UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

IDENTIFICACIÓN DE GENES CODIFICANTES Y ELEMENTOS REPETIDOS REGULADOS POR DAXX MEDIANTE RNA-SEQ

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

B I Ó L O G O

P R E S E N T A:

ERICK ISAAC NAVARRO DELGADO

DIRECTOR DE TESIS:
DR. RODRIGO GONZÁLEZ BARRIOS DE LA PARRA

CO-ASESOR DE TESIS:
DR. NICOLÁS ALCARAZ MILLMAN

Ciudad Universitaria, Cd. Mx., 2021

Dedicatoria

A mi mamá, por siempre confiar en mí y alentar mis sueños en cualquier circunstancia.

A mi papá, por enseñarme el papel irremplazable del trabajo y la perseverancia en la vida.

A mi hermano, por ser mi guía y compañero desde que tengo memoria.

A Edith, por escucharme, apoyarme y siempre estar.

Gracias a todos por ser una fuente inagotable de motivación e inspiración.

Agradecimientos

Al Dr. Rodrigo González Barrios por ayudarme a realizar este proyecto, introducirme al increíble mundo de la epigenética y confiar en mí para colaborar en otros proyectos de su laboratorio.

Al Dr. Nicolás Alcaraz por guiarme en todo lo relacionado al análisis de datos de secuenciación e instruirme en el campo de la bioinformática.

A mis sinodales, la Dra. Thamara Juárez, la Dra. Tzvetanka Dimitrova y el Dr. Ernesto Soto, por sus críticas constructivas que indudablemente contribuyeron a la mejora de este trabajo.

A mis amigos de la preparatoria y universidad por inspirarme, aconsejarme y contribuir a mi formación dentro y fuera de las aulas.

A la Universidad Nacional Autónoma de México, por otorgarme tantas herramientas y oportunidades para mi desarrollo personal y profesional.

Al Taller de Ciencia para Jóvenes del CIMAT y al equipo de las Olimpiadas de Biología de la Academia Mexicana de Ciencias, así como a las personas que participan en su organización durante las etapas estatales, nacionales e internacionales. Mi experiencia en ambos programas fue crucial para que encontrara mi pasión por la biología y decidiera emprender una carrera en el ámbito científico.

Finalmente, agradezco al proyecto CONACYT 290041 por brindar apoyo para la realización de este trabajo.

Índice

Resumen

La célula eucarionte posee mecanismos de regulación epigenética que le permiten modular los genes que se expresan en un momento dado. Con ayuda de estos mecanismos, esta es capaz de responder a los estímulos y generar un perfil de expresión que se adapte al ambiente y preserve su identidad celular. Dentro de los niveles de regulación pre-transcripcional, las variantes de histonas juegan un papel muy importante al formar nucleosomas especializados. Estas estructuras desempeñan funciones específicas en el núcleo celular, tales como mantenimiento de la heterocromatina, apertura transcripcional, formación del cromosoma mitótico, entre otras.

Uno de los elementos más importantes para que las variantes de histonas cumplan su función de manera correcta son las proteínas chaperonas, las cuales se encargan de depositar al nucleosoma variante en las regiones correctas del genoma. A su vez, estas pueden asociarse con factores epigenéticos que modifican el epigenoma circundante y desencadenan la función celular asociada con la presencia de dicho nucleosoma especializado.

En este contexto, la chaperona DAXX se encarga de depositar a la variante de histona H3.3 en el genoma en regiones ricas en elementos repetitivos y genes codificantes. La presencia de la variante de histona H3.3 se asocia principalmente con la represión transcripcional de los loci donde es depositada, así como en el mantenimiento de los dominios de heterocromatina en el núcleo celular. Además, la alteración en la expresión de DAXX está relacionada con inestabilidad cromosómica y fenotipos agresivos in vivo e in vitro en distintos tipos de cáncer. A pesar de esto, pocos estudios han descrito los genes codificantes regulados por esta chaperona, y ninguno ha explorado los elementos repetitivos involucrados en este fenómeno.

En esta tesis, se analizó el transcriptoma de las líneas celulares humanas HCT116, PC3 y GSC23 con el objetivo de describir los principales elementos alterados tras la represión de esta proteína. Al analizar los procesos biológicos desregulados, se encontró que DAXX modula principalmente genes tejido-específico en cada línea celular. Aún así, se encontraron genes sobreexpresados en común que componen una firma basal en todos los tejidos

estudiados. Todos ellos son importantes en el desarrollo del cáncer, como ALDOA, PLAU, MGAT4B y S100A11. Este patrón de poca similitud se repitió en los elementos repetitivos, donde la intersección entre cada experimento fue pequeña o nula. En cuestión de repetidos, cada línea celular presentó repetidos desregulados como ACRO1 en HCT116, TAR1 en PC3 o MSTB2 en GSC23. Sorprendentemente, se encontró que DAXX podría estar involucrada en la activación transcripcional. Los resultados sugieren un mecanismo en el cual DAXX se asocia a factores transcripcionales tejido específico para modular la expresión de genes en la célula; no obstante, algunos pocos son regulados independientemente del tipo celular. Estos resultados son particularmente relevantes para identificar elementos importantes en el fenotipo agresivo ocasionado por la desregulación de DAXX en cáncer.

1. Introducción

El genoma es el conjunto total de DNA que contiene un organismo en sus células; contiene la información necesaria para el desarrollo de la mayoría de las características de un individuo(1). Este conjunto de moléculas posee diversos segmentos que, al ser transcritos, ejercen funciones biológicas en la célula (i.e. genes). A pesar de que aproximadamente el 80% del genoma humano es capaz de ser transcrito (2), solamente una pequeña proporción de él se encuentra transcripcionalmente activa simultáneamente en las células. La distinta combinación de genes expresados en un momento dado, lo cual es conocido como transcriptoma, genera una gran variedad de perfiles que originan la diversidad de tipos celulares que alberga el cuerpo humano. Dado que esta combinación determina la identidad y el estado de una célula, la regulación estricta de los genes activos y reprimidos es un asunto crucial para la correcta función de las células y por ende de los tejidos en un organismo sano, especialmente en seres multicelulares (1). Es debido a esto que existen mecanismos moleculares que determinan la permisividad de una región del DNA a ser transcrita, lo cual es estudiado por la epigenética.

1.1 La cromatina puede regular la expresión de los genes a nivel pre-transcripcional

A nivel pre-transcripcional, los genes suelen ser regulados mediante la asociación del DNA con distintas proteínas, formando un complejo conocido como cromatina. La estructura local y regional de este complejo regula la expresión de los genes, pues su nivel de compactación determina la accesibilidad de proteínas involucradas en la transcripción (1). De esta manera, la estructura de la cromatina es uno de los niveles basales de regulación de la expresión génica en una célula.

La cromatina posee distintos niveles de compactación (**Figura 1.1.1A**), los cuales varían a lo largo del ciclo celular y dependiendo de las regiones genómicas. La estructura básica de la cromatina es el nucleosoma, generalmente compuesto por un octámero de cuatro histonas canónicas diferentes en igual proporción: H2A, H2B, H3 y H4 (**Figura 1.1.1B**). El nucleosoma sirve principalmente para compactar la cromatina, puesto que en él enredan aproximadamente 146 pares de bases de una cadena de DNA mediante interacciones iónicas

(3). Los genes de las histonas se encuentran agrupados en el genoma y son transcritos únicamente en la fase de duplicación del DNA, pues su inserción en el material genético ocurre en la horquilla de replicación inmediatamente después de la síntesis de la cadena complementaria y en sitios de reparación (3).

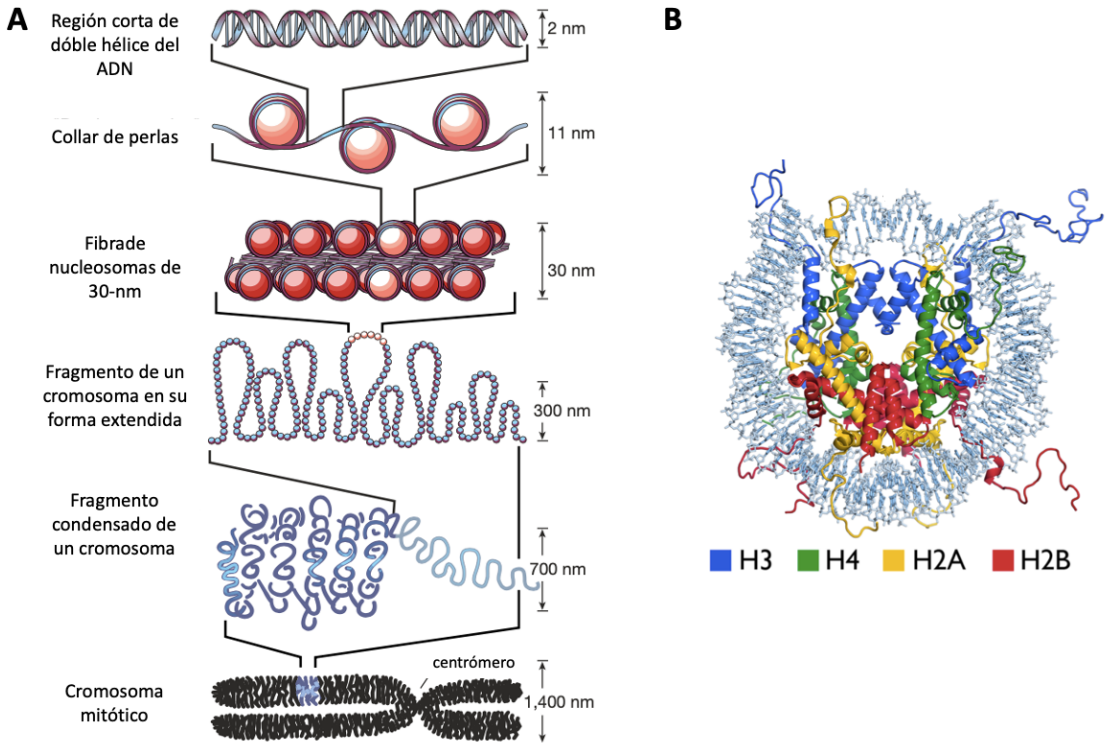


Figura 1.1.1. Niveles de compactación de la cromatina. A) Diferentes estructuras formadas durante la compactación; figura modificada de Felsenfeld & Groudine, 2003 (4). B) Estructura del nucleosoma. Los colores indican las diferentes histonas que componen al complejo canónico; figura tomada de McGinty & Tan, 2015 (5)

Aunque únicamente 4 tipos de histonas componen al nucleosoma canónico, los genes que codifican a estas proteínas presentan múltiples duplicaciones. Particularmente en humanos, se alberga un clúster mayor y menor de histonas, dos arreglos largos en el cromosoma 1 y 6 que contienen 55 y 13 genes de histonas cada uno (6). Estas proteínas estructurales son esenciales para la estabilidad y funcionalidad de la cromatina y componen aproximadamente el 50% de la masa del cromosoma eucarionte; su importancia se ve reflejada en el hecho de son las proteínas más conservadas a lo largo de la evolución de este clado (7).

Las histonas canónicas desempeñan funciones estructurales, de protección y de regulación. Al enrollarse el DNA en las histonas, reduce su volumen ocupado en aproximadamente un 500%, lo cual es esencial para mantener el pequeño tamaño del núcleo y, por ende, de la célula. Además de esta función de compactación, estas proteínas protegen al material genético de la degradación mediada por nucleasas (8) y de la oxidación (9), además de participar en la homeostasis celular de algunos iones gracias a su actividad reductora, como la del Cu^{+1} (10).

Finalmente, las histonas tienen un rol regulatorio en la expresión génica, lo cual sucede por dos mecanismos principalmente. El más sencillo de ellos sucede al reprimir la activación de los genes mediante el impedimento de la interacción del DNA con los activadores transcripcionales. El segundo, y más complejo, es mediante la adición covalente de pequeñas moléculas a los aminoácidos de las histonas que sobresalen del nucleosoma, en las regiones conocidas como colas de histonas (extremos amino terminales). Estas modificaciones postraduccionales son conocidas como "marcas epigenéticas de histonas", y fungen como etiquetas que son reconocidas por proteínas que efectúan alguna acción sobre la cromatina. El efecto puede ser silenciar un gen, activarlo, modular su nivel de expresión, reclutar proteínas específicas a sitios particulares, entre otras cosas. La adición de marcas epigenéticas es un proceso dinámico, y los grupos añadidos varían dependiendo del aminoácido a ser modificado; la identidad del grupo químico, junto a su posición, genera una gran variedad de combinaciones. En su conjunto, estas marcas establecen un código de histonas que interacciona positiva o negativamente con la maquinaria transcripcional y señalizan distintos fenómenos en la cromatina (4).

Ascendiendo en la escala de compactación de la cromatina, entre cada nucleosoma existe un pequeño segmento de DNA de unión de alrededor de 60 pares de bases que interconecta al uno con el otro, formando la estructura conocida como "collar de perlas"; este segmento de DNA de unión se asocia con la histona H1, la cual acumula múltiples segmentos para compactar aún más la cromatina y formar las fibras de 30 nanómetros. Estas fibras pueden generar diversas estructuras de alto orden, como asas y segmentos condensados que generan

una arquitectura nuclear que le dan estabilidad a la cromatina durante la interfase; los mecanismos moleculares precisos involucrados en al formación de estas estructuras aún no se conocen a detalle y son sujetos a investigación activa (4) .

1.2 Las variantes de histonas amplían la diversidad de funciones del nucleosoma

Además de las funciones anteriormente descritas, el nucleosoma se ve involucrado en una mayor gama de procesos biológicos, como es la reparación del DNA, recombinación meiótica, segregación cromosómica, desarrollo, iniciación y terminación de la transcripción, represión o activación de regiones genómicas, entre otros (11). Estas funciones se realizan a través de la presencia de proteínas histonas que difieren en la estructura primaria de las canónicas. Estas proteínas, mejor conocidas como variantes de histonas, poseen características distintas, lo cual diversifica la función de la cromatina en la célula.

Las variantes de histonas pueden estar asociadas a fenómenos biológicos especializados. Los cambios en su secuencia de aminoácidos les permite interactuar con un conjunto distinto de proteínas, presentar interacciones DNA-histonas que desemboquen en nucleosomas más lábiles, o ser modificadas postraduccionalmente en aminoácidos únicos de las variantes. Los genes que codifican a estas proteínas son distintos a los de las histonas canónicas y se encuentran en distintas posiciones a lo largo del genoma; estos también pueden estar presentes en múltiples copias, aunque no se concentran en arreglos cromosómicos. Si bien el cambio en la secuencia de aminoácidos con respecto a la versión canónica de cada variante es distinto, incluso las que poseen diferencias más sutiles son capaces de generar distintos perfiles nucleosómicos queal estar enriquecidas en regiones genómicas específicas, generan distintas conformaciones en la cromatina, lo cual les otorga nuevas propiedades a dichos *loci* (12).

A diferencia de las histonas canónicas, las variantes de histonas pueden ser sintetizadas y posicionadas en el genoma a lo largo de todo el ciclo celular de forma independiente a la replicación (3). De esta manera, la sustitución de histonas canónicas por variantes modifica la composición de la cromatina y es capaz de agregar flexibilidad y dinamismo a la expresión génica. Esto es esencial para que una célula responda y se adapte a su entorno. Así, la

existencia de estas proteínas añade mayor complejidad al papel que desempeña el nucleosoma en la regulación de los procesos nucleares.

1.3 H3.3: la variante de histona mejor descrita de H3

Una de las variantes de histonas mejor descritas es H3.3, cuya versión canónica es la H3.1 o H3.2 (H3). En mamíferos, esta proteína está codificada por los genes H3F3A y H3F3B, ubicados en los cromosomas 1 y 17, respectivamente. La proteína H3.3 difiere de la H3 en 5 aminoácidos localizados en la superficie accesible de la histona. De esta manera, los cambios no afectan la estructura del nucleosoma, sino más bien su interacción con distintas proteínas (13). Las modificaciones más relevantes se encuentran en el inicio de la alfa-hélice 2, donde el motivo SAVM cambia por AAIG, y en la posición 31, donde el residuo de aminoácido A cambia por una S (**Figura 1.3.1**). El primer cambio mencionado es importante porque este motivo le permite a las proteínas que interaccionan específicamente con H3.3 diferenciarla de la H3, mientras que el segundo cambio es relevante porque la serina es susceptible a ser fosforilada (12). Las implicaciones de esta última diferencia en la secuencia se encuentran en fases tempranas de estudio. No obstante, los hallazgos encontrados sugieren que este pequeño cambio tiene grandes implicaciones en la expresión global de los genes durante el desarrollo debido a que esta marca epigenética es necesaria para la acetilación de histonas posicionadas en regiones de *enhancers* y promotores en células troncales embrionarias (14).

En la célula, el correcto funcionamiento de muchas proteínas dependen de su chaperona, la cual es una proteína que se une a ellas para asegurar su plegamiento, ensamblaje y transporte celular (1). Un fenómeno interesante en H3.3 es que su comportamiento puede cambiar en gran medida dependiendo de la chaperona que la deposita en la cromatina y de las proteínas presentes en el núcleo con las que interacciona. A diferencia de la histona H3 canónica, cuya chaperona principal es CAF-1, H3.3 posee dos complejos encargados de proteger a esta variante de interacciones no deseadas, asegurar su correcto plegamiento y posicionarla en las regiones correctas: Histone Cell Cycle Regulator – Ubiniclein 1 (HIRA-UBN1) y alpha-thalassemia/mental retardation X-linked - Death domain-associated protein 6 ATRX-DAXX (16).

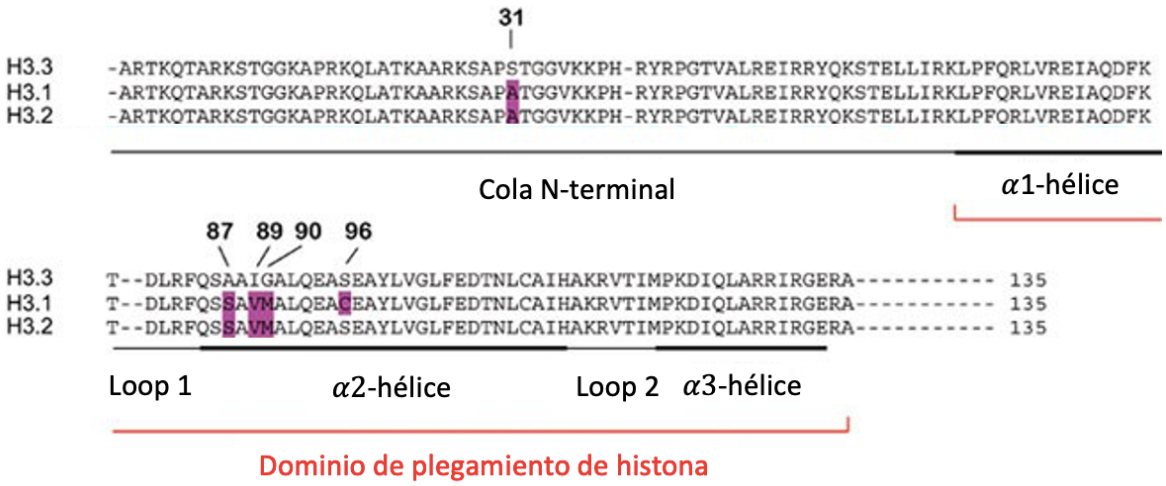


Figura 1.3.1 Comparación de la estructura primaria canónica de H3 y de H3.3. Se resalta la posición de los 5 aminoácidos de la histona H3 (genes H3.1 y H3.2) que difieren de su variante H3.3. El cambio de A por S en la posición 31 es particularmente importante porque le permite a H3.3 adquirir una modificación postraduccional única. Figura modificada de Szenker, *et al.,* 2011 (15).

Los complejos previamente mencionados se asocian a fenómenos biológicos muy distintos. HIRA-UBN1 se asocia al depósito de H3.3 en eucromatina (cromatina poco compacta que usualmente se asocia con actividad transcripcional), especialmente en elementos activos regulatorios y en el cuerpo de genes activos (12). Las histonas depositadas por este complejo se encuentran enriquecidas en marcas epigenéticas de activación como acetilación o H3K36me3 (12). Por otro lado, de manera sorprendente, el depósito de H3.3 mediante el complejo ATRX-DAXX se encuentra asociado a heterocromatina, predominantemente en elementos endógenos retrovirales, telómeros, regiones pericentroméricas, repetidos cortos en tándem y alelos silenciados de genes improntados (17). Esto ayuda a mantener un entorno heterocromático en dichas regiones mediante la asociación del nucleosoma modificado con proteínas represoras como Heterochromatic Protein 1 (HP1), SET Domain Bifurcated Histone Lysine Methyltransferase 1 (SETDB1), KRAB-associated protein 1 (KAP1) y Suppressor Of Variegation 3-9 Homolog 1 (SUV39H1)(3). Las chaperonas de histonas juegan así un papel muy importante en el depósito correcto de H3.3, pues determinan las regiones en las cuales es incorporada esta variante, e influye fuertemente en la función que esta le otorgará al nucleosoma.

El modelo más aceptado del depósito de la histona H3.3 con el complejo ATRX-DAXX en regiones heterocromáticas establece que DAXX reconoce específicamente a la histona H3.3 al interaccionar con la glicina 90 (Gly90), que es única de esta variante, siendo metionina en su versión canónica. Posteriormente, DAXX guía al complejo a los cuerpos nucleares, donde se asocia con ATRX y con otras proteínas represoras, como KAP1 y SETDB1, para formar un complejo. Cabe mencionar que SETDB1 puede colocar la marca epigenética H3.3K9me1 en el dímero H3.3-H4 previo a su incorporación a la cromatina. Después, ATRX reconoce la marca H3K9me3 en la heterocromatina y direcciona al complejo para depositar a H3.3 en regiones adyacentes. Finalmente, SUV39H1/2 y HP1 se encargan de establecer y propagar la marca H3K9me3, manteniendo de esta forma la estructura represiva de la región (3). Como podemos observar en este modelo (**Figura 1.3.2**), DAXX juega un papel central en la correcta funcionalidad de la histona H3.3 en heterocromatina, al ser la proteína que se une directamente a ella y a los demás miembros del complejo. Esto permite que esta variante se deposite en regiones específicas y ejerza un rol represivo.

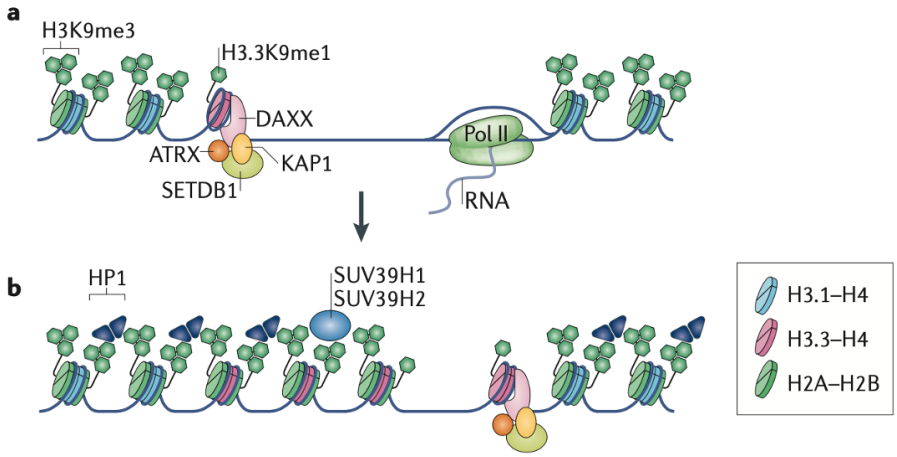


Figura 1.3.2. Represión de regiones genómicas mediadas por el depósito de H3.3 por el complejo DAXX/ATRX. a) Zonas libres de nucleosomas aparecen en regiones adyacentes a la heterocromatina como resultado de distintos procesos, como puede ser la transcripción. Posteriormente, ATRX en complejo con DAXX, el tetrámero con H3.3 y proteínas represoras, reconocen la marca adyacente de H3K9me3 y guía el depósito del nucleosoma no canónico. b) Generación de nuevas marcas epigenéticas represoras y su extensión, mediante SUV39H1 y HP1, promoviendo el mantenimiento del ambiente heterocromático en la región. Figura tomada de Talbert & Henikoff, 2017 (3).

1.4 DAXX y su desregulación en cáncer.

La alteración en la regulación de las variantes de histonas en distintos niveles se ha visto asociada a diversas enfermedades; esta alteración puede presentarse en forma de mutaciones o cambios de expresión en los genes de las variantes o en sus chaperonas. En el caso particular de H3.3, debido a su papel central en el mantenimiento de la heterocromatina y en la regulación de genes durante el desarrollo, tanto la deleción de ambos de sus genes (*H3F3A* y *H3F3B*) como la de sus chaperonas (*ATRX*, *DAXX* o *HIRA*) es letal, lo cual ha sido atribuido principalmente a la inestabilidad genómica (12). Es debido a esto que sus efectos son primariamente visibles en enfermedades donde dichas mutaciones son somáticas, i.e. una vez que el individuo se ha desarrollado.

Una de las formas por las cuales este mecanismo regulatorio puede verse alterado es mediante el incorrecto funcionamiento de las chaperonas, pues de estas depende el depósito efectivo de las variantes de histonas. Por lo tanto, dependiendo de la chaperona afectada, se puede generar una desregulación en diversos mecanismos epigenéticos, los cuales pueden ocasionar alteraciones graves en la célula como lo es el desarrollo de inestabilidad genómica (3). En este sentido, diversos estudios han demostrado la importancia de DAXX en el mantenimiento de la estabilidad del genoma, así como los efectos de su alteración, como se menciona a continuación.

Previamente, algunos estudios han demostrado que DAXX desempeña un papel crucial en el mantenimiento de la estructura cromatínica. En fibroblastos embrionarios de ratón, la ausencia de esta proteína ocasiona una alteración estructural en los dominios enriquecidos con la marca epigenética represiva H3K9me3. Esta marca, propia de heterocromatina constitutiva, deja de asociarse en dichas regiones y se presenta una pérdida de los límites espaciales entre la heterocromatina perinucleolar y el nucleolo. Además, la integridad de los nucleolos y la organización del DNA ribosomal se ve amenazada, lo cual se ve reflejado en un aumento de estructuras anormales llamadas mini-nucleolos (18).

Adicionalmente, la subexpresión de DAXX ha demostrado ser suficiente para generar inestabilidad cromosómica y micronúcleos en células humanas (19). La generación de inestabilidad cromosómica es particularmente importante, pues es muy frecuente en tumores sólidos (70-80%) y se considera esencial en el proceso de transformación de células cancerosas (20). Además, esta pérdida parcial de DAXX promueve una disminución de la metilación de DNA en repetidos pericentroméricos y teloméricos (19), lo cual ha sido asociado con la acumulación de alteraciones cromosómicas (21,22) y la presencia de fenotipos agresivos en cáncer de mama (21). De manera similar, en líneas celulares de cáncer de próstata, la subexpresión de DAXX ocasiona una pérdida en sitios de unión de la DNA metiltransferasa 1 (DNMT1) al genoma (23). Esta enzima se encarga de la metilación de mantenimiento, que es esencial para asegurar la represión de elementos transponibles en las células y conservar la identidad celular (24). Finalmente, la sobreexpresión de DAXX en líneas celulares no neoplásicas de próstata ocasiona números cromosómicos anormales (25), lo cual es una consecuencia clara de inestabilidad genómica.

Por otra parte, en el contexto del cáncer, la expresión anormal de DAXX en tejidos neoplásicos es una característica usual de distintos tumores, y se ha visto asociada a la aparición de características clínicas desfavorables tanto cuando aumenta como cuando disminuye su expresión. Por un lado, la sobreexpresión de DAXX correlaciona con un aumento en tumorigenicidad, progresión de la enfermedad y resistencia a tratamientos para el cáncer (26). Adicionalmente, este aumento de expresión es constante en diversos cánceres, como el de próstata (25,27), ovario (28), gástrico (29), glioblastoma (26) y carcinoma oral de células escamosas (30), entre otros. Finalmente, los niveles de expresión de DAXX son significativamente mayores en metástasis que en tumores primarios en cánceres de colon, mama y próstata (26).

Por otro lado, a pesar de tener una incidencia baja en cánceres comúnmente diagnosticados, hay mutaciones recurrentes en algunos tipos, como es en el carcinoma de células de Hürthle en la tiroides (31) y en los tumores pancreáticos neuroendocrinos (32). Las alteraciones se encuentran típicamente en las regiones que interactúan con ATRX y el dímero H3.3/H4 (17), lo cual se ha asociado a inestabilidad cromosómica y la presencia de telómeros de longitud

anormal (33). Además, la presencia de mutaciones en DAXX o ATRX correlacionan con la etapa del tumor, metástasis y una reducción de tasa de supervivencia en tumores pancreáticos neuroendocrinos (34–36). Lo anterior ha dado pie a que se considere a DAXX como un gen supresor de tumores (26).

En conjunto, estos resultados muestran la asociación de DAXX con la progresión del cáncer y su mal pronóstico. Esto ocurre principalmente por alteraciones en la expresión de esta proteína, lo cual promueve fenotipos patológicos in vitro y el crecimiento del tumor y la progresión de esta enfermedad in vivo. Se ha sugerido previamente que esto podría deberse a alguno de los procesos en los cuales participa DAXX, como la apoptosis, remodelación de la cromatina, o la regulación y reparación del DNA(26).

A pesar de la importancia de DAXX en el mantenimiento de la cromatina y la asociación de su alteración con la progresión del cáncer, la mayoría de los estudios realizados han sido descriptivos o de asociación con características clínicas relevantes. Centrándose en la observación de fenotipos que poseen las células humanas y murinas en ausencia de DAXX, poco se sabe de los genes particulares que son, directa o indirectamente, regulados por esta chaperona. Identificarlos podría significar un gran paso para conocer el o los mecanismos por los cuales los cambios en la concentración de esta proteína desatan la inestabilidad genómica en células humanas y un fenotipo más agresivo en el cáncer.

Dentro de los elementos que se encuentran en el genoma cuya expresión se asocia a la inducción de eventos de inestabilidad genómica y desregulación génica, se encuentran las secuencias repetidas (37–39), las cuales se encuentran mucho menos estudiadas que los genes codificantes (40). El mantenimiento de la represión de estos elementos es un blanco importante de DAXX, y la pérdida de la heterocromatina en estas regiones ha sido asociada con cáncer y la aparición de enfermedades humanas (38,39,41–44).

1.5 Los elementos repetidos y su expresión en cáncer

Los elementos repetidos son secuencias de DNA que componen alrededor de dos tercios del genoma humano (45). De manera general, se agrupan a grandes rasgos en dos

clases principales: repetidos en tándem, que comprenden satélites, microsatélites y minisatélites, y elementos transponibles (ETs, 46). Los ETs conforman alrededor del 45% del genoma humano y se clasifican en retrotransposones, o ET clase I, y transposones de DNA, o ET clase II. Los ET de clase I comprenden los elementos LTR (Long Terminal Repeat), HERVs (Human Endogenous RetroViruses), y no LTR, como lo son los LINEs (Long Interspersed Nuclear Elements) (46).

El estudio de estos elementos es importante porque se ha reportado que influencian la expresión de genes en *cis* y en *trans* modulando incluso redes regulatorias completas (39,41). Un ejemplo de esto ocurre en plantas y animales, donde los ETs pueden originar RNAs pequeños que regulan la expresión de un gran número de genes, modulando procesos esenciales como lo es el desarrollo (41). Esto sucede mediante diversos mecanismos; por ejemplo, elementos repetidos pueden estar embebidos en RNAs largos no codificantes o mensajeros, transcribirse y modular directamente la estabilidad del RNA, su procesamiento o su localización (42). Por otro lado, algunos ET pueden insertarse en diversas regiones del genoma, generando perturbaciones genéticas que desencadenan enfermedades humanas, como lo es el cáncer, desórdenes autoinmunes o neurofibromatosis tipo 1 (38,43,44).

En el contexto del cáncer, la expresión anómala de elementos repetitivos se ha asociado a varios tipos de tumores. Se ha reportado una sobreexpresión en elementos repetitivos centroméricos y pericentroméricos en cáncer testicular, de hígado, ovárico y de pulmón (47). También existe una sobreexpresión del satélite HSATII en cáncer de pulmón, hígado, ovario, próstata, osteosarcoma y en carcinoma pancreático ductal, siendo este último tipo bastante interesante al presentar un aumento en todos los transcritos satelitales (48–50). Por otra parte, distintos HERVs se encuentran asociados a melanoma (51), leucemia y linfoma (52), así como a tumores cancerosos de mama (53,54), testículo (53), ovario (54), próstata (55), riñón (54), útero (54) y a la carcinogénesis colorrectal (56). Finalmente, se ha encontrado una sobreexpresión de los elementos transponibles LINE-1 y algunos SINEs en tumores cancerosos pancreáticos y de próstata (37). Además, se han reportado inserciones *de novo* de esta familia de repetidos en cáncer colorrectal (57); adicionalmente, la sobreexpresión de

estos elementos ha sido propuesta como un evento crucial para que ocurran mutaciones oncogénicas en el carcinoma hepatocelular (58).

En resumen, la expresión de los elementos repetidos puede afectar seriamente la transcripción de genes codificantes y no codificantes y comprometer la estabilidad cromosómica de una célula (39,41,44). Consecuentemente, este fenómeno parece jugar un papel importante en el desarrollo del cáncer. Dado que DAXX se asocia principalmente con la represión de estos elementos, su desregulación podría ocasionar la expresión de los repetidos y desencadenar el fenotipo que se observa en células donde esta chaperona está alterada; este podría ser uno de los mecanismos por los cuales DAXX promueve el desarrollo del cáncer y la adquisición de características de mal pronóstico.

Una de las técnicas más potentes para estudiar el patrón de expresión de los elementos repetitivos en el genoma, así como la de los genes codificantes, es la generación de perfiles de expresión de RNA, mediante secuenciación masiva de cDNA. Este abordaje global permite la detección y cuantificación de la expresión de todos los transcritos presentes en una célula (59). Al acoplar esto con un diseño experimental adecuado, es posible comparar distintas condiciones y encontrar los elementos cuya expresión es diferencial. De esta manera, dicha metodología es capaz de contribuir de manera significativa al entendimiento del papel de DAXX en la regulación de la expresión génica en las células humanas.

1.6 Estudios transcriptómicos como herramienta para conocer los genes modulados por DAXX

El abordaje que generalmente se utiliza para conocer de manera global los genes y vías que son regulados por alguna proteína es al analizar datos de secuenciación del RNA (RNA-seq) en ausencia o disminuyendo la expresión del gen de interés. Esta técnica identifica la expresión de cada elemento del transcriptoma en un grupo de células en un momento dado. De manera general, el conjunto de RNA de una muestra es convertido en una librería de fragmentos de DNA con adaptadores de secuenciación en los extremos. Después de una amplificación, la librería se secuencia y se obtienen secuencias pequeñas de 30 a 400 pares de bases. De esta manera, se obtienen medidas precisas de la expresión génica de cada

transcrito. Con esta metodología, se pueden comparar distintas condiciones y encontrar elementos que cambian en cada una con el fin de entender los mecanismos que subyacen un fenotipo (59).

A pesar de la importancia de DAXX en la estabilidad genómica y su implicación en el aumento de la agresividad en cáncer, únicamente dos estudios se han llevado a cabo para conocer de manera global los genes y vías que son regulados por DAXX en células humanas. El primer trabajo fue en 2015, en un modelo celular de próstata (23), en el que se identificaron genes con cambio en su expresión tras la represión de DAXX. Los autores se centraron en la vía de la autofagia al encontrar que dicho tratamiento aumenta la expresión de reguladores positivos de este proceso, como DAPK1, DAPK3 y ATG8, además de disminuir la expresión de los reguladores negativos como mTOR y Raptor. Así mismo, se observó que esta chaperona se necesita para la unión de la DNMT1 con el genoma en una gran cantidad de casos. Esto sugiere que DAXX reprime la expresión de genes involucrados en la autofagia, entre otros, al reclutar a la DNMT1 a los promotores de estos genes (23).

El segundo estudio fue realizado en 2017, empleando una línea celular proveniente de glioblastoma (60). En él, se reporta que la represión de DAXX afecta la incorporación de H3.3 en la cromatina y la expresión de diversos genes. Entre ellos, se reprimen oncogenes involucrados en el crecimiento de tumores intracraneales como CCND1, MYC, FOS, SOX2 y OLIG2; además, diversos genes supresores de tumores como MAP2K4, KMT2C, EP300 y MLH1 se sobreexpresan. Finalmente, reportan a procesos biológicos involucrados en el desarrollo del sistema nervioso como los más afectados ante esta perturbación. Como resultado, la inhibición de DAXX mejora la supervivencia de ratones con glioblastoma y disminuye la progresión de esta enfermedad en un mecanismo independiente de ATRX (60).

Es interesante observar que, a pesar de que DAXX ha sido asociado exclusivamente con la represión de genes y elementos genómicos (61,62), ambos trabajos observaron una gran cantidad de genes que disminuyeron su expresión al reprimir a DAXX. Esto sugiere que esta chaperona podría también estar involucrada en la activación transcripcional, papel que se le había atribuido a HIRA únicamente (12).

Es sorprendente que aunque DAXX está sobreexpresado en muchos tipos de neoplasias, lo cual les confiere resistencia a tratamientos y agresividad, no existan más trabajos que indaguen en el mecanismo detrás de esto. Es importante recalcar que, a la fecha, no se sabe cuáles son los genes regulados por esta chaperona en otras líneas cancerosas en las cuales DAXX juega un papel relevante en el desarrollo de la enfermedad. Adicionalmente, no existen estudios comparativos que analicen si las vías involucradas en el desarrollo del fenotipo agresivo reportado son las mismas en distintos tejidos cancerosos.

Algo aún más inesperado es que no solo no existen múltiples estudios masivos en los que se analice el impacto de la alteración de DAXX en la expresión global de los genes codificantes en células humanas, sino que no existe ningún trabajo que aborde su impacto en elementos repetidos, siendo la represión de estos la principal función descrita de esta chaperona (3). Esto es consecuencia de que los estudios estándar de RNA-seq ignoran estas secuencias debido a su dificultad de análisis (40). No obstante, es muy importante conocer los elementos cuya expresión es modulada por DAXX , pues el estudio de secuencias repetitivas y genes codificantes modulados por DAXX podría revelar biomarcadores útiles para la detección, pronóstico y seguimiento de los cánceres mencionados en los que se presenta un fenotipo particular asociado a su sobreexpresión.

1.7 Herramientas para analizar datos de RNA-seq

Con los datos de RNA-seq, se puede obtener una gran información sobre el transcriptoma de origen. Las lecturas pueden ser analizadas para descubrir nuevos genes, cuantificar transcritos, hacer análisis de expresión diferencial, pruebas funcionales, identificar isoformas, entre otras cosas (63). Esta metodología es muy popular porque ofrece muchas ventajas con respecto a métodos con objetivos similares, como los microarreglos; el RNA-seq posee una mayor sensibilidad, por lo que es capaz de detectar transcritos poco abundantes con tasas bajas de falsos positivos. Además, dado que este método cuantifica la abundancia absoluta de cada transcrito, los datos producidos pueden re-analizarse de distintas formas y compararse con nuevas muestras. Finalmente, el costo de la secuenciación ha

disminuido considerablemente en los últimos años, haciendo esta tecnología mucho más accesible y usual en experimentos biológicos (64) .

Actualmente existe una gran gama de herramientas bioinformáticas para analizar datos de RNA-seq. Esto se debe a que no hay un pipeline (i.e. flujo de trabajo bioinformático) que funcione bien en todas las condiciones; el uso de cada programa dependerá de las preguntas a resolver en cada proyecto. Uno de los análisis más utilizados y estandarizados es el de la cuantificación y expresión diferencial de genes codificantes (63) (**Figura 1.7.1**). Dado que las lecturas de sus transcritos mapean a regiones únicas del genoma, identificar el gen del cual provienen no presenta muchas complicaciones. Esto se vuelve menos complejo en especies típicamente estudiadas, donde se cuenta con un genoma y un transcriptoma de referencia, como lo es el humano (63).

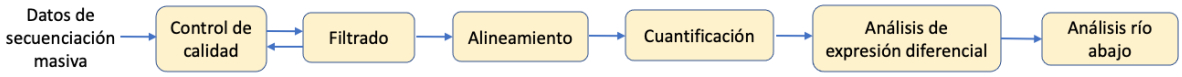


Figura 1.7.1. Etapas estándar del análisis de RNA-seq. De manera general, el análisis en este contexto involucra pruebas de control de calidad de las secuencias, filtrado de bases de mala calidad y adaptadores, alineamiento de las lecturas a un genoma o transcriptoma, cuantificación de transcritos, la prueba de expresión diferencial y finalmente análisis río abajo, como es el análisis de enriquecimiento de Gene Ontology (GO) o la búsqueda de redes de interacción proteína proteína de los genes diferencialmente expresados

Por otro lado, aunque los pasos del pipeline son los mismos, el análisis de elementos repetitivos presenta retos mucho mayores al de genes codificantes; debido a esto, la mayoría de los estudios los ignoran, enfocándose en el 22-51% del genoma que corresponde a regiones no repetitivas (40). La razón de esto se origina en las complicaciones que representa el uso de datos transcriptómicos de secuenciación corta.

El principal problema se encuentra en el mapeo de las lecturas secuenciadas a su región de origen en el genoma o transcriptoma. Las secuencias repetitivas suelen encontrarse en distintos loci, por lo que la única forma certera de identificar la proveniencia de los transcritos

es evaluando la estructura de los repetidos, que puede llegar a ser distinta en diferentes regiones, o contar con fragmentos de las regiones adyacentes. Dado que los fragmentos de secuenciación son más pequeños que la estructura repetitiva y difícilmente capturan las secuencias adyacentes, las lecturas se alinean a múltiples lugares en el genoma; de esta manera, es muy difícil saber de dónde se originó el transcrito al que pertenecen (65). A pesar de esta limitación, distintas herramientas computacionales y estadísticas se han desarrollado para evaluar correctamente la expresión de estas regiones repetitivas en el genoma utilizando lecturas cortas.

De manera general, existen tres abordajes utilizados para asociar lecturas de RNA-seq a regiones del genoma o transcriptoma e identificar su proveniencia. El primero consiste en quedarse solamente con las secuencias que se alineen a regiones únicas en el genoma; este enfoque estricto tiende a descartar una gran cantidad de lecturas y es el menos recomendado para este tipo de análisis. En el segundo abordaje, se alinean las lecturas a secuencias consenso de las familias de elementos repetitivos albergadas en bases de datos. Aunque este enfoque no te indica las regiones físicas de donde proviene un transcrito, es posible identificar clados de elementos de jerarquía relativamente baja si existen secuencias consenso muy definidas. Finalmente, en el tercero, se permite que las lecturas se alineen a múltiples regiones del genoma. Posteriormente, estas lecturas son distribuidas al azar entre las mejores regiones candidatas, o mediante algún algoritmo que, con base en métodos estadísticos, la asigne al locus más probable de origen (40). En resumen, con estos enfoques es posible identificar las familias de elementos repetidos que se expresan en un transcriptoma.

Como ejemplo del segundo abordaje mencionado en el párrafo anterior, se encuentra TETranscripts (66). Esta herramienta bioinformática analiza genes codificantes y elementos repetitivos al mismo tiempo. Después de utilizar un alineador que permite mapear lecturas a múltiples regiones como STAR (67), TETranscripts usa un algoritmo EM (expectation maximization) para encontrar la distribución más probable de las lecturas mapeadas a sitios múltiples, mientras que procede de manera estándar con las que se alinean a sitios únicos. Otro ejemplo que mezcla el abordaje dos y tres es SalmonTE (68). La novedad de esta herramienta es que utiliza el método de quasi-alineamiento implementado en Salmon (69);

este algoritmo cuantifica la abundancia relativa de cada transcrito a partir de datos de secuenciación y un archivo con las secuencias de los elementos repetitivos a cuantificar. Finalmente, esta herramienta hace uso de distintos algoritmos como inferencia bayesiana y EM para distribuir las lecturas y asignarlas a los repetidos de origen más probables.

Estas herramientas bioinformáticas previamente descritas son capaces de identificar genes regulados en común por DAXX en distintas líneas celulares cancerosas en donde su desregulación ocasiona fenotipos agresivos, resistencia a tratamientos e inestabilidad genómica. Por consiguiente, en el presente trabajo, analizamos el transcriptoma de células con una disminución de DAXX mediante un *knockdown* por siRNAs en el modelo celular de HCT116, una línea celular genomicamente estable proveniente de cáncer de colon. Este modelo nos permite conocer el efecto de DAXX en la inducción de inestabilidad genómica y la regulación génica de una manera más aproximada al comportamiento en tejidos no neoplásicos. Cabe señalar que en no existen trabajos sobre el impacto a nivel transcriptómico de la alteración de esta chaperona esta línea celular. Adicionalmente, re-analizamos los datos de secuenciación publicados en artículos previos de cáncer de próstata (23) y glioblastoma (60) con el objetivo de brindar un estudio comparativo e identificar los genes codificantes cuya expresión se altera en cualquiera de estos linaje celulares. Finalmente, analizamos todos los datos mencionados con herramientas desarrolladas para identificar la expresión de elementos repetitivos en el genoma; conforme a nuestro conocimiento, no existen estudios parecidos en el campo, por lo que conocer las secuencias repetidas que se desregulan representará un trabajo innovador que podría revelar patrones no vistos previamente sobre el impacto genómico que ocasiona la desregulación de la proteína DAXX. Para realizar esto, utilizamos herramientas bioinformáticas con diferentes algoritmos como lo son Salmon, SalmonTE y TETranscripts para obtener los genes y elementos repetidos diferencialmente expresados en cada una de las líneas celulares.

Los resultados de este proyecto podrían ser de ayuda para identificar firmas transcriptómicas y mecanismos por los cuales potencialmente la alteración de DAXX desencadena eventos peligrosos para la integridad celular. De esta manera, podremos comprender mejor el alcance de esta chaperona en la regulación de la expresión génica y su importancia en el

mantenimiento de la estructura genómica. Así mismo, esto puede ayudarnos a entender mejor el panorama celular en tumores malignos de pacientes con una desregulación de esta proteína. A futuro, esta información será de gran ayuda para realizar tratamientos personalizados para estos tipos de cánceres, mejorando la esperanza de vida de los pacientes.

2. Hipótesis

La proteína DAXX actuará principalmente como represor génico tanto a nivel de elementos repetidos como de genes codificantes.

3. Objetivo general y particulares

El objetivo general de este trabajo es identificar los genes y elementos repetidos regulados por DAXX en líneas celulares humanas mediante secuenciación de alto rendimiento del RNA (RNA-seq). Los objetivos particulares derivados de este son los siguientes:

1. Identificar los genes codificantes diferencialmente expresados de tres líneas celulares bajo un modelo de DAXX knock-down (*KD*), así como los principales procesos biológicos asociados a ellos.
2. Determinar los genes codificantes diferencialmente expresados compartidos en las tres líneas celulares DAXX *KD*.
3. Identificar cambios en la expresión de elementos repetidos en los tres modelos celulares DAXX *KD.*
4. Determinar los elementos génicos repetitivos diferencialmente expresados en común en las tres líneas celulares cancerosas DAXX *KD*.

4. Metodología

Modelos celulares

Las células HCT116 fueron crecidas en un medio McCoy's 5a. Posteriormente, fueron transfectadas con una mezcla de dos RNAs pequeños de interferencia (siRNAs; Dharmacon, cat. LQ-004420-00-0010) para lograr la subexpresión de DAXX en un medio DharmaFECT 2 (Dharmacon, cat. T-2002-03) en un ensayo de 72 horas con siRNA, como se especifica en Torres-Arciga 2019 (19). Después del tiempo de transfección, las células fueron tratadas con tripsina y divididas para extraer las proteínas y el RNA. Como control, la fracción proteica se utilizó para probar la subexpresión de DAXX a través de un inmunoensayo tipo western blot, demostrado en Torres-Arciga 2019 (19). Finalmente, la integridad del RNA fue comprobada utilizando el TapeStation, mientras que su cuantificación se realizó utilizando el fluorómetro Qubit.

Los datos de las líneas celulares PC3 y GSC23 se obtuvieron de artículos previamente publicados. La línea PC3 fue crecida en el medio RPMI-1640 e infectada con lentivirus recombinantes que contenían RNAs de horquilla pequeña (shRNAs) contra DAXX, como se especifica en Puto *et al.* 2015 (23). Finalmente, las células GSC23 fueron crecidas en un medio DMEM e infectadas con lentivirus recombinantes que contenían shRNAs contra DAXX, como se indica en Benitez *et al.* (60).

Secuenciación de RNA

Las librerías de RNA de HCT116 se generaron con el kit de Illumina Truseq Stranded Total RNA; después se realizó una secuenciación tipo paired-end por triplicado técnico en las muestras DAXX *KD* y en los controles, tras la eliminación del RNA ribosomal; los datos crudos tuvieron una profundidad de secuenciación promedio de 36.7 millones de secuencias.

Las librerías de RNA de PC3 se realizaron con el kit Illumina TruSeq RNA v2; después se realizó una secuenciación single-end con una selección de RNAs con colas de poli A en las muestras tratadas con el shRNA y en los controles por duplicado biológico, como se indica en Puto *et al.* 2015 (23); los datos crudos tuvieron una profundidad de secuenciación promedio de 41.7 millones de secuencias.

Finalmente, las librerías de RNA de GSC23 se generaron utilizando el kit de Illumina TruSeq Stranded mRNA Sample Prep Kit; posteriormente, se realizó una secuenciación single-end con una selección de RNAs con colas de poli A en las muestras tratadas con el shRNA y en los controles por triplicado biológico, como se especifica en Benitez *et al.* (60); los datos crudos tuvieron una profundidad de secuenciación promedio de 32.3 millones de secuencias. Las muestras de los tres estudios fueron secuenciadas en el secuenciador Illumina HiSeq 2500.

Análisis bioinformático

A diferencia de PC3 y GSC23, HCT116 solamente posee réplicas técnicas y carece de biológicas, razón por la cual las conclusiones estadísticas obtenidas a partir de los datos de

dicha muestra describen más la variabilidad técnica que la biológica. Aún así, se realizó el análisis bioinformático al igual que en las otras líneas celulares, tomando las réplicas técnicas como si fueran biológicas, con el objetivo de comparar sus tendencias y explorar su comportamiento ante la represión de DAXX.

La calidad de los datos de secuenciación fue evaluada con la versión 0.11.8 de FastQC (70). Posteriormente, las secuencias fueron filtradas con la versión 0.38 de Trimmomatic (71) con el objetivo de remover los adaptadores de Illumina TruSeq 3 y las bases de baja calidad. Los parámetros utilizados fueron ILLUMINACLIP: 2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. Los archivos resultantes fueron utilizados para obtener una matriz de conteos mediante tres herramientas: TETranscripts (66), Salmon (69) y SalmonTE (68).

Para el pipeline de TETranscripts, los archivos de secuenciación fueron alineados a la versión 31 del ensamble primario del genoma humano obtenido de Gencode (GRCh38.p12) utilizando la versión 2.7.3 de STAR (67). Los parámetros utilizados para esa etapa fueron los sugeridos en el artículo original de TETranscripts (--outFilterMultimapNmax 100 --winAnchorMultimapNmax 100). Después de esto, se utilizó la versión 2.1.4 de TETranscripts (66) en modo -multi . Independientemente, se corrió la versión 0.13.1 de Salmon (69) para estimar la abundancia de los transcritos, mapeando las secuencias a la versión 29 del transcriptoma humano obtenida de Gencode. A la par, se utilizó la versión 0.4 de SalmonTE (68) para obtener la expresión de elementos transponibles; las secuencias fueron mapeadas al archivo humano de referencia provisto por los desarrolladores, como se indica en (68). Al final de cada uno de los tres pipelines mencionados, se obtuvo una matriz de conteos.

La matriz se utilizó para hacer un análisis de expresión diferencial con la versión 1.22.2 de DESeq2 (72). Finalmente, para los genes codificantes diferencialmente expresados (DE), se hizo un análisis de enriquecimiento de términos de Gene Ontology utilizando el paquete de R ClusterProfiler (73). El flujo de trabajo general del análisis bioinformático puede observarse en la **figura 4.1**.

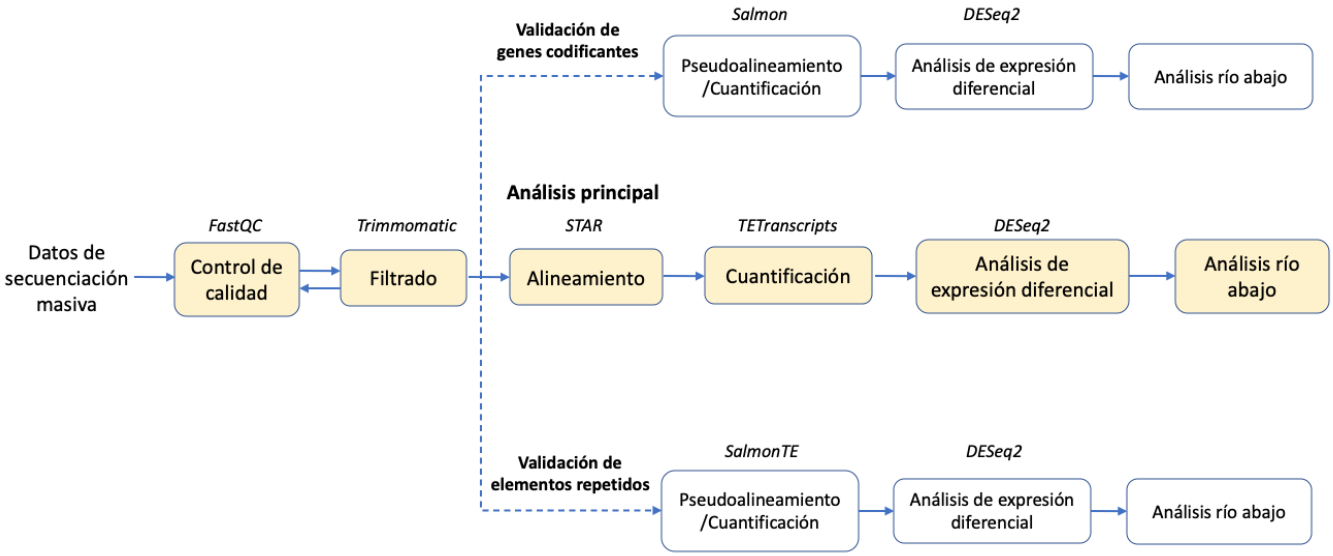


Figura 4.1. Flujo de trabajo en el análisis de RNA-seq. En el centro, en color beige, se pueden observar los pasos del pipeline principal, que consiste en el uso de STAR y TETranscripts. En ambos lados se encuentran los análisis realizados para validar los resultados obtenidos utilizando Salmon y SalmonTE.

Disponibilidad de datos

Los datos de RNA-seq de PC3 y GSC23 se encuentran disponibles públicamente bajo el código de acceso PRJNA283304 y PRJNA345372, respectivamente, en el Sequence Read Archive de NCBI. Los datos de HCT116 se encuentran sin publicar.

5. Resultados

Con el objetivo de identificar los genes codificantes y elementos repetidos regulados por DAXX en células humanas, se utilizaron tres líneas inmortales procedentes de cáncer colon (HCT116), próstata (PC3) y glía (GSC23); cada una de ellas fue sometida a un tratamiento para reprimir la expresión de DAXX utilizando RNAs de interferencia (iRNAs). Tras la eliminación de las secuencias de baja calidad y los adaptadores de los archivos de secuenciación y la obtención de la matriz de conteos con TETranscripts (66), los datos de expresión normalizados con el método Variance Stabilizing Transformation de DESeq2 de cada célula se separaron de acuerdo a lo que esperábamos utilizando un análisis de componentes principales (**figura 5.0.1**); es decir, a excepción de una muestra tratada con siRNA en HCT116, en el componente principal 1, el cual alberga la mayor cantidad de varianza, las muestras control se separan de las tratadas. Esto indica que, dentro de cada conjunto de datos, las condiciones experimentales son la principal fuente de variación entre las muestras y que las réplicas biológicas (o técnicas, en caso de HCT116) son reproducibles.

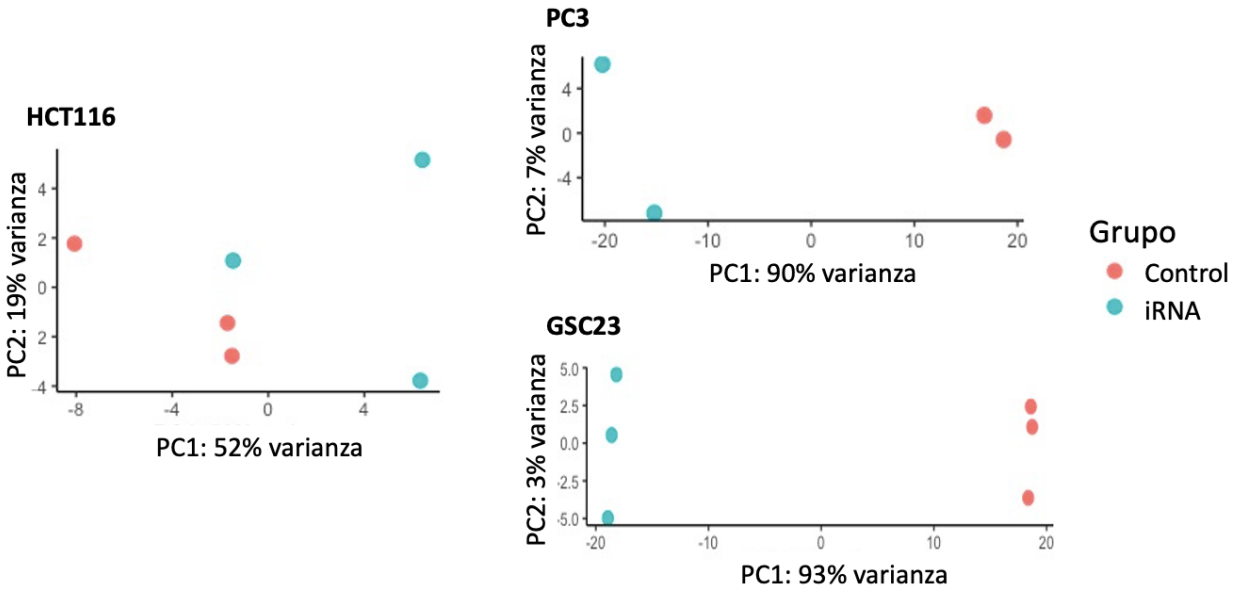


Figura 5.0.1 Análisis de componentes principales. Los paneles corresponden a los conjuntos experimentales de HCT116, PC3 y GSC23.

Posteriormente, se realizó el análisis de expresión diferencial, comparando cada modelo tratado con sus respectivos controles, utilizando DESeq2 (72). Con el objetivo de observar el cambio en los patrones de expresión de las células, se procedió a explorar los resultados en los genes codificantes y en los elementos repetidos de manera separada.

5.1 Análisis de expresión diferencial en genes codificantes

Tras realizar el análisis de expresión diferencial, se encontraron 277, 4704 y 8422 genes codificantes diferencialmente expresados (DEs) en las líneas celulares HCT116, PC3 y GSC23, respectivamente ($p\ adj. < 0.05$). A diferencia de lo esperado por el papel principalmente descrito de represor de DAXX, no se encontró una tendencia clara de los genes a sobreexpresarse en todos los conjuntos de datos (**Figura 5.1.1A**). La línea celular que más clara tiene esta tendencia es HCT116, con un 86% de genes diferencialmente expresados al alza, seguido de PC3 con un 54% y finalmente GSC23 con un 52% (**Figura 5.1.1B**).

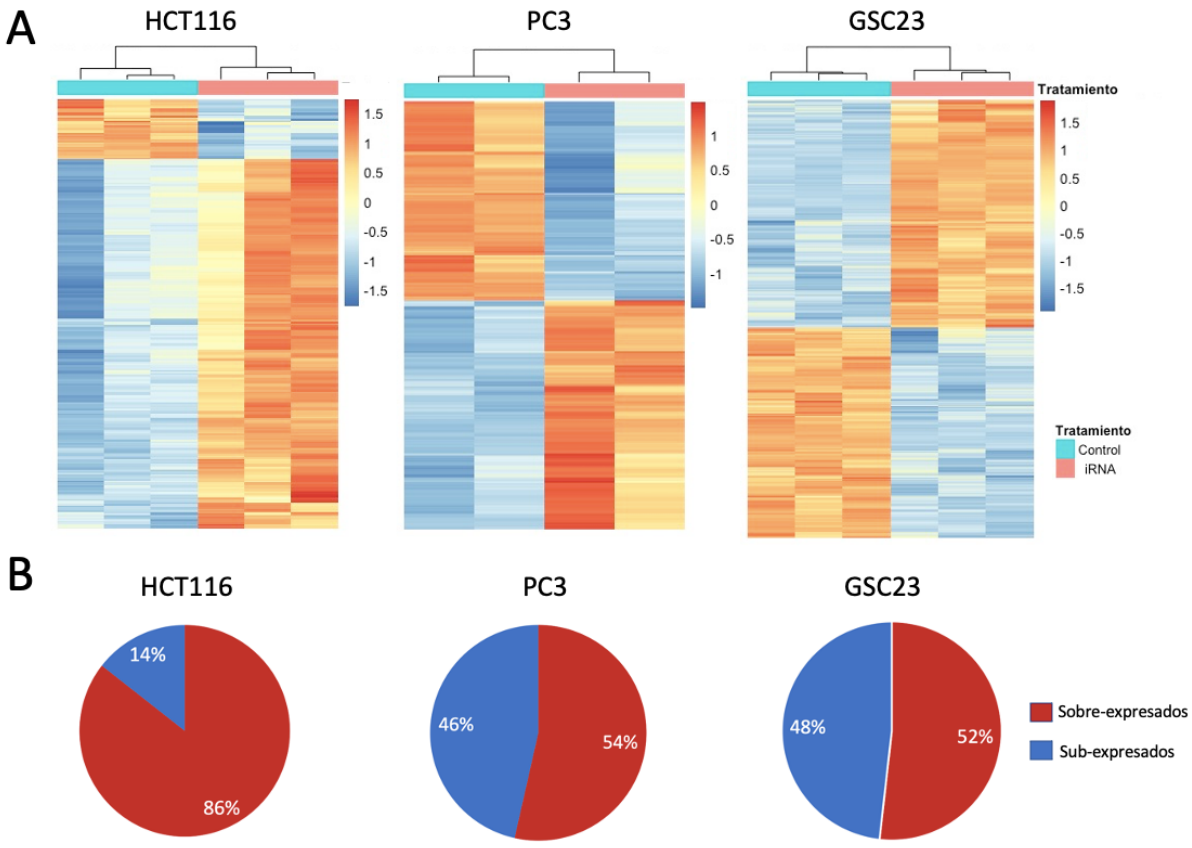


Figura 5.1.1. Genes diferencialmente expresados en las tres líneas celulares. A) Heatmap de los genes diferencialmente expresados (*p adj*. < 0.05) en HCT116, PC3 y GSC23. B) Proporción de genes que están sobre y subexpresados en cada línea celular.

Así mismo, se observó que las consecuencias de la represión de DAXX fueron diferentes para cada línea celular (**Figura 5.1.2A**), siendo HCT116 la que presentó una menor cantidad de genes diferencialmente expresados (277), seguido de PC4 (4704) y GSC23 (8422). Dado que la estrategia que se utilizó para reprimir este gen fue con iRNAs, el efecto del tratamiento es gradual y depende de la eficacia de los oligonucleótidos para silenciar el RNA mensajero de dicho gen. Lo anterior podría explicar la diferencia observada entre las réplicas técnicas, así como la de la cantidad de genes desregulados, pues diferentes niveles de represión podrían generar un cambio de expresión en distintos genes dependiendo del nivel de sensibilidad de los mismos a DAXX. Es decir, entre más fuerte la represión, más grande el efecto. Para explorar lo anterior, en ausencia de datos cuantitativos del cambio de DAXX a nivel proteína después de su represión en PC3 y GSC23, evaluamos su tasa de cambio de expresión en el RNA-seq en los tres conjuntos de datos. Como era de esperarse, DAXX presentó un cambio más pequeño en HCT116 (L_2FC = -0.67), seguido de PC3 (L_2FC = -1.265) y GSC23 (L_2FC = -1.613, **Figura 5.1.2B**). Independientemente de la magnitud, esta fue significativa en todos los casos (*p adj.* < 0.05), lo cual confirma que el cambio de expresión que observamos se debe a la subexpresión de la chaperona. En conclusión, los datos anteriores demuestran que DAXX no juega un papel únicamente represor en la regulación génica, que el efecto de su subexpresión fue distinta en cada conjunto experimental, y que esto podría deberse a la diferente eficiencia de los iRNAs utilizados en cada línea celular.

Después de identificar los genes diferencialmente expresados en cada situación, nos preguntamos cuáles eran los principales procesos biológicos afectados en las distintas líneas celulares. Para resolver esto, se procedió a hacer un análisis de enriquecimiento de términos de Gene Ontology (GO). Este análisis encuentra las categorías que se encuentran sobre-representadas en un conjunto dado de genes diferencialmente expresados mediante una prueba estadística, por lo que es una prueba útil para identificar las principales vías biológicas alteradas en un experimento (73).

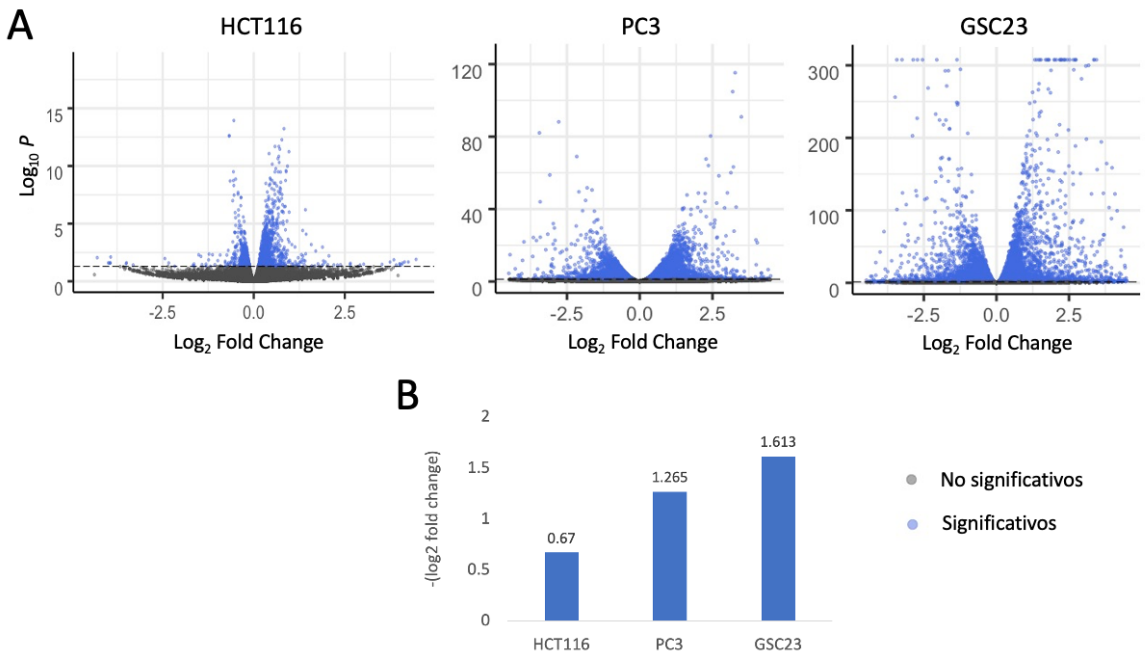


Figura 5.1.2. Efecto del iRNA contra DAXX en las líneas celulares. A) Variación entre el número de genes diferencialmente expresados, y su tasa de cambio entre HCT116, PC3 y GSC23. Se observa un volcano plot por experimento donde los genes diferencialmente expresados se marcan en azul y los no significativos en gris. B) Niveles de disminución de DAXX en cada una de las líneas celulares.

En el caso de HCT116, las categorías de GO enriquecidas correspondieron principalmente a una alteración en vías de regulación génica mediante las modificaciones de la cromatina y de los nucleosomas (**Figura 5.1.3**), procesos relacionados a las funciones que desempeñan enzimas involucradas en mecanismos epigenéticos. Por otro lado, en PC3 las categorías más significativas en procesos biológicos estuvieron relacionados a la síntesis de proteínas, así como a su direccionamiento y exportación (**Figura 5.1.3**). Estos procesos están relacionados a las actividades glandulares que desempeña este tejido de manera normal, pues la principal función de la próstata es producir sustancias esenciales para el semen, por lo que la producción de proteínas en el RE, su movilización y excreción son fenómenos específicamente activos en este tejido. Finalmente, las categorías enriquecidas en GSC23 corresponden principalmente a procesos cruciales en las células nerviosas como la morfogénesis, el aprendizaje y la memoria (**Figura 5.1.3**). Estos procesos se encuentran activos en células nerviosas de manera particular, por lo que corresponden a mecanismos propios de tejido nervioso.

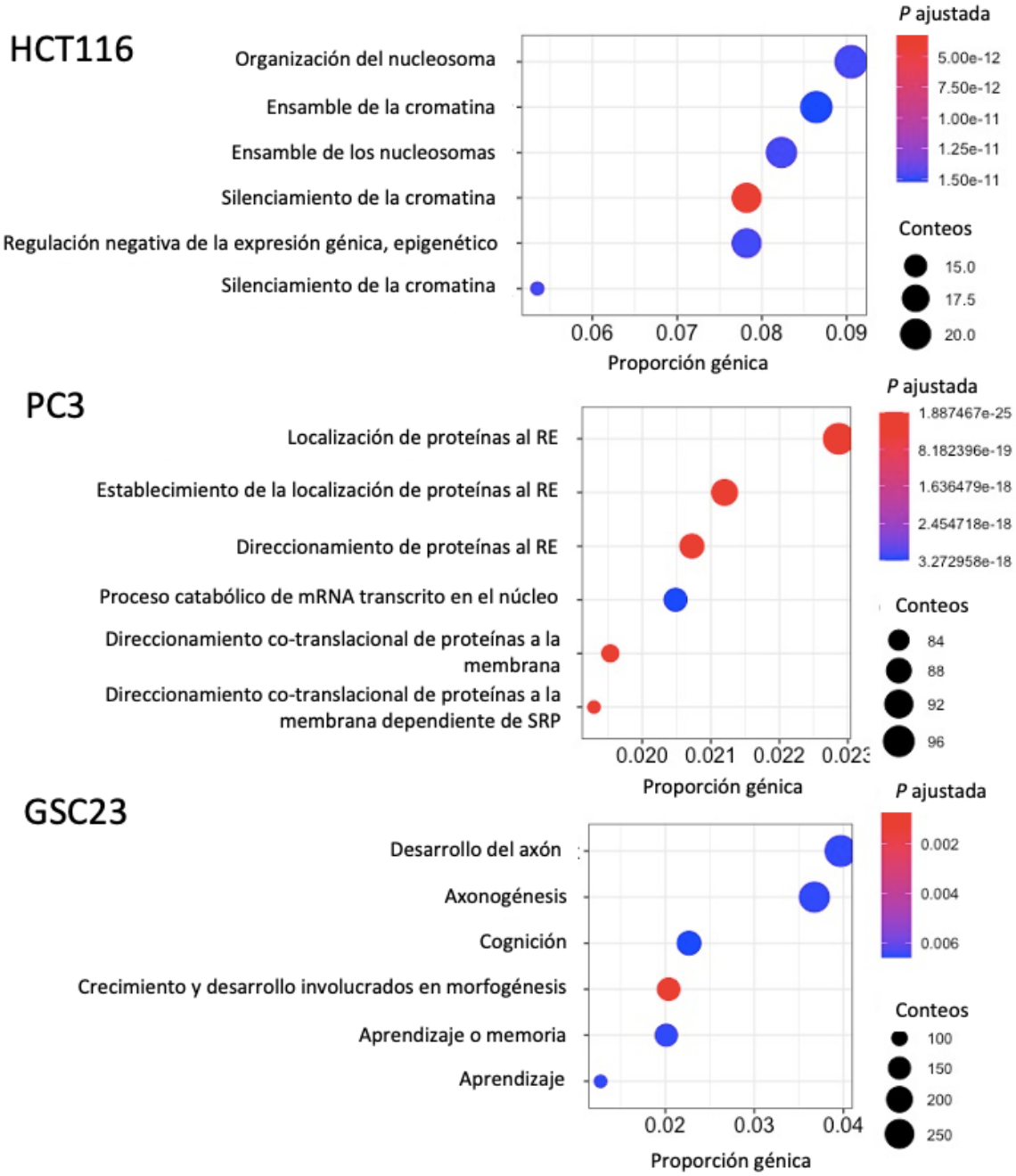


Figura 5.1.3. Primeros 6 términos de procesos biológicos de GO enriquecidos en los genes diferencialmente expresados por experimento. El gradiente de color indica el valor de la p ajustada y el tamaño del círculo la cantidad de genes diferencialmente expresados que pertenecen a dicha categoría. El eje de las x indica la proporción que hay entre los genes DE de cada categoría con respecto al total de genes DE encontrados en ese experimento.

En conclusión, se observó que los genes desregulados en PC3 y GSC23 se encuentran asociados principalmente a procesos biológicos tejido-específico llevados a cabo de manera normal en los tejidos a los que pertenecen estas células. Le atribuimos el hecho de que no hayamos encontrado un proceso biológico específico de colon en HCT116 al menor efecto de represión que observamos en esta línea celular en comparación de las otras dos (**Figura 5.1.2**), sugiriendo que los procesos observados ahí sean los primeros afectados tras la alteración de DAXX. Sin embargo, no se puede descartar la opción de que las diferencias de los genes desregulados en las líneas estudiadas se deban al efecto combinado de la desregulación promovida por el *KD* de *DAXX* y el contexto genómico específico en el que se encontraban previamente dichas líneas celulares antes del *KD*. Si bien la línea celular HCT116 es estable y casi diploide (74), PC3 es casi triploide (75), y desconocemos el cariotipo de GSC23. Sin embargo, se requerirían más estudios para confirmar esta hipótesis.

Tras obtener estos resultados, nos preguntamos si existían genes regulados de forma ubicua por DAXX en células humanas. Esto podría explicar su efecto negativo en distintos tipos de cáncer, proponiendo así un mecanismo basal conservado. Para explorar esta hipótesis, buscamos la intersección de genes diferencialmente expresados en todas las líneas celulares DAXX *KD*. Se encontraron 72 genes codificantes diferencialmente expresados en todos los modelos estudiados (**Figura 5.1.4A**). No obstante, su dirección de cambio no era la misma en todas las ocasiones; es decir, algunos genes presentaban tendencias de alteración distintas en los distintos modelos celulares, como, por ejemplo, *WNT9A* o *SOX9*, los cuales están subexpresados en PC3 y GSC23, pero sobreexpresados en HCT116 (**Figura 5.1.4B**).

Posteriormente, decidimos concentrarnos únicamente en los genes cuya alteración de la expresión ante el *KD* de DAXX tuviera la misma dirección de cambio con el objetivo de identificar los que se asociaran más fuertemente a la represión de esta chaperona; bajo este criterio, encontramos 12 genes, de los cuales 11 estaban sobreexpresados (*ACTG1, ALDOA, AC093512.2, EIF3F, CD81, DAP, PLAU, MGAT4B, MT2A, IFITM3* y *S100A11*). Como era de esperarse, *DAXX* mantiene un patrón de expresión a la baja en las tres líneas celulares, lo cual valida el tratamiento experimental (**Figura 5.1.4C**). Con los datos anteriores, podemos observar que, a excepción de *DAXX*, todos los genes con una dirección de cambio constante

en la intersección de las tres líneas celulares aumentan su expresión, lo cual es interesante y soporta nuestra hipótesis inicial en la que el papel principal que juega DAXX es de represor.

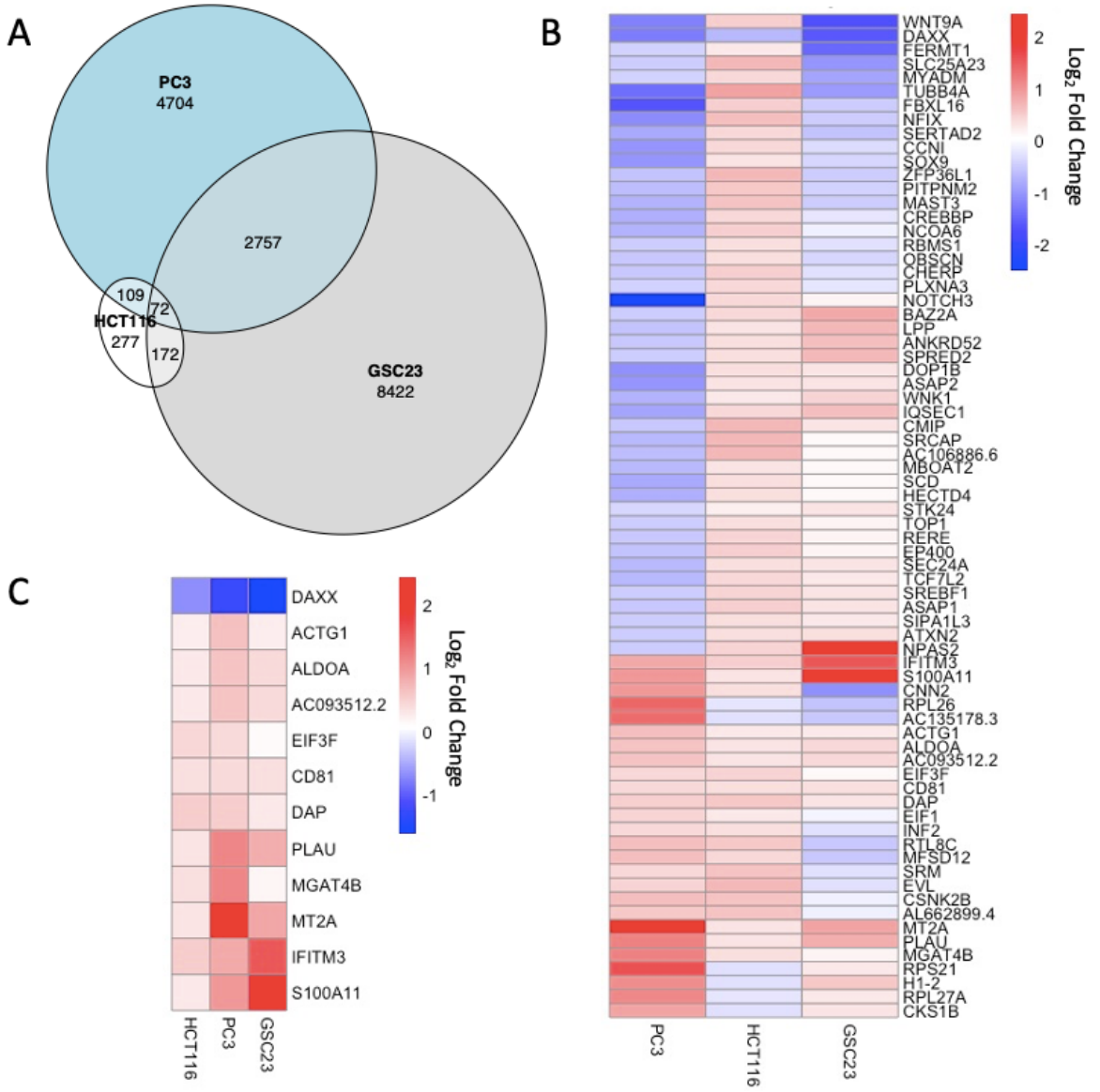


Figura 5.1.4 Genes diferencialmente expresados en común entre HCT116, PC3 y GSC23. A) Diagrama de Venn indicando las intersecciones entre cada uno de los modelos celulares. B) Tasa de cambio de los 62 genes ubicados en la intersección de los tres modelos. C) Tasa de cambio de los genes en la intersección de los tres modelos con una dirección de cambio constante.

Con el objetivo de validar los resultados obtenidos utilizando un pipeline bioinformático distinto, utilizamos el programa de cuantificación y alineamiento rápido Salmon (69). Tras realizar el análisis de expresión diferencial, encontramos el mismo patrón que en TETranscripts: no existe una tendencia clara en las tres líneas celulares a presentar mayoritariamente genes sobreexpresados. La diferencia en genes diferencialmente expresados se mantuvo: HCT fue el que menor cantidad tuvo (311), seguido de PC3 (3291) y GSC23 (6779). De igual manera, el modelo celular con mayor porcentaje de genes sobreexpresados fue HCT116 seguido de PC3 y GSC23 (**Figura 5.1.5A**).

Al comparar la tasa de cambio de los genes diferencialmente expresados en común bajo ambas metodologías, se encontró una correlación fuerte y significativa en HCT ($S = 5.9 \times 10^4$, $p < 0.01$), PC3 ($S = 1.5 \times 10^8$, $p < 0.01$) y GSC23 ($S = 9.6 \times 10^8$, $p < 0.01$; **Figura 5.1.5B**), con un coeficiente de correlación mayor a 0.94 en todos los casos. Adicionalmente, los términos enriquecidos de GO en los genes diferencialmente expresados (p ajustada < 0.05) estuvieron involucrados en los mismos procesos biológicos que en TETranscripts (fenómenos tejido-específico en PC3 y GSC23, y de regulación génica en HCT116). Los datos anteriores muestran que los resultados en el análisis de expresión diferencial de los genes codificantes obtenidos mediante el pipeline de TETranscripts son robustos y reproducibles utilizando un software que utiliza métodos de mapeo y cuantificación completamente distintos.

Posteriormente, nos preguntamos si los genes compartidos en las tres líneas celulares con la misma dirección de cambio (**Figura 5.1.4C**) seguían siendo significativos con el pipeline de Salmon. De los 12 genes previamente identificados con TETranscripts, 8 estuvieron diferencialmente expresados también, cuya tasa de cambio bajo ambos métodos fue casi idéntica (**Figura 5.1.6A**).

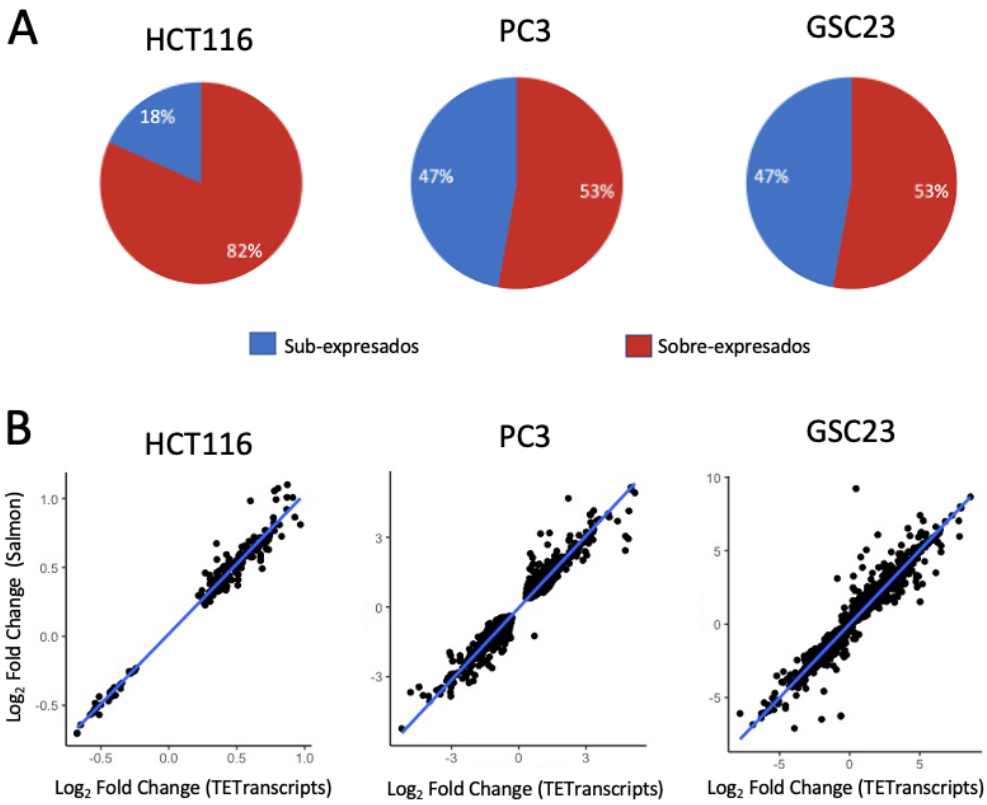


Figura 5.1.5. Validación de TETranscripts utilizando Salmon. A) Genes diferencialmente expresados utilizando la matriz de conteos de Salmon en HCT116, PC3 y GSC23. B) Correlación entre la tasa de cambio de HCT116, PC3 y GSC23 obtenida con el pipeline de TETranscripts y Salmon.

Descartando a DAXX, cuya subexpresión era de esperarse, los siete genes restantes (*CD81, ALDOA, PLAU, MGAT4B, MT2A, IFITM3* y *S100A11*) podrían ser genes regulados por DAXX en cualquier línea celular humana. Con el objetivo de explorar si alguna cascada biológica se veía particularmente afectada de manera basal, evaluamos la interacción proteína-proteína de los productos funcionales de dichos genes utilizando STRING, una base de datos de contactos físicos y asociaciones funcionales entre proteínas (76). No se encontró ninguna interacción significativa, lo cual descarta su participación en cascadas biológicas en común (**Figura 5.1.6A**).

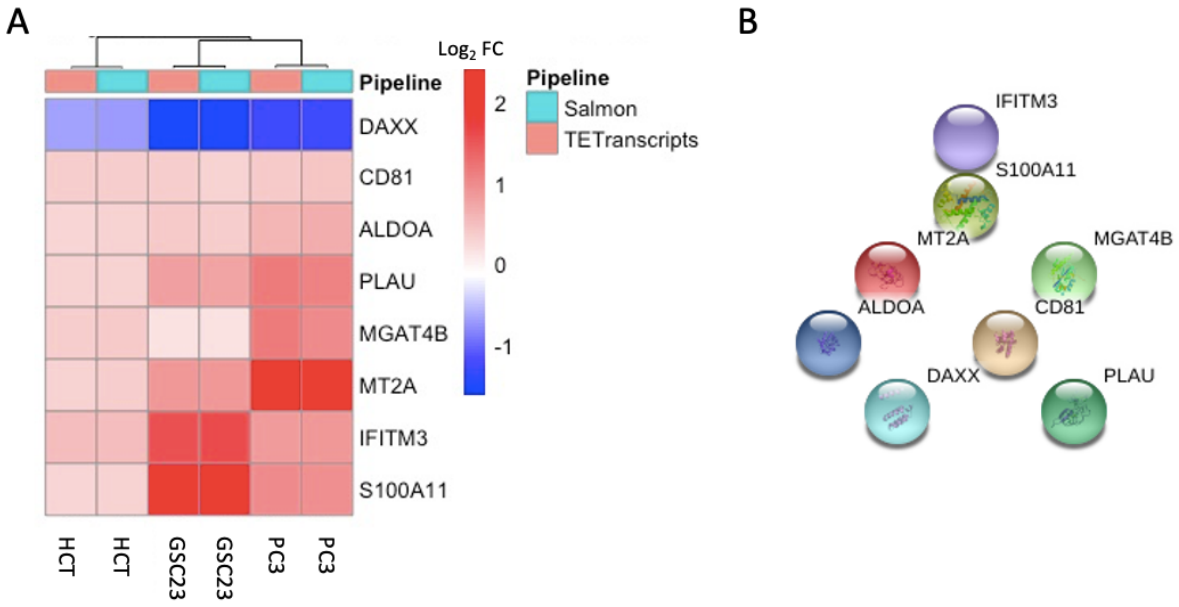


Figura 5.1.6. Genes diferencialmente expresados con una dirección de cambio común en las tres líneas celulares DAXX *KD* en ambos pipelines. A) Tasa de cambio de los 8 genes identificados. B) Resultados de interacción utilizando STRING.

Finalmente, analizamos el efecto de la desregulación de los siete genes sobreexpresados en las líneas celulares con el objetivo de explorar su papel en el fenotipo que se genera tras la alteración en la expresión de DAXX (**Tabla 5.1**). Observamos que todos ellos desempeñan papeles relevantes en cáncer y, de manera particularmente interesante, están asociados con características típicamente observadas en cánceres donde DAXX está desregulado a la alza o a la baja, como lo es el aumento en proliferación y supervivencia (26-38).

Tabla 5.1. Papel de los genes codificantes comunes regulados por DAXX en cáncer. Las descripciones fueron obtenidas de la base de datos GeneCards (77).

Gen / Proteína / GeneCards ID	**Descripción de la proteína**	**Asociación con cáncer**
***ALDOA* /** ALDOA / GC16P030064	Proteína perteneciente a la familia de fructosa-bifosfato aldolasas de clase I. Cataliza la conversión reversible de la	Proteína clave en la reprogramación metabólica del cáncer y la metástasis. Su sobreexpresión se asocia con un incremento en la

	fructosa-1,6-bifosfato a gliceraldehido-3-fosfato y dihidroxiacetona-fosfato. Muy importante en la glicólisis y el mantenimiento homeostático de la glucosa.	migración, invasión y supervivencia de células de cáncer colorrectal (78), pulmonar (79), pancreático (80), entre otros (78). Su alteración en células cancerosas induce apoptosis (81).
CD81 / CD81 / GC11P002377	Glicoproteína de superficie de membrana perteneciente a la familia tetraspanina. Involucrada en la transducción de señales.	En cáncer gástrico, su supresión mediante metilación se asocia con la estimulación de la proliferación celular, el crecimiento y supervivencia (82). En cáncer de próstata y líneas celulares de mama y osteosarcoma, su sobreexpresión se ha relacionado con mayor proliferación, invasión y migración (83–85).
IFITM3 / IFM3 / GC11M000319	Proteína membranal inducida por interferón. Involucrada en procesos de homeostasis intracelular de colesterol, adhesión celular, regulación de células inmunes, desarrollo de células germinales, mineralización, entre otros.	Sobreexpresado en muchos tipos de cánceres como colon, gástrico, astrocitoma, glioma, mama, carcinoma escamoso esofágico. Su silenciamiento inhibe el crecimiento, proliferación y la metástasis en cáncer de colon, próstata, gástrico, carcinoma hepatocelular y glioma (86,87).
MGAT4B / MGAT4B / GC05M179797	Glucosiltransferasa que participa en la transferencia de N-acetilglucosamina (GlcNAc) al núcleo de residuos de manosa de glicanos unidos a nitrógeno. Además, regula la formación de	La presencia elevada de ramificación β1-6 en los N-glicanos de la superficie celular de tumores correlaciona con una regulación positiva de las enzimas ramificadoras de N-glicano

	estructuras ramificadas triantenarias y multiantenarias en cadenas de azúcares del aparato de Golgi.	(MGATs); esto se correlaciona positivamente con el grado histológico y la metástasis en ganglios tumorales (88).
MTA2 / MTA2 / GC11M063413	Miembro de familia de reguladores transcripcionales asociados con metástasis de tumores. Componente central en el complejo de remodelación de nucleosomas y desacetilación de histonas NuRD. Regula la reorganización del citoesqueleto mediante la modulación global de la expresión génica.	Regulador central de la expresión de vías esenciales para la metástasis. Su sobreexpresión se ha asociado con aumento en la agresividad y crecimiento en una gran cantidad de tumores, tales como gástrico (89), renal (90), de mama (91), ovario (92), pulmón (93), entre otros (91).
PLAU / UROK / GC10P073909	Serin-proteasa que convierte al plasminógeno en plasmina.	Sobreexpresado en tejidos de cáncer de próstata y sus líneas celulares invasivas, donde se ha sugerido que se regula mediante metilación; asociado con alta capacidad invasiva *in vitro* y de tumorigénesis *in vivo* (94). Propuesto como biomarcador de pronóstico y blanco terapéutico en cáncer gástrico (95).
S100A11 / S10AB / GC01M152032	También conocida como calgizzarina. Proteína perteneciente a la familia S100, la más grande de unión a calcio con dos motivos EF, las cuales están involucrados en una gran gama de procesos	Sobreexpresado en varios tumores malignos, como adenocarcinoma y carcinoma de células escamosas (96), de tiroides papilar (97) y hepatocelular (98), cáncer de ovario (99), gástrico (100), colorectal (101), pancreático (102) y en mesotelioma pleural (103). Su silenciamiento

	como diferenciación y progresión del ciclo celular.	puede inhibir la proliferación celular, el crecimiento y la invasión en varios contextos cancerosos (96,99,102,103). Subexpresado en cáncer pulmonar de células pequeñas (96).

En resumen, los genes diferencialmente expresados al reprimir a DAXX desarrollan principalmente funciones tejido-específico. Adicionalmente, se identificaron 7 genes candidatos que podrían ser reprimidos por DAXX de manera constante en todas las células humanas, los cuales presentan funciones relevantes en el desarrollo y/o progresión del cáncer.

5.2 Análisis de expresión diferencial en elementos repetitivos

Después de la identificación de los genes codificantes diferencialmente expresados, procedimos a evaluar el cambio en la expresión de los elementos repetidos del genoma tras el silenciamiento de DAXX. Los resultados del análisis de expresión diferencial con DESeq2(72) utilizando la matriz de conteos de TETranscripts (66) identificaron un total de 2, 405 y 111 elementos repetidos diferencialmente expresados en HCT116, PC3 y GSC23, respectivamente (**Figura 5.2.1A**). Al igual que con los genes codificantes, no se encontró un patrón claro de cambio; el porcentaje de elementos repetidos sobreexpresados varió bastante con 100% en HCT116, 1% en PC3 y 59% en GSC23 (**Figura 5.2.1B**).

Las familias de repetidos que se encontraron diferencialmente expresadas de manera predominante fueron los elementos nucleares dispersos; además de esto, todas las líneas presentaron al menos un satélite diferencialmente expresado (**Figura 5.2.2A**). A pesar de que la tendencia de cambio de expresión de algunas familias fue evidente, no encontramos ningún elemento repetido puntual diferencialmente expresado en las tres líneas celulares; las únicas intersecciones encontradas fueron la del elemento transponible LTR7Y entre HCT116 y PC3 y la de 63 elementos más entre GSC23 y PC3 (**Figura 5.2.2B**). Como se mencionó en la sección anterior, la inestabilidad cromosómica de las células no diploides podría ocasionar

que el efecto de DAXX varíe debido a la sinergia que se presenta ante una estructura cromatínica anormal.

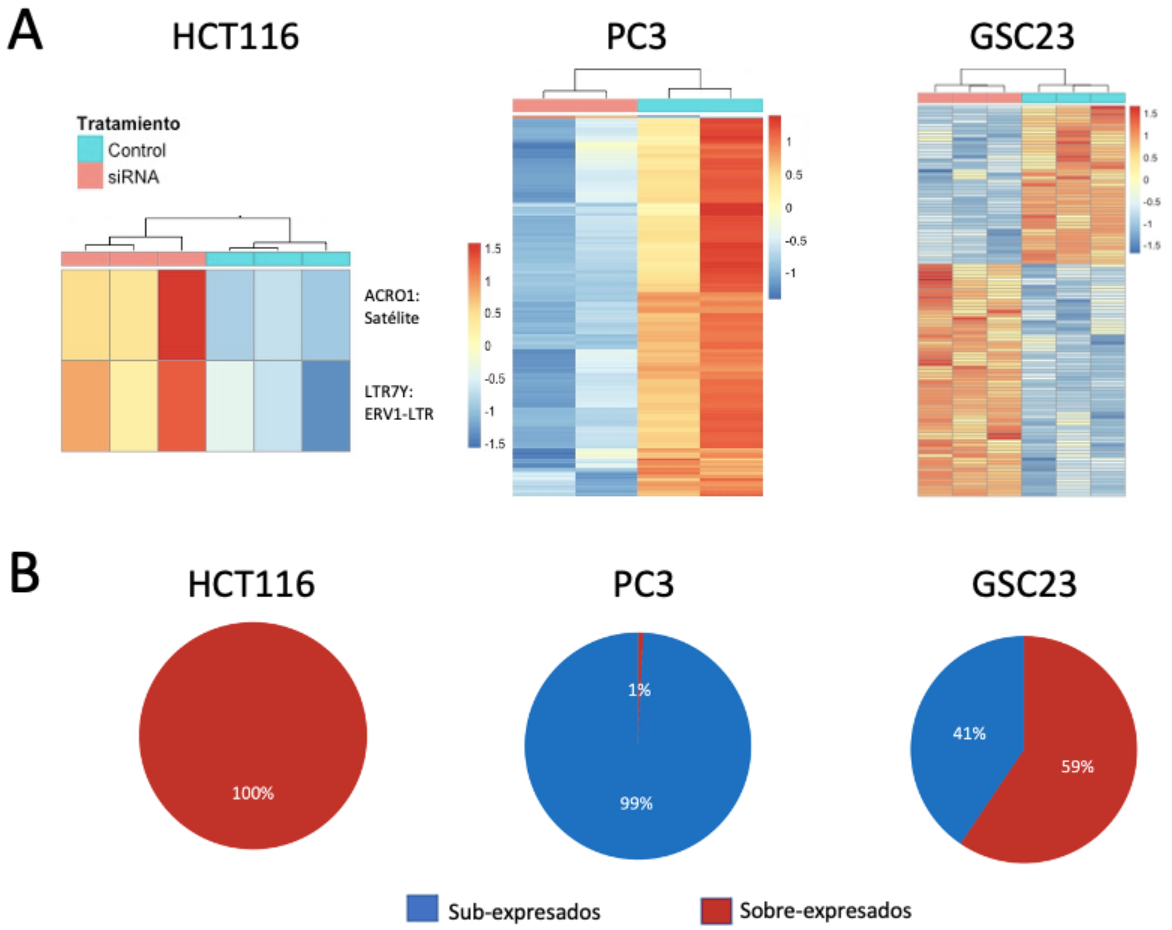


Figura 5.2.1 Elementos repetitivos diferencialmente expresados en DAXX *KD*. A) Expresión de cada elemento repetido diferencialmente expresado en HCT116, PC3 y GSC23. B) Proporción de elementos repetitivos sobre y subexpresados en cada línea celular.

Al igual que con los genes codificantes, os resultados fueron validados utilizando un algoritmo diferente; para esto, utilizamos SalmonTE (68), un programa que se limita al mapeo y cuantificación de elementos transponibles en el RNA-seq, que fue uno de los clados que mayor cantidad de elementos diferencialmente expresados encontramos en las muestras con TETranscripts. De forma interesante, LTR7Y, el elemento repetido DE encontrado en común entre HCT116 y PC3 con TETranscripts, también se encontró DE (p adj. < 0.1) con SalmonTE en dichas líneas celulares. En resumen, al analizar la expresión de elementos

repetidos ante la represión de DAXX, encontramos un patrón menos claro que el de el análisis de genes codificantes; la regulación de DAXX en elementos repetidos parece ser completamente tejido-específico, ejerciendo su función principalmente en elementos nucleares dispersos.

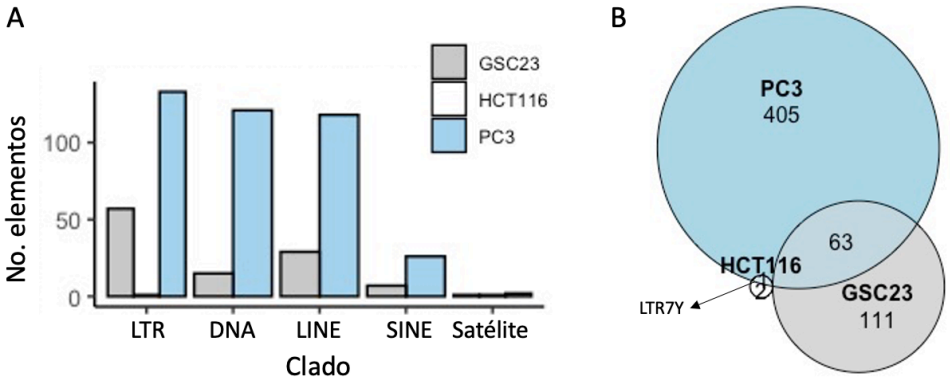


Figura 5.2.2 Clados de elementos repetidos diferencialmente expresados en DAXX *KD*. A) Cantidad de elementos DE por clado en HCT116, PC3 y GSC23. De izquierda a derecha se encuentran los clados que más elementos diferencialmente expresados presentaron de manera general en los experimentos. B) Elementos repetidos DE en común en los tres modelos celulares.

Debido a que DAXX ha sido descrito principalmente como represor de elementos repetidos en las zonas teloméricas y pericentroméricas, y en grupos de elementos retrovirales endógenos (104), nos preguntamos si la ubicación de estos elementos repetidos diferencialmente expresados concordaba con estas regiones. Para esto, mapeamos la secuencia consenso de los que presentaron una tasa de cambio mayor en cada modelo y la de LTR7Y debido a su presencia en HCT116 y PC3 de forma consistente. A pesar de que se encuentran presentes en muchas regiones cromosómicas, en especial los elementos nucleares dispersos, se puede observar que todos ellos se encuentran en las zonas teloméricas y pericentroméricas de algún cromosoma (**Figura 5.2.3**). Si bien es imposible saber en este análisis con precisión que los elementos diferencialmente expresados provienen de estas regiones, se puede hipotetizar que sea así.

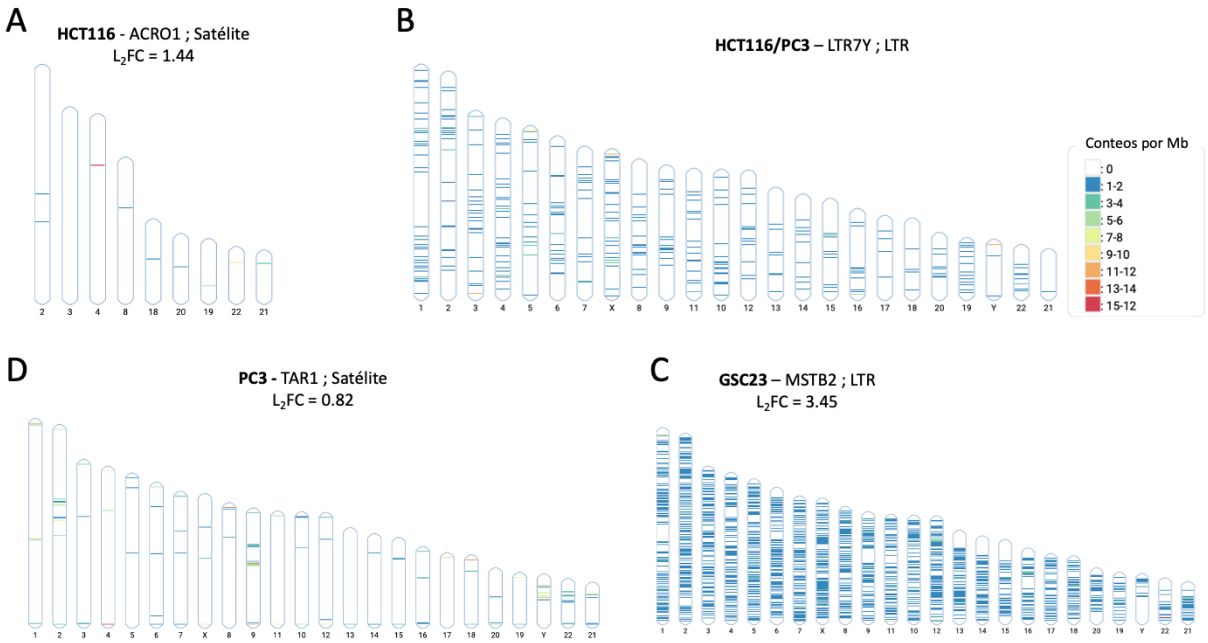


Figura 5.2.3. Ubicación de elementos repetidos diferencialmente expresados. Posición de los elementos repetidos DEs con la tasa de cambio más grande en A)HCT116, C) PC3 y D) GSC23, así como de B) LTR7Y. Las franjas corresponden a segmentos en donde se encuentra la secuencia de dicho repetido de forma no redundante (i.e. segmentos donde no hay otra familia de repetidos que posea una calificación de acierto mayor en el alineamiento). El color indica el número de veces que se presenta el elemento por cada megabase.

VI. Discusión

En este trabajo se estudió el transcriptoma de líneas celulares con y sin un tratamiento *DAXX KD* con el objetivo de evaluar el impacto que tiene su subexrpesión en células humanas. A pesar de la evidencia existente que asocia la alteración de su expresión con la inestabilidad genómica y la presencia de un fenotipo agresivo en cáncer, hasta este estudio, no se habían comparado los datos de distintos trabajos. Además de otorgar una descripción de los elementos del transcriptoma regulados por DAXX en HCT116, un linaje celular no explorado con este abordaje en la literatura, en este trabajo, mediante un análisis comparativo, identificamos genes candidatos cuya expresión es regulada por esta chaperona independientemente del contexto celular.

6.1 Regulación de la expresión mediada por DAXX

De manera general, observamos que DAXX no solamente juega un papel represor en la modulación de genes. De manera consistente se encontraron genes subexpresados tras la represión de esta chaperona que fueron principalmente específicos en cada línea celular, lo cual indica que los niveles normales de esta proteína estimulan la expresión de estos elementos. A pesar de que esto es sorprendente, dado que DAXX ha sido descrito principalmente como represor (7), nuestros resultados concuerdan con hallazgos previos en donde se observa este fenómeno (23,60,105); es interesante que a pesar de que se encuentran genes subexpresados ante la represión de DAXX, ningún autor en esos estudios aborda esto. El mecanismo por el cual lo logra no puede inferirse con los datos provistos en esta tesis, mas es posible hipotetizar mecanismos directos e indirectos por los cuales DAXX podría desempeñar este papel activador, como se desarrolla en los siguientes párrafos.

Además del depósito de H3.3 mediante el modelo clásico explicado en la introducción, en el cual ATRX-DAXX se asocia con el nucleosoma modificado y proteínas represoras para depositar la variante de histona y reprimir el entorno (**Figura 1.3.1**), esta chaperona es capaz de actuar de manera independiente para reprimir ERVs (17). Aunque este fenómeno ha sido pobremente estudiado, se ha reportado que H3.3 no necesita ser depositado en las regiones para que se reprima el entorno; en estos casos, se ha sugerido que la variante de histona

únicamente estabiliza a DAXX y le permite desempeñar su función regulatoria sin necesidad de ser incorporada en la cromatina (17).

Con base en lo anterior, podemos hipotetizar que DAXX altera la expresión de los genes de dos maneras diferentes. De manera directa, en regiones donde funge como represor, la chaperona puede realizar esa función ya sea mediante la incorporación de H3.3 en la cromatina o el reclutamiento por sí sola de remodeladores epigenéticos como SETDB1, KAP1 y HDAC (**Figura 6.1A)**(17). En el caso de los genes y regiones repetidas en los que DAXX desempeña una función activadora, podría deberse a efectos indirectos en donde la chaperona regule negativamente al represor de dicho elemento; de esta manera, al silenciar a DAXX, la represión se elimina; esto es factible dado que hay casos reportados donde sucede, como lo es la supresión mediada por DAXX de la represión transcripcional mediada por Slug en células pulmonares (**Figura 6.1A)**(106). Otra alternativa podría ser mediante un mecanismo descrito recientemente, en el cual se ha visto que la histona H3.3, independientemente de la chaperona que la deposita, puede ser fosforilada en la serina 31(14). La fosforilación de este residuo se asocia con una estimulación de la actividad de la acetiltransferasa p300 (14), con un aumento en la marca epigenética H3K27ac y con un estado cromatínico abierto (107). De esta manera, podría ser que en contextos específicos, DAXX pudiera depositar a la histona H3.3 sin establecer un ambiente heterocromático y que, posteriormente, esta histona sea modificada para ejercer un rol activador (**Figura 6.1B)**.

Para corroborar el mecanismo por el cual DAXX puede modular la expresión, es necesario realizar ensayos de ChIP-seq contra H3.3 en presencia y ausencia de DAXX, o contra DAXX antes y después de la represión. Comparando los datos de dichas condiciones, podríamos observar si el efecto de alteración en la expresión génica que observamos se asocia con la ausencia de la unión de DAXX a la región circundante o del nucleosoma variante con H3.3. Existen dos ensayos así: uno contra DAXX en PC3, donde se observó que esta chaperona se une a enhancers, insulators y promotores de varios genes, entre ellos los de autofagia (23); el segundo es un ChIP-seq contra H3.3 antes y después del silenciamiento de DAXX, donde observan que la sobre y subexpresión de genes se asocia tanto con un enriquecimiento como una disminución de H3.3 en los genes (60). Sin embargo, estos estudios no profundizan o

exploran los datos más allá de sus intereses particulares, por lo que en el futuro podrían re-analizarse para evaluar la distribución de DAXX y H3.3 en el grupo de genes que encontramos diferencialmente expresados en común para saber si el efecto que observamos se debe a cambios en H3.3, DAXX o ambos. Aún así, no existen ensayos de ChIP-seq contra la fosforilación de la serina 31 de H3.3, por lo que esta hipótesis se mantiene aún inexplorada en la literatura.

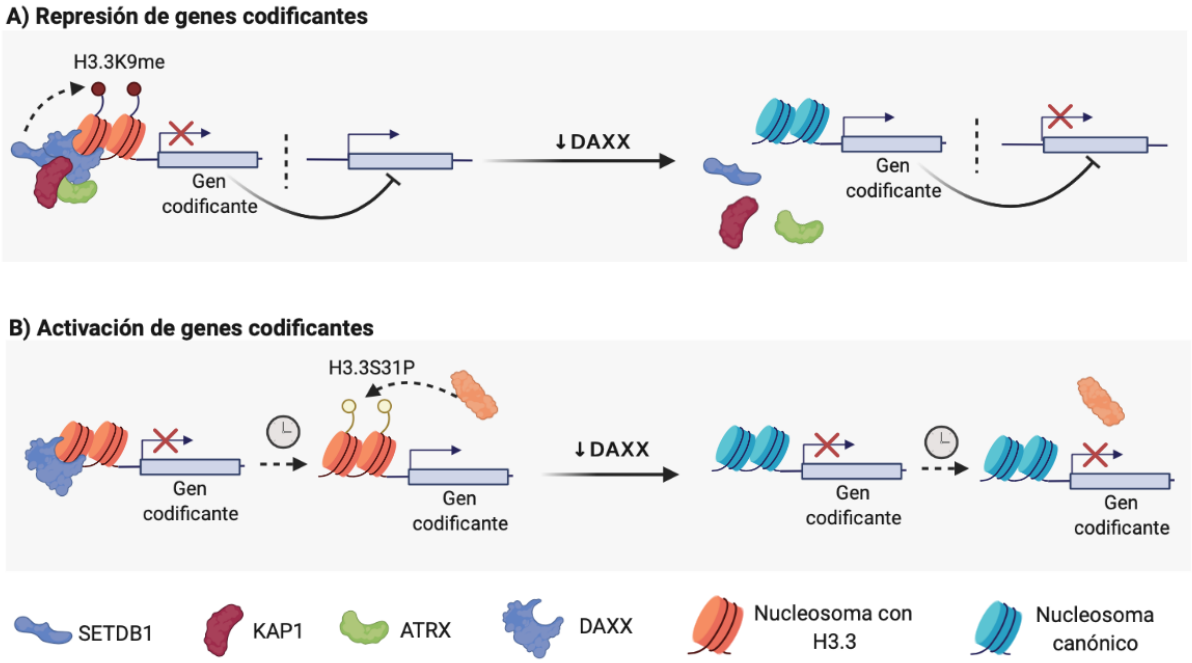


Figura 6.1 Impacto de DAXX en el genes codificantes. Potenciales mecanismos mediante los cuales DAXX podría modular la activación o represión de genes. A) El complejo ATRX-DAXX puede asociarse con proteínas represoras como KAP1 y SETDB1, lo cual genera un ambiente heterocromático. Al depositar a H3.3 en regiones cercanas a un gen codificante, DAXX puede reprimir su expresión. Adicionalmente, en caso de que dicho gen ocasione el silenciamiento de otros genes, el apagarlo ocasiona la activación transcripcional de estos últimos. B) DAXX puede activar la transcripción de forma directa de genes mediante el depósito de H3.3 en regiones cercanas del mismo. La variante de histona puede ser posteriormente fosforilada en la serina 31 y estimular la transcripción.

En resumen, los cambios de expresión que observamos en genes codificantes y elementos repetitivos podrían deberse a su regulación directa por DAXX, o como efecto indirecto tras la represión o activación de un regulador transcripcional que fue alterado por la cascada de

alteraciones en la expresión génica. Cuando actúa de forma directa, DAXX podría reprimir elementos en complejo con ATRX o con otras proteínas epigenéticas como SETDB1, KAP1 y HDACs (17). Por otro lado, DAXX podría mediar la activación transcripcional al depositar a H3.3 en una región genómica, la cual puede ser posteriormente fosforilada y promover la apertura de la cromatina (14,107).

6. 2 Impacto de DAXX en la regulación génica

Independientemente del mecanismo de acción, observamos que DAXX regula principalmente genes asociados a procesos específicos de cada línea celular. Al ser el primer estudio comparativo con este enfoque, nuestros resultados contribuyen al entendimiento de los efectos de la alteracióin de DAXX en células humanas. Estos son consistentes con un estudio realizado en un modelo murino de knock-out condicional de DAXX en páncreas; al analizar el transcriptoma, se encontró que genes ,incluidos factores de transcripción, asociados con tipos celulares específicos de este órgano estaban desregulados en ausencia de DAXX, concluyendo que en pacientes humanos con cáncer pancreático la alteración epigenética ocasionada por esta chaperona podría estar involucrada con la regulación transcripcional isleta-específica (105).

Al observar este comportamiento, parece probable que DAXX forme complejos con factores de transcripción tejido específico y que estos guíen su acción a lo largo del genoma. Esto es particularmente interesante si se toma en cuenta su gran capacidad para formar complejos con una amplia variedad de proteínas involucradas en la transcripción, como CBP, PTEN, RELA o MEN1 tanto *in vivo* como *in vitro* (26). Este mecanismo podría explicar el fenómeno observado en esta nuestros resultados. De ser así, un ChIP-seq contra DAXX en el caso de HCT116 y GSC23, y un reanálisis de los datos existentes en PC3 (23), podría ser útil para realizar un descubrimiento de motivos de DNA *de novo*; este método computacional es capaz de identificar los sitios de unión más abundantes en un conjunto de secuencias (108). Así, tras comparar los motivos encontrados con este método con bases de datos de sitios de unión de factores de transcripción humanos, podrían identificarse potenciales compañeros moleculares en cada una de las líneas celulares, si es que son tejido-específico.

En concordancia con lo anterior, encontramos pocos genes codificantes diferencialmente expresados en común al comparar los resultados de los tres experimentos (DAXX *KD* de HCT116, PC3 y GSC23). A pesar de esto, es muy interesante observar que todos los que están en la intersección desempeñan papeles relevantes en cáncer; estos podrían pertenecer a un conjunto de genes "basales" cuya expresión es afectada por DAXX independientemente del tipo celular. Otra alternativa posible es que la expresión de este grupo se encuentre alterada en muchos tipos de cáncer, y que la disminución *DAXX* potencie su desregulación. Como se puede observar en nuestros resultados, dichos genes no son parte de una misma cascada molecular, dado que no poseen interacciones a nivel proteína-proteína entre ellos. Esto indica que su mecanismo biológico es independiente, lo cual concuerda con la hipótesis de que son regulados por DAXX de manera pre-transcripcional.

La búsqueda de este grupo de genes es muy relevante para encontrar mecanismos conservados en distintas células humanas que podrían explicar la adquisición de características agresivas en varios tipos de cáncer donde DAXX se encuentra desregulado. En esta dirección, sería interesante ampliar el análisis a un mayor número de tipos celulares en donde la alteración de esta chaperona es relevante y el número de muestras. De esta manera, podría definirse mejor este grupo de genes. Esto tendría grandes implicaciones para entender el panorama del cáncer en este contexto y resaltar su importancia como biomarcador, así como en la generación de tratamientos enfocados en medicina de precisión que podrían ayudar a las personas cuyo cáncer posea esta alteración, como ha sido sugerido en los últimos años (26).

Como se mencionó en la introducción, la desregulación de DAXX tanto a la alza como a la baja se asocia con fenotipos agresivos en varios tipos de cáncer. Con base en esto, podemos inferir que los niveles de expresión de esta chaperona son sumamente importantes; cualquier aumento o disminución, genera una desregulación que puede inducir inestabilidad cromosómica (19) o mal pronóstico en cáncer (23-35). En consecuencia, una perspectica de este trabajo es realizar un análisis transcriptómico en el escenario opuesto (i.e. sobreexpresar a DAXX) para identificar los genes cuya expresión es alterada y observar si este grupo de

genes “basales” se mantiene y cómo es que se comporta. Este tipo de ensayos no se han realizado a nuestro conocimiento, por lo que este modelo se encuentra abierto a la exploración.

En resumen, se encontró mayoritariamente una alteración de genes involucrados en procesos tejido-específico en cada una de las líneas celulares. Aún así, encontramos un conjunto de genes que podrían pertenecer a un grupo de respuesta conservado independientemente del linaje. Estos podrían explicar el efecto de mal pronóstico que se asocia con la alteración de DAXX en cáncer.

6. 3. Impacto de DAXX en la regulación de elementos repetidos

Actualmente, lo más aceptado en la función regulatoria realizada por DAXX se enfoca en su rol represor de las regiones heterocromáticas, constituidas principalmente por elementos repetidos. En nuestro trabajo, observamos un comportamiento similar al de las regiones génicas codificantes, en donde se presentó una respuesta principalmente tejido específico. Además, se mantuvo la tendencia a promover principalmente la sobreexpresión de elementos repetidos en dos líneas celulares, reforzando así la idea de DAXX como represor en estas regiones. A pesar de ello, para nuestra sorpresa, en la línea celular PC3 se presentó una predominancia de elementos subexpresados. Esto sugiere que esta chaperona podría desempeñar un comportamiento diferente en esta línea celular a diferencia de las otras. Podría ser algo específico de tejido de próstata o algún fenómeno derivado de la inestabilidad cromosómica presente en PC3. El hecho de que esta línea celular sea casi triploide (75) refleja una gran cantidad de alteraciones presentes en la cromatina. En consecuencia, podría ser que las regiones ya comprometidas por esta desregulación se vean afectadas por la alteración de DAXX de forma diferente a HCT116, una línea celular cromosómicamente estable (74). En este sentido, se podría sugerir que la inestabilidad cromosómica de las células no diploides podría ocasionar un efecto sinérgico con el KD DAXX y siendo así resulten más representados como mayormente alteradas diferentes regiones repetidas y génicas en los modelos celulares.

En resumen, cada línea celular presentó cambios de expresión en elementos repetidos diferentes, lo cual concuerda con nuestra hipótesis de DAXX regulando elementos específicos en distintos linajes celulares. Sería interesante evaluar la expresión de elementos repetidos en una mayor cantidad de muestras para saber si este fenómeno es único de PC3 o es prevalente en otros tejidos. Así mismo, se podrían realizar comparaciones entre diferentes células del mismo linaje para evaluar la similitud entre los elementos repetidos diferencialmente expresados. De esta manera, el rol de DAXX en estas regiones podría ser esclarecido para saber si su desregulación ocasiona efectos opuestos o si esto se debe a la inestabilidad cromosómica particular de algunas líneas.

Independientemente de la tendencia global de cambio de expresión, los principales elementos repetidos que fueron desregulados corresponden al clado LTR en todas las líneas celulares. Dentro de este grupo, se encuentran los elementos retrovirales endógenos (ERVs). Esto coincide con la literatura, ya que su represión por DAXX ha sido reportada previamente (26,104,105). A pesar de eso, los elementos que mayor cambio presentaron pertenecen al clado de satélites. Esto último podría explicarse por la ubicación de estos repetidos, los cuales están principalmente en zonas de heterocromatina constitutiva, como la de las regiones centromérica, pericentromérica, o telomérica (109), que son regiones donde DAXX deposita a H3.3 (3). De esta manera, podría ser que estas regiones cromosómicas sean particularmente sensibles a la alteración de esta chaperona, siendo las más afectadas. Aún así, comprobar la hipótesis de que la expresión de los satélites sea específicamente de estas regiones pericentroméricas y subteloméricas es muy difícil con métodos computacionales debido a la alta cantidad de repetidos en tándem que poseen los satélites, como se describe en la sección 1.6 de la introducción. Aún así, el abordaje utilizado aquí nos permitió identificar elementos diferencialmente expresados aún sin tener la seguridad de la región genómica de la que preceden los transcritos. Esto es sumamente importante, dado que los elementos repetidos influencian la expresión de genes en *cis* y en *trans*, modulando incluso redes regulatorias completas (39,41). Por ende, conocer las principales familias afectadas puede ser crucial para entender el mecanismo de acción de DAXX en cáncer y su contribución a la inestabilidad genómica (19). De esta manera, podemos obtener un panorama de los cambios aún con las limitaciones actuales de la secuenciación corta.

Con los resultados provistos en esta tesis, en el futuro será interesante estudiar la relación que hay entre los patrones observados en genes codificantes y elementos repetitivos con el fin de describir si la alteración de uno ocasiona la del otro grupo, o si es que su desregulación sucede simultáneamente. Recientemente, en un modelo murino, se observó que la alteración que ocasiona DAXX en la expresión génica podría deberse principalmente a la activación de ERVs los cuales, al estar cerca de genes codificantes, generan un impacto global en la regulación del transcriptoma (105). Este modelo está apoyado por una gran variedad de trabajos que describen el efecto que tiene la alteración de elementos repetidos en la expresión génica de organismos eucariontes (38,39,41,42,44,110)

De manera general, la expresión de elementos repetitivos pueden alterar la de genes codificantes al actuar en cis, trans, mediante sus proteínas, o al alterar el entorno epigenético de sus regiones (110). Varios repetidos pueden servir como promotores y activar al expresión de genes en *cis*, como es el caso de ERV1 (111), o como enhancers y actuar en trans, como LTR19B o MER41 en humanos (112). Por otro lado, las proteínas pueden alterar el funcionamiento celular, como es el caso de HERV-K, cuyo producto funcional modula la expresión de citoquinas (113). Finalmente, dado que estos elementos son altamente controlados por la maquinaria epigenética, generalmente mediante el mantenimiento de regiones heterocromáticas que se asocian con marcas represivas como la H3K9me2/3 (114), el entorno represivo generado (o activo, cuando existe una desregulación) modifica la expresión de genes circundantes (110). Así, la expresión de elementos repetitivos impacta la red transcripcional de una célula, la cual al ser alterada puede desembocar en inestabilidad, por lo que sería relevante en estudios posteriores explorar qué tanto contribuyen estos mecanismos al fenotipo observado en células humanas (**Figura 6.2**).

Aún así, el modelo de alteración de expresión de genes codificantes como producto de ERVs propuesto en ratón un un contexto DAXX *KO* (105) explica potencialmente la expresión de menos de la mitad de los genes diferencialmente expresados, lo cual abre puerta proponer mecanismos complementarios que expliquen el fenotipo resultante. En este sentido, el hecho de que la alteración de elementos repetidos sea el principal detonador del efecto observado

en genes codificantes (i.e. perturbación de procesos biológicos tejido-específico) parece poco probable; de ser este mecanismo el prevalente en nuestros modelos, observaríamos alteraciones en vías azarosas. A pesar de esto, este mecanismo de alteración de genes codificantes por expresión de elementos repetidos muy probablemente sucede en nuestro modelo, y podría participar en la activación/represión de varios genes. En resumen, observamos elementos repetidos diferencialmente expresados en todas nuestras líneas celulares, los cuales podrían ocasionar potencialmente la desregulación de varios genes codificantes y contribuir así a la promoción de inestabilidad genómica y agresividad en cáncer.

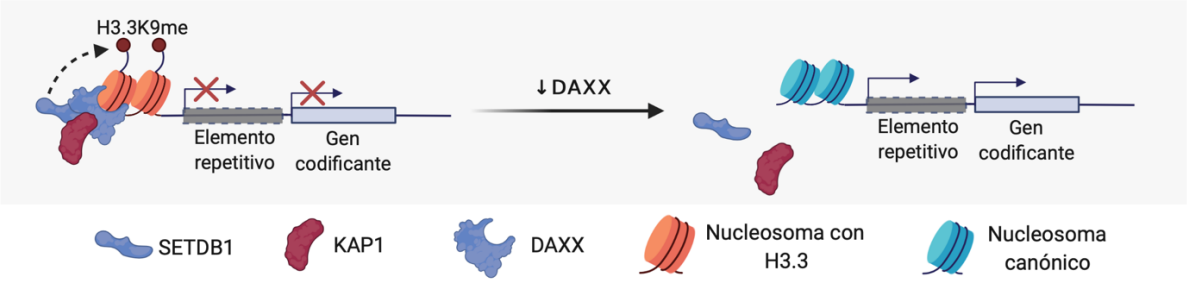


Figura 6.2 Impacto de DAXX en los elementos repetitivos. DAXX, en complejo con proteínas epigenéticas represoras como SETDB1 y KAP1, puede reprimir la transcripción de elementos repetitivos al depositar a H3.3 en las regiones cercanas y establecer un entorno heterocromático. El silenciamiento de DAXX puede ocasionar que este mecanismo falle, resultando en la activación de dichos elementos, los cuales son capaces de alterar la red transcripcional de la célula al activar genes codificantes en *cis* (como se muestra en la figura) o en *trans*.

En conjunto, los resultados de este trabajo contribuyen al entendimiento del impacto que tiene la desregulación de DAXX y es un primer paso para comenzar a entender su mecanismo de acción sobre la regulación génica **(Figura 6.3**). Parece ser que esta chaperona, y su variante de histona, actúan según la teoría del reostato, la cual dice que forman parte de mecanismos finos que le permiten a la célula mantener dominios genómicos a largo plazo y responder a estímulos (13). De esta manera, la perturbación de estos mecanismos podría cambiar este reóstato molecular de estabilidad genómica, contribuyendo a la enfermedad o su progresión.

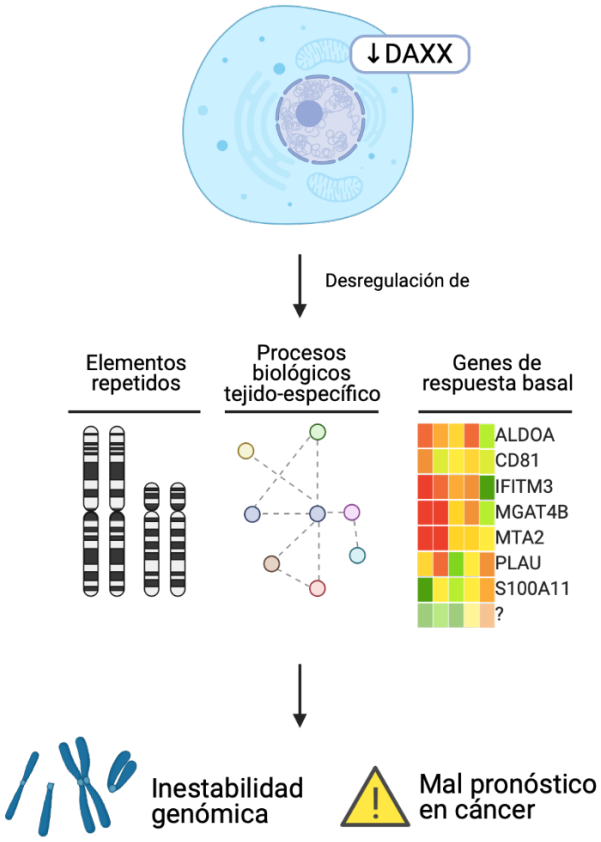


Figura 6.3 Consecuencias de la subexpresión de DAXX en células humanas. La represión de la chaperona de H3.3 ocasiona la desregulación de elementos repetidos y genes codificantes involucrados en procesos tejido-específico, así como un grupo de ellos conservado en todas las líneas celulares independientemente del linaje. Como consecuencia de estos tres fenómenos, se presenta inestabilidad genómica y un mal pronóstico en diversos tipos de cáncer.

VII. Conclusiones

La disminución de DAXX altera principalmente la expresión de genes involucrados en procesos tejido-específico en líneas celulares cancerosas de próstata, colon y glioblastoma. Se identificaron siete genes que potencialmente podrían integrar un grupo de respuesta basal a la alteración de esta chaperona: *ALDOA, CD81, IFITM3, MGAT4B, MTA2, PLAU* y *S100A11*.

La disminución de DAXX promueve alteraciones en la regulación de los elementos repetitivos, donde observamos varios clados diferencialmente expresados en todas líneas celulares, siendo la mayoría de ellos integrantes del grupo LTR. Aún así, los repetidos que presentaron una mayor tasa de cambio son usualmente satélites. Por úlitmo, las tres líneas celulares encontramos comportamientos muy diferentes en cuanto a proporción de elementos sub y sobreexpresados, siendo PC3 y HCT116 las más disímiles entre ellas.

Finalmente, se observó una proporción importante de genes codificantes y elementos repetidos donde la represión de DAXX disminuye su expresión en las tres líneas celulares, con una mayor predominancia en PC3. Esta función podría ser atribuida al depósito de H3.3 en dichas regiones y a su posterior fosforilación, siendo así un fenómeno en el cual DAXX no participa directamente.

Los resultados presentados en esta tesis conforman el primer estudio comparativo realizado a la fecha de transcriptomas humanos en células con un DAXX *KD,* donde se analizan cambios de expresión en genes codificantes y elementos repetitivos. Este trabajo destaca el papel importante de esta chaperona en el mantenimiento de la estabilidad genómica y sienta bases para comprender el mecanismo de acción que subyace el fenotipo agresivo en cáncer y la inestabilidad genómica que se asocia a su desregulación.

VIII. Perspectivas

Para de complementar los resultados de esta tesis, se planea:

- Buscar elementos repetidos cerca de los genes codificantes diferencialmente expresados y analizar si la expresión de su clado está alterada en la misma línea celular tras la represión de DAXX con el objetivo de identificar los genes cuyo cambio de expresión se pueda deber a la desregulación de repetidos.
- Hacer una prueba de enriquecimiento de blancos de factores de transcripción en el conjunto de genes diferencialmente expresados de cada línea celular con el objetivo de identificar proteínas tejido-específico potenciales con las cuales DAXX forme complejo.
- Analizar los datos de ChIP-seq contra DAXX y H3.3 disponibles públicamente de la línea PC3 (23) y GSC23 (60) con el objetivo de evaluar la presencia de estas proteínas antes y después de la represión de DAXX alrededor de los siete genes de respuesta basal y de los clados de elementos repetidos con el objetivo de evaluar si su desregulación se debe a un efecto directo o indirecto de DAXX y H3.3.

Referencias

1. Lewin B, Krebs JE, Goldstein ES, Kilpatrick ST. Lewin's Genes XI. 10th ed. Jones & Bartlett Learning; 2014. 940 p.

2. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.

3. Talbert PB, Henikoff S. Histone variants on the move: Substrates for chromatin dynamics. Nat Rev Mol Cell Biol. 2017;18(2):115–26.

4. Felsenfeld G, Groudine M. Controlling the double helix. Nature. 2003 Jan;421(6921):448–53.

5. McGinty RK, Tan S. Nucleosome Structure and Function. Chem Rev. 2015 Mar 25;115(6):2255–73.

6. Sawyer IA, Dundr M. Nuclear Bodies. In: Nuclear Architecture and Dynamics. Elsevier; 2018. p. 235–56.

7. Henikoff S, Smith MM. Histone variants and epigenetics. Cold Spring Harb Perspect Biol. 2015;7(1):1–25.

8. Van Holde KE, Allen JR, Tatchell K, Weischet WO, Lohr D. DNA-histone interactions in nucleosomes. Biophys J. 1980 Oct;32(1):271–82.

9. Enright HU, Miller WJ, Hebbel RP. Nucleosomal histone protein protects DNA from iron-mediated damage. Nucleic Acids Res. 1992;20(13):3341–6.

10. Attar N, Campos OA, Vogelauer M, Cheng C, Xue Y, Schmollinger S, et al. The histone H3-H4 tetramer is a copper reductase enzyme. Science (80-). 2020 Jul 3;369(6499):59–64.

11. Talbert PB, Henikoff S. Histone variants — ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol. 2010 Apr 3;11(4):264–75.

12. Martire S, Banaszynski LA. The roles of histone variants in fine-tuning chromatin organization and function. Nat Rev Mol Cell Biol. 2020 Sep 14;21(9):522–41.

13. Melters D, Nye J, Zhao H, Dalal Y. Chromatin Dynamics in Vivo: A Game of Musical Chairs. Genes (Basel). 2015 Aug 7;6(3):751–76.

14. Martire S, Gogate AA, Whitmill A, Tafessu A, Nguyen J, Teng Y-C, et al. Phosphorylation of histone H3.3 at serine 31 promotes p300 activity and enhancer acetylation. Nat Genet. 2019 Jun;51(6):941–6.

15. Szenker E, Ray-Gallet D, Almouzni G. The double face of the histone variant H3.3.

Cell Res. 2011 Mar 25;21(3):421–34.

16. Goldberg AD, Banaszynski LA, Noh K-M, Lewis PW, Elsaesser SJ, Stadler S, et al. Distinct Factors Control Histone Variant H3.3 Localization at Specific Genomic Regions. Cell. 2010 Mar;140(5):678–91.

17. Hoelper D, Huang H, Jain AY, Patel DJ, Lewis PW. Structural and mechanistic insights into ATRX-dependent and -independent functions of the histone chaperone DAXX. Nat Commun. 2017 Dec 30;8(1):1193.

18. Rapkin LM, Ahmed K, Dulev S, Li R, Kimura H, Ishov AM, et al. The histone chaperone DAXX maintains the structural organization of heterochromatin domains. Epigenetics Chromatin. 2015 Dec 21;8(1):44.

19. Marisa K, Arciga T. Participación de DAXX en la inducción de inestabilidad cromosómica mediante alteraciones en la heterocromatina constitutiva. Universidad Nacional Autónoma de México; 2019. p. 14080.

20. Giam M, Rancati G. Aneuploidy and chromosomal instability in cancer: a jackpot to chaos. Cell Div. 2015 Dec 20;10(1):3.

21. Tsuda H, Takarabe T, Kanai Y, Fukutomi T, Hirohashi S. Correlation of DNA Hypomethylation at Pericentromeric Heterochromatin Regions of Chromosomes 16 and 1 with Histological Features and Chromosomal Abnormalities of Human Breast Carcinomas. Am J Pathol. 2002 Sep;161(3):859–66.

22. Prada D, González R, Sánchez L, Castro C, Fabián E, Herrera LA. Satellite 2 demethylation induced by 5-azacytidine is associated with missegregation of chromosomes 1 and 16 in human somatic cells. Mutat Res Mol Mech Mutagen. 2012 Jan;729(1–2):100–5.

23. Puto LA, Benner C, Hunter T. The DAXX co-repressor is directly recruited to active regulatory elements genome-wide to regulate autophagy programs in a model of human prostate cancer. Oncoscience. 2015;2(4):362–72.

24. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010 Mar 9;11(3):204–20.

25. Kwan PS, Lau CC, Chiu YT, Man C, Liu J, Tang KD, et al. Daxx regulates mitotic progression and prostate cancer predisposition. Carcinogenesis. 2013 Apr;34(4):750–9.

26. Mahmud I, Liao D. DAXX in cancer: phenomena, processes, mechanisms and regulation. Nucleic Acids Res. 2019 Sep 5;47(15):7734–52.

27. Tsourlakis MC, Schoop M, Plass C, Huland H, Graefen M, Steuber T, et al. Overexpression of the chromatin remodeler death-domain–associated protein in prostate cancer is an independent predictor of early prostate-specific antigen

recurrence. Hum Pathol. 2013 Sep;44(9):1789–96.

28. Pan W-W, Zhou J-J, Liu X-M, Xu Y, Guo L-J, Yu C, et al. Death Domain-associated Protein DAXX Promotes Ovarian Cancer Development and Chemoresistance. J Biol Chem. 2013 May 10;288(19):13620–30.

29. Xu J, Zhao Z, Ye L, Zhuge W, Han Z, Zhang T, et al. Prognostic significance of Daxx NCR (Nuclear/Cytoplasmic Ratio) in gastric cancer. Cancer Med. 2017 Sep;6(9):2063–75.

30. Lin G-J, Huang Y-S, Lin C-K, Huang S-H, Shih H-M, Sytwu H-K, et al. Daxx and TCF4 interaction links to oral squamous cell carcinoma growth by promoting cell cycle progression via induction of cyclin D1 expression. Clin Oral Investig. 2016 Apr 24;20(3):533–40.

31. Gopal RK, Kübler K, Calvo SE, Polak P, Livitz D, Rosebrock D, et al. Widespread Chromosomal Losses and Mitochondrial DNA Alterations as Genetic Drivers in Hürthle Cell Carcinoma. Cancer Cell. 2018 Aug;34(2):242-255.e5.

32. Jiao Y, Shi C, Edil BH, de Wilde RF, Klimstra DS, Maitra A, et al. DAXX/ATRX, MEN1, and mTOR Pathway Genes Are Frequently Altered in Pancreatic Neuroendocrine Tumors. Science (80-). 2011 Mar 4;331(6021):1199–203.

33. Heaphy CM, de Wilde RF, Jiao Y, Klein AP, Edil BH, Shi C, et al. Altered Telomeres in Tumors with ATRX and DAXX Mutations. Science (80-). 2011 Jul 22;333(6041):425–425.

34. Chen S-F, Kasajima A, Yazdani S, Chan MSM, Wang L, He Y-Y, et al. Clinicopathologic significance of immunostaining of α-thalassemia/mental retardation syndrome X-linked protein and death domain–associated protein in neuroendocrine tumors. Hum Pathol. 2013 Oct;44(10):2199–203.

35. Yuan F, Shi M, Ji J, Shi H, Zhou C, Yu Y, et al. KRAS and DAXX/ATRX Gene Mutations Are Correlated with the Clinicopathological Features, Advanced Diseases, and Poor Prognosis in Chinese Patients with Pancreatic Neuroendocrine Tumors. Int J Biol Sci. 2014;10(9):957–65.

36. Marinoni I, Kurrer AS, Vassella E, Dettmer M, Rudolph T, Banz V, et al. Loss of DAXX and ATRX Are Associated With Chromosome Instability and Reduced Survival of Patients With Pancreatic Neuroendocrine Tumors. Gastroenterology. 2014 Feb;146(2):453-460.e5.

37. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics. 2014 Dec 11;15(1):583.

38. Kaul TK, Morales ME, Deininger PL. Repetitive Elements and Human Disorders. In: eLS. Chichester, UK: John Wiley & Sons, Ltd; 2017. p. 1–8.

39. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017 Feb 21;18(2):71–86.

40. Slotkin RK. The case for not masking away repetitive DNA. Mob DNA. 2018 Dec 1;9(1):15.

41. McCue AD, Slotkin RK. Transposable element small RNAs as regulators of gene expression. Trends Genet. 2012 Dec;28(12):616–23.

42. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018 Dec 19;19(1):199.

43. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. Mob DNA. 2016 Dec 6;7(1):9.

44. Chenais B. Transposable Elements in Cancer and Other Human Diseases. Curr Cancer Drug Targets. 2015 May 5;15(3):227–42.

45. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

46. Padeken J, Zeller P, Gasser SM. Repeat DNA in genome organization and stability. Curr Opin Genet Dev. 2015 Apr;31:12–9.

47. Eymery A, Horard B, Atifi-Borel M El, Fourel G, Berger F, Vitte A-L, et al. A transcriptomic analysis of human centromeric and pericentric sequences in normal and tumor cells. Nucleic Acids Res. 2009 Oct;37(19):6340–54.

48. Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, et al. Aberrant Overexpression of Satellite Repeats in Pancreatic and Other Epithelial Cancers. Science (80-). 2011 Feb 4;331(6017):593–6.

49. Bersani F, Lee E, Kharchenko P V., Xu AW, Liu M, Xega K, et al. Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. Proc Natl Acad Sci. 2015 Dec 8;112(49):15148–53.

50. Ho XD, Nguyen HG, Trinh LH, Reimann E, Prans E, Kõks G, et al. Analysis of the Expression of Repetitive DNA Elements in Osteosarcoma. Front Genet. 2017 Nov 30;8.

51. Schiavetti F, Thonnard J, Colau D, Boon T, Coulie PG. A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes. Cancer Res. 2002;62(19):5510–6.

52. Contreras-Galindo R, Kaplan MH, Leissner P, Verjat T, Ferlenghi I, Bagnoli F, et al. Human Endogenous Retrovirus K (HML-2) Elements in the Plasma of People with Lymphoma and Breast Cancer. J Virol. 2008 Oct 1;82(19):9329–36.

53. Pichon JP, Bonnaud B, Cleuziat P. Multiplex degenerate PCR coupled with an oligo sorbent array for human endogenous retrovirus expression profiling. Nucleic Acids Res. 2006 Mar 23;34(6):e46–e46.

54. Wang-Johanning F, Frost AR, Jian B, Azerou R, Lu DW, Chen D-T, et al. Detecting the expression of human endogenous retrovirus Eenvelope transcripts in human prostate adenocarcinoma. Cancer. 2003 Jul 1;98(1):187–97.

55. Gimenez J, Montgiraud C, Pichon J-P, Bonnaud B, Arsac M, Ruel K, et al. Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. Nucleic Acids Res. 2010 Apr;38(7):2229–46.

56. Pérot P, Mullins CS, Naville M, Bressan C, Hühns M, Gock M, et al. Expression of young HERV-H loci in the course of colorectal carcinoma and correlation with molecular subtypes. Oncotarget. 2015 Nov 24;6(37):40095–111.

57. Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, et al. Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res. 2012 Dec 1;22(12):2328–38.

58. Shukla R, Upton KR, Muñoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T, et al. Endogenous Retrotransposition Activates Oncogenic Pathways in Hepatocellular Carcinoma. Cell. 2013 Mar;153(1):101–11.

59. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57–63.

60. Benitez JA, Ma J, D'Antonio M, Boyer A, Camargo MF, Zanca C, et al. PTEN regulates glioblastoma oncogenesis through chromatin-associated complexes of DAXX and histone H3.3. Nat Commun. 2017;8(May 2017):1–14.

61. Wethkamp N, Klempnauer K-H. Daxx Is a Transcriptional Repressor of CCAAT/Enhancer-binding Protein β. J Biol Chem. 2009 Oct 16;284(42):28783–94.

62. Puto LA, Reed JC. Daxx represses RelB target promoters via DNA methyltransferase recruitment and DNA hypermethylation. Genes Dev. 2008 Mar 26;22(8):998–1010.

63. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Dec 26;17(1):13.

64. Check Hyden E. The $1,000 genome. Nature. 2014;507:294–5.

65. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. Nat Rev Genet. 2018 Nov 19;19(11):688–704.

66. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. 2015;31(22):3593–9.

67. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

68. Jeong HH, Yalamanchili HK, Guo C, Shulman JM, Liu Z. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. In: Pacific Symposium on Biocomputing. World Scientific Publishing Co. Pte Ltd; 2018. p. 168–79.

69. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9.

70. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics; 2010.

71. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

72. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):1–21.

73. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi A J Integr Biol. 2012;16(5):284–7.

74. Miao Z-H, Player A, Shankavaram U, Wang Y-H, Zimonjic DB, Lorenzi PL, et al. Nonclassic Functions of Human Topoisomerase I: Genome-Wide and Pharmacologic Analyses. Cancer Res. 2007 Sep 15;67(18):8752–61.

75. Pan Y, Kytölä S, Farnebo F, Wang N, Lui WO, Nupponen N, et al. Characterization of chromosomal abnormalities in prostate cancer cell lines by spectral karyotyping. Cytogenet Genome Res. 1999;87(3–4):225–32.

76. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607–13.

77. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinforma. 2016 Jun 20;54(1).

78. Chang Y-C, Yang Y-C, Tien C-P, Yang C-J, Hsiao M. Roles of Aldolase Family Genes in Human Cancers and Diseases. Trends Endocrinol Metab. 2018 Aug;29(8):549–59.

79. Chang Y-C, Chiou J, Yang Y-F, Su C-Y, Lin Y-F, Yang C-N, et al. Therapeutic Targeting of Aldolase A Interactions Inhibits Lung Cancer Metastasis and Prolongs Survival. Cancer Res. 2019 Jul 29;79(18):4754–66.

80. Ji S, Zhang B, Liu J, Qin Y, Liang C, Shi S, et al. ALDOA functions as an oncogene in the highly metastatic pancreatic cancer. Cancer Lett. 2016 Apr;374(1):127–35.

81. Gizak A, Wiśniewski J, Heron P, Mamczur P, Sygusch J, Rakus D. Targeting a moonlighting function of aldolase induces apoptosis in cancer cells. Cell Death Dis [Internet]. 2019 Oct 26;10(10):712. Available from: http://www.nature.com/articles/s41419-019-1968-4

82. Yoo T-H, Ryu B-K, Lee M-G, Chi S-G. CD81 is a candidate tumor suppressor gene in human gastric cancer. Cell Oncol. 2013 Apr 21;36(2):141–53.

83. Zhang N, Zuo L, Zheng H, Li G, Hu X. Increased Expression of CD81 in Breast Cancer Tissue is Associated with Reduced Patient Prognosis and Increased Cell Migration and Proliferation in MDA-MB-231 and MDA-MB-435S Human Breast Cancer Cell Lines In Vitro. Med Sci Monit. 2018 Aug 17;24:5739–47.

84. Zhang Y, Qian H, Xu A, Yang G. Increased expression of CD81 is associated with poor prognosis of prostate cancer and increases the progression of prostate cancer cells in vitro. Exp Ther Med. 2019 Nov 26;19(1):755–61.

85. Mizoshiri N, Shirai T, Terauchi R, Tsuchida S, Mori Y, Hayashi D, et al. The tetraspanin CD81 mediates the growth and metastases of human osteosarcoma. Cell Oncol. 2019 Dec 7;42(6):861–71.

86. Min J, Feng Q, Liao W, Liang Y, Gong C, Li E, et al. IFITM3 promotes hepatocellular carcinoma invasion and metastasis by regulating MMP9 through p38/MAPK signaling. FEBS Open Bio. 2018 Aug 28;8(8):1299–311.

87. Liu X, Chen L, Fan Y, Hong Y, Yang X, Li Y, et al. IFITM3 promotes bone metastasis of prostate cancer cells by mediating activation of the TGF-β signaling pathway. Cell Death Dis. 2019 Jul 4;10(7):517.

88. Ashkani J, Naidoo KJ. Glycosyltransferase Gene Expression Profiles Classify Cancer Types and Propose Prognostic Subtypes. Sci Rep. 2016 Sep 20;6(1):26451.

89. Zhou C, Ji J, Cai Q, Shi M, Chen X, Yu Y, et al. MTA2 promotes gastric cancer cells invasion and is transcriptionally regulated by Sp1. Mol Cancer. 2013;12(1):102.

90. Chen Y-S, Hung T-W, Su S-C, Lin C-L, Yang S-F, Lee C-C, et al. MTA2 as a Potential Biomarker and Its Involvement in Metastatic Progression of Human Renal Cancer by miR-133b Targeting MMP-9. Cancers (Basel). 2019 Nov 23;11(12):1851.

91. Covington KR, Fuqua SAW. Role of MTA2 in human cancer. Cancer Metastasis Rev. 2014 Dec 14;33(4):921–8.

92. Yuxin J, Ping Z, Yunping L, Ding M. Expression of MTA2 gene in ovarian epithelial cancer and its clinical implication. J Huazhong Univ Sci Technol [Medical Sci. 2006 May;26(3):359–62.

93. Zhang B, Tao F, Zhang H. Metastasis-associated protein 2 promotes the metastasis of non-small cell lung carcinoma by regulating the ERK/AKT and VEGF signaling pathways. Mol Med Rep. 2018 Feb 1;

94. Greene KL, Li L-C, Okino ST, Carroll PR. Molecular Basis of Prostate Cancer. In: The Molecular Basis of Cancer. Elsevier; 2008. p. 431–40.

95. Ai C, Zhang J, Lian S, Ma J, Győrffy B, Qian Z, et al. FOXM1 functions collaboratively with PLAU to promote gastric cancer progression. J Cancer. 2020;11(4):788–94.

96. Hao J, Wang K, Yue Y, Tian T, Xu A, Hao J, et al. Selective expression of S100A11 in lung cancer and its role in regulating proliferation of adenocarcinomas cells. Mol Cell Biochem. 2012 Jan 23;359(1–2):323–32.

97. Anania MC, Miranda C, Vizioli MG, Mazzoni M, Cleris L, Pagliardini S, et al. S100A11 Overexpression Contributes to the Malignant Phenotype of Papillary Thyroid Carcinoma. J Clin Endocrinol Metab. 2013 Oct;98(10):E1591–600.

98. Sobolewski C, Abegg D, Berthou F, Dolicka D, Calo N, Sempoux C, et al. S100A11/ANXA2 belongs to a tumour suppressor/oncogene network deregulated early with steatosis and involved in inflammation and hepatocellular carcinoma development. Gut. 2020 Oct;69(10):1841–54.

99. LIU Y, HAN X, GAO B. Knockdown of S100A11 expression suppresses ovarian cancer cell growth and invasion. Exp Ther Med. 2015 Apr;9(4):1460–4.

100. Wang C, Luo J, Rong J, He S, Zhang L, Zheng F. Distinct prognostic roles of S100 mRNA expression in gastric cancer. Pathol - Res Pract. 2019 Jan;215(1):127–36.

101. Jung Y, Lee S, Choi H-S, Kim S-N, Lee E, Shin Y, et al. Clinical Validation of Colorectal Cancer Biomarkers Identified from Bioinformatics Analysis of Public Expression Data. Clin Cancer Res. 2011 Feb 15;17(4):700–9.

102. Xiao M, Li T, Ji Y, Jiang F, Ni W, Zhu J, et al. S100A11 promotes human pancreatic cancer PANC-1 cell proliferation and is involved in the PI3K/AKT signaling pathway. Oncol Lett. 2017 Oct 31;175–82.

103. Sato H, Sakaguchi M, Yamamoto H, Tomida S, Aoe K, Shien K, et al. Therapeutic potential of targeting S100A11 in malignant pleural mesothelioma. Oncogenesis. 2018 Jan 24;7(1):11.

104. Elsässer SJ, Noh KM, Diaz N, Allis CD, Banaszynski LA. Histone H3.3 is required for endogenous retroviral element silencing in embryonic stem cells. Nature.

2015;522(7555):240–4.

105. Wasylishen AR, Sun C, Moyer SM, Qi Y, Chau GP, Aryal NK, et al. Daxx maintains endogenous retroviral silencing and restricts cellular plasticity in vivo. Sci Adv. 2020 Aug 5;6(32):eaba8415.

106. Lin C-W, Wang L-K, Wang S-P, Chang Y-L, Wu Y-Y, Chen H-Y, et al. Daxx inhibits hypoxia-induced lung cancer cell metastasis by suppressing the HIF-1α/HDAC1/Slug axis. Nat Commun. 2016 Dec 23;7(1):13867.

107. Sitbon D, Boyarchuk E, Dingli F, Loew D, Almouzni G. Histone variant H3.3 residue S31 is essential for Xenopus gastrulation regardless of the deposition pathway. Nat Commun. 2020 Dec 9;11(1):1256.

108. Hashim FA, Mabrouk MS, Al-Atabany W. Review of Different Sequence Motif Finding Algorithms. Avicenna J Med Biotechnol. 2019;11(2):130–48.

109. Miga KH. Completing the human genome: the progress and challenge of satellite DNA assembly. Chromosom Res. 2015 Sep 12;23(3):421–6.

110. Sofuku K, Honda T. Influence of Endogenous Viral Sequences on Gene Expression. In: Gene Expression and Regulation in Mammalian Cells - Transcription From General Aspects. InTech; 2018.

111. Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, et al. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet. 2014 Jun 28;46(6):558–66.

112. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science (80-). 2016 Mar 4;351(6277):1083–7.

113. Morozov VA, Dao Thi VL, Denner J. The Transmembrane Protein of the Human Endogenous Retrovirus - K (HERV-K) Modulates Cytokine Release and Gene Expression. Belshaw R, editor. PLoS One. 2013 Aug 7;8(8):e70399.

114. Szpakowski S, Sun X, Lage JM, Dyer A, Rubinstein J, Kowalski D, et al. Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. Gene. 2009 Dec;448(2):151–67.

Anexos

1. Publicación de artículo original

Título: Comparative transcriptome analysis reveals key epigenetics targets in SARS-CoV-2 infection
Autores: Marisol Salgado-Albarrán, Erick I. Navarro-Delgado, Aylin del Moral-Morales, Nicolas Alcaraz, Jan Baumbach, Rodrigo González Barrios, Ernesto Soto-Reyes.
Revista: npj Systems Biology and Applications
Fecha de publicación: 24 / 05 / 2021.
DOI: 10.1038/s41540-021-00181-x

Title page

Comparative transcriptome analysis reveals key epigenetic targets in SARS-CoV-2 infection

Marisol Salgado-Albarrán[1,2,†], Erick I. Navarro-Delgado[3,†], Aylin Del Moral-Morales[1,†], Nicolas Alcaraz[4,5], Jan Baumbach[6,7] , Rodrigo González-Barrios[3*] and Ernesto Soto-Reyes[1*]

[1] Departamento de Ciencias Naturales, Universidad Autónoma Metropolitana-Cuajimalpa (UAM-C), Mexico City, Mexico.
[2] Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany.
[3] Unidad de Investigación Biomédica en Cáncer, Instituto Nacional de Cancerología, Mexico City, Mexico.
[4] The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen 2200.
[5] National Institute of Genomic Medicine, Periférico sur 4809, Arenal Tepepan, 14610 Mexico City, Mexico.
[6] Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany.
[7] Computational BioMedicine lab, Institute of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark.

* The authors contributed equally as corresponding authors:
Dr. Ernesto Soto-Reyes: Tel: +52(55)58146500 ext. 3879; Email:esotoreyes@cua.uam.mx.
Dr. Rodrigo González-Barrios: Tel:(55)56280400 ext. 70036; Email: rodrigop@ciencias.unam.mx

[†] These authors contributed equally to this work

ABSTRACT

COVID-19 is an infection caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome coronavirus 2), which has caused a global outbreak. Current research efforts are focused on the understanding of the molecular mechanisms involved in SARS-CoV-2 infection in order to propose drug-based therapeutic options. Transcriptional changes due to epigenetic regulation are key host cell responses to viral infection and have been studied in SARS-CoV and MERS-CoV; however, such changes are not fully described for SARS-CoV-2. In this study, we analyzed multiple transcriptomes obtained from cell lines infected with MERS-CoV, SARS-CoV and SARS-CoV-2, and from COVID-19 patient-derived samples. Using integrative analyses of gene co-expression networks and de-novo pathway enrichment, we characterize different gene modules and protein pathways enriched with Transcription Factors or Epifactors relevant for SARS-CoV-2 infection. We identified EP300, MOV10, RELA and TRIM25 as top candidates, and more than 60 additional proteins involved in the

epigenetic response during viral infection that have therapeutic potential. Our results show that targeting the epigenetic machinery could be a feasible alternative to treat COVID-19.

KEYWORDS
SARS-CoV-2, MERS-CoV, SARS-CoV, COVID-19, epigenetics, transcription factors, coronavirus infection, network analysis, co-expression analysis, drug repurposing.

INTRODUCTION

The coronavirus family (CoV) are non-segmented, positive-sense and enveloped RNA viruses that have been identified as the cause of multiple enteric and respiratory diseases in both animals and humans [1]. Three major CoV strains of this family have caused recent human pandemics: Middle East respiratory syndrome coronavirus (MERS-CoV) in 2002-2003 [2], severe acute respiratory syndrome coronavirus 1 (SARS-CoV) in 2012 and SARS-CoV-2 in 2020 [3]. The most recent one was identified in Wuhan, China by the end of 2019 and is the etiological origin of an atypical pneumonia known as Coronavirus Disease 2019 (COVID-19), which has caused a global outbreak and is one of the top sixth public health emergencies of international concern [4] with 98,089,877 confirmed cases and 2,100,404 deaths as of January, 2021, leading to the biggest CoV pandemic in modern times [5].

By being intracellular pathogens, viruses' infection strategy requires the continuous subordination and exploitation of cellular transcriptional machinery and metabolism in order to ensure its own expansion. To do so, the host genome expression must be used and, to be successful, this will depend on chromatin dynamics and transcription regulation, which are principally ruled by epigenetic mechanisms, such as DNA methylation, histone post-translational modifications (HPTM), and transcription factors (TFs) [6]. During a viral infection, it has been reported that epigenetic and transcriptional changes occur for both sides: the infected cell promotes an antiviral environmental response, leading to the induction of pathways to survive, while the virus switches off the expression of critical anti-viral host cell genes [7,8].

Several studies have reported the importance of epigenetic modifications in viral infections. In influenza virus, specific gene promoter DNA methylation [9], decreased H3K4me3 (a hallmark of active chromatin) [10], histone acetylation in H3 and H4 histones, and increased levels of H4K20me2 and unmodified H3K36 and H4K79 have been reported [11]. Interestingly, these HPTMs do not always trigger the same mechanisms and lead to similar phenotypes; for example, depletion of H3K79me2, an epigenetic mark that is usually increased upon viral infections due to an upregulation of DOT1L, results in impaired viral growth in human cytomegalovirus infection [12], while enhancing the replication in influenza virus [11]. However, these mechanisms usually lead to host transcriptional inactivation, which contributes to the altered cellular transcription produced by viral infections.

Regarding CoVs, few experimental studies have been conducted to unravel the epigenetic proteins and marks involved in their infection and pathogenesis in MERS-CoV and SARS-CoV, being especially scarce in SARS-CoV-2 due to its recent appearance. For MERS-CoV and SARS-CoV, different outcomes have been reported, such as the mechanisms used to control the interferon-stimulated genes, which involves H3K27 methylation in MERS-CoV but not in SARS-CoV [13], and the ones used to down-regulate antigen-presenting molecules, which involves DNA methylation in MERS-CoV and not in SARS-CoV [9]. These studies show that epigenetic mechanisms are highly important in the host gene expression control carried out by the virus and that, despite the phylogenetic closeness, these mechanisms can be very different between strains, highlighting the need to understand the epigenetic processes that play a role in SARS-CoV-2 infection.

Integrative computational methods are promising approaches used to generate research hypotheses, generate consensus regulatory networks and describe deregulated processes in SARS-CoV-2 infection [14,15]. Nevertheless, they have overlooked key epigenetic and TFs that underlie the infected phenotype. Since drugs that target the epigenetic landscape of diseased cells have shown great potential and have proved to be game-changing as complementary treatments of complex diseases, such as cancer [16], the identification of these key epigenetic proteins and TFs becomes highly important in our current context, where popular regimen candidates for treating COVID-19, such as Remdesivir, Hydroxychloroquine, Lopinavir and Interferon have shown to have little or no effect on reducing mortality of hospitalized COVID-19 individuals[17].

In this work, we gathered publicly available RNA-seq data from SARS-CoV-2, SARS-CoV and MERS-CoV infected cell lines and patient samples and performed differential expression analyses together with weighted gene co-expression network analysis to identify unique and shared central epigenetic players in SARS-CoV-2, SARS-CoV and MERS-CoV. Candidate genes were further prioritized by integrating differentially expressed genes (DEGs), enrichment tests, gene-coexpression network and viral-host protein-protein interaction network analysis to propose potential key epigenetic proteins involved in SARS-CoV-2 infection. Finally, we identified currently approved drugs that target key epigenetic drivers of SARS-CoV-2 infection, and thus they are potential new therapeutic approaches for COVID-19.

RESULTS

SARS-CoV-2, SARS-CoV and MERS-CoV induce different transcriptional and epigenetic responses during infection in pulmonary cell lines

In order to identify the genes that change their expression in pulmonary cell lines (Calu-3, MRC-5, A549 and NHBE) due to infection of Coronaviruses such as MERS-CoV, SARS-CoV or SARS-CoV-2, differential expression analysis was performed in RNA-seq data (Supplementary Table 1).

As a first approach, we evaluated the overlapping DEGs identified for each virus regardless of the cell type and in common among viruses (Supplementary Table 2). For MERS-CoV and SARS-CoV, the overlap among all cell conditions was considered. For SARS-CoV-2, the overlap among 3 out of the 4 cell conditions was used, since the NHBE cell line showed a small number of DEGs most likely because these cells are derived from normal bronchial epithelial cells [18,19] (Supplementary Figure 1A). We observed that the majority of the virus-associated genes are unique for each virus and a small proportion is shared among them. Specifically, only 3 genes were differentially expressed during infection in cell lines regardless of the virus evaluated (Figure 1A). Furthermore, GO enrichment analysis (Figure 1B) shows that the top 10 enriched GO terms are different for each virus, except “cellular response to lipopolysaccharide”, shared between SARS-CoV-2 and SARS-CoV; however, the three viruses share terms related to immune response processes (Supplementary Figure 1B). The latter shows that, despite their phylogenetic relationship, the main changes in gene expression driven by MERS-CoV, SARS-CoV-2 and SARS-CoV infection are divergent at both levels: at the DEGs and the cellular processes, suggesting that each virus uses specific molecular strategies during infection.

Subsequently, we inspected the DEGs with epigenetic or transcriptional regulatory function present among viruses, hereinafter referred as epigenes. A comparative analysis of the DEGs among the three viruses revealed that only *INO80D*, a regulatory component of the chromatin remodeling INO80 complex, is shared among them. MERS-CoV and SARS-CoV only share the histone deacetylase *HDAC9*; while MERS-CoV and SARS-CoV-2 share *DUSP1*, *KDM6B*, *CHD2* and *GADD45A*. Between SARS-CoV and SARS-CoV-2, we found *PBK MYSM1*, *ZNF711* and *PCGF5* (Figure 1C, Supplementary Figure 1C). In addition, given that TFs are also key elements in gene remodeling and regulation, we evaluated the ones differentially expressed across viruses and none of them was affected in all conditions. However, MERS-CoV and SARS-CoV share *ZNF484* and *CEBPD*;

SARS-CoV and SARS-CoV-2 share *ZEB1*, *ZEBTB20*, *NR4A1* and *FOXN2,* and between MERS-CoV and SARS-CoV-2 15 shared TFs were found, including *RELB*, *JUN*, *FOSB*, *E2F8*, among others (Figure 1C, Supplementary Figure 1D).

Furthermore, the analysis showed that the differentially expressed epifactors belong to a wide range of functional categories, such as histone writers, histone readers, histone erases, Polycomb group proteins, chromatin remodeling, DNA modifications, among others (Figure 1D). In addition,regarding differentially expressed TFs, cell lines infected with SARS-CoV-2 show differential expression of TFs of the STAT (mediators of the cellular response to cytokine) and IRF (interferon-regulatory factor) family, which are not differentially expressed in MERS-CoV and SARS-CoV (Figure 1E). We noted that most of the TFs that are differentially expressed and shared between two or more of the Coronaviruses infected cells are members of the Znf TF family (ZNF436, 448, 543, 597, 773, XSCAN12, ZEB1, ZDTB20, KLF10 and HIVEP1) bHLH family (MXD1 and MXD4), involved in CCAAT/Enhancer Binding Protein (C/EBP) (DDIT3 and CEPPD), NF-κB complex (RELB), AP-1 complex (FOSB, JUN), ETS family (SPDEF) and E2F TF, among others (Supplementary Figure 1C).

Since the repeated elements contained in the genome and their expression can also be regulated by epigenetic components, we evaluated the changes in genes expression of repeat elements after viral infection (Supplementary Table 3). We found that 47, 22 and 319 repeat elements are differentially expressed in SARS-CoV-2, SARS-CoV and MERS-CoV infected cell lines, respectively. In SARS-CoV-2, the repeat elements belong predominantly to the Long Interspersed Nuclear Elements (LINE; 13 elements), Long Terminal Repeat (LTR; 17 elements) and DNA repeat elements (17 elements) families. Similarly, for SARS-CoV and MERS-CoV we found LINE (5 and 59 elements), LTR (10 and 179 elements) and DNA repeat (5 and 65 elements). Interestingly, Short Interspersed Nuclear Elements (SINE) elements are not differentially expressed (only 3 elements found in MERS-CoV) and few Satellite elements were identified (3, 1, and 10 in SARS-CoV-2, SARS-CoV and MERS-CoV, respectively) The elements L1MA4:L1:LINE, L1PA8A:L1:LINE and LTR54:ERV1:LTR are shared among all viruses. Notably, the L1 or LINE-1 elements are the only autonomous transposons that remain active in the human genome and are mainly repressed by epigenetic mechanisms such as HPTMs (H3K9me3) [20]. The latter, along with the fact that SARS-CoV-2 infected cells overexpress the histone demethylases KDM7A and KDM6B that target the heterochromatin histone marks such as H3K9me and H3K27me [21] (Supplementary Figure 1C), suggest that SARS-CoV-2 infection could promote an open chromatin conformation, thus affecting the transcriptional expression and the derepression of the repeated sequences.

SARS-CoV-2 transcriptional effect in COVID-19 patient-derived samples

Afterwards, we evaluated the transcriptional response in patient samples infected with SARS-CoV-2 to assess their resemblance to the previously observed results in cell lines. For this purpose, we obtained datasets from samples of bronchoalveolar lavage fluid (BALF) and lung. From the differential expression analysis, we found 389 DEGs shared among both samples (Figure 2A). In this geneset, we identified 28 epigenes whose fold change direction was consistent in most of the cases (Figure 2B); GO enrichment analysis shows that most of the DEGs were related to the immune response to viral infection such as leukocyte mediated immunity and humoral immune response (Figure 2C). Furthermore, they were involved in a wide range of epigenetic processes, such as histone modification, chromatin remodeling, DNA modification and TF (Figure 2D). Following, we evaluated the similarity of these results with the data observed for SARS-CoV-2 in infected cell lines by comparing the overlap between the virus-associated genes with the DEGs present the patients samples (hereinafter referred as patient-DEGs, 389 genes). We found 46, 10 and 22 genes in common with SARS-CoV-2, SARS-CoV and MERS-CoV respectively. In particular, for SARS-CoV-2 infected cell lines and patients, 5 TFs (*STAT5A, MAFF, IRF9, MXD1* and *STAT4)* and no epifactors were identified. The shared DEG between

samples of SARS-CoV-2 infected patients and MERS-CoV and SARS-CoV infected cell lines, which were found at a lesser extent, are likely to be non-specific viral-responding immune genes. Finally, we found that 804 and 20 repeat elements are differentially expressed in BALF and LUNG samples, respectively, being LTR elements the most differentially expressed in both samples (Supplementary Table 3). Collectively, these results show that the gene expression changes promoted by SARS-CoV-2 infection in patients are similar in respiratory tract samples, where immune response processes are the main ones affected.

SARS-CoV-2 and MERS-CoV infection induce different transcriptional fold changes in shared gene co-expression modules, which recapitulate the expression profiles in COVID-19 patient-derived samples

So far, our analyses showed that cell lines and patient samples infected with SARS-CoV-2 exhibited DEGs related to immunological processes, which has been previously described by Blanco-Melo et al. [22] and is congruent with our results. However, differential expression analysis often overlooks the subtle differences in several genes that altogether can be responsible for major changes in global transcriptional regulation. Weighted gene co-expression network analysis overcomes this limitation by studying the expression of thousands of genes in the same analysis [23]. Thus, we expanded our previous results by including a co-expression analysis to identify gene modules associated with each viral infection and genes that play central roles within them.

We constructed the co-expression network with the log2 fold changes of each sample compared to its controls. After identifying the modules, we calculated the correlation between each module and the different traits, where we found that out of the 24 total modules identified, 13 were significantly correlated to the infection of any of the three viruses (Supplementary Figure 2; Supplementary Table 4). Specific modules are associated with MERS-CoV (module 1), SARS-CoV (module 7), and SARS-CoV-2 (module 9 and 10) (Figure 3); shared modules were also identified. Notably, more than half of the modules (7 out of 13) are significantly associated with both SARS-CoV-2 and MERS-CoV; contrary to SARS-CoV-2 and SARS-CoV, which have only one module jointly associated. Remarkably, SARS-CoV-2 and MERS-CoV share a higher number of modules that are significantly associated with each virus. Even though the transcriptional profile is not the same, as it can be seen by the opposing response, the same genes from the shared modules are involved in both infections. Therefore, those infections share more players relevant to the infection than SARS-CoV-2 and SARS-CoV, as it would be expected.

Further analysis with GO enrichment analysis (Supplementary Figure 3), shows that genes in modules 11 and 12 are involved in the host cell response to viral infection, while modules 4, 5, 7, 8, 10 and 13 were associated with intracellular processes used by the viruses during the infection, such as DNA replication, translation, ribosome biogenesis and protein folding. Interestingly, module 6 was found related to epigenetic processes, particularly, transcriptional activation of promoters. These results show that, in addition to the immunological processes identified in our previous differential expression analysis, the transcriptional response to SARS-CoV-2 and MERS-CoV infection contain genes that also participate in RNA translation, DNA replication and epigenetic regulation.

Following, in order to determine the modules that might be more relevant to SARS-CoV, MERS-CoV, or SARS-CoV-2 infection in terms of epigenetic regulation, we conducted enrichment analyses to identify an over-representation of epigenes, virus-associated gene sets, DEGs found in patients (Figure 3).

When integrating these analyses, we found module 1 relevant for MERS-CoV and module 7 for SARS-CoV infection since they are exclusive for these viruses according to the co-expression analysis (Figure 3). To determine the most important epigenes for each of these modules, we evaluated the eigengene-based connectivity (Module Membership, MM) to find hub genes, and the gene significance (GS) of each gene.

Subsequently, we prioritized them by identifying drugs targeting them. After looking for candidate drugs that targeted these central players, we found approved medication for CDK7 and PCNA for SARS-CoV and NCOA1, NR1H2, PRKAB2, CLOCK, KDM1B and ATF2 (Supplementary Figure 4).

Regarding SARS-CoV-2, module 9 was uniquely associated with it, but also other modules standed out. First, module 4 had a consistent behavior between cell lines and patients, since we found it negatively correlated to SARS-CoV-2, while being enriched in downregulated genes in BALF and LUNG samples; in addition, it was enriched in epifactors, suggesting an important role in viral-related epigenetic modifications carried out by these genes. A similar phenomenon is observed for module 12, which was positively correlated with SARS-CoV-2 infection and enriched with upregulated genes in patients and with SARS-CoV-2 DEGs. Additionally, modules 10 and 11 were positively correlated to SARS-CoV-2 and enriched in SARS-CoV-2 DEGs, with module 11 being also enriched with patient's DEGs. Module 6 was negatively associated with SARS-CoV-2 in the co-expression network and enriched with epigenes, and module 8 was enriched with upregulated genes in PBMC and negatively associated with SARS-CoV-2 in the co-expression network, while being enriched with epifactors and SARS-CoV-2-DEGs (Figure 3). Finally, the enrichment of TFs targets in each module was evaluated to identify the ones that could explain the co-expression patterns of the genes within the module. With this analysis, it was found that Module 4 is enriched in the target genes of the transcriptional factors MTA1, MORC2, and RBM34 that belong to the same module (Supplementary Table 5), being MTA1 and RBMC34 differentially expressed in BALF samples. It is worth to mention that in most of these modules epigenes showed a higher MM than the rest of the genes (Module 1: $W = 164249$, $p < 0.05$, Module 2: $W = 1497$, $p < 0.05$, Module 4: $W = 51768$, $p < 0.05$, Module 6: $W = 1359$, $p < 0.05$, Module 7: $W = 3741$, $p < 0.05$, Module 12: $W = 113820$, $p < 0.05$, Module 13: $W = 2182049$, $p < 0.05$), evincing their central role within their modules.

Collectively, these results show that the transcriptional response to infection of SARS-CoV-2 and MERS-CoV involve a higher similarity regarding gene modules but with a different extent of transcriptional change in host cells during infection, which extends our previous observations in the differential expression analysis. Therefore, the same genes in the shared modules play a potential role in both infections, despite presenting a different transcriptional behaviour. Importantly, the virus-correlated co-expression modules either recapitulate the changes in gene expression observed in different COVID-19 patient sample types or are enriched with epifactors, and also contain genes involved in several biological processes related to viral infection, suggesting that the data obtained in the cell lines could recapitulate what was found in infected patients

Protein-protein interaction network analysis provides additional therapeutic alternatives and new targets for drug development for COVID-19

To prioritize epigenes that play a key function in each co-expression module relevant for SARS-CoV-2 (modules 4, 6, 8, 9, 10, 11 and 12), we examined them at the protein-protein interaction (PPI) level in the context of SARS-CoV-2 infection. We constructed a PPI network containing all experimentally validated human protein interactions [24] and the reported virus-host protein interactions from Gordon et al. 2020 and Stukalov et al 2020 [25,26]. Using the virus-human PPI network, we performed *de novo* pathway enrichment analysis with KeyPathwayMiner [27] to extract the largest network using a selection of epigenes as input, while also taking into account the SARS-CoV-2-DEGs and patient-DEGs previously identified. The selection of epigenes for each module was based on their shortest path length with viral proteins, their expression correlation with viral genes and their MM.

All genes contained in the networks identified (Figure 4) provide insights about the molecular machinery involved in SARS-CoV-2 infection, since the genes are either differentially expressed in infected cell lines or patients, or they are hub-epigenes in the co-expression analysis.

For module 4, the network obtained contains mainly epifactors. Notably, DNMT1 directly interacts with the viral protein ORF8 and with TRIM28, to which it is also highly co-expressed. Other relevant epigenes in the networks are SIRT6 (highly co-expressed and interactor of TRIM28), SENP3, MTA1 (a TF whose targets are also enriched in module 4; Supplementary Table V) and BAP1 (differentially expressed in patients). Furthermore, MEPCE, a snRNA methyl phosphate capping enzyme, is differentially expressed in patients and interacts with viral protein NSP8. Module 6 contains BRD4, which directly interacts with E viral protein and is highly co-expressed with EP300, a histone acetyltransferase. Another relevant epigene is SETD1B (related the trimethylation of H3K4[3], a unique epigenetic histone mark related to transcriptional activation), which interacts with TRIM28, present in module 4 [28]. For module 8, notable epigenes are CENPF, differentially expressed in SARS-CoV and SARS-CoV-2 and directly interacting with NSP13; and TOP2A, differentially expressed in SARS-CoV infected cell lines and patients. EP300, an exception node in module 9, interacts with several TFs such as NR2F2, HOXB9, NR3C2 and SOX9. In module 10, RELA (also known as nuclear factor NF-κB p65 subunit) and MOV10 are exception nodes that interact with the TFs SMAD3, ZNF277 and UBE2D3. Also, viral proteins ORF7B and ORF3 interact with FXYD2, STEAP1B and TMEM156, which are differentially expressed in SARS-CoV-2 infected cell lines. For module 11, IRF7 (interferon regulatory factor 7) and STAT5A interact with EP300. Module 12, also contains epigenes of interest, such as MOV10 (Putative helicase MOV-10) that interacts with the N protein, TRIM25, RELA and TLE1, which has a direct interaction with viral protein NSP13. Further genes which are classified as hub-epigenes and are also differentially expressed in cell lines or patients are FOS, CEBPD, NR4A1, PRDM1, PCGF5, ZNF652, IRF2, and ZEB2.

Briefly, we identified relevant TFs known to participate in Coronavirus infection and support the veracity of our results, such as TFs from the STAT family (STAT1, STAT2, STAT5A), interferon regulatory factors (IRF7, IRF2), cytokines (CCL3, CCL4, IL1B), and FOS and JUND, members of the AP-1 complex [29]. However, we also identify important genes that appear to be drivers of SARS-CoV-2 infection; such as the epifactors MOV10 and EP300 and the TF RELA, since they are exception connectors (genes that do not belong to the specific module, but are important in the protein pathway found) in more than one module (modules 6, 9, 10 and 11), and belong to modules enriched in genes that participate in histone H3-K4 methylation and in the response to interferon gamma. EP300 is a histone acetyltransferase that was also identified in SARS-CoV-2 infected cell lines [14]. Additionally, we found that MOV10, a putative helicase, also participates in SARS-CoV-2 infection. The TF RELA has been increasingly recognized as a crucial modulator of the response to SARS-CoV-2 infection [14,30] and is part of the NF-κB complex, along with RELB [31], which is differentially expressed in MERS-CoV and SARS-CoV-2 infected cell lines (Supplementary Figure 1D). TRIM25 is a ubiquitin ligase required for production of INF-1 and is inhibited by Nucleocapsid of SARS-CoV [32]. Finally TRIM28 (also known as KAP1) has been shown to interfere with viral integration into host genome [33] and represses the expression of repeat elements of the LINE family, in particular L1NA4 [21], which was previously identified as differentially expressed in cell lines infected with the three Coronaviruses.

Afterwards, we evaluated whether the proteins in the networks had annotated drugs targeting them. We found drugs for 69 out of the total 260 proteins, being PLAU, RELA, NEK6, NR1H4, PTGS2, PRKDC, ESR1, NR3C2, TTK, TOP2A, ADRB2, HDAC4, TRIM25, STK10, RPS6KA5 and EP300 the ones with the most drugs identified (more than 20). A total of 799 drugs were found, where Erlotinib, Imatinib, Lapatinib, Sunitinib, S-adenosyl-L-homocysteine, Quercetin, Tandutinib, RAF-265, Pictilisib, Neratinib and Fedratinib are the drugs with more targets (more than 5; Supplementary Table 6). Relevant epigenes that have associated medication are shown in Table I.

Most notably, RELA is targeted by SC-236, Bortezomib, Indoprofen (an anti-inflammatory) and Betulinic Acid, whose derivatives show anti-HIV activity [34,35]. EP300 is targeted by curcumin, a molecule with anti-inflammatory

properties [36]. The latter proposes RELA and EP300 as new potential drug target candidates for SARS-CoV-2 infection, not only because they participate in immune-related processes, but also because they belong to the cellular epigenetic machinery used by the virus during infection. Furthermore, self-evident immune-related targets STAT5A, STAT1 and FOXN2 are also good candidates for treatment. Finally, the proteins MOV10, TRIM25 and TRIM28 do not have associated drugs, thus they are good candidates for drug development, as well as other relevant epigenes shown in Table II.

Together, network analysis at the protein level allowed the identification of several epigenes that are part of the molecular machinery used by the virus during infection (Figure 5). Epigenes that participate in immune response through different mechanisms (response to interferon or NF-kB complex) are among the main genes identified and are evident drug target candidates for COVID-19 because they already have associated drugs targeting them (such as STAT5A and STAT1). Furthermore, new candidate druggable epigenes were also identified, notable examples are EP300 and RELA, which are targeted by drugs with anti-inflammatory or antiviral properties; and TRIM25, TRIM28 and MOV10, which are good candidates for drug development.

DISCUSSION

Cells are in constant adaptation with their environment, in fact they can sense and respond to different stimuli by changing their transcriptional patterns. This cellular plasticity allows cells to adapt almost immediately to insults, including virus infections [37]. Epigenetic proteins and TFs are one of the main elements involved in the transcriptional response of cells during viral infection. These elements can be used as protein targets for drug identification and treatment. In this work we aimed to identify key TFs and proteins involved in the epigenetic response to viral infection of SARS-CoV-2, SARS-CoV and MERS-CoV by integrating co-expression and de *novo* pathway enrichment analyses. Therefore, our study focused on the infection part of COVID-19, which is relevant mostly during the early stages of the disease, in contrast to the immune pathologies seen in the later ones.

One of our main findings is that the transcriptional response (regarding DEGs and significantly co-expressed modules) induced by SARS-CoV-2 and MERS-CoV involves a higher similarity regarding gene players and biological processes than SARS-CoV-2 and SARS-CoV, despite presenting a different transcriptional behaviour. However, it is interesting to notice that regarding the transcriptional trend of the modules (i.e. correlation sign), SARS-CoV-2 and SARS-CoV behave more similarly despite many modules not being significantly associated with SARS-CoV. Nevertheless, unique modules, patterns and DEG were found in each CoV. Despite they belong to the coronavirus family, each one has unique characteristics that could influence its pathogenicity and virulence. This finding agrees with a recent study that has found specific biological process deregulations in SARS-CoV-2 infected cell lines, which are not found in other CoVs [15]. In addition, different transcriptional change patterns have been observed between MERS-CoV and SARS-CoV during the infection; these changes are not recapitulated by phylogenetic relationships since, in some groups of genes, MERS-CoV-infected transcriptional behavior appears to be more similar to the more remotely related influenza H5N1 virus infection [13].

Furthermore, the contrasting transcriptional response induced by the infection of SARS-CoV-2 and MERS-CoV in several modules suggests that genes in those modules participate in both viral infections but with a different mechanism, which leads to distinct pathways of infection that could explain the dissimilar phenotypes observed in both diseases. Divergent fold change trends, such as the ones described in this study, have been previously described in MERS-CoV and SARS-CoV infections to limit the host type I interferon (IFN-I) response, where predominant active and repressive epigenetic marks in involved genes are the opposite between both CoVs [13]. In our study, we present a list of epigenes and biological processes whose fold change trend is the

opposite between MERS-CoV and SARS-CoV-2; further investigation on them could shed light on the mechanisms responsible for the differences in pathogenesis and outcome of both viral infections.

We further identify at the protein interaction level, that several TFs take part mainly in the immunological response to viral infection. One example is NF-κB, whose p65 subunit (also known as *RELA*) is a central part in the protein interaction network for SARS-CoV-2. NF-κB induces the expression of several pro-inflammatory cytokines, including IL-6, CCL2 and CCL3 [38], which had been found in high levels in COVID19 patients [39]. On the other hand, TRIM25, an ubiquitinase, is essential for the activation of NF-κB and the production of IL-6 [40]. TRIM25 is over-expressed in cell lines infected with SARS-CoV-2 but not in those with MERS-CoV, which furthermore suggest that NF-κB could be a medullary part of the host immune response against SARS-CoV-2. The previous observation is reinforced by the fact that it was observed that RELA directly interacts with histone acetyltransferase EP300, and both proteins interact with various components of the AP-1 complex such as FOS, JUND, and FOSL1. AP-1. EP300 and NF-κB regulate chromatin accessibility in the proximal promoter region of IL-6 and CCL2, both pro-inflammatory cytokines [41,42]. The p300/CBP complex is one of the best characterized cofactors of NF-κB and specifically binds RELA and acetylates it along with the surrounding histones [31]. It is known that adults older than 65 years have higher NF-κB levels compared to younger adults [43] and some authors had suggested that this may be one reason older adults are more susceptible to develop the severe form of COVID-19 [44].

According to our results, SARS-CoV-2 infection modifies the expression of several TFs of the interferon regulatory factor (IRF) and STAT families, which are primarily involved in the immune response against pathogens. STAT1 and STAT2 are key elements of the signaling induced by type I interferons, these proteins form a dimer upon interferon mediated phosphorylation and, together with IRF9, form the complex ISGF3 that activates the transcription of interferon stimulated genes [45]. Our results also showed that IRF9 is upregulated in cell lines infected with SARS-CoV-2; however, module 12's interactome showed that STAT2 and STAT1 interact with IRF2. IRF2 is a negative regulator of IFNα and its inhibition causes an increase in the antiviral response induced by IFNα [46]. This fact further suggests an impairment of interferon type I stimulated genes activation, as previously described as a hallmark of SAR-CoV-2 infection [47]. On the other hand, IRF1 and IRF7 were also upregulated in SARS-CoV-2 infected cell lines. IRF7 is a key TF for IFNα expression, and it has been previously identified as a hub gene for SARS-CoV-2 infection together with IFR9 and STAT1 [48]. It is also interesting that IRF7 loss of function mutations were associated with severe COVID-19 patients [49] and with the development of life-threatening influenza in children [50] which suggest that inhibition of IRF7 activity is crucial for SARS-CoV-2 pathology.

Viruses have been reported to use epigenetic machinery to take advantage of the cell and hijack its regulatory capacity for their own benefit [13]. The epigenetic machinery can be affected by coronaviruses in this same sense, and this can happen either by promoting alterations in the epigenetic code, such as DNA methylation and post-translational modifications of histones, or directly by promoting the dysregulation of enzymes and other proteins associated with the epigenome.

We found that among the deregulated epifactors with histone acetylation function are HDAC9 and SIRT1 enzymes. In this sense, it has recently been reported that the SIRT1 protein (a class 3 HDAC) was positively regulated in the lung of patients with severe COVID-19 comorbidities [51]. Likewise, another work demonstrated that under conditions of cellular energy stress, SIRT1 can epigenetically regulate the ACE2 receptor [52]. Also, it has been observed that treatment with non-steroidal anti-inflammatory drugs can inhibit SIRT1 activity, which in turn could affect ACE2 expression [53]. Accordingly, it has been postulated that in some diseases where the epigenetic dysregulation is implicit (such as lupus) the entrance of SARS-CoV-2 into the host cells may be facilitated [54].

Interestingly, the enzymes HAT1, HDAC2 and KDM5B have been reported to also potentially regulate ACE2 in human lungs. KDM5B has gained interest, because it is associated with other viral infections such as the hepatitis B virus [55], and potentially with SARS-CoV-2 [51]. Remarkably, in breast cancer cells, it has been shown that inhibition of this enzyme triggers a robust interferon response that results in resistance to infection by DNA and RNA viruses [56]. In this regard, we observed several deregulated KDMs in the different coronavirus infections, in which KDM6B stands out by being deregulated in both MERS-CoV and SARS-CoV2 infection. KDM6B is a specific demethylase of H3K27me3, which acts as a repressive histone mark. Although it remains to be fully studied, it is associated with the regulation of a wide range of genes involved in inflammatory agents, development, cancer, viral infection response, senescence and is an important host response against environmental, cellular stress [57]. Therefore, adding to the above, it is suggested that demethylases, such as KDM6B, are potential epigenes that are affected during SARS-CoV-2 infection and can be presented as potential targets for the treatment of COVID-19. However, this should be further studied.

Several epigenes previously involved in response to viral infections stood out in our protein interaction analysis, such as BRD4, TOP2A, and TRIM28. Bromodomain protein 4 (BRD4) is a histone acetylation reader and writer that plays an important role in DNA replication, transcription, and DNA repair [58]. This epigene is critical for the maintenance of the higher-order chromatin structure, since its inhibition leads to chromatin decondensation and fragmentation, and it also can stimulate innate antiviral immunity [58]. BRD4 complexes with RELA and CDK9 and is functionally required for effective activation of NF-kB-dependent immediate early cytokine genes in response to viral patterns. In this sense, our results show a protein-protein interaction with EP300, which involves the p300 / CBP complex, one of the best characterized cofactors of NF-kB and binds specifically to RELA [59], validating the possible importance of this system in infection with SARS-CoV2. Examples like this suggest that the virus, through these epigenetic remodelers, promotes chromatin remodeling that could lead to opening, both at the local and global level. Accordingly, an indicator of global changes is the increased expression of transcripts from repeated sequences such as LINE1. If this is so, then the virus is manipulating the chromatin aperture to promote the expression of genes that support its invasion. In this regard, other work has suggested the importance of LINE1 elements. Where these types of repetitive elements are very relevant in gene regulation, especially when these elements are in proximity to neighboring genes, since they could alter their expression. Therefore, the dysregulation of repeated elements such as LINE1 could indirectly change the cellular transcriptome [60].

Furthermore, we find epigenes that interact with the viral proteins directly or very closely. This connection suggests a virus-promoted modulation to affect the epigenome of the host cell's interactome. Which reinforces the idea that the virus strategy is partly to take advantage of the epigenetic machinery. In general, our data suggest that the SARS-CoV-2 infection deregulates the epigenetic master machinery of the host cell. One of the points that should be taken into consideration in the future is that if this epigenetic machinery is not re-established after disease courses it could generate other diseases such as cancer in the long term. This is based on the fact that many of the genes that we found in our study have been proposed as epigenetic hallmarks in various neoplasms.

Our last key finding is the identification of driver epigenetic proteins and TFs involved in SARS-CoV-2 infection that can be targeted by existing drugs. We identified S-adenosyl-L-homocysteine (SAH) targeting several epigenetic components of the host response to SARS-CoV-2 infection. SAH is the product of the chemical reaction performed by methyltransferases using nucleic acids or proteins as substrates, and has been previously suggested as a potential treatment for viral infections such as ZIKA, MERS-CoV and SARS-CoV [61–64] due to its inhibitory activity of the viral RNA cap 2′-O-methyltransferase, formed by the NSP16-NSP10 complex [65,66]. Furthermore, given the interaction between DNMT1 and ORF8 at the protein level, SAH could potentially work against SARS-CoV-2 infection, not only by inhibiting the methyltransferase activity of NSP16-NSP10, but also by directly modulating the activity of the key host proteins involved in the

transcriptional response to infection or by interfering with the interactions observed between ORF8 and DNMT1.

Furthermore, as anticipated, many proteins with epigenetic functions involved in SARS-CoV-2 infection have kinase activity and can be targeted by kinase inhibitors. One important example is imatinib, which we identified as a potential drug for SARS-CoV-2 and SARS-CoV, and is currently undergoing clinical trials to evaluate its efficacy in COVID-19 patients (NCT04394416, NCT04422678, NCT04346147 and NCT04357613; www.clinicaltrials.gov). Similarly, we found quercetin targeting several epifactors with kinase activity. Quercetin is a plant-derived compound with anti-inflammatory and antiviral effects [67,68] that has been evaluated in clinical trials as a dietary supplement or prophylaxis for COVID-19 (NCT04578158, NCT04377789 and NCT0446813). Even though some independent studies show no clear evidence of its effectiveness, preliminary data shows that it could be effective to decrease the frequency and duration of respiratory tract infections [69–71]. It is worth mentioning that these drugs are being tested in clinical trials based on their described inhibitory activity of enzymes related to the activation of immune response and inflammation, such as growth receptors [72]. The latter, together with our results, suggests that drugs targeting epigenetic mechanisms could be also effective to treat SARS-CoV-2 by modulating their kinase activity.

Finally, we also identified Bortezomib and betulinic acid associated with RELA. Bortezomib is a proteasome inhibitor that has been proposed as COVID-19 therapy given its capacity to inhibit (although only marginally) the papain-like protease (NSP3) of SARS-CoV, which also has deubiquitinase activity [73–75]. Likewise, betulinic acid has been proposed as a target of NSP3 in SARS-CoV-2 [76].

Together, we have supporting evidence that current drug-based therapies to treat COVID-19 also target the transcriptional response to infection by the modulation of the epigenetic proteins identified in this study. Furthermore, we provide additional new potential drug targets and drug candidates which could be effective and whose potential use has not been exploited yet. These results provide comprehensive evidence that epigenetic therapy could aid in restoring the transcriptional changes observed during infection. By using epigenetic drugs, a therapeutic effect can be achieved due to their systemic effects, which can be advantageous to treat a disease that targets different tissues and cellular mechanisms, as observed in COVID-19.

In this study, we used a blend of bioinformatic approaches to comparatively analyze transcriptomic data from SARS-CoV-2, SARS-CoV and MERS-CoV infected pulmonary cell lines and COVID-19 patient-derived samples. In particular, we focused on the epigenetic processes and transcriptional factors, since these have been widely proposed as the master regulators of the expression of most genes. We found that the transcriptional response to infection of SARS-CoV-2 and MERS-CoV is more similar to that observed for SARS-CoV regarding shared significantly associated gene modules; however, the transcriptional change elicited by MERS-CoV and SARS-CoV seems to be opposite. At the same time, we identified specific altered modules in the response to infection with SARS-CoV2 that could serve as a guide for the proposal of different therapeutic strategies based on epigenetic therapy. Thus, our results add a piece to the puzzle of the strategies used by the different coronaviruses to manipulate the gene regulation capacity of the cell. Although the pathways are differential between them, the virus objective is to take advantage of the TFs and various chromatin remodelers to avoid being detected and prevail in the invasion. This is a very fine strategy that the virus uses and it has been poorly studied in both its biological importance and its future therapeutic application. This could open a new window of opportunities for treatment and thus close the chapter on this pandemic disease.

METHODS

Data processing and differential expression analysis

Raw sequencing data was trimmed with Trimmomatic version 0.39 [77] using the parameters ILLUMINACLIP 2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36; and the quality of reads was evaluated with FastQC version 0.11.9 [78]. Technical replicates (when existing) were merged and each biological replicate was aligned to the GRCh38 v33 human genome with STAR version 2.7.3 [79] using the mapping parameters suggested in Jin et al. [80]: (--outFilterMultimapNmax 100 --winAnchorMultimapNmax 100). To estimate the abundance of the transcripts accounting for coding and non-coding genes as well as repetitive elements, we used TETranscripts version 2.1.4 [80] with the multi-mode. Raw count tables were used for differential expression analysis using DESeq2 [81]. DEGs were identified with a p adj. < 0.05 and abs($\log_2$ fold change) > $\log_2(1.5)$.

Viral transcripts quantification

Viral transcriptome was constructed with the 11 gene sequences reported in SARS-CoV-2 genome (NCBI Reference Sequence NC_045512.2). Viral transcript expression was quantified in each trimmed RNA-seq file of SARS-CoV-2 infected samples with Salmon v 1.3.0 [82].

Virus and patient DEGs

Virus-associated gene sets were obtained with the intersection of DEGs identified in all the cell lines infected with the corresponding virus, except for SARS-CoV-2. For SARS-CoV, the intersection between the cell lines infected consisted of 182 genes (SARS-CoV-DEGs); for MERS-CoV the intersection was 1139 genes (MERS-CoV-DEGs); and for SARS-CoV-2, the intersection between at least 3 out of the 4 cell lines was used instead and consisted in 909 genes (SARS-CoV-2-DEGs) (Supplementary Table2). Patient-associated gene set was obtained with the shared DEGs in lung and bronchoalveolar lavage fluid (BALF) conditions (389 genes, patient-DEGs) (Supplementary Table 2).

Epigenes catalogue

To build the Epigenes catalogue, 4 different databases were used: EpiFactors [83], Histome [84], dbEM [85] and the manually curated TF list from Lambert et al. [86]. TFs' functional annotation was taken from Lambert et al. [86]. The final list consisted of 2161 genes (776 epifactors, 1348 TFs and 41 categorized as both TF and epifactor).

Co-expression analysis

Count matrices of the analyzed cell lines were filtered to remove low-expressed genes using the function filterByExpr from edgeR [87] while accounting for the treatment (i.e. virus infection) and cell type in the filtering design. Following, normalization of gene counts was performed with vst function from DESeq2[81] (treatment and cell type of each sample were included in the design matrix and accounted for these effects with the blind argument). The gene co-expression network was built with the $\log_2$ fold changes ($\log_2$FC) of each biological sample compared with the controls of the same biological condition by applying the formula (1).

(1) $\log_2 FC_i = \log_2(SC_i / ACC_i)$

Where *SC* and *ACC* correspond to the normalized counts of gene *i* in the infected and controls samples respectively. The resulting matrix containing the log2FoldChanges per sample was used to construct the weighted gene co-expression network with the WGCNA package [23]. A soft threshold of 9 was used to construct the network and modules were identified with a minimum size of 20. Modules whose expression was similar were merged using a dissimilarity threshold of 0.25, resulting in a total of 24 modules. Finally, the module-eigengene pearson correlation of each module with the viruses was tested.

Enrichment analysis

Gene Ontology (GO) enrichment analyses were performed using clusterProfiler [88] in virus associated and patient gene sets. For the differential expression analyses of infected cell lines, the enrichment of GO terms in DEGs was tested using the expressed genes on each particular comparison as background. For the co-expression network, the enrichment of GO terms was tested in each module using the genes of the full network as background.
Epigenes, virus-associated DEGs and TF-target enrichment analyses were performed with gProfiler2 [89] using a custom gmt file or the TRANSFAC database included in the package for TF-target enrichment. The correction method used was g:SCS and an adjusted p-value significance threshold of 0.05. As background, all the genes annotated in the co-expression network were used for epigenes and TF-target enrichment and the expressed genes in each virus for virus-associated DEG enrichment.

Co-expression module selection
SARS-CoV-2 modules were selected from the co-expression analysis based on whether they were uniquely and significantly associated with SARS-COV-2 in the co-expression analysis. If they were not uniquely associated with SARS-CoV-2, the modules enriched with at least one dataset (DEG, patient-DEG or Epigenes) were selected. Based on these criteria, modules 4, 6, 8, 9, 10, 11 and 12 were selected. MERS-CoV and SARS-CoV modules were selected on whether they were uniquely associated with each specific virus in the co-expression analysis. Module 1 was selected for MERS-CoV and module 7 for SARS-CoV. SARS-CoV-2 selected modules were further analyzed, as described in the following sections.

Virus-host network construction
Virus-human interactions were obtained from Gordon et al. [25] and Stukalov et al. [26]. The human protein-protein interaction network (PPI) was obtained from IID version 2018-11 [24] using only the experimentally validated interactions (“exp”, "exp;ortho", "exp;ortho;pred" or "exp;pred"). After homogenizing the viral protein nomenclature, the three sources of interactions were merged to create the entire virus-human PPI, followed by the removal of duplicated edges and self-loops. The final integrated network contained 30 viral nodes, 17524 human nodes and 329054 edges. The mapping of viral transcript counts to viral proteins in the PPI was based on the reference sequence annotation (NCBI Reference Sequence NC_045512.2) and the data provided in Supplementary Data from Gordon et al. 2020 [25].

Epigene selection
For co-expression modules 4, 6, 8, 10, 11 and 12, relevant epigenes were selected based on whether they satisfied at least one of the following criteria: (1) its shortest path length with viral proteins, (2) the correlation value between its expression and the expression of viral proteins and (3) its module membership (MM) value, a measure of the correlation between a gene expression profile and the module eigengene, which is highly related to the intramodular connectivity, and gene significance (GS) the correlation of a gene with an external trait (viral infection) [90].

1. The shortest path length was calculated between all pairs of viral proteins and human proteins in the PPI network with the igraph package version 1.0.0 [91]. The retained epigenes were the ones whose shortest path length with at least one viral protein was less than 3.
2. Pearson’s correlation coefficient was computed between the count values of viral transcripts and count values of epigenes in infected cell lines. Epigenes with p value < 0.05 and abs(correlation_estimate) > 0.5 with at least one viral transcript were selected.
3. Epigenes with abs(MM) > 0.8 in the corresponding module of the co-expression network were retained.

For modules 1 and 7, epigenes with abs(MM) > 0.8 and abs(GS) > 0.3 were selected.

***De novo* pathway enrichment**

De novo pathway enrichment analysis for co-expression modules 4, 6, 8, 10, 11 and 12 was performed with KeyPathwayMiner [27], the built virus-human PPI network, the full list of viral proteins as positive nodes and a customized input indicator matrix for each module containing as active genes those which belonged to any of the following categories: (1) it was a SARS-CoV-2-DEG, (2) it was a patient-DEG or (3) it was an epigene selected as described above. The parameters used for all the analyses were the Greedy search algorithm, INES search strategy, remove border exception nodes, L=0, and K=0 for modules 4 and 12, K=2 for module 6, and K=3 for modules 8, 9, 10 and 11.

Drug identification

All approved and non-approved drugs targeting the genes/proteins contained in each network were obtained with CoVex[92] by mapping the gene names to uniprot IDs, using the closeness centrality algorithm and the following parameters: result size= 50000, disabled hub penalty, disabled max degree, include indirect drugs=FALSE and include non-approved drugs=TRUE. The latter parameters ensure the retrieval of all drugs associated with the input genes. A total of 265 out of 277 genes mapped to the CoVex database.

DATA AVAILABILITY STATEMENT

Raw RNA-seq data was obtained from the Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) of the National Center for Biotechnology Information (NCBI), U.S. National Library of Medicine, and the Genome Sequence Archive in BIG Data Center (bigd.big.ac.cn/), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (Supplementary Table 1).

ACKNOWLEDGEMENTS

This work was supported by Apoyo para proyectos de investigación científica, desarrollo tecnológico e innovación en salud ante la contingencia por COVID-19, CONACyT [00312021 to ESR], Fondo CB-SEP-CONACyT [284748 to ESR] and DSA-SEP (PRODEP) [47310681, id-250690 to ESR]. MSA and ADMM are doctoral students in the "Programa de Doctorado en Ciencias Bioquímicas, UNAM" and received a fellowship funding from CONACYT (MSA CVU659273 and ADMM CVU894530). MSA was awarded by the German Academic Exchange Service, DAAD (ref. 91693321). NA would like to acknowledge the Independent Research Fund Denmark (6108-00038B). ESR was supported by the Departamento de Ciencias Naturales, UAM-Cuajimalpa. JB's work was supported by his VILLUM Young Investigator Grant nr. 13154. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

COMPETING INTERESTS

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

MSA, EIND and ADMM equally contributed to the data collection, bioinformatic analyses and manuscript writing. NA and JB provided critical feedback and helped to improve the manuscript. RGB and ESR were in charge of overall direction, planning, and supervision. MSA, EIND and ADMM contributed equally to this work.

REFERENCES

1. Payne, S. Chapter 17 - Family Coronaviridae. in *Viruses* (ed. Payne, S.) 149–158 (Academic Press, 2017).

2. Memish, Z. A., Perlman, S., Van Kerkhove, M. D. & Zumla, A. Middle East respiratory syndrome. *Lancet* **395**, 1063–1077 (2020).
3. Hui, D. S. C. & Zumla, A. Severe Acute Respiratory Syndrome: Historical, Epidemiologic, and Clinical Features. *Infect. Dis. Clin. North Am.* **33**, 869–889 (2019).
4. Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J. & Hsueh, P.-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *Int. J. Antimicrob. Agents* **55**, 105924 (2020).
5. World Health Organization. Coronavirus disease (COVID-19) pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (2020).
6. Youssef, N., Budd, A. & Bielawski, J. P. Introduction to Genome Biology and Diversity. *Methods Mol. Biol.* **1910**, 3–31 (2019).
7. Marazzi, I. *et al.* Suppression of the antiviral response by an influenza histone mimic. *Nature* **483**, 428–433 (2012).
8. Flanagan, J. M. Host epigenetic modifications by oncogenic viruses. *Br. J. Cancer* **96**, 183–188 (2007).
9. Menachery, V. D. *et al.* MERS-CoV and H5N1 influenza virus antagonize antigen presentation by altering the epigenetic landscape. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1012–E1021 (2018).
10. Marcos-Villar, L., Pazo, A. & Nieto, A. Influenza Virus and Chromatin: Role of the CHD1 Chromatin Remodeler in the Virus Life Cycle. *J. Virol.* **90**, 3694–3707 (2016).
11. Marcos-Villar, L. *et al.* Epigenetic control of influenza virus: role of H3K79 methylation in interferon-induced antiviral response. *Sci. Rep.* **8**, 1230 (2018).
12. O'Connor, C. M., DiMaggio, P. A., Jr, Shenk, T. & Garcia, B. A. Quantitative proteomic discovery of dynamic epigenome changes that control human cytomegalovirus (HCMV) infection. *Mol. Cell. Proteomics* **13**, 2399–2410 (2014).
13. Menachery, V. D. *et al.* Pathogenic influenza viruses and coronaviruses utilize similar and contrasting approaches to control interferon-stimulated gene responses. *MBio* **5**, e01174–14 (2014).
14. Ochsner, S. A., Pillich, R. T. & McKenna, N. J. Consensus transcriptional regulatory networks of coronavirus-infected human cells. *Sci Data* **7**, 314 (2020).
15. Singh, K. *et al.* Network Analysis and Transcriptome Profiling Identify Autophagic and Mitochondrial Dysfunctions in SARS-CoV-2 Infection. *Preprint at [https://www.biorxiv.org/content/10.1101/2020.05.13.092536v2]* (2020).
16. Ganesan, A., Arimondo, P. B., Rots, M. G., Jeronimo, C. & Berdasco, M. The timeline of epigenetic drug discovery: from reality to dreams. *Clin. Epigenetics* **11**, 174 (2019).
17. WHO Solidarity Trial Consortium. Repurposed Antiviral Drugs for Covid-19 - Interim WHO Solidarity Trial Results. *N. Engl. J. Med.* **384**, 497–511 (2021).
18. Rayner, R. E., Makena, P., Prasad, G. L. & Cormet-Boyaka, E. Optimization of Normal Human Bronchial Epithelial (NHBE) Cell 3D Cultures for in vitro Lung Model Studies. *Sci. Rep.* **9**, 500 (2019).
19. Davis, A. S. *et al.* Validation of normal human bronchial epithelial cells as a model for influenza A infections in human distal trachea. *J. Histochem. Cytochem.* **63**, 312–328 (2015).
20. Bulut-Karslioglu, A. *et al.* Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. *Mol. Cell* **55**, 277–290 (2014).
21. Castro-Diaz, N. *et al.* Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev.* **28**, 1397–1409 (2014).
22. Blanco-Melo, D. *et al.* Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181**, 1036–1045.e9 (2020).
23. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
24. Kotlyar, M., Pastrello, C., Malik, Z. & Jurisica, I. IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* **47**, D581–D589 (2019).

25. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
26. Stukalov, A., Girault, V., Grass, V., Bergant, V. & Karayel, O. Multi-level proteomics reveals host-perturbation strategies of SARS-CoV-2 and SARS-CoV. *Preprint at [https://www.biorxiv.org/content/10.1101/2020.06.17.156455v1]* (2020).
27. Alcaraz, N. *et al.* Robust de novo pathway enrichment with KeyPathwayMiner 5. *F1000Res.* **5**, 1531 (2016).
28. Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J., 3rd. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* **16**, 919–932 (2002).
29. Hess, J., Angel, P. & Schorpp-Kistner, M. AP-1 subunits: quarrel and harmony among siblings. *J. Cell Sci.* **117**, 5965–5973 (2004).
30. Fagone, P. *et al.* Transcriptional landscape of SARS-CoV-2 infection dismantles pathogenic pathways activated by the virus, proposes unique sex-specific differences and predicts tailored therapeutic strategies. *Autoimmun. Rev.* **19**, 102571 (2020).
31. Bhatt, D. & Ghosh, S. Regulation of the NF-κB-Mediated Transcription of Inflammatory Genes. *Front. Immunol.* **5**, 71 (2014).
32. Hu, Y. *et al.* The Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Inhibits Type I Interferon Production by Interfering with TRIM25-Mediated RIG-I Ubiquitination. *J. Virol.* **91**, e02143–16 (2017).
33. Fehervari, Z. Putting a KAP on infection. *Nat. Immunol.* **12**, 816 (2011).
34. Pavlova, N. I., Savinova, O. V., Nikolaeva, S. N., Boreko, E. I. & Flekhter, O. B. Antiviral activity of betulin, betulinic and betulonic acids against some enveloped and non-enveloped viruses. *Fitoterapia* **74**, 489–492 (2003).
35. Aiken, C. & Chen, C. H. Betulinic acid derivatives as HIV-1 antivirals. *Trends Mol. Med.* **11**, 31–36 (2005).
36. Gupta, S. C., Patchva, S. & Aggarwal, B. B. Therapeutic roles of curcumin: lessons learned from clinical trials. *AAPS J.* **15**, 195–218 (2013).
37. Bollati, V. & Baccarelli, A. Environmental epigenetics. *Heredity* **105**, 105–112 (2010).
38. Beacon, T. H., Su, R.-C., Lakowski, T. M., Delcuve, G. P. & Davie, J. R. SARS-CoV-2 multifaceted interaction with the human host. Part II: Innate immunity response, immunopathology, and epigenetics. *IUBMB Life* **72**, 2331–2354 (2020).
39. Chu, H. *et al.* Comparative Replication and Immune Activation Profiles of SARS-CoV-2 and SARS-CoV in Human Lungs: An Ex Vivo Study With Implications for the Pathogenesis of COVID-19. *Clinical Infectious Diseases* vol. 71 1400–1409 (2020).
40. Liu, Y. *et al.* TRIM25 Promotes TNF-α–Induced NF-κB Activation through Potentiating the K63-Linked Ubiquitination of TRAF2. *The Journal of Immunology* **204**, 1499–1507 (2020).
41. Wolter, S. *et al.* c-Jun Controls Histone Modifications, NF-κB Recruitment, and RNA Polymerase II Function To Activate the ccl2 Gene. *Mol. Cell. Biol.* **28**, 4407–4423 (2008).
42. Ndlovu, M. N. *et al.* Hyperactivated NF-{kappa}B and AP-1 transcription factors promote highly accessible chromatin and constitutive transcription across the interleukin-6 gene promoter in metastatic breast cancer cells. *Mol. Cell. Biol.* **29**, 5488–5504 (2009).
43. Bektas, A. *et al.* Age-associated changes in basal NF-κB function in human CD4+ T lymphocytes via dysregulation of PI3 kinase. *Aging* **6**, 957–974 (2014).
44. Do, L. A. H., Anderson, J., Mulholland, E. K. & Licciardi, P. V. Can data from paediatric cohorts solve the COVID-19 puzzle? *PLoS Pathog.* **16**, e1008798 (2020).
45. Martinez-Moczygemba, M., Gutch, M. J., French, D. L. & Reich, N. C. Distinct STAT Structure Promotes Interaction of STAT2 with the p48 Subunit of the Interferon-α-stimulated Transcription Factor ISGF3. *J. Biol. Chem.* **272**, 20070–20076 (1997).
46. Robichon, K. *et al.* Identification of Interleukin1β as an Amplifier of Interferon alpha-induced Antiviral Responses. *PLoS Pathog.* **16**, e1008461 (2020).

47. Hadjadj, J. *et al.* Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* **369**, 718–724 (2020).
48. Prasad, K. *et al.* Targeting hub genes and pathways of innate immune response in COVID-19: A network biology perspective. *Int. J. Biol. Macromol.* **163**, 1–8 (2020).
49. Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, eabd4570 (2020).
50. Ciancanelli, M. J. *et al.* Infectious disease. Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency. *Science* **348**, 448–453 (2015).
51. Pinto, B. G. G. *et al.* ACE2 Expression Is Increased in the Lungs of Patients With Comorbidities Associated With Severe COVID-19. *J. Infect. Dis.* **222**, 556–563 (2020).
52. Clarke, N. E., Belyaev, N. D., Lambert, D. W. & Turner, A. J. Epigenetic regulation of angiotensin-converting enzyme 2 (ACE2) by SIRT1 under conditions of cell energy stress. *Clin. Sci.* **126**, 507–516 (2014).
53. Dell'Omo, G. *et al.* Inhibition of SIRT1 deacetylase and p53 activation uncouples the anti-inflammatory and chemopreventive actions of NSAIDs. *Br. J. Cancer* **120**, 537–546 (2019).
54. Sawalha, A. H., Zhao, M., Coit, P. & Lu, Q. Epigenetic dysregulation of ACE2 and interferon-regulated genes might suggest increased COVID-19 susceptibility and severity in lupus patients. *Clin. Immunol.* **215**, 108410 (2020).
55. Wang, X. *et al.* Hepatitis B virus X protein induces hepatic stem cell-like features in hepatocellular carcinoma by activating KDM5B. *World J. Gastroenterol.* **23**, 3252–3261 (2017).
56. Wu, L. *et al.* KDM5 histone demethylases repress immune response via suppression of STING. *PLoS Biol.* **16**, e2006134 (2018).
57. Zhang, X., Liu, L., Yuan, X., Wei, Y. & Wei, X. JMJD3 in the regulation of human diseases. *Protein Cell* **10**, 864–882 (2019).
58. Wang, J. *et al.* BRD4 inhibition exerts anti-viral activity through DNA damage-dependent innate immune responses. *PLoS Pathog.* **16**, e1008429 (2020).
59. Tian, B. *et al.* BRD4 Couples NF-κB/RelA with Airway Inflammation and the IRF-RIG-I Amplification Loop in Respiratory Syncytial Virus Infection. *J. Virol.* **91**, e00007–17 (2017).
60. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41**, 563–571 (2009).
61. Bray, M., Driscoll, J. & Huggins, J. W. Treatment of lethal Ebola virus infection in mice with a single dose of an S-adenosyl-L-homocysteine hydrolase inhibitor. *Antiviral Res.* **45**, 135–147 (2000).
62. Coutard, B. *et al.* Zika Virus Methyltransferase: Structure and Functions for Drug Design Perspectives. *J. Virol.* **91**, e02202–16 (2017).
63. Aouadi, W. *et al.* Binding of the Methyl Donor S-Adenosyl-l-Methionine to Middle East Respiratory Syndrome Coronavirus 2'-O-Methyltransferase nsp16 Promotes Recruitment of the Allosteric Activator nsp10. *J. Virol.* **91**, e02217–16 (2017).
64. Li, G. & De Clercq, E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nat. Rev. Drug Discov.* **19**, 149–150 (2020).
65. Mahalapbutr, P., Kongtaworn, N. & Rungrotmongkol, T. Structural insight into the recognition of S-adenosyl-L-homocysteine and sinefungin in SARS-CoV-2 Nsp16/Nsp10 RNA cap 2'-O-Methyltransferase. *Comput. Struct. Biotechnol. J.* **18**, 2757–2765 (2020).
66. Krafcikova, P., Silhan, J., Nencka, R. & Boura, E. Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. *Nat. Commun.* **11**, 3717 (2020).
67. Uchide, N. & Toyoda, H. Antioxidant therapy as a potential approach to severe influenza-associated complications. *Molecules* **16**, 2032–2052 (2011).
68. Nair, M. P. N. *et al.* The flavonoid, quercetin, differentially regulates Th-1 (IFNγ) and Th-2 (IL4) cytokine gene expression by normal peripheral blood mononuclear cells. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1593**, 29–36 (2002).

69. Karunakaran, K. B., Balakrishnan, N. & Ganapathiraju, M. Potentially repurposable drugs for COVID-19 identified from SARS-CoV-2 Host Protein Interactome. *Preprint at [https://www.researchsquare.com/article/rs-30363/v1]* (2020).
70. Colunga Biancatelli, R. M. L., Berrill, M., Catravas, J. D. & Marik, P. E. Quercetin and Vitamin C: An Experimental, Synergistic Therapy for the Prevention and Treatment of SARS-CoV-2 Related Disease (COVID-19). *Front. Immunol.* **11**, 1451 (2020).
71. Aucoin, M. *et al.* The effect of quercetin on the prevention or treatment of COVID-19 and other respiratory tract infections in humans: A rapid review. *Adv Integr Med* **7**, 247–251 (2020).
72. Luo, W. *et al.* Targeting JAK-STAT Signaling to Control Cytokine Release Syndrome in COVID-19. *Trends Pharmacol. Sci.* **41**, 531–543 (2020).
73. Lindner, H. A. *et al.* The papain-like protease from the severe acute respiratory syndrome coronavirus is a deubiquitinating enzyme. *J. Virol.* **79**, 15199–15208 (2005).
74. Schneider, M. *et al.* Severe acute respiratory syndrome coronavirus replication is severely impaired by MG132 due to proteasome-independent inhibition of M-calpain. *J. Virol.* **86**, 10112–10122 (2012).
75. Longhitano, L. *et al.* Proteasome Inhibitors as a Possible Therapy for SARS-CoV-2. *Int. J. Mol. Sci.* **21**, 3622 (2020).
76. Chojnacka, K., Witek-Krowiak, A., Skrzypczak, D., Mikula, K. & Młynarz, P. Phytochemicals containing biologically active polyphenols as an effective agent against Covid-19-inducing coronavirus. *J. Funct. Foods* **73**, 104146 (2020).
77. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
78. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. (2010).
79. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
80. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
81. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
82. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
83. Medvedeva, Y. A. *et al.* EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database* **2015**, bav067 (2015).
84. Khare, S. P. *et al.* HIstome—a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res.* **40**, D337–D342 (2011).
85. Singh Nanda, J., Kumar, R. & Raghava, G. P. S. dbEM: A database of epigenetic modifiers curated from cancerous and normal genomes. *Sci. Rep.* **6**, 19340 (2016).
86. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
87. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
88. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* **16**, 284–287 (2012).
89. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
90. Horvath, S. & Langfelder, P. Tutorials for the WGCNA package for R: WGCNA Background and glossary. https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/ (2011).
91. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, complex systems* **1695**, 1–9 (2006).
92. Sadegh, S. *et al.* Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat. Commun.* **11**, 3518 (2020).

FIGURE LEGENDS

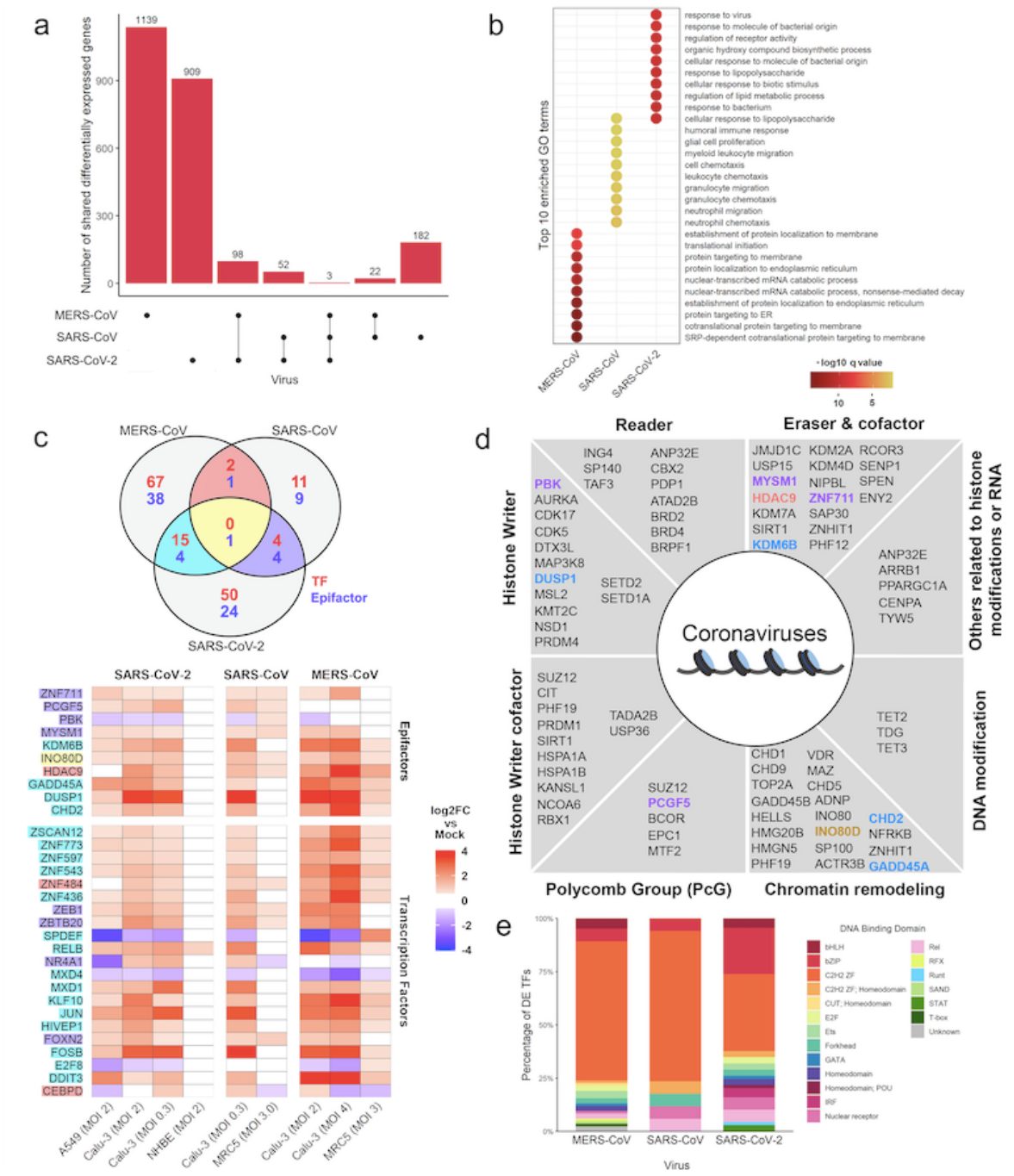


Figure 1. Differential expression analysis of coronavirus-infected cell lines. a Intersection size of the DEGs common to each viral infection represented as single dots (virus-associated gene sets) and the size of their intersections with the other sets (multiple vertical dots). **b** Top 10 simplified enriched Gene Ontology terms of biological process in the virus-associated gene sets ordered by q-value. **c** Shared differentially expressed epigenes between virus-associated gene sets; text color corresponds to the gene classification as either TF (red) or epifactor (blue) (upper panel). Log2 fold change of shared differentially expressed epifactors in each cell line are also shown as a heatmap (lower panel); blank color represents non-significant differential expression, text

highlight corresponds to the intersections shown in the Venn diagram. **d** Functional classification of the identified epifactors; text color corresponds to the intersection color of subsection **c**. **e** Characterization of the DNA-binding domain (DBDs) of human transcription factors (TFs) altered by the viral infection of coronaviruses.

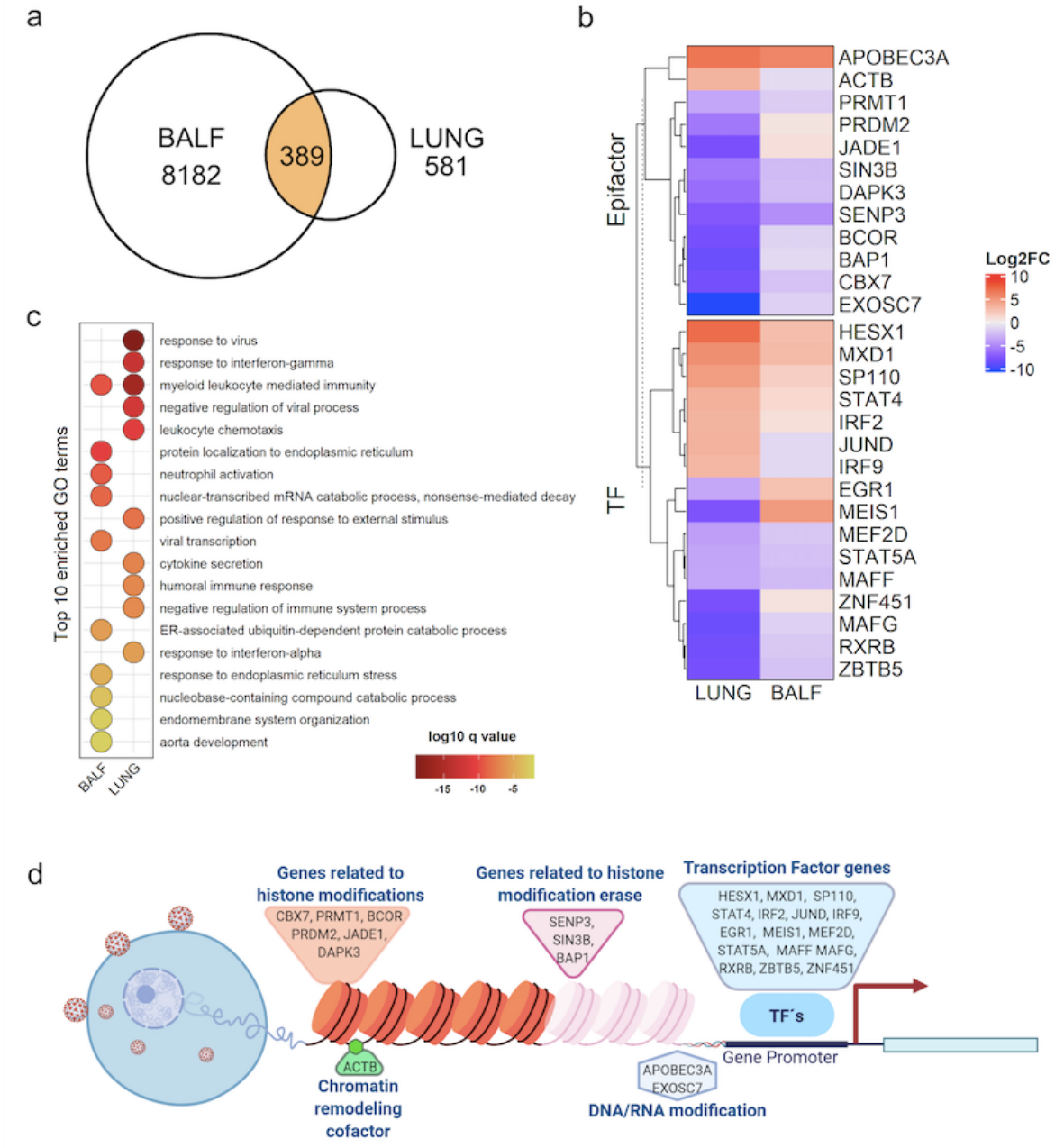


Figure 2. Differential expression analysis of COVID-19 patient samples. **a** Number of shared differentially expressed genes between the samples. **b** Log2 fold change of shared differentially expressed epigenes in patients' samples. **c** Top 10 simplified Gene Ontology enriched terms belonging to the biological process sub-ontology; ordered by q-value. **d** Epigenetic processes associated with the shared differentially expressed epigenes between patient samples. Created with BioRender.com

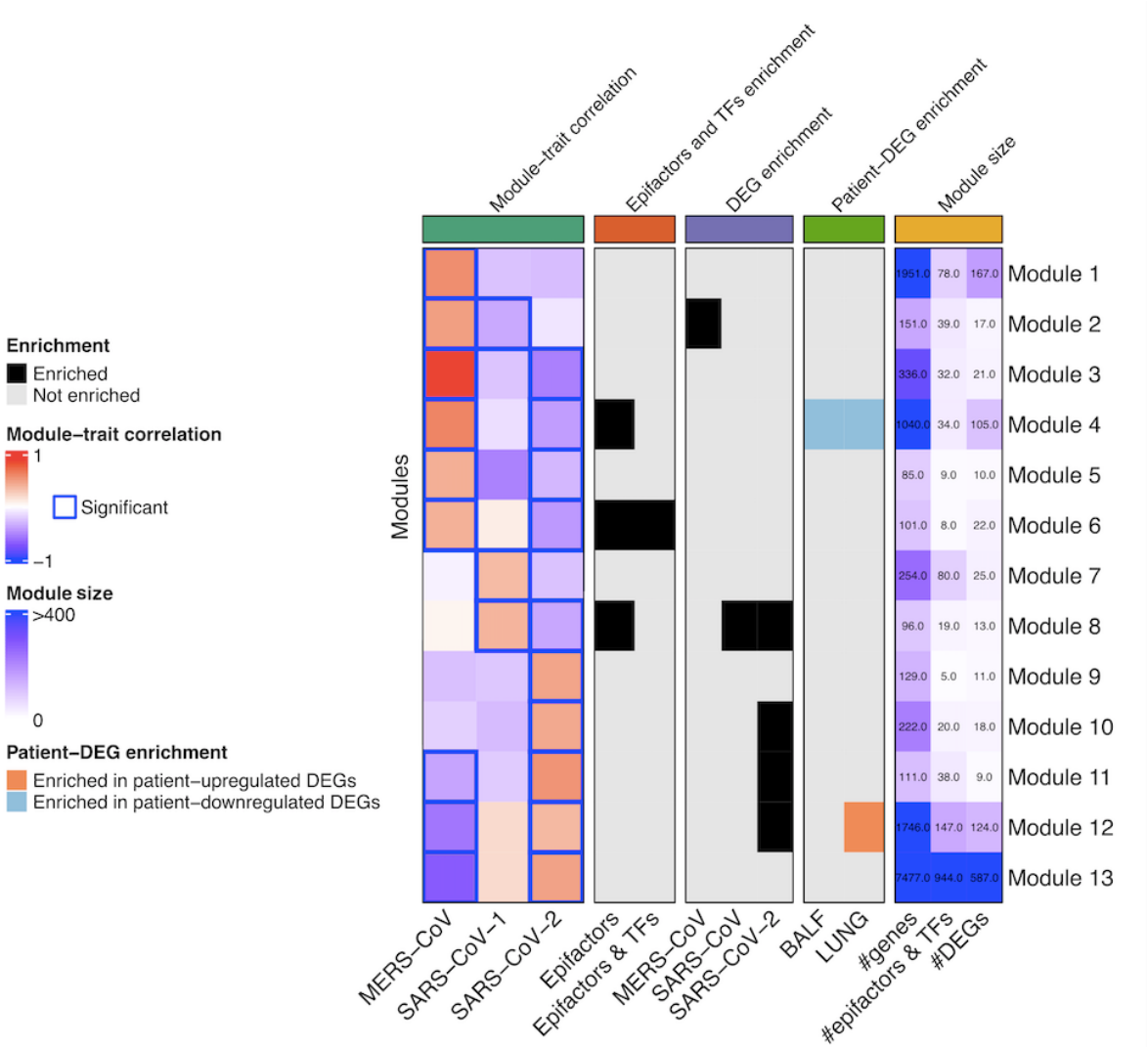


Figure 3. Relevant modules for coronaviruses infection. Summary of the analyses used to identify relevant modules for each infection. From left to right, grids show the module-trait correlation, the enrichment of epigenes, the enrichment of DEGs found in cell lines, enrichment of DEGs found in patients' samples and information of the module size.

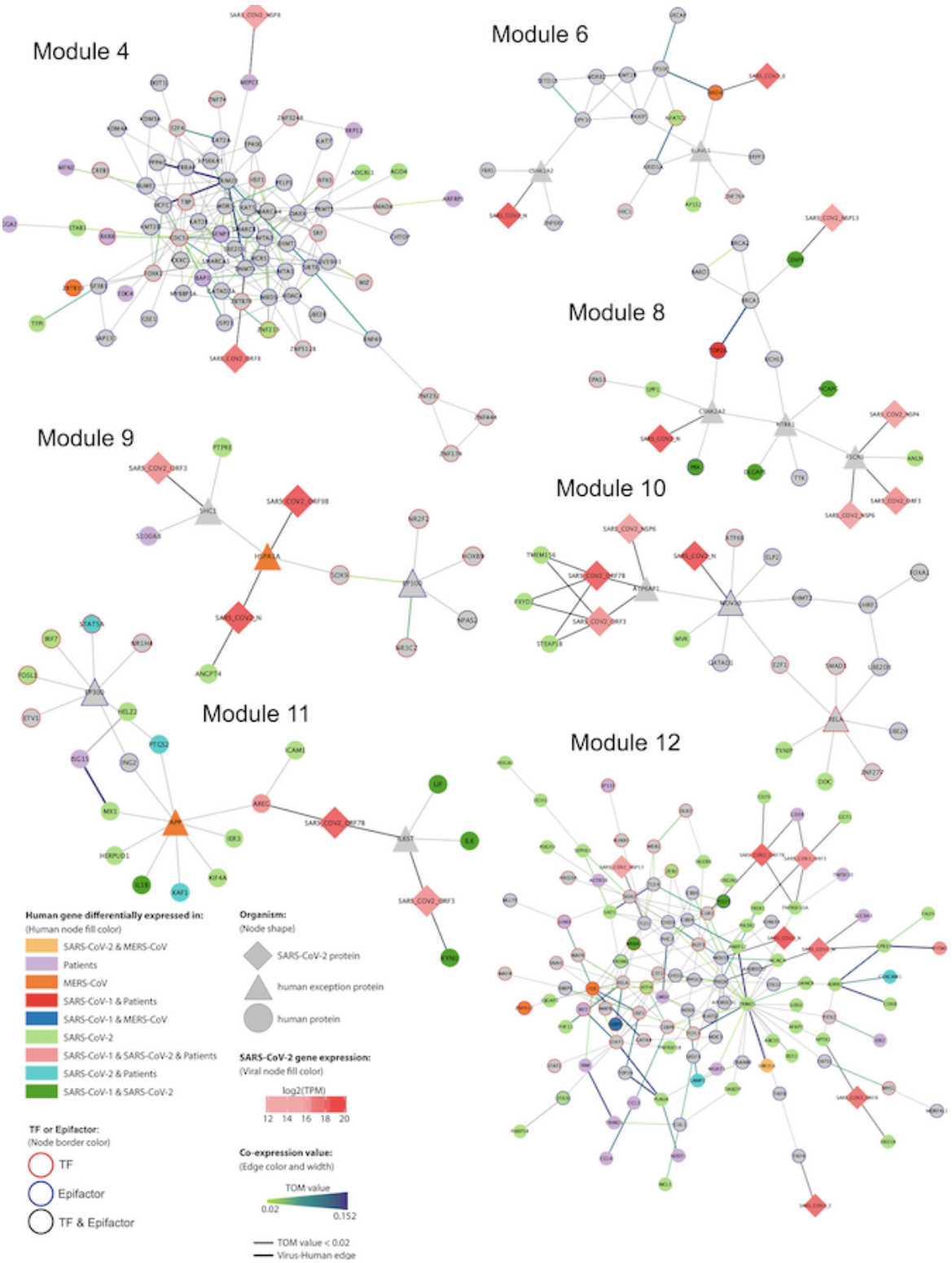


Figure 4. Protein-protein interactions network containing SARS-CoV-2-DEGs, patient-DEGs or selected epigenes for modules 4, 6, 8, 9, 10, 11 and 12. Nodes and edges represent proteins and the interaction between them, respectively. The node and edge color's meaning is indicated in the annotation panel.

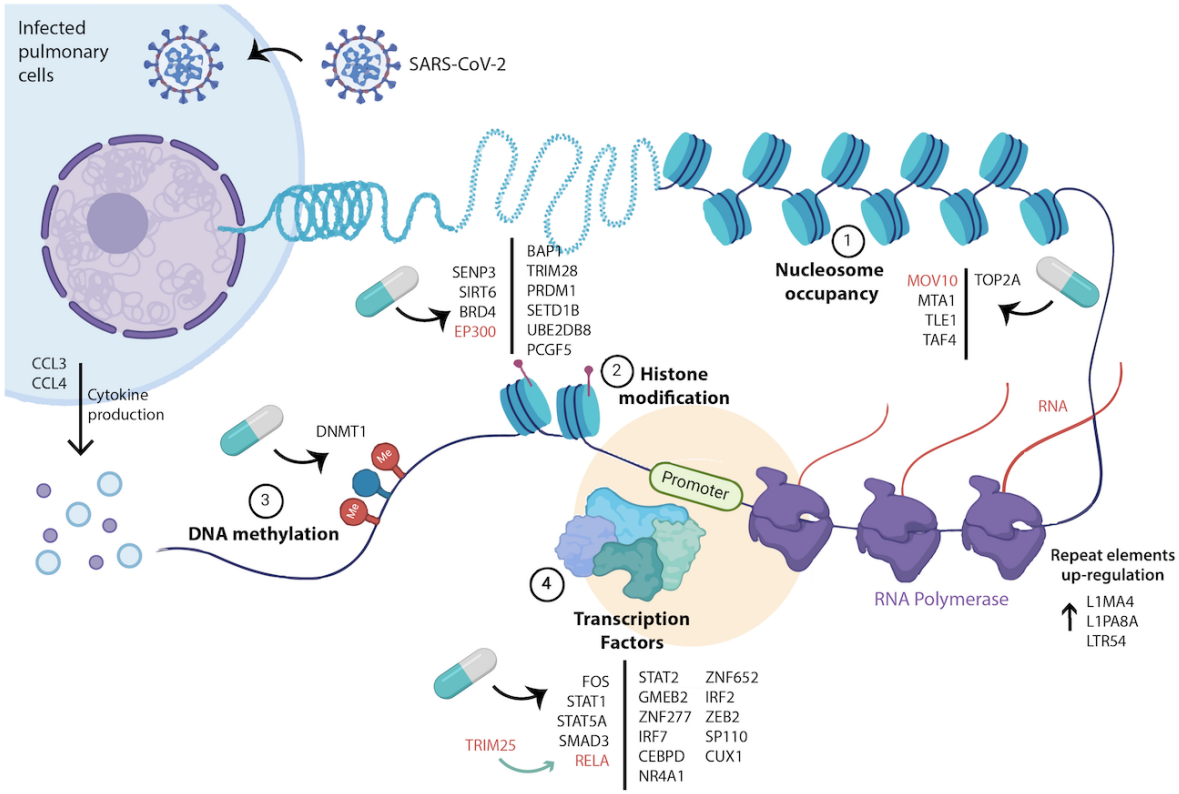


Figure 5. Relevant epigenes in SARS-CoV-2 infection with therapeutic potential. Epigenetic targets are indicated in different processes such as nucleosome occupancy (1), histone modification (2), DNA methylation (3) and also TFs (4). Top gene candidate targets are highlighted in red. Created with BioRender.com

TABLE LEGENDS

Table 1. Drugs targeting candidate epigenes from selected relevant modules for SARS-CoV-2 infection.

Target protein	Drug Name	Module	Function
TOP2A	Genistein, Fluorouracil, Intoplicine, Enoxacin, Sparfloxacin, Amrubicin, Etoposide, Epirubicin, Ciprofloxacin, Myricetin, Mitoxantrone, Trovafloxacin, RTA 744, Daunorubicin, Norfloxacin, Finafloxacin, Dexrazoxane, 13-deoxydoxorubicin, Idarubicin, Lomefloxacin, Lucanthone, Pefloxacin, Valrubicin, Amsacrine, Levofloxacin, Doxorubicin, Declopramide, Annamycin, Banoxantrone, ZEN-012, Podofilox, Aldoxorubicin, Teniposide, Moxifloxacin, SP1049C, Amonafide, Dactinomycin, Fleroxacin, Becatecarin, Ofloxacin, Elsamitrucin	module 8	Epifactor (Chromatin remodeling)
BRD4	Fedratinib, Panobinostat, Romidepsin, Birabresib, Alprazolam, Vorinostat, Volasertib, Alobresib, Belinostat, Apabetalone	module 6	Epifactor (Histone modification read)
EP300	Curcumin	module 6	Epifactor (Histone modification write)

DNMT1	S-adenosyl-L-homocysteine, Procainamide, Palifosfamide, Cefalotin, Decitabine, Azacitidine, Flucytosine, Epigallocatechin gallate, Hydralazine	module 4	Epifactor (DNA methylation)
SENP3	Methylphenidate	module 4	Epifactor (Histone modification erase, Histone modification write cofactor)
SIRT6	7-[4-(Dimethylamino)Phenyl]-N-Hydroxy-4,6-Dimethyl-7-Oxo-2,4-Heptadienamide	module 4	Epifactor (Histone modification erase)
FOS	Pseudoephedrine, Nadroparin	module 12	TF
RELA	SC-236, Betulinic Acid, Bortezomib, Dimethyl fumarate, PHENYL-5-(1H-PYRAZOL-3-YL)-1,3-THIAZOLE, Indoprofen	module 12	TF
STAT1	Epigallocatechin gallate	module 12	TF
STAT5A	AZD-1480	module 11	TF
SMAD3	Ellagic Acid	module 10	TF

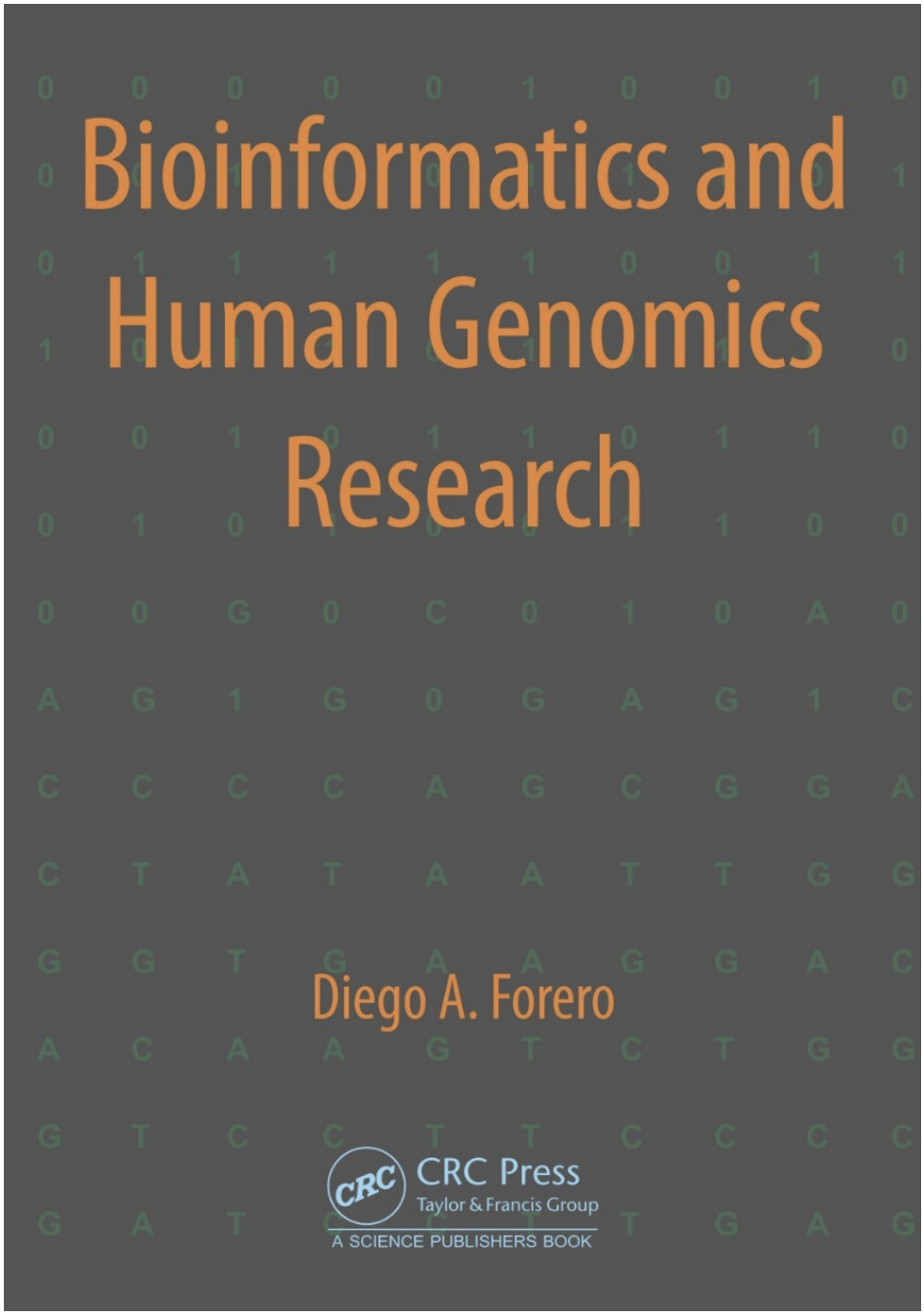
Table 2. Candidate epigenes for drug development in selected relevant modules for SARS-CoV-2 infection.

Target protein	Module	Function
MOV10	module 12	Epifactor (Chromatin remodeling)
MTA1	module 4	Epifactor (Chromatin remodeling cofactor)
TLE1	module 12	Epifactor (Chromatin remodeling, Histone modification cofactor)
TAF4	module 12	Epifactor (Histone chaperone)
BAP1	module 4	Epifactor (Histone modification erase, Polycomb group (PcG) protein)
TRIM28	module 4	Epifactor (Histone modification read)
PRDM1	module 12	Epifactor (Histone modification write cofactor)
SETD1B	module 6	Epifactor (Histone modification write)
UBE2D3	module 10	Epifactor (Histone modification write)
PCGF5	module 12	Epifactor (Polycomb group (PcG) protein)

STAT2	module 12	TF
GMEB2	module 6	TF
ZNF277	module 10	TF
IRF7	module 11	TF
CEBPD	module 12	TF
NR4A1	module 12	TF
ZNF652	module 12	TF
IRF2	module 12	TF
ZEB2	module 12	TF
SP110	module 12	TF
CUX1	module 12	TF
TRIM25	module 12	E3 ubiquitin ligase

2. Publicación de capítulo de libro internacional

Título: Bioinformatics of transcription factor binding prediction
Autores: Erick I. Navarro-Delgado, Marisol Salgado-Albarrán, Karla Torres-Arciga, Nicolas Alcaraz, Ernesto Soto-Reyes, Luis A. Herrera, Rodrigo González-Barrios.
Libro: Bioinformatics and Human Genomics Research
Autor del libro: Diego A. Forero
Editorial: CRC Press, Taylor & Francis Group.
Fecha de aceptación: 24 / 06 / 2020
Fecha de publicación: En prensa



Bioinformatics of transcription factor binding prediction

Erick I. Navarro-Delgado[1,*], Marisol Salgado-Albarrán[2,3,*], Karla Torres-Arciga[1], Nicolas Alcaraz[4], Ernesto Soto-Reyes[2], Luis A. Herrera[1,5], Rodrigo González-Barrios[1].

[1] Unidad de Investigación Biomédica en Cáncer, Instituto Nacional de Cancerología-Instituto de Investigaciones Biomédicas, UNAM, Avenida San Fernando No. 22, Colonia Sección XVI, Tlalpan, CP 14080, Mexico City, Mexico.
[2] Natural Sciences Department, Universidad Autónoma Metropolitana-Cuajimalpa (UAM-C), Mexico City, 05300, Mexico.
[3] Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich.
[4] The Bioinformatics Centre, Department of Biology, University of Copenhagen, 2200 Copenhagen.
[5] Instituto Nacional de Medicina Genómica, Periférico Sur 4809, Arenal Tepepan, Tlalpan, CP 14610, Mexico City, Mexico.
* These authors contributed equally to this work.

Introduction

Transcription factors (TFs) are proteins directly involved in interpreting the genome. These proteins are crucial to the cell, performing the first step in decoding the DNA sequence, which leads to chromatin remodeling and ultimately transcription. TFs belong to a wide number of proteins that are involved in different molecular machineries that regulate the transcriptional control of the cell. The diverse functions of TFs regulate the language of the cell, which directs the development, differentiation, specialization, and response to the environment.

Historically, the term TF has been applied to describe any protein involved in transcription and/or capable of altering gene-expression levels. However, nowadays the term is applied to proteins that directly perform transcriptional control. The key components to understand such transcriptional control were established by Jacob and Monod, more than half a century ago (1961), with their groundbreaking genetic and biochemical experiments in bacterial systems. Their findings shed light to two major concepts in gene regulation: 1) protein-binding regulatory sequences are present in the DNA and 2) proteins bind to such DNA sequences to activate or repress transcription. Through their pioneering work and many subsequent studies, it was established that TFs recognize and occupy those specific DNA sequences, regulating the transcriptional machinery and the outcome of genes (1–3). Due to the importance that TFs exert to the control of gene expression; an intense study has been carried out for decades to understand their functions. This led to the discovery of many general TFs and cofactors in eukaryotic organisms, as well as various chromatin regulators and the mechanisms by which they control gene expression (4,5).

Since TFs depend on DNA sequences and its specific location on chromosomes to carry out their function, it is important to emphasize that these proteins cannot be functionally understood without a detailed knowledge of the DNA sequences to which they bind. These specific TF DNA binding sites (TFBSs) are often referred to as "motifs", which are templates representing the set of related short DNA sequences which are recognized by a given TF (6,7). These sequences can be used to scan longer sequences, such as genetic promoters and enhancers in order to identify possible binding sites (BS). Identifying a DNA binding motif is often the first step towards a detailed understanding of the function of a given TF; knowing the possible BS of a protein provides a gateway for further analysis.

Due to recent advances in biotechnology and especially after the advent of DNA massive sequencing, our knowledge of the mammal regulatory elements, as well as the transcription and chromatin regulators that operate at these sites, has increased considerably in the last decade. Currently there are enormous amounts of data of TFs and the sequences they are associated to. However, being able to predict the expression pattern of a gene based only on its regulatory sequence, turns out to be more complicated when studying the cell; it is generally highly context-specific, depending on the cell type and intracellular factors (2,8). Also, the regulatory regions are not necessarily organized in discrete, easily identifiable regions of the genome and can exert their

influence on genes at great genomic distances (9). Furthermore, even experimentally determined BSs are relatively poor predictors of genes that the TF actually regulates (6).

To date, genomic studies are trying to elucidate regulatory elements, as well as to identify and/or predict the regulatory sequences of the TFs in different species. They have taken two main paths: 1) those studies that identify specific TF binding sites using experimental techniques such as ChIP-seq, SELEX-seq, ChIP-on-chip, CUT&RUN or CUT&Tag; or 2) those studies focused on predicting the possible regulatory elements and their sites in the genome through computational reconstructions and genetic regulatory networks (1,2,5,8,10–13).

Due to the biological importance and implications of understanding gene regulatory machinery, many groups have dedicated themselves to develop various catalogs of TFs, their binding sites and their associated gene elements, as well as various tools for their analysis, visualization, and prediction. Knowing these elements will have important implications for understanding the cell, its development and differentiation, as well as their implications in human medicine. In this chapter, the two different approaches to TF binding analysis will be reviewed. We will also show an overview of the different tools used and their pros and cons in practice in order to understand the general workflow of TFBS studies, as well as the experimental basis, which are necessary to identify the limitations of the field.

1. High-throughput experimental approaches to detect TF-DNA interactions

The study of TFBS has been approached by multiple methodologies through time, each of them with particular advantages and disadvantages. The experimental methodologies can be divided into low and high throughput methods. In this chapter, we will focus on high throughput methodologies, which include Systematic Evolution of Ligands by EXponential enrichment-sequencing (SELEX-seq), Chromatin Immunoprecipitation and DNA microarrays (ChIP-on-chip), Chromatin Immunoprecipitation and sequencing (ChIP-seq and ChIP-exo), and Cleavage Under Targets (CUT&RUN and CUT&Tag). Their basis is briefly explained in the following paragraphs.

Systematic Evolution of Ligands by EXponential enrichment (SELEX, **Figure 1A**) consists of finding the TFBS by creating a pool of random double-stranded DNA sequences (or aptamers) and incubating it with the TF of interest. Following, an immunoprecipitation against this protein is performed, resulting in a selection of DNA fragments containing potential TFBS. The aptamers usually have adapters in the 5' and 3' ends to allow primer hybridization (14), which makes them suitable for amplification and sequencing (SELEX-seq). The main limitation of SELEX is the fact that it is performed completely *in vitro*; thus, several factors that can be important for a TF binding in a living cell, such as transcriptional co-factors, epigenetic modifications or DNA accessibility are absent. (15)

Chromatin Immunoprecipitation (ChIP) allows the identification of DNA fragments bound to a TF. The first step in ChIP is the fixation of DNA-protein complexes with formaldehyde, followed by fragmentation of the DNA. Next, only the DNA bound to the TF of interest is immunoprecipitated using specific antibodies and isolated. The advantage of ChIP is that it captures the DNA-protein interaction *in vivo*; however, it depends on the formaldehyde fixation and the efficiency of the antibody used to immunoprecipitate. The DNA obtained from ChIP can be evaluated by multiple methods to identify the DNA fragments that contain a TFBS; for instance, it can be evaluated by DNA microarrays (ChIP-on-chip) or by high throughput sequencing (ChIP-seq, **Figure 1B**) (15,16). ChIP-exo is a variant of ChIP-seq which uses exonucleases to reduce the length of the DNA fragments used for sequencing, which improves resolution and the identification of TFBSs (17). These methods offer several advantages, such as the low number of cells needed, the high amount of information generated, the reliability and the higher signal to noise ratio (18).

Cleavage Under Targets and Release Using Nuclease (CUT&RUN, **Figure 1C**) is a strategy that utilizes TF-specific antibodies and a micrococcal nuclease (MNase) to produce and select the specific DNA fragments bound to the TF *in situ*. Briefly, in this approach, TFs of interest are recognized by a specific antibody coupled with the MNase, which then cleaves the DNA in the surrounding nucleotides of the TFBS, releasing the TF-DNA complexes. Finally, DNA is extracted, amplified and used for high throughput sequencing (19). Cleavage Under Targets and Tagmentation (CUT&Tag) is an *in situ* methodology derived from CUT&RUN. Instead of MNase, it utilizes a transposase that cleaves the DNA and integrates sequencing adapters at the same time.

DNA fragments bound to TFs are purified and sequenced (20). The advantages of these methodologies are the absence of crosslinking, the reduction of background noise, cost and time compared to ChIP-seq, the low number of required starting cells and the improved resolution of TFBSs identification. Due to all of their advantages, CUT protocols are quickly establishing themselves over ChIP-seq as the standard methods for obtaining genome-wide TFBS.

Finally, data obtained from all the methodologies described above require specific pre-processing to generate a set of selected sequences that contain a potential TFBS that one aims to identify. These final sequences are used as input in the *de novo* motif discovery tools (*Section 3*).

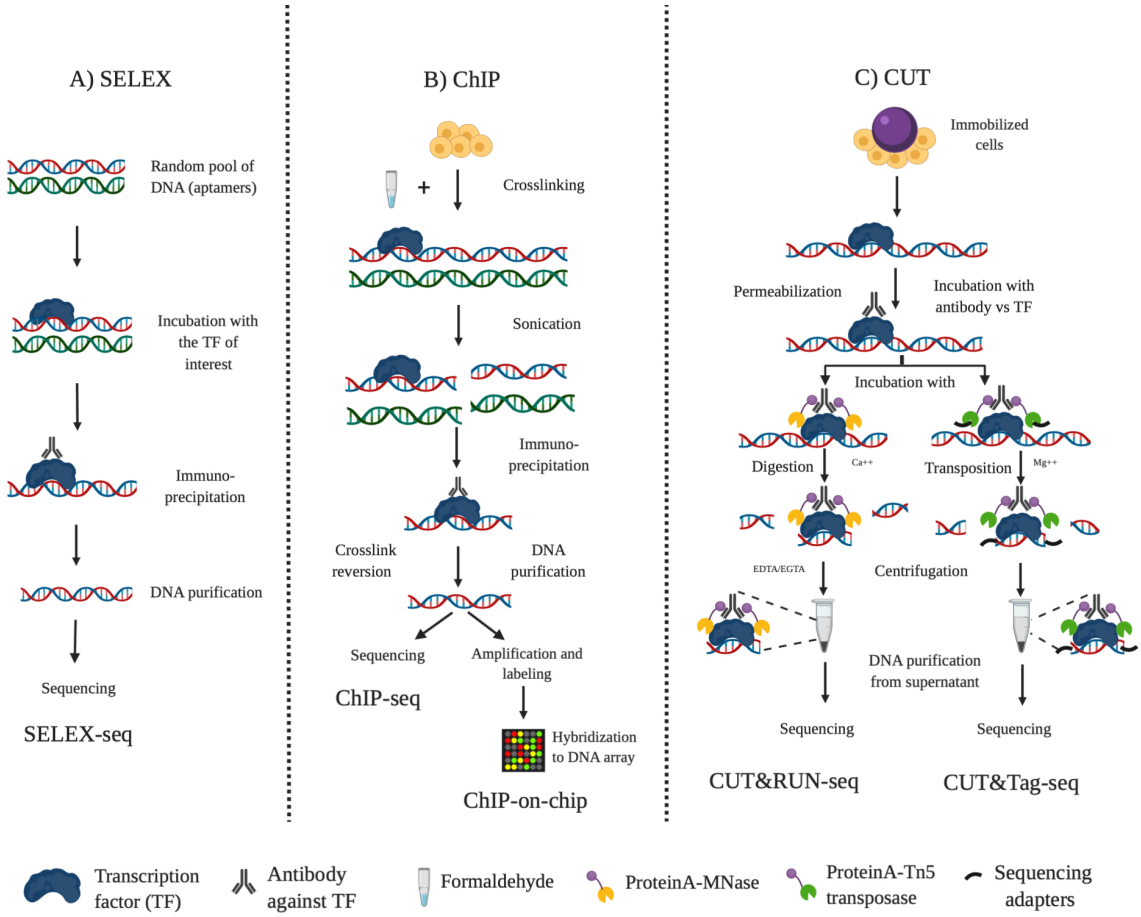


Figure 1: High-throughput experimental approaches to identify TFBSs.
A) SELEX-seq: A random pool of DNA fragments are incubated with a TF of interest *in vitro*, the DNA-TF complexes are immunoprecipitated and sequenced. B) ChIP: the DNA-TF interactions in a cell are fixated with formaldehyde and sonicated to break the strands, then TF-DNA complexes are immunoprecipitated and the DNA fragments are evaluated through DNA microarrays (ChIP-on-chip) or sequenced (ChIP-seq). C) CUT: cells are permeabilized to allow the entrance of the reagents and immobilized on magnetic beads, then the antibody binds to the TF and is recognized by an A-MNase (CUT&RUN) or by a transposase (CUT&Tag), which break the surrounding DNA and allow the isolation of the DNA-TF complexes, finally isolated DNA is sequenced.

2. TF binding motif representations

TFBS can be represented (or modeled) in different ways, which provide different levels of information about the motif recognized by the TF. For instance, the consensus string is the most basic and simple representation, since it depicts the most frequent nucleotide in a motif (i.e. CTCF binding motif 5'-TGGCCACCAGGGGGCGCTA-3') (21–23). Other simple representations exist, such as mismatch strings

(MM) and IUPAC strings, which are consensus representations that permit mismatches or include IUPAC degenerate base symbols, respectively (24).

However, TFs do not recognize fixed and invariable sequences; instead, the nucleotides in each position of a binding site are variable to some extent and the consensus representation doesn't capture the complexity of TFBS recognition. To address these issues, other representations have been proposed, such as Position Weight Matrix (PWM) (25), Dinucleotide Weight Matrix (DWM) (26) and Transcription Factor Flexible Models (TFFM) (27). In this section, we will describe the different representation models for TFBSs, focusing on PWMs since their use in the study of TFs is widespread.

2.1 Position Weight Matrix

A Position Weight Matrix (PWM), also referred by some authors as Position Specific Scoring Matrix (PSSM), is the most common representation of TFBS. It is constructed from a group of aligned sequences recognized by a TF (**Figure 2A**). It consists of a matrix where the probability of appearance of the bases at each position is given, taking into account the background genome frequencies (28). A PWM is defined as a matrix of numbers [$M(b,i)$] for each base ($b = A, C, G, T$) in any position ($i = 1$ *to* l) of a TFBS of length L . It provides an additive score system that reflects the contribution of each position to the TF binding (**Figure 2B**). PWMs offer the following advantages (25) :

1. It depicts the nucleotide frequencies in each position in the motif, which could reflect their importance for the TF binding.
2. It includes position-specific penalties for a mismatch; thus, mismatches at different positions are not treated equally.
3. It employs a logical, easy to understand mathematical model.
4. It is flexible, since it can be modified to incorporate additional characteristics in order to improve the representation accuracy.

The most common visualization of a PWM is via a logo representation, which shows the contribution of each position to the binding of the TF, as well as the base frequency associated with each of them. In a typical logo representation, the x-axis shows each position of the motif, and the y-axis the information content (IC) measured in bits (**Figure 2C**). When the frequency of each nucleotide at a given position is random (taking into account the specific composition of each base in the genome), the IC equals 0. In the opposite case, if a particular nucleotide is found in that position in 100% of the sequences, the IC at that position would be 2. This measure indicates the importance of that position to the specificity of the TF. Consequently, positions with the highest IC are the most critical to the binding, while the ones with the lowest values can have variations without having big effects on the binding (25). Further information on the computation of the IC, as well as the equations and concepts, can be found in Stormo, 2013 (25).

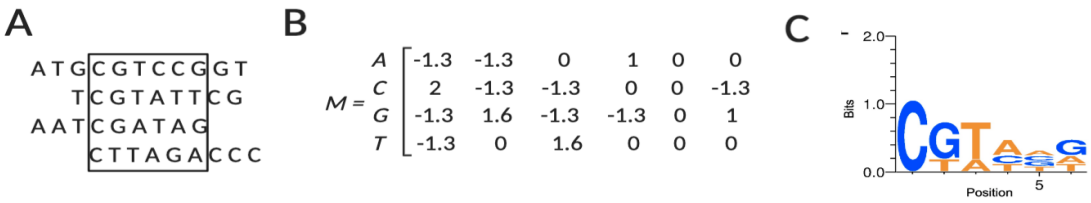


Figure 2. PWM and its representation.

A) Aligned sequences containing the 6 nucleotides where a hypothetical TF binds. B) PWM of the TFBS; each element in the matrix represents the score of every base at any position of the motif. C) Logo representation of the hypothetical TFBS.

Even though this model is the most used nowadays, it is important to remember that it is just a way to approximate and represent the real specificity of a protein, and it has limitations as any model. First, this model assumes that each position contributes independently and additively to the binding of the TF, which is not always true (29). This simplification is preferred because a complete dependence model would require to estimate a joint distribution that grows exponentially with the size of the motif and becomes computationally

intractable (30). Also, when trying to scan PWM to the genome, in order to find new TFBS, the false-positive rate is usually high. This could happen because some TFBS might be in a non-permissive locus, being inaccessible to the TF (22). Finally, it is worth mentioning that the representation accuracy of the TFBS is highly dependent on the algorithms that generate the PWM, as well as the parameters that they use. Therefore, they can lead to poor models that result from the limitations of the tools, not from the PWM approach *per se (25)*. Nevertheless, due to their simplicity and interpretability PWMs remain a popular and standard representation of TFBS.

2.2. Other representations

Given the limitations of PWMs, which assume independence among nucleotides in a TFBS, there have been several efforts to develop alternative representations capable of incorporating additional information that allow the improvement of TFBS predictions in new locations.

As an example, one way to improve the consensus and PWM representations is by incorporating the inter-position dependence between the nucleotides in a motif. As an example, given a group of sequences known to be recognized by a TF, Osada *et al*. 2004 also considered the number of shared bases and the pairwise nucleotide dependencies within the sequence to construct their model (31). The work of Osada *et al*., along with other studies, show that the incorporation of the inter-position dependencies improves the prediction of new sites in the genome (32–34).

Another approach was proposed by Hannenhalli and Wang, who used mixture models to search for subclasses of a given TFBS. The rationale behind this approach is that one TF can have different binding preferences depending on the biological context (i.e. cell type or high and low-affinity sites); thus, classifying one PWM into subclasses of PWMs can provide better TFBS predictions (35).

Finally, an important representation model is the one introduced by Mathelier and Wasserman, named Transcription Factor Flexible Model (TFFM,) which is based on Hidden Markov Models. The main advantage of this model is its capacity to capture nucleotide inter-position dependencies and variable lengths in a motif, which has led this model to outperform PWMs in several contexts. Furthermore, the TFBS database Jaspar (*Section 4*) contains TFFM in addition to the PWMs (27,36).

3. *De novo* motif discovery: obtaining TFBSs from a set of sequences

A number of methods with different approaches have been developed to find the optimal model with accurate weights to generate the best representation of the TFBS. The problem of *de novo* motif discovery can be stated as following: given a group of sequences, one must infer both the binding motif and the position (which can be different for each sequence) at the same time. Solving this problem involves applying some form of algorithm to find the most likely TFBS motif and constructing its corresponding PWM or another representation model. Finally, the resultant motifs can be evaluated with different tools (Section 4) to filter the best TFBS candidates that could play a potential biological role.

The main challenges in *de novo* motif discovery are that, given a set of sequences with different lengths and unknown motifs at unknown positions, we have to obtain an accurate motif representation. Furthermore, the motifs are usually not identical to each other, since some nucleotide positions might not be critical for the binding of the TF (37). With over a hundred publications to date and countless software tools, *de novo* motif discovery has been one of the oldest core computational problems tackled in the field of bioinformatics. They can be classified according to the approach of their algorithms in enumerative, probabilistic, nature-inspired, deep learning-based and ensemble (**Figure 3**). For the purposes of this book, only the most representative and widely used methods will be described (**Table 1**).

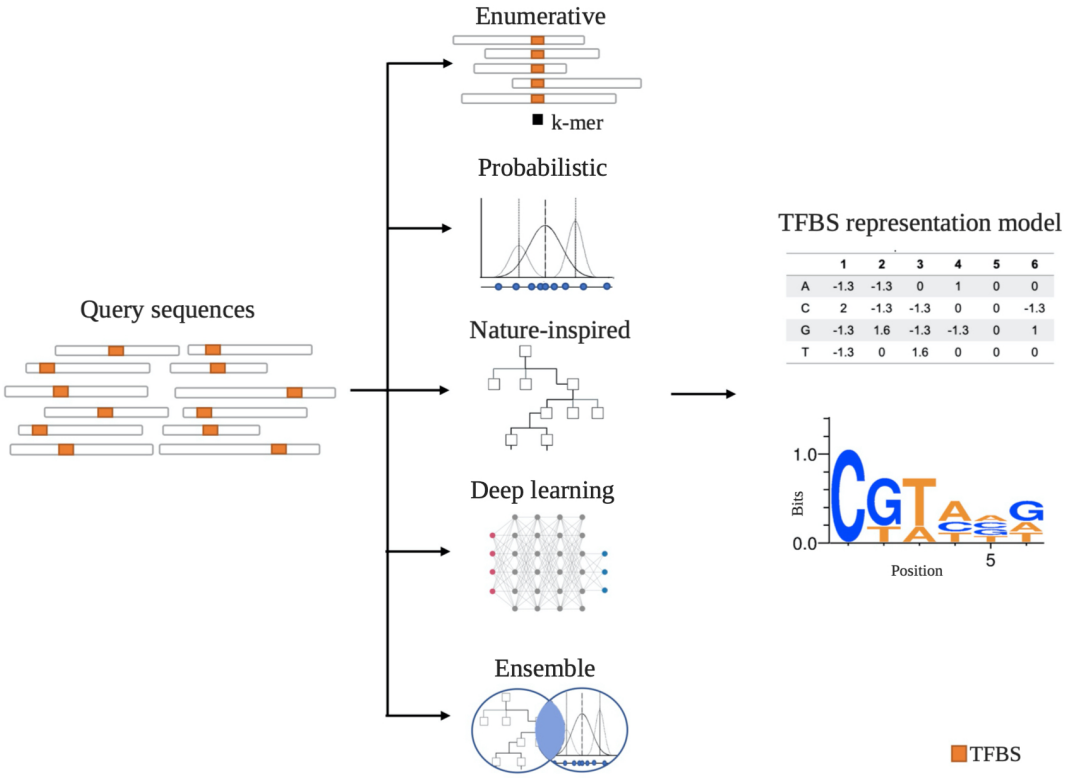


Figure 3. *De novo* motif discovery.
Sequences with a potential TFBS are used to perform the *de novo* motif discovery analysis, which can be done with enumerative (simple-word enumeration), probabilistic (EM), nature-inspired (GA), deep learning-based (CNN) or ensembl approaches. After this step, a TFBS representation is obtained, such as a PWM, which can be represented as a sequence logo.

3.1. Enumerative approaches

Enumerative approaches perform an exhaustive search for a consensus motif by the comparison and computation of the similarity between oligonucleotides. Therefore, this method is more likely to find the global optimum (i.e. the best solution in the whole search space). It works well for short motifs, as the ones found in eukaryotes, and is suitable for finding totally constrained motifs (i.e. where all the instances are identical) (38). Some of the programs that use this approach are DREME (38) and CisFinder (39).

Since this approach analyzes the frequencies of all the DNA strings in order to generate a PWM from the overrepresented oligonucleotides identified (28), it has an exponential time complexity. Thus, it has problems handling big data or finding long motifs. Also, it needs a high amount of parameters specified by the user, such as motif length, mismatches allowed and a certain number of sequences where the motif appears. In addition, because most of the TF in eukaryotes have weak constrained positions, the results of this approach can be problematic, so they need to be post-processed with clustering systems (37). In order to adapt these methods to high-throughput sequencing data, parallel processing and optimized data structures have been implemented to accelerate the algorithms.

DREME is a simple word enumeration method developed to find multiple short, non-redundant and statistically significant eukaryotic motifs in an optimized way using regular expression words. The speed is achieved in part by limiting the search to short motifs (4-8 bp). Also, it is exhaustive for exact words and heuristic for words with wildcards. The general algorithm workflow starts with generating a set of short oligonucleotides, which are tested with Fisher's exact test using a threshold to calculate the significance of each k-mer. This test looks for an overrepresentation of the sequence identified as a motif in the data. To perform it, two datasets are used:

the enriched regions obtained from an experimental technique (Section 2) and unrelated regions, which could be shuffled sequences produced from the same dataset (38). The most significant motif is used in an inner loop, where it becomes a seed regular expression to conduct a beam search that identifies the most significant generalization (30). Then, a PWM is created by aligning the sequences that match with the suggested motif. Finally, the best motif is erased in order to find multiple non-redundant motifs, and the previous steps are repeated iteratively until there are no more motifs found with an E-value less than the specified significance threshold (38).

An alternative approach is CisFinder, a word clustering-based method that detects short motifs as well, but with a more efficient processing speed. This method is based on clustering short Position Frequency Matrices (PFMs) that are overrepresented in the dataset using a hypergeometric probability distribution, similar to the method described above. Briefly, PFMs are estimated from 8 base pairs sequences with and without gaps. Then, the flanking regions of the overrepresented motifs are extended, generating PFM for the sequences in the gaps and on the sides. If these regions are not informative, they are trimmed. Then, these matrices are clustered based on their similarity using Pearson correlations. After single-linkage clustering, each group is evaluated for homogeneity and separated if they are not similar enough. Separated motifs are used as seeds for adding more motifs, which are later evaluated. These processes are repeated iteratively until all the motifs are separated in homogeneous groups. Finally, the PWM of each entire cluster is estimated, giving as result several non-redundant TF binding motifs. The advantage of this method is its capacity to discover multiple and weak motifs in a single run, even with a low level of enrichment and the ability to process large sequences (39).

3.2. Probabilistic approaches

Probabilistic approaches are the most used currently. They test PWM parameters with probabilistic methods while doing multiple local sequence alignment (40). These algorithms have some improvements in comparison to the enumeration ones: they are faster, require fewer parameters, remain unaffected by motif length, can handle big datasets and are able to find weak constrained motifs (i.e. motifs where not all the instances are identical). Nevertheless, these algorithms scale poorly with dataset size and converge to a locally optimal solution (37). These methods typically use Expectation Maximization (EM), like MEME, and Gibbs sampling, like Align ACE (41).

EM is a deterministic approach that works under the assumption that each sequence of the data has at least one common motif. It works in two phases: the expectation and the maximization. During the expectation phase, the score for different motifs in all the sequences is estimated based on the entries in the PWM and the base pair composition of the genome (which reflects the background probability of each nucleotide). In the second step, those estimated values are used to refine the PWM through several iterations (37). This approach uses oligonucleotides from the data as starting points to increase the probability of getting to a global optimum (22). In other words, the goal in this method is to find an initial motif and then use the described phases to improve it until it converges to a locally optimal solution. This approach has been widely used in various software tools, being MEME (Multiple EM for Motif Elicitation) (42) the most popular. The main disadvantages are that this algorithm is very sensitive to the initial conditions and it assumes only one TF binding motif per sequence (37). Also, it is very time consuming, so a usual strategy when using this algorithm in large datasets is to run it on a small subset of the data.

Another algorithm that is widely used is Gibbs sampling. This is a Markov Chain Monte Carlo (MCMC) approach, where the results of every step depend only on the immediately previous state. This is rather a stochastic model since each step is based on random sampling. In this approach, the mutual segments within the sequences are analyzed. The goal is to find the best common pattern, which is obtained by localizing the alignment with the highest ratio of pattern probability to background one (43). It is less dependent on the initial conditions, but more dependent on the input sequences. Align ACE (41) is a program that uses Gibbs sampling. This program evaluates the motifs with the MAP (maximum *a priori* log-likelihood) score, which judges the motifs that are obtained through the course of the program. Briefly, this score takes into account the direct relationship between the number of aligned sites and the degree of overrepresentation of the TF binding motif in the input data (30). Some of Align ACE's advantages are that the base frequencies are fixed according to the source genome, both strands of the input sequence are considered without allowing overlaps and multiple motifs can be found by masking iteratively single motifs (41).

3.3. Nature-inspired approaches

This category includes algorithms that have been inspired by natural phenomena. They are typically based on swarm intelligence, as well as biological, chemical and physical systems. Some of them have been created to solve complex and dynamic problems, offering low-time and optimal-cost solutions. Even though not all of them are very efficient or widely used, some have offered new approaches to the field with different advantages over the other categories. However, approaches solely based on these algorithms are rare; they are more frequently used in this field in combination with other methods. Popular algorithms used in *de novo* motif discovery include Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) (44) and Artificial Bee Colony (ABC) (37).

Genetic Algorithms are optimization procedures that iteratively improve a set of solutions (population). The main goal of these types of algorithms is the production of "offspring" results by mutation and recombination. The starting point is a set of random individuals, which are used to produce offspring. After each step, a set of new solutions is generated and evaluated, keeping only the ones with the best fitness for further exploration (i.e. the ones with the highest score). Throughout several generations, local optimum solutions are found (37). Some methods, like rGADEM (45), which will be explained in a following subsection, use this algorithm.

Alternative methods that have been explored in the field encompass PSO, which simulates the behavior of social animals like birds to find resources. This algorithm consists in a population of candidate solutions that keep moving in the search space. Each solution can communicate with the other ones to influence their movement toward the best positions, so that the swarm can explore and find local optimal solutions. This algorithm has been used to generate seeds that are used afterward by the EM method (37). In a similar way, both ACO and ABC algorithms simulate the social behavior of ants and bees respectively when trying to find food. In the case of ACO, artificial ants randomly search the solution space and leave "pheromone" over their search paths for other ants to use as "memory" of good solutions and moving towards better ones. In the case of ABC, three roles of artificial bees exist: employees, onlookers and scouts. The candidate solutions can communicate as well and inform the other ones about the best positions to explore these search spaces. The evaluation of the new possible solutions are done based on the similarity values of the consensus sequences (37).

3.4. Deep learning

Artificial Neural Networks (ANN) are mathematical models that attempt to mimic how the biological brain learns to solve problems by training on a set of examples. In ANNs, artificial neurons are interconnected in sets of at least three layers: the input layer, one or more intermediate (hidden) layers that learn the features driving the prediction, and an output layer providing the final predicted value or values. When presented with an input example, each neuron can activate and "fire" a value to its output connections in the next layer if a certain threshold is reached. The value compared to this threshold is the weighted sum function of the values of their firing input connections in the previous layer. On each training phase, the weights for each connection are updated by backpropagating the error from the final predicted value compared to the actual value; these networks are iteratively trained until convergence to a minimum error.

Although the origin of ANNs can be traced back to the 1950s with the perceptron model (46), until recently, several challenges made them difficult to apply to complex problems: large computational requirements, need for very large datasets to train with and low performance due to numerical optimization problems in the backpropagation step. However, the last two decades have seen major algorithmic breakthroughs and technological advancements that have enabled to train "deep" ANNs with many layers and millions of neurons. This new Deep Learning paradigm has enabled researchers to solve complex problems and greatly outperform conventional algorithms in a wide range of fields.

One of the first models used to predict TFBS from sequence was DeepBind (46), which predicted *in-vivo* and *in-vitro* binding affinities of various proteins. Other methods soon followed, such as DeepSea (49) and Basset (47) which in addition to TFBS predict other features such as histone modifications and chromatin accessibility, DeepSite (48) which uses protein structure in addition to sequence, DESSO (45) that incorporates DNA shape, and many others. The success of the first Deep Learning methods have inspired a growing number of models that try to improve them in different ways, such as using more complex architectures (e.g. combining other types of networks with CNNs), solving other problems (e.g. enhancer prediction, Protein-RNA binding site

prediction) or using other types of information as input in addition to the sequences (e.g. DNA methylation, phylogenetic conservation).

Although Deep Learning is making substantial gains in the genomics fields and is quickly outperforming traditional methods for TFBS, it is still considered to need some time until it reaches a mature state where non-expert users are able to make use of them on a daily basis. A standardized protocol to develop, train, report, validate and share the models has yet to be defined by the scientific community, making most models difficult to use or adapt to other more specific problems. Some challenges also remain to be tackled, such as training with a low number of examples, improving the interpretability of more complex architectures and defining good sets of "negative" examples to train with. Nevertheless, given their superior performance and multi-tasking capabilities it is just a matter of time until these drawbacks are surpassed and Deep Learning becomes the standard method for TFBS prediction.

3.5. Ensemble approaches

The programs described here combine different approaches from the above-described algorithms. Therefore, there is not an established algorithm or a set of characteristics for the methods here mentioned. Rather, their strengths and limitations depend on the hybrid algorithm that results from the integration of other approaches. Examples of this type of methods are MEME-ChIP (50), HOMER (51), rGADEM (45) and DeepFinder (52).

MEME-ChIP (50) is a web-based tool that mixes 4 algorithms for *de novo* motif discovery, comparison and visualization. This tool mixes the advantages of two already explained approaches; the probabilistic algorithm MEME (42) and the enumerative one DREME (38). After using both approaches, CentriMo (53) evaluates the enrichment of the candidate motifs in the input data. The last component of this tool is Tomtom (54) which helps to identify TFs that could mediate an indirect and cooperative binding in the source protein (50).

HOMER (Hypergeometric Optimization of Motif EnRichment) (51) is a tool that combines enumeration and probabilistic approaches. The algorithm is composed of 2 stages: the first one is an exhaustive search for overrepresented putative motifs, which correspond to the enumerative stage. To speed this process, a sequence tree is used to optimize the comparison between words and the consensus motif. Then, a modified version of the Fisher exact test is used to identify the enriched motifs. Afterward, the top results are converted into probability matrices to start the probabilistic step, where the putative motifs are refined using a local hill-climbing approach, an iterative local optimization algorithm. Whenever the local optimization algorithm finds a solution, the motif is reported and the matching sequences are removed. This step is then repeated in order to find multiple motifs. In the end, the best threshold and probability matrices are reported (51).

Alternatively, rGADEM (45) is an R package that combines GA with EM. Briefly, short candidate words of 4-6 nucleotides are used to construct spaced dyads (i.e. motifs with gaps in between). These spaced dyads are sorted according to their enrichment in the input dataset, followed by a conversion to PWM. The matrices are then optimized with EM and passed to the GA as starting points in order to increase the probability of finding the best local optimums. Afterward, the GA runs with the generated population, where the fitness score is based on the logarithm of the E-value. The motif with a fitness value less or equal to a specified cutoff value is reported, followed by the mask of its binding sites in the input dataset. This step is repeated iteratively to find multiple motifs until no new ones with the required fitness value are found. This tool can process big datasets, identifies multiple dimer and monomer motifs and adjusts motif widths, offering a fast and efficient framework (45).

Finally, DeepFinder (52) utilizes other tools for identification of initial candidate motifs (MEME, MotifSampler, Bioprospector and MDSCAN), followed by a stacked-autoencoder neural network learning step which is used to predict the associated TFBS in the input sequences.

Tool	Main method(s)	Reference

Enumerative		
DREME	Simple-word enumeration	(38)
CisFinder	Word-clustering based method	(39)
Probabilistic		
MEME	Expectation Maximization	(42)
Align ACE	Gibbs sampling	(41)
Deep learning-based		
DESSO	Convolutional Neural Networks	(55)
DeepBind	Convolutional Neural Networks	(56)
DeepSite	Convolutional Neural Networks	(57)
Basset	Convolutional Neural Networks	(58)
Combinatorial		
MEME-ChIP	Expectation Maximization and simple-word enumeration	(50)
HOMER	Enumeration and hill-climbing approach	(51)
rGADEM	Genetic Algorithm and Expectation Maximization	(45)
DeepFinder	Probabilistic algorithms and neural networks	(52)

Table 1. Non-exhaustive list of tools used for *de novo* motif discovery.

4. Motif prediction: identifying candidate TFBS in the genome

As mentioned in the previous sections, the approaches for *de novo* motif discovery yield a high number of false-positive sites because TFBS are short and variable (59). To address this problem, PWMs obtained from the discovery phase or from a database containing TFBS (**Box A**) can be further filtered to retain only the candidates with a potential biological function. Furthermore, we can use PWMs to know if a particular TFBS is contained in a sequence of interest, such as a promoter region. In this section, we refer as "prediction" to the process of scanning a region of interest for a TFBS.

The first step of the prediction process is the search for occurrences of one or multiple PWMs in a sequence of interest (**Figure 4A**) (60). Several tools are available for this purpose which can search for individual sites or for several TFBS (clusters) (28). For a detailed review on the topic, see Aerts 2012 (61), Hannenhalli 2008 (30), Das and Dai 2007 (43) and Bulyk 2003 (59). The principle behind the prediction of TFBS is the search for the number of occurrences (or matches) in a sequence of interest, given one PWM. Several tools are available for this purpose which rely on nucleotide sequence information only and use a pattern matching method to identify an occurrence. The degree of match can be represented in different ways (p-value, percentage, etc), depending on the method used (30,61). Some example tools are: FIMO (62), MATCH (63) and Matrix-Scan (64).

Once the starting match-based search has been performed, tools are available for predicting and/or filtering potential false positive matches by incorporating extra layers of biological information, such as evolutionary conservation, gene expression or epigenetic data (61), which will be briefly described in the following subsections.

4.1. TFBS clusters

This approach is based on searching for clusters of TFBS in a region (**Figure 4B**), instead of looking for individual TFBSs. The premise is that transcriptional regulation is not controlled by one TF, but by a combination of several ones and that regulatory regions with a higher density of TFBS (clusters) can be biologically relevant. Most of the tools developed, search for clusters of TFBS regardless of the order, strand or the separation between the sites (61). Representative examples of tools are Cluster-Buster (65), MCast (66) and BayCis (67).

4.2. Phylogenetic footprinting

The premise of phylogenetic footprinting is that TFBSs located in conserved regions among different species (orthologous) are more likely to be biologically relevant, in contrast to TFBSs located in non-conserved regions (**Figure 4C**). Phylogenetic footprinting is capable of identifying potential TFBSs for a single region in the genome, provided that it is conserved across other species (orthologous). For a detailed description of phylogenetic footprinting see Hannenhalli, 2008 (30).

Several tools have been developed and can significantly improve the discovery of relevant motifs (43). These tools usually involve two phases: 1) global multiple alignment of the orthologous sequences and 2) identification of the conserved region in the alignment. Some tools also incorporate a third phase which includes the search for matching PWMs, only if the TF binding site is conserved (43); (61). Some examples of tools are: TargetOrtho (68), rVISTA (69), MONKEY (70) and TFLOC (71).

4.3. Co-expression

These approaches work under the assumption that genes with similar patterns of expression (co-regulated) can contain some similarities in their regulatory regions, including TFBSs (**Figure 4D**). Thus, the purpose is to find PWM matches enriched or overrepresented in co-expressed genes (43). This is usually done with the integration of RNA-seq or expression microarray data.

4.4. Multiple evidence

These approaches take advantage of the high number of genome-wide data available, such as gene expression and epigenetic modifications, to identify combinatorial codes and to better predict TFBSs (**Figure 4E**) (61). As an example, PriorsEditor (72) is a tool that can combine different data, such as phylogenetic conservation, DNA melting temperatures, nucleosome-positioning, GC content, DNA bendability and DNA duplex-free energy to better identify functional TFBS in a cell type of interest. Other examples are CHROMIA (73), CENTIPEDE (74) and MotifLab (75)

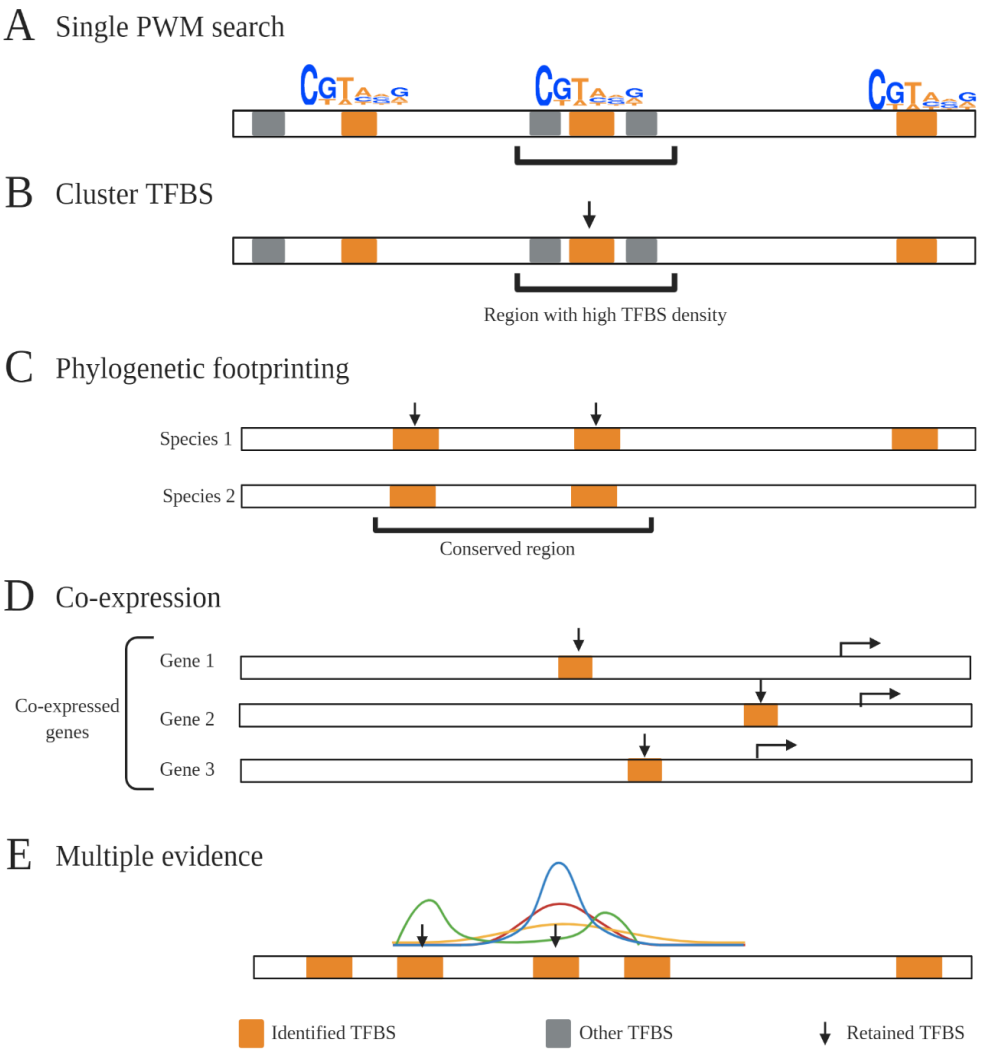


Figure 4. Approaches to predict TFBS in the genome from a PWM.
A) Initial matching of a PWM in a region provides potential TFBSs. Further methods can be used to reduce the false-positives based on B) the density of TFBS (clusters), C) conservation of the matching region across different species, D) the presence of co-expressed genes in the sample, E) integration of several layers of information such as phylogenetic conservation, GC content and physicochemical information of the protein, among others.

Box A. Databases containing human PWM

There are different databases where PWMs can be obtained, being TRANSFAC (13) and JASPAR (36) the most popular. However, other ones have information of human binding sites as well, like HOCOMOCO (76), HOMER (51) and CIS-BP (77) (Table 2).

TRANSFAC is a commercial database that contains TFBS of several species, with a focus on model organisms. It has more than 67,000 manually annotated TF site interactions, over 7,000 PWM derived from experimental evidence, and more than 2,000 TFBS ChIP-seq experiment reports. Additionally, it offers additional tools and data, like pathway visualization for regulatory networks, promoter reports and TF reports (13).

JASPAR is an open-access database of curated, non-redundant TF binding profiles derived from experimental evidence and ChIP-seq data. These representations are stored as PWMs and TFFMs and cover a wide range of

species grouped in six main taxonomic groups (vertebrates, plants, insects, nematodes, fungi and Urochordata). It is the most complete free resource, collecting over 1,700 TF binding profiles (36).

Finally, some research groups have made alternative databases with public data analyzed or curated in a different way, which is typically utilizing their own software tools. HOCOMOCO (76) (Homo sapiens comprehensive Model Collection) provides TF binding models for 680 human and 453 mouse TFs. All of these models are generated with ChIPMunk, a probabilistic de novo motif discovery tool that mixes greedy optimization with bootstrapping (78). On the other hand, HOMER is a database maintained as part of the HOMER software. It is based on the analysis of public datasets using their suggested approach, collecting over 400 TF binding representations (51). Finally, CIS-BP is a public database that incorporates data of more than 390,000 TFs data from around 700 species. It collects data from other databases like TRANSFAC and JASPAR. The novelty of this database is that it includes inferred motifs, which are TF binding motifs that are inferred from related species with the known TFBS of the ortholog protein (77). It is important to mention that several other databases that focus on specific organisms exist. However, since they do not contain human-related content, they are not mentioned.

Database	Description	Link
TRANSFAC(13)	Focused on model organisms >67,000 manually annotated TF interactions ≈ 7000 TF binding profiles	http://genexplain.com/transfac/
JASPAR(36)	Focused on 6 main taxonomic groups ≈ 1700 TF binding profiles	http://jaspar.genereg.net/
HOCOMOCO(76)	Focused on mice and humans ≈ 680 human and 453 mouse TF binding profiles	https://hocomoco11.autosome.ru/
HOMER(51)	Focused on humans ≈ 400 TF binding profiles	http://homer.ucsd.edu/homer/motif/motifDatabase.html
CIS-BP(77)	Wide range of species (> 700 organisms) > 165,000 TF binding profiles Collects data from >70 sources	http://cisbp.ccbr.utoronto.ca/

Table 2. TFBS databases.

5. Final remarks

Studies performing comparisons between methods for discovering TFBS have not had conclusive results (i.e. no method consistently outperforms the others in all data sets). Interestingly, most algorithms work better in simpler organisms' data, like yeasts, than in similarly created datasets from higher organisms, like mice and humans (43). Therefore, no standard methodology exists; the accuracy and performance of the approaches change with different input data.

Additionally, it is very difficult to evaluate the performance of *de novo* motif discovery and PWM scanning tools. In order to do so, one should have complete annotations of precise validated sets of TFBS in the DNA for specific proteins, which would be used as a gold standard reference. This kind of information is usually missing or limited in the majority of situations. However, in human ChIP-seq data, rGADEM has shown to be one of the best-performing tools, outperforming HOMER and MEME-ChIP (28). Regarding PWM scanning tools, MCAST and FIMO perform better than some of its competitors (28).

It is worth to mention that some TFs interact with other partners in the cell, which might change their binding motif either by indirect binding (the partners are the ones that bind to the DNA) or cooperative binding (the protein binds to a different motif when interacting with its partner). Furthermore, we must remember that, in a cell, there are multiple factors that influence the specific binding of a TF. Just to mention some, the methylation status of a sequence can change the affinity of the binding of the protein, as well as the DNA shape, features of the sequence, the GC content of the surrounding regions, the concentration of other molecules and other context variables (79). These factors add additional layers of complexity that are not being totally captured in the developed methods, and need to be taken into account when interpreting the motif discovery and prediction results.

By understanding the limitations of the protocols and tools, we can explore further hypotheses and build models that could explain the biological phenomenon we are interested in. Therefore, advances toward integrating different types of data offer very promising approaches that will very likely increase the accurateness of the current TFBS models.

Acknowledgements

We thank CONACyT FOSISS (290041) for their support in the present work. MSA is grateful for a PhD fellowship funding from CONACYT (CVU659273) and the German Academic Exchange Service, DAAD (ref. 91693321). KTA is thankful for a Masters scholarship from CONACYT (CVU1009360). NA would like to acknowledge the Independent Research Fund Denmark (6108-2700038B) .

References

1. Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, et al. TFCat: the curated catalog of mouse and human transcription factors. Genome Biol. 2009 Mar 12;10(3):R29.

2. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009 Apr;10(4):252–63.

3. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013 Mar 14;152(6):1237–51.

4. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018 Feb 8;172(4):650–65.

5. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013 Jan 17;152(1-2):327–39.

6. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation in transcription factor binding. PLoS Genet. 2014 Mar;10(3):e1004226.

7. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved elements in the human genome. Science. 2004 May 28;304(5675):1321–5.

8. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57–74.

9. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. Scanning human gene deserts for long-range enhancers. Science. 2003 Oct 17;302(5644):413.

10. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010 Feb;2010(2):db.prot5384.

11. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012 Feb 5;482(7385):390–4.

12. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014 Apr 29;111(17):6131–8.

13. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34(Database issue):D108–10.

14. Smaczniak C, Angenent GC, Kaufmann K. SELEX-Seq: A method to determine DNA binding specificities of plant transcription factors. Methods Mol Biol. 2017;1629:67–82.

15. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009 Oct;10(10):669–80.

16. Collas P. The current state of chromatin immunoprecipitation. Mol Biotechnol. 2010 May;45(1):87–100.

17. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011 Dec 9;147(6):1408–19.

18. Fattori J, de Carvalho Indolfo N, de Oliveira Campos JCL, Videira NB, Bridi AV, Doratioto TR, et al. Investigation of Interactions between DNA and Nuclear Receptors: A Review of the Most Used Methods
. Nuclear Receptor Research. 2014;1:1–20.

19. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife [Internet]. 2017 Jan 16;6. Available from: http://dx.doi.org/10.7554/eLife.21856

20. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun. 2019 Apr 29;10(1):1930.

21. Matrix profile: CTCF - MA0139.1 - from JASPAR 2018 [Internet]. [cited 2020 May 12]. Available from: http://jaspar.genereg.net/matrix/MA0139.1/

22. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000 Jan;16(1):16–23.

23. Schneider TD. Consensus sequence Zen. Appl Bioinformatics. 2002;1(3):111–9.

24. Sandve GK, Abul O, Walseng V, Drabløs F. Improved benchmarks for computational motif discovery. BMC Bioinformatics. 2007 Jun 8;8:193.

25. Stormo GD. Modeling the specificity of protein-DNA interactions. Quant Biol. 2013 Jun;1(2):115–30.

26. Siddharthan R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. PLoS One. 2010 Mar 22;5(3):e9722.

27. Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. PLoS Comput Biol. 2013 Sep 5;9(9):e1003214.

28. Jayaram N, Usvyat D, R Martin AC. Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics. 2016 Nov 2;17(1):547.

29. Man T-K. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res. 2001;29(12):2471–8.

30. Hannenhalli S. Eukaryotic transcription factor binding sites--modeling and integrative search methods. Bioinformatics. 2008 Jun 1;24(11):1325–31.

31. Osada R, Zaslavsky E, Singh M. Comparative analysis of methods for representing and searching for transcription factor binding sites. Bioinformatics. 2004 Dec 12;20(18):3516–25.

32. Barash Y, Elidan G, Friedman N, Kaplan T. Modeling dependencies in protein-DNA binding sites. In: Proceedings of the seventh annual international conference on Research in computational molecular biology. New York, NY, USA: Association for Computing Machinery; 2003. p. 28–37. (RECOMB '03).

33. Bulyk ML, Johnson PLF, Church GM. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. 2002 Mar 1;30(5):1255–61.

34. King OD, Roth FP. A non-parametric model for transcription factor binding sites. Nucleic Acids Res. 2003 Oct 1;31(19):e116.

35. Hannenhalli S, Wang L-S. Enhanced position weight matrices using mixture models. Bioinformatics. 2005 Jun;21 Suppl 1:i204–12.

36. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020 Jan 8;48(D1):D87–92.

37. Hashim FA, Mabrouk MS, Al-Atabany W. Review of different sequence motif finding algorithms. Avicenna J Med Biotechnol. 2019;11(2):130–48.

38. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27(12):1653–9.

39. Sharov AA, Ko MSH. Exhaustive search for over-represented DNA sequence motifs with CisFinder. DNA Res. 2009;16(5):261–73.

40. Narlikar L, Ovcharenko I. Identifying regulatory elements in eukaryotic genomes. Brief Funct Genomic Proteomic. 2009;8(4):215–30.

41. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol. 2000 Mar;296(5):1205–14.

42. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28.

43. Das MK, Dai H-K. A survey of DNA motif finding algorithms. BMC Bioinformatics. 2007 Nov 1;8 Suppl 7:S21.

44. Dorigo M, Du Fnrs Marco Dorigo D de R, Stützle T. Ant Colony Optimization. MIT Press; 2004. 305 p.

45. Mercier E, Droit A, Li L, Robertson G, Zhang X, Gottardo R. An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. PLoS One. 2011;6(2):e16432.

46. Freund Y, Schapire RE. Large Margin Classification Using the Perceptron Algorithm. Mach Learn. 1999 Dec 1;37(3):277–96.

47. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. 800 p.

48. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. JMLR.org; 2017. p. 3145–53. (ICML'17).

49. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015 Oct;12(10):931–4.

50. Ma W, Noble WS, Bailey TL. Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nat Protoc. 2014;9(6):1428–50.

51. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. 2010;38(4):576–89.

52. Lee NK, Azizan FL, Wong YS, Omar N. DeepFinder: An integration of feature-based and deep learning approach for DNA motif discovery. Biotechnol Biotechnol Equip. 2018 May 4;32(3):759–68.

53. Bailey TL, Machanick P. Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res. 2012;40(17):e128.

54. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8(2):R24.

55. Yang J, Ma A, Hoppe AD, Wang C, Li Y, Zhang C, et al. Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. Nucleic Acids Res. 2019 Sep 5;47(15):7809–24.

56. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015 Aug;33(8):831–8.

57. Zhang Y, Qiao S, Ji S, Li Y. DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding. International Journal of Machine Learning and Cybernetics. 2020 Apr 1;11(4):841–51.

58. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016 Jul;26(7):990–9.

59. Bulyk ML. Computational prediction of transcription-factor binding site locations. Genome Biol. 2003 Dec 23;5(1):201.

60. Korhonen JH, Palin K, Taipale J, Ukkonen E. Fast motif matching revisited: high-order PWMs, SNPs and indels. Bioinformatics. 2017 Feb 15;33(4):514–21.

61. Aerts S. Chapter five - Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets. In: Plaza S, Payre F, editors. Current Topics in Developmental Biology. Academic Press; 2012. p. 121–45.

62. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011 Apr 1;27(7):1017–8.

63. Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 2003 Jul 1;31(13):3576–9.

64. Turatsinze J-V, Thomas-Chollier M, Defrance M, van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc. 2008;3(10):1578–88.

65. Frith MC, Li MC, Weng Z. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic Acids Res. 2003 Jul 1;31(13):3666–8.

66. Grant CE, Johnson J, Bailey TL, Noble WS. MCAST: scanning for cis-regulatory motif clusters. Bioinformatics. 2016 Apr 15;32(8):1217–9.

67. Lin T-H, Ray P, Sandve GK, Uguroglu S, Xing EP. BayCis: A Bayesian Hierarchical HMM for Cis-Regulatory Module Decoding in Metazoan Genomes. In: Research in Computational Molecular Biology. Springer Berlin Heidelberg; 2008. p. 66–81.

68. Glenwinkel L, Wu D, Minevich G, Hobert O. TargetOrtho: a phylogenetic footprinting tool to identify transcription factor targets. Genetics. 2014 May;197(1):61–76.

69. Loots GG, Ovcharenko I. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W217–21.

70. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol. 2004 Nov 30;5(12):R98.

71. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, et al. The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 2011 Jan;39(Database issue):D876–82.

72. Klepper K, Drabløs F. PriorsEditor: a tool for the creation and use of positional priors in motif discovery. Bioinformatics. 2010 Sep 1;26(17):2195–7.

73. Cheng C, Shou C, Yip KY, Gerstein MB. Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors. Genome Biol. 2011 Nov 7;12(11):R111.

74. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011 Mar;21(3):447–55.

75. Klepper K, Drabløs F. MotifLab: a tools and data integration workbench for motif discovery and regulatory sequence analysis. BMC Bioinformatics. 2013 Jan 16;14:9.

76. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018;46(D1):D252–9.

77. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158(6):1431–43.

78. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ. Deep and wide digging for binding motifs in ChIP-Seq data. Bioinformatics. 2010;26(20):2622–3.

79. Inukai S, Kock KH, Bulyk ML. Transcription factor–DNA binding: beyond binding site motifs. Curr Opin Genet Dev. 2017;43:110–9.