30 positive controls synthesized by Integrated DNA Technologies (IDT, Coralville, IA, USA), with the same detection protocol as the clinical samples. The Ct values obtained from each dilution were used to interpolate the Ct value of each gene from the samples and to calculate viral copy numbers.

Stability assay

The stability of viral RNA in saliva for the detection of SARS-CoV-2 over time after sampling was assessed. A second RNA and an RT-qPCR extraction were performed from 150 SARS-CoV-2-positive saliva samples (stored at 4 °C) at 5, 10, and 15 days after the first positive result.

Sample pooling

SARS-CoV-2 detection in the pooling strategy was performed using the DAAN-Gene Kit following the manufacturer's instructions. Briefly, the kit detects the ORF1ab and N genes of the virus. Five microliters of total RNA were used in the RT-qPCR reaction, and Ct values less than 40 were considered positive.

The evaluation of sample pools was conducted for both viral transport medium with NPS and saliva samples collected in parallel. To test the sensitivity of the pooling strategy, several pools were prepared from saliva and NPS. In the first pooling approach, one positive saliva sample (Ct values of 19.6 and 28.0 for the N viral gene) and its paired positive NPS (Ct values of 18.0 and 36.3 for the N viral gene) were mixed with five and nine known negative samples, respectively. Five hundred microliters of saliva was used from each sample to obtain the pool. This volume was necessary to obtain a homogeneous mixture in the saliva pool, given the differences in viscosity between different samples.

Based on the results obtained from the 10 sample pools and with the premise that asymptomatic individuals might have lower Ct values, which might result in false-negatives in the 10-sample pool, we generated five-sample pools from NPS and saliva from asymptomatic non health-related workers. For NPS, 51 pools made out of 255 individuals were evaluated. For saliva, 26 pools made out of 130 individuals were evaluated.

Statistical analysis

The accuracy of SARS-CoV-2 saliva detection, including sensitivity, sensitivity, predictive values, and likelihood ratios, was determined using RT-qPCR in NPS as the 'gold standard'. Other statistical analyses were performed using GraphPad Prism 7.0 and IBM SPSS Statistics version 24 software. One-tailed parametric (Student *t*-test) and non-parametric (Mann–Whitney *U*-test) statistical tests were used to determine the significance of the data, considering a statistically significant value of $p \leq 0.05$. The kappa coefficient was used to estimate the concordance between saliva and NPS results (McHugh, 2012).

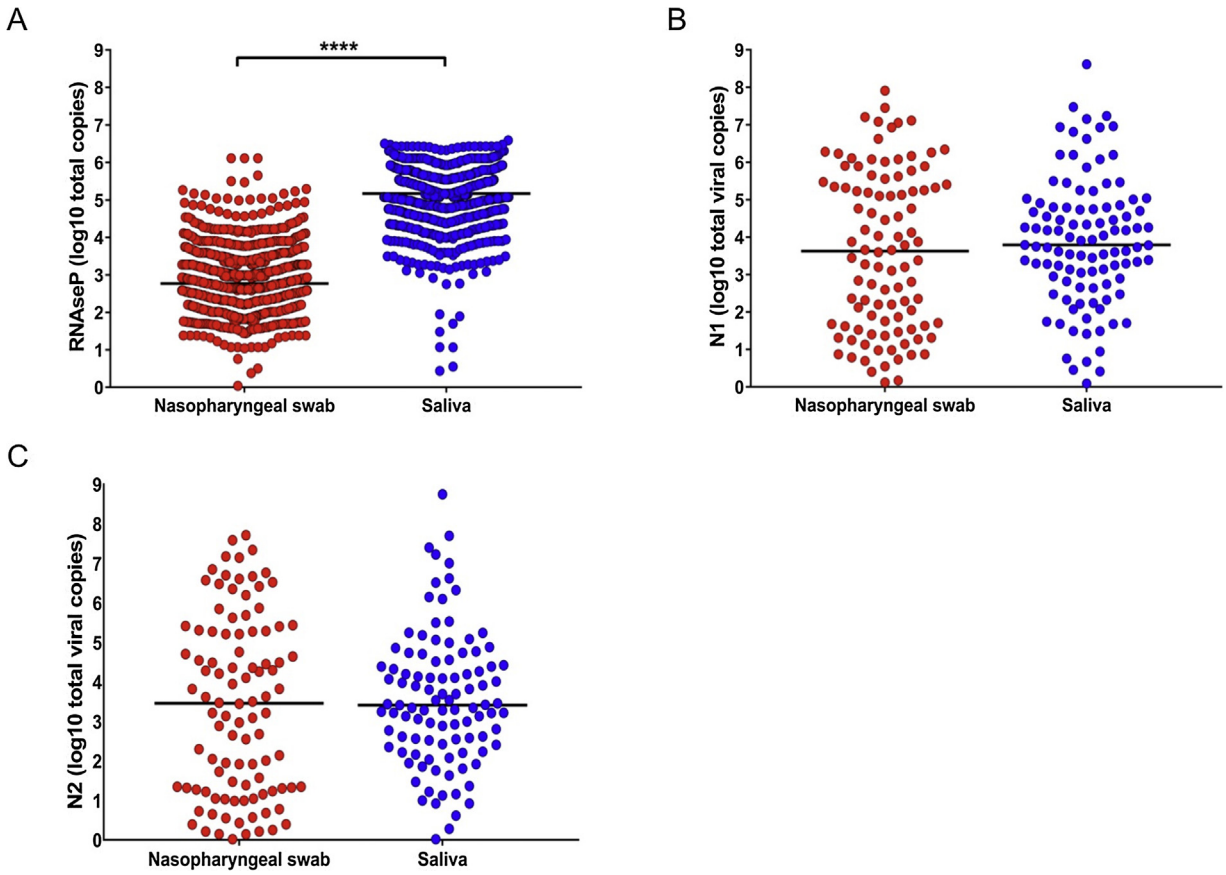


Figure 3. Total viral copies in nasopharyngeal swab versus saliva. (A) RNAse P gene in all samples; (B) N1 in SARS-CoV-2-positive samples; (C) N2 in SARS-CoV-2-positive samples. RNAse P had significantly higher copies in saliva than nasopharyngeal swab; $p < 0.00001$ (Mann–Whitney *U*-test).

Results and discussion

The design of this study was intended to compare the reproducibility, accuracy, and feasibility of saliva sampling followed by RT-qPCR to identify SARS-CoV-2 and to evaluate the use of saliva in sample pooling strategies. It was a priori accepted that the use of NPS followed by RT-qPCR is the gold standard for identification of the virus, despite current studies showing marked variation in the accuracy of this test.

A total of 2107 paired NPS and saliva samples were included in the analysis. The distribution of the results is described in Figure 1. Concordance between saliva and NPS results was statistically significant (Cohen's kappa 0.727, standard error 0.025; $p = 0.0001$; Table 1A). Concordance improved when inconclusive samples were removed from the analysis (Cohen's kappa 0.852, standard error = 0.022; $p = 0.0001$). Overall, 2006 out of 2050 tests (98%) showed the same results in both saliva and NPS (Table 1B). Saliva had a lower number of inconclusive results than NPS (0.9% vs 1.9%) (Table 1 and Fig. 1).

The concordance between the RT-qPCR results from viral RNA obtained from saliva and NPS was statistically significant, indicating that saliva is at least as sensitive as NPS for SARS-CoV-2 detection. Comparison of both the Ct values and the transcript copies of RNAse P showed a significantly higher total RNA concentration in saliva than in NPS (Figures 2A and 3 A). However, when the two viral genes in the positive samples were analyzed, saliva and NPS did not show significant differences in viral load (Figure 2B, C and Figure 3B, C). Spearman correlation analysis of viral copies confirmed that saliva and NPS are both reliable sources for SARS-CoV-2 detection (N1: $r = 0.4217$, $p = 0.0001$; N2: $r = 0.4261$, $p = 0.0001$).

Saliva and paired NPS, which were previously analyzed in our laboratory (60 paired samples), were sent to two independent laboratories for extraction and SARS-CoV-2 detection and processed using different extraction and detection kits. Each laboratory processed 30 paired samples. A 100% concordance was observed in the results between our laboratory and the Instituto de Investigaciones Biomédicas (27 negative and three positive both in saliva and NPS), while 96.7% of the samples sent to Facultad de Química had the same result as in our laboratory (28 negative, one positive, and one discordant). This independent validation is an initial and exploratory assessment.

The accuracy of the saliva test is useful for clinical purposes. The positive likelihood ratio strongly supports its use as a reliable clinical test. A statistically significant correlation and concordance of the RT-qPCR detection of the virus in the saliva samples compared to NPS was identified, and a high concordance between the two types of samples was observed (Table 2A). Given the high number of paired samples analyzed, the results clearly indicate that saliva is as good as NPS for viral detection in the diagnosis of COVID-19, as it has been shown in other studies in hospitalized patients (Table 2B) The data also demonstrated that saliva is stable even without the use of any preservative during sample collection and that a positive SARS-CoV-2 sample can be detected 5, 10, and 15 days after collection when the sample is stored at 4 °C: variation in Ct values in the viral N gene was 0.88 ± 1.92 at 5 days, -0.93 ± 3.01 at 10 days, and -0.76 ± 2.12 at 15 days. Other studies have also demonstrated the stability of saliva for the detection of SARS-CoV-2, with storage for 10–25 days at room temperature (Uwamino et al., 2021) without buffers or stabilizers (Ott et al., 2020).

Table 2
Detection of SARS-CoV-2 in samples of saliva and nasopharyngeal swabs: (A) paired samples; (B) saliva samples only; (C) saliva only in the present study.

A	Paired samples				
Country	Study population	Paired samples	Viral genes	Concordance %	Reference
Australia	Ambulatory patients	522	ORF1a, ORF8	84.6	(Williams et al., 2020)
Canada	Hospitalized patients	91	RdRp, E, N	61.0	(Jamal et al., 2020)
China	Ambulatory patients	229	E	76.0	(Cheuk et al., 2020)
China	Hospitalized patients	58	RdRp/Hel, E, N2	84.5	(Chen et al., 2020)
China	Patients from 12 independent cohorts	944	S, E, ORF1ab, N, RdRp, 5′UTR	92.1	(Zhu et al., 2020)
China	Hospitalized patients	95	E, RdRp	78.9	(Leung Chi-man et al., 2021)
Japan	Ambulatory patients	76	N	97.4	(Iwasaki et al., 2020)
Mexico	Ambulatory patients	253	E	78.6	(Moreno-Contreras et al., 2020)
Thailand	Hospitalized patients	200	ORF1ab, N	97.5	(Pasomsub et al., 2020a)
USA	Hospitalized patients and asymptomatic healthcare workers	29	N1, N2	79.0	(Wyllie et al., 2020)
USA	Ambulatory patients	91	N1, N2	94.0	(Miller et al., 2020)
Mexico	Asymptomatic healthcare and office workers	2107	N1, N2	97.9	Our study

B	Saliva samples only				
Country	Study population	Saliva samples	Viral genes	Positivity %	Reference
China	Hospitalized patients	12	S	91.7	(To et al., 2020a)
China	Hospitalized patients	18	E	84.0	(Hung et al., 2020)
Italy	Hospitalized patients	25	5′UTR	100.0	(Azzi et al., 2020)
Japan	Hospitalized patients	103	N1, N2, ORF1, E	93.4	(Ikeda et al., 2020)

C	Saliva only in the present study					
Setting	Total samples	Number of tests	Positive samples	Positivity (%)	Reduction in testing costs (%)[a]	Reduction in total sample collection direct costs (USD)[b,c]
Asymptomatic office workers	3983	1032	26	0.65	74	$10 754.50
Asymptomatic healthcare personnel	2126	870	98	4.6	59	$5740.20
Symptomatic office workers	846	846	67	7.9	0	$2284.20

RdRp, RNA-dependent RNA polymerase; RdRp/Hel, RNA -dependent RNA polymerase/helicase; ORF1, open reading frame 1 (a,b); ORF8, open reading frame 8; E, envelope; N, nucleocapsid.
[a] Cost reduction was calculated considering the number of tests necessary to identify the positive individuals in the positive pools.
[b] Sample collection direct cost: 3 USD vs 0.3 USD, nasopharyngeal and saliva, respectively.
[c] The sample cost includes both direct and indirect costs.

Sample pooling

Positive samples were selected according to their RT-qPCR results, representing low and high Ct values, to evaluate the effect of a 1:10 pooling with negative samples in the detection capacity of the test. In the first set of saliva samples, the initial Ct values for the positive sample were 22.3 for ORF1ab, 19.6 for N, and 19.8 for RNAse P. As expected, after pooling with the additional nine negative samples, the Ct values increased to 23.8 for ORF1ab, 22.4 for N, and 21.6 for RNAse P, showing that pooling did not affect the detection capacity of the test. A similar situation was observed in the NPS sample pool. In the second saliva pool, the positive sample had higher Ct values (31.9 ORF1ab, 28 for N, and 19.1 for RNAse P). After pooling, an increase in four Ct values in both viral genes was observed. Even though this result is still within the acceptable range for detecting the positive sample in the pool (Figure 4), samples with a higher Ct value might become false-negatives if analyzed by pooling; for this reason, the subsequent experiments were focused on the analysis of five-sample pools.

A total of 130 individuals were tested in 26 NPS pools with five samples each, identifying 20 positive cases (15.4%). All positive cases identified in the pools were confirmed through the analysis of the individual samples used to generate the pool. In the case of saliva, 255 individuals were grouped into 26 pools with five samples each. In this case, two positive cases were identified (7.7%), which were also confirmed through analysis of the individual samples.

Additionally, asymptomatic office and healthcare personnel were tested, as well as office workers presenting mild symptoms who were suspected of being SARS-CoV-2 carriers. Only saliva was used and five samples were pooled in the first two groups. Table 2C shows the positivity among the three groups, which was increased in healthcare personnel and symptomatic office workers. Substantial reductions in direct costs for sampling compared with NPS and in the costs by testing pools instead of individual samples were observed.

These results showed that it is feasible to apply pooling strategies using saliva. However, some considerations should be taken into account, including the use of 500 μL of saliva to generate the pool to obtain a homogeneous mixture. Dilution of one positive sample with nine negative samples showed that, even though positive results can still be obtained in the pool, samples with a low viral load might become difficult to detect. Therefore, we suggest pooling no more than five samples, even though other reports indicate that pooling strategies of 16 and 24 samples are useful in high prevalence populations (≥10%) (Verwilt et al., 2020).

Concluding remarks

The study data indicate that saliva is a reliable source for the detection of SARS-CoV-2 infection. However, several aspects must be addressed to successfully use saliva testing: (1) Sample collection: even though saliva self-collection might be easier than NPS sampling, proper biosafety and risk evaluation protocols must be followed by medical personnel to minimize contagions due to the production of potential aerosols during saliva collection. (2) Sample handling: the application of proven and standardized methods for the inactivation and handling of a saliva sample

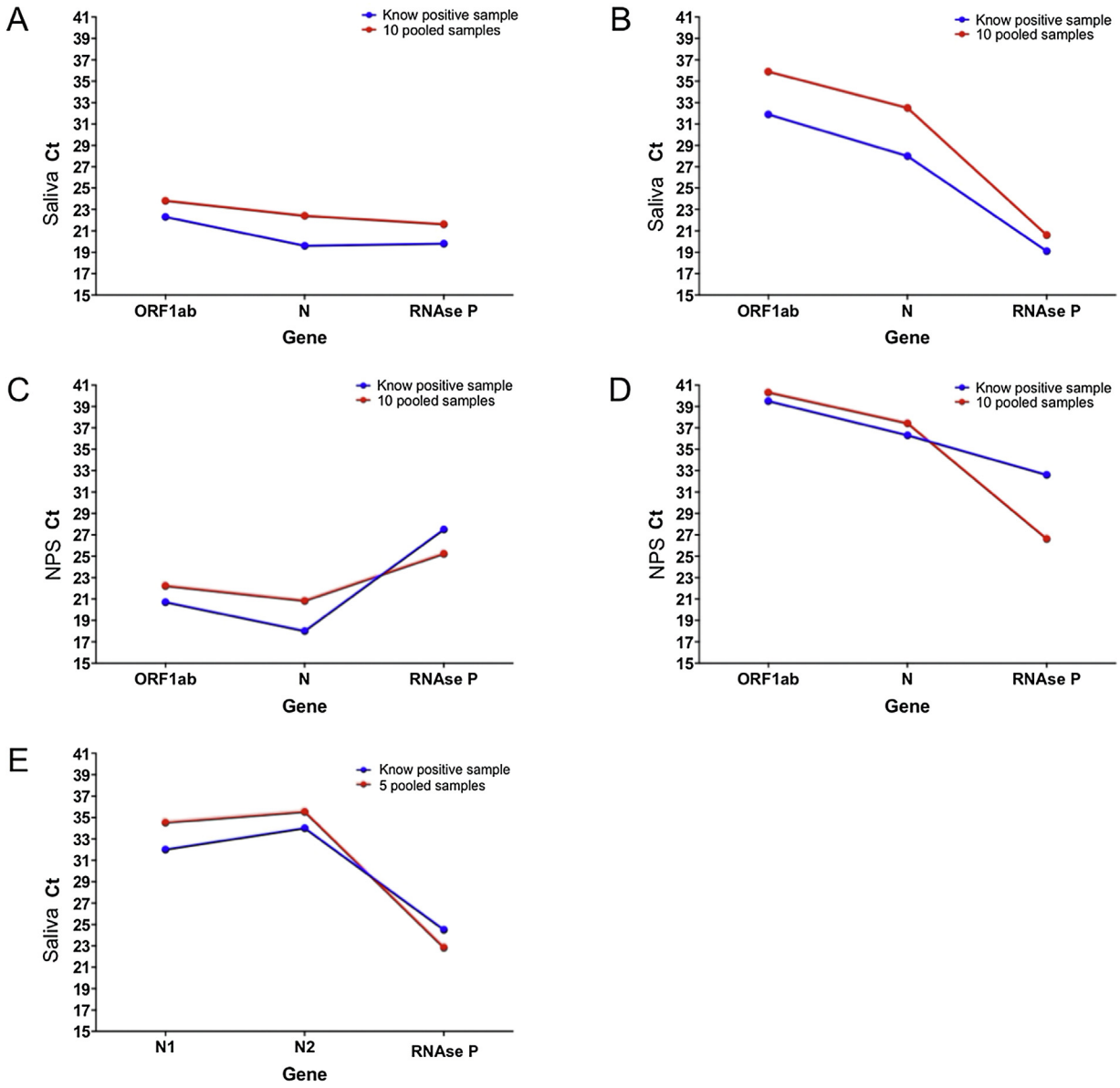


Figure 4. Analysis of pooled (1:5 and 1:10) saliva and swab paired samples. Saliva and nasopharyngeal swab pools were generated by mixing one positive sample with four/nine known negative samples. Positive samples with early and late Ct of the N gene were selected to evaluate the impact of dilution on its detection. (A) and (B) show saliva pooled 1:10; (C) and (D) show nasopharyngeal swab pooled 1:10; (E) shows saliva pooled 1:5.

should be considered, and saliva samples must always be regarded as potentially infected. The packaging and cold-chain protocols used for NPS samples must be followed. (3) RNA extraction and RT-qPCR: it has been well documented that several components of saliva can inhibit PCR, highlighting the importance of using viral RNA extraction systems that have been tested and approved by regulatory agencies that generate pure and high-quality RNA for RT-qPCR analysis. We did not use any preservative for saliva samples and suggest that samples should be stored at 4 °C after collection and processed within 4 days post collection.

Given the situations mentioned above, the use of saliva represents a viable option for SARS-CoV-2 detection. Thus, saliva and the pooling strategy presented here are effective options for the analysis of samples in well-controlled cohorts, which provide a cost-effective screening tool in asymptomatic populations. The cost reduction was calculated considering the number of tests necessary to identify the positive individuals in the positive pool. This is particularly suitable, for example, in office workers, faculty, or other groups where testing is necessary on a periodic basis to identify and isolate infected individuals. The implementation of testing for SARS-CoV-2 infection by RT-qPCR using saliva as a source for viral RNA constitutes an easy, non-invasive, inexpensive, and less risky option compared to NPS, without compromising the accuracy of the test. The combination of saliva sampling and pooling represents a viable and useful method for population-based studies that will be necessary for a safe return to economic activities.

Funding

This work was funded by the Secretaría de Educación, Ciencia, Tecnología e Innovación de la Ciudad de México (SECTEI).

Conflict of interest

The authors do not have an association that might pose a conflict of interest.

Acknowledgments

The authors acknowledge M.C. Isabel Gracia Mora for providing BSL2 facilities at UNAM; Facultad de Química-UNAM, Patronato de la Facultad de Química-UNAM for providing all materials used in the RT-qPCR analyses; and CONACyT 314298 for project funding.

References

Abdalhamid B, Bilder CR, McCutchen EL, Hinrichs SH, Koepsell SA, Iwen PC. Assessment of specimen pooling to conserve SARS CoV-2 testing resources. Am J Clin Pathol 2020;153:, doi:http://dx.doi.org/10.1093/ajcp/aqaa064.

Azzi L, Carcano G, Gianfagna F, Grossi P, Gasperina DD, Genoni A, et al. Saliva is a reliable tool to detect SARS-CoV-2. J Infect 2020;81:, doi:http://dx.doi.org/10.1016/j.jinf.2020.04.005.

Ben-Ami R, Klochendler A, Seidel M, Sido T, Gurel-Gurevich O, Yassour M, et al. Large-scale implementation of pooled RNA extraction and RT-PCR for SARS-CoV-2 detection. Clin Microbiol Infect 2020;26:, doi:http://dx.doi.org/10.1016/j.cmi.2020.06.009.

Callahan CJ, Lee R, Zulauf KE, Tamburello L, Smith KP, Previtera J, et al. Open development and clinical validation of multiple 3D-printed sample-collection swabs: rapid resolution of a critical COVID-19 testing bottleneck. MedRxiv 2020;, doi:http://dx.doi.org/10.1101/2020.04.14.20065094.

Centers for Disease Control and Prevention. (2019-nCoV), CDC 2019-Novel Coronavirus Panel, Real-Time RT-PCR Diagnostic. 2020. https://www.fda.gov/media/134922/download.

Chen JHK, Yip CCY, Poon RWS, Chan KH, Cheng VCC, Hung IFN, et al. Evaluating the use of posterior oropharyngeal saliva in a point-of-care assay for the detection of SARS-CoV-2. Emerg Microbes Infect 2020;9:, doi:http://dx.doi.org/10.1080/22221751.2020.1775133.

Cheuk S, Wong Y, Tse H, Siu HK, Kwong TS, Chu MY, et al. Posterior oropharyngeal saliva for the detection of SARS-CoV-2. Clin Infect Dis 2020;71:2939–46, doi:http://dx.doi.org/10.1093/cid/ciaa797.

Hogan CA, Garamani N, Sahoo MK, Huang CH, Zehnder J, Pinsky BA. Retrospective screening for SARS-CoV-2 RNA in California, USA, late 2019. Emerg Infect Dis 2020;26:, doi:http://dx.doi.org/10.3201/eid2610.202296.

Hung DL-L, Li X, Chiu KH-Y, Yip CC-Y, To KK-W, Chan JF-W, et al. Early-morning vs spot posterior oropharyngeal saliva for diagnosis of SARS-CoV-2 infection: implication of timing of specimen collection for community-wide screening. Open Forum Infect Dis 2020;7:, doi:http://dx.doi.org/10.1093/ofid/ofaa210.

Ikeda M, Imai K, Tabata S, Miyoshi K, Murahara N, Mizuno T, et al. Clinical evaluation of self-collected saliva by RT-qPCR, direct RT-qPCR, RT-LAMP, and a rapid antigen test to diagnose COVID-19. MedRxiv 2020;58:e01438-20, doi:http://dx.doi.org/10.1101/2020.06.06.20124123.

Iwasaki S, Fujisawa S, Nakakubo S, Kamada K, Yamashita Y, Fukumoto T, et al. Comparison of SARS-CoV-2 detection in nasopharyngeal swab and saliva. J Infect 2020;81:, doi:http://dx.doi.org/10.1016/j.jinf.2020.05.071.

Jamal AJ, Mozafarihashjin M, Coomes E, Powis J, Liu AX, Paterson A, et al. Sensitivity of nasopharyngeal swabs and saliva for the detection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). MedRxiv 2020;, doi:http://dx.doi.org/10.1101/2020.05.01.20081026.

Leung Chi-man E, Chow Chi-ying V, Lee Kin-ping M, Lai Wai-man R. Deep throat saliva as an alternative diagnostic specimen type for the detection of SARS-CoV-2. J Med Virol 2021;93:, doi:http://dx.doi.org/10.1002/jmv.26258.

Lohse S, Pfuhl T, Berkó-Göttel B, Rissland J, Geißler T, Gärtner B, et al. Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. Lancet Infect Dis 2020;20:, doi:http://dx.doi.org/10.1016/S1473-3099(20)30362-5.

McHugh ML. Interrater reliability: The kappa statistic. Biochem Medica 2012;22:, doi:http://dx.doi.org/10.11613/bm.2012.031.

Miller M, Jansen M, Bisignano A, Mahoney S, Wechsberg C, Albanese N, et al. Validation of a Self-administrable, Saliva-based RT-qPCR Test Detecting SARS-CoV-2. MedRxiv 2020;, doi:http://dx.doi.org/10.1101/2020.06.05.20122721.

Moreno-Contreras J, Espinoza MA, Sandoval-Jaime C, Cantú-Cuevas MA, Barón-Olivares H, Ortiz-Orozco OD, et al. Saliva sampling and its direct lysis, an excellent option to increase the number of SARS-CoV-2 diagnostic tests in settings with supply shortages. J Clin Microbiol 2020;58:, doi:http://dx.doi.org/10.1128/JCM.01659-20.

Noah K, Fred T, Vlad S, Agatha B, Laura D, Siri K, et al. Self-collected oral fluid and nasal swabs demonstrate comparable sensitivity to clinician collected nasopharyngeal swabs for Covid-19 detection. MedRxiv 2020;, doi:http://dx.doi.org/10.1101/2020.04.11.20062372.

Ott IM, Strine MS, Watkins AE, Boot M, Kalinich CC, Harden CA, et al. Simply saliva: Stability of SARS-CoV-2 detection negates the need for expensive collection devices. MedRxiv 2020;, doi:http://dx.doi.org/10.1101/2020.08.03.20165233.

Pasomsub E, Watcharananan SP, Boonyawat K, Janchompoo P, Wongtabtim G, Suksuwan W, et al. Saliva sample as a non-invasive specimen for the diagnosis of coronavirus disease 2019: a cross-sectional study. Clin Microbiol Infect 2020a;, doi:http://dx.doi.org/10.1016/j.cmi.2020.05.001.

Pasomsub Ekawat, Watcharananan SP, Watthanachockchai T, Rakmanee K, Tassaneetrithep B, Kiertiburanakul S, et al. Saliva sample pooling for the detection of SARS-CoV-2. J Med Virol 2020b;, doi:http://dx.doi.org/10.1002/jmv.26460.

SalivaClear by Mirimus Clinical. The SalivaClear Solution. 2020. https://www.salivaclear.com/.

To KKW, Tsang OTY, Leung WS, Tam AR, Wu TC, Lung DC, et al. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. Lancet Infect Dis 2020a;20:, doi:http://dx.doi.org/10.1016/S1473-3099(20)30196-1.

To KKW, Tsang OTY, Yip CCY, Chan KH, Wu TC, Chan JMC, et al. Consistent detection of 2019 novel coronavirus in saliva. Clin Infect Dis 2020b;71:, doi:http://dx.doi.org/10.1093/cid/ciaa149.

U.S. Food and Drug Administration. Coronavirus (COVID-19) update: FDA authorizes first diagnostic test using at-home collection of saliva specimens. 2020. https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-first-diagnostic-test-using-home-collection-saliva.

U.S. Food and Drug Administration. Coronavirus (COVID-19) update: FDA authorizes first diagnostic test for screening of people without known or suspected COVID-19 infection authorization is also second to allow testing of pooled samples. 2020.

Uwamino Y, Nagata M, Aoki W, Fujimori Y, Nakagawa T, Yokota H, et al. Accuracy and stability of saliva as a sample for reverse transcription PCR detection of SARS-CoV-2. J Clin Pathol 2021;74:, doi:http://dx.doi.org/10.1136/jclinpath-2020-206972.

Verwilt J, Mestdagh P, Vandesompele J. Evaluation of efficiency and sensitivity of 1D and 2D sample pooling strategies for diagnostic screening purposes. MedRxiv 2020;, doi:http://dx.doi.org/10.1101/2020.07.17.20152702.

Watkins AE, Fenichel EP, Weinberger DM, Vogels CBF, Brackney DE, Casanovas-Massana A, et al. Pooling saliva to increase SARS-CoV-2 testing capacity. MedRxiv 2020;, doi:http://dx.doi.org/10.1101/2020.09.02.20183830.

Williams E, Bond K, Zhang B, Putland M, Williamson DA. Saliva as a noninvasive specimen for detection of sars-cov-2. J Clin Microbiol 2020;58:, doi:http://dx.doi.org/10.1128/JCM.00776-20.

Wyllie AL, Fournier J, Casanovas-Massana A, Campbell M, Tokuyama M, Vijayakumar P, et al. Saliva is more sensitive for SARS-CoV-2 detection in COVID-19 patients than nasopharyngeal swabs. MedRxiv 2020;, doi:http://dx.doi.org/10.1101/2020.04.16.20067835.

Yelin I, Aharony N, Tamar ES, Argoetti A, Messer E, Berenbaum D, et al. Evaluation of COVID-19 RT-qPCR Test in Multi sample Pools. Clin Infect Dis 2020;71:, doi:http://dx.doi.org/10.1093/cid/ciaa531.

Zhu J, Guo J, Xu Y, Chen X. Viral dynamics of SARS-CoV-2 in saliva from infected patients. J Infect 2020;81:, doi:http://dx.doi.org/10.1016/j.jinf.2020.06.059.

Revista del Centro de Investigación de la Universidad La Salle
Vol. 14, No. 56, Julio-Diciembre, 2021: 37-62
DOI: http://doi.org/10.26457/recein.v14i56.2856

La hologenómica y su relación con el cáncer

Hologenomics and its relationship to cancer

Cristian Arriaga-Canon
Instituto Nacional de Cancerología-Instituto de Investigaciones Biomédicas, UNAM (México)

Laura Contreras-Espinosa
Instituto Nacional de Cancerología-Instituto de Investigaciones Biomédicas, UNAM (México)

Rogelio Montiel-Manríquez
Instituto Nacional de Cancerología-Instituto de Investigaciones Biomédicas, UNAM (México)

Daniela Venegas-Suárez
Instituto Nacional de Medicina Genómica (México)

Luis Alonso Herrera Montalvo*
Instituto Nacional de Cancerología-Instituto de Investigaciones Biomédicas, UNAM y Instituto Nacional de Medicina Genómica (México)

Recibido: 14 de enero de 2021
Aceptado: 05 de mayo de 2021
Publicado: 13 de septiembre de 2021

Resumen

Un holobionte es considerado una entidad biológica asociada en simbiosis con microorganismos que complementan sus vías metabólicas, sus funciones fisiológicas y su variación genética. El término hologenoma se refiere a todo el contenido genético del holobionte, es decir, a la suma del genoma del hospedero, sus organelos junto con los genomas del microbioma que lo componen. En la actualidad se ha establecido que existe una relación entre el desarrollo de enfermedades y el microbioma de los

*Email: herreram@biomedicas.unam.mx

humanos, por lo que en esta revisión se describirá el papel del hologenoma en el cáncer y las técnicas de secuenciación masiva en paralelo aplicadas en la hologenómica. El estudio de la hologenómica junto con las tecnologías de secuenciación y bioinformática proporcionará información relevante para el desarrollo de nuevas herramientas diagnósticas y su posterior aplicación en la práctica clínica.

Palabras clave: Holobionte, Hologenómica, Cáncer, Genómica, Microbioma.

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

Abstract

A holobiont is considered a biological entity associated in symbiosis with microorganisms that complement its metabolic pathways, physiological functions, and genetic variation. The term hologenome refers to the entire genetic content of the holobiont, i.e. the sum of the host's genome, its organelles along with the microbial genomes that comprise it. It has now been established that there is a relationship between disease development and the human microbiome, so this review will describe the role of the hologenome in cancer and the parallel mass sequencing techniques applied in hologenomics. The study of hologenomics with sequencing and bioinformatics technologies will provide relevant information for developing new diagnostic tools and their subsequent application in clinical practice.

Keywords: Holobiont; Hologenomics; Cancer; Genomics; Microbiome.

Introducción

En la actualidad a los organismos vivos como las plantas y los animales no se les puede considerar como individuos o entidades biológicas independientes definidas mediante criterios anatómicos o fisiológicos, debido a que una gran diversidad de organismos simbiontes se encuentra habitando cada una de estas entidades biológicas para completar sus vías metabólicas y así, cumplir funciones fisiológicas. Por otra parte, los organismos simbiontes también constituyen un segundo modo de herencia genética que proporciona y complementa la variación genética del organismo hospedero, el cual finalmente es sometido a los mecanismos de la selección natural (Gilbert et al., 2012).

Entre los microorganismos que podemos encontrar como simbiontes se encuentran las bacterias, las arqueas, los virus y los organismos protistas que, en conjunto con el hospedero, forman el holobionte (Raina et al., 2018). El holobionte se considera, de manera general y conceptual, como un organismo eucarionte, planta o animal, que incluye a todos sus microorganismos simbióticos asociados que forman unidades anatómicas, fisiológicas, inmunológicas y evolutivas en la naturaleza (Simon et al., 2019a). El termino holobionte deriva del griego antiguo *ὅλος* (hólos) que significa «todos» y *βιός,* que significa "organismo", "ser vivo" (*Revista Ciencia y Desarrollo*, 2018). El concepto de holobionte fue introducido por Lynn Margulis en 1991, sin embargo, se ha propuesto que dicho término fue usado desde 1943 por el biólogo teórico alemán Adolf Meyer-Abich (Baedke et al., 2020), lo que sugiere que este concepto ha sido utilizado por la comunidad científica desde hace más de 50 años.

De la misma manera, el concepto de holobionte ha surgido en el marco teórico y experimental para estudiar las interacciones entre hospederos y sus comunidades microbianas, las cuales evolucionan y se desarrollan en todo tipo de ecosistemas. Actualmente, se han establecido varios modelos biológicos de estudio, en donde se han incluido a los humanos, las plantas, las esponjas y los insectos, para determinar el papel del holobionte en la biología, la ecología y la evolución del hospedero (Simon et al., 2019b; van de Guchte et al., 2018). Asimismo, se ha observado que en el proceso evolutivo del holobionte pueden ocurrir cambios en el genoma del hospedero como en cualquiera de los genomas microbianos asociados al mismo, y que depende de la cooperación entre los genomas dentro del holobionte, así como de la competencia con otros holobiontes para que se desarrollen en un nicho ecológico (Theis et al., 2016).

A partir de las evidencias científicas antes mencionadas, Richard Jefferson en 1994 propuso el término de hologenoma, mismo que en 2007 fue propuesto de forma independiente por Eugene Rosenberg y Llana Zilber-Rosenberg (Bordenstein & Theis, 2015). El concepto de hologenoma se refiere al contenido genético completo del hospedero, de sus organelos y de todo su microbioma, el cual puede ser transferido entre ellos y varía entre cada holobionte. En este sentido, el hologenoma es todo el sistema genético del holobionte, que aumenta o disminuye la salud física del hospedero. Además, dentro de las subunidades genómicas contenidas en el hologenoma constantemente surgen las mutaciones a sus propias tasas finitas, tal es el caso del genoma nuclear en el que la selección natural distingue entre variantes genéticas favorables y no favorables, determinando cuales de estas mutaciones servirán para mejorar la aptitud del holobionte. Por ejemplo, en el microbioma la selección favorece la propagación de microbios benéficos involucrados en la nutrición, la defensa y su reproducción (Douglas, 2015), mientras que los microorganismos patógenos son eliminados a través de la selección natural, aunque en

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

ocasiones estos mismos desarrollan adaptaciones para mejorar su transmisión en la próxima generación del hospedero (LePage & Bordenstein, 2013; Ma et al., 2014). En resumen, el hospedero puede adquirir nuevos microorganismos provenientes del ambiente o ampliar la abundancia microbiana por medio de la transferencia genética horizontal entre organismos del microbioma del hospedero, lo cual genera un efecto directo sobre la variabilidad genética del hologenoma.

Finalmente, se ha establecido que existe una estrecha relación entre el desarrollo de enfermedades y el microbioma de los seres humanos. El microbioma de los humanos el cual se compone en su mayoría por bacterias, ha llamado la atención en los últimos años debido a que existen varios reportes científicos que han sugerido que el microbioma puede contribuir al riesgo de desarrollar cáncer (Gopalakrishnan et al., 2018). En esta revisión nos enfocaremos en dar un panorama general sobre los últimos estudios del hologenoma y su relación con el desarrollo del cáncer a través del microbioma. Además, se mencionarán las distintas técnicas de secuenciación de ácidos nucleicos que han sido utilizadas actualmente para tratar de entender al holobionte a través del estudio de la hologenómica para finalmente hacer especial énfasis en los análisis bioinformáticos los cuales han apoyado para entender la hologenómica y su relación con el cáncer para el desarrollo de nuevas herramientas diagnósticas y su aplicación en la práctica clínica.

1. Concepto de microbioma

Durante más de un siglo las enfermedades se han clasificado en infecciosas y no infecciosas, con base en los postulados de Robert Koch, con la finalidad de establecer una relación causal entre un agente infeccioso específico y su enfermedad particular. Este acercamiento permitió avanzar en la identificación y comprensión de distintas enfermedades y sus tratamientos, sin embargo, contribuyó a mantener un panorama incompleto de los factores involucrados en el inicio y desarrollo de una enfermedad (Pitlik & Koren, 2017). Durante esos años, el estudio del microbioma estaba restringido a las técnicas dependientes del cultivo celular, que modificaban la composición original del microbioma y limitaban el análisis únicamente a especies con características particulares que les permitían crecer en cultivo (Shanahan et al., 2020), excluyendo especies difíciles de cultivar que incluían bacterias intracelulares como *Coxiella burnetiior* o *Tropheryma whipplei* las cuales requieren células hospedadoras para poder sobrevivir y dividirse (Bonnet et al., 2020).

Durante la última década, el estudio del microbioma ha adquirido gran importancia y se le define como aquellos microorganismos que viven, habitan e interactúan entre sí en un mismo entorno. En este mismo concepto también se incluyen los genomas y distintos metabolitos que genera cada especie del microbioma, así como el ambiente donde se desarrollan y su interacción con el organismo hospedero. Entre las diversas comunidades de microorganismos que habitan el cuerpo humano, en su mayoría corresponde a una gran diversidad de bacterias, principalmente del filo Firmicutes, Actinobacteria y Bacteroidetes (The Human Microbiome Project Consortium, 2012), así como una gran diversidad de hongos, dentro de los que destacan los géneros Candida, Saccharomyces y Cladosporium (Underhill & Iliev, 2014). Además, se ha visto que la composición y distribución del microbioma es heterogénea en el humano (Schwabe & Jobin, 2013). Por otro lado, el estudio del microbioma se puede abordar desde diferentes perspectivas, y técnicas, entre ellas, se encuentran la secuenciación masiva en paralelo del DNA,

la cual ha permitido estudiar el conjunto de los diferentes genomas que componen al microbioma (Lynch & Pedersen, 2016). El acercamiento, desde un punto de vista genómico, ha permitido realizar un análisis más amplio sobre la composición y distribución del microbioma en humanos, así como realizar el análisis de metadatos sobre el parentesco genético y el éxito reproductivo relativo dentro de las poblaciones en el microbioma, con lo cual se han podido estudiar e identificar diferentes condiciones extrínsecas que influyen sobre el microbioma y, a su vez, el efecto que tienen estos cambios sobre la salud humana. Parte de la gran importancia que ha adquirido el estudio del microbioma ha sido el hecho de que se han encontrado asociaciones entre los cambios en el microbioma y diversas enfermedades, lo cual ha cambiado la visión general que se tenía sobre la relación entre enfermedad, salud y microbioma.

Actualmente podríamos considerar a cualquier enfermedad como la interacción disfuncional entre los diferentes factores ambientales tales como la dieta, el estilo de vida, la composición del microbioma y la genética del hospedero (Apidianakis & Ferrandon, 2014). Esta definición se sustenta en la creciente evidencia que relaciona diversas patologías con las alteraciones en la composición, la distribución y la función del microbioma, dichas alteraciones son conocidas como disbiosis (AlHilli & Bae-Jump, 2020). A la fecha, se han descrito tres tipos: disbiosis por proliferación de patobiontes en donde se incluyen las bacterias de los géneros *Klebsiella y Enterobacter* (Kitamoto, s. f.), disbiosis por pérdida de microorganismos comensales que incluye diferentes especies de *Lactobacillus* (Brusselaers, s. f.) y disbiosis por pérdida de diversidad, que implica la pérdida de especies de *Lactobacillus* provocando disbiosis vaginal (van de Wijgert & Jespers, 2017). Estos tres tipos de disbiosis no son mutuamente excluyentes y puede haber situaciones en las que se presente más de un tipo a la vez (Levy et al., 2017). Asimismo, el estudio de los distintos tipos de disbiosis y su relación con el desarrollo de diversas enfermedades multifactoriales (J. Wang & Jia, 2016), ha sido impulsado en gran parte mediante el análisis de datos de microbioma incluso en enfermedades consideradas como no infecciosas, como las metabólicas, la diabetes y la obesidad (Fan & Pedersen, 2020); así como enfermedades neurológicas que incluyen el autismo y la depresión (Ezra-Nevo et al., 2020; Marx et al., 2020) e incluso el cáncer (Elinav et al., 2019; Vogtmann & Goedert, 2016) enfermedad donde el estudio de los factores que la promueven ha sido tema central en su investigación durante los últimos años y donde existe una estrecha relación de su desarrollo con la composición de microbioma.

2. Cáncer

El cáncer se ha definido como un conjunto de enfermedades heterogéneas caracterizadas por la presencia de alteraciones a nivel genético y epigenético que afectan los mecanismos celulares regulatorios y de homeostasis, tales como el crecimiento, la división y la diferenciación celular, entre otros (Ginsburg et al., 2020). Los factores que originan el cáncer han sido motivo de estudio durante varios años y se sabe que existen diferentes factores de riesgo involucrados en su desarrollo, tales como la edad (Rizzi et al., 2020), las infecciones virales (Golrokh Mofrad et al., 2020; Nelson & Benson, 2017), la exposición a radiación (Akbani et al., 2015), la exposición a contaminantes ambientales (Hwang et al., 2020), y la dieta (AlHilli & Bae-Jump, 2020; Lien & Vander Heiden, 2019), entre otros, siendo esta última un factor que ha adquirido gran importancia en los últimos años (Steck & Murphy, 2020), debido a que se ha observado que los cambios en la dieta pueden originar predisposición al desarrollo de cáncer, o por el contrario, ofrecer

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

efectos protectores contra este padecimiento. Por tal motivo, el estudio sobre la influencia que tiene la dieta en el desarrollo de cáncer ha cobrado gran importancia actualmente.

3. Asociación de la alimentación y el desarrollo de cáncer

Como se mencionó anteriormente, a través de diversos estudios se ha comprobado la asociación entre la dieta y el desarrollo de cáncer, donde se ha observado que la dieta puede tener efectos benéficos y ayudar en la prevención del desarrollo o progresión de la enfermedad (Bail et al., 2016) así como potenciar el efecto de las terapias contra el cáncer (Kanarek et al., 2020). Sin embargo, también se ha establecido que la dieta puede tener un efecto adverso, convirtiéndose en un factor de riesgo para el desarrollo de esta enfermedad (Lien & Vander Heiden, 2019; Steck & Murphy, 2020). Se ha comprobado, que a nivel molecular, la dieta puede tener un efecto en la producción de especies reactivas de oxígeno (Sun et al., 2020; Tobore, 2020), variaciones en la disponibilidad de nutrientes (Lien & Vander Heiden, 2019), cambios en el metabolismo celular (Bose et al., 2020) y cambios importantes en el microbioma (AlHilli & Bae-Jump, 2020; Bisanz et al., 2019; Polo et al., 2020). El estudio del microbioma y su asociación con el desarrollo del cáncer ha sido de vital importancia ya que se ha comprobado la regulación de la homeostasis de manera importante (Baquero & Nombela, 2012). Por lo tanto, los cambios en la composición y distribución del microbioma pueden estar asociados al desarrollo de enfermedades multifactoriales como el cáncer.

Como se mencionó anteriormente, la disbiosis, es uno de los factores predisponentes a cáncer (Biragyn & Ferrucci, 2018; Sharma et al., 2020); se sabe que al menos un 20% de todos los casos de cáncer a nivel mundial están asociados a esta condición (Pevsner-Fischer et al., 2016). Además de los factores como la dieta, la actividad física, el estilo de vida y la genética del hospedero (Manor et al., 2020; Schmidt et al., 2018; J. Wang & Jia, 2016), la edad es otro de los factores importantes que provocan cambios en el microbioma (Biragyn & Ferrucci, 2018), los cuales tienen un gran impacto sobre el metabolismo, la respuesta inmune, así como en procesos inflamatorios, siendo estas últimas tres parte de las marcas distintivas del cáncer (Hanahan & Weinberg, 2011).

4. Relación del microbioma y el cáncer

Uno de los principales tipos de cáncer asociado al microbioma es el cáncer del tracto gastrointestinal. Por ejemplo, las infecciones bacterianas por la bacteria *Helicobacter pylori*, están fuertemente asociadas al desarrollo de cáncer gástrico (Malfertheiner et al., 2014; Noh et al., 2020). Se ha visto que esta bacteria patógena promueve estrés replicativo en las células de la mucosa gástrica, promoviendo el daño al DNA (Bauer et al., 2020), la inestabilidad genómica, y por consiguiente el desarrollo de cáncer gástrico. (Cassidy & Venkitaraman, 2012; Pikor et al., 2013). Sin embargo, este ejemplo representa una enfermedad asociada a una única bacteria patógena presente en el tracto gastrointestinal, pero existen otros ejemplos como el cáncer colorrectal (CCR) donde se han caracterizado, mediante el análisis del microbioma, cambios en la abundancia de distintas especies bacterianas presentes en tejido tumoral. El CCR presenta una disbiosis característica en la que se ha encontrado una abundancia mayor de microorganismos pro-carcinogénicos como *Fusobacterium nucleatum* y *Escherichia coli*, y una menor cantidad de bacterias consideradas "protectoras" como *Roseburia y Clostridium* (Janney et al., 2020). Como resul-

tado, el microbioma se ha propuesto como un biomarcador para el CCR, siendo mayormente asociados los géneros *Fusobacterium, Porphyromonas* (Shindo et al., 2019), y *Peptostreptococcus* (Ternes et al., 2020). Dentro del género *Fusobacterium*, el efecto de *F. nucleatum* ha sido ampliamente estudiado a nivel molecular y se ha visto asociado a la proliferación celular, una de las marcas distintivas del cáncer, esta bacteria puede promover la proliferación de células cancerosas por distintas vías de señalización intracelular, por ejemplo, se ha comprobado que activa al factor nuclear kappa B (NFkB) induciendo la sobreexpresión del microRNA-21 (Yang et al., 2017) el cual suprime la expresión del gen *RASA1* (RAS P21 *Protein Activator 1*), y en donde la disminución en sus niveles de expresión origina la activación de la vía de MAPK (*RAS-mitogen-activated protein kinase*) induciendo un aumento en la proliferación celular. Además de esta vía, también se ha comprobado que puede promover la proliferación celular a través de la vía de señalización de Wnt, como se verá más adelante a detalle (Rubinstein et al., 2013). Por tal motivo, en este ejemplo de CCR, se puede observar la influencia del microbioma a nivel molecular en el desarrollo de cáncer.

Otro de los ejemplos de cáncer, donde se sabe que la disbiosis promueve el desarrollo de otra de sus marcas distintivas, la inflamación, es el carcinoma oral de células escamosas (COCE). La inflamación crónica es una condición que se ha asociado a cáncer durante varios años (Hanahan & Weinberg, 2011), ya que promueve la liberación de factores de crecimiento que inducen la proliferación celular y también la liberación de enzimas de matriz extracelular que facilitan la invasión, la metástasis y la angiogénesis (Dias et al., 2021). En COCE, se ha observado en saliva que un aumento en el conteo de las bacterias *Porphyromonas gingivalis*, *Fusobacterium nucleatum* y *Prevotella intermedia,* mediante la técnica de *checkerboard DNA-DNA hybridization* (Mager et al., 2005) están asociados a periodontitis que es un tipo de inflamación en las encías (Gholizadeh et al., 2016). En este caso, se ha propuesto al microbioma como un posible biomarcador de diagnóstico ya que en las muestras de saliva se ha observado una gran abundancia de las bacterias *Capnocytophaga gingivalis*, *Prevotella melaninogenica* y *Streptococcus mitis*, por lo que podría ser un indicador diagnóstico de COCE (Mager et al., 2005).

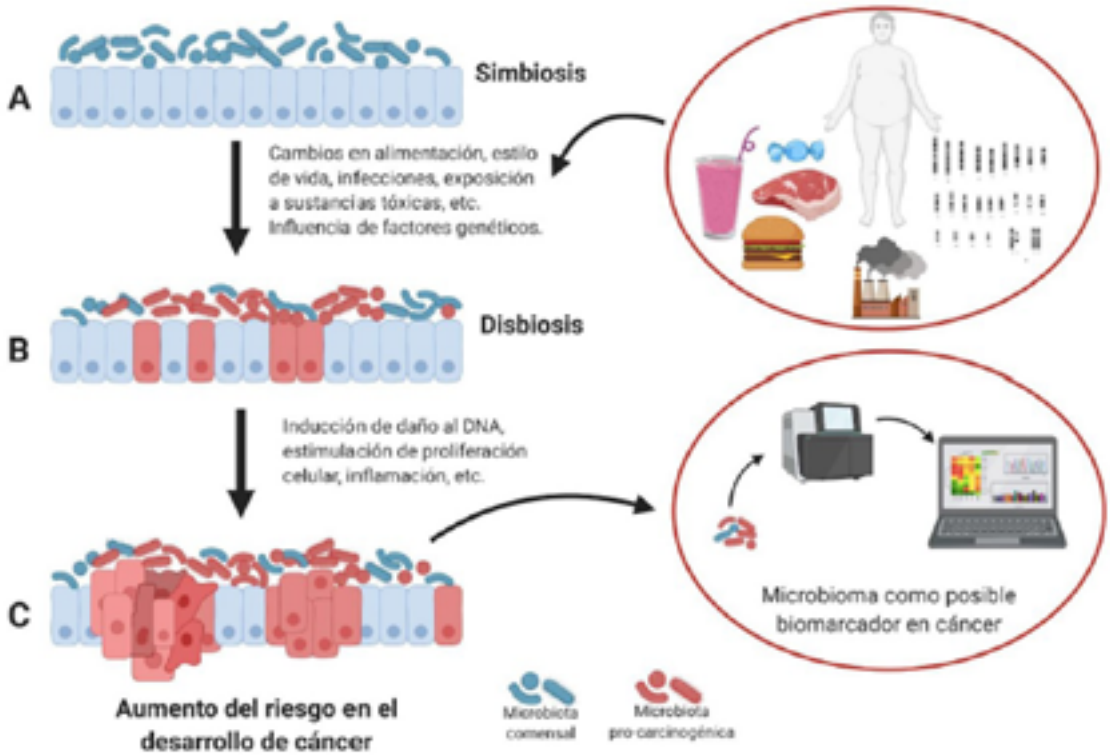
Hasta ahora se ha comprobado la relación entre el microbioma y el desarrollo de algunas neoplasias, en general se sabe que existen diversos factores intrínsecos del hospedero como la edad y la genética, o cambios ambientales como la dieta y el estilo de vida que pueden originar variaciones en la distribución y diversidad del microbioma, y en algunos casos originar disbiosis (Figura 1A), una condición que puede promover inflamación o estrés oxidativo (Figura 1B), estando estas dos condiciones asociadas al desarrollo de cáncer como marcas distintivas (Figura 1C). La identificación de los cambios en la composición del microbioma se puede identificar de diversas maneras, inicialmente se pueden utilizar técnicas como el cultivo en agares sólidos que permiten la identificación y aislamiento de colonias específicas de bacterias u hongos, pero a su vez dificultan la identificación de microorganismos de difícil crecimiento en cultivo como bacterias intracelulares. Además de estas técnicas de cultivo, las técnicas para la identificación de microorganismos basadas en el análisis de ácidos nucleicos como PCR o *checkerboard DNA-DNA hybridization,* permiten identificar microorganismos sin la necesidad de ser cultivados, por ejemplo, bacterias de interés clínico para la detección de COCE como *Capnocytophaga gingivalis*, *Prevotella melaninogenica* y S*treptococcus mitis* (Mager et al., 2005). Adicionalmente, el desarrollo de técnicas de secuenciación masiva de DNA han hecho posible identificar diversas poblaciones de microorganismos

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

presentes en el cuerpo humano y el análisis de su abundancia y distribución, por ejemplo, mediante el análisis de RNA ribosomal 16s (Takahashi et al., 2019)because the mortality rate has recently decreased in other developed countries. The impact of microbiota in carcinogenesis, especially in the digestive tract has been reported. This study aimed to clarify the relationship between oral cancer and oral microbiota in Japanese patients.\nMethods: DNA was extracted from salivary samples of 60 oral cancer patients and 80 non-cancer individuals as controls. We performed metagenomic analysis using 16S rRNA amplicon sequencing. Statistical analysis in this study was performed using R (version 3.5.0 o el análisis de metagenoma (*Shotgun metagenomics*) (S. Yu et al., 2021)three with moderately severe acute pancreatitis (MSAP, técnicas que permiten un acercamiento más amplio para el estudio de las diferentes comunidades de microorganismos presentes en una muestra y no solo un enfoque a una única especie o un género de microorganismo.

Figura 1

Influencia del microbioma en el desarrollo y progresión del cáncer



A) En un estado de simbiosis hay un equilibrio adecuado del microbioma comensal en el cuerpo humano, sin embargo, cambios en factores intrínsecos como influencia de factores genéticos, o extrínsecos como cambios en la dieta, estilo de vida, infecciones o exposición a contaminantes ambientales, entre otros, pueden inducir la pérdida de este equilibrio y favorecer un estado de disbiosis. B) En un estado de disbiosis hay una pérdida del equilibrio en el microbioma, se puede observar una o varias de las siguientes situaciones: proliferación de patobiontes, pérdida de microbioma comensal o pérdida de diversidad. Algunas bacterias patógenas pueden favorecer la aparición de marcas distintivas del cáncer, *Helicobacter pylori* puede inducir daño al DNA o, en el caso de la bacteria *Fusobacterium nucleatum*, un aumento en la proliferación. Otras bacterias como *Porphyromonas*

gingivalis pueden inducir inflamación, estas tres bacterias están asociadas al desarrollo de cáncer gástrico, cáncer colorrectal y carcinoma oral de células escamosas, respectivamente. C) Se ha comprobado que la disbiosis puede inducir el desarrollo de cáncer mediante diferentes vías, por ende se ha propuesto al microbioma como un posible biomarcador para el diagnóstico de diferentes neoplasias entre las que destacan el cáncer colorrectal donde se puede ver una abundancia mayor de bacterias como *F. nucleatum,* también en cáncer gástrico donde se observa la presencia de *H. pylori* o tipos de cáncer como el COCE donde el conjunto de varias especies de bacterias (*Capnocytophaga gingivalis*, *Prevotella melaninogenica* y *Streptococcus mitis*) se pueden utilizar como un biomarcador de diagnóstico.

5. Microbioma y su relación con la regulación de las características distintivas del cáncer (*Hallmarks*)

Se ha propuesto la presencia de 8 rasgos distintivos del cáncer que describen las capacidades que adquieren las células tumorales en su evolución hacia el desarrollo de una neoplasia y que les permite sobrevivir, proliferar y propagarse. Estas características son: la auto señalización de proliferación sostenida, la insensibilidad ante inhibidores de crecimiento, la evasión de la apoptosis, la angiogénesis sostenida, la capacidad de invasión tisular, así como la generación de metástasis, la adquisición de potencial replicativo ilimitado, la reprogramación del metabolismo energético y la evasión de destrucción por el sistema inmune (Pevsner-Fischer et al., 2016). Recientemente, se ha encontrado evidencia que sugiere la participación del microbioma para promover la adquisición de estos rasgos distintivos del cáncer, como veremos a continuación (Schmidt et al., 2018).

Los procesos de proliferación y muerte celular están regulados por diversas señales que deben mantener un balance para conservar la arquitectura y función de los tejidos. En el caso de la mucosa intestinal se ha comprobado que diversas bacterias comúnmente encontradas en el tracto digestivo tienen un efecto en este balance, como *Bacteroides fragilis* y *Fusobacterium nucleatum*, que mediante la interacción con la proteína de adhesión celular E-Caderina, pueden inducir la proliferación de células epiteliales al activar la vía intracelular de señalización Wnt/Beta-Catenina y llevar a cabo la transcripción del proto-oncogen c-Myc (Gholizadeh et al., 2016; Yang et al., 2017). Asimismo, se demostró que las cepas de la bacteria *E. coli* portadoras de la isla pks (*pks+ E. coli*), comúnmente encontradas en tumores colorrectales, inducen a las células senescentes a secretar el factor de crecimiento de hepatocitos (HGF). La presencia de HGF lleva a la proliferación de células epiteliales mediante la expresión del microRNA-20a-5p que tiene como blanco a SENP1, que regula la SUMOilación de la proteína P53, uno de los principales reguladores del ciclo celular y la apoptosis (Mager et al., 2005). Esta relación entre el microbioma y la proteína de P53 evidencia un posible mecanismo para la adquisición de algunas de las características distintivas del cáncer, como son la proliferación celular sostenida y la evasión de señales de apoptosis.

Por consiguiente, al haber una mayor proliferación y menor apoptosis en tejidos neoplásicos hay un incremento general del número de células, por lo que requieren un mayor suministro de nutrientes. Por este motivo, la angiogénesis sostenida es esencial para el mantenimiento, el crecimiento y la propagación de células tumorales, siendo otra de las características distintivas del cáncer que puede ser inducida por el microbioma. La evidencia experimental ha demostrado que durante un proceso infeccioso las bacterias, hongos, virus y protozoarios pueden inducir la angiogénesis ya sea mediante la estimulación directa de moléculas del patógeno sobre el hospedero, como en el caso de los lipopolisacáridos propios de bacterias

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

gramnegativas, que son reconocidos por el sistema inmune innato y estimulan la producción de citocinas proinflamatorias; o de manera indirecta en respuesta a un proceso inflamatorio o daño celular (Ding & Schloss, 2014). Por otra parte, en ratones criados en ausencia de microorganismos se observa un arresto en el desarrollo de redes capilares, que puede ser reactivado y corregido en tan solo 10 días después de la inoculación de microbioma de ratones sanos o cultivos de *Bacteroides thetaiotaomicron* (Limborg et al., 2018), demostrando que la interacción del microbioma con su hospedero no sólo influye, sino que es esencial para la angiogénesis durante el desarrollo del intestino y probablemente en procesos neoplásicos.

También se ha estudiado la interacción del microbioma con el sistema inmune en las mucosas. Se ha demostrado que la presencia de microorganismos, como son *B. fragilis* y varias especies del género *Alistipes* pueden inducir una respuesta inmune con presencia abundante de células T colaboradoras productoras de interleucina-17 (Th17), que se asocian a la activación de vías proinflamatorias como la de Stat3, mayor crecimiento tumoral y metástasis en cáncer colorrectal (Tierney et al., 2019). Asimismo, el microbioma puede interactuar también de manera directa con las células del sistema inmune, como es el caso de *F. nucleatum,* que interactúa mediante proteínas de adhesión Fap2 con los receptores de células T, inhibiendo su capacidad citotóxica y con esto su respuesta antitumoral (Plaza Oñate et al., 2019). Esta evidencia demuestra que el microbioma puede regular al sistema inmune ayudando a las células neoplásicas a evadir su destrucción, que es otra de las características distintivas del cáncer. Sin embargo, el microbioma no siempre tiene un efecto protumoral, los microorganismos específicos como los del género *Bifidobacterium,* pueden estimular las capacidades citotóxicas de las células T, promoviendo su actividad antitumoral (Forslund et al., 2015). De este modo, el microbioma puede regular al sistema inmune hacia una respuesta proinflamatoria o antiinflamatoria dependiendo su composición, lo que sugiere el uso de posibles tratamientos basados en la generación de una respuesta inmune antitumoral mediante la manipulación específica del microbioma.

Otra posible línea de tratamiento consiste en utilizar el microbioma para generar un ambiente antitumoral aprovechando la desregulación del metabolismo energético que es una de las características de las células neoplásicas. A diferencia de la mayoría de las células en el organismo que de manera general basan su metabolismo en la fosforilación oxidativa, las células tumorales exhiben un consumo acelerado de glucosa que oxidan parcialmente a lactato mediante glucólisis aerobia, lo que se conoce como efecto Warburg (Zolfo et al., 2017). Las células epiteliales en el colon, por otra parte, pueden basar su metabolismo hasta en un 70% en oxidación del butirato que se encuentre presente en el tracto digestivo. El butirato es producido por el microbioma a través de la fermentación de fibra dietaria, que se define como los componentes de las plantas que no pueden ser metabolizadas por las enzimas digestivas de los humanos. Los polisacáridos no amiláceos que contiene la fibra dietaria son procesados por el microbioma en el tracto digestivo, produciendo ácidos grasos de cadena corta como el propio butirato, que ha demostrado tener un efecto anti-tumoral (Truong et al., 2015). El efecto antitumoral del butirato no consiste únicamente en no servir a las células tumorales como fuente de carbono, sino que al acumularse funciona como inhibidor de las desacetilasas de histonas y promueve la

expresión de genes proapoptóticos como *FAS* y reguladores del ciclo celular, como *p21* y *p27,* lo que en conjunto limita la proliferación celular y la progresión tumoral(Fan & Pedersen, 2020). De esta manera, se evidencia una estrecha relación entre la dieta y el microbioma en la generación de un microambiente antitumoral basado en el metabolismo energético característico de las células tumorales.

La evidencia experimental aquí revisada demuestra que hay una relación entre la composición del microbioma y la regulación de los mecanismos moleculares de cada una de las características distintivas del cáncer. En el caso de la mucosa intestinal, se ha sugerido que esta relación es establecida por la interacción directa e indirecta de los microorganismos en el tracto digestivo con diversos tipos celulares del hospedero, como son las células epiteliales y las células T del sistema inmune, para producir un efecto protumoral o antitumoral, dependiendo de la composición del microbioma. Asimismo, la necesidad de microorganismos en el tracto digestivo para el correcto desarrollo del intestino es una de las evidencias más sólidas para demostrar que la evolución del hospedero y su microbioma ha ocurrido de forma conjunta y, por lo tanto, el cáncer y otros padecimientos podrán ser comprendidos en su totalidad únicamente en el contexto del holobionte.

6. Flujos de trabajo y técnicas utilizadas en el análisis bioinformático de hologenomas

Como se ha mencionado antes, se ha descubierto que algunas funciones biológicas asociadas al microbioma contribuyen con el desarrollo de enfermedades consideradas como no infecciosas, por lo que ahora debemos entender cómo se han hecho este tipo de acercamientos y asociaciones. Sabemos que el holobionte es una entidad biológica constituida por microorganismos simbiontes y patógenos, cuya identidad biológica así como sus interacciones moleculares pueden ser descritas mediante el análisis integral de las biomoléculas que los componen, lo que se conoce como estudios ómicos, a través de técnicas como la secuenciación masiva en paralelo de DNA y RNA (Limborg et al., 2018).

Dada la complejidad de la muestra analizada, ya que contiene un conjunto de biomoléculas (ya sea DNA, RNA, proteínas o metabolitos) que pertenecen a los organismos asociados en comunidad en el holobionte, se han diseñado diferentes flujos de trabajo bioinformáticos (del inglés *bioinformatic pipelines*) para determinar si existen relaciones biológicas (como intercambios de metabolitos) entre dichos microorganismos y si éstas se asocian a funciones específicas tanto del hospedero como de los demás organismos comprendidos en el holobionte. Actualmente, la evidencia científica señala la existencia de relaciones funcionales (como las funciones metabólicas que aportan nutrientes esenciales) entre el hospedero y los otros organismos a nivel de epigenoma, genoma, transcriptoma, proteoma y metaboloma, que pudieran estar asociadas al desarrollo de enfermedades como el cáncer gastrointestinal, así como otros padecimientos crónico-degenerativos (Apidianakis & Ferrandon, 2014), como la diabetes (Forslund et al., 2015). En consecuencia, se han desarrollado metodologías para el análisis de las asociaciones funcionales descritas para los holobiontes de estos padecimientos, así como de su aplicación clínica en el desarrollo de fármacos y biomarcadores (Limborg et al., 2018).

Las asociaciones funcionales que existen entre el hospedero y los otros organismos que componen el holobionte se determinan calculando coeficientes de correlación que permiten identificar los componen-

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

tes moleculares con los que cada organismo contribuye al hologenoma, así como las vías de señalización en las que se involucran, para lo cual se utilizan genomas de referencia de los diferentes componentes del holobionte. Inicialmente, el análisis del hologenoma incluye el flujo de trabajo de alineamiento, mapeo e identificación de unidades taxonómicas mediante el reconocimiento de regiones conocidas como marcos abiertos de lectura que son únicos para cada componente del holobionte (Tierney et al., 2019). El resultado de este proceso permite determinar la abundancia conjunta de los grupos de genes que deriva en la identificación de las diferentes especies metagenómicas presentes en el holobionte, así como los niveles taxonómicos a los que éstas pertenecen, mediante el uso de herramientas como *MSPminer*, que a partir de perfiles genómicos, lleva a cabo el análisis de co-abundancia de genes y permite identificar los organismos presentes en la muestra, tanto los ya caracterizados como los organismos no descritos antes para esa condición, lo que ha permitido la identificación de biomarcadores para el diagnóstico de enfermedades gastrointestinales así como nuevas especies en el hologenoma que no habían sido descritas (Plaza Oñate et al., 2019).

Asimismo, existen herramientas como *StrainPhlAn*, *MetaMLST* (Zolfo et al., 2017) y *MetaPanPhlAn* (Truong et al., 2015) que permiten hacer el análisis por tipo de organismo presente en el holobionte, aumentando la precisión al establecer relaciones funcionales, como se ha visto en el estudio del microbioma asociado a estómago en población europea que permitió la identificación de un nuevo perfil molecular asociado a *Helycobacter pylori*, lo que es relevante en el estudio de los padecimientos gastrointestinales y el desarrollo de biomarcadores novedosos (Fan & Pedersen, 2020). Además, se ha propuesto introducir el uso de clasificadores supervisados basados en aprendizaje de máquina (del inglés *machine learning*), que consiste en usar los diferentes lenguajes de programación para enseñarle a las computadoras cómo analizar y aprender de los datos para poder realizar predicciones o inferencias. Esto se hace con diferentes estrategias, como el algoritmo de bosques aleatorios que emplea la herramienta *AUC-RF* (Calle et al., 2011) y otros métodos como el descrito por Thompson y colaboradores (Thompson et al., 2019), que mejora la identificación de los organismos del holobionte, permitiendo además el descubrimiento de nuevas características del hologenoma que pueden tener implicaciones clínicas dentro del holobionte, lo que a su vez ha permitido la identificación de nuevos biomarcadores de diagnóstico, como se ha visto en el estudio de los componentes microbiológicos presentes cuando existen lesiones de colon pre-neoplásicas, demostrando los beneficios de la aplicación de estos nuevos enfoques bioinformáticos al estudio del hologenoma (Baxter et al., 2016; Flemer et al., 2018).

En cuanto a la información que utilizan estos flujos de trabajo, ésta proviene principalmente del análisis de datos de secuenciación, para lo cual se utilizan diferentes estrategias dependiendo de las características del holobionte que se pretenda estudiar. Particularmente, en el caso del hologenoma, éste es analizado comúnmente con secuenciación por fuerza bruta (o *shotgun* en inglés) y estudios de asociación de genoma completo (GWAS, por sus siglas en inglés), de cuyo análisis se deriva la determinación de la identidad de los organismos presentes en el holobionte (Fatkhullina et al., 2018). Sin embargo, se ha demostrado que los estudios de mayor utilidad corresponden a aquellos que usan los meta-GWAS (MGWAS), ya que aportan información relevante acerca de los cambios fenotípicos del hospedero asociados a genes microbianos específicos. Para ello, se requiere un genoma de referencia y conocer previamente el microbioma específico con el que se interactúa, lo que representa una dificultad si éstos no se

conocen previamente o no han sido secuenciados. Por ejemplo, si se trata de un organismo cuyo genoma ya ha sido secuenciado, pero que no se había identificado previamente como parte del holobionte, el estudio se puede auxiliar de otros componentes del holobionte, como el holo-transcriptoma (Nyholm et al., 2020; Qin et al., 2012)

Por otro lado, con respecto al holo-transcriptoma asociado al holobionte éste, a diferencia del hologenoma, se define como el conjunto total de transcritos presente en el holobionte, y su estudio se puede abordar con experimentos de secuenciación masiva en paralelo de RNA pareada (del inglés *paired-end*) utilizando librerías del gen ribosomal 16s, cuya secuencia es específica entre los diferentes organismos y ha demostrado ser útil en la identificación de los microorganismos presentes en el tracto biliar que se relacionan con la condición neoplásica de este tejido (Lee et al., 2020). La secuenciación de tipo pareada del gen 16s se lleva a cabo con plataformas como MiSeq de Illumina (Kozich et al., 2013), para lo cual es necesario filtrar las secuencias para identificar los genes 16s rRNA con paquetes como *Mothur* (Schloss et al., 2009), y alineando con paquetes especiales como SILVA 16S rRNA. La identificación de los organismos del holobionte se puede realizar con los clasificadores bayesianos, como el descrito por Wang y colaboradores (Q. Wang et al., 2007), para finalmente identificar las unidades taxonómicas operacionales y con ello a los diferentes organismos presentes en la muestra evaluada, como se ha descrito en el estudio de Baxter y colaboradores, donde identificaron las unidades taxonómicas operacionales asociadas a *Pophyromonas assaccharolytica*, así como algunas otras bacterias presentes en el tracto gastrointestinal y cuya presencia puede utilizarse como biomarcador diagnóstico en padecimientos como el cáncer colorectal (Baxter et al., 2016). Además, se ha mostrado que no todos los componentes metagenómicos identificados a partir de la comparación de los rRNA 16S contribuyen en composición significativamente al hologenoma, por lo que es necesario el desarrollo de nuevas herramientas y metodologías bioinformáticas que permitan discriminar las relaciones significativas presentes entre los componentes del holobionte (Douglas & Werren, 2016)

Asimismo, en la determinación de las asociaciones estadísticamente significativas entre los componentes del holobionte, es decir, si estos componentes interactúan entre ellos como un sistema biológico integral, se utilizan de manera rutinaria análisis estadísticos como la prueba *PERMANOVA*, que se basa en el análisis multivariado llevado a cabo mediante permutaciones para determinar la respuesta simultánea de una o varias variables en un modelo de estudio con varios factores (Anderson, 2017). Incluso, ésta y otras herramientas se encuentran incluidas en una variedad de paqueterías informáticas diseñadas para análisis ecológico que se pueden implementar en el análisis de muestras clínicas en el estudio del hologenoma asociado a patologías (J. Yu et al., 2017), como es el caso de *VEGAN* (Dixon, 2003), que ha servido en la identificación de biomarcadores del hologenoma para la detección de lesiones cancerosas en el colon en estados tempranos (detectable desde la etapa temprana II (J. Yu et al., 2017) a través del uso de *PERMANOVA*, y *DAME* (Ding & Schloss, 2014), que es una plataforma en línea que permite analizar archivos de secuenciación de metagenomas de manera interactiva, lo que sugiere un avance práctico en las técnicas de diagnóstico para este padecimiento, ya que lo hace accesible a usuarios que no son expertos en bioinformática (Baxter et al., 2016). Entonces, la implementación de estas herramientas bioinformáticas en flujos de trabajo de investigación clínica, cuya funcionalidad ya ha sido comprobada en análisis ecológicos, permite ampliar el conocimiento del hologenoma para establecer su utilidad en

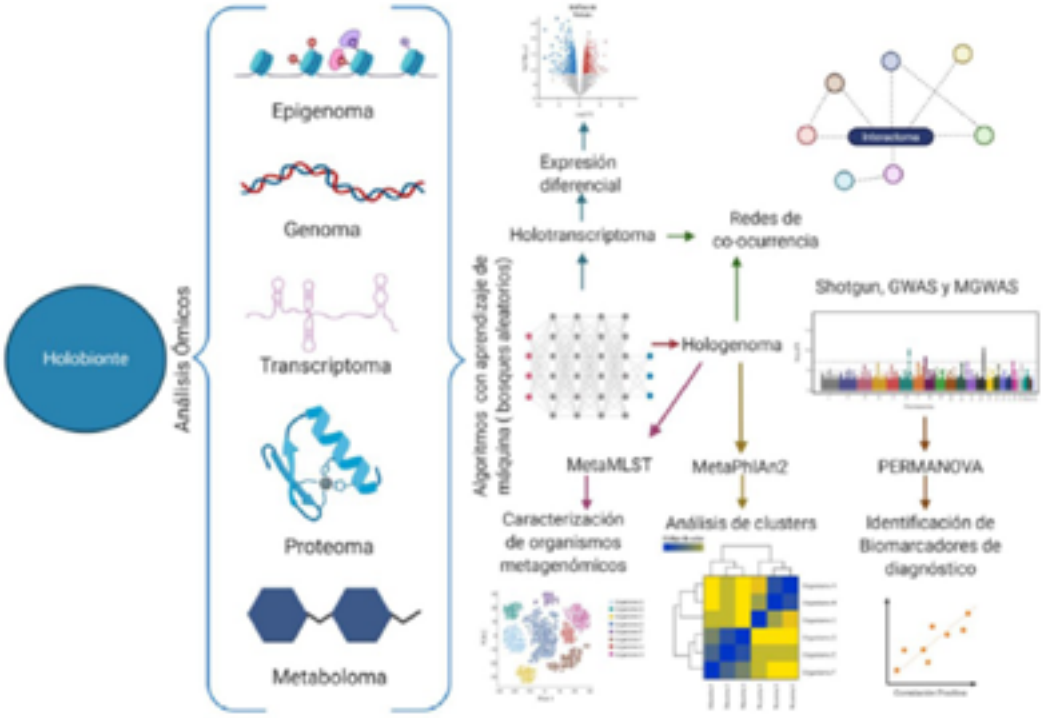
ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

el desarrollo de nuevos biomarcadores para padecimientos crónico degenerativos como la diabetes mellitus (Forslund et al., 2015; Qin et al., 2012), en donde también se ha incluido el cáncer (Apidianakis & Ferrandon, 2014; Flemer et al., 2018; Lee et al., 2020; J. Yu et al., 2017), lo que lleva a un avance en la medicina personalizada y llevando el manejo clínico de los pacientes de una manera óptima.

En resumen, el avance en las tecnologías de secuenciación, así como el desarrollo de nuevas herramientas bioinformáticas han permitido avanzar en la caracterización de los holobiontes, esto incluye desde mejorar la precisión con la cual se identifican los organismos presentes y la información genómica que puede adquirirse de ellos, hasta la caracterización de nuevos organismos no identificados previamente al conjuntar el análisis del hologenoma con el estudio ómico de otros componentes moleculares, como el transcriptoma. . Un ejemplo, es la identificación de una nueva cepa de *Helicobacter pylori* en el tracto gastrointestinal en población europea descrito por Fan y colaboradores (Fan & Pedersen, 2020) (Figura 2). Además, la implementación de herramientas bioinformáticas ya probadas en otras áreas como la ecología pueden ampliar la capacidad analítica del estudio del hologenoma y permitir su aplicación al ámbito clínico, como es el caso de los biomarcadores de diagnóstico en cáncer colorectal identificados por Baxter y colaboradores(Baxter et al., 2016), lo que representa un avance en el descubrimiento de nuevos biomarcadores moleculares, y lo cual contribuye con el desarrollo de la medicina de precisión en padecimientos como el cáncer y otras enfermedades crónico-degenerativas como la diabetes.

Figura 2

Análisis holo-ómicos



El estudio integrativo de los holobiontes permite establecer las asociaciones funcionales entre el hospedero y los otros organismos que componen este sistema biológico. Los flujos de trabajo bioinformáticos pueden mejorarse implementando el aprendizaje de máquina mediante el uso de bosques aleatorios para mejorar, por ejemplo, la identificación de los organismos presentes y sus características genómicas mediante diferentes tecnologías de secuenciación como *shotgun*, GWAS y MGWAS, para posteriormente identificar los grupos taxonómicos mediante análisis de *clusters* con herramientas como *MetaPhlAn2* o mediante análisis multivariados como el utilizado con la herramienta *MetaMLST*, que se ha demostrado puede tener aplicaciones clínicas en el desarrollo de nuevos biomarcadores diagnósticos. Además, la integración del análisis de otros componentes moleculares del holobionte como el holotranscriptoma permiten la identificación de transcritos expresados en el holosistema mediante análisis de expresión diferencial, lo cual, en conjunto con la información del hologenoma, permite conocer las asociaciones funcionales entre los diferentes componentes del holobionte, lo que puede denominarse como el interactoma del holobionte.

7. Conclusiones

El holobionte es una entidad biológica en donde la asociación entre el hospedero y los simbiontes afectan la aptitud del holobionte dentro de su entorno, donde la variación o mutaciones en el hologenoma puede ser provocada por cambios en los genomas del hospedador o del microbioma y bajo estrés ambiental, la comunidad microbiana simbiótica puede cambiar rápidamente y es donde la visión del microbioma ha evolucionado bastante desde el desarrollo de las técnicas de secuenciación masiva, donde anteriormente el estudio del microbioma era realizado mediante la identificación de microorganismos con técnicas dependientes de cultivo, restringiendo este estudio únicamente a microorganismos con capacidad de crecimiento en cultivo. Al poder identificar y estudiar la composición y distribución del microbioma en el cuerpo humano mediante técnicas de secuenciación masiva en paralelo de ácidos nucleicos y la consecuente introducción del término "microbioma" se empezó a ver la compleja asociación del ser humano y su microbioma, por lo tanto, ahora se ve al ser humano como un holobionte el cual depende, para mantener su estado de homeostasis, de una adecuada distribución y diversidad en el microbioma que lo compone.

Es importante resaltar que, aunque algunos de los efectos del microbioma se deben a interacciones directas de los microorganismos con las células del intestino, una parte de los efectos se deben a la estimulación de una respuesta inflamatoria por parte del hospedero. En respuesta al daño en la mucosa intestinal, el Receptor del Factor de Crecimiento Epidérmico (EGFR) activa cascadas de señalización de algunas subfamilias de cinasas de proteínas activadas por mitógenos (MAPK), como ERK1/2, que estimula la proliferación celular y bloquea la apoptosis, lo que promueve la formación de tumores colorrectales. Una parte considerable de la investigación actual en prevención y tratamiento de cáncer colorrectal se centra en comprender los diversos factores que participan en activación de esta respuesta inflamatoria. Se ha demostrado, por ejemplo, que una dieta abundante en colesterol y grasas estimula una mayor producción de ácidos biliares, que son metabolizados por el microbioma en el intestino en ácidos biliares secundarios, como el ácido desoxicólico, que estimula la activación de EGFR y la proliferación celular. Algunas bacterias del género *Clostridium*, por otra parte, metabolizan los ácidos

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

biliares en ácido ursodesoxicólico, que ha demostrado tener un papel protector contra el cáncer de colon y su administración a pacientes ha tenido resultados alentadores al prevenir la reaparición de tumores colorrectales.

Por otro lado, el metabolito producido por el microbioma que más extensamente ha sido estudiado por su efecto antitumoral es el butirato. Se ha demostrado que el butirato inhibe la desacetilasa de histona HDAC3, lo que lleva a la inactivación de las vías de Akt1 y Erk1/2, inhibiendo la proliferación y migración celular. Se descubrió de manera más reciente, que el butirato inhibe la transcripción del microRNA MiR-92a, lo que aumenta la expresión de genes supresores de tumores como la fosfatasa PTEN, que antagoniza la vía de PI3K, disminuyendo la proliferación celular y promoviendo la apoptosis. Asimismo, se vio que el butirato aumenta la transcripción de MiR-203, lo que disminuye la proliferación de la ubiquitina ligasa "Hakai", disminuyendo la transición epitelio-mesénquima y la metástasis de células neoplásicas. La comprensión de estas nuevas vías de regulación conlleva a pensar en tratamientos no solo basados en la regulación de la alimentación o manipulación de la composición del microbioma, sino en el suministro directo de inhibidores, como el ácido ursodexicólico o incluso terapias génicas dirigidas. El uso de diversos anti-mirs, por ejemplo, para disminuir los niveles de MiR-92a ha mostrado resultados prometedores en modelos animales y podría dirigir el futuro de los tratamientos para cáncer de colon en los próximos años.

Entonces, los flujos de trabajo y técnicas bioinformáticas en el análisis del hologenoma, el desarrollo de nuevas herramientas tecnológicas de secuenciación, como la secuenciación de nueva generación, que actualmente permite el análisis de células individuales (del inglés *single cell*), permitirá la identificación individual con mayor precisión respecto a la resolución actual que proveen las técnicas de bulto (del inglés *bulk*) para el análisis de muestras complejas. No obstante, algunas limitaciones como el tamaño celular y las interacciones físicas de los microorganismos presentes en el holobionte pueden significar un obstáculo en la preparación de las librerías individuales, por lo que el desarrollo de nuevas metodologías de separación celular de muestras es uno de los retos actuales en el estudio de los holobiontes. No obstante, en las últimas dos décadas el análisis del hologenoma y del holotranscriptoma ha podido perfeccionarse, como se ha discutido en este texto, con el desarrollo de nuevas herramientas bioinformáticas capaces de identificar un mayor número de organismos presentes en los holobiontes, los cuales se incluyen en las bases de datos y que finalmente enriquecen con la adición de nuevas especies a este tipo de bases de acceso público. Actualmente, en la era de la secuenciación masiva, una de las premisas de esta área de estudio es el enriquecimiento de las bases de datos ya conocidas, ya que éstas son la fuente de referencias para la identificación de los organismos en el holobionte, aunque es preciso mencionar que uno de los mayores retos actuales para los investigadores y desarrolladores es el perfeccionamiento de herramientas bioinformáticas con una mayos sensibiliad y especificidad de sus métodos de análisis, que son los parámetros que permiten establecer la confiabilidad de los nuevos descubrimientos. En este sentido, uno de los aportes al perfeccionamiento de las herramientas bioinformáticas y los flujos de trabajo, ha sido la incorporación de algoritmos que incluyen el aprendizaje de máquina, dado que elimina el error humano y ha permitido la identificación de nuevos organismos que componen los holobiontes, como se ve descrito en el trabajo de Baxter y colaboradores donde se identificaron nuevos biomarcadores en cáncer colorectal, y lo cual permite el establecimiento de asociaciones funcionales de relevancia

clínica en padecimientos como en el caso de cáncer de colon, y lo cual podría tener utilidad en el desarrollo de nuevos blancos terapéuticos así como la identificación de nuevos biomarcadores moleculares. Sin embargo, actualmente existe un número limitado de paqueterías específicas para el análisis de datos de secuenciación de hologenoma y holotranscriptoma con información clínica, y aunado a ello, existen pocos trabajos en los que se describa la incorporación de algoritmos con aprendizaje de máquina, por lo que una perspectiva para el desarrollo de flujos de trabajo bioinformático sería desarrollar nuevas plataformas o paqueterías amigables para usuarios no expertos en programación que permitan hacer análisis confiables de datos hologenómicos utilizando metadatos con información clínica para establecer asociaciones de utilidad clínica dentro del holobionte. Por ejemplo, las herramientas en línea que provee la plataforma Galaxy para análisis metagenómicos, o el caso de la aplicación web *DAME*. En ambos casos, ninguna de las plataformas es específica para el uso de datos clínicos, sin embargo, los flujos de trabajo pueden ser adaptados para obtener resultados clínicos relevantes y confiables en la identificación de organismos presentes en una muestra clínica, así como la identificación de nuevos biomarcadores o blancos terapéuticos. La recopilación específica de la información antes presentada también tiene como finalidad invitar al lector interesado en el análisis ómico del holobionte a que explore las nuevas posibilidades que le proporcionan las herramientas bioinformáticas descritas en este trabajo y que las identifique como nuevas oportunidades de análisis en el estudio de holobiontes con relevancia clínica. Asimismo, la incorporación rutinaria de algoritmos que utilicen el aprendizaje de máquina representa una alternativa analítica novedosa que permitirá la identificación de asociaciones funcionales estadísticamente significativas que no han sido descritas, lo cual contribuye con la descripción molecular y funcional de los sistemas holobiontes y proporcionará información de utilidad en el desarrollo de nuevas herramientas diagnósticas y su aplicación en la práctica clínica, lo cual también es un área de oportunidad para los desarrolladores de paqueterías y herramientas bioinformáticas, ya que como hemos mencionado antes, no existen paqueterías específicas, ni plataformas, para el análisis de muestras de holobiontes de relevancia clínica que sean amigables con usuarios no expertos en programación.

Finalmente, es indispensable seguir estudiando la relación entre el ser humano y su microbioma, así como su influencia con el desarrollo de cáncer con ayuda de nuevas plataformas de secuenciación y nuevas paqueterías bioinformáticas, para tener una visión global sobre el desarrollo de esta enfermedad y en un futuro utilizar al microbioma como una herramienta clínica a nuestro favor, ya sea como un posible biomarcador o un blanco terapéutico para el manejo clínico del paciente oncológico.

Agradecimientos

Agradecemos el apoyo de la Dra. María del Rocío Arellano Llamas y la M. en C. Clementina Castro Hernández por la lectura crítica y revisión de este artículo. Se agradece al Instituto Nacional de Cancerología (INCan) por su apoyo. Las figuras se crearon con BioRender.com.

Fuentes de Financiamiento

Laura Contreras-Espinosa es estudiante de maestría en el "Programa de Posgrado en Ciencias Biológicas, UNAM" y recibió una beca del Consejo Nacional de Ciencia y Tecnología (CONACYT) con CVU 1003211. Rogelio Montiel-Manríquez es estudiante de doctorado en el "Programa de Doctorado en Ciencias Biomédicas, UNAM con CVU: 581151.

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

Referencias

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., Arachchi, H., Arora, A., Auman, J. T., Ayala, B., Baboud, J., Balasundaram, M., Balu, S., Barnabas, N., Bartlett, J., Bartlett, P., Bastian, B. C., Baylin, S. B., Behera, M., … Zou, L. (2015). Genomic Classification of Cutaneous Melanoma. *Cell*, *161*(7), 1681-1696. https://doi.org/10.1016/j.cell.2015.05.044

AlHilli, M. M., & Bae-Jump, V. (2020). Diet and gut microbiome interactions in gynecologic cancer. *Gynecologic Oncology*, S0090825820338282. https://doi.org/10.1016/j.ygyno.2020.08.027

Anderson, M. J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). En *Wiley StatsRef: Statistics Reference Online* (pp. 1-15). American Cancer Society. https://doi.org/10.1002/9781118445112.stat07841

Apidianakis, Y., & Ferrandon, D. (2014). Modeling hologenome imbalances in inflammation and cancer. *Frontiers in Cellular and Infection Microbiology*, *4*. https://doi.org/10.3389/fcimb.2014.00134

Baedke, J., Fábregas-Tejeda, A., & Nieves Delgado, A. (2020). The holobiont concept before Margulis. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*, *334*(3), 149-155. https://doi.org/10.1002/jez.b.22931

Bail, J., Meneses, K., & Demark-Wahnefried, W. (2016). Nutritional Status and Diet in Cancer Prevention. *Seminars in Oncology Nursing*, *32*(3), 206-214. https://doi.org/10.1016/j.soncn.2016.05.004

Baquero, F., & Nombela, C. (2012). The microbiome as a human organ. *Clinical Microbiology and Infection*, *18*, 2-4. https://doi.org/10.1111/j.1469-0691.2012.03916.x

Bauer, M., Nascakova, Z., Mihai, A.-I., Cheng, P. F., Levesque, M. P., Lampart, S., Hurwitz, R., Pfannkuch, L., Dobrovolna, J., Jacobs, M., Bartfeld, S., Dohlman, A., Shen, X., Gall, A. A., Salama, N. R., Töpfer, A., Weber, A., Meyer, T. F., Janscak, P., & Müller, A. (2020). The ALPK1/TIFA/NF-κB axis links a bacterial carcinogen to R-loop-induced replication stress. *Nature Communications*, *11*(1), 5117. https://doi.org/10.1038/s41467-020-18857-z

Baxter, N. T., Ruffin, M. T., Rogers, M. A. M., & Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*, *8*(1), 37. https://doi.org/10.1186/s13073-016-0290-3

Biragyn, A., & Ferrucci, L. (2018). Gut dysbiosis: A potential link between increased cancer risk in ageing and inflammaging. *The Lancet Oncology*, *19*(6), e295-e304. https://doi.org/10.1016/S1470-2045(18)30095-0

Bisanz, J. E., Upadhyay, V., Turnbaugh, J. A., Ly, K., & Turnbaugh, P. J. (2019). Meta-Analysis Reveals Reproducible Gut Microbiome Alterations in Response to a High-Fat Diet. *Cell Host & Microbe*, *26*(2), 265-272. e4. https://doi.org/10.1016/j.chom.2019.06.013

Bonnet, M., Lagier, J. C., Raoult, D., & Khelaifia, S. (2020). Bacterial culture through selective and non-selective conditions: The evolution of culture media in clinical microbiology. *New Microbes and New Infections*, *34*, 100622. https://doi.org/10.1016/j.nmni.2019.100622

Bordenstein, S. R., & Theis, K. R. (2015). Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. *PLoS Biology*, *13*(8), e1002226. https://doi.org/10.1371/journal.pbio.1002226

Bose, S., Allen, A. E., & Locasale, J. W. (2020). The Molecular Link from Diet to Cancer Cell Metabolism. *Molecular Cell*, *78*(6), 1034-1044. https://doi.org/10.1016/j.molcel.2020.05.018

Brusselaers, N. (s. f.). Vaginal dysbiosis and the risk of human papillomavirus and cervical cancer: Systematic review and meta-analysis. *Systematic Reviews*, 18.

Calle, M. L., Urrea, V., Boulesteix, A.-L., & Malats, N. (2011). AUC-RF: A new strategy for genomic profiling with random forest. *Human Heredity*, *72*(2), 121-132. https://doi.org/10.1159/000330778

Cassidy, L. D., & Venkitaraman, A. R. (2012). Genome instability mechanisms and the structure of cancer genomes. *Current Opinion in Genetics & Development*, *22*(1), 10-13. https://doi.org/10.1016/j.gde.2012.02.003

Dias, T. R., Santos, J. M. O., Gil da Costa, R. M., & Medeiros, R. (2021). Long non-coding RNAs regulate the hallmarks of cancer in HPV-induced malignancies. *Critical Reviews in Oncology/Hematology*, *161*, 103310. https://doi.org/10.1016/j.critrevonc.2021.103310

Ding, T., & Schloss, P. D. (2014). Dynamics and associations of microbial community types across the human body. *Nature*, *509*(7500), 357-360. https://doi.org/10.1038/nature13178

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927-930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x

Douglas, A. E. (2015). Multiorganismal insects: Diversity and function of resident microorganisms. *Annual Review of Entomology*, *60*, 17-34. https://doi.org/10.1146/annurev-ento-010814-020822

Douglas, A. E., & Werren, J. H. (2016). Holes in the Hologenome: Why Host-Microbe Symbioses Are Not Holobionts. *MBio*, *7*(2). https://doi.org/10.1128/mBio.02099-15

Elinav, E., Garrett, W. S., Trinchieri, G., & Wargo, J. (2019). The cancer microbiome. *Nature Reviews Cancer*, *19*(7), 371-376. https://doi.org/10.1038/s41568-019-0155-3

Ezra-Nevo, G., Henriques, S. F., & Ribeiro, C. (2020). The diet-microbiome tango: How nutrients lead the gut brain axis. *Current Opinion in Neurobiology*, *62*, 122-132. https://doi.org/10.1016/j.conb.2020.02.005

Fan, Y., & Pedersen, O. (2020). Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-020-0433-9

Fatkhullina, A. R., Peshkova, I. O., Dzutsev, A., Aghayev, T., McCulloch, J. A., Thovarai, V., Badger, J., Vats, R., Sundd, P., Tang, H.-Y., Kossenkov, A. V., Hazen, S. L., Trinchieri, G., Grivennikov,

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

S. I., & Koltsova, E. K. (2018). An interleukin-23- interleukin-22 axis regulates intestinal microbial homeostasis to protect from diet-induced atherosclerosis. *Immunity*, *49*(5), 943-957.e9. https://doi.org/10.1016/j.immuni.2018.09.011

Flemer, B., Warren, R. D., Barrett, M. P., Cisek, K., Das, A., Jeffery, I. B., Hurley, E., O'Riordain, M., Shanahan, F., & O'Toole, P. W. (2018). The oral microbiota in colorectal cancer is distinctive and predictive. *Gut*, *67*(8), 1454-1463. https://doi.org/10.1136/gutjnl-2017-314814

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., Prifti, E., Vieira-Silva, S., Gudmundsdottir, V., Krogh Pedersen, H., Arumugam, M., Kristiansen, K., Yvonne Voigt, A., Vestergaard, H., Hercog, R., Igor Costea, P., Roat Kultima, J., Li, J., Jørgensen, T., … Pedersen, O. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature*, *528*(7581), 262-266. https://doi.org/10.1038/nature15766

Gholizadeh, P., Eslami, H., Yousefi, M., Asgharzadeh, M., Aghazadeh, M., & Kafil, H. S. (2016). Role of oral microbiome on oral cancers, a review. *Biomedicine & Pharmacotherapy*, *84*, 552-558. https://doi.org/10.1016/j.biopha.2016.09.082

Gilbert, S. F., Sapp, J., & Tauber, A. I. (2012). A symbiotic view of life: We have never been individuals. *The Quarterly Review of Biology*, *87*(4), 325-341. https://doi.org/10.1086/668166

Ginsburg, O., Ashton-Prolla, P., Cantor, A., Mariosa, D., & Brennan, P. (2020). The role of genomics in global cancer prevention. *Nature Reviews Clinical Oncology*. https://doi.org/10.1038/s41571-020-0428-5

Golrokh Mofrad, M., Taghizadeh Maleki, D., & Faghihloo, E. (2020). The roles of programmed death ligand 1 in virus-associated cancers. *Infection, Genetics and Evolution*, *84*, 104368. https://doi.org/10.1016/j.meegid.2020.104368

Gopalakrishnan, V., Helmink, B. A., Spencer, C. N., Reuben, A., & Wargo, J. A. (2018). The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy. *Cancer Cell*, *33*(4), 570-580. https://doi.org/10.1016/j.ccell.2018.03.015

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646-674. https://doi.org/10.1016/j.cell.2011.02.013

Hwang, J., Bae, H., Choi, S., Yi, H., Ko, B., & Kim, N. (2020). Impact of air pollution on breast cancer incidence and mortality: A nationwide analysis in South Korea. *Scientific Reports*, *10*(1), 5392. https://doi.org/10.1038/s41598-020-62200-x

Janney, A., Powrie, F., & Mann, E. H. (2020). Host–microbiota maladaptation in colorectal cancer. *Nature*, *585*(7826), 509-517. https://doi.org/10.1038/s41586-020-2729-3

Kanarek, N., Petrova, B., & Sabatini, D. M. (2020). Dietary modifications for enhanced cancer therapy. *Nature*, *579*(7800), 507-517. https://doi.org/10.1038/s41586-020-2124-0

Kitamoto, S. (s. f.). *The Intermucosal Connection between the Mouth and Gut in Commensal Pathobiont-Driven Colitis*. 31.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Applied and Environmental Microbiology*, *79*(17), 5112-5120. https://doi.org/10.1128/AEM.01043-13

Lee, H., Lee, H. K., Min, S. K., & Lee, W. H. (2020). 16S rDNA microbiome composition pattern analysis as a diagnostic biomarker for biliary tract cancer. *World Journal of Surgical Oncology*, *18*(1), 19. https://doi.org/10.1186/s12957-020-1793-3

LePage, D., & Bordenstein, S. R. (2013). Wolbachia: Can we save lives with a great pandemic? *Trends in Parasitology*, *29*(8), 385-393. https://doi.org/10.1016/j.pt.2013.06.003

Levy, M., Kolodziejczyk, A. A., Thaiss, C. A., & Elinav, E. (2017). Dysbiosis and the immune system. *Nature Reviews Immunology*, *17*(4), 219-232. https://doi.org/10.1038/nri.2017.7

Lien, E. C., & Vander Heiden, M. G. (2019). A framework for examining how diet impacts tumour metabolism. *Nature Reviews Cancer*, *19*(11), 651-661. https://doi.org/10.1038/s41568-019-0198-5

Limborg, M. T., Alberdi, A., Kodama, M., Roggenbuck, M., Kristiansen, K., & Gilbert, M. T. P. (2018). Applied Hologenomics: Feasibility and Potential in Aquaculture. *Trends in Biotechnology*, *36*(3), 252-264. https://doi.org/10.1016/j.tibtech.2017.12.006

Lynch, S. V., & Pedersen, O. (2016). The Human Intestinal Microbiome in Health and Disease. *New England Journal of Medicine*, *375*(24), 2369-2379. https://doi.org/10.1056/NEJMra1600266

Ma, W.-J., Vavre, F., & Beukeboom, L. W. (2014). Manipulation of arthropod sex determination by endosymbionts: Diversity and molecular mechanisms. *Sexual Development: Genetics, Molecular Biology, Evolution, Endocrinology, Embryology, and Pathology of Sex Determination and Differentiation*, *8*(1-3), 59-73. https://doi.org/10.1159/000357024

Mager, D., Haffajee, A., Devlin, P., Norris, C., Posner, M., & Goodson, J. (2005). The salivary microbiota as a diagnostic indicator of oral cancer: A descriptive, non-randomized study of cancer-free and oral squamous cell carcinoma subjects. *Journal of Translational Medicine*, *3*, 27. https://doi.org/10.1186/1479-5876-3-27

Malfertheiner, P., Link, A., & Selgrad, M. (2014). Helicobacter pylori: Perspectives and time trends. *Nature Reviews Gastroenterology & Hepatology*, *11*(10), 628-638. https://doi.org/10.1038/nrgastro.2014.99

Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., Gibbons, S. M., & Magis, A. T. (2020). Health and disease markers correlate with gut microbiome composition across thousands of people. *Nature Communications*, *11*(1), 5206. https://doi.org/10.1038/s41467-020-18871-1

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

Marx, W., Lane, M., Hockey, M., Aslam, H., Berk, M., Walder, K., Borsini, A., Firth, J., Pariante, C. M., Berding, K., Cryan, J. F., Clarke, G., Craig, J. M., Su, K.-P., Mischoulon, D., Gomez-Pinilla, F., Foster, J. A., Cani, P. D., Thuret, S., … Jacka, F. N. (2020). Diet and depression: Exploring the biological mechanisms of action. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-020-00925-x

Nelson, V. M., & Benson, A. B. (2017). Epidemiology of Anal Canal Cancer. *Surgical Oncology Clinics of North America*, *26*(1), 9-15. https://doi.org/10.1016/j.soc.2016.07.001

Noh, C.-K., Lee, G. H., Park, J. W., Roh, J., Han, J. H., Lee, E., Park, B., Lim, S. G., Shin, S. J., Cheong, J. Y., Kim, J. H., & Lee, K. M. (2020). Diagnostic accuracy of "sweeping" method compared to conventional sampling in rapid urease test for Helicobacter pylori detection in atrophic mucosa. *Scientific Reports*, *10*(1), 18483. https://doi.org/10.1038/s41598-020-75528-1

Nyholm, L., Koziol, A., Marcos, S., Botnen, A. B., Aizpurua, O., Gopalakrishnan, S., Limborg, M. T., Gilbert, M. T. P., & Alberdi, A. (2020). Holo-Omics: Integrated Host-Microbiota Multi-omics for Basic and Applied Biological Research. *IScience*, *23*(8), 101414. https://doi.org/10.1016/j.isci.2020.101414

Pevsner-Fischer, M., Tuganbaev, T., Meijer, M., Zhang, S.-H., Zeng, Z.-R., Chen, M.-H., & Elinav, E. (2016). Role of the microbiome in non-gastrointestinal cancers. *World Journal of Clinical Oncology*, *7*(2), 200-213. https://doi.org/10.5306/wjco.v7.i2.200

Pikor, L., Thu, K., Vucic, E., & Lam, W. (2013). The detection and implication of genome instability in cancer. *Cancer Metastasis Reviews*, *32*(3-4), 341-352. https://doi.org/10.1007/s10555-013-9429-5

Pitlik, S. D., & Koren, O. (2017). How holobionts get sick—Toward a unifying scheme of disease. *Microbiome*, *5*. https://doi.org/10.1186/s40168-017-0281-7

Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A. C. L., Gauthier, F., Magoulès, F., Ehrlich, S. D., & Pichaud, M. (2019). MSPminer: Abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics*, *35*(9), 1544-1552. https://doi.org/10.1093/bioinformatics/bty830

Polo, A., Arora, K., Ameur, H., Di Cagno, R., De Angelis, M., & Gobbetti, M. (2020). Gluten-free diet and gut microbiome. *Journal of Cereal Science*, *95*, 103058. https://doi.org/10.1016/j.jcs.2020.103058

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., … Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, *490*(7418), 55-60. https://doi.org/10.1038/nature11450

Raina, J.-B., Eme, L., Pollock, F. J., Spang, A., Archibald, J. M., & Williams, T. A. (2018). Symbiosis in the microbial world: From ecology to genome evolution. *Biology Open*, *7*(2). https://doi.org/10.1242/bio.032524

Revista Ciencia y Desarrollo. (2018, agosto 19). https://web.archive.org/web/20180819204101/http://www.cyd.conacyt.gob.mx/272/articulos/planta-animal-u-holobionte.html

Rizzi, S., Wensink, M., Ahrenfeldt, L. J., Christensen, K., & Lindahl-Jacobsen, R. (2020). Age-specific cancer rates: A bird's-eye view on progress. *Annals of Epidemiology*, *48*, 51-54.e1. https://doi.org/10.1016/j.annepidem.2020.04.007

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., & Han, Y. W. (2013). Fusobacterium nucleatum Promotes Colorectal Carcinogenesis by Modulating E-Cadherin/β-Catenin Signaling via its FadA Adhesin. *Cell Host & Microbe*, *14*(2), 195-206. https://doi.org/10.1016/j.chom.2013.07.012

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Horn, D. J. V., & Weber, C. F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, *75*(23), 7537-7541. https://doi.org/10.1128/AEM.01541-09

Schmidt, T. S. B., Raes, J., & Bork, P. (2018). The Human Gut Microbiome: From Association to Modulation. *Cell*, *172*(6), 1198-1215. https://doi.org/10.1016/j.cell.2018.02.044

Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, *13*(11), 800-812. https://doi.org/10.1038/nrc3610

Shanahan, F., Ghosh, T. S., & O'Toole, P. W. (2020). The Healthy Microbiome (What Is the Definition of a Healthy Gut Microbiome?). *Gastroenterology*, S0016508520355086. https://doi.org/10.1053/j.gastro.2020.09.057

Sharma, V. R., Singh, M., Kumar, V., Yadav, M., Sehrawat, N., Sharma, D. K., & Sharma, A. K. (2020). Microbiome dysbiosis in cancer: Exploring therapeutic strategies to counter the disease. *Seminars in Cancer Biology*, S1044579X20301619. https://doi.org/10.1016/j.semcancer.2020.07.006

Shindo, Y., Hazama, S., Tsunedomi, R., Suzuki, N., & Nagano, H. (2019). Novel Biomarkers for Personalized Cancer Immunotherapy. *Cancers*, *11*(9). https://doi.org/10.3390/cancers11091223

Simon, J.-C., Marchesi, J. R., Mougel, C., & Selosse, M.-A. (2019a). Host-microbiota interactions: From holobiont theory to analysis. *Microbiome*, *7*(1), 5. https://doi.org/10.1186/s40168-019-0619-4

Simon, J.-C., Marchesi, J. R., Mougel, C., & Selosse, M.-A. (2019b). Host-microbiota interactions: From holobiont theory to analysis. *Microbiome*, *7*(1), 5. https://doi.org/10.1186/s40168-019-0619-4

Steck, S. E., & Murphy, E. A. (2020). Dietary patterns and cancer risk. *Nature Reviews Cancer*, *20*(2), 125-138. https://doi.org/10.1038/s41568-019-0227-4

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico

Sun, Y., Ge, X., Li, X., He, J., Wei, X., Du, J., Sun, J., Li, X., Xun, Z., Liu, W., Zhang, H., Wang, Z.-Y., & Li, Y. C. (2020). High-fat diet promotes renal injury by inducing oxidative stress and mitochondrial dysfunction. *Cell Death & Disease*, *11*(10), 914. https://doi.org/10.1038/s41419-020-03122-4

Takahashi, Y., Park, J., Hosomi, K., Yamada, T., Kobayashi, A., Yamaguchi, Y., Iketani, S., Kunisawa, J., Mizuguchi, K., Maeda, N., & Ohshima, T. (2019). Analysis of oral microbiota in Japanese oral cancer patients using 16S rRNA sequencing. *Journal of Oral Biosciences*, *61*(2), 120-128. https://doi.org/10.1016/j.job.2019.03.003

Ternes, D., Karta, J., Tsenkova, M., Wilmes, P., Haan, S., & Letellier, E. (2020). Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? *Trends in Microbiology*, *28*(5), 401-423. https://doi.org/10.1016/j.tim.2020.01.001

The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207-214. https://doi.org/10.1038/nature11234

Theis, K. R., Dheilly, N. M., Klassen, J. L., Brucker, R. M., Baines, J. F., Bosch, T. C. G., Cryan, J. F., Gilbert, S. F., Goodnight, C. J., Lloyd, E. A., Sapp, J., Vandenkoornhuyse, P., Zilber-Rosenberg, I., Rosenberg, E., & Bordenstein, S. R. (2016). Getting the Hologenome Concept Right: An Eco-Evolutionary Framework for Hosts and Their Microbiomes. *MSystems*, *1*(2). https://doi.org/10.1128/mSystems.00028-16

Thompson, J., Johansen, R., Dunbar, J., & Munsky, B. (2019). Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLOS ONE*, *14*(7), e0215502. https://doi.org/10.1371/journal.pone.0215502

Tierney, B. T., Yang, Z., Luber, J. M., Beaudin, M., Wibowo, M. C., Baek, C., Mehlenbacher, E., Patel, C. J., & Kostic, A. D. (2019). The Landscape of Genetic Content in the Gut and Oral Human Microbiome. *Cell Host & Microbe*, *26*(2), 283-295.e8. https://doi.org/10.1016/j.chom.2019.07.008

Tobore, T. O. (2020). Towards a comprehensive theory of obesity and a healthy diet: The causal role of oxidative stress in food addiction and obesity. *Behavioural Brain Research*, *384*, 112560. https://doi.org/10.1016/j.bbr.2020.112560

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, *12*(10), 902-903. https://doi.org/10.1038/nmeth.3589

Underhill, D. M., & Iliev, I. D. (2014). The mycobiota: Interactions between commensal fungi and the host immune system. *Nature Reviews Immunology*, *14*(6), 405-416. https://doi.org/10.1038/nri3684

van de Guchte, M., Blottière, H. M., & Doré, J. (2018). Humans as holobionts: Implications for prevention and therapy. *Microbiome*, *6*(1), 81. https://doi.org/10.1186/s40168-018-0466-8

van de Wijgert, J. H. H. M., & Jespers, V. (2017). The global health impact of vaginal dysbiosis. *Research in Microbiology*, *168*(9-10), 859-864. https://doi.org/10.1016/j.resmic.2017.02.003

Vogtmann, E., & Goedert, J. J. (2016). Epidemiologic studies of the human microbiome and cancer. *British Journal of Cancer*, *114*(3), 237-242. https://doi.org/10.1038/bjc.2015.465

Wang, J., & Jia, H. (2016). Metagenome-wide association studies: Fine-mining the microbiome. *Nature Reviews Microbiology*, *14*(8), 508-522. https://doi.org/10.1038/nrmicro.2016.83

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261-5267. https://doi.org/10.1128/AEM.00062-07

Yang, Y., Weng, W., Peng, J., Hong, L., Yang, L., Toiyama, Y., Gao, R., Liu, M., Yin, M., Pan, C., Li, H., Guo, B., Zhu, Q., Wei, Q., Moyer, M.-P., Wang, P., Cai, S., Goel, A., Qin, H., & Ma, Y. (2017). Fusobacterium nucleatum Increases Proliferation of Colorectal Cancer Cells and Tumor Development in Mice by Activating Toll-Like Receptor 4 Signaling to Nuclear Factor−κB, and Up-regulating Expression of MicroRNA-21. *Gastroenterology*, *152*(4), 851-866.e24. https://doi.org/10.1053/j.gastro.2016.11.018

Yu, J., Feng, Q., Wong, S. H., Zhang, D., Liang, Q. yi, Qin, Y., Tang, L., Zhao, H., Stenvang, J., Li, Y., Wang, X., Xu, X., Chen, N., Wu, W. K. K., Al-Aama, J., Nielsen, H. J., Kiilerich, P., Jensen, B. A. H., Yau, T. O., … Wang, J. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, *66*(1), 70-78. https://doi.org/10.1136/gutjnl-2015-309800

Yu, S., Xiong, Y., Fu, Y., Chen, G., Zhu, H., Mo, X., Wu, D., & Xu, J. (2021). Shotgun metagenomics reveals significant gut microbiome features in different grades of acute pancreatitis. *Microbial Pathogenesis*, *154*, 104849. https://doi.org/10.1016/j.micpath.2021.104849

Zolfo, M., Tett, A., Jousson, O., Donati, C., & Segata, N. (2017). MetaMLST: Multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Research*, *45*(2), e7-e7. https://doi.org/10.1093/nar/gkw837

ISSN 1405-6690 impreso
ISSN 1665-8612 electrónico