UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

ESTUDIO DE LA COMPENSACIÓN DE LA EXPRESIÓN GENÉTICA DE LOS CROMOSOMAS SEXUALES EN LA LAGARTIJA VERDE (*Anolis carolinensis*)

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
BIOL. MARIELA TENORIO PEREZ

DIRECTOR DE TESIS:
DR. CLAUDIO DIEGO CORTEZ QUEZADA
BIOLOGÍA DE SISTEMAS

COMITÉ TUTOR:

DR. ARTURO CARLOS II BECERRA BRACHO
ORIGEN DE LA VIDA, DEPARTAMENTO DE BIOLOGÍA EVOLUTIVA

DR. LUIS DAVID ALCARAZ PERAZA
LABORATORIO DE GENÓMICA AMBIENTAL

CUERNAVACA, MORELOS. OCTUBRE DE 2023

ÍNDICE

RESUMEN

Los cromosomas sexuales son un par de cromosomas asociados con la determinación del sexo de un individuo. En muchas especies los cromosomas sexuales son del tipo XY, donde los machos poseen un cromosoma único y diferente, el cromosoma Y. La falta de recombinación homóloga entre el cromosoma Y con su contraparte el X ha impedido la purga de mutaciones deletéreas, inversiones y la pérdida de grandes fragmentos de material genético en el cromosoma Y. Esta pérdida acelerada de material genético en el cromosoma Y provocó un desbalance de dosis génica entre ambos sexos ya que las hembras se quedaron con dos cromosomas X activos, mientras que los machos se quedaron sólo con X, más un Y altamente degenerado. En la actualidad, en diferentes especies se han descrito mecanismos que equilibran el desbalance de expresión entre cromosomas X. En mamíferos placentarios y marsupiales, por ejemplo, se silencia transcripcionalmente un cromosoma X en hembras, fenómeno que se conoce como “Inactivación del cromosoma X”. Por otro lado, en *Drosophila*, el complejo proteínico MLS incrementa la transcripción del cromosoma X en los machos. Se sabe que los RNA largos no codificantes (long non-coding RNA, lncRNA) son factores que regulan directamente la compensación de dosis de los cromosomas sexuales en todas estas especies. El lncRNA *XIST* (mamíferos placentarios) y *RSX* (marsupiales) participan en el silenciamiento del cromosoma sexual X en hembras. Por el contrario, el lncRNA *ROX2* (*Drosophila melanogaster*), participa en la sobreexpresión del cromosoma X de los machos. En esta tesis describo un cuarto sistema orquestado por un nuevo lncRNA el cual nombré *MAYEX*, por Male-specific long non-coding RNA AmplifYing the Expression of the X. Este lncRNA está localizado en el cromosoma X de la lagartija verde, *Anolis carolinensis*. Este es el primer lncRNA descrito en reptiles que regula por completo el cromosoma X de los machos, aumentando al doble su nivel de expresión. Las similitudes entre los sistemas descritos en los mamíferos, la mosca de la fruta y ahora en la lagartija verde, indican que los lncRNA han sido seleccionados en especies distantes durante la evolución resolviendo los mecanismos de regulación en *cis* a lo largo del cromosoma X, controlando los niveles de expresión del cromosoma completo y restableciendo el equilibrio de la expresión entre machos y hembras. Este nuevo estudio abre una amplia gama de oportunidades para estudiar qué otros factores ayudan a regular el cromosoma X de *A. carolinensis*.

ABSTRACT

Sex chromosomes are a pair of chromosomes associated with determining the sex of an individual. In many species the sex chromosomes are of the XY type, where males have a unique and different chromosome, the Y chromosome. The lack of homologous recombination between the Y chromosome with its counterpart the loss of large fragments of genetic material on the Y chromosome. This accelerated loss of genetic material on the Y chromosome caused an imbalance of gene dosage between both sexes since females were left with two active X chromosomes, while males were left with only one plus a highly degenerate Y. Currently, several mechanisms that balance the expression imbalance between X chromosomes have been described in different species. In placental mammals and marsupials, for example, an X chromosome is transcriptionally silenced in females, a phenomenon known as “X chromosome inactivation”. On the other hand, in *Drosophila*, the MLS protein complex increases the transcription of the X chromosome in males. Long non-coding RNAs (lncRNAs) are known to be factors that directly regulate dosage compensation of sex chromosomes in all these species. The lncRNA *XIST* (placental mammals) and *RSX* (marsupials) participate in the silencing of the X sex chromosome in females. On the contrary, the lncRNA *ROX2* (Drosophila melanogaster) participates in the overexpression of the X chromosome in males. In this thesis I describe a fourth system orchestrated by a new lncRNA which I named *MAYEX*, for Male-specific long non-coding RNA Amplifying the Expression of the X. This lncRNA is located on the X chromosome of the green lizard, *Anolis carolinensis*. This is the first lncRNA described in reptiles that completely regulates the X chromosome of males, doubling its expression level. The similarities between the systems described in mammals, the fruit fly and now in the green lizard, indicate that lncRNAs have been selected in distant species during evolution that help to resolve cis-regulation mechanisms along along the X chromosome, controlling expression levels of the entire chromosome and restoring the balance of expression between males and females. This new study opens a wide range of opportunities to study what other factors help regulate the *A. carolinensis* X chromosome.

INTRODUCCIÓN

La predicción de Susumu Ohno

Susumu Ohno fue uno de los grandes genetistas moleculares de la era moderna. Su vasto trabajo también abarcó a los cromosomas sexuales. Ohno se dio cuenta de que el tamaño de los cromosomas sexuales variaba entre machos y hembras. En el caso de las hembras, éstas presentan dos cromosomas X del mismo tamaño, pero en los machos, el cromosoma Y es de un tamaño menor al del cromosoma X (Figura 1). Debido a esta variación en cuanto a tamaño, Ohno planteó una serie de interrogantes: ¿A las hembras les sobra material genético?, o ¿a los machos les falta material genético? En su libro “Sex Chromosomes and Sex-Linked Genes” publicado en 1967, propuso que los cromosomas inicialmente fueron un par de cromosomas autosomales y subsecuentemente – en el caso de los machos – se diferenciaron a X y Y como los conocemos hoy en día (Ohno, 1967; Beutler, 1998). Ohno predijo que debería de haber un mecanismo que compensara el desbalance en el número de genes de los cromosomas sexuales entre machos y hembras. Ohno propuso que probablemente un cromosoma X en hembras se inactivaba.

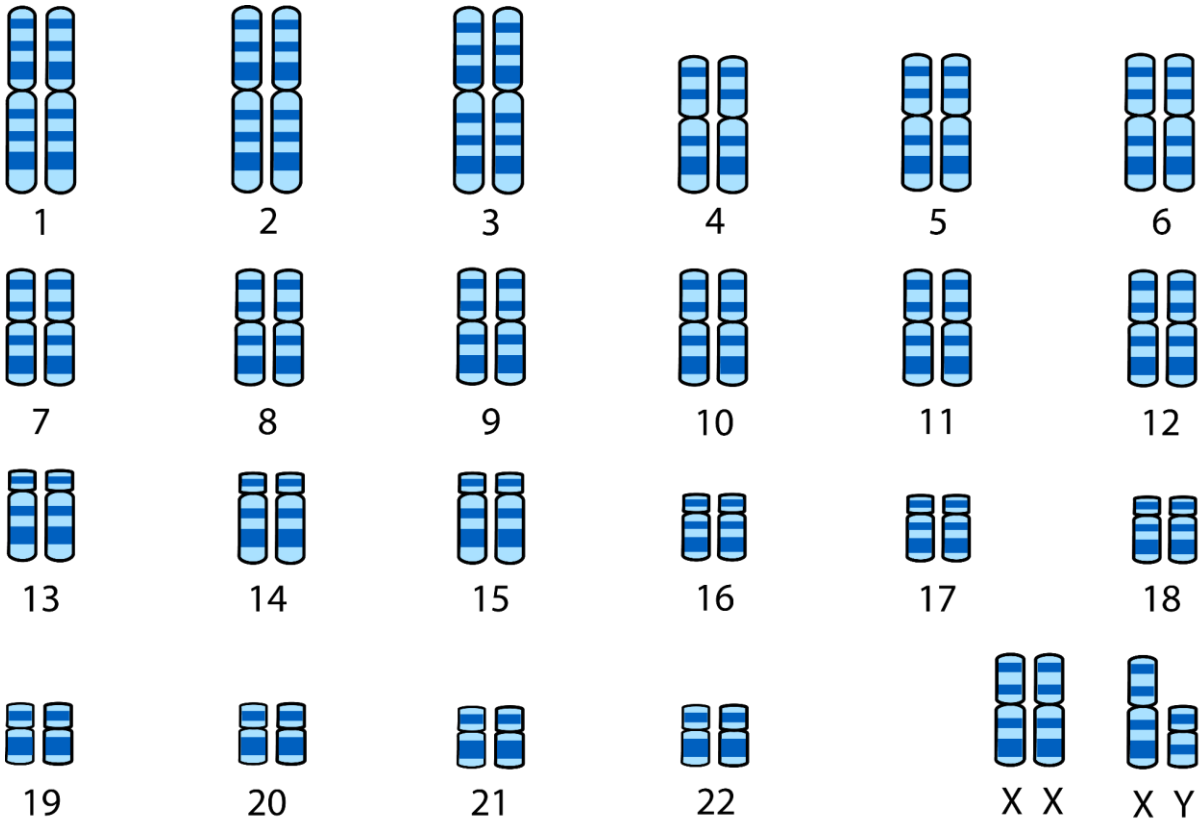


Figura 1. Cariotipo del humano con su pareja de cromosomas sexuales XY.

La teoría de Müller: el origen de los cromosomas sexuales

Los cromosomas sexuales son un par de cromosomas que determinan el sexo de un individuo. La teoría de Hermann Joseph Müller sugiere que, en un principio, un cromosoma autosomal adquirió un gen maestro capaz de activar la cascada de señalización determinante del sexo en machos (Bachtrog, 2013). La adquisición de este nuevo gen determinante del sexo produjo que un cromosoma autosomal pasara a ser un cromosoma sexual. En el caso de los mamíferos, el gen responsable de que el individuo se desarrolle como un macho se conoce como *SRY* y está localizado en el cromosoma Y (Marais y Galtier, 2003).

Una vez que se originaron los cromosomas sexuales (XY), la teoría postula que hubo un rearreglo cromosomal en el cromosoma Y, como sería el caso de la inversión o la translocación de una región dentro del mismo cromosoma. Este rearreglo causó que los cromosomas X y Y dejaran de recombinar, al menos en esta región, ya que el orden de los genes dentro del cromosoma Y habría cambiado respecto al cromosoma X (Bachtrog, 2013). Hoy se sabe que los cromosomas X y Y recombinan únicamente en las regiones “pseudoautosomales”, ubicados en las regiones terminales de los brazos p y q, donde todavía mantienen la misma secuencia sinténica (Marais y Galtier 2003).

La recombinación homóloga entre dos cromosomas permite que ambos eliminen mutaciones deletéreas y corrijan deleciones (Bachtrog, 2013; Snell y Turner 2018). La falta de recombinación en el cromosoma Y condujo a que este cromosoma no pudiera corregir fielmente las rupturas de doble hebra y tendiera a perder una gran cantidad de material genético (Gatler, 2014).

Podemos hablar entonces de que los cromosomas Y son una versión degenerada de los cromosomas X. Además, los cromosomas Y degenerados están enriquecidos en una gran cantidad de DNA repetido. En el caso de los humanos, el cromosoma Y mide 57 Mb, contiene 63 genes que codifican para proteínas y el 70% de su secuencia está formada por elementos repetidos (https://www.ensembl.org/Homo_sapiens/Location/Chromosome?r=Y%3A1-1000). Mientras que el cromosoma X mide 156 Mb y contiene 859 genes que codifican para proteínas

(https://www.ensembl.org/Homo_sapiens/Location/Chromosome?r=X:86964745-87064745). En el caso de *Drosophila melanogaster*, el cromosoma Y mide 40 Mb y contiene únicamente 13 genes que codifican para proteínas, mientras que el cromosoma X mide 22 Mb y contiene 2200 genes (Bachtrog, 2013).

La pérdida acelerada de material genético en el cromosoma Y provocó un desbalance de dosis génica entre ambos sexos: las hembras se quedaron con dos cromosomas X activos, mientras que los machos se quedaron sólo con uno. Concomitantemente aparecieron mecanismos de compensación de dosis que restablecieron el balance en los niveles de expresión genética de los cromosomas X entre machos y hembras (Mank, 2013).

La compensación de dosis entre los cromosomas sexuales

La compensación de dosis es un mecanismo regulador que afecta a todo el cromosoma (Mank, 2013). Su función es igualar la expresión de genes entre machos y hembras.

La dosificación de un cromosoma hace referencia al número de copias de un gen en el genoma. La duplicación o eliminación de pequeñas regiones cromosomales conlleva a presentar variaciones en el número de copias (CNV por sus siglas en inglés) (Ercan, 2015). Los cambios en el número de copias de un gen pueden causar problemas en el organismo, ya que los niveles de RNA mensajero (RNAm) y la abundancia de las proteínas son proporcionales al número de copias de cada gen (Brockdorff y Turner, 2015). De hecho, la alteración en el número de copias de un gen es un mecanismo de regulación de su función (Ercan, 2015).

En la actualidad, se han descrito varios mecanismos que equilibran el desbalance de expresión entre cromosomas X en diferentes especies (Figura 2) (Payer y Lee, 2008; Conrad y Akhtar, 2012).

En mamíferos placentarios, por ejemplo, se silencia transcripcionalmente un cromosoma X en hembras; fenómeno que se conoce como “Inactivación del cromosoma X”. Es un proceso que ocurre en el desarrollo temprano en las hembras y que proporciona una equivalencia de dosificación entre machos y hembras (Figura 2) (Payer y Lee, 2008).

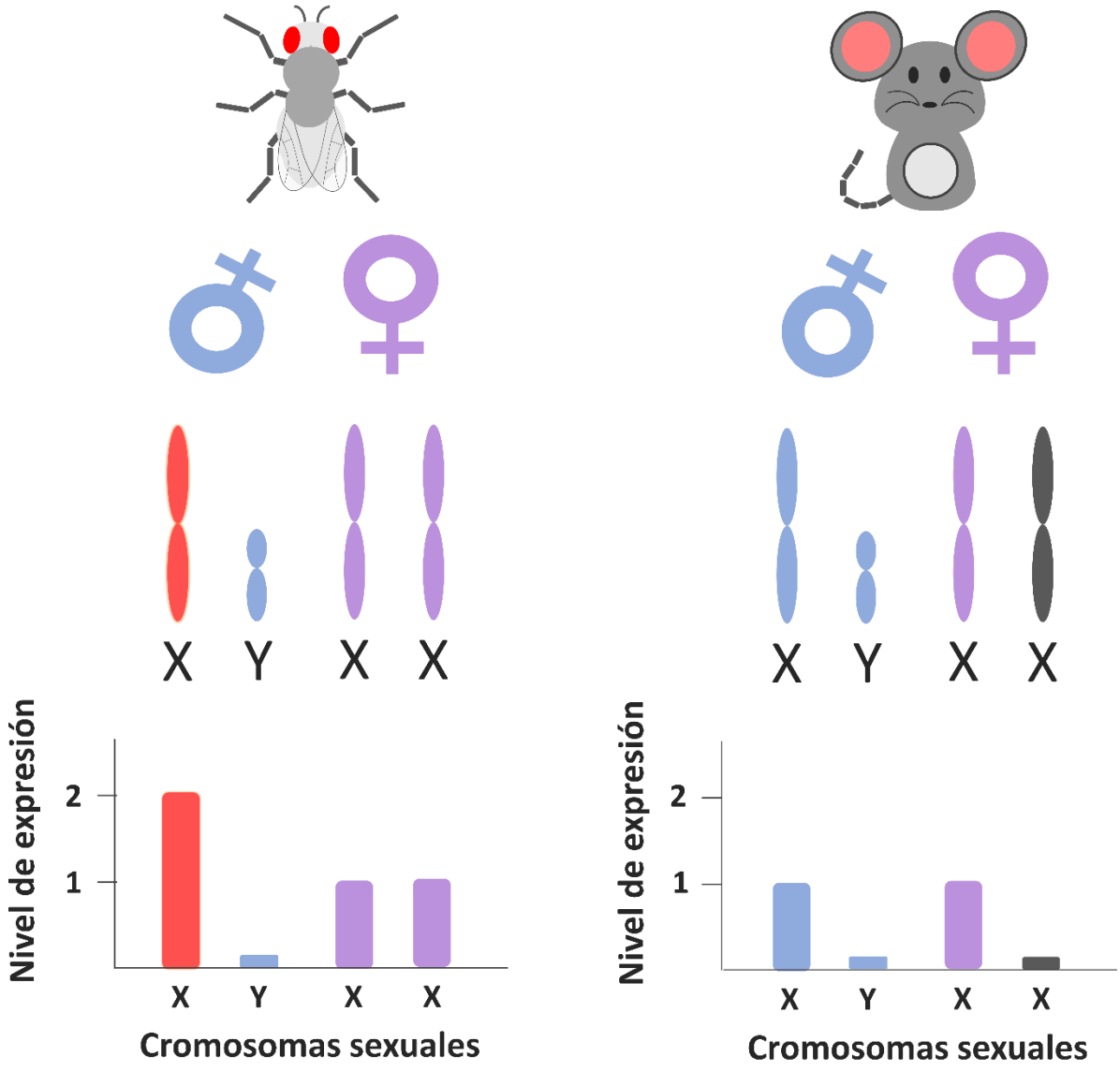


Figura 2. Tipos de compensación de dosis en sistemas XY.

En 1949, Barr y Bertram describieron en mamíferos una estructura heterocromática en el núcleo de las hembras; ellos la denominaron “Corpúsculo de Barr” (Barr body) (Barr y Bertram, 1949). Años más tarde, Susumu Ohno propuso que esta estructura pudiera ser un cromosoma altamente condensado e inactivo (Beutler, 1998). Hoy sabemos que esta estructura es en efecto un cromosoma X silenciado.

En el caso de los machos, el Y transcribe algunos genes que están compartidos con el cromosoma X; estos genes son llamados gametólogos. Este nuevo problema de dosis se resuelve cuando los gametólogos dentro del X inactivo escapan parcialmente de la inactivación. Esto permite mantener un equilibrio en los niveles de transcripción

de genes tanto en el X como en el Y (Brown, et al. 1997; Richardson, 2010; Shapiro, et al. 1979).

Otro modelo ampliamente estudiado es la mosca de la fruta. En *Drosophila*, el complejo proteínico MLS incrementa la transcripción del cromosoma X en los machos (Figura 2). La sobreexpresión del cromosoma X es un proceso de desarrollo temprano en los machos (XY) de *D. melanogaster* que activa transcripcionalmente la mayoría de los genes a lo largo del cromosoma X lo que hace que coincida con la producción de las hembras con dos cromosomas (XX). El cromosoma X, adquiere una estructura que permite una mejor accesibilidad de la maquinaria de transcripción, lo que determina una doble tasa de transcripción por fragmento de DNA en el cromosoma X de los machos (Conrad y Akhtar, 2012).

Por otra parte, se sabe que los RNA largos no codificantes (long non-coding RNA, lncRNA) son elementos genéticos que regulan directamente la compensación de dosis de los cromosomas sexuales en mamíferos placentarios y en la mosca de la fruta (Ercan, 2015).

Los RNA largos no codificantes

Los lncRNA se definen por ser transcritos superiores a los 200 nucleótidos que no codifican para proteínas, se transcriben gracias a la polimerasa II, se empalman y en ocasiones se poliadenilan (Fatica y Bozzoni, 2014). Hasta la fecha, en el genoma humano se han descrito aproximadamente 14,880 lncRNA, y de todos, únicamente 298 presentan una función conocida (Derrien et al., 2012). Estos están implicados en la regulación postraduccional de genes a través del control de procesos como la síntesis de proteínas, la maduración del RNA, el transporte de proteínas y la activación o silenciamiento de genes a través de la regulación de la estructura de la cromatina (Marchese et al., 2017).

Varios mecanismos de acción están involucrados en la función de los lncRNA los cuales se clasifican en cuatro categorías: señales (marcadores), señuelos, guías y andamios (Wang y Chang 2011).

En la primera categoría, cuando actúan como señales, los lncRNA intervienen en algunas vías de señalización. Al igual que los demás genes, su transcripción ocurre

en un momento y lugar específico lo que les permite integrar señales de desarrollo, interpretar el contexto celular o responder a diversos estímulos. Algunos lncRNA que actúan como señales son: *XIST*, presente en mamíferos, que actúa en mecanismos de regulación epigenética, silenciando alelos e interviniendo en la modificación de histonas; *HOTAIR* y *HOTTIP*, presentes en los mamíferos, están implicados en la señalización de la posición anatómica; *PANDA*, que desempeña un papel regulador en la respuesta transcripcional de p53; *COLDAIR* y *COOLAIR* en plantas están implicados en la vernalización silenciando epigenéticamente los represores florales (Wang y Chang 2011; Engreitz et al., 2013; Rao 2017).

En la segunda categoría, cuando actúan como señuelos, los lncRNA pueden regular negativamente la transcripción de un efector, atraer al DNA proteínas que pueden ser factores de transcripción, modificadores de la cromatina u otros factores reguladores. Algunos lncRNA que actúan como señuelos son: *TERRA*, que protege la longitud de los telómeros; *PANDA*, que inhibe la expresión de genes apoptóticos y *MALAT1*, que al bajar su nivel de expresión induce la formación de sinapsis (Wang y Chang 2011; Engreitz et al., 2013; Rao 2017).

En la tercera categoría, cuando actúan como guías, los lncRNA pueden reclutar enzimas modificadoras de la cromatina hacia genes específicos, ya sea en *cis* como los lncRNA *XIST*, *COLDAIR* y *HOTTIP* o en *trans* como el lncRNA *HOTAIR* (Wang y Chang 2011; Engreitz et al., 2013; Rao 2017).

Por último, en la cuarta categoría, cuando actúan como andamios, los lncRNA pueden reunir proteínas múltiples para formar complejos ribonucleoprotéicos. Estos complejos pueden actuar sobre la cromatina y afectar las modificaciones de las histonas. Algunos lncRNA que actúan como andamios son: *TERC*, que desempeña un papel fundamental en el mantenimiento de la estabilidad del genoma mediante la adición de repeticiones de DNA en las regiones teloméricas y *HOTAIR*, que se une al complejo Polycomb y promueve la represión genética (Wang y Chang 2011; Engreitz et al., 2013; Rao 2017).

Hasta la fecha, se conocen algunos lncRNA que activan la cascada de señalización que desencadena la compensación de dosis de los cromosomas sexuales entre ambos sexos; éstos son: *XIST* (mamíferos placentarios), *RSX* (marsupiales) y *ROX*-

2 (la mosca de la fruta) (Quinn y Chang, 2015). A continuación, se describe lo que se conoce sobre la función de *XIST* y *ROX-2*.

LncRNA XIST

El lncRNA *XIST* es un RNA de 17 a 19 kb de largo, presenta patrones de repetición y consiste en una región A que contiene 8 repeticiones separadas por espaciadores ricos en Uracilo, la cual presenta dos estructuras largas en forma de tallo y bucle. El gen de *XIST*, se encuentra en el brazo largo del cromosoma X (Sado, 2017).

XIST participa en el silenciamiento del cromosoma sexual X en hembras. El proceso está regulado por varios factores, incluido el centro de inactivación del X (X inactivation center, XIC), donde se encuentran presentes todos los genes que participan en el silenciamiento del cromosoma (Engreitz *et al*., 2013; Sado, 2017).

Un factor requerido antes del silenciamiento es la proteína YY1, la cual se une a la región promotora del gen *XIST* y activa su transcripción. Una vez que *XIST* adopta su estructura, llega la proteína SF2 y se une con alta especificidad a la región A del lncRNA con lo que procesa eficientemente a *XIST* (Figura 3). Después, *XIST* recluta directa e indirectamente al grupo Polycomb, el cual realiza modificaciones en histonas. Existen dos grupos, Complejo Represivo Polycomb 1 y 2 (Polycomb Repressive Complex, PRC1 y PRC2) (Figura 3). Ambos presentan varios complejos de proteínas que se unen a *XIST*. La región A interacciona con SPEN, el complejo WTAP-RBM15-RBM15B y el receptor de la lámina B. Las proteínas RBM15 y RBM15B interaccionan con la proteína METTL3 para la metilación de adeninas en RNA. El complejo PCR2 por medio de la histona-lisina metiltransferasa EZH2 son los responsables en la trimetilación de la lisina 27 de la histona 3 (H3K27me3) (Figura 3) (Fiskus et al, 2006; Brockdorff 2017; Creamer y Lawrence, 2017; Lu et al., 2017; Sado 2017; Kaufmann y Wutz, 2023).

Algo que debemos de tomar en cuenta es que cada transcrito de *XIST* atrae a las proteínas anteriormente mencionadas y recubre completamente el cromosoma X inactivo a partir de XIC. Al final, este cromosoma se compacta y se reorganiza estructuralmente en la periferia nuclear (Figura 3) (Brockdorff 2017; Creamer y Lawrence, 2017; Lu et al., 2017; Sado 2017).

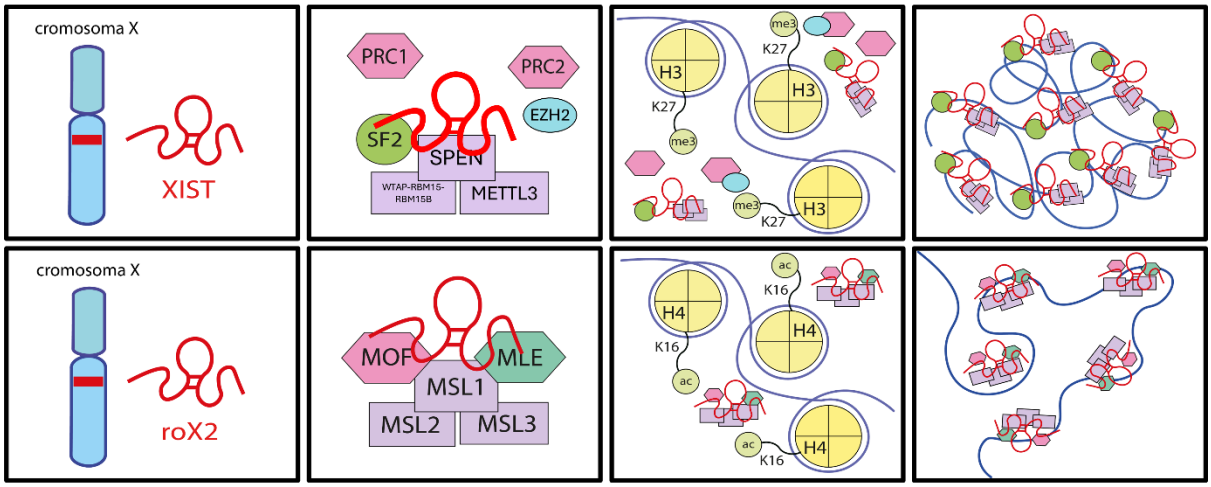


Figura 3. Mecanismo de acción de *XIST* y *roX2* en mamíferos placentarios y la mosca de la fruta respectivamente.

LncRNA ROX2

El lncRNA *ROX2* participa en la sobreexpresión del cromosoma X. Este proceso ocurre durante el desarrollo temprano de los machos de *D. melanogaster*, donde se activan transcripcionalmente los genes del cromosoma X que tengan el mismo nivel de expresión que tienen en las hembras (XX). En este proceso están involucrados el complejo de compensación de dosis (Dosage Compensation Complex, DCC), el cual está formado por las proteínas: Male-Specific Lethal (MSL) 1, 2 y 3, la proteína Male Absent On the First (MOF) y por último la proteína Maleess (MLE) y por los lncRNA: *ROX*-1 y -2 (Figura 3) (Straub et al., 2005; Conrad y Akhtar, 2012).

Primero, la proteína MSL2 se une y estabiliza a MSL1. Esta última proteína contiene un dominio llamado PEHE, que sirve como andamio de ensamblaje para reclutar tanto a MOF como a la proteína MSL3. Posteriormente, gracias a este ensamblaje y a la helicasa MLE, MSL2 activa la transcripción de los genes *ROX*1 y 2. La incorporación de ambos lncRNA depende de MLE, la cual remodela la estructura de los lncRNAs a una configuración repetitiva de bucle-tallo-bucle logrando una mayor afinidad de unión hacia MLE y MSL2. Por último, el complejo y los lncRNA se extienden en cis, recubriendo todo el cromosoma X y acetilando la lisina 16 de la histona 4 (Figura 3). En este punto el cromosoma adquiere una estructura que permite una mejor accesibilidad de la maquinaria de transcripción mediada por la RNA polimerasa II para transcribir los genes, lo que determina una doble tasa de transcripción por fragmento

de DNA en el cromosoma X de los machos (Morales et al., 2005; Straub et al., 2005; Long, et al. 2017).

Recientemente se desarrolló un protocolo experimental capaz de atrapar e identificar los RNAs que están asociados a la cromatina; a esta técnica se le conoce como Chromatin-Associated RNA sequencing, o por sus siglas, ChAR-seq (Bell et al., 2018; Juckam et al., 2019).

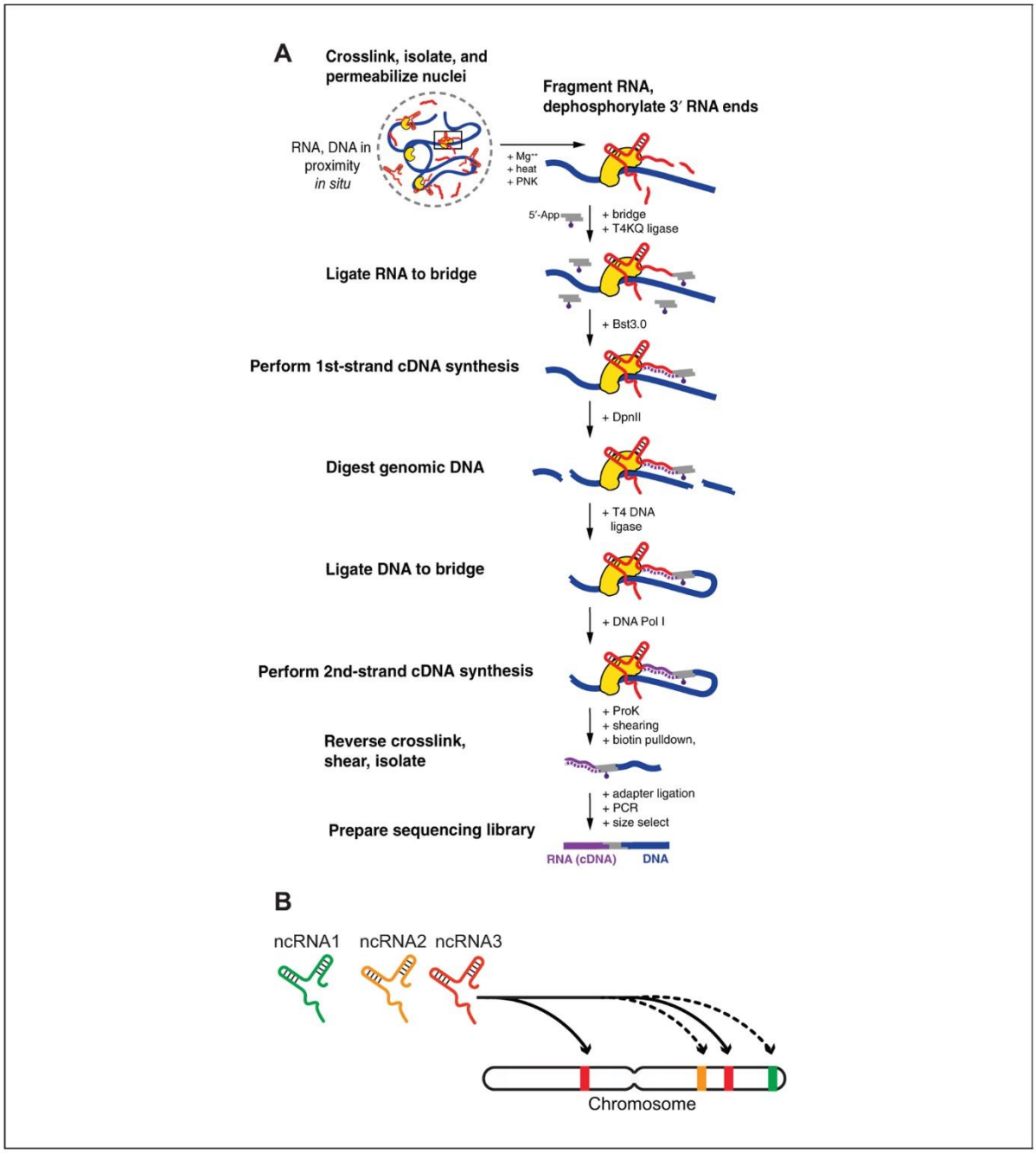


Figura 4. Esquema del experimento ChAR-seq.
(imagen obtenida de Jukam et al., 2019)

Chromatin-Associated RNA sequencing (ChAR-seq)

La secuenciación de RNA asociado a la cromatina es un método de captura que mapea los contactos de RNA con DNA de todo el genoma (Bell et al., 2018; Juckam et al., 2019) y sirve para identificar a gran escala los RNA asociados a la cromatina (Figura 4) (Bell et al., 2018). Este método fija todas las interacciones de la cromatina y posteriormente, liga *in situ* los extremos de los RNAs y DNAs que interactúan directamente o muy cerca en el espacio nuclear. La ligación se hace por medio de una secuencia de doble cadena, la cual denominan “puente” (Figura 4) (Juckam et al., 2019). El extremo 5’ del puente contiene una secuencia monocatenaria adenilada, que se liga a extremos 3’ de RNA y el extremo 3’ del puente contiene un sitio de reconocimiento para fragmentos de DNA que han sido digeridos por la enzima DpnII. Esta enzima corta fragmentos GATC, los cuales empalman perfectamente en el “puente” (Figura 4). Después de la ligación de ambos fragmentos, se forma la molécula quimérica (RNA-puente-DNA) (Figura 4). La asimetría de la secuencia del puente permite identificar claramente el RNA y DNA originales después de la secuenciación. Una de las ventajas de ChAR-seq, es que preserva la organización tridimensional del genoma (Jukam et al., 2019).

La organización del genoma

Se sabe que el genoma se mantiene condensado dentro del núcleo y que está organizado jerárquicamente en diferentes niveles de empaquetamiento (Finn y Misteli 2019). Si comenzamos a desempaquetar cada nivel desde lo más complejo hasta lo más simple, encontraremos primero a los cromosomas dentro de su propio territorio, aunque en las fronteras de cada uno estarán interactuando con otros cromosomas (Misteli 2008). Dentro de cada cromosoma, existen diferentes plegamientos, los cuales constituyen el siguiente nivel de empaquetamiento y se denominan compartimientos. Estos dependen del tipo de modificación (metilación o acetilación) que tengan las histonas (Rice y Allis 2001). Las regiones transcripcionalmente inactivas conforman el compartimiento B y las regiones transcripcionalmente activas conforman el compartimiento A (Gibcus y Dekker 2013; Rowley y Corcer, 2018). A su vez, dentro de cada compartimiento se presenta un nivel más denominado como Dominios de Asociación Topológica (Topologically Associating Domains, TADs) (Gibcus y Dekker 2013; Lu et al., 2019). Normalmente, los TADs miden entre 200 kb

a 1 Mb y están compuestos de varios loops que conectan genes y enhancers a regiones que están linealmente lejos pero espacialmente cerca (Gibcus y Dekker 2013; Finn y Misteli 2019). El siguiente nivel sería la estructura de un solo loop, la cual está conformada por varios nucleosomas y en cada nucleosoma encontramos al DNA enrollado sobre las histonas. El último nivel es la doble cadena del DNA (Finn y Misteli 2019).

Un método experimental que puede identificar los fragmentos de DNA que estén espacialmente cerca pero linealmente lejos se conoce como Hi-C (Lieberman-Aiden, et al., 2009).

Hi-C: Método de captura de conformación de la cromatina

El método de captura de conformación del genoma es el primer método molecular que captura las interacciones físicas de la cromatina. Se han desarrollado variantes de este método y la mayoría sólo captura patrones específicos de interacción a pequeña escala. En cambio, el método Hi-C examina todas las interacciones de todo el genoma en un solo experimento (Dekker, et al., 2002). Hi-C introduce nucleótidos biotinilados en las uniones de ligadura que permiten la purificación específica de estas uniones (Lajoie, et al., 2015).

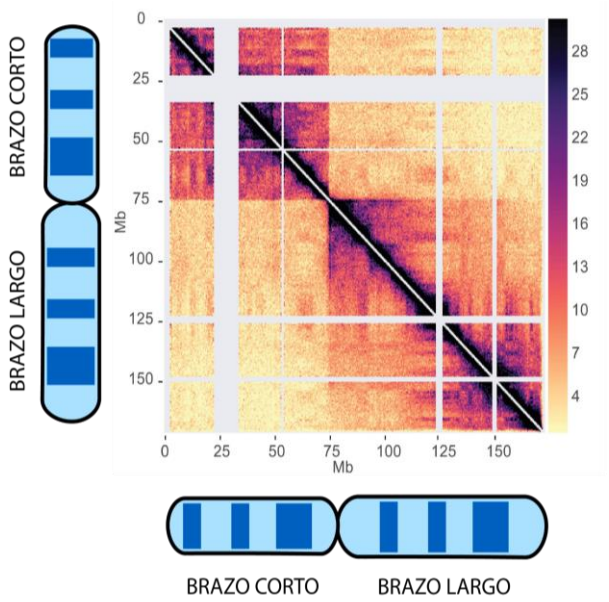


Figura 5. El cromosoma X inactivo del humano está estructurado en dos enormes TADs (modificado de Fang, et al., 2019)

Da información de la organización espacial del genoma y cómo los cromosomas controlan la expresión de genes. Utilizando la técnica de Hi-C en células Patski de ratón podemos observar que el cromosoma X en hembras se condensa en dos grandes super dominios (Figura 4) (Fang et al., 2019). Como se mencionó anteriormente, el cromosoma se compacta, lo que evita que las RNA polimerasas puedan transcribir los genes (Engreitz et al. 2013).

Modelo de estudio: Anolis carolinensis

A. carolinensis pertenece a la familia Iguanidae y su nombre común es Anolis verde (Green anole) (Figura 5) (Álvarez-Romero, et al., 2005). Es una lagartija pequeña con una cola y garras largas. Los machos miden de 12.5 a 20 cm, mientras que las hembras miden 12.5 cm aproximadamente (Smith, 2001). Una de las características distinguibles de los machos de la lagartija verde es su hocico puntiagudo. Otra es que presentan un saco gular de color rosa, mientras que las hembras no, las cuales presentan la garganta de color rosa pálido (Figura 5) (Álvarez-Romero, et al., 2005). La coloración de su piel varía según la temperatura, humedad, salud y estado de ánimo, variando entre verde, café y grisáceo (Smith, 2001). Su distribución va desde el sureste de Estados Unidos, específicamente en el este de Texas hasta el sur de Virginia (Smith, 2001) aunque también hay reportes en una localidad del estado de Tamaulipas (Álvarez-Romero, et al., 2005). Esta especie está adaptada en bosques templados o tropicales y se encuentra comúnmente perchada en postes, bardas y troncos de árboles con la cabeza hacia abajo. Es un animal solitario, diurno, que defiende su territorio agresivamente y es la única especie de anolis de clima templado y nativa de Estados Unidos.

Anolis presenta un sistema de determinación sexual ligado a cromosomas el que hembras son homogaméticas (XX) y los machos son heterogaméticos (XY) (Alföldi, et al., 2011; Marín, Cortez *et al*., 2017). El sistema de determinación sexual es muy antiguo, ya que data de entre 160 y 170 millones de años (Marín, Cortez *et al*., 2017). Con el paso del tiempo, el cromosoma sexual Y se ha degenerado y hoy parece tener solo 7 genes de los 350 genes que tenía originalmente (Marín, Cortez *et al*., 2017).

Figura 6. Hembra (izquierda) y macho (derecha) *A. carolinensis* respectivamente.

(fotos obtenidas de White Python y Zooplus Magazine).

JUSTIFICACIÓN

En la mayoría de las especies con cromosomas sexuales que presentan un sistema XY, las hembras tienen un par de cromosomas X idénticos, mientras que, en los machos, la pareja está formada por un cromosoma X y un cromosoma Y degenerado. De esta forma, en comparación con el X, el Y es un cromosoma pequeño. Por ello, si el cromosoma Y tiene pocos genes, entonces faltan genes en los machos y si el X es muy grande, entonces sobran genes en las hembras.

En mamíferos, como ya se mencionó con anterioridad, el problema se resuelve por la inactivación de un cromosoma X en hembras mediada por el lncRNA *XIST* que sólo se expresa en el cromosoma que se inactiva (Lu et al., 2017). En *D. melanogaster*, el lncRNA *ROX2* sirve de andamio para el complejo MSL con lo que cromosoma X en los machos aumente su expresión al doble (Conrad and Akhtar 2012).

El reptil *A. carolinensis* tiene cromosomas XY, con un Y altamente degenerado. En un trabajo previo, se especuló que los *Anolis* presentan un mecanismo de compensación de dosis resultado de la pérdida masiva de genes en el cromosoma Y (Marín, Cortez *et al*., 2017). El Dr. Diego Cortez y sus colegas encontraron que *A. carolinensis* coincide a la perfección con las predicciones de Ohno. Es decir, los cromosomas X se expresan de igual manera en ambos sexos, pese a que los machos tienen solo un cromosoma X. Esta lagartija consigue esto expresando al doble el cromosoma X en el macho a través de la hiper-acetilación de la lisina 16 de la histona 4, lo que representa un maravilloso ejemplo de evolución convergente con la mosca de la fruta (Marín, Cortez *et al*., 2017). Sin embargo, se desconocen por completo los factores ligados a la posible regulación del cromosoma X. Este sería el primer caso de regulación cromosomas descrito en reptiles. Sus similitudes y diferencias con otros sistemas nos permitirán entender la evolución de los procesos de compensación de dosis.

HIPÓTESIS

Existe un lncRNA en el cromosoma X de *A. carolinensis* cuya actividad en *cis* promueve la sobreexpresión del cromosoma X específicamente en los machos restableciendo con ello el balance de los niveles de expresión entre ambos sexos.

CAPÍTULO 1

El artículo “MAYEX is an old long non-coding RNA recruited for X chromosome dosage compensation in lizards” se centra en la identificación de un lncRNA en *A. carolinensis* que regula la expresión del cromosoma X en machos. A este lncRNA lo denominamos “Male-specific long non-coding RNA AmplifYing the Expression of the X” (*MAYEX*)”. Su expresión en el cromosoma X es detectada desde los 2 o 3 días del desarrollo del embrión hasta su etapa adulta en todos los tejidos. Utilizando y analizando datos de ChIP-seq se detectaron elevadas marcas de acetilación en todo el cromosoma X y logramos identificar que el locus de *MAYEX* mostró los más altos niveles de acetilación de todo el genoma. En este artículo, se utilizó la técnica Hi-C para obtener información sobre la estructura de la cromatina del cromosoma X. No logramos ver una diferencia entre el cromosoma X de los machos con respecto a las hembras, pero detectamos una gran frecuencia de contactos río abajo del locus de *MAYEX*. En este proyecto, se realizó la técnica ChAR-seq diseñada para capturar moléculas de RNA asociadas a la cromatina, la cual nos mostró que *MAYEX* está en contacto directo con múltiples regiones de DNA distantes en el cromosoma X. Por último, encontramos secuencias ortólogas de *MAYEX* en el cromosoma sexual X de otras especies. Este trabajo está bajo revisión en la revista Science.

MAYEX is an old long non-coding RNA recruited for X chromosome dosage compensation in lizards.

Mariela Tenorio[1], Samantha Cruz-Ruiz[2], Jose Antonio Corona-Gomez[3], Mario Zurita[2], Fania Santiago[1], Fausto Méndez-de-la-Cruz[4], Joanna Serwatowska[5], Selene L. Fernandez-Valverde[3], Katarzyna Oktaba[5], Diego Cortez[1]

1. Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México (UNAM), CP62210, Cuernavaca, México.

2. Instituto de Biotecnología, Universidad Nacional Autónoma de México (UNAM), CP62210, Cuernavaca, México.

3. Centro de Investigación y de Estudios Avanzados del IPN, Unidad Irapuato, Irapuato, México.

4. Instituto de Biología, Universidad Nacional Autónoma de México (UNAM), CU, CP04510, Ciudad de México, México.

5. Unidad de Genómica Avanzada, Centro de Investigación y de Estudios Avanzados del IPN, Irapuato, México.

Keywords: X chromosome; Long non-coding RNAs; Dosage compensation mechanisms; *Anolis carolinensis*; Reptiles.

Abstract

Long non-coding RNAs (lncRNAs) are important regulatory elements of sex chromosomes. For example, the *XIST* and *RSX* genes regulate the silencing of an X chromosome in females of placental and marsupial mammals, respectively. In addition, *roX2* triggers the overexpression of the X chromosome in male fruit flies. Recently, it was found that the green anole (*Anolis carolinensis*) contains a perfect X chromosome dosage compensation system. In a remarkable case of evolutionary convergence with the fruit fly, the green anole increases the levels of H4K16 acetylation to up-regulate the expression of the male X. In this study, we continued exploring the attributes of the only known dosage compensation mechanism that regulates full X chromosomes in reptiles. We found that an ancient lncRNA, *MAYEX*, gained male-specific expression more than 89-million-year-old. The acetylation machinery is retained at the *MAYEX* locus and *MAYEX* transcription is co-regulated

with an upstream lncRNA specific to females. *MAYEX* nucleotide sequence, together with a neighboring cluster of repeats and an X-enriched DNA motif evolved to be recognized by a still unknown protein complex capable of connecting multiple regions on the X chromosome with the *MAYEX* locus to increase the acetylation levels.

Main text

In placental mammals, marsupials, and the fruit fly (*Drosophila melanogaster*), females carry two copies of the X chromosome whereas males carry a single copy of the X chromosome and a degenerated copy of the Y chromosome. In 1967, Susumu Ohno predicted that females with two X chromosomes should not have more transcriptionally active genes than males with one X chromosome (1,2). Ohno proposed that one of the X chromosomes in females could be silenced to achieve dosage compensation between the two sexes (1,2). Later, it was confirmed that placental females follow a random inactivation of one X chromosome (1,3). An analogous X-inactivation system was found in marsupials, a group of mammals that share the same XY chromosomes with placental mammals (4). In contrast, an alternative strategy evolved in the fruit fly, where the single X in males is overexpressed to match the expression output of the two X chromosomes in females (5). Interestingly, in these three lineages, long non-coding RNA (lncRNA) are central players in the mechanisms that balance gene expression between males and females. In placental mammals, X chromosome inactivation happens early during female development by activity of the lncRNA *XIST* (X inactive specific transcript) (6,8). The *XIST* locus is regulated by the proteins YY1 (9,10) and KDM5C (11). Once *XIST* is transcribed, it is upregulated by SPEN (12), and then acts in *cis* to recruit the SF2 protein, Polycomb complexes, and other chromatin-modifying proteins like RBM15, RBM15B, and METTL3 that will progressively induce chromatin silencing by histone H3 lysine 27 trimethylation (H3K27me3) and subsequently chromatin compaction (7). Analogously, in marsupials, the lncRNA *RSX* (RNA-on-the-silent X) acts in *cis* to trigger the X-inactivation signaling pathway (4). This lncRNA is thought to interact with Polycomb complexes and other potential chromatin-modifying proteins to silence transcription (13). Finally, in the fruit fly, the lncRNA *ROX2* (RNA on the X) is bound by the male-specific lethal (MSL) complex that hyper-acetylates histone 4 at the lysine 16 (H4K16ac) to increase the transcription output of the X chromosome in males

(5,14). As in mammals, the regulation of the X chromosome in *Drosophila* occurs early during embryonic development (15).

The green anole, *Anolis carolinensis*, has 160-million-year-old sex chromosome (16,17) with a highly degenerated Y chromosome and an X chromosome which shows balanced expression levels between males and females (18). In this lizard, the dosage compensation mechanism up-regulates the expression of genes on the male X by increasing the levels of the H4K16 acetylation mark (18). In this study, we further explored the specificities of this dosage compensation system.

Results

Two neighboring sex-specific lncRNAs on the X chromosome of *A. carolinensis*

To identify potential candidates regulating the expression levels of the X chromosome in *A. carolinensis*, we performed differential expression analyzes of RNA-seq data using annotated genes. We did not detect sex-specific genes on the X chromosome (Supplemental Table 1), however, we speculated that some elements, such as lncRNA genes, might be poorly annotated in the reference genome. Therefore, we divided the X chromosome into thousands of small windows (50 base pairs -bp- long) and performed differential expression analyses between male and female samples. We verified that this approach retrieved the *XIST* locus in human and pig X chromosomes, and the *RSX* locus in the opossum X chromosome (Fig. 1, A-F).

We then applied the method to the X chromosome of *A. carolinensis*. We found two neighboring loci on this chromosome, one with strong male expression bias and one with strong female expression bias (Fig. 1 G and H and fig. S1). Sequences searches in the GenBank database indicated that the male-specific locus was a lncRNA (XR_001730858.1) whereas the female-specific locus did not report significant matches.

Next, we performed a detailed mapping of strand-specific RNA-seq data from male and female adult and embryonic samples. We found that the male-specific lncRNA is 3,327 bp long and is located on the forward strand (GL343417:269276-272592; the scaffold GL343417 is part of the X chromosome) (17) in a large intergenic region, opposite to the *MORC2* gene (ENSACAG00000009903) that lays on the reverse

strand (Fig. 2 B, C and E); both genes share the same CpG promoter (Fig. 2 F). We named the male-specific lncRNA *MAYEX* for, “Male-specific long non-coding RNA AmplifYing the Expression of the X”. *MAYEX* has a ubiquitous expression: detected in whole male embryos, from as early 2-3 days of development, and in all embryonic and adult tissues (brain, heart, kidney, liver, and testis; Fig 2 B and C, E and F). *MAYEX* has two exons and has an average expression level of ~30 RPKM. In a few female samples, *MAYEX* showed basal expression levels (Fig. 2 D –heart tissues-). The female-specific locus (named *FERX* for “Female Expressed Region on the X”) is located upstream *MAYEX*, also on the forward strand and in the same intergenic region (GL343417:348846-359846). *FERX* is only expressed in female embryonic tissues (mainly during the early stages) and adult brain (Fig. 2, A y D). Oddly, *FERX* varies in length across samples (3,000-50,000 bp; Fig. 2 A and D). *FERX* expression profile is also unusual, instead of the typical peak of RNA-seq reads, characteristic of most genes, exhibiting lower read coverage at its 5’ end and higher read coverage at its 3’ end; a distinctive pattern of rapidly-degrading transcripts (Fig. 2 A and D).

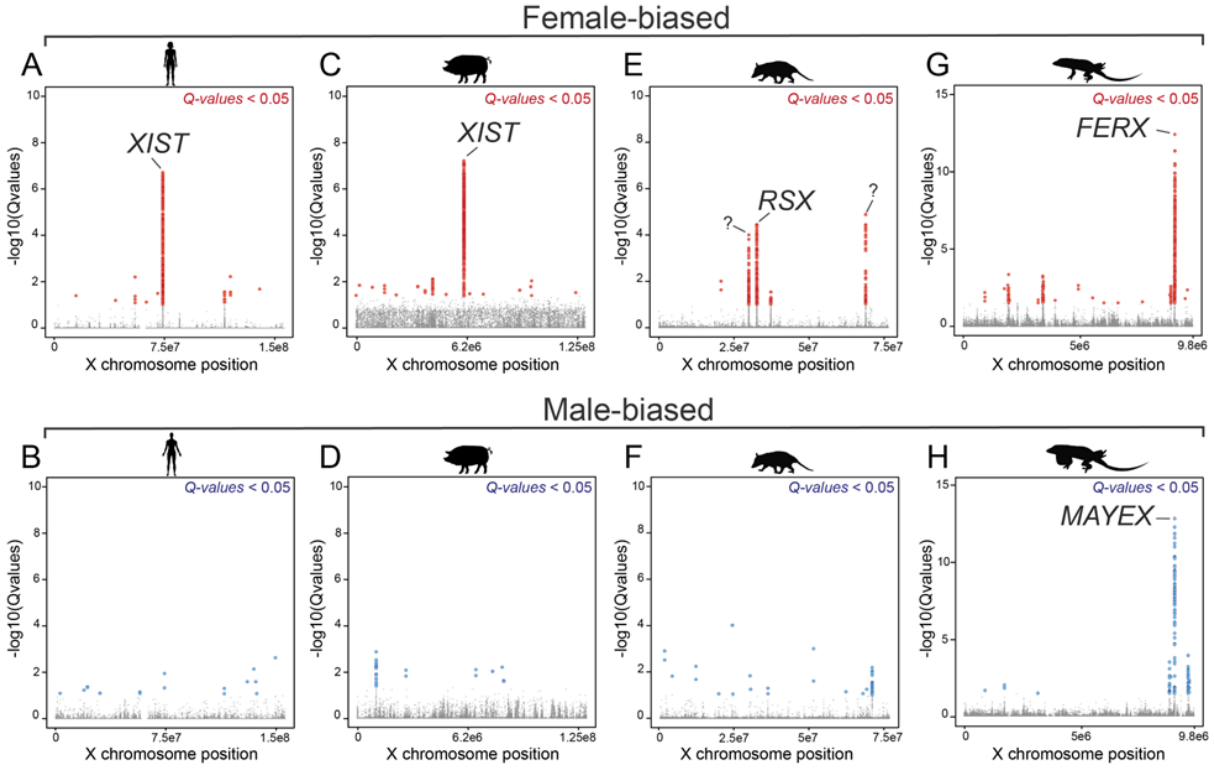


Fig. 1. Genomic loci with male or female expression bias. (**A**) In human and (**C**) pig, the *XIST* locus concentrates multiple 50 bp windows with significant female expression bias. (**B**) The X chromosome of human and (**D**) pig does not have loci with male expression bias. (**E**) In the opossum, the *RSX* locus and two unknown loci concentrate on multiple 50 bp windows with significant female expression bias. (**F**) The X chromosome of the opossum does not have

loci with male expression bias. (**G**) In *A. carolinensis*, the *FERX* locus concentrates multiple 50 bp windows with significant female expression bias. (**I**) In *A. carolinensis*, the *MAYEX* locus concentrates multiple 50 bp windows with significant male expression bias. Red dots in females and blue dots in males represent 50 bp windows with *Q*-values < 0.05.

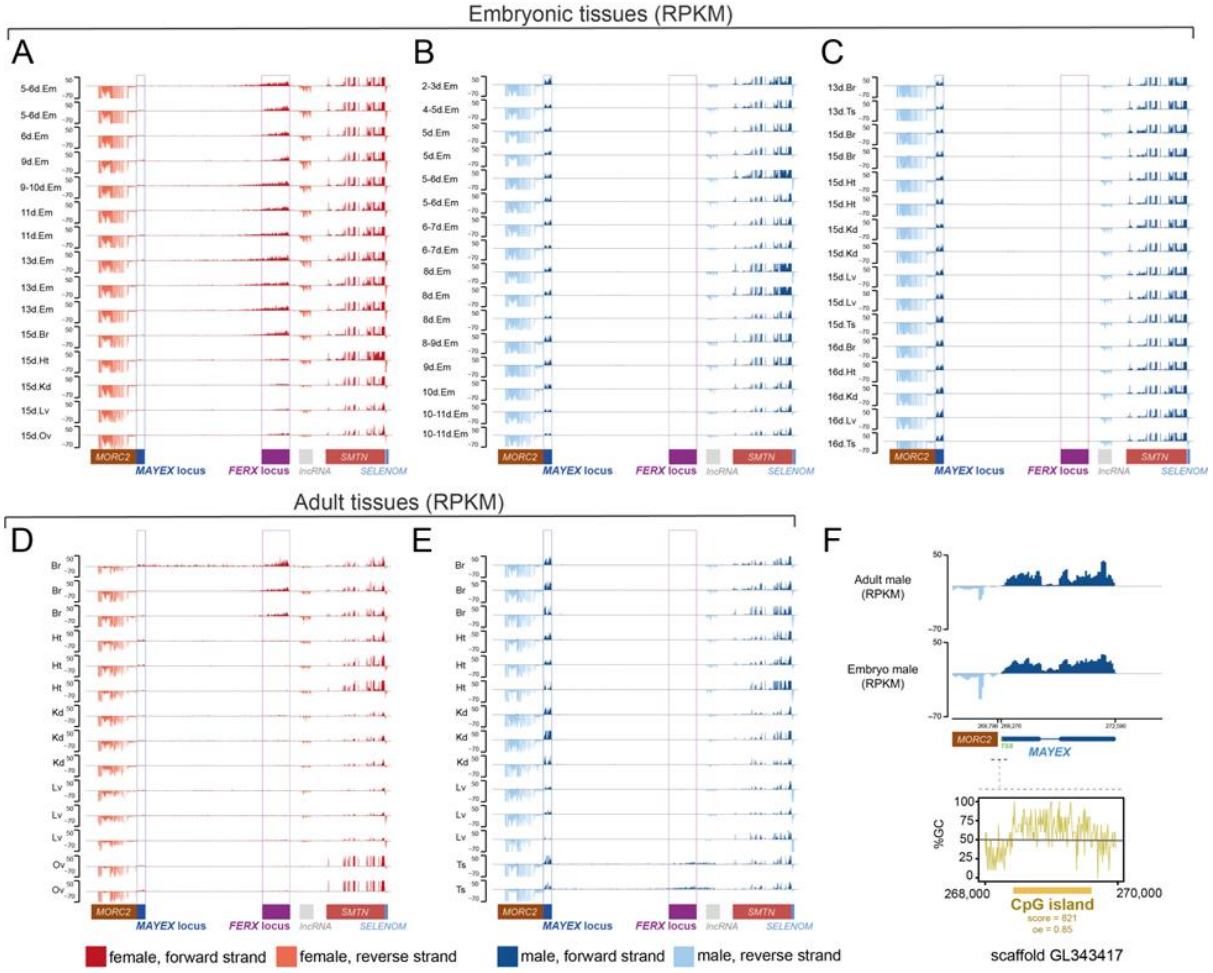


Fig. 2. Expression profile of the *MAYEX* and *FERX* region. Track panels showing the expression levels (RPKMs) of the *MAYEX* and *FERX* region across (**A**) female embryonic samples, (**B** to **C**) male embryonic samples, (**D**) female adult samples and (**E**) male adult samples. (**F**) Average expression levels (RPKMs) of *MAYEX* in male embryos and adults. *MAYEX* has two exons and its CpG promoter is shared with *MORC2*. (**A** to **E**) Dark and light orange represent expression levels in the forward and reverse strand of female samples, respectively. Dark and light blue represent expression levels in the forward and reverse strand of male samples, respectively. Genes located near the *MAYEX* locus are indicated at the bottom of each panel.

High levels of histone acetylation at the *MAYEX* locus

A mechanism of dosage compensation is active on the entire X chromosome since genes, regardless of location, showed male-to-female expression ratios close to zero (Fig. 3 A and D). Moreover, using differential coverage analysis of ChIP-seq data for the H4K16ac epigenetic mark, we detected male-specific hyper-acetylation along the entire X chromosome (Fig. 3 B and D). The *MAYEX* locus brain and liver of adult males

consistently shows the highest enrichment of H4K16ac signal across the X chromosome (Fig 3 B and C) and the rest of the genome (Fig S2).H4K16ac at the *MAYEX* locus in males is significantly higher than at the hyper-acetylated X chromosome or the autosomes (Mann Whitney U test, $P < 0.00001$; Fig. 3 E and F). The expression level of *MAYEX* is 30 times higher in males compared to females (male/female ratio = 5), whereas its acetylation levels are 60 times greater in males compared to females (male/female ratio = 6; Fig. 3 B and C).

MAYEX **and neighboring repeats organize an intra-chromosomal regulatory domain**

To get insight into the chromatin structure of the X chromosome and the potential role of *MAYEX*, we performed Hi-C experiments in male and female lizards. The Hi-C data analysis showed that there are not many differences in chromatin topology between males and females (Fig. S3). On the X chromosome, we detected a higher frequency of contacts and more topologically associating domains (TADs) in males compared to females (Fig. 4 A -TADs1-2-; Fig. S4). These results are consistent with a more open chromatin state of the hyper-acetylated male X. Importantly, we noted that a particular region close to the end of the X chromosome shows multiple long-range interactions with the rest of the chromosome (Fig. 4 A -red arrows, histograms, and interaction plots of Hi-C contacts-). The frequency of long-range contacts in this region is higher in males compared to females, and also greater than that observed for other loci on the X (Fig. 4 B). The long-range Hi-C contacts are enriched in a ~30kb region that begins at the *MAYEX* locus and stretches upstream into the intergenic region (Fig. 4 C). *MAYEX* locus is at the edge of a TAD in males but lies in the middle of a TAD in females (Fig. 4 A -TADs1-2-), thus, indicating that this region is open and transcriptionally active in males buy heterochromatic in females. Data obtained from a ChAR-seq experiment, designed to capture RNA molecules associated with chromatin, shows that the transcribed form of *MAYEX* is in direct contact with multiple distant DNA regions on the X chromosome in males (Fig. 4 A -interaction plots of ChAR-seq contacts-). This pattern is not observed for other lncRNA on the X (Fig. S5). We also found that the promoter region of *MAYEX* has differential contacts in males and females (Fig. S6), which may affect the regulation of *MAYEX* differential expression.

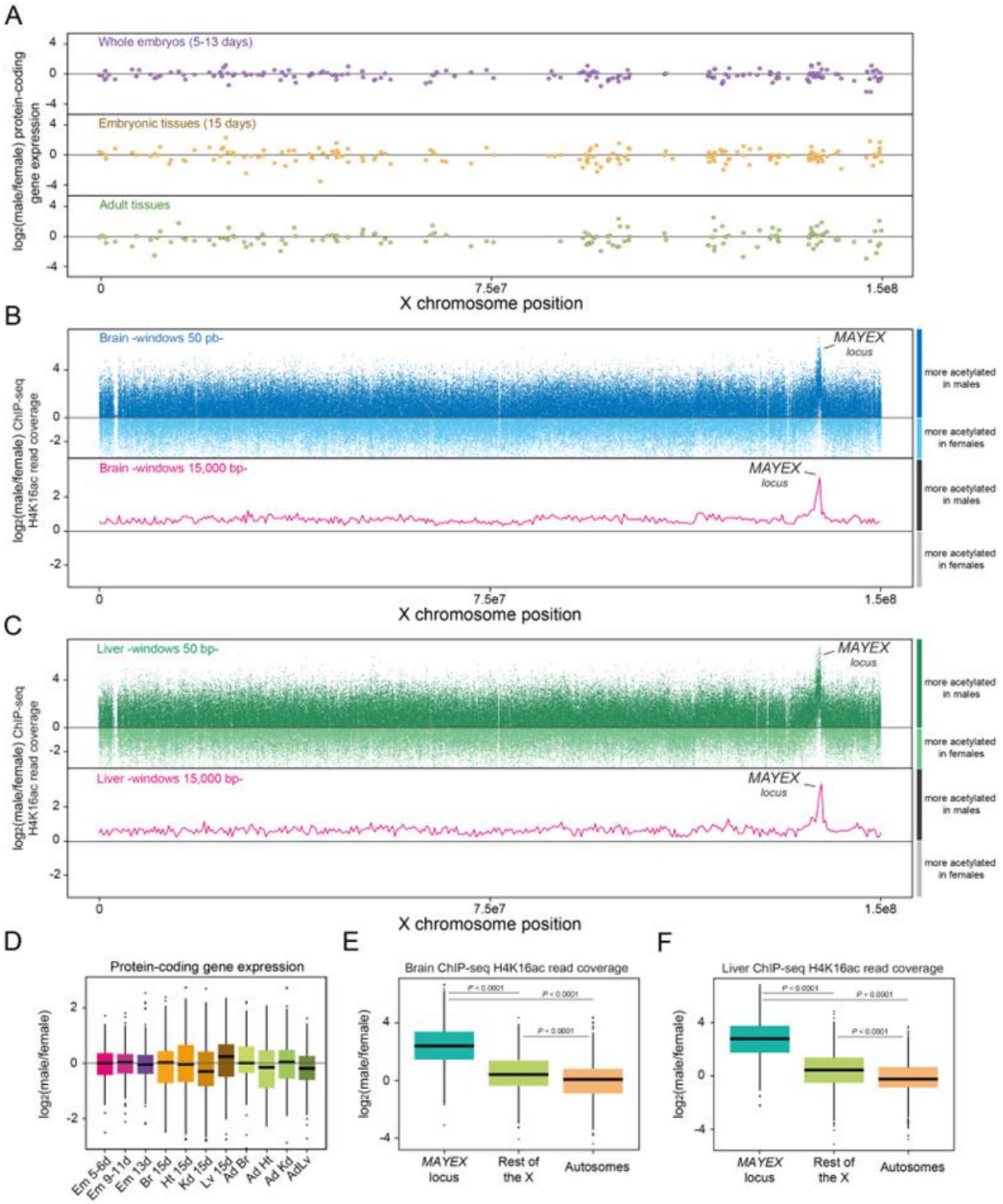


Fig. 3. Gene expression and acetylation levels of the X chromosome. (**A**) Log_2 ratio of the male to female expression levels of protein-coding genes in whole embryos (purple dots, top), embryonic tissues (yellow dots, middle), and adult tissues (green dots, bottom). Values are median expression levels across tissues. N = 315. (**B**) Log_2 ratio of the male to female H4K16ac read coverage in brain samples; blue dots (top) indicate median values every 50 bp and the pink line (bottom) indicates median values every 15,000 bp. (**C**) Same as (B) for the liver. (**D**) Boxplots of the log_2 ratio of the male to female expression levels (RNA-seq data) of protein-coding genes in embryonic and adult somatic tissues; Em, whole embryos at a given day of development; Br, brain, Ht, heart; Kd, kidney; Lv, liver; 15d, day fifteen of development; Ad, adults. (**E**) Boxplots of the log_2 ratio of the male to female H4K16ac read coverage at the

MAYEX locus compared to the rest of the X chromosome and autosomes 1 to 6 for the brain. (**F**) same as (E) for the liver. (**E** to **F**) N = 1300 for the *MAYEX* locus, and N = 1300 of randomly subsampled values for the X chromosome and autosomes. Significant differences (Mann-Whitney *U* test): Benjamin Hochberg-corrected $P < 0.05$. Error bars, maximum and minimum values, excluding outliers.

Next, we examined the intergenic region upstream of *MAYEX* for genomics signatures that could explain its higher frequency of long-range contacts. We found that this region is enriched with a 72 bp-long repeat (~42 copies, 85-98% identity between copies: Fig 4 D), which is not present anywhere else in the genome. At position 2900, *MAYEX* exhibits the first half of the repeat (Fig. 4 D). We postulate that this cluster of unique repeats and *MAYEX* might be recognized by a hypothetical complex of proteins that establishes and stabilizes long-range contacts in sequence of the long-range interacting loci on the X detected a $(TTA)_5$ repeated motif (Fig. 4 E). This motif is present in the vicinity of ~40% of the long-range Hi-C contacts that interact with *MAYEX* locus. We observed a large cluster of $(TTA)_5$ in the intergenic region upstream of *MAYEX* (Fig. 4 E) that could also be important for establishing these long-range contacts. We found that the X chromosome has the highest frequency of $(TTA)_5$ genome-wide, with 3.5-4 times more $(TTA)_5$ repeats than the autosomes. Finally, RNA-fluorescence in situ hybridization (FISH) experiments using blood cells from males showed that *MAYEX* forms distinctive aggregates (Fig. 4 F; Fig. S7).

Origin and evolution of *MAYEX*

We found the orthologous sequence of *MAYEX* in the male transcriptomes of the anoles *A. porcatus* (99% of sequence identity), *A. allisoni* (99% of sequence identity), and *A. distichus favillarum* (83% of sequence identity; Fig. 5; Fig. S8). We hypothesize that the sequence and male-specific expression of *MAYEX* were likely following the origin of the XY chromosome. Given this premise, we analyzed the genome of the spiny lizard *Sceloporus undulatus* (the eastern fence lizard; Phrynosomatidae family) (19). We found *MAYEX* sequence in this species (66% of sequence identity; Fig. S9), despite the 89 million years of divergence of anoles and spiny lizard. *MAYEX* is located opposite the *MORC* gene on the X chromosome.

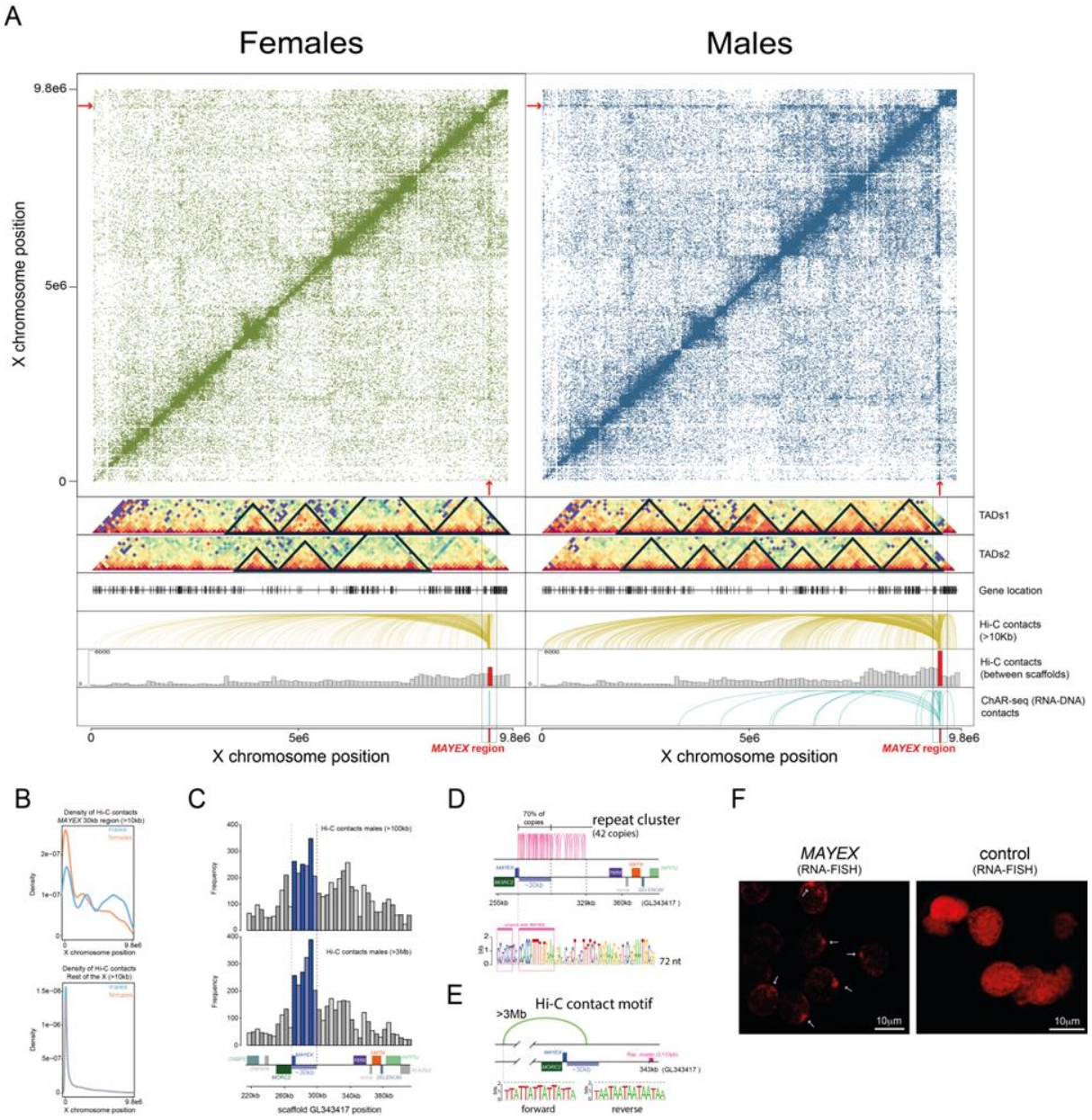


Fig. 4. Chromatin interactions on the X chromosome. (**A**) Hi-C contact plots for females (dark green) and males (dark blue) for the X chromosome; contacts >10kb were plotted; red arrows point to the region enriched in long-range contacts. Lower panels, from top to bottom: *TADs1* and *TADs2* show heatmaps of contacts and projected TADs for two male and two female replicates; colors indicate the frequency of contacts: from 1 (blue squares) to 200 (red square); bin size = 2e+05. *Gene locations* show the TSS positions for protein-coding genes. The *>10kb Hi-C contacts* plot the contacts (>10kb) between the locus of *MAYEX* and the rest of the X chromosome (N = 1074 contacts in males and N = 488 contacts in females). *Histograms of Hi-C contacts* (bin size = 8e+04 bp) show the frequency of long-range contacts between two different X-linked scaffolds. *Char-seq contacts* show the contacts between the RNAm of *MAYEX* and the chromatin. (**B**) Density plots of the frequency of contacts versus the distance of contacts for the 30kb locus upstream of *MAYEX* and the rest of the X chromosome. (**C**) Histograms representing the frequency of long-range contacts (>100kb and >3Mb) between the 30kb locus upstream of *MAYEX* and other regions of the X chromosome. (**D**) Top, location of the 42 copies in the cluster of repeated sequences. 70% of copies are located in the 30kb region next to *MAYEX*. Bottom, consensus sequence of the 72 bp repeat; the pattern present in *MAYEX* is highlighted by pink squares (fig. S11). (**E**) The $(TTA)_5$ motif

enriched in the long-range Hi-C contacts. Pink square, position of the $(TTA)_5$ cluster relative to the *MAYEX* locus. (**C** to **E**) The 30kb region next to *MAYEX* is highlighted in purple. The neighboring genes are indicated with squares of different colors. Scaffold GL343417 is part of the X chromosome. Positions at GL343417 instead of the full X chromosome are shown to facilitate browsing in the current genome assembly. (**F**) RNA-FISH using a *MAYEX*-specific probe (left) and a control gene (right), *COL1A1*, that has ubiquitous expression; white arrows signal the aggregates that *MAYEX* forms inside the nucleus. See also fig. S7.

Gene synteny around the neighboring region of *MORC2/MAYEX* has been broken in *Sceloporus* (Fig. 5), however, we found that *MAYEX* fulfills the expected pattern and is only expressed in male tissues. (Fig. 5). Gene synteny of the neighboring region to *MORC2* is well conserved in five outgroups species (the bearded dragon, Indian cobra, tiger snake, common wall lizard, and the chicken). These outgroups species harbor four independent ZW chromosomes (Fig. 5) and, in agreement without hypothesis, the nucleotide sequences of the orthologous loci to *MAYEX* are not conserved (Fig. S10) and these loci are transcribed in tissues from both sexes (Fig. 5). These results indicate that an old lncRNA gained male-specific expression following the origin of the Pleurodont XY chromosomes. We also found that the cluster of unique repeats located upstream of *MAYEX* (Fig. 4 D) is present in *S. undulatus* (Fig. S11), but not in the other reptiles. Finally, we used the orthologous sequences of *MAYEX* in the anoles to calculate its most likely secondary structure. We found a complex structure that can be divided into six contiguous domains (Fig. S12). Four of these domains have conserved loops, with positions not changing or co-evolving to maintain the secondary structure. The two domains with low sequence conservation are unstructured.

Discussion

In this study, we explored the dosage compensation system of the X chromosome in the green anole. We described the lncRNA *MAYEX* which is located on the X chromosome in *A. carolinensis* and *S. undulatus*. This is the first lncRNA that regulates full X chromosomes described in reptiles. The similarities between mammals, the fruit fly, and now the green anole, indicate that lncRNAs have been concurrently recruited in distant species during evolution to establish chromosome-wide *cis*-regulatory mechanisms that control the expression levels of full sex chromosomes and restore balance expression ratios between males and females.

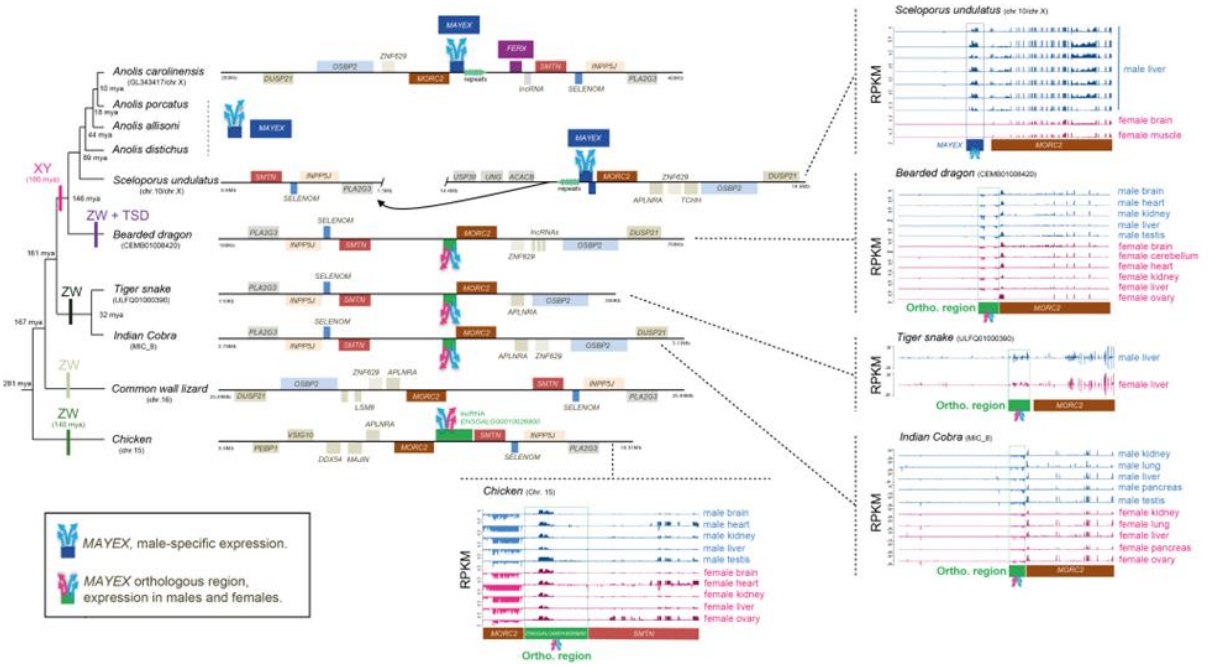


Fig. 5. Evolution of *MAYEX*. Gene synteny around the neighboring region of *MAYEX* in *A. carolinensis* and other reptile species. The position of *MAYEX* is indicated by blue rectangles in *A. carolinensis* and *S. undulatus*. Blue vertical arrows indicate expression in male tissues. Green horizontal arrows indicate the position of the cluster of repeated sequences. For *A. carolinensis*, positions at GL343417 scaffold instead of the full X chromosome are shown to facilitate browsing in the current genome assembly. Positions of orthologous lncRNAs to *MAYEX* are indicated by green rectangles. Pink/blue vertical arrows indicate expression in male and female tissues. The phylogenetic tree is scaled based on divergence estimates and sex chromosomes are displayed at their approximate origin; ZW systems have independent origins. Track panels indicate expression levels (RPKMs) in male (blue tracks) and female (pink tracks) tissues for *MAYEX* in *S. undulatus* and its orthologous loci in the bearded dragon, the tiger snake, the Indian cobra, and chicken. Expression profile of *MORC2* gene (brown square) and *SMTN* (in chicken, light brown square) are also shown. Type and sex of tissues are indicated at the end of each track.

Our current model (Fig. 6) suggests that *MAYEX* and *FERX* are mutually exclusive lncRNA; when the expression of *FERX* is dominant (in females) the *MAYEX* locus is switched off, and viceversa. We think that *MAYEX* and *FERX* could be different isoforms of the same gene. Indeed, the chicken orthologous intergenic region to *MAYEX* and *FERX* contains a single lncRNA (ENSGALG00010026900; Fig. 5 A). The *MAYEX* system operates in all tissues during development and in adults, suggesting that hyper-acetylation of the X chromosomes needs to be actively maintained. However, it is unclear why *FERX* expression is no longer required in adult tissues. *MAYEX* is an old lncRNA transcribed in multiple tissues of both males and females that gained male-specific expression more than 89 million years ago, likely after the origin of the XY chromosomes in pleurodonts. Although *MAYEX* shares the same promoter region with *MORC2*, this gene does not have male-specific expression but

rather shows ubiquitous expression in males and females, which indicates *MAYEX* expression may have a complex regulation. We found that the acetylation mark is remarkably high at the *MAYEX* locus, suggesting a close interaction between *MAYEX* and a potential acetylation complex. We hypothesize that a protein complex loops the entire X chromosome to the *MAYEX* locus and given the close proximity of the acetylation machinery to this locus, it could result in significant increases in H4K16ac levels in the X chromosome. The X chromosome in *Anolis* is 16 times smaller than the X chromosome in placental mammals and 2.4 times smaller than the one from *Drosophila*. This could explain the emergence of the looping mechanism instead of the scaffolding of lncRNAs around the chromosome carried out by *XIST*, *RSX*, and *ROX2*. The X chromosome is enriched in the $(TTA)_5$ repeat, which could explain why the system is active only on the X chromosome. This repeat evokes the $(GA)_4$ motif associated with the DNA binding sites of *ROX2*/MSL3-TAP in *Drosophila* (20).

This work presents a novel lncRNA that is involved in dosage compensation of the X chromosomes in lizards. Our model, however, has raised several questions. Future studies should focus on identifying the hypothetical proteins associated with *MAYEX* and its neighboring repeat-rich region, which could reveal the molecular mechanisms that loops together regions of the X chromosome with the *MAYEX* locus, and the dynamics of this process during development.

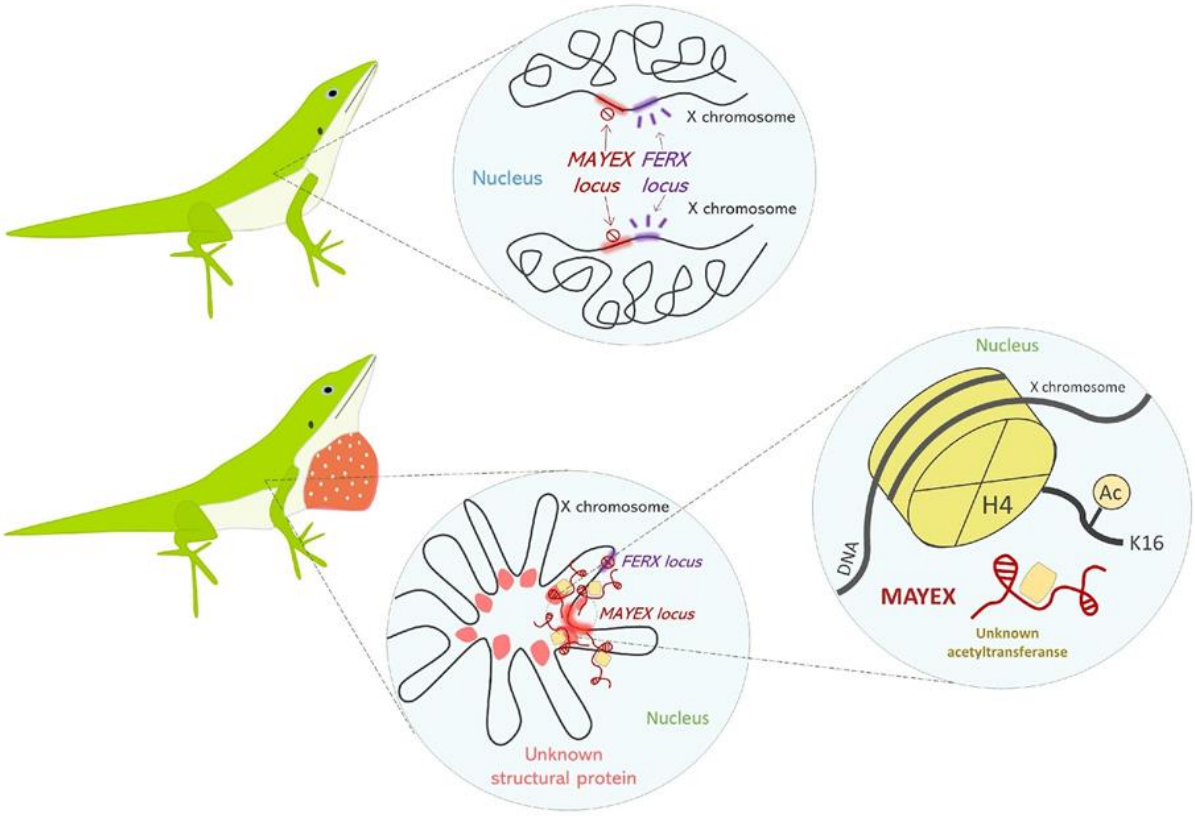


Fig. 6. The current model of dosage compensation of the X chromosome in *A. carolinensis*. The upper panel shows the model in females and the lower panel shows the model in males.

References

1. E. Beutler, Susumu Ohno: the father of X-inactivation. Cytogenet Cell Genet 80, 16-17 (1998).
2. S. Ohno, Sex chromosomes and sex linked genes. (Springer Berlin, Heidelberg, ed. 1, 1967).
3. B. P. Balaton, T. Dixon-McDougall, S. B. Peeters, C. J. Brown, The eXceptional nature of the X chromosome. Hum Mol Genet 27, R242-R249 (2018).
4. J. Grant et al., Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. Nature 487, 254-258 (2012).
5. M. E. Gelbart, E. Larschan, S. Peng, P. J. Park, M. I. Kuroda, Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. Nat Struct Mol Biol 16, 825-832 (2009).
6. C. Patrat, J. F. Ouimette, C. Rougeulle, X chromosome inactivation in human development. Development 147, (2020).

7. B. Payer, J. T. Lee, X chromosome dosage compensation: how mammals keep the balance. Annu Rev Genet 42, 733-772 (2008).
8. K. Plath, S. Mlynarczyk-Evans, D. A. Nusinow, B. Panning, Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet 36, 233-278 (2002).
9. J. D. Kim et al., Identification of clustered YY1 binding sites in imprinting control regions. Genome Res 16, 901-911 (2006).
10. M. Makhlouf et al., A prominent and conserved role for YY1 in Xist transcriptional activation. Nat Commun 5, 4878 (2014).
11. M. K. Samanta et al., Activation of Xist by an evolutionarily conserved function of KDM5C demethylase. Nat Commun 13, 2602 (2022).
12. T. Robert-Finestra et al., SPEN is required for Xist upregulation during initiation of X chromosome inactivation. Nat Commun 12, 7000 (2021).
13. D. Sprague et al., Nonlinear sequence similarity between the Xist and Rsx long noncoding RNAs suggests shared functions of tandem repeat domains. RNA 25, 1004-1019 (2019).
14. J. J. Quinn, H. Y. Chang, In situ dissection of RNA functional subunits by domain-specific chromatin isolation by RNA purification (dChIRP). Methods Mol Biol 1262, 199-213 (2015).
15. T. Conrad, A. Akhtar, Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. Nat Rev Genet 13, 123-134 (2012).
16. J. Alfoldi et al., The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477, 587-591 (2011).
17. M. Rovatsos, M. Altmanova, M. J. Pokorna, L. Kratochvil, Novel X-linked genes revealed by quantitative polymerase chain reaction in the green anole, Anolis carolinensis. G3 (Bethesda) 4, 2107-2113 (2014).
18. R. Marin et al., Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage. Genome Res 27, 1974-1987 (2017).
19. A. K. Westfall et al., A chromosome-level genome assembly for the eastern fence lizard (Sceloporus undulatus), a reptile model for physiological and evolutionary ecology. Gigascience 10, (2021).

20. M. D. Simon et al., The genomic binding sites of a noncoding RNA. Proc Natl Acad Sci U S A 108, 20497-20502 (2011).

Supplementary Materials

Materials and Methods

Samples

Two males and two females of *A. carolinensis* were captured in Tampico, Tamaulipas, Mexico (170 m.a.s.l.; SEMARNAT Scientific Collector Permit 08-043). Animals were sacrificed and we flash-frozen in liquid nitrogen the brains, livers, hearts, lungs, kidneys, muscles, and gonads. Organs were stored in 1.5 ml tubes at -80°C until use.

Analysis of RNA-seq and ChIP-seq data

RNA-seq data for 15 embryonic and 14 adult tissues from females and 32 embryonic and 14 adult tissues from males were downloaded from the NCBI-SRA database (https://www.ncbi.nlm.nih.gov/sra; PRJNA381064). We also downloaded ChIP data for H4K16ac and the Input data for the brain and liver of two female replicates and two male replicates from the NCBI-SRA database (PRJNA381064). We downloaded RNA-seq data from the NCBI-SRA database for adult male and female tissues for humans and the opossum (PRJNA381064), and pig (PRJNA580502). We downloaded the reference genomes of *A. carolinensis*, human, pig, and opossum from the Ensembl database (release 104; https://www.ensembl.org). RNA-seq and ChIP-seq data were trimmed for adaptors and low-quality positions using trim_galore (v0.6.2) (https://github.com/FelixKrueger/TrimGalore). RNA-seq data was aligned to the reference genomes with HISAT2 (v2.1.0; parameters: -q --threads 16 -N 1 -L 18 -i S,1,0.50 -D 20 -R 3 --pen-noncansplice 15 --mp 1,0) (*21*). Specifically for *A. carolinensis,* we ordered the 13 X-linked scaffolds (*22, 23*) based on the Hi-C density contacts across scaffolds (see Hi-C analysis below). We divided the nucleotide sequences of the X chromosomes in human, pig, opossum and *A. carolinensis* into windows of 50 bp. We calculated the coverage of these windows using BEDtools (v2.27.1) (*24*). Read counts for all windows and every tissue (excluding testis that has a great number of tissue-specific genes (*25*)) were added together into a data frame using standard R libraries (v4.1.0) (https://www.R-project.org/). We ran differential

expression analyses for all windows in males compared to females using *edgeR* R library (normalization using TMM -Trimmed Mean of M-values-, FDR -False Discovery Rate- set at 0.05) (*26*). The resulting FDR values (i.e., *Q*-values, which are the *P*-values that have been corrected for multiple tests) were plotted for every window against the position of the windows using the *ggplots2* R library (*27*). Mapped RNA-seq reads from *A. carolinensis* were sorted in a BAM file for each tissue using SAMtools (v1.9) (*28*). BAM files were indexed and strand-specific RPKMs for windows of 50 bp were calculated using deepTools2 (v3.3.1) (bamCoverage tool) (*29*). Bedgraph files containing the coverage for each tissue were plotted using *trackViewer* R library (*30*). We downloaded for *A. carolinensis* the positions of the protein-coding and non-coding genes from the Ensemble database (release 104; https://www.ensembl.org). We calculated read counts per gene using htseq-count (v0.9.1) (*31*) and using the lengths of the genes and the total number of mapped read per sample, we estimated gene expression levels (TPMs). Brain and liver ChIP-seq and their corresponding Input data were mapped to the *A. carolinensis* reference genome using Bowtie2 (v2.3.4.1) (*32*). BAM files were indexed and we estimated the coverage for windows of 50 bp using deepTools2 (v3.3.1) (bamCoverage tool) (*29*). We added one to the coverage of each window to avoid having -infinite values during the $\log_2$ transformation step. We then divided the coverage per window into the ChIP and Input files by their corresponding total number of mapped reads in the library. We calculated the median coverage per window in the ChIP data using the two male or the two female replicates. Finally, we $\log_2$ transformed, per window, the ChIP coverage divided by their corresponding Input coverage. We plotted the $\log_2$ values for every window (and the median of 30 windows) against the position of the windows using the *ggplots2* R library (*27*). The positions of TSS and CpG islands were obtained from the Ensemble database (release 106; https://www.ensembl.org).

Generation and analysis of Hi-C data

The Hi-C method allows the examination of genome-wide interactions (*33*). Hi-C introduces biotinylated nucleotides at ligation junctions to specifically purify these junctions (*34*). The detailed experimental protocol is described in the Supplementary Note (see below). Sequencing of four libraries, two male replicates, and two female replicates, was performed in an Illumina NovaSeq 6000 machine in Novogene,

California. We sequenced 1,036,627,136 & 1,029,732,950 (male libraries) and 942,303,252 & 946,529,692 (female libraries) paired-end reads. Reads were trimmed using the list of adaptors used in the experimental protocol with trimmomatic (v0.36; parameters: ILLUMINACLIP:illuminaClipping_main.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15) (*35*). Valid Hi-C pairs were obtained using HiCUP (v0.7.3) (*36*); HiCUP finds Hi-C junctions (hicup_truncater), maps paired-end reads to the *A. carolinensis* reference genome (release 104; https://www.ensembl.org) using Bowtie2 (v2.3.4.1) (*32*) (hicup_mapper), removes experimental artifacts using an in silico digested genome based on the enzyme's binding site (DpnII that cuts at GATC; hicup_filter), and finally removes PCR duplicates (hicup_deduplicator). We performed post-pipeline analysis using HiCExplorer (v3.6) (*37*); matrix building (hicBuildMatrix), correction of matrices (hicCorrectMatrix), merging of data into larger bins (hicMergeMatrixBins), normalize matrices across samples (hicNormalize), plot of Hi-C contacts for specific chromosomes (hicPlotMatrix), detection and plotting of TADs for specific genomic regions (hicFindTADs and hicPlotTADs). We used *trackViewer* (*30*), and *InteractionSet* (*38*) R libraries to generate contact plots between different genomic regions. We used unique mapped contacts to avoid redundant or complex structures due to multimapers. We also divided the X chromosome of *A. carolinensis* in windows of 10kb. We measured the number of long-distant contacts for each window at 10kb, 100kb, 1Mb, and >3Mb. We used the HiCUP-validated read pairs between the 13 X-linked scaffolds to order the scaffolds (fig. S13). We calculated the density of contacts between pairs of scaffolds and ordered the scaffolds following the best reciprocal 5'-3' density of contacts. The order used for the 13 X-linked scaffolds was GL343913.1 (length = 147151 bp), GL345060.1 (length = 12621 bp), GL343550.1 (length = 526944 bp), AAWZ02039360 (length = 8757 bp), GL343423.1 (length = 834740 bp), GL343282.1 (length = 1779868 bp), AAWZ02041299 (length = 5850 bp), AAWZ02040114 (length = 7379 bp), GL343364.1 (length = 1083274 bp, reversed), b (length = 3271537 bp), GL343338.1 (length = 1258094 bp, reversed), GL343417.1 (length = 831895 bp, reversed), GL343947.1 (length = 117443 bp, reversed). More details about X-linked scaffolds can be found in *(18)*.

Generation and analysis of ChAR-seq data

Chromatin-associated RNA sequencing is a chromosome conformation capture method that traps RNA-to-chromatin associations (*39, 40*). This method fixes all chromatin interactions and subsequently ligates the interacting RNA and DNA ends. The ligation is done through a double-stranded sequence, called "bridge" (*40*). At its 5' end, the bridge contains an adenylated single-stranded sequence, where it binds to the 3' ends of an RNA, and at its 3' end, the bridge contains a recognition site for the DNA fragments that have been digested by the DpnII enzyme. This enzyme cuts at GATC, which fit perfectly into the "bridge". After the ligation of both fragments, the chimeric molecule (RNA-bridge-DNA) is converted to DNA and isolated by biotin pulldown. The detailed experimental protocol is described in the Supplementary Note (see below). One male and one female library from the liver were sequenced using an Illumina NovaSeq 6000 machine in Novogene, California. We sequenced 1,020,074,230 (male library) and 9,96,050,284 (female library) paired-end reads. Reads were trimmed using the list of adaptors used in the experimental protocol with trimmomatic (v0.36; parameters: ILLUMINACLIP:illuminaClipping_main.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15) (*35*). Valid reads were obtained by extracting sequences that contained the bridge sequence (ACCGGCGTCCAAG or CTTGGACGCCGGT). The orientation of the bridge sequence allows the identification of the RNA and DNA fragments; the 5’ end before the bridge corresponds to the RNA whereas the 3’ end after the bridge corresponds to the DNA. DNA fragments were mapped onto the *A. carolinensis* reference genome (release 104; https://www.ensembl.org) using Bowtie2 (v2.3.4.1; parameters: -p 6 -D 20 -R 3 -N 1 -L 18 -i S,1,0.50 --no-unal --no-head --no-sq) (*32*). RNA fragments were mapped to the *A. carolinensis* transcriptome (release 104; https://www.ensembl.org) using Bowtie2 (v2.3.4.1; parameters: -p 6 -D 20 -R 3 -N 1 -L 18 -i S,1,0.50 --no-unal --no-head --no-sq) (*32*), mapped to a *de novo* male/female transcriptome generated using Trinity (v2.8.5, parameters: --seqType fq –single –SS_lib_type F –CPU 15 –max_memory 150G) (*41*) also using Bowtie2, and mapped to the reference genome using HISAT2 (v2.1.0; parameters: -q --threads 16 -N 1 -L 18 -i S,1,0.50 -D 20 -R 3 --pen-noncansplice 15 --mp 1,0) (*21*). Identical genome mapping positions were merged. We used unique mapped contacts to avoid redundant or complex structures due to multimapers. IDs from the paired-end reads were used to match the RNA and DNA

mapping positions. The quality of ChAR-seq experiments was established by studying the top genes with the highest frequency of contacts with the DNA. These genes corresponded to some of the most common RNA molecules found in the nucleus, such as ribosomal RNAs, small nuclear RNAs of the spliceosome (*i.e*., U2), or the RNA metazoan signal recognition particle (fig. S14). We also analyzed the male/female contact patterns of other lncRNAs annotated on the *Anolis* X chromosome.

FISH-RNA

The FISH-RNA method allows the visual inspection of the cellular localization of an RNA of interest. We designed *MAYEX*-specific primers (3' end; forward: ACACTGGAAAGATAATGATGGC and reverse: CAAGGAAGAATGCCACACTTAC) and primers for a control gene *COL1A1* (exon four; forward: CACCTAGCGGTGGCTTTGACTT and reverse: AGTGCGGGCTGGGTTCTTAC), the collagen, that has average expression levels. Expression of these genes in male tissues was verified by PCR using cDNA: RNA was purified from tissues using Qiagen RNeasy Mini Kit (Cat. No. 74104); cDNA was obtained using the ThermoFisher RT (Cat. No. 28025013) kit. Standard 30-cycle PCRs were performed. A ~200 bp region of the *MAYEX* and *COL1A1* genes was amplified (fig. S15) and used to prepare the constructs for the hybridization protocol. FISH-RNA was performed on blood cells from males. The detailed experimental protocol is described in the Supplementary Note (see below).

Identification of *MAYEX* in other species

We downloaded RNA-seq data from three species of *Anolis*: *Anolis allisoni* (PRJDB9984), *A. porcatus* (PRJDB9984), and *A. distichus favillarum* (PRJEB41750). RNA-seq was trimmed for adaptors and low-quality positions using trim_galore (v0.6.2) (https://github.com/FelixKrueger/TrimGalore). Transcriptome assemblies were obtained using Trinity (v2.8.5; default parameters) (*41*). We searched for *MAYEX* using BLASTN (*42*). We downloaded the *S. undulatus* genome from the NCBI database (GCF_019175285.1_SceUnd_v1; PRJNA746303). Reference genomes from the central bearded dragon (*Pogona vitticeps*), the mainland tiger snake (*Notechis scutatus*), the Indian cobra (*Naja naja*), and chicken (*Gallus gallus*) were downloaded from the Ensembl database (release 106; https://www.ensembl.org).

RNA-seq data for these species were retrieved from the NCBI-SRA database (https://www.ncbi.nlm.nih.gov/sra): *S. undulatus* (PRJNA371829, PRJNA437943, PRJNA605699, and PRJNA437943), the tiger snake (PRJNA170152), the Indian cobra (PRJNA527614), and chicken (PRJNA381064). RNA-seq data for the central bearded dragon was downloaded from European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/; ERR753524-ERR753530 and ERR413064-ERR413076). RNA-seq was trimmed for adaptors and low-quality positions using trim_galore (v0.6.2) (https://github.com/FelixKrueger/TrimGalore). RNA-seq data was aligned to the reference genomes with HISAT2 (v2.1.0; parameters: -q --threads 16 -N 1 -L 18 -i S,1,0.50 -D 20 -R 3 --pen-noncansplice 15 --mp 1,0) (*21*). For each species, mapped RNA-seq reads were sorted in a BAM file (one per tissue) using SAMtools (v1.9) (*28*). BAM files were indexed and strand-specific RPKMs for windows of 50 bp were calculated using deepTools2 (v3.3.1) (bamCoverage tool) (*29*). Bedgraph files containing the coverage for each tissue were plotted using *trackViewer* R library (*30*). We examined gene synteny around the *MORC2* genomic region using the Ensembl browser (release 106; https://www.ensembl.org). We included the Common wall lizard (*Podarcis muralis*) in this analysis. The *MORC2* gene was not annotated in the Indian cobra genome. We located the gene using the BLASTN engine in the Ensembl browser (https://www.ensembl.org) and the nucleotide sequence from the tiger snake.

RNA structure prediction

MAYEX sequences from four *Anolis* species (*A. porcatus*, *A. allisoni*, and *A. distichus favillarum*) were aligned using muscle (v3.8.1551) (*43*). The alignment was divided into 500 pb windows every 100 bp, producing a total of 26 windows. For each window, two structural predictions were performed using CMfinder (v0.4.1; parameters: -c 1 -m1 25 -M1 200 -m2 25 -M2 200 -s1 5 -s2 5 -combine) (*44*) and locARNA (v1.9.2; parameters: --sparse -stockholm) (*45*). A covariance model for the two predicted structures was generated with Infernal (v1.1.3) (*46*) with the *cmbuild* command; the values of the model were extracted with the *cmstat* command (Supplementary Table 2). The structures of the best six contiguous windows that cover the entire sequence of *MAYEX* were plotted using the RNAplot function (parameters: -t 4 -a --covar) of the

Vienna package (v2.5.1) (*47*). The sequence alignment was plotted with Geneious (v2021.2.2) (*48*).

Identification of repeats and DNA motifs

To find repeated sequences, we performed BLASTN Campo (42) searches of the 100kb upstream of the MORC2 gene against itself. We also retrieved 500 bp around those regions having contacts with the 30 kb region upstream of *MAYEX*. We ran the DNA motif-enrichment program HOMER2 (http://homer.ucsd.edu/homer/motif/) on these sequences using background sets of 50,000 sequences of the same length from each of the autosomes. We selected only the motif that was significantly enriched in the X regions in all comparisons against autosomal backgrounds. We also used HOMER2 (http://homer.ucsd.edu/homer/motif/) to estimate the frequency of this motif in the X chromosome and the autosomes.

References

21. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12, 357-360 (2015).
22. J. Alfoldi et al., The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477, 587-591 (2011).
23. M. Rovatsos, M. Altmanova, M. J. Pokorna, L. Kratochvil, Novel X-linked genes revealed by quantitative polymerase chain reaction in the green anole, Anolis carolinensis. G3 (Bethesda) 4, 2107-2113 (2014).
24. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842 (2010).
25. M. Soumillon et al., Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep 3, 2179-2190 (2013).
26. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140 (2010).
27. H. Wickham, ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York, 2009).
28. P. Danecek et al., Twelve years of SAMtools and BCFtools. Gigascience 10, (2021).

29. F. Ramirez et al., deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic acids research 44, W160-165 (2016).
30. J. Ou, L. J. Zhu, trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. Nat Methods 16, 453-454 (2019).
31. G. H. Putri, S. Anders, P. T. Pyl, J. E. Pimanda, F. Zanini, Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics 38, 2943-2945 (2022).
32. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-359 (2012).
33. J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing chromosome conformation. Science 295, 1306-1311 (2002).
34. B. R. Lajoie, J. Dekker, N. Kaplan, The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods 72, 65-75 (2015).
35. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120 (2014).
36. S. Wingett et al., HiCUP: pipeline for mapping and processing Hi-C data. F1000Res 4, 1310 (2015).
37. F. Ramirez et al., High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun 9, 189 (2018).
38. A. T. Lun, M. Perry, E. Ing-Simmons, Infrastructure for genomic interactions: Bioconductor classes for Hi-C, ChIA-PET and related experiments. F1000Res 5, 950 (2016).
39. J. C. Bell et al., Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. Elife 7, (2018).
40. D. Jukam et al., Chromatin-Associated RNA Sequencing (ChAR-seq). Curr Protoc Mol Biol 126, e87 (2019).
41. M. G. Grabherr et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644-652 (2011).
42. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. J Mol Biol 215, 403-410 (1990).
43. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32, 1792-1797 (2004).
44. Z. Yao, Z. Weinberg, W. L. Ruzzo, CMfinder--a covariance model based RNA motif finding algorithm. Bioinformatics 22, 445-452 (2006).

45. S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, R. Backofen, LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. RNA 18, 900-914 (2012).
46. E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933-2935 (2013).
47. R. Lorenz et al., ViennaRNA Package 2.0. Algorithms Mol Biol 6, 26 (2011).
48. M. Kearse et al., Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647-1649 (2012).

Supplementary Figures

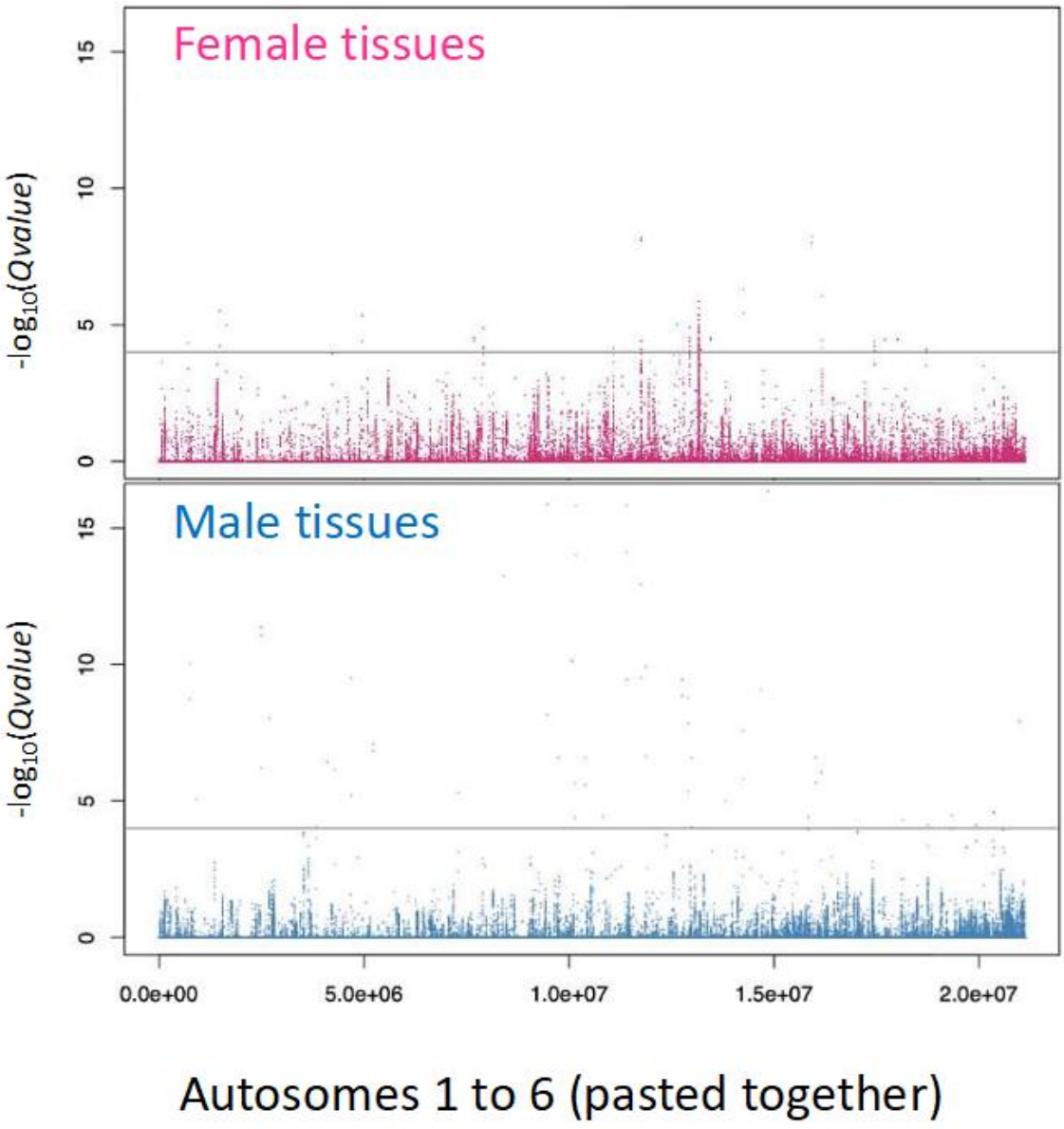


Fig. S1. Analysis of differential expression bias using windows of 50 base pairs for autosomes (1 to 6, pasted together) in *A. carolinenesis*. We did not detect other loci having similar differential expression biases as the ones shown by *MAYEX* in males or *FERX* in females.

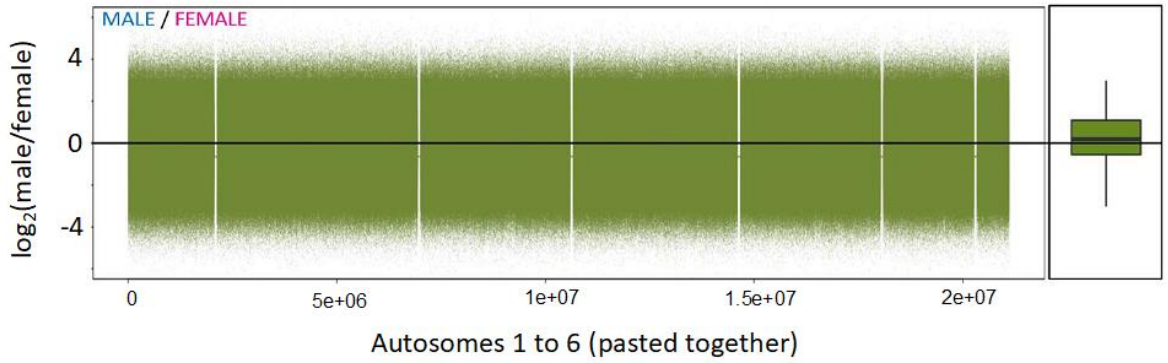


Fig. S2. Analysis of differential coverage for H4K16ac levels using windows of 50 base pairs for autosomes (1 to 6, pasted together) in *A. carolinenesis*. Autosomes show similar H6K16ac levels in males and females. We did not detect other loci having similar H4K16ac levels as the one shown by *MAYEX* in males.

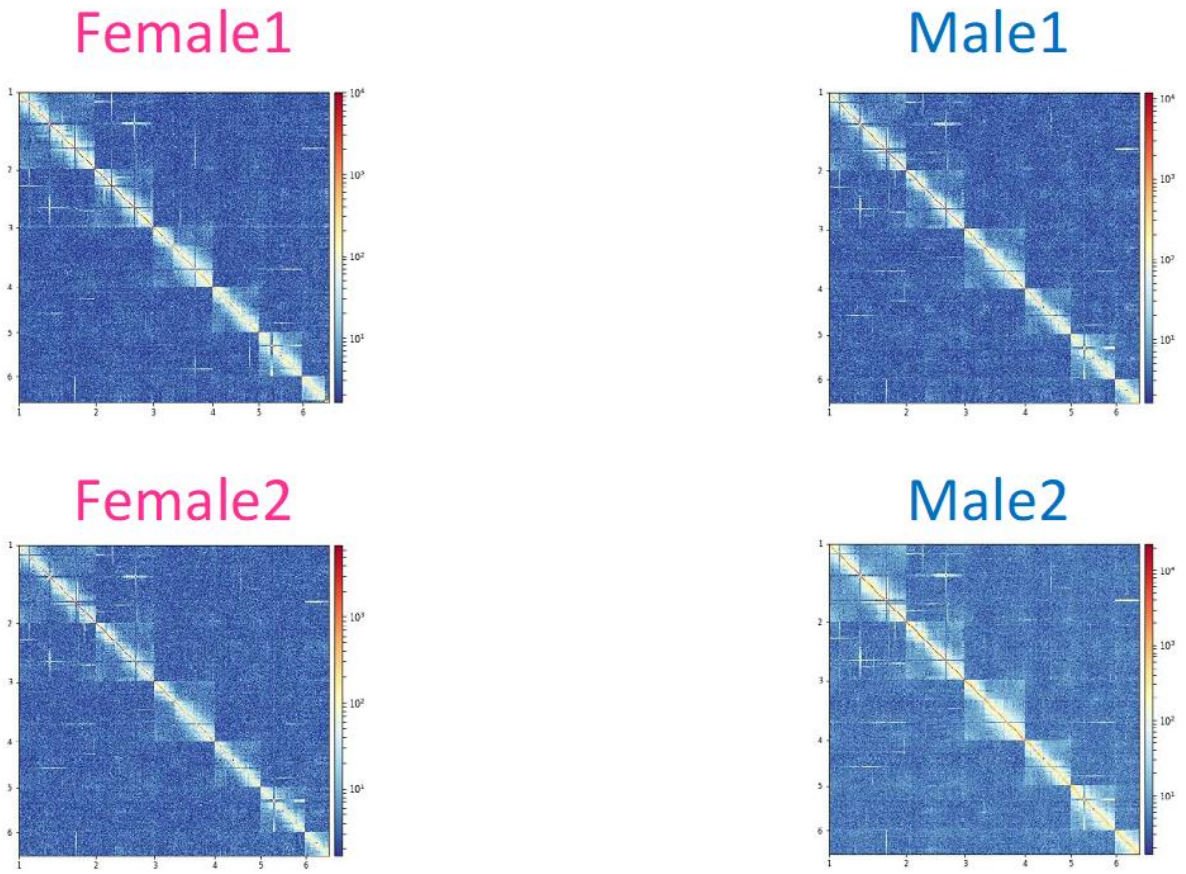


Fig. S3. Normalized and corrected Hi-C contact plots (bins = 500kb) for autosomes 1 to 6 in two male and two female replicates in *A. carolinenesis*. The Hi-C data showed little differences in chromatin topology between males and females.

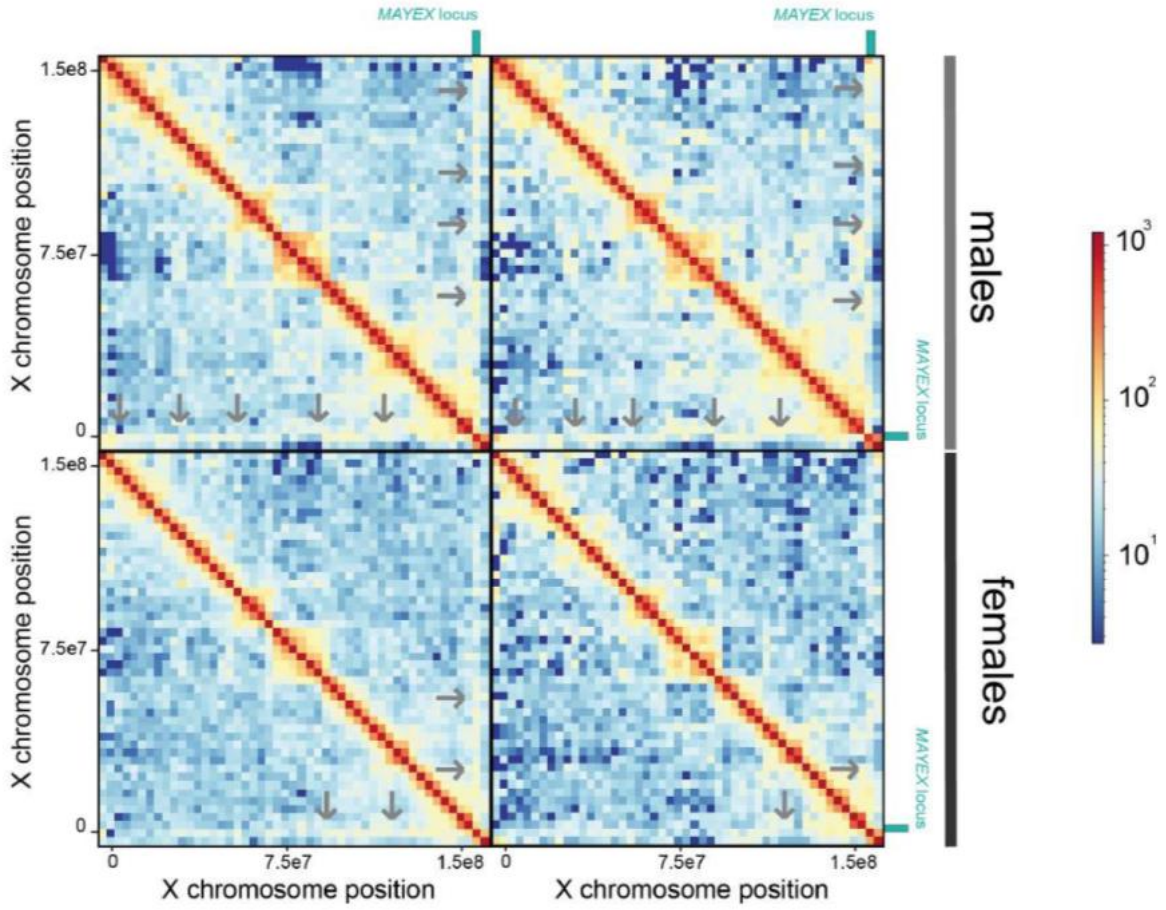


Fig. S4. Normalized and corrected Hi-C contact plots (bins = 500kb) for the X chromosome in two male and two female replicates in *A. carolinenesis*. Grey arrows point at the high density of long-range contacts between the *MAYEX* locus and the rest of the X chromosome.

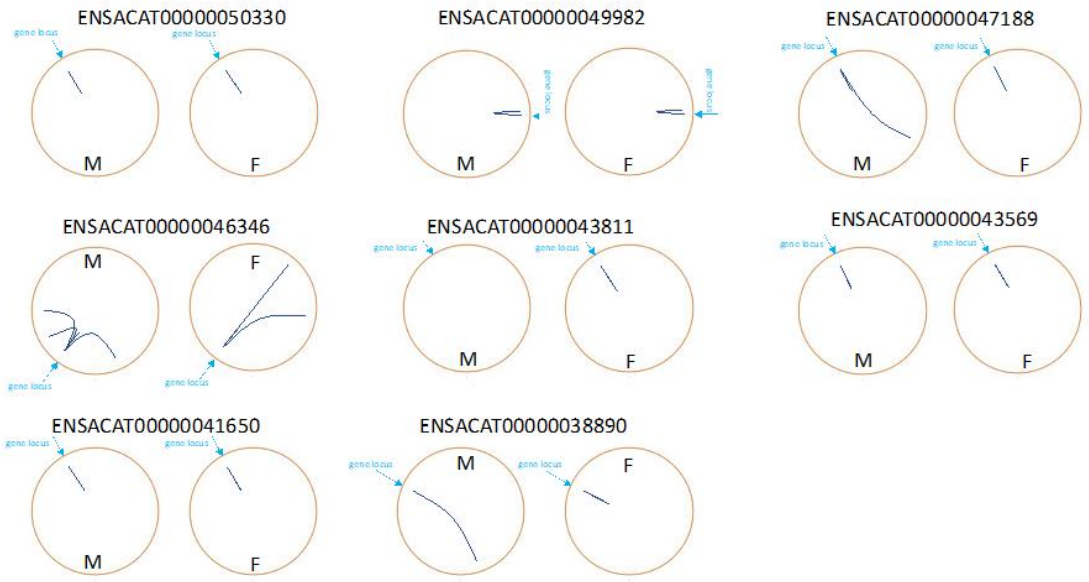


Fig. S5. Results from the ChAR-seq experiment, designed to capture RNA molecules associated with chromatin. We show the contacts between the chromatin and mRNAs of annotated lncRNAs on the X chromosome in *A. carolinenesis*. Blue arrows indicate the position of the lncRNA on the X chromosome. The dark blue lines indicate RNA-DNA contacts. M is the male sample. F is the female sample. In most cases, RNA-DNA contacts were mapped to the same locus, that is, the method captured the mRNAs at their transcription site.

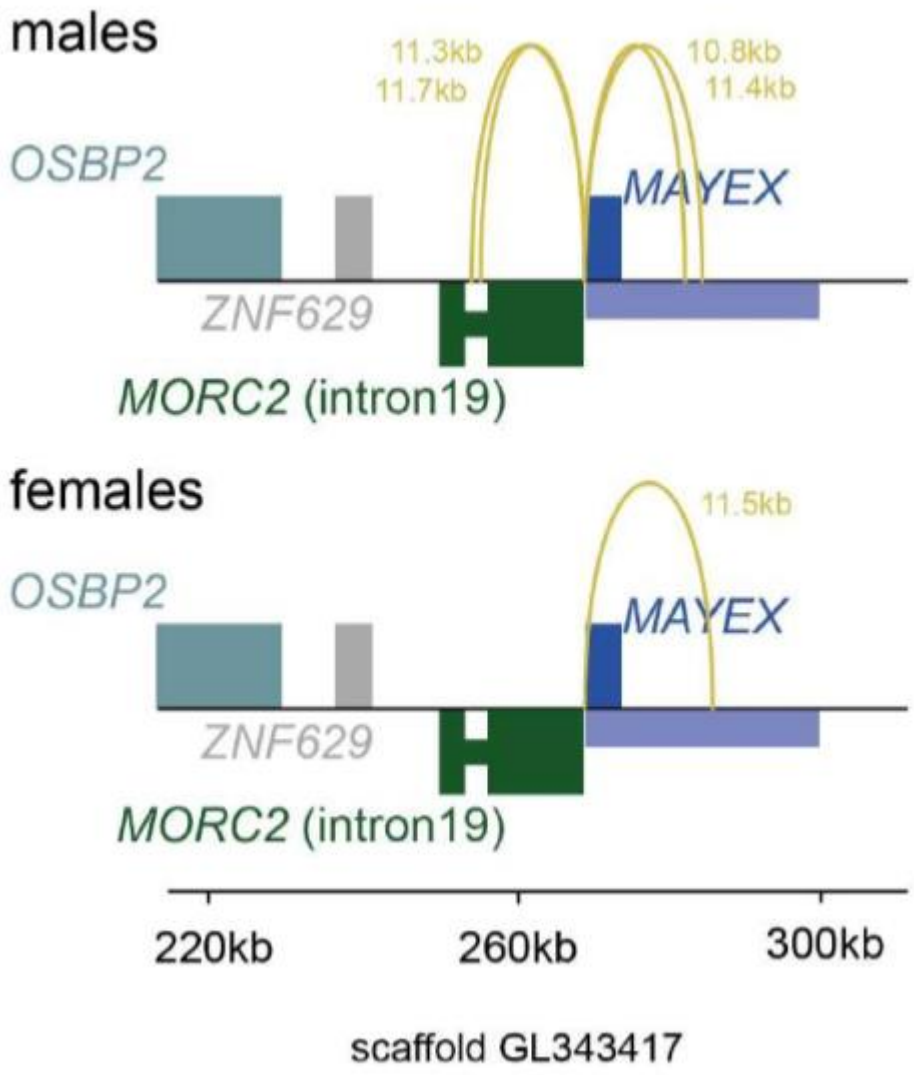


Fig. S6. Hi-C contacts between the CpG promoter region of *MAYEX* and its neighboring region. The promoter region of *MAYEX* shows contacts with the cluster of repeated sequences in the 30Kb region upstream *MAYEX*. In males, in particular, the promoter region is also in contact with intron 19 of the *MORC2* gene.

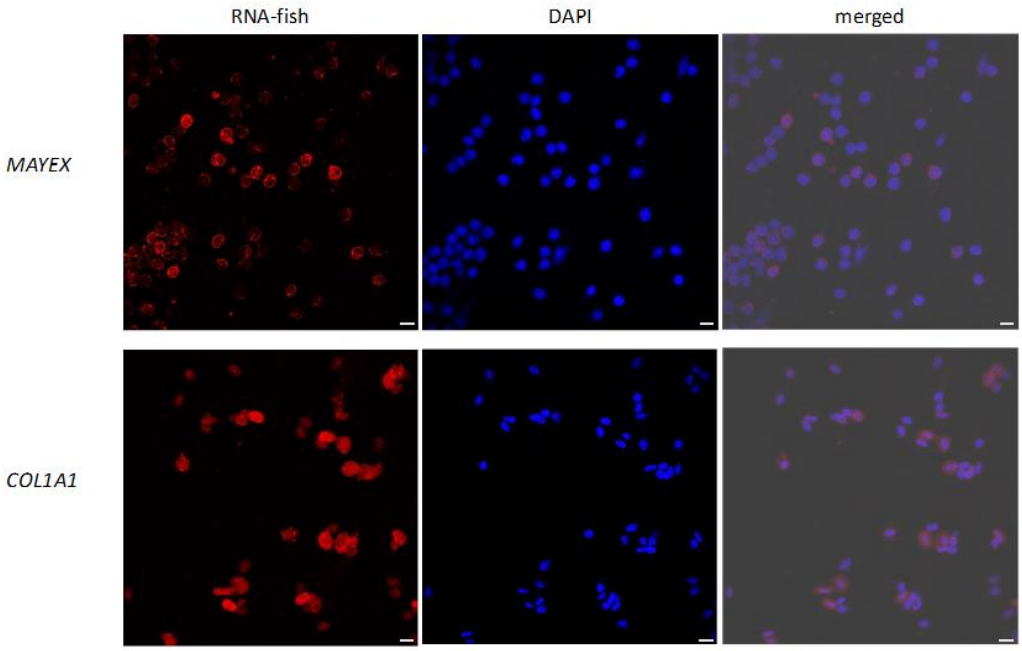


Fig. S7. RNA-fluorescence in situ hybridization (FISH) experiments using blood cells from males. Top images: RNA-fish using a *MAYEX*-specific probe shows distinctive aggregates inside the nucleus in ~70% of the cells, left panel. DAPI staining is shown in blue, middle panel. FISH and DAPI merged images are shown on the right panel. Bottom images: RNA-fish using a probe for a control gene, *COL1A1*, shows homogeneous signal (aggregates in less than 10% of the cells), left panel. DAPI staining is shown in blue, middle panel. FISH and DAPI merged images are shown on the right panel. Scale bar: 10mm.

Fig. S8. High sequence conservation of *MAYEX* in the four species of *Anolis*, representing 44 million years of divergence. Nucleotides shared with the 72-bp cluster repeat found upstream *MAYEX* are highlighted in blue.

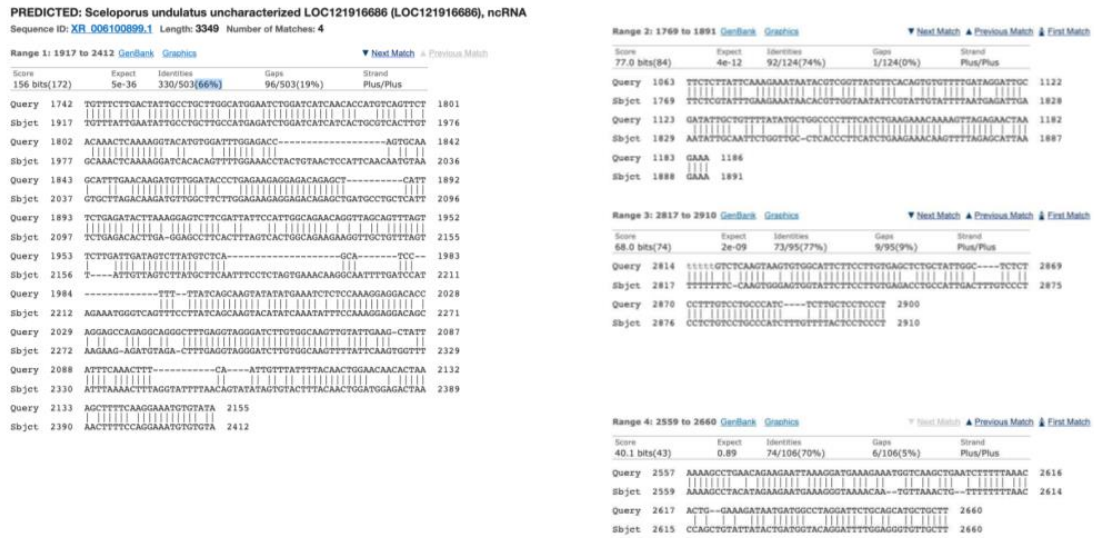


Fig. S9. High conservation of *MAYEX* regions in *S. undulatus*. Conservation is about 66-77% identity, representing 89 million years of divergence.

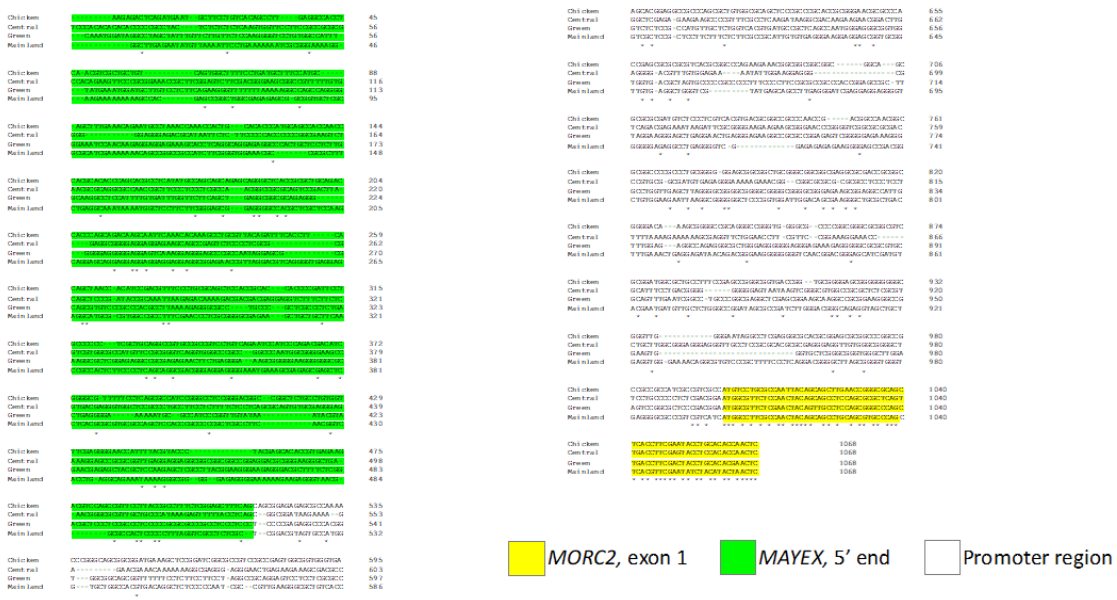
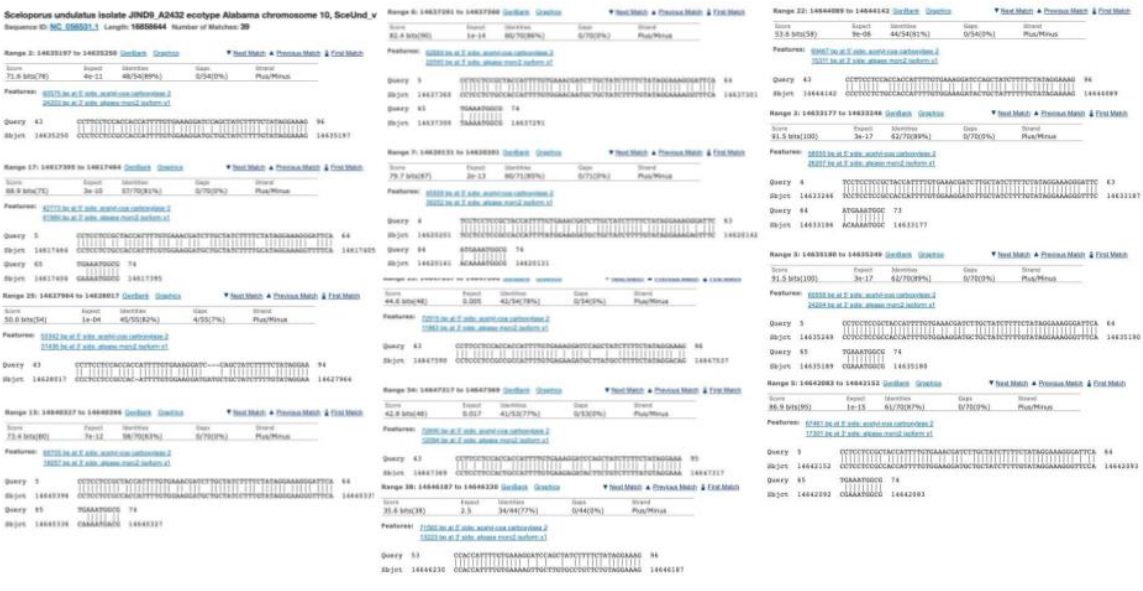


Fig. S10. High sequence conservation of exon 1 of the *MORC2* gene (in yellow) but lack of sequence conservation of the promoter (in white) and *MAYEX* (in green) across reptiles. We analyzed the chicken (Chicken), the central bearded dragon (Central), the mainland tiger snake (Mainland), and the green anole (Green).



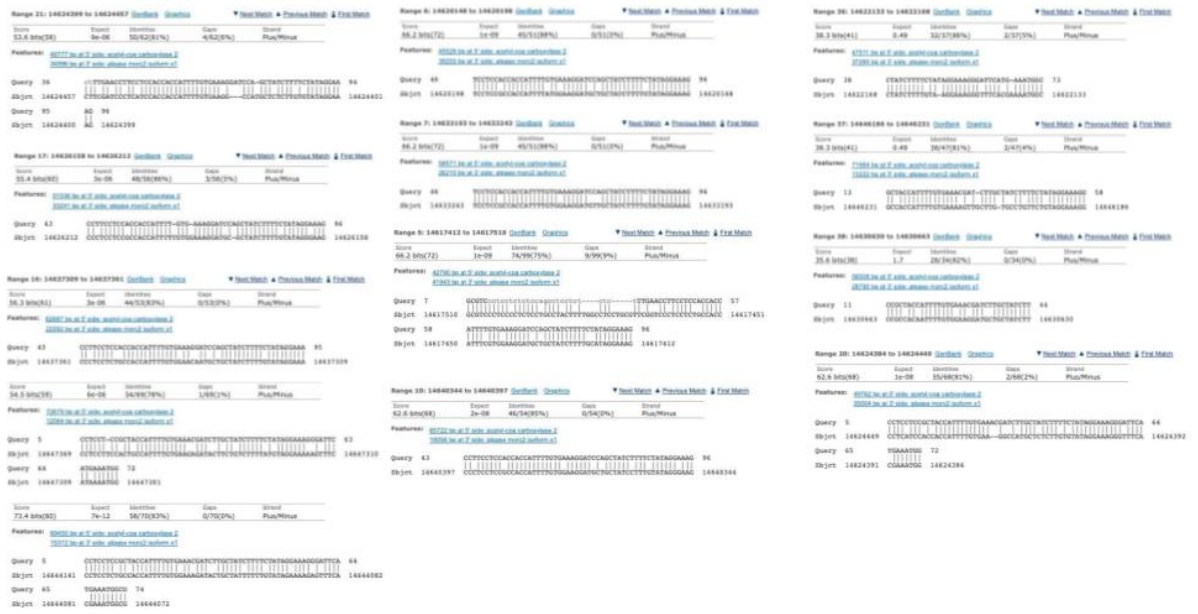


Fig. S11. Copies of the 72-bp repeat found next to *MAYEX* in *S. undulatus* chrX/chr10. *MORC2* chromosome positions are, chrX: 14,659,453-14,690,356; *MAYEX* chromosome positions are, chrX: 14,658,698-14,654,479. Repeats are on chrX: < 14,654,479, in the intergenic region next to *MAYEX*, exactly as in *A. carolinenesis*.

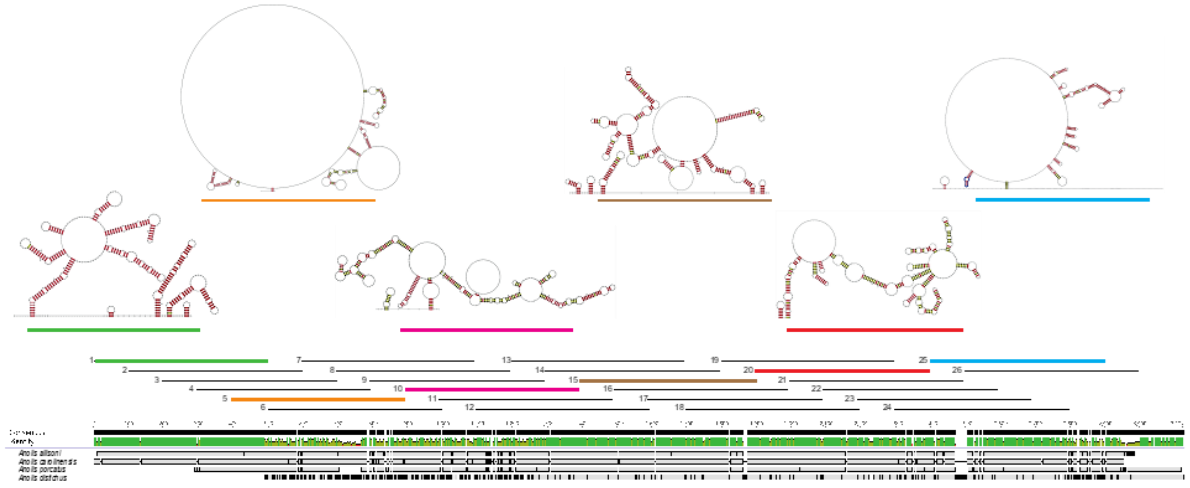


Fig. S12. Secondary structure of *MAYEX* divided into six contiguous domains. Orthologous sequences of *MAYEX* in the four anoles were used to calculate the secondary structure. Top: Four of the domains have conserved loops (green, magenta, brown and red tags), with positions not changing (in red) or co-evolving (in

green) to maintain the secondary structure. Two domains with low sequence conservation are also unstructured (orange and blue tags). Middle: Representation of the 26 windows used for prediction of structure. Bottom: Sequence conservation of the four *MAYEX* genes, their identity (in green), and consensus sequence (in black).

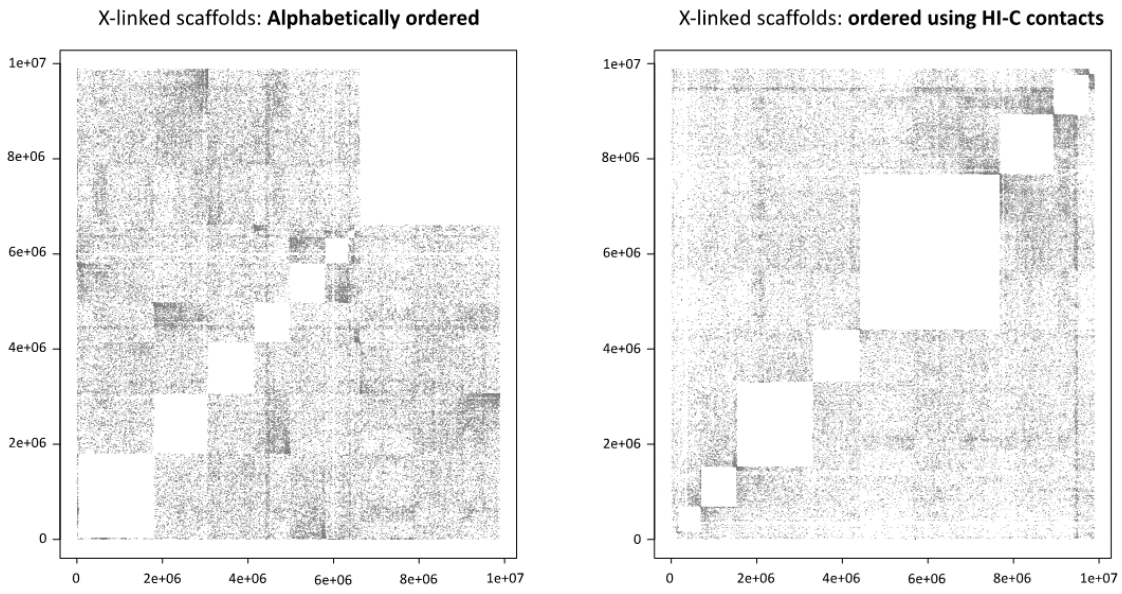


Fig. S13. The X chromosome in *A. carolinensis* is divided into 13 scaffolds. The left panel shows contacts between different scaffolds when the scaffolds were arranged in no particular order. The right panel shows the scaffolds ordered based on their density of contacts. The order used was GL343913.1 (length = 147151 bp), GL345060.1 (length = 12621 bp), GL343550.1 (length = 526944 bp), AAWZ02039360 (length = 8757 bp), GL343423.1 (length = 834740 bp), GL343282.1 (length = 1779868 bp), AAWZ02041299 (length = 5850 bp), AAWZ02040114 (length = 7379 bp), GL343364.1 (length = 1083274 bp, reversed), b (length = 3271537 bp), GL343338.1 (length = 1258094 bp, reversed), GL343417.1 (length = 831895 bp, reversed), GL343947.1 (length = 117443 bp, reversed).

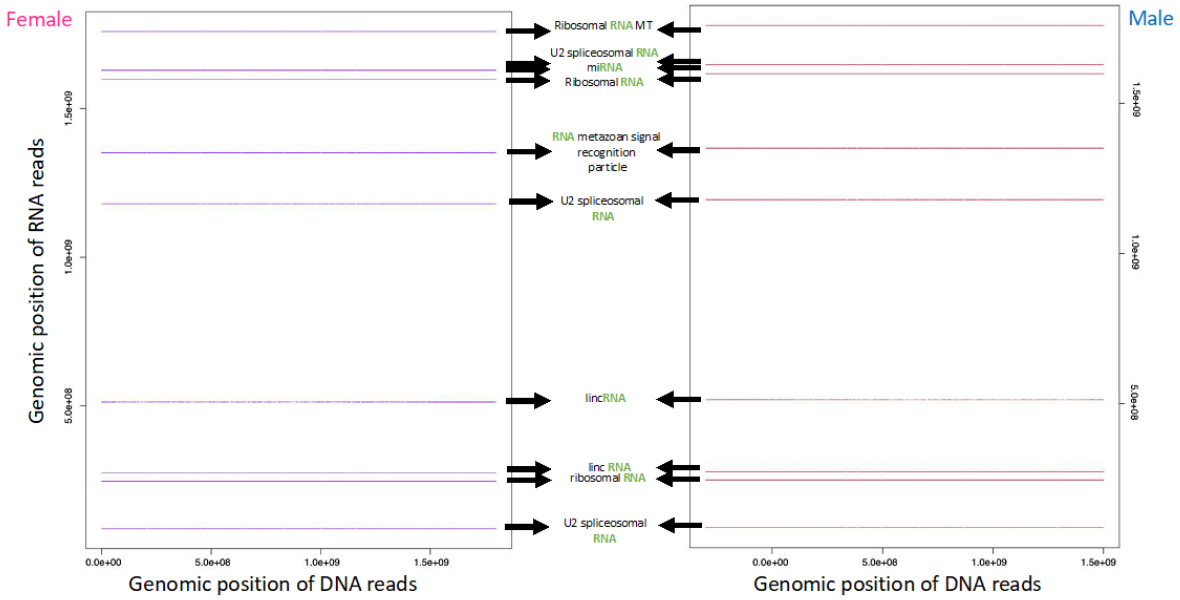


Fig. S14. The quality of ChAR-seq experiments was established by studying the top genes with the highest frequency of contact with the DNA (chromatin). These genes corresponded to some of the most common RNA molecules found in the nucleus, such as ribosomal RNAs, small nuclear RNAs of the spliceosome (i.e., U2), or the RNA metazoan signal recognition particle. In the figure, we observe the top 10 genes with the most frequent contact with chromatin. In all of these cases, the RNA read mapped to a single gene locus and the DNA read mapped to numerous locations across the genome, which is why we see straight horizontal lines.

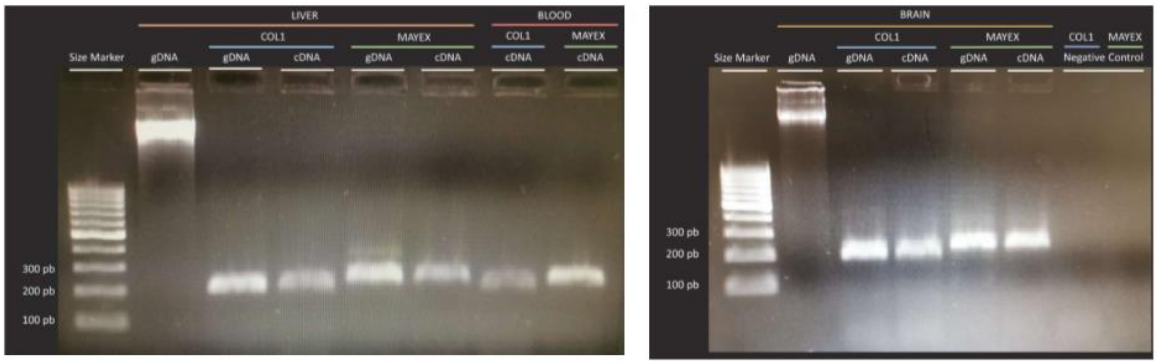


Fig. S15. The probes used for RNA-FISH were obtained from gene-specific amplicons using cDNAs. Primers used to amplify the probes by PCR showed unique ~200 bp amplicons for *MAYEX* and *COL1A1* in male tissues. Probes were designed inside an

exon; this is why amplifications using genomic DNA (gDNA) or cDNA resulted in amplicons of the same size. Amplifications of *MAYEX* and *COL1A1* using cDNA showed that these genes are expressed in blood (tissue used in the FISH experiments), liver, and brain. Standard 30-cycle PCRs were performed. Negative controls did not have gDNA or cDNA.

Supplementary Tables

Supplementary Table 1. Results from a differential expression analyzes of RNA-seq data using annotated genes (male v.s. female samples)

Ensembl ID	logFC	logCPM	LR	PValue	FDR	Gene symbol	Chromosome /scaffold	Function
ENSACAG00000042836	-4.381058	1.99028	68.43132	1.31E-16	3.50E-13	NA	GL343594.1	NA
ENSACAG00000043805	-3.086469	3.7942606	57.93985	2.70E-14	3.60E-11	NA	chr3	lncRNA.
ENSACAG00000010018	-8.204262	1.7935932	55.63665	8.72E-14	7.75E-11	ALPK2	GL343213.1	Protein serine/threonine kinase activity. Involved in heart development and cardiomyocyte differentiation through downregulation of the Wnt/beta-catenin signaling pathway.
ENSACAG00000005247	-3.966321	1.3393136	50.63874	1.11E-12	7.40E-10	LOC100552508	chr2	This gene codes for the transforming protein RhoA, which belongs to the family of homologous genes of Ras (member a). It regulates a signal transduction pathway that links plasma membrane receptors with the assembly of focal adhesions and actin stress fibers. Involved in a microtubule-dependent signal that is required for myosin contractile ring formation during cell cycle cytokinesis.
ENSACAG00000044488	-1.992818	2.4001033	42.41952	7.37E-11	3.93E-08	NA	chr3	NA
ENSACAG00000041946	-3.160352	0.9519483	41.51573	1.17E-10	5.20E-08	NA	GL343438.1	NA
ENSACAG00000007854	-3.299808	-1.744908 3	35.65809	2.35E-09	8.96E-07	GPD2	GL343254.1	Calcium-sensitive mitochondrial glycerol-3-phosphate dehydrogenase activity. This protein is involved in pathways related to glycerophospholipid biosynthesis and triglyceride metabolism.

ENSACAG00000007625	-4.853694	-1.8430286	34.45753	4.36E-09	1.45E-06	GJA1	chr1	The encoded protein is a member of the connexin family and a major component of gap junctions in the heart; which are believed to play a crucial role in the synchronized contraction of the heart and in embryonic development.
ENSACAG00000017750	-1.53595	0.9469	34.20477	4.96E-09	1.47E-06	TKFC	GL343235.1	Glycerone kinase activity and triokinase activity.
ENSACAG00000004594	-6.480361	-1.3194072	33.23449	8.17E-09	2.18E-06	ZAR1L	chr3	The protein encoded by this gene belongs to the ZAR1 family and is predominantly expressed in oocytes and early embryos. It can function as an RNA regulator in early embryos.
ENSACAG00000002413	-5.49398	-1.3090472	32.7834	1.03E-08	2.50E-06	RARA	chr6	Essential for the regulation of germ cell development (induced by retinoic acid) during spermatogenesis, and also for the survival and development of early spermatocytes at the beginning of meiotic prophase. The gene codes for retinoic acid receptor alpha, which regulates transcription in a ligand-dependent manner.
ENSACAG00000006333	-1.72813	0.2766757	32.12505	1.45E-08	3.21E-06	NOP16	GL343548.1	The expression of this gene is induced by estrogens and Myc protein. It is a marker of low survival of the patient with breast cancer. The gene codes for a protein located in the nucleolus.
ENSACAG00000003744	-3.681112	-1.8755903	31.76363	1.74E-08	3.57E-06	LOC100554317	GL343465.1	Voltage-gated potassium channel activity. This gene codes for member 1 of the subfamily C of voltage-gated potassium channels.
ENSACAG00000017835	-2.051775	-0.3871923	29.61296	5.28E-08	9.30E-06	MAP3K6	GL343464.1	This gene encodes a protein serine/threonine kinase that forms a component of the TCR and p38 MAPK signaling pathways. It activates the JNK kinase pathways, but not ERK or p38. One of the diseases associated with MAP3K6 includes gastric breast cancer syndrome.
ENSACAG00000005937	3.460274	1.2878402	29.60776	5.29E-08	9.30E-06	EIF3C	GL343287.1	It contributes to RNA-binding and translation initiation factor activity. This gene codes for one of the components of the eukaryotic translation initiation factor 3 (eIF-3) complex, which initiates the translation of a subset of mRNAs involved in cell proliferation.
ENSACAG00000043728	-8.758891	-0.2859689	29.50399	5.58E-08	9.30E-06	NA	chr2	lncRNA.
ENSACAG00000043959	-2.03139	0.3643597	29.1513	6.69E-08	1.05E-05	NA	chr2	lncRNA.
ENSACAG00000013502	-2.306759	1.2534027	28.54627	9.15E-08	1.36E-05	TRA2A	chr6	This gene codes for a sequence-specific RNA-binding protein that participates in the regulation of pre-mRNA splicing.

ENSACAG00000017957	3.487222	-0.376644	28.38192	9.96E-08	1.40E-05	KCNH4	chr6	Voltage-gated potassium channel activity. The gene is brain-specific and encodes a member of the H subfamily of voltage-gated potassium channels.
ENSACAG00000005001	-3.432682	-2.2343311	28.12973	1.13E-07	1.48E-05	LOC100557381	GL343286.1	Ecotropic viral integration site similar to 5.

Supplementary Table 2. Details of all windows analized for the secondary structure prediction.

id	accession	nseq	eff_nseq	clen	W	bps	bifs	model	cm	hmm	method	name	window	step
F_1-500-best_isoforms_alignment_2.out	0	3	0.85	494	553	148	8	cm	0.59	0.306	locARNA	1	1	500
F_1-500-best_isoforms_alignment_2.out	0	0	0	0	0	0	0	cm	0	0	CMfinder	1	1	500
F_1-500-best_isoforms_alignment_2.out	0	0	0	0	0	0	0	cm	0	0	CMfinder	2	2	500
F_2-500-best_isoforms_alignment_2.out	0	3	0.85	494	517	141	7	cm	0.59	0.319	locARNA	2	2	500
F_3-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	3	0.91	87	103	24	0	cm	0.653	0.401	CMfinder	3	3	500
F_3-500-best_isoforms_alignment_2.out	0	4	1.57	405	1515	101	5	cm	0.589	0.353	locARNA	3	3	500
F_4-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	3	0.91	87	103	24	0	cm	0.653	0.401	CMfinder	4	4	500

F_4-500-best_isoforms_alignment_2.out	0	4	1.54	475	537	116	7	cm	0.59	0.362	locARNA	4	4	500
F_5-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	4	4	33	47	10	0	cm	1.196	1.004	CMfinder	5	5	500
F_5-500-best_isoforms_alignment_2.out	0	4	1.64	385	2072	103	5	cm	0.589	0.333	locARNA	5	5	500
F_6-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	4	4	33	47	10	0	cm	1.196	1.004	CMfinder	6	6	500
F_6-500-best_isoforms_alignment_2.out	0	4	3.94	393	3975	57	5	cm	0.59	0.461	locARNA	6	6	500
F_7-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	3	3	39	54	9	0	cm	1.291	1.172	CMfinder	7	7	500
F_7-500-best_isoforms_alignment_2.out	0	4	2	430	732	87	7	cm	0.59	0.402	locARNA	7	7	500
F_8-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	3	3	39	54	9	0	cm	1.291	1.172	CMfinder	8	8	500
F_8-500-best_isoforms_alignment_2.out	0	4	1.82	440	757	104	5	cm	0.591	0.371	locARNA	8	8	500
F_9-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	4	4	39	76	11	0	cm	0.907	0.65	CMfinder	9	9	500
F_9-500-best_isoforms_alignment_2.out	0	4	1.5	463	633	125	11	cm	0.589	0.337	locARNA	9	9	500
F_10-500-best_isoforms_alignment_2	0	4	0.79	108	127	31	1	cm	0.589	0.32	CMfinder	10	10	500

.fasta.motif.h2_1														
F_10-500-best_isoforms_alignment_2.out	0	4	1.4	474	916	141	7	cm	0.59	0.311	locARNA	10	10	500
F_11-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	4	0.79	108	127	31	1	cm	0.589	0.32	CMfinder	11	11	500
F_11-500-best_isoforms_alignment_2.out	0	4	1.27	478	589	135	8	cm	0.591	0.33	locARNA	11	11	500
F_12-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	4	3.18	44	59	10	1	cm	1.249	1.137	CMfinder	12	12	500
F_12-500-best_isoforms_alignment_2.out	0	4	0.89	480	582	152	10	cm	0.59	0.295	locARNA	12	12	500
F_13-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	4	3.18	44	59	10	1	cm	1.249	1.137	CMfinder	13	13	500
F_13-500-best_isoforms_alignment_2.out	0	4	0.87	496	582	158	10	cm	0.591	0.291	locARNA	13	13	500
F_14-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	4	3.18	44	59	10	1	cm	1.249	1.137	CMfinder	14	14	500
F_14-500-best_isoforms_alignment_2.out	0	4	0.82	497	560	144	9	cm	0.59	0.318	locARNA	14	14	500
F_15-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	4	1.79	59	77	24	0	cm	0.947	0.644	CMfinder	15	15	500
F_15-500-best_isoforms_alignment_2.out	0	4	0.82	485	588	145	9	cm	0.59	0.307	locARNA	15	15	500

F_16-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	4	3.56	40	55	11	1	cm	1.367	1.233	CMfinder	16	16	500
F_16-500-best_isoforms_alignment_2.out	0	4	0	0	0	0	0	cm	0	0	locARNA	16	16	500
F_17-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	4	3.56	40	55	11	1	cm	1.367	1.233	CMfinder	17	17	500
F_17-500-best_isoforms_alignment_2.out	0	4	0.76	486	522	153	9	cm	0.591	0.297	locARNA	17	17	500
F_18-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	4	3.56	40	55	11	1	cm	1.367	1.233	CMfinder	18	18	500
F_18-500-best_isoforms_alignment_2.out	0	4	0.86	485	563	136	13	cm	0.589	0.329	locARNA	18	18	500
F_19-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	4	3.56	40	55	11	1	cm	1.367	1.233	CMfinder	19	19	500
F_19-500-best_isoforms_alignment_2.out	0	4	1.41	485	792	129	9	cm	0.59	0.339	locARNA	19	19	500
F_20-500-best_isoforms_alignment_2.fasta.motif.h1_1.h2_1	0	4	0.93	174	195	45	3	cm	0.591	0.358	CMfinder	20	20	500
F_20-500-best_isoforms_alignment_2.out	0	4	1.25	489	669	147	8	cm	0.59	0.309	locARNA	20	20	500
F_21-500-best_isoforms_alignment_2.fasta.motif.h1_1.h2_1	0	4	0.93	174	195	45	3	cm	0.591	0.358	CMfinder	21	21	500
F_21-500-best_isoforms_alignment_2	0	4	1.6	450	1397	120	5	cm	0.591	0.34	locARNA	21	21	500

.out														
F_22-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	4	4	28	42	9	0	cm	1.17	0.98	CMfinder	22	22	500
F_22-500-best_isoforms_alignment_2.out	0	4	1.33	443	616	121	8	cm	0.59	0.338	locARNA	22	22	500
F_23-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	4	4	28	42	9	0	cm	1.17	0.98	CMfinder	23	23	500
F_23-500-best_isoforms_alignment_2.out	0	4	1.11	442	1363	115	6	cm	0.59	0.351	locARNA	23	23	500
F_24-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	4	4	28	42	9	0	cm	1.17	0.98	CMfinder	24	24	500
F_24-500-best_isoforms_alignment_2.out	0	4	1.36	439	1712	137	7	cm	0.589	0.297	locARNA	24	24	500
F_25-500-best_isoforms_alignment_2.fasta.motif.h1_1	0	4	4	28	42	9	0	cm	1.17	0.98	CMfinder	25	25	500
F_25-500-best_isoforms_alignment_2.out	0	4	1.21	421	1019	105	10	cm	0.591	0.36	locARNA	25	25	500
F_26-500-best_isoforms_alignment_2.fasta.motif.h2_1	0	4	4	41	61	15	1	cm	1.333	1.171	CMfinder	26	26	500
F_26-500-best_isoforms_alignment_2.out	0	4	1.75	476	666	109	9	cm	0.59	0.382	locARNA	26	26	500
F_27-500-best_isoforms_alignment_2.out	0	4	2.1	482	573	87	7	cm	0.59	0.429	locARNA	27	27	500

F_28-500-best_isoforms_alignment_2.out	0	4	3.18	344	1473	14	1	cm	0.59	0.563	locARNA	28	28	500

CAPÍTULO 2 (Artículo requisito)

El artículo “Genome-wide analysis of RNA-chromatin interactions in lizards as a mean for functional lncRNA identification” se centra en los lncRNA presentes en *A. carolinensis*. Los lncRNAs pueden actuar en *cis* regulando la expresión genética en los genes cercanos, o pueden actuar en *trans* regulando genes ubicados en otros cromosomas. En este proyecto utilizamos la técnica Chromatin Associated RNA sequencing (ChAR-seq), la cual captura la conformación cromosómica de contactos entre el RNA y el DNA en todo el genoma. Analizamos la frecuencia de contactos de DNA en diferentes clases de RNA como rRNAs (70% de interacción), lncRNAs (14% de interacción), mRNAs (13% de interacción), snRNAs (2% de interacción), srpRNA (1% de interacción) y snoRNAs (0.04% de interacción). También se determinó la interacción de los RNAs con la cromatina intra-cromosomal o inter-cromosomal. Esto reveló que los rRNA, srpRNA, snRNA y snoRNA presentan contactos en todo el genoma, en cambio, los lncRNA y los mRNA presentan contactos inter-cromosomales. En nuestro análisis se identificaron 2,282 lncRNAs, lo que representa el 71.8% de los 3,176 lncRNAs anotados en el genoma de *A. carolinensis*. Analizamos los tipos de interacción *cis* y *trans* de los lncRNAs y aproximadamente el 87.5% de los lncRNAs mostraron más del 50% de sus contactos en la categoría *cis*-proximal, los cuales interactúan en estrecha proximidad, mientras que el 12.5% exhibieron más del 50% de sus interacciones en la categoría de *trans*-acting. Además, se observó que los lncRNA que presentan *trans*-acting, exhiben una mayor cantidad de interacciones de cromatina en comparación con los *cis*-acting. Utilizando y analizando datos de ChIP-seq identificamos un lncRNA (ENSACAG00000036367) el cual mostró interacciones más frecuentes con loci enriquecidos en H4K16ac. También, identificamos otros dos lncRNAs (ENSACAG00000045045 y ENSACAG00000044053) los cuales mostraron menos interacciones en sitios acetilados. Investigando más a fondo, descubrimos que estos tres lncRNAs se expresan en múltiples tejidos tanto en embriones como en adultos. Este trabajo fue publicado en la revista BMC Genomics 2023 Aug 7;24(1):444. doi: 10.1186/s12864-023-09545-5.

Genome-wide analysis of RNA-chromatin interactions in lizards as a mean for functional lncRNA identification

Mariela Tenorio[1], Joanna Serwatowska[2], Selene L. Fernandez-Valverde[2], Katarzyna Oktaba[2], Diego Cortez[1]

1. Center for Genome Sciences, National Autonomous University of Mexico (UNAM), Cuernavaca, Mexico.

2. Center for Research and Advanced Studies (Cinvestav), Irapuato, Mexico

Background

The eukaryotic cell is home to a plethora of non-coding RNAs, among which ribosomal RNAs, small nuclear RNAs, and small nucleolar RNAs are the most abundant. Ribosomal RNAs play a crucial role in translating messenger RNAs, while small nuclear RNAs are essential for gene splicing, and small nucleolar RNAs guide chemical modifications of other RNA molecules. The most enigmatic group of non-coding RNAs is long-non coding RNAs (lncRNAs) [1], encompassing RNA molecules longer than 200 nucleotides that lack protein-coding potential. Transcriptomic studies have identified thousands of lncRNAs that may possess functional roles in humans and mice [2-10]. Large-scale screenings have associated many of these lncRNAs with regulatory functions [11].

LncRNAs can be categorized into two groups based on their regulatory activity. *Cis*-acting lncRNAs regulate gene expression on the same chromosome from which they are transcribed, whereas *trans*-acting lncRNAs regulate gene transcription on different chromosomes. Some extensively studied *cis*-acting lncRNAs in vertebrates include *XIST* [12-14], *RSX* [15], and *ROX2* [16], which regulate the expression levels of entire

X chromosomes in placental mammals, marsupials, and the fruit fly, respectively. Other characterized lncRNAs can regulate genetic imprinting [17], recruit protein complexes that modify chromatin [18], or influence the expression of remote genes [19]. Although a few *trans*-acting lncRNAs have been experimentally studied, their number remains limited. Notably, *HOTAIR* [20, 21], a lncRNA known to silence the *HOXD* gene by recruiting the Polycomb Repressive Complex 2 has faced challenges regarding its *trans*-activity following a recent study that analyzed *HOTAIR* knockout mice [22]. Another example is FIRRE, a trans-acting lncRNA involved in hematopoiesis [23].

For more than two decades, hybridization capture methods have been the standard technique for identifying the DNA and proteins associated with specific lncRNA [24]. These methods, known as one-to-all approaches, employ biotinylated DNA probes to selectively purify a lncRNA that has been cross-linked to their adjacent DNA and binding proteins. The most renowned techniques include Chromatin Isolation by RNA Purification (ChIRP) [25], Capture Hybridization Analysis of RNA Targets (CHART) [26], and RNA antisense purification (RAP) [27].

Recently, four all-to-all approaches have emerged to capture all possible interactions between RNA molecules and chromatin. These methodologies are designed to provide comprehensive insights into RNA-genome interactions. The four methods are MARGI (Mapping RNA–Genome Interactions) [28], ChAR-seq (Chromatin-Associated RNA sequencing) [29, 30], GRID-seq (Global RNA Interaction with DNA sequencing) [31], and RADICL-seq (RNA And DNA Interacting Complexes Ligated and sequenced) [32]. These methodologies involve capturing RNAs in contact with DNA by employing specific short linkers that ligate an RNA fragment to an adjacent DNA fragment. Both

MARGI and CHAR-seq enable the sequencing of long RNA-DNA tags. In a successful application of MARGI, researchers used human cells to demonstrate that *XIST* exhibits long-range binding sites along the female X chromosome [28]. Similarly, ChAR-seq was employed in *Drosophila* to unveil the detailed map of RNA-DNA contacts of *ROX2* along the X chromosome of males [30].

While these all-to-all techniques have proven successful in model species such as humans, mice, and *Drosophila*, RNA-DNA contact maps have yet to be explored in other species. In recent years, the number of reptile genomes deposited in public databases has increased by over 600% (from 17 genomes before 2018 to 123 genomes between 2018 and 2023). However, gene annotations in these genomes typically rely on automated modeling of gene predictions based on protein-coding genes from species with curated annotations. In some cases, such as the reference genome of the green anole lizard, *Anolis carolinensis*, transcriptomic data was utilized to enhance the annotations of coding and non-coding genes [33]. The current version of the genome of *A. carolinensis* contains 3,176 lncRNAs (https://www.ensembl.org/Anolis_carolinensis/Info/Annotation), yet most of them lack functional information. In this study, we hypothesized that ChAR-seq-like methods could aid in identifying potential regulatory lncRNAs in genomes where they have been predicted. It should be noted that the ChAR-seq method can provide information about RNA molecules that interact with DNA but is unable to report interactions with other molecules, such as proteins. Therefore, we applied the ChAR-seq method in *A. carolinensis* to investigate the overall map of interactions between RNA molecules and chromatin. We characterized the frequencies of contact for different classes of RNAs and annotate *cis*- and *trans*-acting lncRNAs. By correlating the ChAR-seq results with ChIP-seq data for the acetylation of lysine 16 on histone H4 (H4K16ac) epigenetic

mark, we identified three lncRNAs with *trans*-activity that likely play a role in gene expression regulation in *A. carolinensis*.

Results

Variations in the frequency of associations between RNA and chromatin

To investigate the interactions between RNA molecules and chromatin in *A. carolinensis*, we employed the ChAR-seq method on two adult liver samples. By utilizing a specialized short linker, we captured RNA molecules in contact with chromatin and sequenced a total of 1,020,074,230 and 9,96,050,284 reads from the two biological replicates. The paired reads were then mapped to generate a comprehensive genome-wide map of RNA-chromatin interactions.

We analyzed unique interactions for each class of RNA present in the cell. As expected, highly abundant RNA molecules, such as ribosomal RNAs (rRNAs), were overrepresented in our results. Based on the gene annotations available for *A. carolinenesis*, we found that ribosomal RNAs (rRNAs) accounted for 70% of the interactions, whereas long non-coding RNAs (lncRNAs) represented 14%, messenger RNAs (mRNAs) 13%, small nuclear RNAs (snRNAs) 2%, the metazoan signal recognition particle RNA (metazoan srpRNA) 1%, and small nucleolar RNAs (snoRNAs) 0.04% (Figure 1a,b). The rRNA genes exhibited interactions with numerous chromatin sites (average contacts per gene = 1,132,775), whereas the three annotated metazoan srpRNAs showed an average of 55,000 contacts with chromatin across the genome. The frequency of contacts for other RNA types ranged between 50 and 80,000 (Figure 1a,b), with consistent patterns across the two replicates (Figure 1a,b). Given the potential presence of unannotated non-coding

genes in the *A. carolinensis* genome, we analyzed our RNA-chromatin interactions dataset using a sliding window of 10Kb. This analysis revealed 9,000 transcribed regions that do not overlap with known coding or non-coding genes, exhibiting between 500 to 87,000 chromatin interactions (Figure 1a,b; unkRNA).

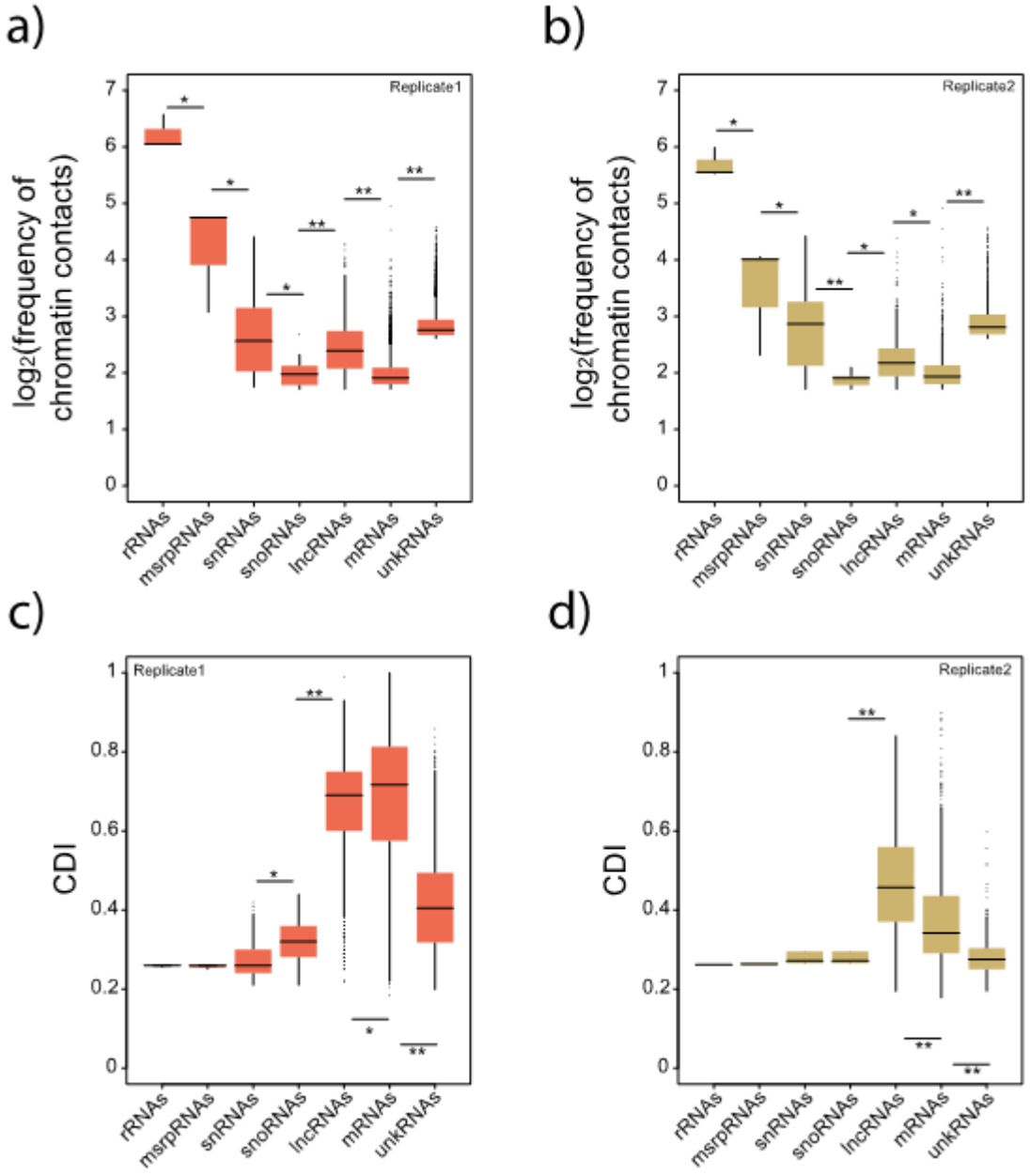


Fig. 1. Chromatin contacts for seven different classes of RNAs. (**a**) Boxplots representing the log_2-transformed ratio of the frequency of chromatin contacts for seven different types of RNA molecules. (**b**) Same as (a) for replicate 2. (**c**) Boxplot representing the values of the Contact Distribution Index (CDI) for seven different classes of RNA molecules. (**d**) Same as (c) for replicate 2. N values for replicate 1 are ribosomal RNAs, 3; metazoan signal recognition particle RNAs, 3; small RNAs, 70; small nucleolar RNAs, 33; long non-coding RNAs, 1188,

messenger RNAs, 3724, unannotated RNAs, 3338. N values for replicate 2 are ribosomal RNAs, 3; metazoan signal recognition particle RNAs, 3; small RNAs, 6; small nucleolar RNAs, 5; long non-coding RNAs, 756, messenger RNAs, 769, unannotated RNAs,756. Data for lncRNAs, mRNAs, and unkRNAs are limited to chromosomes 1-6. Error bars, maximum and minimum values, excluding outliers. Significant differences, Mann-Whitney U test; * represents $P < 0.01$, ** represents $P < 0.001$. *P*-values were corrected using the Benjamini-Hochberg method.

Subsequently, we determined whether the RNA-chromatin contacts were localized within the same chromosomes (intra-chromosomal) or involved contacts across different chromosomes (inter-chromosomal). To assess this, we introduced a Contact Distribution Index (CDI), which was calculated by dividing the number of contacts on the chromosome with the highest interaction count by the total number of contacts across all chromosomes; CDI values around 0.2-0.3 indicated interactions spread across many chromosomes, while CDI >0.4 indicated a bias toward fewer chromosomes. To ensure the reliability of the interactions in *cis* and *trans*, we focused on genes located on the assembled macro-chromosomes (1 to 6) and discarded the short fragments (scaffolds) in the *A. carolinensis* reference genome.

Our analysis revealed that rRNAs, metazoan srpRNAs, snRNAs, and snoRNAs displayed lower CDI values (Figure 1c,d), indicating widespread contacts across the entire genome (Figure 2a,b). This finding aligns with expectations, considering that ribosomal RNAs are the most abundant type of RNA molecules within eukaryotic cell nucleoli. Similarly, snRNAs and snoRNAs, involved in mRNA splicing or RNA chemical modifications, respectively, interact with numerous genomic regions. In contrast, lncRNAs and mRNAs exhibited higher CDI values (Figure 1c,d), suggesting a predominance of intra-chromosomal contacts. Notably, unkRNAs showed lower CDI values (Figure 1c,d), implying the absence of lncRNAs or mRNAs within these transcribed regions. In total, our analysis identified 2,282 lncRNAs in replicates 1 and

2 combined, representing 71.8% of the 3,176 lncRNAs annotated in the *A. carolinensis* genome.

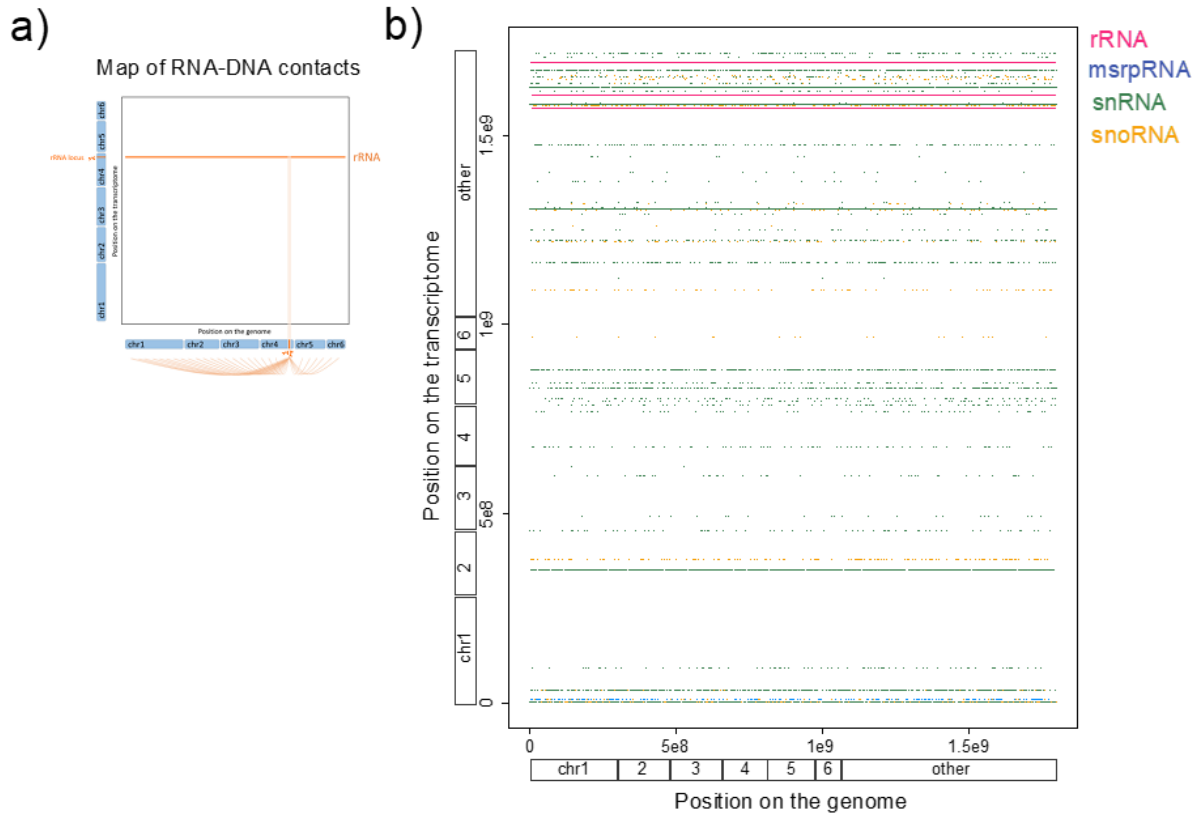


Fig. 2. Contact map of abundant non-coding RNAs between the transcriptome and the genome. (**a**) Illustration to help explain the horizontal lines in panel b; these lines represent the contacts between a single locus on the transcriptome with multiple loci on the genome. (**b**) Dot-plot representation of specific non-coding RNAs loci on the transcriptome (Y-axis) and their multiple genomic contacts (X-axis). The positions on the genome and transcriptome correspond to the concatenated chromosomes 1 to 6 (indicated), followed by the linkage groups and the unassembled scaffolds ordered alphabetically (indicated as other). Ribosomal RNAs are in pink, the metazoan signal recognition particle RNAs in blue, small RNAs in green, and small nucleolar RNAs in orange.

C*is*-acting and *trans*-acting lncRNAs

ChAR-seq data provides valuable insights into distinguishing between lncRNAs that interact with loci in close proximity (*cis*-proximal) or distantly (*cis*-distal) on the same chromosome from which they are transcribed, as well as lncRNAs that have interactions with other chromosomes (*trans*-acting). To examine these types of interactions, we analyzed the lncRNAs on chromosomes 1 to 6 and estimated the

percentage of *cis*-proximal, *cis*-distant, and *trans*-acting contacts. Notably, lncRNAs exhibited a gradient distribution between *cis*-proximal and *trans*-acting interactions, with a substantial bias towards *cis*-proximal interactions (Figure 3a). This distinctive sets lncRNAs apart from other classes of RNA molecules (Figure 3b). Specifically, approximately, 87.5% (n = 1040) of lncRNAs displayed over 50% of their contacts in the *cis*-proximal category (Figure 3c,e; Supplementary Table 1), while 12.5% (n = 148) exhibited over 50% of their interactions in the *trans*-acting category (Figure 3d,f; Supplementary Table 1).

For 98% of the *cis*-acting lncRNAs*,* the majority of their chromatin contacts clustered within 20 Kb around the transcription locus of the lncRNA (Figure 4a,b), aligning with the gene body of the lncRNA. However, upon narrowing our analysis to the top twenty *cis*-acting lncRNAs with the highest number of contacts, the range of interactions increased to approximately 40 Kb around the lncRNA locus (Figure 4c), extending beyond the boundaries of the gene body. Notably, only when examining the top five *cis*-acting lncRNAs with the most contacts, the range of interactions further increased to 100 to 300 Kb around the lncRNA locus (Figure 4d-f), encompassing nearby genes. Overall, our data revealed a substantial number of lncRNAs with contacts at their transcription sites, likely representing nascent transcripts.

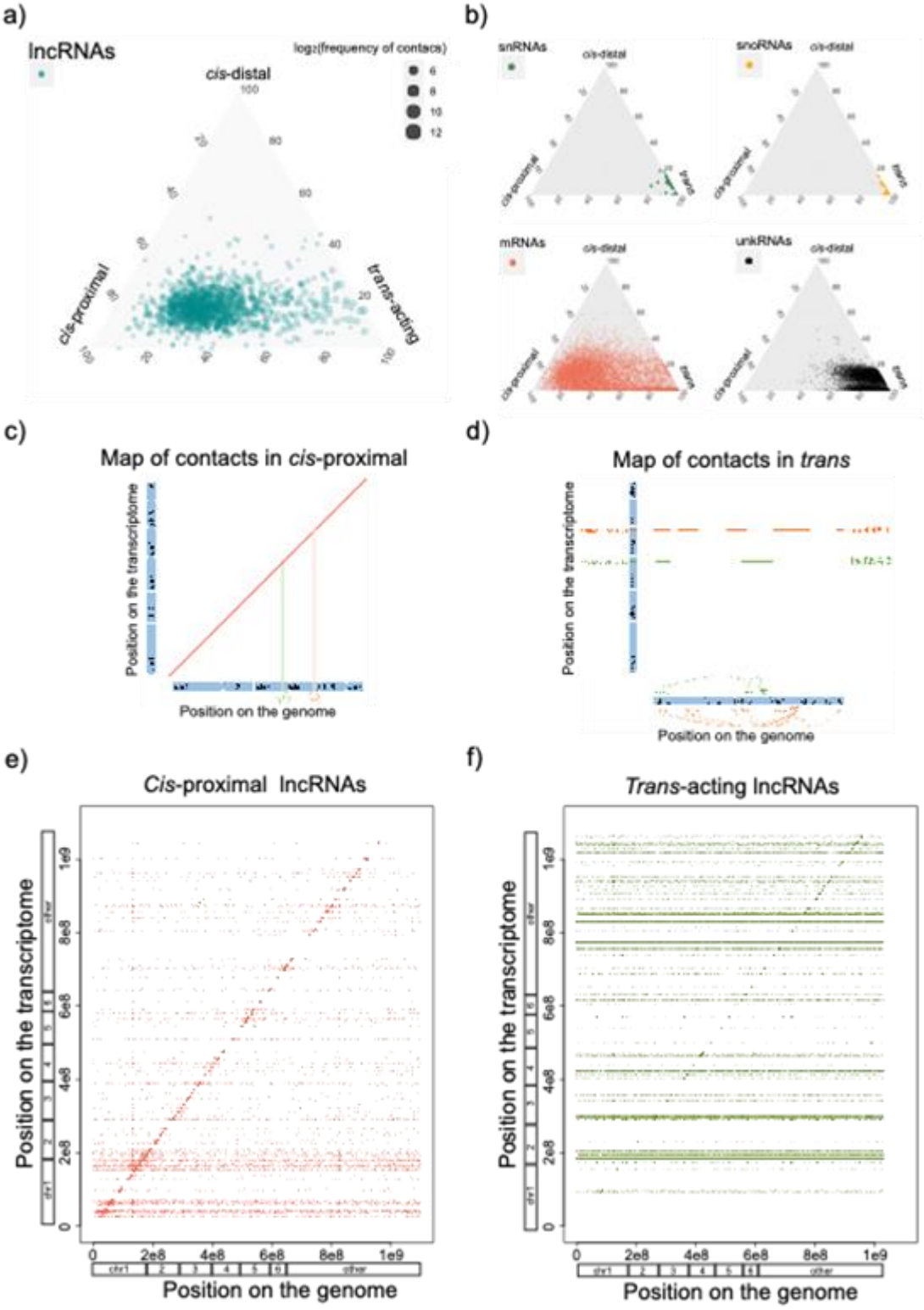


Fig. 3. Contact map of *cis*-acting and *trans*-acting lncRNAs between the transcriptome and the genome. (**a**) Ternary plots representing the type of contacts of lncRNAs; *cis*-proximal (<10 Kb around the gene loci), *cis*-distal (>10 Kb on the same chromosome), *trans*-acting (in other chromosomes). Dot sizes are defined by log2 of the frequency of contacts. (**b**) Same as (a) for small RNAs, small nucleolar RNAs, messenger RNAs, and unannotated RNAs; rRNAs and msrpRNAs are not shown since their contacts are >99.9% *trans*-acting. (**c**) Illustration to help explain the map of RNA-DNA contacts in *cis*-proximal. (**d**) Illustration to help explain the

map of RNA-DNA contacts in *trans* (**e**) *Cis*-proximal lncRNAs have most of their RNA-DNA contacts within their loci, which explains the diagonal line on the dot plot. The dot-plot represents specific lncRNAs loci on the transcriptome (Y-axis) and their multiple genomic contacts (X-axis). (**f**) *Trans*-acting lncRNAs have most of their contacts on other chromosomes, which explains the horizontal lines on the dot plot. Same as (c) for *trans*-acting lncRNAs. The positions on the genome and transcriptome correspond to the concatenated chromosomes 1 to 6 (indicated), followed by the linkage groups and the unassembled scaffolds ordered alphabetically (indicated as other).

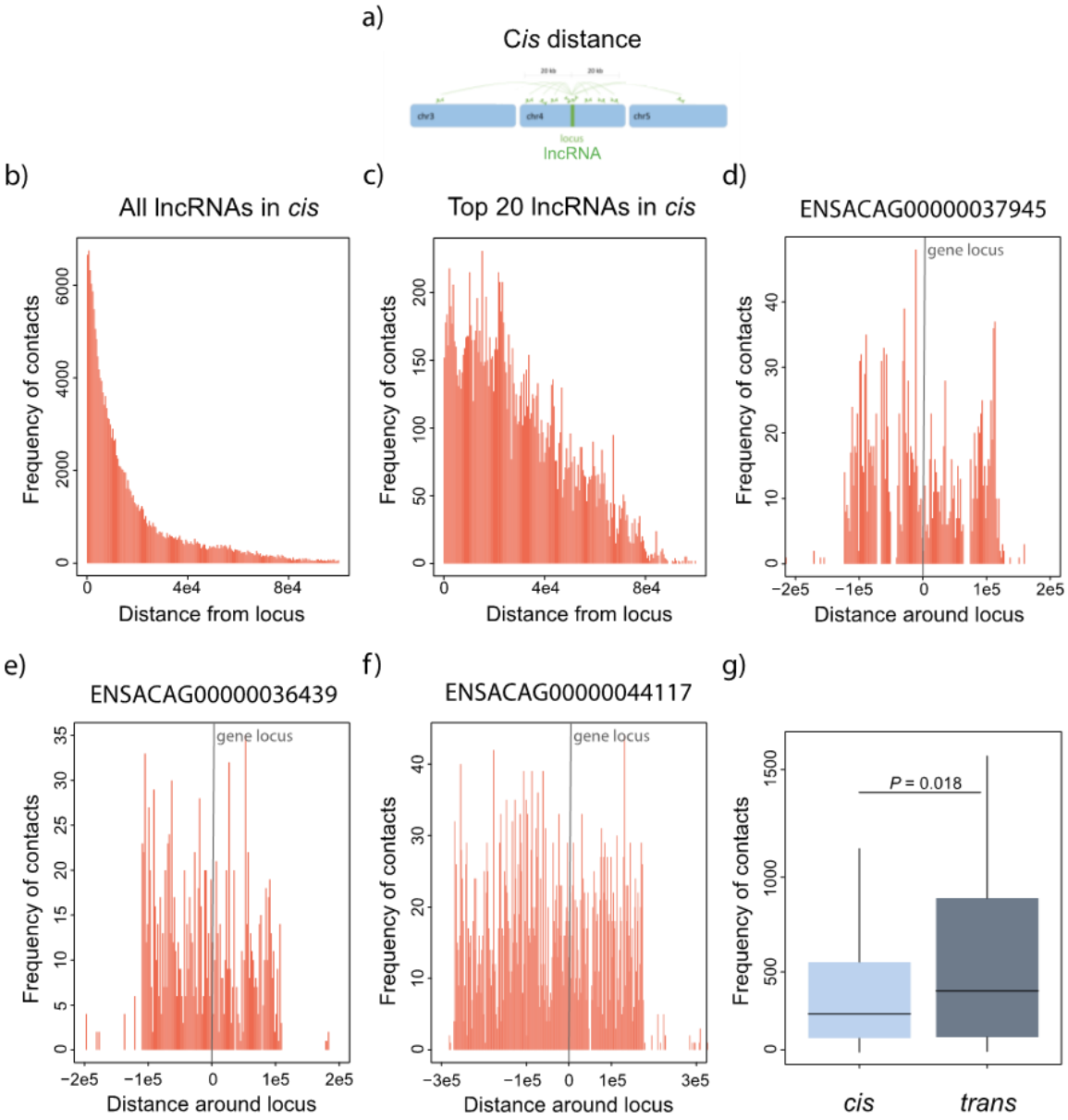


Fig. 4. Range and frequency of RNA-DNA contacts for *cis*-acting lncRNAs . (**a**) Illustration to help explain the RNA-DNA contacts in *cis*. (**b**) Frequency and range of the RNA-DNA contacts (histograms in orange) for all *cis*-acting lncRNAs. (**c**) Same as in (b) but for the top twenty lncRNAs. (**d-f**) Three examples of lncRNAs and their frequency and range of contacts (histograms in orange) around their loci. (**g**) Frequency of contacts between *cis*-proximal and *trans*-acting lncRNAs. Significant differences, Mann-Whitney U test. Error bars, maximum and minimum values, excluding outliers. N values: *cis*-proximal, 1040; *trans*-acting 148.

Furthermore, we observed that *trans*-acting lncRNAs exhibited a larger number of chromatin interactions compared to *cis*-acting lncRNAs (Figure 4g). Focusing on the top ten *trans*-acting lncRNAs, we discovered that they have interactions with all chromosomes, in addition to displaying a peak of interactions at their locus (Figure 5a-k). Some of these *trans*-acting lncRNAs showed interactions throughout the genome, while others displayed interactions with specific genomic regions (Figure 5g-k). Notably, examples such as ENSACAG00000045045 (Figure 5i) and ENSACAG00000030666 (Figure 5j) exhibited an enrichment of chromatin contacts at unassembled scaffolds. Conversely, ENSACAG00000036367 (Figure 5h) and ENSACAG00000039554 (Figure 5k) displayed discrete peaks of interactions distributed along the genome.

Three *trans*-acting lncRNAs exhibit significant associations with the H4K16 acetylation signal

To investigate the potential role of trans-acting lncRNAs in gene expression regulation, we examined the top *trans*-acting lncRNAs and compared their chromatin interaction profiles against ChIP-seq data for the H4K16ac. In the green anole, H4K16ac is known to be enriched at transcription start sites and associated with active transcription [34]. We employed ChIP-seq data generated for both liver (the same tissue used for ChAR-seq) and brain [34]. We assessed whether the RNA-DNA contact regions of lncRNAs displayed a higher coverage of the H4K16ac mark compared to a randomized set of RNA-DNA contacts. Conversely, if a lncRNA is not associated with the H4K16ac signal, the enrichment for H4K16ac will not differ significantly from a randomized set of RNA-DNA contacts.

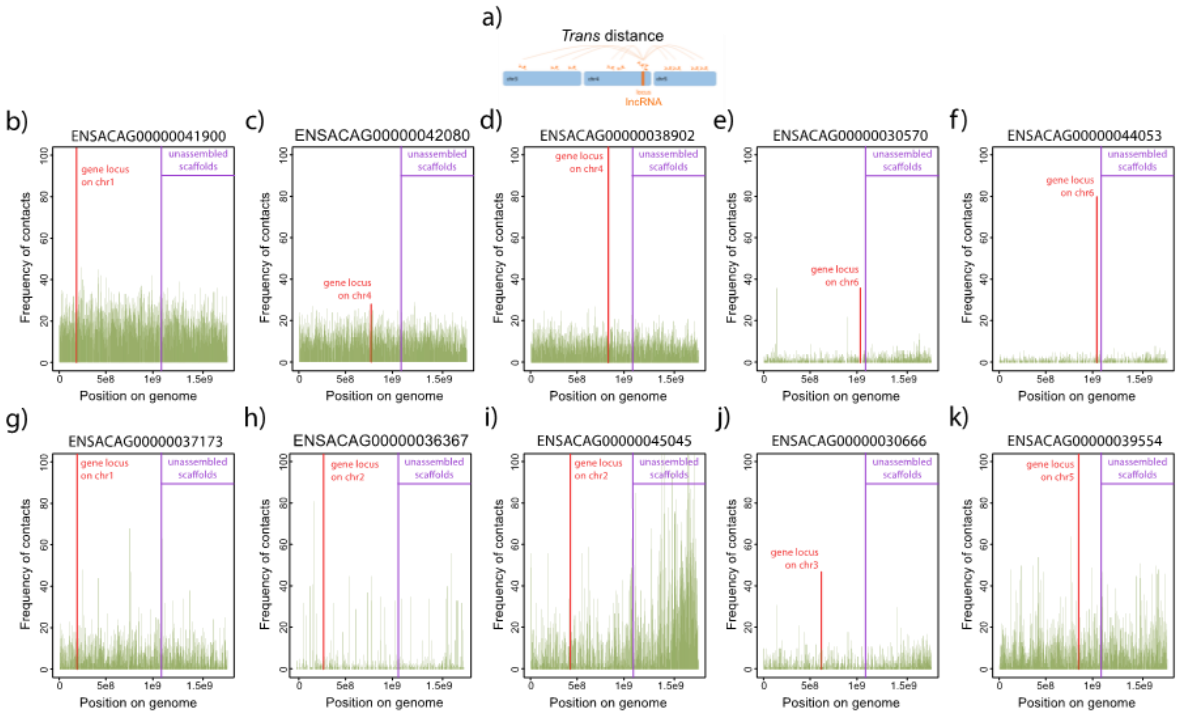


Fig. 5. Frequency of RNA-DNA contacts for *trans*-acting lncRNAs. (**a**) Illustration to help explain the RNA-DNA contacts in *trans*. (**b-k**) Frequency of contacts (histograms in green) across the genome for the top ten lncRNAs with the largest number of contacts in *trans*. The positions on the genome (X-axis) correspond to the concatenated chromosomes 1 to 6, followed by the linkage groups and unassembled scaffolds ordered alphabetically. Purple lines indicate the boundary between the assembled and unassembled parts of the genome. The frequency of contacts at the lncRNAs locus is indicated by the orange histograms; the chromosome where each lncRNA is found is also indicated.

We identified notable associations by analyzing the trans-acting lncRNAs with the highest number of chromatin contacts. One lncRNA, ENSACAG00000036367, exhibited more frequent interactions with loci enriched in H4K16ac (Figure 6a-i). Conversely, two other lncRNAs, ENSACAG00000045045 and ENSACAG00000044053, displayed fewer interactions with H4K16ac sites than expected across samples (Figure 6a-i). We further investigated the transcription profile of these three lncRNAs and found that they are expressed in multiple tissues in both embryos and adults (Figure 6j).

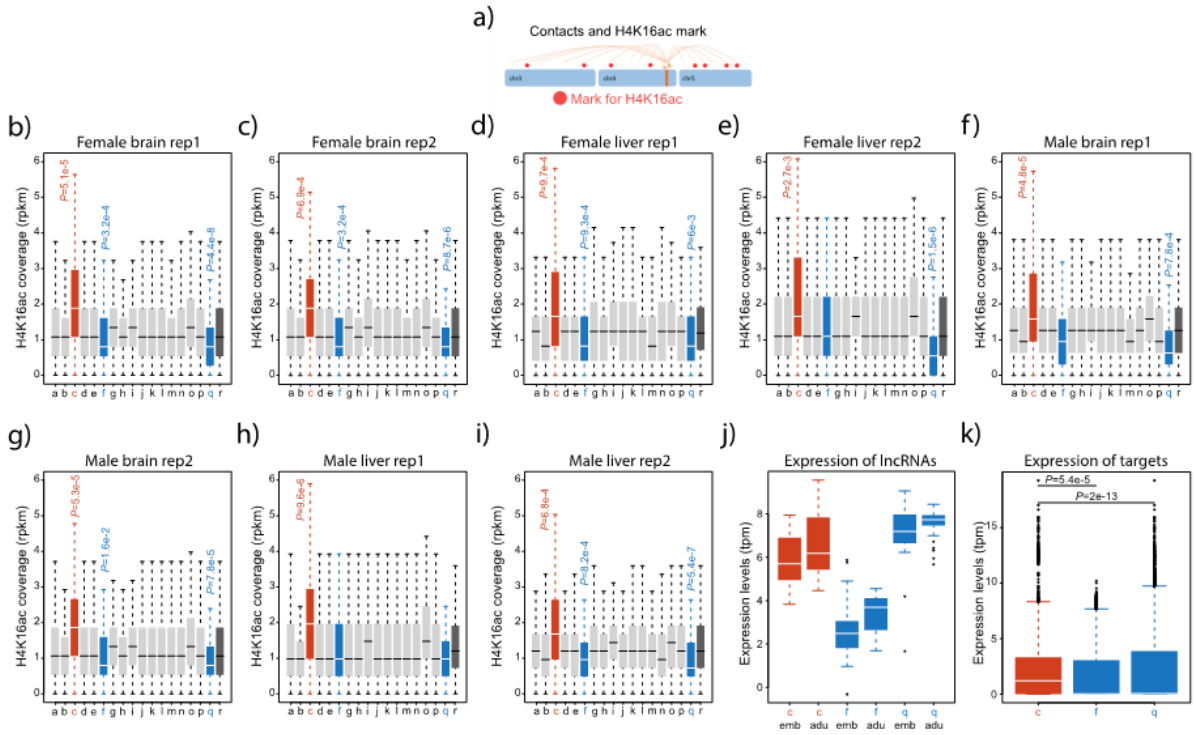


Fig. 6. H4K16ac coverage at chromatin contacts for the top *trans*-acting lncRNAs. (**a**) Illustration to help explain the RNA-DNA contacts and their association with the H4K16ac mark. (**b-e**) Normalized coverage of the H4K16ac mark (RPKM) in female and male livers at the positions where lncRNAs interact with chromatin. The identity of the *trans*-acting lncRNAs is as follows: a is ENSACAG00000041900 (locus on chromosome 1), b is ENSACAG00000037173 (locus on chromosome 1), c is ENSACAG00000036367 (locus on chromosome 2), d is ENSACAG00000031293 (locus on chromosome 2), e is ENSACAG00000034833 (locus on chromosome 2), f is ENSACAG00000045045 (locus on chromosome 2), g is ENSACAG00000041129 (locus on chromosome 2), h is ENSACAG00000030666 (locus on chromosome 3), i is ENSACAG00000040072 (locus on chromosome 4), j is ENSACAG00000044525 (locus on chromosome 4), k is ENSACAG00000042080 (locus on chromosome 4), l is ENSACAG00000038902 (locus on chromosome 4), m is ENSACAG00000039554 (locus on chromosome 5), n is ENSACAG00000032218 (locus on chromosome 5), o is ENSACAG00000042324 (locus on chromosome 6), p is ENSACAG00000030570 (locus on chromosome 6), and q is ENSACAG00000044053 (locus on chromosome 6). The lncRNA in the red boxplot (c) is significantly associated with higher coverage of the H4K16ac signal, whereas the lncRNAs in the blue boxplots (f and q) are significantly associated with lower coverage of the H4K16ac signal. Boxplots in dark grey represent H4K16ac coverage from 100,000 random positions. Significant differences, Mann-Whitney U test. Error bars, maximum and minimum values, excluding outliers. *P*-values were corrected using the Benjamin Hochberg correction. N values for the lncRNAs are 18567, 7525, 2133, 8323, 1176, 14767, 1489, 2842, 873, 3543, 11509, 9044, 10440, 1160, 3159, 1732, 1152. (**f-i**) Same as in (a-d) for female and male brains. (**j**) Expression levels (TPM) from 47 embryonic and 28 adult tissues for ENSACAG0000003636 (in red) associated with a higher signal of H4K16ac and ENSACAG00000044053 and ENSACAG00000045045 (in blue) associated with a lower signal of H4K16ac. (**k**) Expression levels (TPM) from 47 embryonic and 28

adult tissues for the gene targets of ENSACAG0000003636 (in red) and ENSACAG00000044053 and ENSACAG00000045045 (in blue). Significant differences, Mann-Whitney U test. Error bars, maximum and minimum values, excluding outliers. *P*-values were corrected using the Benjamin Hochberg correction. N values are 8112 (c), 15210 (f), and 35646 (q).

Additionally, we explored the putative gene targets of these three lncRNAs: ENSACAG00000036367 exhibited contacts with promoter regions of 86 protein-coding genes and 11 lncRNAs, while ENSACAG00000044053 contacted 368 protein-coding genes and 32 lncRNAs. Similarly, ENSACAG00000045045 had interactions with 768 protein-coding genes and 114 lncRNAs. Although functional enrichment analyses did not reveal any overrepresented biological process or metabolic pathway, an intriguing finding emerged. The expression levels of target genes associated with ENSACAG00000036367, which displayed enrichment in H4K16ac, were significantly higher than those of target genes associated with ENSACAG00000044053 and ENSACAG00000045045, which did not exhibit enrichment in H4K16ac (Figure 6k).

Discussion

In this study, we explored the ChAR-seq methodology to investigate the contact map of chromatin-interacting RNA molecules in a non-traditional model species, *A. carolinesis*. We encountered several challenges associated with the lack of annotations for many non-coding elements and the need to restrict our analyses to chromosomes 1 to 6 due to incomplete genome assembly. Despite these obstacles, we discovered intriguing patterns regarding the RNAs in *A. carolinesis*, which should inspire future studies into RNA-DNA interactions in other non-traditional model species.

Notably, we observed that certain RNAs, such as ribosomal RNAs, snRNAs, and snoRNAs exhibited multiple interactions across the entire genome. This observation

is expected from a successful ChAR-seq experiment and serves as a positive control to validate the reliability of the results. These interactions are likely non-specific, arising from highly transcribed RNA molecules diffusing within the eukaryotic nucleus and being captured in connection with accessible chromatin. It is worth also mentioning that the ChAR-seq method captured significant amounts of nascent transcripts [24], which can also serve as positive controls of the ChAR-seq experiments because they confirm that the method truly trapped RNA molecules that were in contact or in close proximity to adjacent DNA. A recently published technique [32], RADICL-seq, attempts to mitigate the number of nascent transcripts by inhibiting RNA Polymerase II using actinomycin D before cell fixations. Although this technique works effectively with cell cultures, adapting it to bulk tissues from non-model species without available cell lines could pose challenges. Nascent transcripts attached to chromatin may represent regulatory elements within their respective loci. However, differentiating between mature RNAs that regulate their own locus and nascent transcripts attached to chromatin is challenging for *cis*-acting lncRNAs.

Despite the methodological difficulties encountered in working with *A. carolinenesis*, we successfully characterized the type and frequency of contacts made by lncRNAs. The majority of these contacts are either *cis*-proximal or *trans*-acting, exhibiting a pattern distinct from other classes of RNAs, such as ribosomal RNAs, snRNAs, and snoRNAs, which mostly exhibit *trans*-acting interactions. Although lncRNAs may have contacts in *trans* that represent spurious interactions, combining ChAR-seq and ChIP-seq data allowed us to uncover statistical associations that could help differentiate lncRNAs with potential regulatory functions. While our conclusions are based on ChAR-seq data generated from two individuals and may be limited in terms of predicting processes active in a population, the consistency of patterns observed

across different types of data (ChAR-seq, ChIP-seq, and RNA-seq) supports the notion that our findings represent general active processes in *A. carolinensis*. Three lncRNAs are of particular interest due to their significant enrichment in chromatin contacts with the H4K16ac epigenetic mark. ENSACAG00000036367 may be involved in gene activation, while ENSACAG00000045045 and ENSACAG00000044053 could play a role in gene silencing. Further work using gene-specific techniques could reveal associated proteins, such as acetyltransferases or methyltransferases complexes.

The results presented in this study provide a partial glimpse into the interaction map between RNA-DNA molecules in *A. carolinensis*. Our data is limited to lncRNAs with broad expression patterns or specifically expressed in the liver. Currently, the liver is the most suitable tissue for ChAR-seq in lizards due to the considerable amount of starting material required and the small size of the organs. Ideally, future studies will integrate ChAR-seq, ChIP-seq, and RNA-seq analyses using the same sample. To further elucidate the functional characterization of annotated lncRNAs, additional data encompassing various tissues and developmental stages, along with other epigenetic modifications, would be necessary. It is worth noting that many lncRNAs are tissue-specific [35-37], and their characterization would require a broader range of experimental data. Moreover, our focus was solely on lncRNAs interacting with DNA, while numerous lncRNAs may interact with other molecules within the cell. Therefore, investigating lncRNAs associated with the proteome would necessitate an all-to-all protein-RNA protocol.

Methods

Samples

Two adult *A. carolinensis* individuals, one male and one female, were captured in Tampico, Tamaulipas, Mexico (170 m.a.s.l.; SEMARNAT Scientific Collector Permit 08-043). The animals were housed under controlled conditions, with an ambient temperature of 22 ± 2 °C, relative humidity of 55 ± 15%, and a day/night cycle of 12 h/12 h with live food and water *ad libitum*. The animals were housed together in a large terrarium (50.8 cm width, 40.64 cm depth, 20.32 cm height) equipped with UV light. Prior to the experimental procedure, animals were euthanized using a guillotine. The procedure was performed by an experienced technician. All animal procedures were conducted in accordance with the ethical guidelines of the Bioethical Committee of the Universidad Nacional Autónoma de México. The livers were immediately flash-frozen in liquid nitrogen and stored in 1.5 ml tubes at -80°C until use. Due to the requirements in starting material, the liver was the most suitable option for the study because it is the largest organ in lizards. The inclusion/exclusion criteria, randomization, blinding/masking, and outcome measures do not apply to this study since the experiment was carried out with two individuals as biological replicates.

Generation of ChAR-seq data

ChAR-seq is a recent capture method that traps RNA/chromatin interactions [29, 30]. We rapidly homogenized the frozen livers with a mortar and pestle before performing the protein cross-link with formaldehyde (16%, 10 minutes at room temperature, Thermo Scientific, Cat. No. 28908). We then ligated a specific biotinylated linker to the 3’ ends of the RNA molecules using an RNA ligase (T4 RNA Ligase 2, truncated KQ,

NEB, Cat. No. M0373L). The top strand of the linker is a 5′-adenylated ssDNA, HPLC purified, ordered at IDT (https://www.idtdna.com/site/home/) as follows: /5rApp/AANNNAAACCGGCGTCCAA GGATCTTTAATTAAGTCGCAG/3SpC3/. The bottom strand is a biotinylated ssDNA, HPLC purified, ordered at IDT as follows: /5Phos/GATCTGCGACTTAATTA AAGATCCTTGGACGCCGG/iBiodT/T). RNA molecules were reverse transcribed (Bst 3.0 DNA Polymerase, NEB, Cat. No. M0374L), the genomic DNA was cut with a restriction enzyme (DpnII, NEB, Cat. No. R0543L), and the 5' of the adjacent genomic DNA was ligated to the other end of the linker using a DNA ligase (T4 DNA Ligase, HC, Thermo Scientific, Cat. No. EL0013). Proteins were removed using Proteinase K (Thermo Scientific, Cat. No. EO049) and the 2nd strand of the cDNA-linker-DNA molecules was synthesized with Escherichia coli DNA polymerase I (NEB, Cat. No. M0209L). cDNA-linker-DNA molecules were purified with magnetic beads coated with streptavidin (Dynabeads MyOne Streptavidin T1, ThermoFisher, Cat. No. 65601) and prepared for sequencing. The biotinylated linker, which is the key to a successful ChAR-seq experiment, has a 5' end that contains an adenylated single-stranded sequence that can bind RNA and a 3' end that contains a recognition site for adjacent DNA fragments that have been digested by DpnII. The linker's short sequence serves as a molecular tool to control the ligation of RNA and DNA molecules that are in direct contact or in close proximity. More details about the experimental protocol can be found in [29].

Analysis of ChAR-seq data

One male and one female Illumina TruSeq stranded RNA library were sequenced using an Illumina NovaSeq 6000 machine in Novogene, California. We sequenced 1,020,074,230 (male library) and 9,96,050,284 (female library) 150 nucleotides long

paired-end reads. Reads were trimmed using the list of adaptors used in the experimental protocol with trimmomatic (v0.36; parameters: ILLUMINACLIP:illuminaClipping_main.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15) [38]. Valid reads were obtained by extracting sequences that contained the tag sequence (ACCGGCGTCCAAG) present in the linker or its reverse complement (CTTGGACGCCGGT). The orientation of the tag sequence allowed the identification of the RNA and DNA fragments; the 5' end before the tag sequence corresponded to the RNA whereas the 3' end after the tag sequence corresponded to the DNA. Since we sequenced paired-end reads, the orientation of the paired read with the tag sequence relative to the paired read without the tag sequence indicates if the latter represented an RNA or DNA molecule, and was mapped accordingly. DNA fragments longer than 17 base pairs were mapped onto the *A. carolinensis* reference genome (release 104; https://www.ensembl.org) using Bowtie2 (v2.3.4.1; parameters: -p 6 -a -D 20 -R 3 -N 1 -L 18 -i S,1,0.50 --no-unal --no-head --no-sq) [39]. RNA fragments longer than 17 base pairs were mapped to the *A. carolinensis* transcriptome (release 104; https://www.ensembl.org) using Bowtie2 (v2.3.4.1; parameters: -p 6 -a -D 20 -R 3 -N 1 -L 18 -i S,1,0.50 --no-unal --no-head --no-sq) [39]. Since the annotated transcriptome from the Ensembl database could be incomplete, we also mapped using Bowtie2 the RNA fragments to a *de novo* male/female transcriptome generated using Trinity (v2.8.5, parameters: --seqType fq –single –SS_lib_type F –CPU 15 –max_memory 150G) [40]. Finally, RNA fragments longer than 50 base pairs were also mapped to the reference genome using HISAT2 (v2.1.0; parameters: -q --threads 16 -a -N 1 -L 18 -i S,1,0.50 -D 20 -R 3 --pen-noncansplice 15 --mp 1,0) [41]. RNA or DNA fragments that were less than 17 base pairs were not utilized and the paired reads were discarded. We also discarded reads where the RNA or DNA fragments showed

two or more top alignments with the same score (multimappers). We verified the redundancy of the genomic coordinates of fragments that mapped to the transcriptome from Ensembl, the *de novo* transcriptome, and the genome using HISAT2. When a fragment was mapped to the same location in the different databases, we chose the coordinates from the Ensembl transcriptome. IDs from the paired-end reads were used to match the RNA and DNA mapping positions. Valid RNA-DNA contacts were defined as fragments that mapped a single time to their respective databases. We obtained 15,520,949 and 45,816,652 valid contacts for the two replicates. We assigned the contacts to specific genes using the annotations from *A. carolinensis* genome (release 104; https://www.ensembl.org). We reorganized the contacts based on the different classes of RNA molecules. We plotted the number of contacts against the type of RNA molecules and using dot plots we plotted the specific positions of the RNA contacts mapped to the transcriptome against the positions of the DNA contacts mapped to the genome using R [42]. For plotting purposes, we concatenated the chromosomes and unassembled scaffolds. We assigned continuous positions starting with chromosomes 1 to 6, then the linkage groups alphabetically, and finally the unassembled scaffolds alphabetically. We also calculated a Contact Distribution Index for each RNA class based on the chromosome with the maximum number of contacts divided by the total number of contacts in all chromosomes. We plotted the values of the index against the type of RNA molecules using R [42]. *Cis*-acting lncRNAs were defined as those having >50% of their contacts on the same chromosome from which they are transcribed. *Trans*-acting lncRNAs were defined as those having >50% of their interactions in other chromosomes. We calculated the distance from the locus for the *cis*-acting lncRNA as the absolute difference between the middle point of the genomic position of a lncRNA and all the genomic positions of the start of the contacts on the same chromosome.

We also estimated the difference between the middle point of the genomic position of a lncRNA and the genomic positions of its contacts restricted to 20 Kb, 40 Kb, 100 Kb, 200 Kb, and 300 Kb around the lncRNA locus. For *trans*-acting lncRNAs, we limited the analysis to those annotated on the assembled chromosomes 1 to 6 (other scaffolds are too small and tend to have an overestimated number of contacts in *trans*). We mapped the frequency of contacts along the genome, using the positions of the concatenated genome.

Analysis of ChIP-seq data

We downloaded the reference genome and transcriptome of *A. carolinensis* from the Ensembl database (release 104; https://www.ensembl.org). The results regarding the validity and robustness of the ChIP-seq data were published previously in [34]. We downloaded ChIP data for H4K16ac data for brain and liver of two female replicates and two male replicates from the NCBI-SRA database (PRJNA381064). Liver was also the tissue used for ChAR-seq. ChIP-seq data were trimmed for adaptors and low-quality positions using trim_galore (v0.6.2) (https://github.com/FelixKrueger/TrimGalore). ChIP-seq data were mapped to the *A. carolinensis* reference genome using Bowtie2 (v2.3.4.1) [39]. BAM files were indexed and we estimated RPKM values of their coverage along the entire genome for windows of 50 bp using deepTools2 (v3.3.1) (bamCoverage tool) [43]. We then obtained the RPKM values for the positions where a particular lncRNA was in contact with chromatin. We plotted the RPKM values for all contact sites for the top 17 *trans*-acting lncRNAs using R [42]. Significant differences were estimated using the Mann-Whitney U test. *P*-values were corrected using the Benjamin Hochberg correction. Statistical analyzes were conducted in R [42].

Analysis of RNA-seq data

We downloaded RNA-seq data for 15 embryonic and 14 adult tissues from females (including liver samples) and 32 embryonic and 14 adult tissues from males (including liver samples) from the NCBI-SRA database (https://www.ncbi.nlm.nih.gov/sra; PRJNA381064). The results regarding the validity and robustness of the RNA-seq data were published previously in [34]. RNA-seq data were trimmed for adaptors and low-quality positions using trim_galore (v0.6.2) (https://github.com/FelixKrueger/TrimGalore). RNA-seq data were aligned to the reference transcriptome using Kallisto [44] to estimate gene expression levels (TPM). We obtained the TPM values for specific lncRNAs and plotted the TPMs from the 15 embryonic and 14 adult tissues using R [42]. We verified that chromatin contacts were at TSS (+- 5Kb). We identified these genes and carried out functional enrichment analyses using the webgestalt platform (http://www.webgestalt.org/). We used chicken as reference species, and we performed over-representation analyses of geneontology, focusing on biological processes (no-redundant) and pathways (KEGG [45]). We used the Ensembl IDs as input data and the genome protein-coding as a reference set. We also used the string-db platform (https://string-db.org/), standard settings, to explore potential interactions among selected genes. We obtained the TPM values of these target genes and plotted their TPMs from the 15 embryonic and 14 adult tissues using R [42]. Significant differences were estimated using the Mann-Whitney U test. Statistical analyzes were conducted in R [42].

References

1. Salama SR: **The Complexity of the Mammalian Transcriptome**. *Adv Exp Med Biol* 2022, **1363**:11-22.

2. Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H: **Developmental dynamics of lncRNAs across mammalian organs and species**. *Nature* 2019, **571**(7766):510-514.
3. Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J *et al*: **An atlas of human long non-coding RNAs with accurate 5' ends**. *Nature* 2017, **543**(7644):199-204.
4. Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M, Hanna JH, Regev A, Garber M: **Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs**. *Genome Biol* 2016, **17**:19.
5. Bu D, Luo H, Jiao F, Fang S, Tan C, Liu Z, Zhao Y: **Evolutionary annotation of conserved long non-coding RNAs in major mammalian species**. *Sci China Life Sci* 2015, **58**(8):787-798.
6. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H: **The evolution of lncRNA repertoires and expression patterns in tetrapods**. *Nature* 2014, **505**(7485):635-640.
7. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG *et al*: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. *Genome Res* 2012, **22**(9):1775-1789.
8. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A *et al*: **Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression**. *Proc Natl Acad Sci U S A* 2009, **106**(28):11667-11672.

9. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP *et al*: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals**. *Nature* 2009, **458**(7235):223-227.
10. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C *et al*: **The transcriptional landscape of the mammalian genome**. *Science* 2005, **309**(5740):1559-1563.
11. Yu F, Zhang G, Shi A, Hu J, Li F, Zhang X, Zhang Y, Huang J, Xiao Y, Li X *et al*: **LnChrom: a resource of experimentally validated lncRNA-chromatin interactions in human and mouse**. *Database (Oxford)* 2018, **2018**.
12. Patrat C, Ouimette JF, Rougeulle C: **X chromosome inactivation in human development**. *Development* 2020, **147**(1).
13. Payer B, Lee JT: **X chromosome dosage compensation: how mammals keep the balance**. *Annu Rev Genet* 2008, **42**:733-772.
14. Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B: **Xist RNA and the mechanism of X chromosome inactivation**. *Annu Rev Genet* 2002, **36**:233-278.
15. Grant J, Mahadevaiah SK, Khil P, Sangrithi MN, Royo H, Duckworth J, McCarrey JR, VandeBerg JL, Renfree MB, Taylor W *et al*: **Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation**. *Nature* 2012, **487**(7406):254-258.
16. Conrad T, Akhtar A: **Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription**. *Nat Rev Genet* 2012, **13**(2):123-134.

17. Sleutels F, Zwart R, Barlow DP: **The non-coding Air RNA is required for silencing autosomal imprinted genes**. *Nature* 2002, **415**(6873):810-813.
18. Li X, Fu XD: **Chromatin-associated RNAs as facilitators of functional genomic interactions**. *Nat Rev Genet* 2019, **20**(9):503-519.
19. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q *et al*: **Long noncoding RNAs with enhancer-like function in human cells**. *Cell* 2010, **143**(1):46-58.
20. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E *et al*: **Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs**. *Cell* 2007, **129**(7):1311-1323.
21. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL *et al*: **Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis**. *Nature* 2010, **464**(7291):1071-1076.
22. Amandio AR, Necsulea A, Joye E, Mascrez B, Duboule D: **Hotair Is Dispensible for Mouse Development**. *PLoS Genet* 2016, **12**(12):e1006232.
23. Lewandowski JP, Lee JC, Hwang T, Sunwoo H, Goldstein JM, Groff AF, Chang NP, Mallard W, Williams A, Henao-Meija J *et al*: **The Firre locus produces a trans-acting RNA molecule that functions in hematopoiesis**. *Nat Commun* 2019, **10**(1):5137.
24. Kato M, Carninci P: **Genome-Wide Technologies to Study RNA-Chromatin Interactions**. *Noncoding RNA* 2020, **6**(2).
25. Chu C, Quinn J, Chang HY: **Chromatin isolation by RNA purification (ChIRP)**. *J Vis Exp* 2012(61).

26. Simon MD: **Capture hybridization analysis of RNA targets (CHART)**. *Curr Protoc Mol Biol* 2013, **Chapter 21**:Unit 21 25.
27. Engreitz J, Lander ES, Guttman M: **RNA antisense purification (RAP) for mapping RNA interactions with chromatin**. *Methods Mol Biol* 2015, **1262**:183-197.
28. Sridhar B, Rivas-Astroza M, Nguyen TC, Chen W, Yan Z, Cao X, Hebert L, Zhong S: **Systematic Mapping of RNA-Chromatin Interactions In Vivo**. *Curr Biol* 2017, **27**(4):602-609.
29. Jukam D, Limouse C, Smith OK, Risca VI, Bell JC, Straight AF: **Chromatin-Associated RNA Sequencing (ChAR-seq)**. *Curr Protoc Mol Biol* 2019, **126**(1):e87.
30. Bell JC, Jukam D, Teran NA, Risca VI, Smith OK, Johnson WL, Skotheim JM, Greenleaf WJ, Straight AF: **Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts**. *Elife* 2018, **7**.
31. Li X, Zhou B, Chen L, Gou LT, Li H, Fu XD: **GRID-seq reveals the global RNA-chromatin interactome**. *Nat Biotechnol* 2017, **35**(10):940-950.
32. Bonetti A, Agostini F, Suzuki AM, Hashimoto K, Pascarella G, Gimenez J, Roos L, Nash AJ, Ghilotti M, Cameron CJF *et al*: **RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions**. *Nat Commun* 2020, **11**(1):1018.
33. Eckalbar WL, Hutchins ED, Markov GJ, Allen AN, Corneveaux JJ, Lindblad-Toh K, Di Palma F, Alfoldi J, Huentelman MJ, Kusumi K: **Genome reannotation of the lizard Anolis carolinensis based on 14 adult and embryonic deep transcriptomes**. *BMC Genomics* 2013, **14**:49.

34. Marin R, Cortez D, Lamanna F, Pradeepa MM, Leushkin E, Julien P, Liechti A, Halbert J, Bruning T, Mossinger K *et al*: **Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage**. *Genome Res* 2017, **27**(12):1974-1987.
35. Chen L, Zhang YH, Pan X, Liu M, Wang S, Huang T, Cai YD: **Tissue Expression Difference between mRNAs and lncRNAs**. *Int J Mol Sci* 2018, **19**(11).
36. Gloss BS, Dinger ME: **The specificity of long noncoding RNA expression**. *Biochim Biophys Acta* 2016, **1859**(1):16-22.
37. Jiang C, Li Y, Zhao Z, Lu J, Chen H, Ding N, Wang G, Xu J, Li X: **Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs**. *Oncotarget* 2016, **7**(6):7120-7133.
38. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
39. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.
40. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nat Biotechnol* 2011, **29**(7):644-652.
41. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nat Methods* 2015, **12**(4):357-360.
42. Team RC: **R: A language and environment for statistical computing.** . *R Foundation for Statistical Computing, Vienna, Austria URL https://wwwR-projectorg/* 2021.

43. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T: **deepTools2: a next generation web server for deep-sequencing data analysis**. *Nucleic acids research* 2016, **44**(W1):W160-165.
44. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification**. *Nat Biotechnol* 2016, **34**(5):525-527.
45. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic acids research* 2000, **28**(1):27-30.

CAPÍTULO 3

El artículo “X Chromosome Genomics” enfatiza los grandes e interesantes roles que posee el cromosoma X. Después de la aparición del *SRY* en el cromosoma Y, y después de haber sufrido varias inversiones, ambos cromosomas sexuales dejaron de recombinar por su falta de sintenia. La enorme diferencia de contenido de genes entre ambos cromosomas provocó un desbalance en los niveles de expresión entre ambos sexos, lo que desencadenó la evolución de un mecanismo que restableciera la equivalencia de dosis del cromosoma X entre hembras y machos. Por este motivo, en *D. melanogaster* el cromosoma X en los machos se sobre-expresa logrando la misma expresión que los dos cromosomas X de las hembras. Por el contrario, en los mamíferos placentarios y los marsupiales, uno de los dos cromosomas X en las hembras está casi completamente silenciado lo que mantiene el equilibrio de dosis con los machos. La metilación y acetilación de las histonas a lo largo del cromosoma X provoca que mantenga dos estructuras completamente diferentes. Cuando las histonas presentan marcas de silenciamiento encontramos un cromosoma X con muchos más pliegues, en cambio cuando presentan marcas de activación tenemos un cromosoma X mucho más abierto lo que permite a la maquinaria de transcripción entrar fácilmente. Estas dos grandes conformaciones cromosómicas que encontramos en el cromosoma X son reguladas por lncRNAs los cuales residen en el mismo cromosoma que se regula (regulación en *cis*). Estos transcritos participan en el reclutamiento de proteínas modificadoras de la cromatina, que son las encargadas de agregar las marcas de metilación y acetilación, respectivamente. El cromosoma X también está implicado en varias enfermedades y problemas de infertilidad, especialmente entre los hombres. Dado que los hombres solo tienen un cromosoma X, esto los hace más propensos a desarrollar mutaciones de pérdida de función que afectan a la producción del esperma. Además, a medida que envejecemos, algunas células de nuestro cuerpo muestran mosaicismo, lo que significa que pueden llegar a perder el cromosoma X masculino o femenino, o también el cromosoma Y masculino. Este trabajo fue publicado en la revista Reference Module in Life Sciences 2022 Elseiver doi:10.1016/B978-0-12-822563-9.00072-X

X Chromosome Genomics

Mariela Tenorio[1] and Diego Cortez[1]

1. Centro de Ciencias Genómicas. Universidad Nacional Autónoma de México (UNAM). CP62210. Cuernavaca. México

Introduction

Within the vertebrate genome, the most dynamic chromosomes are the sex chromosomes since they are subjected to particular evolutionary forces that can trigger massive genetic loss, the deactivation of large chromosomal regions, the emergence of complex dosage compensation mechanisms that balance expression ratios between males and females through chromatin remodeling, the mass movement of genes, etc. (Bachtrog et al. 2014; Balaton et al. 2018). Moreover, the study of sex chromosomes can be a particularly useful tool to understand basic biological processes related to the appearance of new genes, the regulation of gene expression, the dynamics of epigenetic markers, the effect of sexual selection, the relationship between the environment and gene function, as well as many other equally fascinating topics. This chapter will review some of the most interesting findings related to X chromosome genomics.

Origin of the X chromosome

In humans and many vertebrates, female cells carry two copies of the X chromosome, whereas male cells have one X and one Y chromosome; the Y chromosome is male-specific, highly heterochromatic, and frequently contains only a limited set of genes. XY chromosomes derive from a pair of ancestral autosomes following the emergence of one or multiple sex-determining genes (Bachtrog 2013). For example, sex chromosomes in marsupial and placental mammals originated ~180 million years ago (Cortez et al. 2014) when the Y-linked genes *SRY* (Sinclair et al. 1990) diverged from the X-linked gene *SOX3*. The protein coded by *SRY* could start the testis developmental pathway (Sekido and Lovell-Badge 2008; Li et al. 2014). Soon after the emergence of *SRY*, the Y chromosome underwent a series of large inversions that halted recombination with the X chromosome, resulting in massive gene loss due to the combined effect of lack of homologous recombination, which corrects errors and

deletions, and accumulation of transposable elements that increase the frequency of genetic loss through non-homologous recombination (Furman et al. 2020). Currently, Y chromosomes contain less than 10% of the genetic material of the X chromosome (Bachtrog 2013). The Y-specific genes are located in the male-specific region of the Y chromosome (MSY) (Charlesworth and Charlesworth 2000; Skaletsky et al. 2003). The difference in gene content between the X and Y chromosomes led to an imbalance of expression levels among both sexes, triggering the evolution of a mechanism that could restore dosage equivalence of the X chromosome between males and females. The X chromosome is the carrier of long non-coding RNAs (lncRNAs) that, together with specialized proteins, can reshape the structure and transcription activity (Plath et al. 2002).

Gene expression of the X chromosome

The level at which the X chromosome is expressed is very important for dosage compensation among males and females. Dosage compensation is a regulatory mechanism that acts on entire chromosomes or at a gene-by-gene level and its objective is to equalize the expression levels of genes between both sexes (Mank 2013). For most genes in a genome, males and females have two copies. However, sex-specific duplications or deletions of small chromosomal regions can affect their copy number (Ercan 2015). Particularly, following the decay of the Y sex chromosome, the levels of messenger RNAs and the abundance of the coded proteins differ between sexes, producing different stoichiometries of protein complexes that may not be functional in males causing problems to the organism (Brockdorff and Turner 2015). Some genes are more sensitive to changes in their copy number because a minimal amount of protein is required to achieve a biological function. These genes are known as dosage-sensitive and are the first to be regulated by dosage compensation mechanisms to maintain cellular homeostasis.

Dosage compensation mechanisms acting on sex chromosomes can be limited to regulating dosage-sensitive genes, when a few of these genes are present on the sex chromosomes, as in the case of chicken and platypus (Julien et al. 2012) or they can regulate the expression levels of entire X chromosomes. In *Drosophila melanogaster* (the fruit fly), for example, the X chromosome in males becomes overexpressed to achieve the same expression output as the two X chromosomes in females (Figure 1)

(Conrad and Akhtar 2012). In contrast, in placental mammals and marsupials, one of the two X chromosomes in female cells is almost completely silenced to maintain dosage balance with male cells (Figure 1) (Payer and Lee 2008).

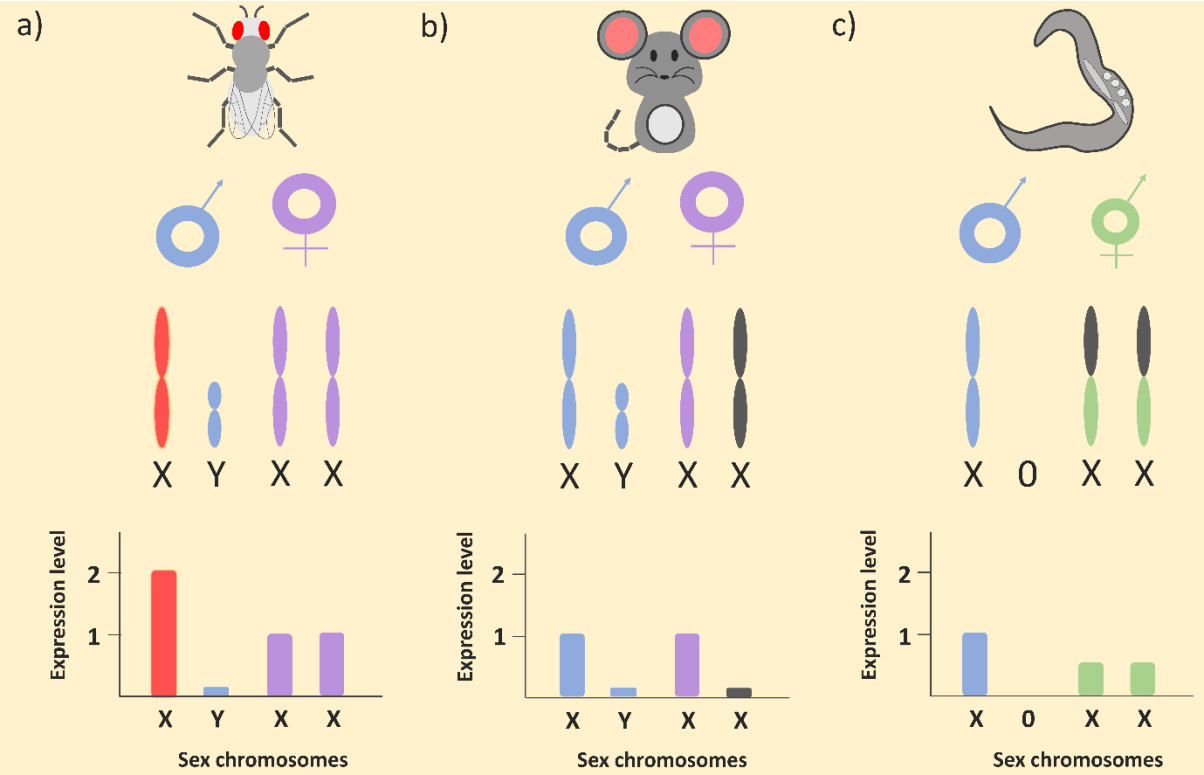


Fig. 1 Expression levels of sex chromosomes. (a) Expression levels of X and Y chromosomes in males and females of the fruit fly. (b) Expression levels of X and Y chromosomes in males and females of placental mammals. (c) Expression levels of X chromosomes in hermaphrodites (with two X chromosomes) and males (with only one X chromosome) in C. elegans.

The chromatin of the inactivated X chromosome in placental mammals is organized into two large super-domains (Figure 2) (Fang et al. 2019), a feature unique to this chromosome since other chromosomes present multiple domains. In humans and mice, several components allow the structure of the two super-domains to be maintained (Balaton et al. 2018). Among them, we have the *dxz4* microsatellite, which is located right at the division of the two super-domains and its presence is critical to forming the structure (Figure 2) (Balaton et al. 2018), and the lncRNA FIRRE that helps to stabilize the super-domains by retaining the repressive histone 3-lysine 27-trimethylation mark and guiding the inactivated X chromosome near the nuclear envelope (Figure 2) (Balaton et al. 2018). The cellular machinery is actively maintaining the repression of gene expression on the inactivated X chromosome, but

even so, there are genes that escape the expression silencing because they need higher expression levels for cellular homeostasis (Berletch et al. 2011; Balaton et al. 2015). Some genes that escape from X inactivation are tissue-specific, although the majority of escaping genes show the same expression levels across multiple human tissues (Tukiainen et al. 2017). Also, genes escaping X inactivation are enriched in gametologs (genes with a paralogue on the Y chromosome) (Bellott et al. 2014). In many species, the X and Y chromosomes share an identical region that continues to have homologous recombination, which is necessary for the correct segregation of XY chromosomes during mitosis and meiosis. This region is known as the pseudoautosomal region (PAR), which is typically located at the edge of the sex chromosomes (Bachtrog 2013); most of the genes found in the PAR region escape inactivation, although they are expressed less than in the active X chromosome (Tukiainen et al. 2017).

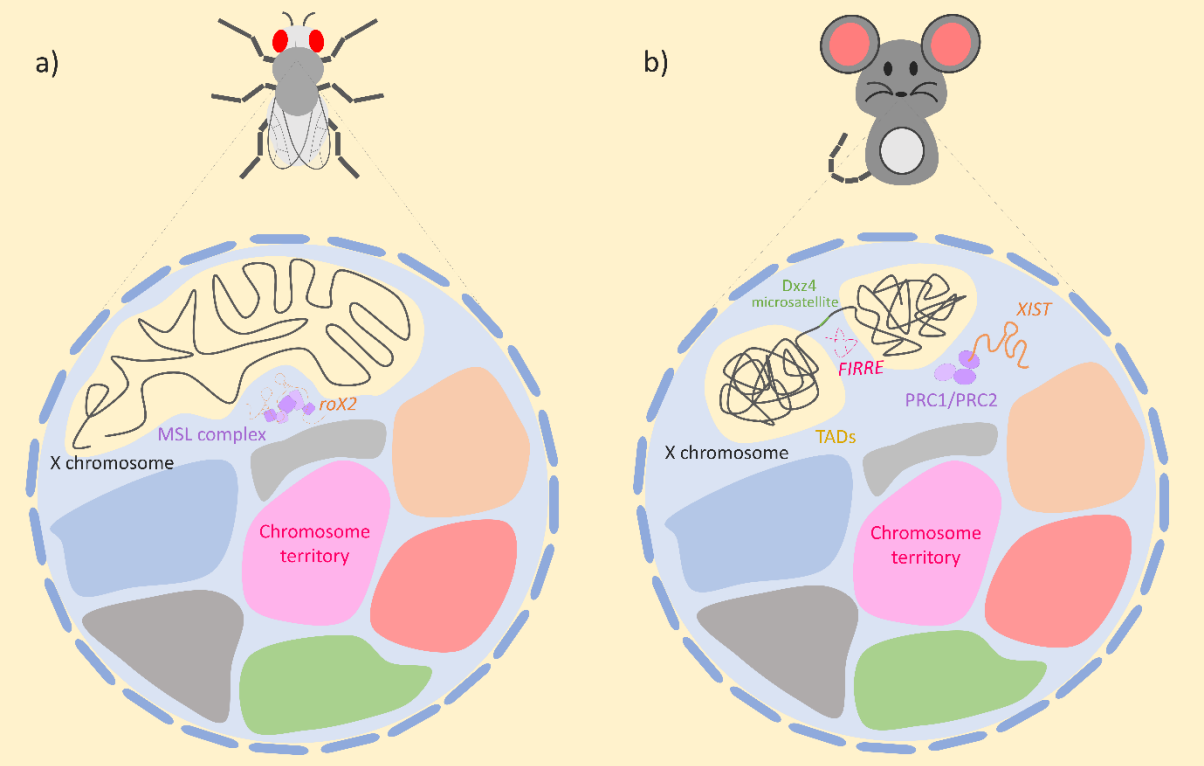


Fig. 2 Nuclear localization, domains, and genetic factors in X chromosome regulation. (a) The X chromosome in the fruit fly is located near the nuclear envelope, with a structure similar to other chromosomes. Overexpression of this chromosome is triggered by the joint activity of the lncRNA ROX2 and the MSL complex. (b) The X chromosome in placental mammals is located near the nuclear envelope, structured into two super-domains (TADs) that are maintained by the dxz4 microsatellite and the lncRNA FIRRE. Inactivation of the X chromosome is controlled by the lncRNA XIST that interacts with the Polycomb repressive

complex (PRC). Chromosome territories are regions of the nucleus preferentially occupied by particular chromosomes in interphase.

X chromosome dosage compensation

X chromosome dosage compensation is driven by lncRNAs that reside on the X chromosome that will be either inactivated or overexpressed. X-specific lncRNAs are involved in the active recruitment of chromatin-modifying proteins that change specific histones by adding acetylation or methylation marks (Figure 3). In placental mammals, the lncRNA XIST (X-inactive specific transcript) is activated by the protein YY1 early during female development (Payer and Lee 2008). XIST can then recruit chromatin-modifying proteins (Polycomb repressive complex -PRC-) that will trimethylate the lysine 27 on histone 3, H3K27me3 (Figure 2-3) (Patrat et al. 2020), therefore, progressively inducing chromatin compaction and transcription silencing (Plath et al. 2002). The XIST gene is highly conserved in humans and mice, and the tandem repetitive regions show some degree of conservation, for example, the A repeats are very well conserved, while the B to F repeats differ between the species. It has also been shown that the internal region of exon 7 of XIST is functionally essential to establishing X chromosome inactivation (Balaton et al. 2018). Similarly, in marsupials, the lncRNA RSX (RNA-on-the-silent X) inactivates the X chromosome probably by interacting with the PRC (Polycomb repressive complex) to inhibit the transcription of the X chromosome (Figure 3) (Sprague et al. 2019). In the fruit fly, the lncRNA ROX2 (RNA on the X) is associated with the male-specific lethal (MSL) protein complex that can hyper-acetylate the lysine 16 on histone 4 (H4K16ac) to increase the transcription output of the X chromosome in males (Figure 2-3) (Gelbart et al. 2009). The CLAMP protein has been shown to promote three-dimensional aggregation of the male-specific lethal dosage compensation complex (MSLc) to promote the three-dimensional molding of the X chromosome so that its active chromatin regions interact with other insulating proteins (Jordan and Larschan 2021). Lastly, in hermaphrodites (XX) of *Caenorhabditis elegans*, dosage compensation is active on both X chromosomes, where the expression levels are lowered by half (Figure 3), rather than completely inhibiting one sex chromosome, to attain the expression levels shown by males (with a single X chromosome) (Strome et al. 2014). The activity of the DC complex is important in this process (Figure 3) (Strome et al. 2014). In *C. elegans*,

however, it is not yet known whether a lncRNA participates in regulating the X chromosomes (Figure 3).

The reasons why in some species dosage compensation is achieved through overexpression and in others through repression of the X chromosome is still debated. It has been proposed that the strength of X and Y cis-regulatory elements may play an important role in the evolution of dosage compensation systems (Lenormand and Roze 2022). Hence, it could be hypothesized that the presence of strong X cis-regulatory elements could lead to X overexpression whereas weak X cis-regulatory elements could be associated with X inactivation.

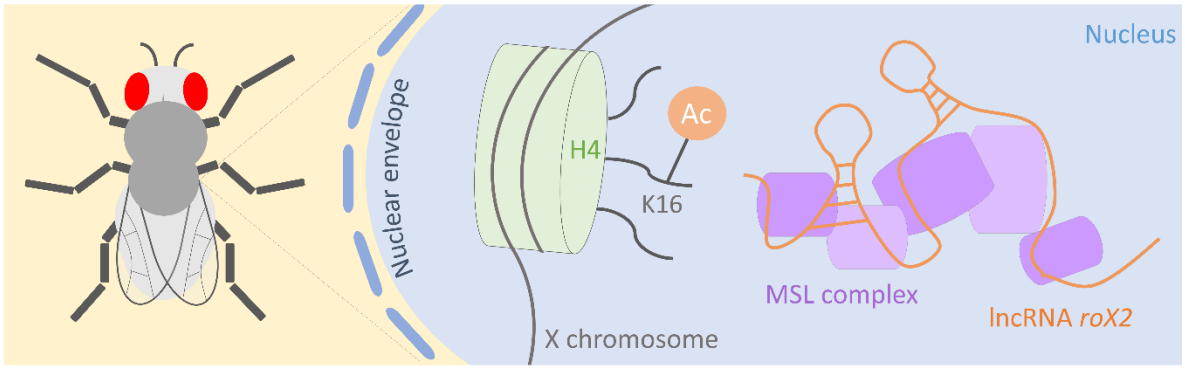
3D structure of the X chromosome

Chromatin-remodeling proteins are one of the main actors in the mechanisms that bring dosage equivalence between sex chromosomes. The lncRNAs indirectly or directly attract these proteins to modify or re-organize specific regions on the X chromosomes (Makki and Meller 2021). The inactive X chromosome in mammals is organized differently from the autosomal and active X chromosomes. As mentioned, the inactive X chromosome shows two topologically associated macrodomains (TADs) that are stabilized near the periphery of the nuclear envelope (Figure 2) (Fang et al. 2019). In *C. elegans*, the chromatin architecture of both X chromosomes in hermaphrodites changes, and the sex chromosomes are relocated closer to the periphery, distinctly from autosomes, which strongly interact with the nuclear lamina (Makki and Meller 2021). Finally, in the fruit fly, the topology of the X chromosome in males is stretched and opts a similar shape to that of the autosomes, which are composed of multiple TADs (Figure 2) (Quinn and Chang 2015; Schauer et al. 2017).

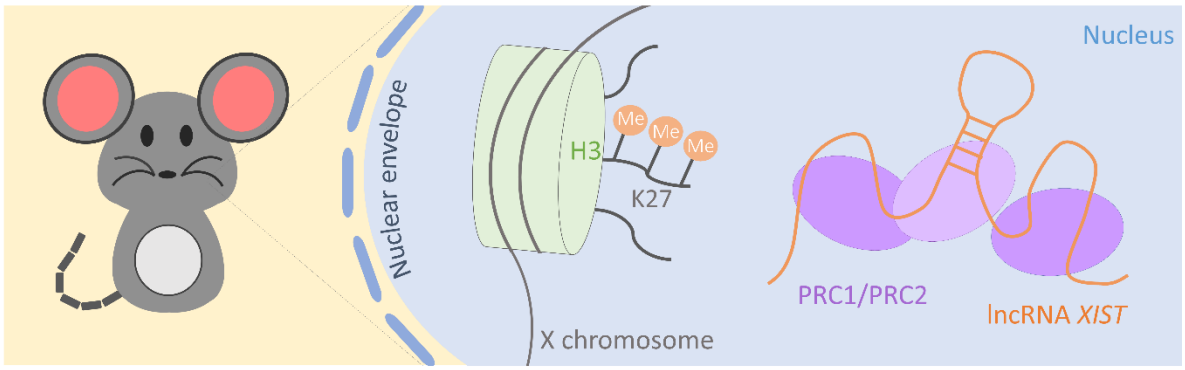
Infertility, Diseases, and Mosaicism

The X chromosome is involved in several diseases and infertility problems, particularly among males. For example, male infertility is commonly caused by defects of sperm. It has recently been revealed that the X chromosome of mammals is enriched in genes that are expressed during spermatogenesis (Vockel et al. 2021). Since males only have one X chromosome, this makes them more likely to develop loss-of-function mutations affecting sperm production. It has also been shown that the X chromosome

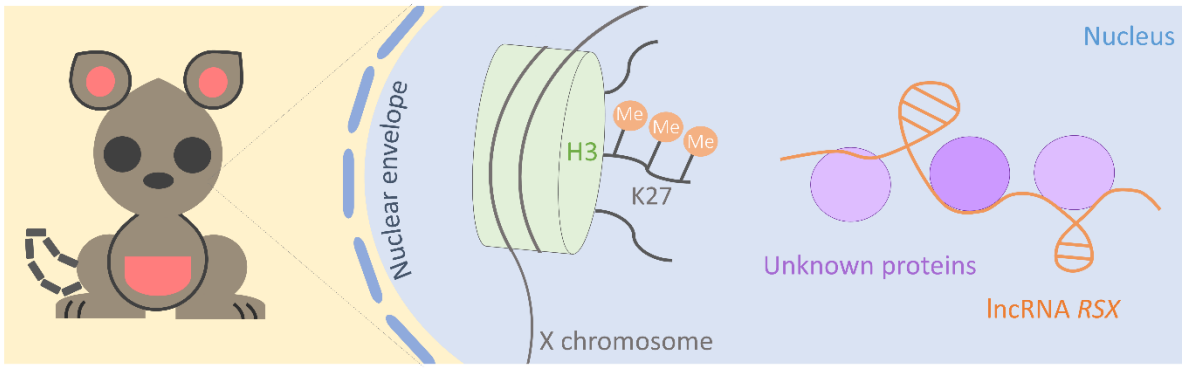
a)


Nucleus
Nuclear envelope
H4
Ac
K16
X chromosome
MSL complex
lncRNA roX2

b)


Nucleus
Nuclear envelope
H3
Me
Me
Me
K27
X chromosome
PRC1/PRC2
lncRNA XIST

c)


Nucleus
Nuclear envelope
H3
Me
Me
Me
K27
X chromosome
Unknown proteins
lncRNA RSX

d)

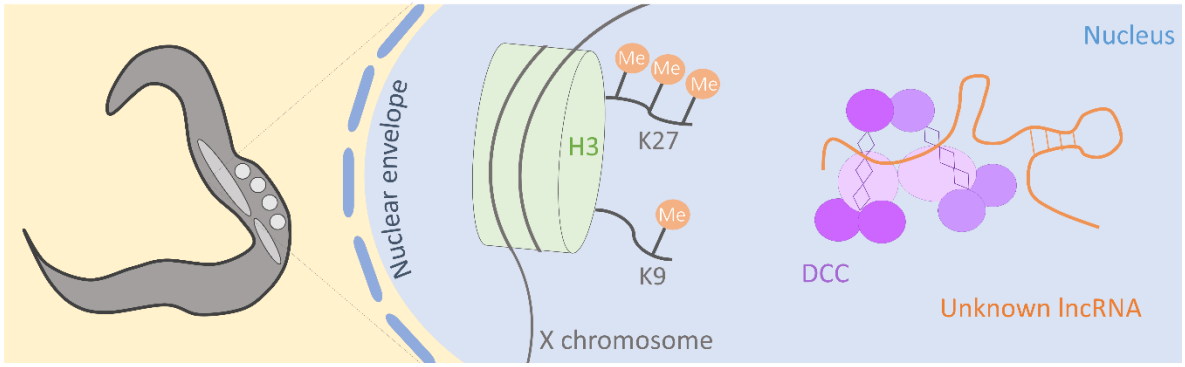

Nucleus
Nuclear envelope
H3
Me
Me
Me
K27
Me
K9
X chromosome
DCC
Unknown lncRNA

Fig. 3 Histone modifications following X inactivation or X overexpression. (a) In the fruit fly, the lncRNA ROX2 interacts with the MSL complex to specifically acetylate the lysine 16 on histone 4 (a transcription activation mark), which initiates the overexpression of the X chromosome. (b) In placental mammals, the lncRNA XIST interacts with the Polycomb repressive complex (PRC) to specifically trimethylate the lysine 27 on histone 3 (a transcription repression mark), triggering the inactivation of the X chromosome. (c) In marsupials, the lncRNA RSX is hypothesized to interact with the Polycomb repressive complex (PRC) to trimethylate the lysine 27 on histone 3 (a transcription repression mark) and induce the inactivation of the X chromosome. (d) In C. elegans, the two X chromosomes in hermaphrodites are downregulated by the tri-methylation of lysine 27 on histone 3 and the methylation of lysine 9 on histone 3 (two transcription repression marks). This process is catalyzed by the DC complex and, potentially, by still unknown lncRNAs.

expression levels in individuals with Turner syndrome (having a single X chromosome) and Klinefelter syndrome (XXY) are more balanced, thus causing the individuals to have milder diseases compared to other trisomies (e.g., trisomy of chromosome 21) that have more serious consequences (Pereira and Doria 2021). Also, aberrations in *XIST* expression and, in some cases, disruption of X-chromosome inactivation as a whole, are related to Alzheimer's disease (Chanda and Mukhopadhyay 2020). Moreover, as we age, some cells in our bodies become mosaic, meaning they can lose either the male or female X chromosome or the male Y. However, recent studies showed that the frequency of loss of the male X chromosome in leukocytes is rare relative to the female X chromosome (Zhou et al. 2021).

Conclusions

Ever since Susumu Ohno proposed that one X chromosome in females of placental mammals could be silenced (Ohno 1967; Beutler 1998), researchers have been fascinated by this process that involves both sex-specific and chromosome-specific genetic signals that regulate the epigenetic landscape and the expression output of entire chromosomes to restore cellular homeostasis between males and females. Similar processes have been investigated in marsupials, the fruit fly, and *C. elegans*. In all of these cases, chromatin-modifying proteins and lncRNAs are the major players in molecular mechanisms. Recent advances in genomics and molecular biology have revealed the structure of the sex chromosomes during the epigenetic changes, and have identified new genetic elements involved in the complex regulatory pathways that control gene expression levels of the X chromosome. Although many advances have been made in the field over the past few years, numerous aspects of the dosage compensation mechanisms and their relationship with human health and development

remain unanswered. Also, probably many other species of animals with sex chromosomes have evolved interesting but still unknown dosage mechanisms to solve potential expression unbalances between males and females. Sex chromosomes are influenced by selection forces that lead to the evolution of sex-specific genes, the fixation of sex-antagonistic genes, and the emergence of sex-beneficial genes. The study of these genes has been important in understanding the regulation of chromosomal-wide gene expression levels and their role in health, sexual selection, and the evolution of the species. Finally, analyses of the structure of the X chromosome have opened a wide array of opportunities for the study of chromatin dynamics and epigenetic markers that will be the focus of future work for many decades to come.

References

1. Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* **14**: 113-124.
2. Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman TL, Hahn MW, Kitano J, Mayrose I, Ming R et al. 2014. Sex determination: why so many ways of doing it? *PLoS Biol* **12**: e1001899.
3. Balaton BP, Cotton AM, Brown CJ. 2015. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol Sex Differ* **6**: 35.
4. Balaton BP, Dixon-McDougall T, Peeters SB, Brown CJ. 2018. The eXceptional nature of the X chromosome. *Hum Mol Genet* **27**: R242-R249.
5. Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, Koutseva N, Zaghlul S, Graves T, Rock S et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**: 494-499.
6. Berletch JB, Yang F, Xu J, Carrel L, Disteche CM. 2011. Genes that escape from X inactivation. *Hum Genet* **130**: 237-245.
7. Beutler E. 1998. Susumu Ohno: the father of X-inactivation. *Cytogenet Cell Genet* **80**: 16-17.
8. Brockdorff N, Turner BM. 2015. Dosage compensation in mammals. *Cold Spring Harb Perspect Biol* **7**: a019406.
9. Chanda K, Mukhopadhyay D. 2020. LncRNA Xist, X-chromosome Instability and Alzheimer's Disease. *Curr Alzheimer Res* **17**: 499-507.

10. Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* **355**: 1563-1572.
11. Conrad T, Akhtar A. 2012. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet* **13**: 123-134.
12. Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grutzner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**: 488-493.
13. Ercan S. 2015. Mechanisms of x chromosome dosage compensation. *J Genomics* **3**: 1-19.
14. Fang H, Disteche CM, Berletch JB. 2019. X Inactivation and Escape: Epigenetic and Structural Features. *Front Cell Dev Biol* **7**: 219.
15. Furman BLS, Metzger DCH, Darolti I, Wright AE, Sandkam BA, Almeida P, Shu JJ, Mank JE. 2020. Sex Chromosome Evolution: So Many Exceptions to the Rules. *Genome Biol Evol* **12**: 750-763.
16. Gelbart ME, Larschan E, Peng S, Park PJ, Kuroda MI. 2009. Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. *Nat Struct Mol Biol* **16**: 825-832.
17. Jordan W, 3rd, Larschan E. 2021. The zinc finger protein CLAMP promotes long-range chromatin interactions that mediate dosage compensation of the Drosophila male X-chromosome. *Epigenetics Chromatin* **14**: 29.
18. Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, Schutz F, Daish T, Grutzner F, Kaessmann H. 2012. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol* **10**: e1001328.
19. Lenormand T, Roze D. 2022. Y recombination arrest and degeneration in the absence of sexual dimorphism. *Science* **375**: 663-666.
20. Li Y, Zheng M, Lau YF. 2014. The sex-determining factors SRY and SOX9 regulate similar target genes and promote testis cord formation during testicular differentiation. *Cell Rep* **8**: 723-733.
21. Makki R, Meller VH. 2021. When Down Is Up: Heterochromatin, Nuclear Organization and X Upregulation. *Cells* **10**.
22. Mank JE. 2013. Sex chromosome dosage compensation: definitely not for everyone. *Trends Genet* **29**: 677-683.

23. Ohno S. 1967. *Sex chromosomes and sex linked genes*. Springer Berlin, Heidelberg.
24. Patrat C, Ouimette JF, Rougeulle C. 2020. X chromosome inactivation in human development. *Development* **147**.
25. Payer B, Lee JT. 2008. X chromosome dosage compensation: how mammals keep the balance. *Annu Rev Genet* **42**: 733-772.
26. Pereira G, Doria S. 2021. X-chromosome inactivation: implications in human disease. *J Genet* **100**.
27. Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B. 2002. Xist RNA and the mechanism of X chromosome inactivation. *Annu Rev Genet* **36**: 233-278.
28. Quinn JJ, Chang HY. 2015. In situ dissection of RNA functional subunits by domain-specific chromatin isolation by RNA purification (dChIRP). *Methods Mol Biol* **1262**: 199-213.
29. Schauer T, Ghavi-Helm Y, Sexton T, Albig C, Regnard C, Cavalli G, Furlong EE, Becker PB. 2017. Chromosome topology guides the Drosophila Dosage Compensation Complex for target gene activation. *EMBO Rep* doi:10.15252/embr.201744292.
30. Sekido R, Lovell-Badge R. 2008. Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer. *Nature* **453**: 930-934.
31. Sinclair AH, Berta P, Palmer MS, Hawkins JR, Griffiths BL, Smith MJ, Foster JW, Frischauf AM, Lovell-Badge R, Goodfellow PN. 1990. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346**: 240-244.
32. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**: 825-837.
33. Sprague D, Waters SA, Kirk JM, Wang JR, Samollow PB, Waters PD, Calabrese JM. 2019. Nonlinear sequence similarity between the Xist and Rsx long noncoding RNAs suggests shared functions of tandem repeat domains. *RNA* **25**: 1004-1019.
34. Strome S, Kelly WG, Ercan S, Lieb JD. 2014. Regulation of the X chromosomes in Caenorhabditis elegans. *Cold Spring Harb Perspect Biol* **6**.

35. Tukiainen T, Villani AC, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L, Fleharty M, Kirby A et al. 2017. Landscape of X chromosome inactivation across human tissues. *Nature* **550**: 244-248.
36. Vockel M, Riera-Escamilla A, Tuttelmann F, Krausz C. 2021. The X chromosome and male infertility. *Hum Genet* **140**: 203-215.
37. Zhou W, Lin SH, Khan SM, Yeager M, Chanock SJ, Machiela MJ. 2021. Detectable chromosome X mosaicism in males is rarely tolerated in peripheral leukocytes. *Sci Rep* **11**: 1193.

CAPÍTULO 4

El artículo “Regulación de los cromosomas sexuales por RNAs largos no codificantes” explica el papel importante que tienen los RNAlnc en la regulación de los niveles de expresión que presentan los cromosomas sexuales, específicamente en el RNAlnc *XIST* y *RSX* en la inactivación del cromosoma X (en hembras, mamíferos placentarios y marsupiales respectivamente), en el lncRNA *ROX2* en la hiper-acetilación del cromosoma X (*D. melanogaster*) y por último un cuarto RNAlnc identificado en esta tesis de doctorado el cual lo denominamos como “Male-specific long non-coding RNA AmplifYing the Expression of the X” (*MAYEX*)”. El cual está fuertemente asociado con la maquinaria de acetilación logrando así la sobre-regulación del cromosoma X. Este trabajo fue publicado en la Gaceta Biomédicas de la UNAM en agosto de 2023, Pág. 8-9. Número 8, ISSN 1607-6788

Regulación de los cromosomas sexuales por RNAs largos no codificantes

Mariela Tenorio y Diego Cortez

Centro de Ciencias Genómicas de la UNAM

Los cromosomas sexuales son los responsables de la determinación del sexo en múltiples especies, incluyendo al humano. A raíz de la evolución de esta función, estos cromosomas han tenido cambios masivos a tal punto de apagar o activar casi por completo su expresión según el tipo de células en las que están presentes (Mank, 2013). Este fenómeno es entendible bajo la luz de la teoría del origen de los cromosomas sexuales propuesta por el genetista Hermann Müller (Bachtrog, 2013). La teoría de Müller propone que los cromosomas sexuales se originan a partir de un par de cromosomas autosomales. Por ejemplo, en los mamíferos, hace aproximadamente 180 millones de años, un cromosoma autosomal adquirió un gen capaz de activar la cascada de señalización que desarrolla el testículo (Bachtrog, 2013; Marais y Galtier, 2003). Este gen lo conocemos ahora como *SRY*. La aparición de este gen produjo que un autosoma se transformara es un cromosoma sexual específico de machos, el cromosoma Y; a la pareja de este cromosoma la llamamos X. Rápidamente la región alrededor del gen *SRY* fue aislada de la recombinación por una gran inversión cromosomal, lo que ocasionó que los cromosomas X y Y dejaran de recombinar y *SRY* se fijara en los machos de la población. La falta de recombinación del Y condujo a la acumulación de mutaciones, de secuencias repetidas y, posteriormente, a la pérdida masiva de material genético (Bachtrog, 2013; Gatler, 2014). Así, los machos se quedaron con un cromosoma X y un cromosoma Y degenerado, mientras que las hembras conservaron dos cromosomas X. La expresión de aquellos genes que anteriormente conservaba el Y se perdió, provocando un desbalance de expresión génica entre machos y hembras. Casi de manera simultánea a la degeneración del cromosoma Y, evolucionó un mecanismo que pudiera restablecer el balance de expresión génica entre los dos sexos (Ercan, 2015).

Los mecanismos de compensación de dosis génica ayudan a igualar la expresión de los cromosomas X en machos y hembras (Gatler, 2014). En estos procesos hay cambios en las marcas epigenéticas del cromosoma X que regulan sus niveles de

expresión. Los mecanismos de compensación de dosis génica son variados y pueden ocurrir tanto en machos como en hembras. En mamíferos, por ejemplo, las hembras apagan casi por completo uno de sus cromosomas X a través de la metilación de las histonas, específicamente tri-metilan la lisina 27 de la histona 3 (H3K27me3) (Payer y Lee, 2008). En cambio, en *Drosophila*, el cromosoma X de los machos aumenta sus niveles de expresión porque se hiper-acetila, específicamente se acetila la lisina 16 de la histona 4 (H4K16ac) (Conrad y Akhtar, 2012).

En los mecanismos que se conocen en mamíferos, marsupiales y en la mosca de la fruta, el sistema es orquestado por RNAs largos no codificantes (RNAlnc), que son transcritos que no se traducen a proteínas y miden más de 200 nucleótidos (Wang y Chang 2011). Los RNAlnc son capaces de responder a diversos estímulos, reclutan enzimas modificadoras de la cromatina hacia genes específicos y funcionan como andamios al formar complejos ribonucleoproteicos para actuar sobre las histonas de los cromosomas X (Wang y Chang 2011). Los RNAlnc que desencadenan la inactivación del cromosoma X de las hembras son *XIST* (mamíferos placentarios) y *RSX* (en marsupiales). Por otro lado, el RNAlnc que participa en la hiper-acetilación del cromosoma X en machos de la mosca de la fruta *ROX2* (Quinn y Chang, 2015). Con nuestro trabajo hemos podido añadir un RNAlnc a la regulación de los cromosomas X. Hablamos de la regulación del cromosoma X de la lagartija verde, *Anolis carolinensis*.

Los cromosomas sexuales XY de *A. carolinensis* aparecieron aproximadamente hace 160 millones de años (Marin *et al*., 2017). Al igual que en mamíferos, el cromosoma Y de la lagartija verde se degeneró a tal grado que sólo conserva 7 genes de los 350 que tenía cuando era un autosoma (Marin *et al*., 2017). De forma similar a lo que ocurre en *Drosophila*, el cromosoma X en los machos de *A. carolinensis* presenta una hiper-acetilación (H4K16ac) del cromosoma X en machos; primer caso de regulación del X de los machos en vertebrados.

Durante el trabajo de tesis de doctorado de Mariela Tenorio, identificamos un RNAlnc que sólo está activo en el cromosoma X de los machos al que denominamos como *MAYEX* por “Male-specific long non-coding RNA AmplifYing the Expression of the X”. *MAYEX* está fuertemente asociado con la maquinaria de acetilación y logra crear un dominio que permite que diferentes regiones del cromosoma X se plieguen hacia el

locus de MAYEX y se sobre-acetilen, logrando así la sobre-regulación del cromosoma X.

Aún quedan muchas preguntas sin respuesta, pues aún no conocemos las proteínas con las que interactúa *MAYEX*, ni los factores específicos de machos que pudieran estar regulando la expresión de *MAYEX* para que se active únicamente en machos.

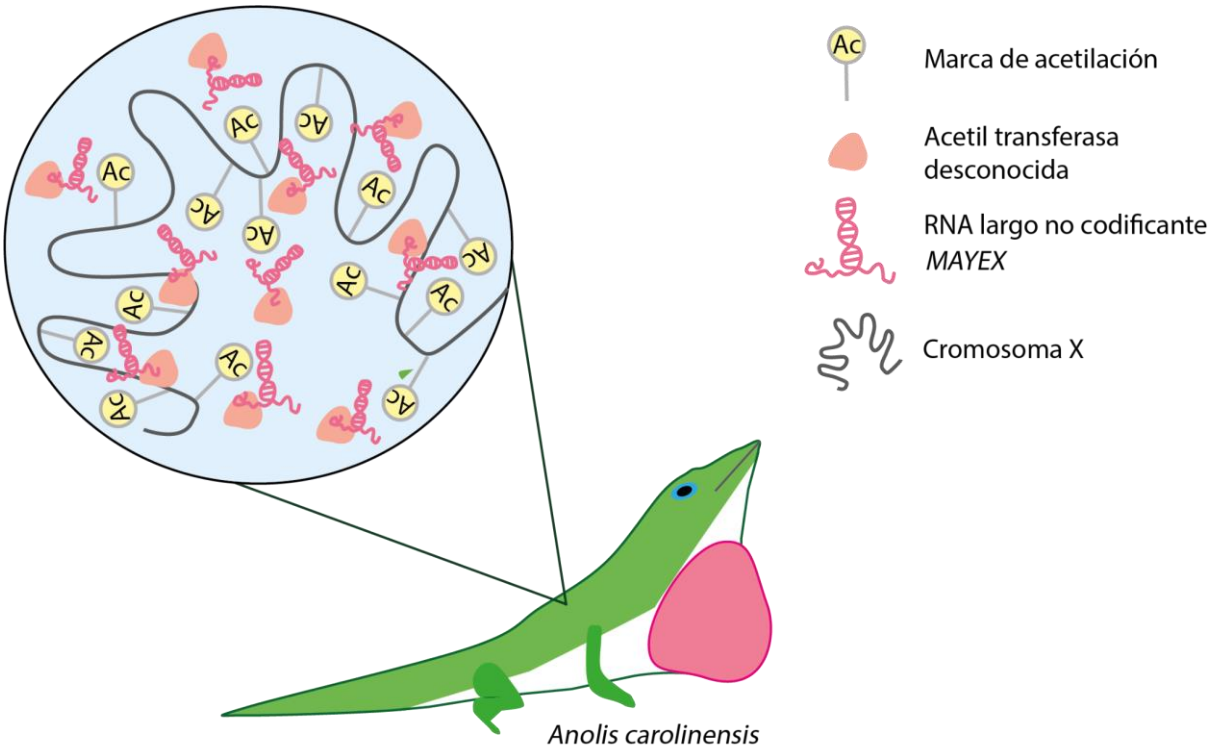


Figura 1. Modelo propuesto para la regulación del cromosoma X por el RNAlnc *MAYEX* en *A. carolinensis.*

DISCUSIÓN

En las especies que presentan un sistema de determinación sexual XY, las hembras tienen un par de cromosomas X homólogos, mientras que los machos presentan cromosomas heterólogos, formados por un cromosoma X y un cromosoma Y. Cuando Susumu Ohno observó este patrón, se preguntó si a las hembras les sobra material genético o a los machos les falta. Ohno propuso que uno de los cromosomas X de las hembras en los mamíferos placentarios pudiera estar silenciado (Ohno 1967; Beutler 1998).

Esto ocasionó grandes rearreglos cromosómicos complejos a lo largo del cromosoma Y. Fascinantemente, estos rearreglos han sido observados tanto en vertebrados como en invertebrados. En mamíferos placentarios y en marsupiales este problema se resolvió con la inactivación parcial del cromosoma X en hembras y en *D. melanogaster* la sobreexpresión del X en machos ayudó a restablecer el balance (Payer y Lee, 2008; Conrad y Akhtar, 2012).

Hasta fechas recientes, únicamente conocíamos una sola manera en la que ocurre la compensación de dosis en los vertebrados, la inactivación del cromosoma X. Sin embargo, gracias a los hallazgos de este trabajo, pudimos describir una especie vertebrada en la cual la compensación que ocurre con un mecanismo similar al de un invertebrado.

A. carolinensis es un reptil que posee el sistema de determinación sexual por los cromosomas sexuales XY y el cromosoma Y se encuentra altamente degenerado (Marin, et al., 2017). El simple hecho de conocer que en la lagartija verde el problema de compensación de dosis de los cromosomas sexuales se equilibra sobreacetilando el cromosoma X de los machos, dejó varias incógnitas y una de las principales fue conocer ¿cuáles son los factores que promueven la sobre-expresión del cromosoma X en los machos de *A. carolinensis* y que median el balance de los niveles de expresión entre ambos sexos?

En los modelos mencionados anteriormente, la compensación de dosis de los cromosomas sexuales es llevada a cabo por RNAs.

Los lncRNA son moléculas que participan en varios procesos, como la síntesis de proteínas, la maduración del RNA, el transporte de proteínas y la activación o el silenciamiento de genes transcripcionales, a través de la regulación de la estructura de la cromatina (Marchese et al., 2017). A pesar de que los lncRNA son capaces de interactuar con otras moléculas dentro de la célula, en este estudio solo nos enfocamos en los que interactúan con el DNA, específicamente con el cromosoma X. Los lncRNA pueden actuar tanto en *cis* como en *trans*. En *cis*, los transcritos regulan la expresión genética de los genes que están cercanos. En *trans*, regulan a los genes ubicados en otros cromosomas. En nuestro análisis utilizando la técnica ChAR-seq, la cual mapea los contactos de RNA contra DNA, pudimos identificar 2,283 lncRNA de los 3,176 que se conocen actualmente. Combinando los datos anteriores con datos de ChIP-seq que utilizaron la marca de acetilación H4K16ac, encontramos tres lncRNAs con posibles funciones reguladoras. Uno de ellos fue ENSACAG00000036367 el cual puede estar involucrado en la activación genética, mientras que ENSACAG00000045045 y ENSACAG00000044053 podrían desempeñar un papel en el silenciamiento de genes. También, encontramos que la mayoría de los transcritos actúan en *cis*-proximal y muy pocos actúan en *trans*.

El locus de los lncRNA que orquestan los rearreglos cromosomales complejos a lo largo del cromosoma X se encuentra en este mismo, lo que significa que actúan en *cis*. En los mamíferos placentarios se encuentra el lncRNA *XIST*, en marsupiales se encuentra *RSX* y en *Drosophila melanogaster* encontramos a *ROX2* (Conrad and Akhtar 2012; Lu et al., 2017).

En este trabajo describimos un cuarto lncRNA el cual lo denominamos como *MAYEX* por Male-specific long non-coding RNA AmplifYing the Expression of the X. Este se localiza en el cromosoma X y es el primer lncRNA descrito que regula por completo el cromosoma X en un reptil. Las similitudes encontradas en los mamíferos, la mosca de la fruta y ahora en la lagartija verde, indica que los lncRNA se comparten en especies totalmente distantes durante la evolución, participando los mecanismos de regulación en *cis* a lo largo del cromosoma X, controlando los niveles de expresión del cromosoma completo y restableciendo el equilibrio de la expresión entre machos y hembras.

El sistema de *MAYEX* opera en todos los tejidos durante el desarrollo embrionario y en los adultos, lo que significa que es necesario mantener activamente la hiperacetilación del cromosoma X. *MAYEX* es un antiguo transcrito con expresión en los machos desde hace más de 89 millones de años. Este lncRNA comparte la región promotora con el gen *MORC2*. Encontramos que tanto en machos como en hembras la expresión de *MORC2* es activa a pesar de compartir la región promotora con *MAYEX*. Esto indica que *MAYEX* podría tener una regulación compleja.

El análisis de datos de ChIP-seq en *A. carolinensis* reveló que el locus de *MAYEX* presenta un alto enriquecimiento de la marca de acetilación H4K16ac por lo que creemos que existe una interacción directa entre *MAYEX* y el complejo de acetilación. Nuestra hipótesis es que un complejo proteico envuelve todo el cromosoma X con el locus de *MAYEX* y, dada la proximidad de la maquinaria de acetilación a este locus, podría provocar aumentos significativos en los niveles de H4K16ac en el cromosoma X. Esto explicaría un nuevo tipo de compensación en los vertebrados. También recordemos que el cromosoma X de *A. carolinensis* es 16 veces más pequeño que el cromosoma X de los mamíferos placentarios y 2.4 veces más pequeño que el de *D. melanogaster*. Esto sugiere la existencia de un andamiaje de todo el cromosoma X cerca del locus de *MAYEX*. Otro tema interesante es que el cromosoma X está enriquecido de repeticiones (TTA)5, lo que podría explicar el por qué el sistema está activo únicamente en el cromosoma X. Esta repetición evoca el motif (GA)4 el cual es asociado con los sitios de unión al DNA de *ROX2*/MSL3-TAP en *Drosophila* (Simon, et al., 2011).

También localizamos un lncRNA propio de las hembras, al cual lo denominamos como Female Expressed Region on the X. La expresión de *FERX* únicamente se encuentra activa en el desarrollo embrionario, en los tejidos de adultos se encuentra apagado. Aún no nos queda claro por qué la expresión de *FERX* no es requerida en los adultos. Por esto, nuestro modelo actual propone que *MAYEX* Y *FERX* son completamente excluyentes, ya que, la expresión de *MAYEX* domina sobre la de *FERX* en machos y viceversa en las hembras. También creemos que *MAYEX* y *FERX* pueden ser isoformas diferentes del mismo gen, porque la región intergénica ortóloga del pollo para *MAYEX* y *FERX* contiene un solo lncRNA (ENSGALG00010026900).

Como habíamos mencionado anteriormente, los lncRNA lideran los rearreglos cromosómicos complejos del cromosoma X. Estos transcritos ayudan a crear estructuras totalmente diferentes. El cromosoma X de los mamíferos placentarios se compacta en dos grandes TADs, mientras que el cromosoma de *Drosophila* posee más TADs pequeños que son más accesibles a la maquinaria de acetilación (Schauer, et al., 2017; Fang et al., 2019). En *A. carolinensis* detectamos una gran frecuencia de TADs en machos comparados con las hembras. Finalmente, los análisis de la estructura del cromosoma X han abierto una amplia gama de oportunidades para el estudio de la dinámica de la cromatina y los marcadores epigenéticos.

PERSPECTIVAS

Este trabajo presenta un nuevo lncRNA que participa en la compensación de dosis de los cromosomas X en lagartos. Estudios futuros podrían centrarse en identificar las proteínas hipotéticas asociadas con *MAYEX* y su región vecina rica en repeticiones, lo que podría revelar los mecanismos moleculares que unen regiones del cromosoma X con el locus *MAYEX*, y la dinámica de este proceso durante el desarrollo.

También encontramos varios desafíos asociados con la falta de anotaciones para muchos elementos no codificantes y la necesidad de restringir nuestros análisis a los cromosomas 1 a 6 debido al ensamblaje incompleto del genoma.

Para dilucidar aún más la caracterización funcional de los lncRNA anotados, serían necesarios datos adicionales que abarquen varios tejidos y etapas de desarrollo, junto con otras modificaciones epigenéticas.

Además, nos centramos únicamente en los lncRNA que interactúan con el ADN, mientras que numerosos lncRNA pueden interactuar con otras moléculas dentro de la célula. Por lo tanto, la investigación de los lncRNA asociados con el proteoma requeriría un protocolo de proteína-ARN integral.

BIBLIOGRAFÍA

Alföldi, J., Di Palma, F., Toh, L., et al., (2011) The genome of the green anole lizards and a comparative analysis with birds and mammals. Nature. Research Letter.

Álvarez-Romero, J., Medellín, R. A., Gómez de Silva, H., Oliveras de Ita, A. (2005) Anolis carolinensis. Vertebrados superiores exóticos en México: diversidad, distribución y efectos potenciales. Instituto de Ecología, Universidad Nacional Autónoma de México. Bases de datos SNIB-CONABIO. Proyecto U020. México. D.F.

Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nature Reviews.

Barr, M.L. and Bertram, E.G. (1949) A morphological distinction between neurones of the male and female, and the behavior of the nucleolar satellite during accelerated nucleoprotein synthesis. Nature, 163, 676–677

Bell, J. C., Jukam, D., Teran, N. A., Straight, A. F., et al. (2018) Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts

Beutler E. (1998) Susumu Ohno: The father of X-inactivation. Cytogenet Cell Genet 80:16-17

Brockdorff, N. (2017) Polycomb complexes in X chromosome inactivation. Current Biology Review. Cell Press

Brown C. J., Carrel L., et al (1997) Expression of genes from the human active and inactive X chromosomes. Am J Hum Genet 60:1333-1343

Cancino-Bello, A., Oktaba K. (2019) Análisis de la unión del cofactor APBB1 a genes hiperexpresados en machos de Anolis carolinensis. (Tesis de pregrado). Universidad Autónoma Metropolitana, Unidad Cuajimalpa

Corad, T., Akhtar, A. (2012). Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. Nature Reviews Genetics.

Creamer KM, Lawrence JB. (2017) XIST RNA: a window into the broader role of RNA in nuclear chromosome architecture. Philos Trans R Soc Lond B Biol Sci.

Dekker J, Rippe K, Dekker M, Kleckner N. (2002) Capturing chromosome conformation. Science

Derrien, T., Johnson, R., Guigo, R., et al. (2012) The GENCODE v7 catalog human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Cold Spring Harbor Laboratory Press.

Eckalbar, W., Hutchins, E. D., Markov, G. J., et al. (2013) Genome reannotation of the lizard Anolis carolinensis based on 14 adult and embryonic deep transcriptomes. BMC Genomics

Engreitz, J.M, Pandya-Jones, A., McDonel, P., Shiskin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E. S., Plath, K. Gluttman, M. (2013). The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the Chromosome. Research Article.

Ercan S. (2015). Mechanisms of x chromosome dosage compensation. J Genomics.

Fang H., Disteche C. M., Berletch, J. B. (2019) X inactivation and escape: epigenetic and structural. Features. Front. Cell Dev.

Fatica, A., Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. Nature Reviews.

Finn E. H., Misteli, T. (2019) Molecular basis and biological function of variability in spatial genome organization. Molecular Biology. Science

Fiskus, W., Pranpat, M., Balasis, M., Herger, B., Rao, R., Chinnaiyan, A., ... & Bhalla, K. (2006). Histone deacetylase inhibitors deplete enhancer of zeste 2 and associated polycomb repressive complex 2 proteins in human acute leukemia cells. Molecular cancer therapeutics, 5(12), 3096-3104.

Gatler, s. m. (2014) A brief history of dosage compensation. Journal of Genetics. Vol 93 No. 2

Gibcus J. H., Dekker, J. (2013). The hierarchy of the 3D genome. Molecular Cell Review.

Gorman, G. C. (1973) Cytotaxonomy and Vertebrate Evolution (eds Chriarelli, A. B. & Canpanna, E.) Ch 349-424. Academic

Jukam, D., Limose, C., Smith, O. K., Risca, V. I., Bell, J. C., Straight, A. F. (2019). Chromatin-Associated RNA Sequencing (ChAR-seq). Current Protocols in Molecular Biology

Kaufmann, C., & Wutz, A. (2023). IndiSPENsable for X Chromosome Inactivation and Gene Silencing. Epigenomes, 7(4), 28.

Lajoie, B. R., Dekker, J., Kaplan, N. (2015). The Hitchhiker´s guide to Hi-C analysis: practical guidelines. Elsevier Methods.

Lieberman-Aiden, E., Berkum, N. L. V., Dekker J., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science.

Long, Y., Wang, X., Youmans, D. T., Cech, T. R. (2017) How do lncRNA regulate transcription? Gene Expression. Science Advance

Lu, A., Carte, A. C., Chang, H. Y. (2017) Mechanistic insights in X-chromosome inactivation. Philosophical Transactions B. Royal Society Publishing

Marchese, F.P., Raimondi, I., Huarte, M. (2017). The multidimensional mechanisms of long noncoding RNA function. Genome Biology.

Mariais G., Galtier, N. (2003). Sex chromosomes: how X-Y recombination stops. Current Biology. Vol. 13, R641, R643

Marin, R. et al. (2017) Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage. Genome Res 27, 1974-1987.

Mank, J. E. (2013). Sex chromosome dosage compensation: definitely not for everyone. Cell Press

Misteli, T. (2008) Chromosome territories: The arrangement of chromosomes in the nucleus. Nature Education 1(1): 167

Morales, V., Regnard, C., Izzo, A., Vetter, I., Becker, P. B. (2005) The MRG domain mediates the functional integration of MSL3 into the dosage compensation complex

Muller H. J. Altenburg E. (1919) The rate of change of hereditary factors in Drosophila. Proc Soc. Exper. Biol and Med. 17:10-14

Muller H. J. (1932) Further studies on the nature and causes of gene mutations. International Congress of Genetics.

Ohno S. (1967) Ohno S: Monographs on Endocrinology. Sex chromosomes and sex-linked genes, 1st edition Heidelber, Springer-Verlag

Payer B, Lee JT. (2008) X chromosome dosage compensation: how mammals keep the balance. Annu Rev Genet.

Quinn J.J., Chang, H.Y. (2015). Unique features of long non-coding RNA biogenesis and function. Nature Reviews.

Raha, D., Hong, M., Snyder, M. (2010) ChIP-seq: A method for global identification of regulatory elements in the genome. Curr. Protoc. Mol. Biol.

Rao M. R. S. (2017) Long noncoding RNA biology. Advances in Experimental Medicine and Biology. Springer

Rice J. C., Allis C. D. (2001) Histone methylation versus histone acetylation: new insights into epigenetics regulation. Elsevier Science

Richardson, S. S. (2010) Sexes, species and genomes: why males and females are not like humans and chimpanzees. Biol Philos.

Rowley M. J., Corces V. G. (2018) Organizational principles of 3D genome architecture. Opinion Nature Reviews Genetics.

Sado, T. (2017) What makes the maternal X chromosome resistant to undergoing imprinted X inactivation? Philosophical Transactions B. Royal Society Publising

Schauer, T., Ghavi-Helm, Y., Sexton, T., Albig, C., Regnard, C., Cavalli, G., Furlong, E. EM., Becker, P. B. (2017) Chromosome topology guides the Drosophila dosage compensation complex for target gene activation. EMBO Reports

Shapiro LJ, Mohandas T et al (1979) Non-inactivation of an X-chromosome locus in man. Science

Simon et al., M. D. (2011). The genomic binding sites of a noncoding RNA. Proc Natl Acad Sci U S A 108, 20497-20502.

Smith, R. (2001) Anolis carolinensis, Green Anole. Animal Diversity

Snell D. M. y Turner J. M. A. (2018) Sex chromosomes effects on male-female differences in mammals. Current Biology Review. Cell Press

Straub, T., Gilfillan, G. D., Maier, V. K., Becker, P. B. (2005) The Drosophila MSL complex activates the transcription of target genes. Research Communication. Genes y Development. Cold Spring Harbor Laboratory Press

Wang K. C., Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. Cell Press. DOI 10.1016/j.molcel.2011.08.018

ANEXOS

A continuación, se muestran los tres artículos publicados durante mi tesis de doctorado.

BMC Genomics

RESEARCH

Open Access

Genome-wide analysis of RNA-chromatin interactions in lizards as a mean for functional lncRNA identification

Mariela Tenorio[1], Joanna Serwatowska[2], Selene L. Fernandez-Valverde[2,3], Katarzyna Oktaba[2] and Diego Cortez[1*]

Abstract

Background Long non-coding RNAs (lncRNAs) are defined as transcribed molecules longer than 200 nucleotides with little to no protein-coding potential. LncRNAs can regulate gene expression of nearby genes (*cis*-acting) or genes located on other chromosomes (*trans*-acting). Several methodologies have been developed to capture lncRNAs associated with chromatin at a genome-wide level. Analysis of RNA-DNA contacts can be combined with epigenetic and RNA-seq data to define potential lncRNAs involved in the regulation of gene expression.

Results We performed Chromatin Associated RNA sequencing (ChAR-seq) in *Anolis carolinensis* to obtain the genome-wide map of the associations that RNA molecules have with chromatin. We analyzed the frequency of DNA contacts for different classes of RNAs and were able to define *cis*- and *trans*-acting lncRNAs. We integrated the ChAR-seq map of RNA-DNA contacts with epigenetic data for the acetylation of lysine 16 on histone H4 (H4K16ac), a mark connected to actively transcribed chromatin in lizards. We successfully identified three *trans*-acting lncRNAs significantly associated with the H4K16ac signal, which are likely involved in the regulation of gene expression in *A. carolinensis*.

Conclusions We show that the ChAR-seq method is a powerful tool to explore the RNA-DNA map of interactions. Moreover, in combination with epigenetic data, ChAR-seq can be applied in non-model species to establish potential roles for predicted lncRNAs that lack functional annotations.

Keywords Chromatin, Long non-coding RNA, *Anolis carolinensis*, Chromatin-associated RNA sequencing

Background

The eukaryotic cell is home to a plethora of non-coding RNAs, among which ribosomal RNAs, small nuclear RNAs, and small nucleolar RNAs are the most abundant. Ribosomal RNAs play a crucial role in translating messenger RNAs, while small nuclear RNAs are essential for gene splicing, and small nucleolar RNAs guide chemical modifications of other RNA molecules. The most enigmatic group of non-coding RNAs is long-non coding RNAs (lncRNAs) [1], encompassing RNA molecules longer than 200 nucleotides that lack protein-coding potential. Transcriptomic studies have identified thousands of

*Correspondence:
Diego Cortez
dcortez@ccg.unam.mx
[1]Center for Genome Sciences, National Autonomous University of Mexico (UNAM), Cuernavaca, Mexico
[2]Center for Research and Advanced Studies (Cinvestav), Irapuato, Mexico
[3]Present address: School of Biotechnology and Biomolecular Sciences and the RNA Institute, The University of New South Wales, Sydney, NSW 2052, Australia

lncRNAs that may possess functional roles in humans and mice [2–10]. Large-scale screenings have associated many of these lncRNAs with regulatory functions [11].

LncRNAs can be categorized into two groups based on their regulatory activity. *Cis*-acting lncRNAs regulate gene expression on the same chromosome from which they are transcribed, whereas *trans*-acting lncRNAs regulate gene transcription on different chromosomes. Some extensively studied *cis*-acting lncRNAs in vertebrates include *XIST* [12–14], *RSX* [15], and *ROX2* [16], which regulate the expression levels of entire X chromosomes in placental mammals, marsupials, and the fruit fly, respectively. Other characterized lncRNAs can regulate genetic imprinting [17], recruit protein complexes that modify chromatin [18], or influence the expression of remote genes [19]. Although a few *trans*-acting lncRNAs have been experimentally studied, their number remains limited. Notably, *HOTAIR* [20, 21], a lncRNA known to silence the *HOXD* gene by recruiting the Polycomb Repressive Complex 2 has faced challenges regarding its *trans*-activity following a recent study that analyzed *HOTAIR* knockout mice [22]. Another example is FIRRE, a trans-acting lncRNA involved in hematopoiesis [23].

For more than two decades, hybridization capture methods have been the standard technique for identifying the DNA and proteins associated with specific lncRNA [24]. These methods, known as one-to-all approaches, employ biotinylated DNA probes to selectively purify a lncRNA that has been cross-linked to their adjacent DNA and binding proteins. The most renowned techniques include Chromatin Isolation by RNA Purification (ChIRP) [25], Capture Hybridization Analysis of RNA Targets (CHART) [26], and RNA antisense purification (RAP) [27].

Recently, four all-to-all approaches have emerged to capture all possible interactions between RNA molecules and chromatin. These methodologies are designed to provide comprehensive insights into RNA-genome interactions. The four methods are MARGI (Mapping RNA–Genome Interactions) [28], ChAR-seq (Chromatin-Associated RNA sequencing) [29, 30], GRID-seq (Global RNA Interaction with DNA sequencing) [31], and RADICL-seq (RNA And DNA Interacting Complexes Ligated and sequenced) [32]. These methodologies involve capturing RNAs in contact with DNA by employing specific short linkers that ligate an RNA fragment to an adjacent DNA fragment. Both MARGI and CHAR-seq enable the sequencing of long RNA-DNA tags. In a successful application of MARGI, researchers used human cells to demonstrate that *XIST* exhibits long-range binding sites along the female X chromosome [28]. Similarly, ChAR-seq was employed in *Drosophila* to unveil the detailed map of RNA-DNA contacts of *ROX2* along the X chromosome of males [30].

While these all-to-all techniques have proven successful in model species such as humans, mice, and *Drosophila*, RNA-DNA contact maps have yet to be explored in other species. In recent years, the number of reptile genomes deposited in public databases has increased by over 600% (from 17 genomes before 2018 to 123 genomes between 2018 and 2023). However, gene annotations in these genomes typically rely on automated modeling of gene predictions based on protein-coding genes from species with curated annotations. In some cases, such as the reference genome of the green anole lizard, *Anolis carolinensis*, transcriptomic data was utilized to enhance the annotations of coding and non-coding genes [33]. The current version of the genome of *A. carolinensis* contains 3,176 lncRNAs (https://www.ensembl.org/Anolis_carolinensis/Info/Annotation), yet most of them lack functional information. In this study, we hypothesized that ChAR-seq-like methods could aid in identifying potential regulatory lncRNAs in genomes where they have been predicted. It should be noted that the ChAR-seq method can provide information about RNA molecules that interact with DNA but is unable to report interactions with other molecules, such as proteins. Therefore, we applied the ChAR-seq method in *A. carolinensis* to investigate the overall map of interactions between RNA molecules and chromatin. We characterized the frequencies of contact for different classes of RNAs and annotate *cis*- and *trans*-acting lncRNAs. By correlating the ChAR-seq results with ChIP-seq data for the acetylation of lysine 16 on histone H4 (H4K16ac) epigenetic mark, we identified three lncRNAs with *trans*-activity that likely play a role in gene expression regulation in *A. carolinensis*.

Results

Variations in the frequency of associations between RNA and chromatin

To investigate the interactions between RNA molecules and chromatin in *A. carolinensis*, we employed the ChAR-seq method on two adult liver samples. By utilizing a specialized short linker, we captured RNA molecules in contact with chromatin and sequenced a total of 1,020,074,230 and 9,96,050,284 reads from the two biological replicates. The paired reads were then mapped to generate a comprehensive genome-wide map of RNA-chromatin interactions.

We analyzed unique interactions for each class of RNA present in the cell. As expected, highly abundant RNA molecules, such as ribosomal RNAs (rRNAs), were overrepresented in our results. Based on the gene annotations available for *A. carolinenesis*, we found that ribosomal RNAs (rRNAs) accounted for 70% of the interactions, whereas long non-coding RNAs (lncRNAs) represented 14%, messenger RNAs (mRNAs) 13%, small nuclear RNAs (snRNAs) 2%, the metazoan signal recognition

particle RNA (metazoan srpRNA) 1%, and small nucleolar RNAs (snoRNAs) 0.04% (Fig. 1a,b). The rRNA genes exhibited interactions with numerous chromatin sites (average contacts per gene=1,132,775), whereas the three annotated metazoan srpRNAs showed an average of 55,000 contacts with chromatin across the genome. The frequency of contacts for other RNA types ranged between 50 and 80,000 (Fig. 1a,b), with consistent patterns across the two replicates (Fig. 1a,b). Given the potential presence of unannotated non-coding genes in the *A. carolinensis* genome, we analyzed our RNA-chromatin interactions dataset using a sliding window of 10Kb. This analysis revealed 9,000 transcribed regions that do not overlap with known coding or non-coding genes, exhibiting between 500 and 87,000 chromatin interactions (Fig. 1a,b; unkRNA).

Subsequently, we determined whether the RNA-chromatin contacts were localized within the same chromosomes (intra-chromosomal) or involved contacts across different chromosomes (inter-chromosomal). To assess this, we introduced a Contact Distribution Index (CDI), which was calculated by dividing the number of contacts on the chromosome with the highest interaction count by the total number of contacts across all chromosomes;

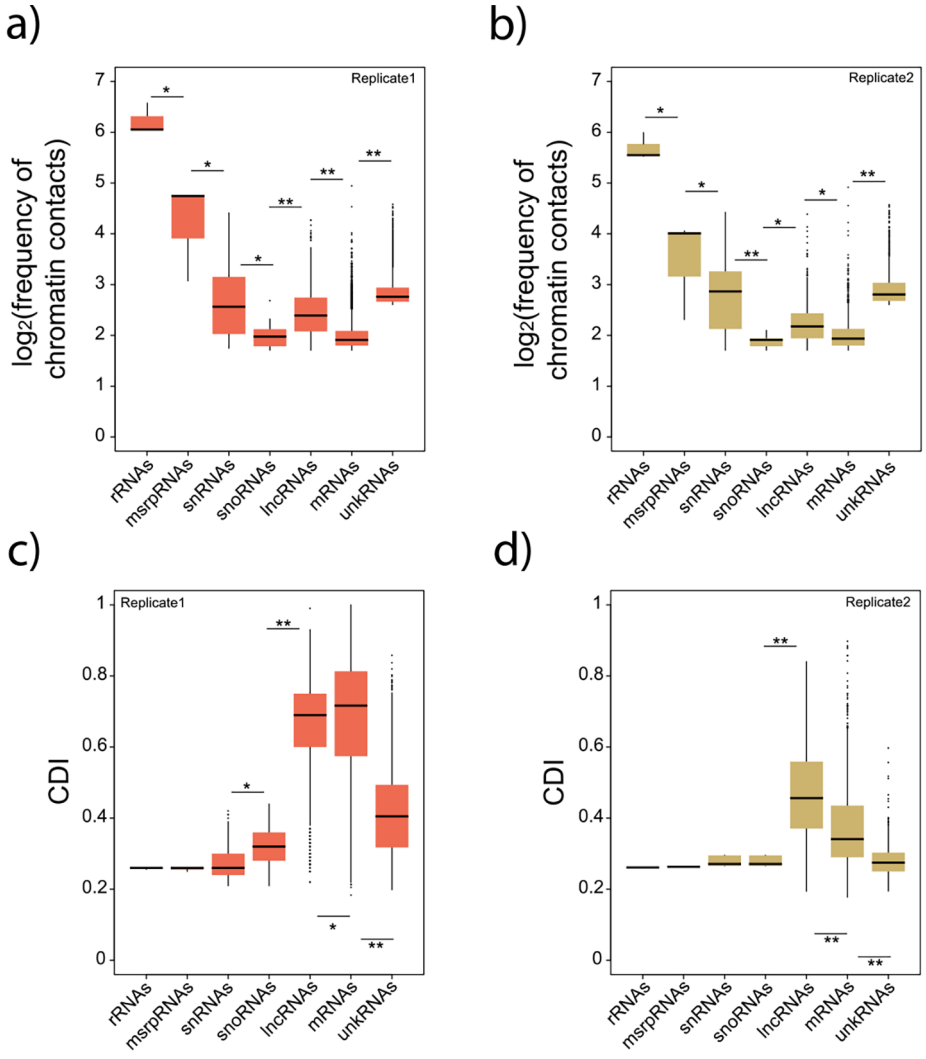


Fig. 1 Chromatin contacts for seven different classes of RNAs. (**a**) Boxplots representing the $\log_2$-transformed ratio of the frequency of chromatin contacts for seven different types of RNA molecules. (**b**) Same as (a) for replicate 2. (**c**) Boxplot representing the values of the Contact Distribution Index (CDI) for seven different classes of RNA molecules. (**d**) Same as (c) for replicate 2. N values for replicate 1 are ribosomal RNAs, 3; metazoan signal recognition particle RNAs, 3; small RNAs, 70; small nucleolar RNAs, 33; long non-coding RNAs, 1188, messenger RNAs, 3724, unannotated RNAs, 3338. N values for replicate 2 are ribosomal RNAs, 3; metazoan signal recognition particle RNAs, 3; small RNAs, 6; small nucleolar RNAs, 5; long non-coding RNAs, 756, messenger RNAs, 769, unannotated RNAs,756. Data for lncRNAs, mRNAs, and unkRNAs are limited to chromosomes 1–6. Error bars, maximum and minimum values, excluding outliers. Significant differences, Mann-Whitney U test; * represents $P<0.01$, ** represents $P<0.001$. *P*-values were corrected using the Benjamini-Hochberg method

CDI values around 0.2–0.3 indicated interactions spread across many chromosomes, while CDI>0.4 indicated a bias toward fewer chromosomes. To ensure the reliability of the interactions in *cis* and *trans*, we focused on genes located on the assembled macro-chromosomes (1 to 6) and discarded the short fragments (scaffolds) in the *A. carolinensis* reference genome.

Our analysis revealed that rRNAs, metazoan srpRNAs, snRNAs, and snoRNAs displayed lower CDI values (Fig. 1c,d), indicating widespread contacts across the entire genome (Fig. 2a,b). This finding aligns with expectations, considering that ribosomal RNAs are the most abundant type of RNA molecules within eukaryotic cell nucleoli. Similarly, snRNAs and snoRNAs, involved in mRNA splicing or RNA chemical modifications, respectively, interact with numerous genomic regions. In contrast, lncRNAs and mRNAs exhibited higher CDI values (Fig. 1c,d), suggesting a predominance of intra-chromosomal contacts. Notably, unkRNAs showed lower CDI values (Fig. 1c,d), implying the absence of lncRNAs or mRNAs within these transcribed regions. In total, our analysis identified 2,282 lncRNAs in replicates 1 and 2 combined, representing 71.8% of the 3,176 lncRNAs annotated in the *A. carolinensis* genome.

C *is*-acting and *trans*-acting lncRNAs

ChAR-seq data provides valuable insights into distinguishing between lncRNAs that interact with loci in close proximity (*cis*-proximal) or distantly (*cis*-distal) on the same chromosome from which they are transcribed, as well as lncRNAs that have interactions with other chromosomes (*trans*-acting). To examine these types of interactions, we analyzed the lncRNAs on chromosomes 1 to 6 and estimated the percentage of *cis*-proximal, *cis*-distant, and *trans*-acting contacts. Notably, lncRNAs exhibited a gradient distribution between *cis*-proximal and *trans*-acting interactions, with a substantial bias towards *cis*-proximal interactions (Fig. 3a). This distinctive sets lncRNAs apart from other classes of RNA molecules (Fig. 3b). Specifically, approximately, 87.5% (n=1040)

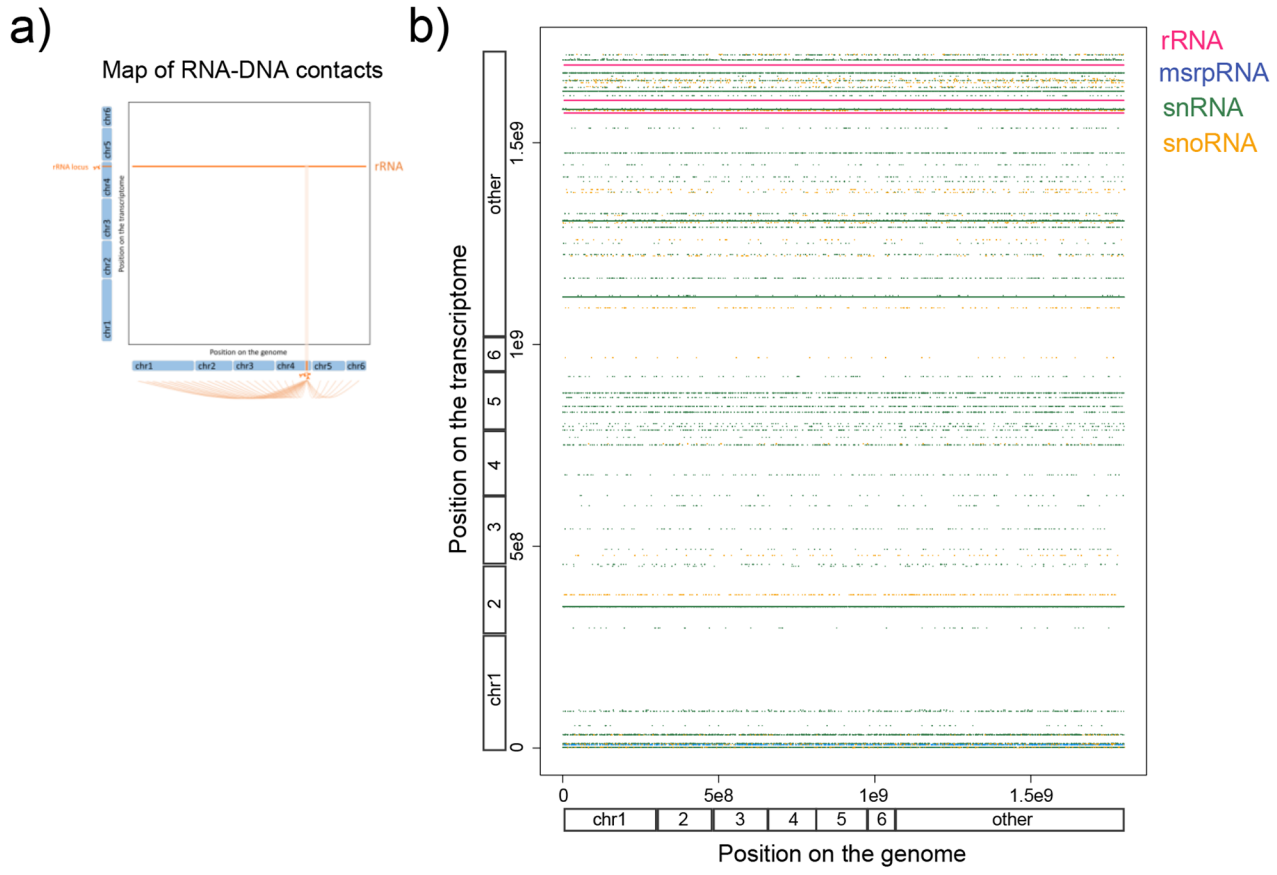


Fig. 2 Contact map of abundant non-coding RNAs between the transcriptome and the genome. (**a**) Illustration to help explain the horizontal lines in panel b; these lines represent the contacts between a single locus on the transcriptome with multiple loci on the genome. (**b**) Dot-plot representation of specific non-coding RNAs loci on the transcriptome (Y-axis) and their multiple genomic contacts (X-axis). The positions on the genome and transcriptome correspond to the concatenated chromosomes 1 to 6 (indicated), followed by the linkage groups and the unassembled scaffolds ordered alphabetically (indicated as other). Ribosomal RNAs are in pink, the metazoan signal recognition particle RNAs in blue, small RNAs in green, and small nucleolar RNAs in orange

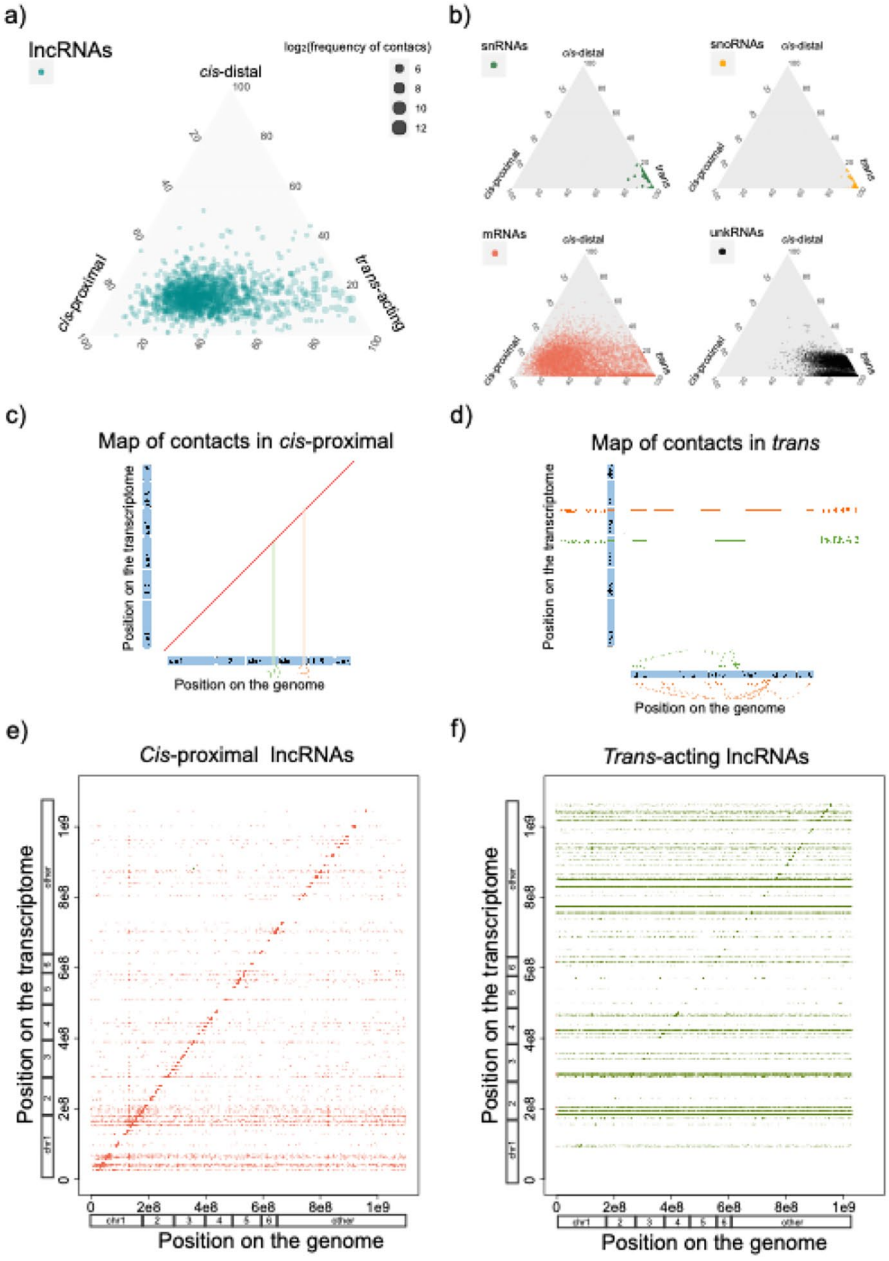


Fig. 3 Contact map of*cis*-acting and*trans*-acting lncRNAs between the transcriptome and the genome. (**a**) Ternary plots representing the type of contacts of lncRNAs; *cis*-proximal (< 10 Kb around the gene loci), *cis*-distal (> 10 Kb on the same chromosome), *trans*-acting (in other chromosomes). Dot sizes are defined by log2 of the frequency of contacts. (**b**) Same as (a) for small RNAs, small nucleolar RNAs, messenger RNAs, and unannotated RNAs; rRNAs and msrpRNAs are not shown since their contacts are > 99.9% *trans*-acting. (**c**) Illustration to help explain the map of RNA-DNA contacts in *cis*-proximal. (**d**) Illustration to help explain the map of RNA-DNA contacts in *trans* (**e**) *Cis*-proximal lncRNAs have most of their RNA-DNA contacts within their loci, which explains the diagonal line on the dot plot. The dot-plot represents specific lncRNAs loci on the transcriptome (Y-axis) and their multiple genomic contacts (X-axis). (**f**) *Trans*-acting lncRNAs have most of their contacts on other chromosomes, which explains the horizontal lines on the dot plot. Same as (c) for *trans*-acting lncRNAs. The positions on the genome and transcriptome correspond to the concatenated chromosomes 1 to 6 (indicated), followed by the linkage groups and the unassembled scaffolds ordered alphabetically (indicated as other)

of lncRNAs displayed over 50% of their contacts in the *cis*-proximal category (Fig. 3c,e; Supplementary Table 1), while 12.5% (n = 148) exhibited over 50% of their interactions in the *trans*-acting category (Fig. 3d,f; Supplementary Table 1).

For 98% of the *cis*-acting lncRNAs, the majority of their chromatin contacts clustered within 20 Kb around the transcription locus of the lncRNA (Fig. 4a,b), aligning with the gene body of the lncRNA. However, upon narrowing our analysis to the top twenty *cis*-acting lncRNAs

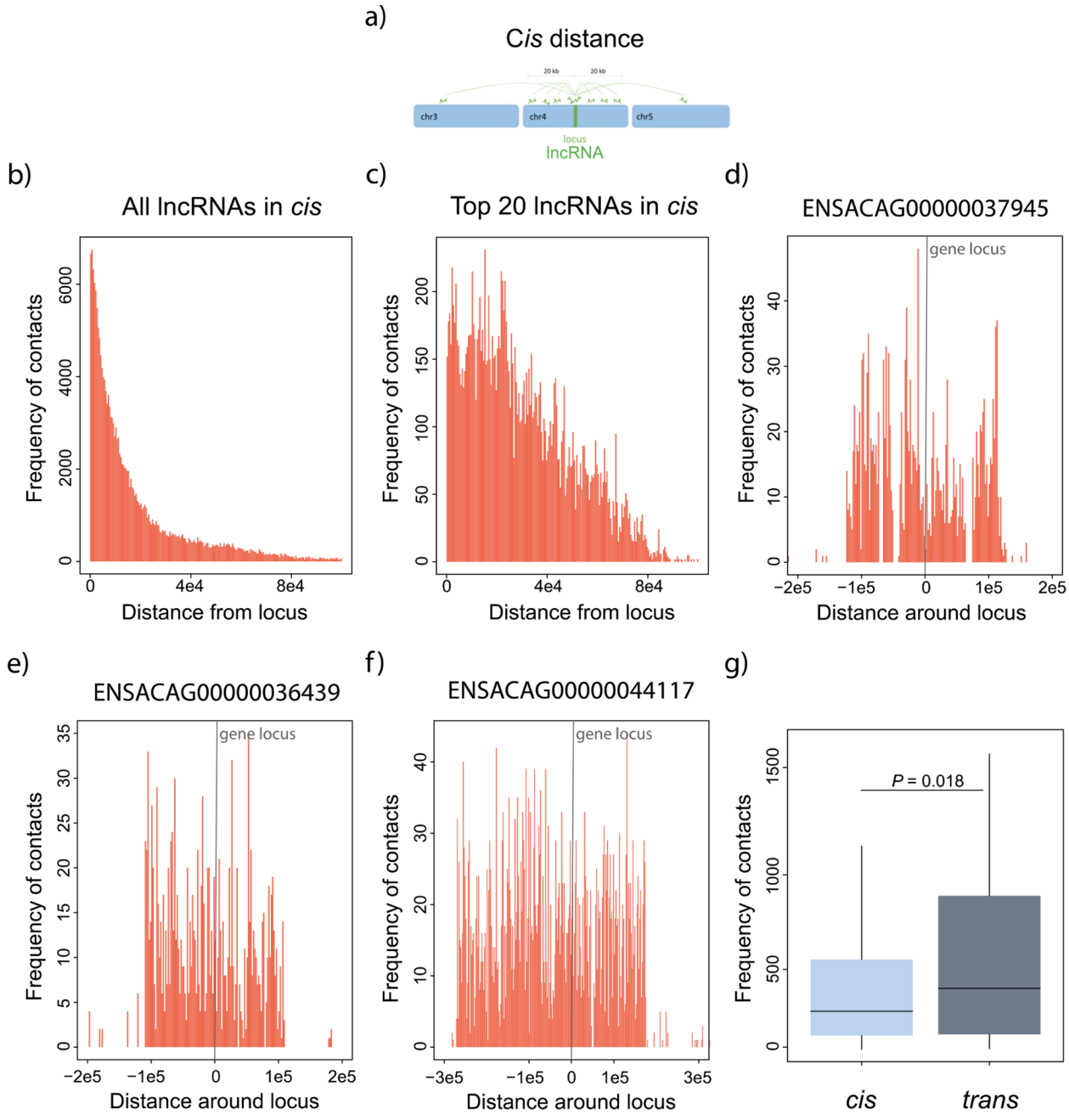


Fig. 4 Range and frequency of RNA-DNA contacts for*cis*-acting lncRNAs. (**a**) Illustration to help explain the RNA-DNA contacts in *cis*. (**b**) Frequency and range of the RNA-DNA contacts (histograms in orange) for all *cis*-acting lncRNAs. (**c**) Same as in (b) but for the top twenty lncRNAs. (**d-f**) Three examples of lncRNAs and their frequency and range of contacts (histograms in orange) around their loci. (**g**) Frequency of contacts between *cis*-proximal and *trans*-acting lncRNAs. Significant differences, Mann-Whitney U test. Error bars, maximum and minimum values, excluding outliers. N values: *cis*-proximal, 1040; *trans*-acting 148

with the highest number of contacts, the range of interactions increased to approximately 40 Kb around the lncRNA locus (Fig. 4c), extending beyond the boundaries of the gene body. Notably, only when examining the top five *cis*-acting lncRNAs with the most contacts, the range of interactions further increased to 100 to 300 Kb around the lncRNA locus (Fig. 4d-f), encompassing nearby genes. Overall, our data revealed a substantial number of lncRNAs with contacts at their transcription sites, likely representing nascent transcripts.

Furthermore, we observed that *trans*-acting lncRNAs exhibited a larger number of chromatin interactions compared to *cis*-acting lncRNAs (Fig. 4g). Focusing on the top ten *trans*-acting lncRNAs, we discovered that they

have interactions with all chromosomes, in addition to displaying a peak of interactions at their locus (Fig. 5a-k). Some of these *trans*-acting lncRNAs showed interactions throughout the genome, while others displayed interactions with specific genomic regions (Fig. 5g-k). Notably, examples such as ENSACAG00000045045 (Fig. 5i) and ENSACAG00000030666 (Fig. 5j) exhibited an enrichment of chromatin contacts at unassembled scaffolds. Conversely, ENSACAG00000036367 (Fig. 5h) and ENSACAG00000039554 (Fig. 5k) displayed discrete peaks of interactions distributed along the genome.

Three *trans*-acting lncRNAs exhibit significant associations with the H4K16 acetylation signal

To investigate the potential role of trans-acting lncRNAs in gene expression regulation, we examined the top *trans*-acting lncRNAs and compared their chromatin interaction profiles against ChIP-seq data for the H4K16ac. In the green anole, H4K16ac is known to be enriched at transcription start sites and associated with active transcription [34]. We employed ChIP-seq data generated for both liver (the same tissue used for ChAR-seq) and brain [34]. We assessed whether the RNA-DNA contact regions of lncRNAs displayed a higher coverage of the H4K16ac mark compared to a randomized set of RNA-DNA contacts. Conversely, if a lncRNA is not associated with the H4K16ac signal, the enrichment for H4K16ac will not differ significantly from a randomized set of RNA-DNA contacts.

We identified notable associations by analyzing the trans-acting lncRNAs with the highest number of chromatin contacts. One lncRNA, ENSACAG00000036367, exhibited more frequent interactions with loci enriched in H4K16ac (Fig. 6a-i). Conversely, two other lncRNAs, ENSACAG00000045045 and ENSACAG00000044053, displayed fewer interactions with H4K16ac sites than expected across samples (Fig. 6a-i). We further investigated the transcription profile of these three lncRNAs and found that they are expressed in multiple tissues in both embryos and adults (Fig. 6j).

Additionally, we explored the putative gene targets of these three lncRNAs: ENSACAG00000036367 exhibited contacts with promoter regions of 86 protein-coding genes and 11 lncRNAs, while ENSACAG00000044053 contacted 368 protein-coding genes and 32 lncRNAs. Similarly, ENSACAG00000045045 had interactions with 768 protein-coding genes and 114 lncRNAs. Although functional enrichment analyses did not reveal any

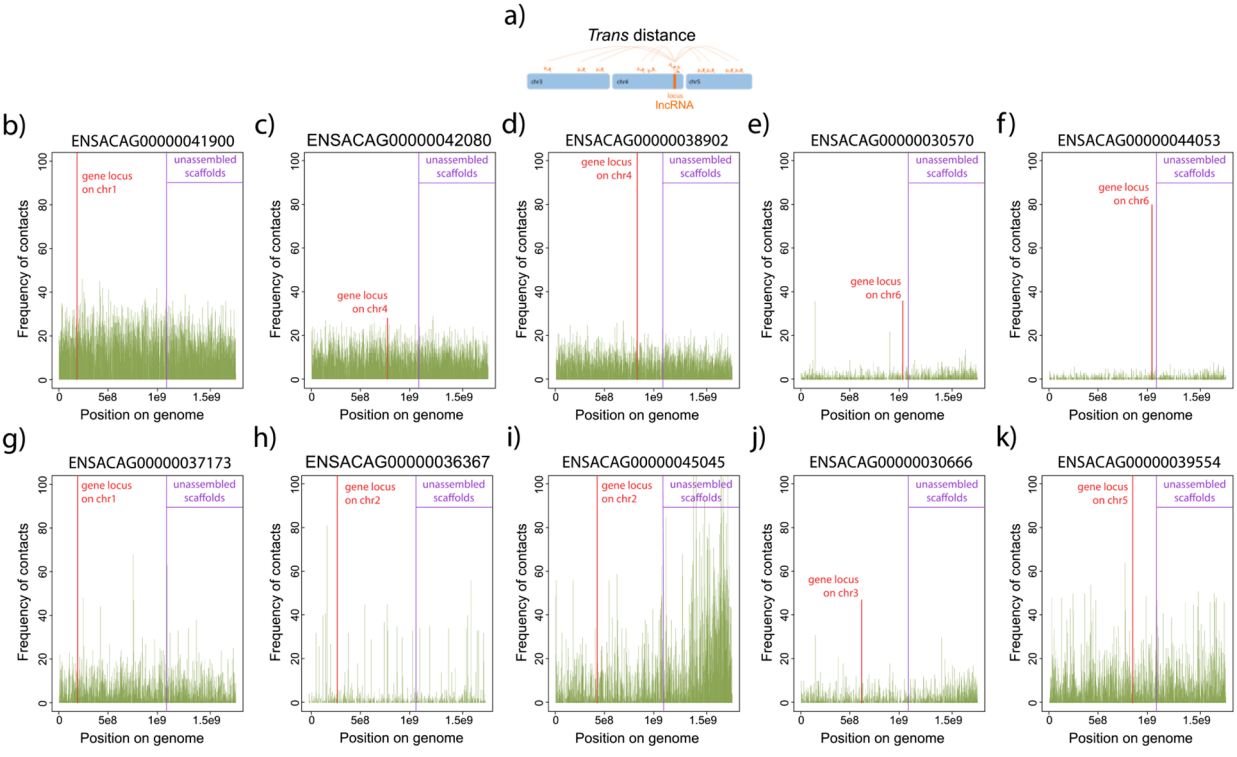


Fig. 5 Frequency of RNA-DNA contacts for*trans*-acting lncRNAs. (**a**) Illustration to help explain the RNA-DNA contacts in *trans*. (**b-k**) Frequency of contacts (histograms in green) across the genome for the top ten lncRNAs with the largest number of contacts in *trans*. The positions on the genome (X-axis) correspond to the concatenated chromosomes 1 to 6, followed by the linkage groups and unassembled scaffolds ordered alphabetically. Purple lines indicate the boundary between the assembled and unassembled parts of the genome. The frequency of contacts at the lncRNAs locus is indicated by the orange histograms; the chromosome where each lncRNA is found is also indicated

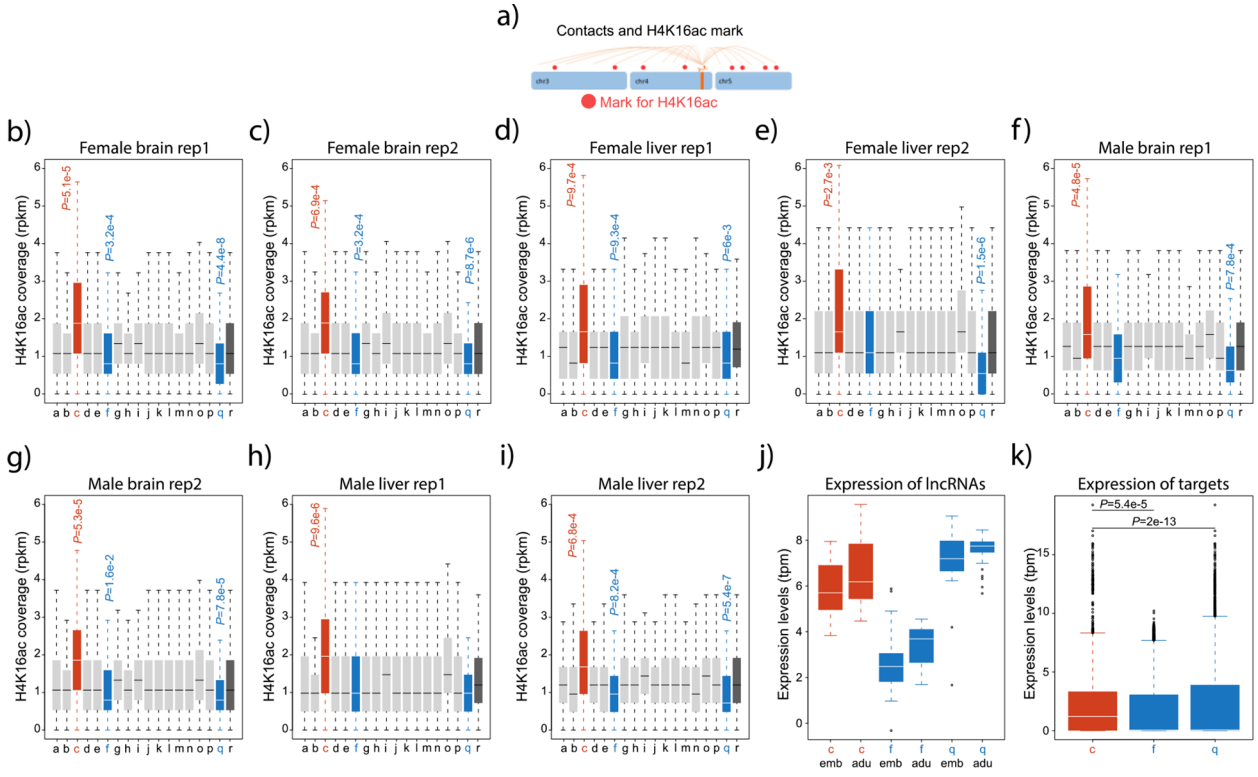


Fig. 6 H4K16ac coverage at chromatin contacts for the top*trans*-acting lncRNAs. (**a**) Illustration to help explain the RNA-DNA contacts and their association with the H4K16ac mark. (**b-e**) Normalized coverage of the H4K16ac mark (RPKM) in female and male livers at the positions where lncRNAs interact with chromatin. The identity of the *trans*-acting lncRNAs is as follows: a is ENSACAG00000041900 (locus on chromosome 1), b is ENSACAG00000037173 (locus on chromosome 1), c is ENSACAG00000036367 (locus on chromosome 2), d is ENSACAG00000031293 (locus on chromosome 2), e is ENSACAG00000034833 (locus on chromosome 2), f is ENSACAG00000045045 (locus on chromosome 2), g is ENSACAG00000041129 (locus on chromosome 2), h is ENSACAG00000030666 (locus on chromosome 3), i is ENSACAG00000040072 (locus on chromosome 4), j is ENSACAG00000044525 (locus on chromosome 4), k is ENSACAG00000042080 (locus on chromosome 4), l is ENSACAG00000038902 (locus on chromosome 4), m is ENSACAG00000039554 (locus on chromosome 5), n is ENSACAG00000032218 (locus on chromosome 5), o is ENSACAG00000042324 (locus on chromosome 6), p is ENSACAG00000030570 (locus on chromosome 6), and q is ENSACAG00000044053 (locus on chromosome 6). The lncRNA in the red boxplot (c) is significantly associated with higher coverage of the H4K16ac signal, whereas the lncRNAs in the blue boxplots (f and q) are significantly associated with lower coverage of the H4K16ac signal. Boxplots in dark grey represent H4K16ac coverage from 100,000 random positions. Significant differences, Mann-Whitney U test. Error bars, maximum and minimum values, excluding outliers. *P*-values were corrected using the Benjamin Hochberg correction. N values for the lncRNAs are 18,567, 7525, 2133, 8323, 1176, 14,767, 1489, 2842, 873, 3543, 11,509, 9044, 10,440, 1160, 3159, 1732, 1152. (**f-i**) Same as in (a-d) for female and male brains. (**j**) Expression levels (TPM) from 47 embryonic and 28 adult tissues for ENSACAG0000003636 (in red) associated with a higher signal of H4K16ac and ENSACAG00000044053 and ENSACAG00000045045 (in blue) associated with a lower signal of H4K16ac. (**k**) Expression levels (TPM) from 47 embryonic and 28 adult tissues for the gene targets of ENSACAG0000003636 (in red) and ENSACAG00000044053 and ENSACAG00000045045 (in blue). Significant differences, Mann-Whitney U test. Error bars, maximum and minimum values, excluding outliers. *P*-values were corrected using the Benjamin Hochberg correction. N values are 8112 (c), 15,210 (f), and 35,646 (q)

overrepresented biological process or metabolic pathway, an intriguing finding emerged. The expression levels of target genes associated with ENSACAG00000036367, which displayed enrichment in H4K16ac, were significantly higher than those of target genes associated with ENSACAG00000044053 and ENSACAG00000045045, which did not exhibit enrichment in H4K16ac (Fig. 6k).

Discussion

In this study, we explored the ChAR-seq methodology to investigate the contact map of chromatin-interacting RNA molecules in a non-traditional model species, *A. carolinesis*. We encountered several challenges associated with the lack of annotations for many non-coding elements and the need to restrict our analyses to chromosomes 1 to 6 due to incomplete genome assembly. Despite these obstacles, we discovered intriguing patterns regarding the RNAs in *A. carolinesis*, which should inspire future studies into RNA-DNA interactions in other non-traditional model species.

Notably, we observed that certain RNAs, such as ribosomal RNAs, snRNAs, and snoRNAs exhibited multiple interactions across the entire genome. This observation is expected from a successful ChAR-seq experiment and serves as a positive control to validate the reliability of the results. These interactions are likely non-specific,

arising from highly transcribed RNA molecules diffusing within the eukaryotic nucleus and being captured in connection with accessible chromatin. It is worth also mentioning that the ChAR-seq method captured significant amounts of nascent transcripts [24], which can also serve as positive controls of the ChAR-seq experiments because they confirm that the method truly trapped RNA molecules that were in contact or in close proximity to adjacent DNA. A recently published technique [32], RADICL-seq, attempts to mitigate the number of nascent transcripts by inhibiting RNA Polymerase II using actinomycin D before cell fixations. Although this technique works effectively with cell cultures, adapting it to bulk tissues from non-model species without available cell lines could pose challenges. Nascent transcripts attached to chromatin may represent regulatory elements within their respective loci. However, differentiating between mature RNAs that regulate their own locus and nascent transcripts attached to chromatin is challenging for *cis*-acting lncRNAs.

Despite the methodological difficulties encountered in working with *A. carolinenesis*, we successfully characterized the type and frequency of contacts made by lncRNAs. The majority of these contacts are either *cis*-proximal or *trans*-acting, exhibiting a pattern distinct from other classes of RNAs, such as ribosomal RNAs, snRNAs, and snoRNAs, which mostly exhibit *trans*-acting interactions. Although lncRNAs may have contacts in *trans* that represent spurious interactions, combining ChAR-seq and ChIP-seq data allowed us to uncover statistical associations that could help differentiate lncRNAs with potential regulatory functions. While our conclusions are based on ChAR-seq data generated from two individuals and may be limited in terms of predicting processes active in a population, the consistency of patterns observed across different types of data (ChAR-seq, ChIP-seq, and RNA-seq) supports the notion that our findings represent general active processes in *A. carolinensis*. Three lncRNAs are of particular interest due to their significant enrichment in chromatin contacts with the H4K16ac epigenetic mark. ENSACAG00000036367 may be involved in gene activation, while ENSACAG00000045045 and ENSACAG00000044053 could play a role in gene silencing. Further work using gene-specific techniques could reveal associated proteins, such as acetyltransferases or methyltransferases complexes.

The results presented in this study provide a partial glimpse into the interaction map between RNA-DNA molecules in *A. carolinensis*. Our data is limited to lncRNAs with broad expression patterns or specifically expressed in the liver. Currently, the liver is the most suitable tissue for ChAR-seq in lizards due to the considerable amount of starting material required and the small size of the organs. Ideally, future studies will integrate ChAR-seq, ChIP-seq, and RNA-seq analyses using the same sample. To further elucidate the functional characterization of annotated lncRNAs, additional data encompassing various tissues and developmental stages, along with other epigenetic modifications, would be necessary. It is worth noting that many lncRNAs are tissue-specific [35–37], and their characterization would require a broader range of experimental data. Moreover, our focus was solely on lncRNAs interacting with DNA, while numerous lncRNAs may interact with other molecules within the cell. Therefore, investigating lncRNAs associated with the proteome would necessitate an all-to-all protein-RNA protocol.

Methods

Samples

Two adult *A. carolinensis* individuals, one male and one female, were captured in Tampico, Tamaulipas, Mexico (170 m.a.s.l.; SEMARNAT Scientific Collector Permit 08–043). The animals were housed under controlled conditions, with an ambient temperature of 22±2 °C, relative humidity of 55±15%, and a day/night cycle of 12 h/12 h with live food and water *ad libitum*. The animals were housed together in a large terrarium (50.8 cm width, 40.64 cm depth, 20.32 cm height) equipped with UV light. Prior to the experimental procedure, animals were euthanized using a guillotine. The procedure was performed by an experienced technician. All animal procedures were conducted in accordance with the ethical guidelines of the Bioethical Committee of the Universidad Nacional Autónoma de México. The livers were immediately flash-frozen in liquid nitrogen and stored in 1.5 ml tubes at -80 °C until use. Due to the requirements in starting material, the liver was the most suitable option for the study because it is the largest organ in lizards. The inclusion/exclusion criteria, randomization, blinding/masking, and outcome measures do not apply to this study since the experiment was carried out with two individuals as biological replicates.

Generation of ChAR-seq data

ChAR-seq is a recent capture method that traps RNA/chromatin interactions [29, 30]. We rapidly homogenized the frozen livers with a mortar and pestle before performing the protein cross-link with formaldehyde (16%, 10 minutes at room temperature, Thermo Scientific, Cat. No. 28908). We then ligated a specific biotinylated linker to the 3' ends of the RNA molecules using an RNA ligase (T4 RNA Ligase 2, truncated KQ, NEB, Cat. No. M0373L). The top strand of the linker is a 5′-adenylated ssDNA, HPLC purified, ordered at IDT (https://www.idtdna.com/site/home/) as follows: /5rApp/AANNNAAACCGGCGTCCAAGGATCTTTAATTAAGTCGCAG/3SpC3/. The bottom

strand is a biotinylated ssDNA, HPLC purified, ordered at IDT as follows: /5Phos/GATCTGCGACTTAATTA AAGATCCTTGGACGCCGG/iBiodT/T). RNA molecules were reverse transcribed (Bst 3.0 DNA Polymerase, NEB, Cat. No. M0374L), the genomic DNA was cut with a restriction enzyme (DpnII, NEB, Cat. No. R0543L), and the 5' of the adjacent genomic DNA was ligated to the other end of the linker using a DNA ligase (T4 DNA Ligase, HC, Thermo Scientific, Cat. No. EL0013). Proteins were removed using Proteinase K (Thermo Scientific, Cat. No. EO049) and the 2nd strand of the cDNA-linker-DNA molecules was synthesized with Escherichia coli DNA polymerase I (NEB, Cat. No. M0209L). cDNA-linker-DNA molecules were purified with magnetic beads coated with streptavidin (Dynabeads MyOne Streptavidin T1, ThermoFisher, Cat. No. 65601) and prepared for sequencing. The biotinylated linker, which is the key to a successful ChAR-seq experiment, has a 5' end that contains an adenylated single-stranded sequence that can bind RNA and a 3' end that contains a recognition site for adjacent DNA fragments that have been digested by DpnII. The linker's short sequence serves as a molecular tool to control the ligation of RNA and DNA molecules that are in direct contact or in close proximity. More details about the experimental protocol can be found in [29].

Analysis of ChAR-seq data

One male and one female Illumina TruSeq stranded RNA library were sequenced using an Illumina NovaSeq 6000 machine in Novogene, California. We sequenced 1,020,074,230 (male library) and 9,96,050,284 (female library) 150 nucleotides long paired-end reads. Reads were trimmed using the list of adaptors used in the experimental protocol with trimmomatic (v0.36; parameters: ILLUMINACLIP:illuminaClipping_main.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15) [38]. Valid reads were obtained by extracting sequences that contained the tag sequence (ACCGGCGTCCAAG) present in the linker or its reverse complement (CTTGGACGCCGGT). The orientation of the tag sequence allowed the identification of the RNA and DNA fragments; the 5' end before the tag sequence corresponded to the RNA whereas the 3' end after the tag sequence corresponded to the DNA. Since we sequenced paired-end reads, the orientation of the paired read with the tag sequence relative to the paired read without the tag sequence indicates if the latter represented an RNA or DNA molecule, and was mapped accordingly. DNA fragments longer than 17 base pairs were mapped onto the *A. carolinensis* reference genome (release 104; https://www.ensembl.org) using Bowtie2 (v2.3.4.1; parameters: -p 6 -a -D 20 -R 3 -N 1 -L 18 -i S,1,0.50 --no-unal --no-head --no-sq) [39]. RNA fragments longer than 17 base pairs were mapped to the *A. carolinensis* transcriptome (release 104; https://www.ensembl.org) using Bowtie2 (v2.3.4.1; parameters: -p 6 -a -D 20 -R 3 -N 1 -L 18 -i S,1,0.50 --no-unal --no-head --no-sq) [39]. Since the annotated transcriptome from the Ensembl database could be incomplete, we also mapped using Bowtie2 the RNA fragments to a *de novo* male/female transcriptome generated using Trinity (v2.8.5, parameters: --seqType fq –single –SS_lib_type F –CPU 15 –max_memory 150G) [40]. Finally, RNA fragments longer than 50 base pairs were also mapped to the reference genome using HISAT2 (v2.1.0; parameters: -q --threads 16 -a -N 1 -L 18 -i S,1,0.50 -D 20 -R 3 --pen-noncansplice 15 --mp 1,0) [41]. RNA or DNA fragments that were less than 17 base pairs were not utilized and the paired reads were discarded. We also discarded reads where the RNA or DNA fragments showed two or more top alignments with the same score (multimappers). We verified the redundancy of the genomic coordinates of fragments that mapped to the transcriptome from Ensembl, the *de novo* transcriptome, and the genome using HISAT2. When a fragment was mapped to the same location in the different databases, we chose the coordinates from the Ensembl transcriptome. IDs from the paired-end reads were used to match the RNA and DNA mapping positions. Valid RNA-DNA contacts were defined as fragments that mapped a single time to their respective databases. We obtained 15,520,949 and 45,816,652 valid contacts for the two replicates. We assigned the contacts to specific genes using the annotations from *A. carolinensis* genome (release 104; https://www.ensembl.org). We reorganized the contacts based on the different classes of RNA molecules. We plotted the number of contacts against the type of RNA molecules and using dot plots we plotted the specific positions of the RNA contacts mapped to the transcriptome against the positions of the DNA contacts mapped to the genome using R [42]. For plotting purposes, we concatenated the chromosomes and unassembled scaffolds. We assigned continuous positions starting with chromosomes 1 to 6, then the linkage groups alphabetically, and finally the unassembled scaffolds alphabetically. We also calculated a Contact Distribution Index for each RNA class based on the chromosome with the maximum number of contacts divided by the total number of contacts in all chromosomes. We plotted the values of the index against the type of RNA molecules using R [42]. *Cis*-acting lncRNAs were defined as those having > 50% of their contacts on the same chromosome from which they are transcribed. *Trans*-acting lncRNAs were defined as those having > 50% of their interactions in other chromosomes. We calculated the distance from the locus for the *cis*-acting lncRNA as the absolute difference between the middle point of the genomic position of a lncRNA and all the genomic positions of the start of the contacts

on the same chromosome. We also estimated the difference between the middle point of the genomic position of a lncRNA and the genomic positions of its contacts restricted to 20 Kb, 40 Kb, 100 Kb, 200 Kb, and 300 Kb around the lncRNA locus. For *trans*-acting lncRNAs, we limited the analysis to those annotated on the assembled chromosomes 1 to 6 (other scaffolds are too small and tend to have an overestimated number of contacts in *trans*). We mapped the frequency of contacts along the genome, using the positions of the concatenated genome.

Analysis of ChIP-seq data

We downloaded the reference genome and transcriptome of *A. carolinensis* from the Ensembl database (release 104; https://www.ensembl.org). The results regarding the validity and robustness of the ChIP-seq data were published previously in [34]. We downloaded ChIP data for H4K16ac data for brain and liver of two female replicates and two male replicates from the NCBI-SRA database (PRJNA381064). Liver was also the tissue used for ChAR-seq. ChIP-seq data were trimmed for adaptors and low-quality positions using trim_galore (v0.6.2) (https://github.com/FelixKrueger/TrimGalore). ChIP-seq data were mapped to the *A. carolinensis* reference genome using Bowtie2 (v2.3.4.1) [39]. BAM files were indexed and we estimated RPKM values of their coverage along the entire genome for windows of 50 bp using deepTools2 (v3.3.1) (bamCoverage tool) [43]. We then obtained the RPKM values for the positions where a particular lncRNA was in contact with chromatin. We plotted the RPKM values for all contact sites for the top 17 *trans*-acting lncRNAs using R [42]. Significant differences were estimated using the Mann-Whitney U test. *P*-values were corrected using the Benjamin Hochberg correction. Statistical analyzes were conducted in R [42].

Analysis of RNA-seq data

We downloaded RNA-seq data for 15 embryonic and 14 adult tissues from females (including liver samples) and 32 embryonic and 14 adult tissues from males (including liver samples) from the NCBI-SRA database (https://www.ncbi.nlm.nih.gov/sra; PRJNA381064). The results regarding the validity and robustness of the RNA-seq data were published previously in [34]. RNA-seq data were trimmed for adaptors and low-quality positions using trim_galore (v0.6.2) (https://github.com/FelixKrueger/TrimGalore). RNA-seq data were aligned to the reference transcriptome using Kallisto [44] to estimate gene expression levels (TPM). We obtained the TPM values for specific lncRNAs and plotted the TPMs from the 15 embryonic and 14 adult tissues using R [42]. We verified that chromatin contacts were at TSS (+- 5Kb). We identified these genes and carried out functional enrichment analyses using the webgestalt platform (http://www.webgestalt.org/). We used chicken as reference species, and we performed over-representation analyses of geneontology, focusing on biological processes (no-redundant) and pathways (KEGG [45]). We used the Ensembl IDs as input data and the genome protein-coding as a reference set. We also used the string-db platform (https://string-db.org/), standard settings, to explore potential interactions among selected genes. We obtained the TPM values of these target genes and plotted their TPMs from the 15 embryonic and 14 adult tissues using R [42]. Significant differences were estimated using the Mann-Whitney U test. Statistical analyzes were conducted in R [42].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09545-5.

Supplementary Material 1

Acknowledgements

Mariela Tenorio Perez is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 814927 from CONAHCYT. We thank PAPIIT-UNAM (No. IN201920) for the support of this work.

Authors' contributions

Conceptualization: MT, DC. Methodology: MT, JS, SLFV, KO, DC. Investigation: MT, JS, SLFV, KO, DC. Funding acquisition: DC. Project administration: DC. Supervision: MT, KO, DC. Writing – original draft: MT, DC. Writing – review & editing: MT, SLFV, KO, DC.

Funding

This work was supported by a grant from PAPIIT-UNAM (IN201920).

Data Availability

All data are available in the main text or the supplementary materials. Sequencing data have been deposited in the NCBI-SRA under BioProject PRJNA880637.

Declarations

Ethics approval and consent to participate

The Animal Care Ethics Committee of the Universidad Nacional Autónoma de México, approved this study and the Mexican Government issued research and sampling permits, SEMARNAT Scientific Collector Permit 08–043. All methods comply with relevant institutional, national, and international guidelines and legislation. All methods are reported in accordance with ARRIVE guidelines for the reporting of animal experiments.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflict of interest.

Received: 30 March 2023 / Accepted: 29 July 2023
Published online: 07 August 2023

References

1. Salama SR. The complexity of the mammalian transcriptome. Adv Exp Med Biol. 2022;1363:11–22.
2. Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H. Developmental dynamics of lncRNAs across mammalian organs and species. Nature. 2019;571(7766):510–4.
3. Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature. 2017;543(7644):199–204.
4. Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M, Hanna JH, Regev A, Garber M. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol. 2016;17:19.
5. Bu D, Luo H, Jiao F, Fang S, Tan C, Liu Z, Zhao Y. Evolutionary annotation of conserved long non-coding RNAs in major mammalian species. Sci China Life Sci. 2015;58(8):787–98.
6. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505(7485):635–40.
7. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775–89.
8. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 2009;106(28):11667–72.
9. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458(7235):223–7.
10. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. The transcriptional landscape of the mammalian genome. Science. 2005;309(5740):1559–63.
11. Yu F, Zhang G, Shi A, Hu J, Li F, Zhang X, Zhang Y, Huang J, Xiao Y, Li X et al. LnChrom: a resource of experimentally validated lncRNA-chromatin interactions in human and mouse. *Database (Oxford)* 2018, 2018.
12. Patrat C, Ouimette JF, Rougeulle C. X chromosome inactivation in human development. Development 2020, 147(1).
13. Payer B, Lee JT. X chromosome dosage compensation: how mammals keep the balance. Annu Rev Genet. 2008;42:733–72.
14. Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B. Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet. 2002;36:233–78.
15. Grant J, Mahadevaiah SK, Khil P, Sangrithi MN, Royo H, Duckworth J, McCarrey JR, VandeBerg JL, Renfree MB, Taylor W, et al. Rsx is a metatherian RNA with xist-like properties in X-chromosome inactivation. Nature. 2012;487(7406):254–8.
16. Conrad T, Akhtar A. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. Nat Rev Genet. 2012;13(2):123–34.
17. Sleutels F, Zwart R, Barlow DP. The non-coding air RNA is required for silencing autosomal imprinted genes. Nature. 2002;415(6873):810–3.
18. Li X, Fu XD. Chromatin-associated RNAs as facilitators of functional genomic interactions. Nat Rev Genet. 2019;20(9):503–19.
19. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell. 2010;143(1):46–58.
20. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007;129(7):1311–23.
21. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464(7291):1071–6.
22. Amandio AR, Necsulea A, Joye E, Mascrez B, Duboule D. Hotair is dispensible for Mouse Development. PLoS Genet. 2016;12(12):e1006232.
23. Lewandowski JP, Lee JC, Hwang T, Sunwoo H, Goldstein JM, Groff AF, Chang NP, Mallard W, Williams A, Henao-Meija J, et al. The firre locus produces a trans-acting RNA molecule that functions in hematopoiesis. Nat Commun. 2019;10(1):5137.
24. Kato M, Carninci P. Genome-wide Technologies to study RNA-Chromatin interactions. Noncoding RNA 2020, 6(2).
25. Chu C, Quinn J, Chang HY. Chromatin isolation by RNA purification (ChIRP). J Vis Exp 2012(61).
26. Simon MD. Capture hybridization analysis of RNA targets (CHART). *Curr Protoc Mol Biol* 2013, Chap. 21:Unit 21 25.
27. Engreitz J, Lander ES, Guttman M. RNA antisense purification (RAP) for mapping RNA interactions with chromatin. Methods Mol Biol. 2015;1262:183–97.
28. Sridhar B, Rivas-Astroza M, Nguyen TC, Chen W, Yan Z, Cao X, Hebert L, Zhong S. Systematic mapping of RNA-Chromatin interactions in vivo. Curr Biol. 2017;27(4):602–9.
29. Jukam D, Limouse C, Smith OK, Risca VI, Bell JC, Straight AF. Chromatin-Associated RNA sequencing (ChAR-seq). Curr Protoc Mol Biol. 2019;126(1):e87.
30. Bell JC, Jukam D, Teran NA, Risca VI, Smith OK, Johnson WL, Skotheim JM, Greenleaf WJ, Straight AF. Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. Elife 2018, 7.
31. Li X, Zhou B, Chen L, Gou LT, Li H, Fu XD. GRID-seq reveals the global RNA-chromatin interactome. Nat Biotechnol. 2017;35(10):940–50.
32. Bonetti A, Agostini F, Suzuki AM, Hashimoto K, Pascarella G, Gimenez J, Roos L, Nash AJ, Ghilotti M, Cameron CJF, et al. RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions. Nat Commun. 2020;11(1):1018.
33. Eckalbar WL, Hutchins ED, Markov GJ, Allen AN, Corneveaux JJ, Lindblad-Toh K, Di Palma F, Alfoldi J, Huentelman MJ, Kusumi K. Genome reannotation of the lizard Anolis carolinensis based on 14 adult and embryonic deep transcriptomes. BMC Genomics. 2013;14:49.
34. Marin R, Cortez D, Lamanna F, Pradeepa MM, Leushkin E, Julien P, Liechti A, Halbert J, Bruning T, Mossinger K, et al. Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage. Genome Res. 2017;27(12):1974–87.
35. Chen L, Zhang YH, Pan X, Liu M, Wang S, Huang T, Cai YD. Tissue expression difference between mRNAs and lncRNAs. Int J Mol Sci 2018, 19(11).
36. Gloss BS, Dinger ME. The specificity of long noncoding RNA expression. Biochim Biophys Acta. 2016;1859(1):16–22.
37. Jiang C, Li Y, Zhao Z, Lu J, Chen H, Ding N, Wang G, Xu J, Li X. Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs. Oncotarget. 2016;7(6):7120–33.
38. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
40. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
41. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
42. Team RC. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria URL https://wwwR-projectorg/* 2021.
43. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44(W1):W160–165.
44. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7.
45. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.

Publisher's Note

X Chromosome Genomics

Mariela Tenorio and Diego Cortez, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, México

Abstract

The X chromosome is one of the more dynamic chromosomes in our genomes. This sex chromosome has evolved numerous specific mechanisms that regulate its expression levels and epigenetic signals to maintain cellular homeostasis between males and females despite the genetic decay of the Y chromosome. In this article, we will review the main findings related to the origin of the X chromosome, the complex regulation of its expression levels, and the genetic and genomic actors involved in the dosage compensation mechanisms that have so far been studied in placental mammals, marsupials, the fruit fly and the worm *Caenorhabditis elegans*. Finally, we will discuss recent genomic data on the 3D structure of this chromosome and the relationship between gene content and gene expression of the X chromosome and human diseases and male infertility.

Glossary

Cellular homeostasis State of steady internal, physical, and chemical variables maintained by cells for optimal functioning.
X chromosome compensation An active regulatory mechanism that acts on the X chromosome to equalize the expression levels of genes between both sexes.
Gene copy number Number of copies that a gene has on a particular genome; in diploid organisms, such as humans, genes are found in two copies.
Chromosome territory Regions of the nucleus preferentially occupied by particular chromosomes in interphase.
Gene expression dosage balance To attain similar gene expression levels in males and females by a gene-by-gene or a whole chromosome molecular mechanism.
Gametologs Paralog genes found one copy on the X chromosome and one copy on the Y chromosome.
Heterochromatic DNA tightly packed form of DNA (chromatin) with no transcription activity (silenced).
Klinefelter syndrome is a condition that occurs in males with an extra X chromosome (XXY).
Long non-coding RNAs Genes that produce messenger RNAs of 200 or more nucleotides with no obvious open reading frame and no similarity with known proteins.
MSY, the male-specific region of the Y chromosome The non-recombinant part of the Y chromosome that is only found in males.
MSLc, the male-specific lethal dose compensation complex A chromatin-modifying complex composed of five protein subunits and two non-coding RNAs found in the fruit fly (*Drosophila melanogaster*); knock out of the MSLc complex produces a lethal phenotype.
Mosaicism Possession of two or more genetically different cell lines in a single organism.
PAR, pseudoautosomal region Chromosomal regions shared between X and Y chromosomes where they still recombine and are necessary for chromosomal segregation during mitosis and meiosis.
PRC, Polycomb repressive complex Protein complex with histone methyltransferase activity that primarily methylates histone H3 on lysine 27 (H3K27me3), a mark of transcriptionally silent chromatin.
Sex-specific genes Genes only found in one sex.
Sex-antagonistic genes Genes that are beneficial for one sex but detrimental to the other sex.
Sex-beneficial genes Genes with an important sex-specific function, such as spermatogenesis.
SOX3 gene The "SRY-Box Transcription Factor 3" gene represents the ancestral version of the *SRY* gene and it is located on the X chromosome of placental and marsupial mammals.
SRY gene The "Sex-determining Region of the Y chromosome" gene was discovered in 1990 and it is a sex-determination gene in marsupial and placental mammals located on the short arm of the Y chromosome.
TADs, topologically associated macrodomains Fundamental units of three-dimensional (3D) nuclear organization.
Turner syndrome Genetic disorder that affects the development of girls. The cause is a missing or incomplete X chromosome (XO).

 doi:10.1016/B978-0-12-822563-9.00072-X

Key Points

- The evolutionary events that led to the origin of the X chromosome.
- The highly-regulated expression levels of the X chromosome and the main genetic and genomic actors involved.
- The balance of X chromosome expression outputs between males and females to maintain cellular homeostasis; is a process that has evolved in different species of animals with remarkable similarities.
- Recent genomic data has allowed the study of the structural changes followed by the X chromosome during dosage compensation.
- Genes on the X chromosome and changes in the epigenetic landscape are related to infertility and other diseases in humans.

Introduction

Within the vertebrate genome, the most dynamic chromosomes are the sex chromosomes since they are subjected to particular evolutionary forces that can trigger massive genetic loss, the deactivation of large chromosomal regions, the emergence of complex dosage compensation mechanisms that balance expression ratios between males and females through chromatin remodeling, the mass movement of genes, etc., (Bachtrog *et al.*, 2014; Balaton *et al.*, 2018). Moreover, the study of sex chromosomes can be a particularly useful tool to understand basic biological processes related to the appearance of new genes, the regulation of gene expression, the dynamics of epigenetic markers, the effect of sexual selection, the relationship between the environment and gene function, as well as many other equally fascinating topics. This article will review some of the most interesting findings related to X chromosome genomics.

Origin of the X Chromosome

In humans and many vertebrates, female cells carry two copies of the X chromosome, whereas male cells have one X and one Y chromosome; the Y chromosome is male-specific, highly heterochromatic, and frequently contains only a limited set of genes. XY chromosomes derive from a pair of ancestral autosomes following the emergence of one or multiple sex-determining genes (Bachtrog, 2013). For example, sex chromosomes in marsupial and placental mammals originated ~180 million years ago (Cortez *et al.*, 2014) when the Y-linked genes *SRY* (Sinclair *et al.*, 1990) diverged from the X-linked gene *SOX3*. The protein coded by *SRY* could start the testis developmental pathway (Sekido and Lovell-Badge, 2008; Li *et al.*, 2014). Soon after the emergence of *SRY*, the Y chromosome underwent a series of large inversions that halted recombination with the X chromosome, resulting in massive gene loss due to the combined effect of lack of homologous recombination, which corrects errors and deletions, and accumulation of transposable elements that increase the frequency of genetic loss through non-homologous recombination (Furman *et al.*, 2020). Currently, Y chromosomes contain less than 10% of the genetic material of the X chromosome (Bachtrog, 2013). The Y-specific genes are located in the male-specific region of the Y chromosome (MSY) (Charlesworth and Charlesworth, 2000; Skaletsky *et al.*, 2003). The difference in gene content between the X and Y chromosomes led to an imbalance of expression levels among both sexes, triggering the evolution of a mechanism that could restore dosage equivalence of the X chromosome between males and females. The X chromosome is the carrier of long non-coding RNAs (lncRNAs) that, together with specialized proteins, can reshape the structure and transcription activity (Plath *et al.*, 2002).

Gene Expression of the X Chromosome

The level at which the X chromosome is expressed is very important for dosage compensation among males and females. Dosage compensation is a regulatory mechanism that acts on entire chromosomes or at a gene-by-gene level and its objective is to equalize the expression levels of genes between both sexes (Mank, 2013). For most genes in a genome, males and females have two copies. However, sex-specific duplications or deletions of small chromosomal regions can affect their copy number (Ercan, 2015). Particularly, following the decay of the Y sex chromosome, the levels of messenger RNAs and the abundance of the coded proteins differ between sexes, producing different stoichiometries of protein complexes that may not be functional in males causing problems to the organism (Brockdorff and Turner, 2015). Some genes are more sensitive to changes in their copy number because a minimal amount of protein is required to achieve a biological function. These genes are known as dosage-sensitive and are the first to be regulated by dosage compensation mechanisms to maintain cellular homeostasis.

Dosage compensation mechanisms acting on sex chromosomes can be limited to regulating dosage-sensitive genes, when a few of these genes are present on the sex chromosomes, as in the case of chicken and platypus (Julien *et al.*, 2012) or they can regulate the expression levels of entire X chromosomes. In *Drosophila melanogaster* (the fruit fly), for example, the X chromosome in males becomes overexpressed to achieve the same expression output as the two X chromosomes in females (**Fig. 1**) (Conrad and Akhtar, 2012). In contrast, in placental mammals and marsupials, one of the two X chromosomes in female cells is almost completely silenced to maintain dosage balance with male cells (**Fig. 1**) (Payer and Lee, 2008). The chromatin of the inactivated X chromosome in placental mammals is organized into two large super-domains (**Fig. 2**) (Fang *et al.*, 2019), a feature unique to this

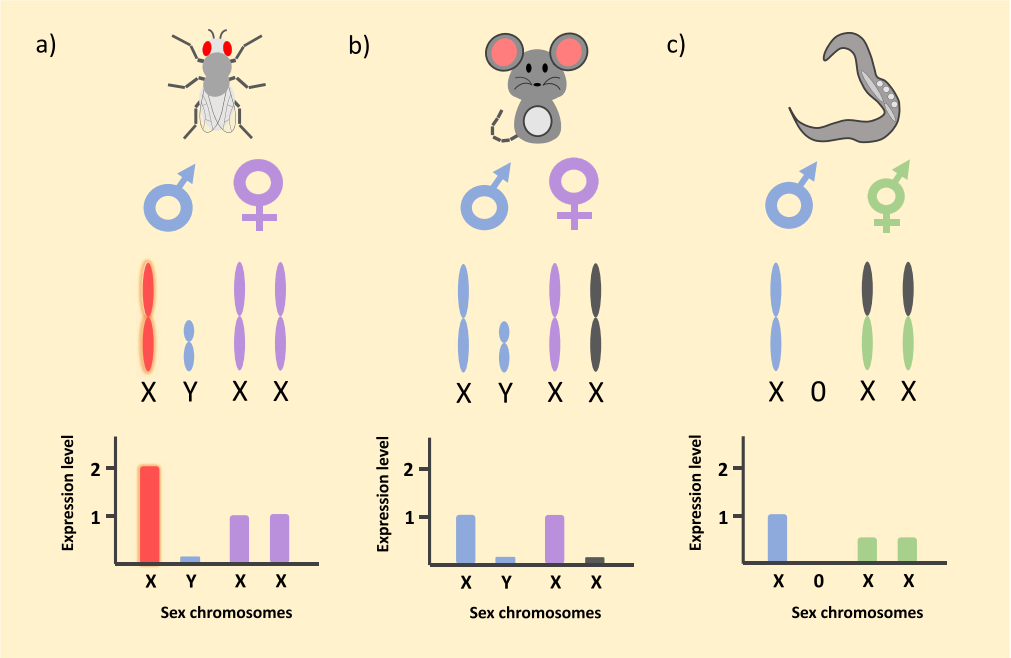


Fig. 1 Expression levels of sex chromosomes. (a) Expression levels of X and Y chromosomes in males and females of the fruit fly. (b) Expression levels of X and Y chromosomes in males and females of placental mammals. (c) Expression levels of X chromosomes in hermaphrodites (with two X chromosomes) and males (with only one X chromosome) in *C. elegans*.

chromosome since other chromosomes present multiple domains. In humans and mice, several components allow the structure of the two super-domains to be maintained (Balaton *et al.*, 2018). Among them, we have the *dxz4* microsatellite, which is located right at the division of the two super-domains and its presence is critical to forming the structure (**Fig. 2**) (Balaton *et al.*, 2018), and the lncRNA FIRRE that helps to stabilize the super-domains by retaining the repressive histone 3-lysine 27-trimethylation mark and guiding the inactivated X chromosome near the nuclear envelope (**Fig. 2**) (Balaton *et al.*, 2018). The cellular machinery is actively maintaining the repression of gene expression on the inactivated X chromosome, but even so, there are genes that escape the expression silencing because they need higher expression levels for cellular homeostasis (Berletch *et al.*, 2011; Balaton *et al.*, 2015). Some genes that escape from X inactivation are tissue-specific, although the majority of escaping genes show the same expression levels across multiple human tissues (Tukiainen *et al.*, 2017). Also, genes escaping X inactivation are enriched in gametologs (genes with a paralogue on the Y chromosome) (Bellott *et al.*, 2014). In many species, the X and Y chromosomes share an identical region that continues to have homologous recombination, which is necessary for the correct segregation of XY chromosomes during mitosis and meiosis. This region is known as the pseudoautosomal region (PAR), which is typically located at the edge of the sex chromosomes (Bachtrog, 2013); most of the genes found in the PAR region escape inactivation, although they are expressed less than in the active X chromosome (Tukiainen *et al.*, 2017).

X Chromosome Dosage Compensation

X chromosome dosage compensation is driven by lncRNAs that reside on the X chromosome that will be either inactivated or over-expressed. X-specific lncRNAs are involved in the active recruitment of chromatin-modifying proteins that change specific histones by adding acetylation or methylation marks (**Fig. 3**). In placental mammals, the lncRNA *XIST* (X-inactive specific transcript) is activated by the protein YY1 early during female development (Payer and Lee, 2008). *XIST* can then recruit chromatin-modifying proteins (Polycomb repressive complex -PRC-) that will trimethylate the lysine 27 on histone 3, H3K27me3 (**Figs. 2–3**) (Patrat *et al.*, 2020), therefore, progressively inducing chromatin compaction and transcription silencing (Plath *et al.*, 2002). The *XIST* gene is highly conserved in humans and mice, and the tandem repetitive regions show some degree of conservation, for example, the A repeats are very well conserved, while the B to F repeats differ between the species. It has also been shown that the internal region of exon 7 of *XIST* is functionally essential to establishing X chromosome inactivation (Balaton *et al.*, 2018). Similarly, in marsupials, the lncRNA *RSX* (RNA-on-the-silent X) inactivates the X chromosome probably by interacting with the PRC (Polycomb repressive complex) to inhibit the transcription of the X chromosome (**Fig. 3**) (Sprague *et al.*, 2019). In the fruit fly, the lncRNA *ROX2* (RNA on the X) is associated with the male-specific lethal (MSL) protein complex that can hyper-acetylate the lysine 16 on histone 4 (H4K16ac) to increase the transcription output of the X chromosome in males (**Figs. 2–3**) (Gelbart *et al.*, 2009). The CLAMP protein has been shown to promote three-dimensional aggregation of the male-specific

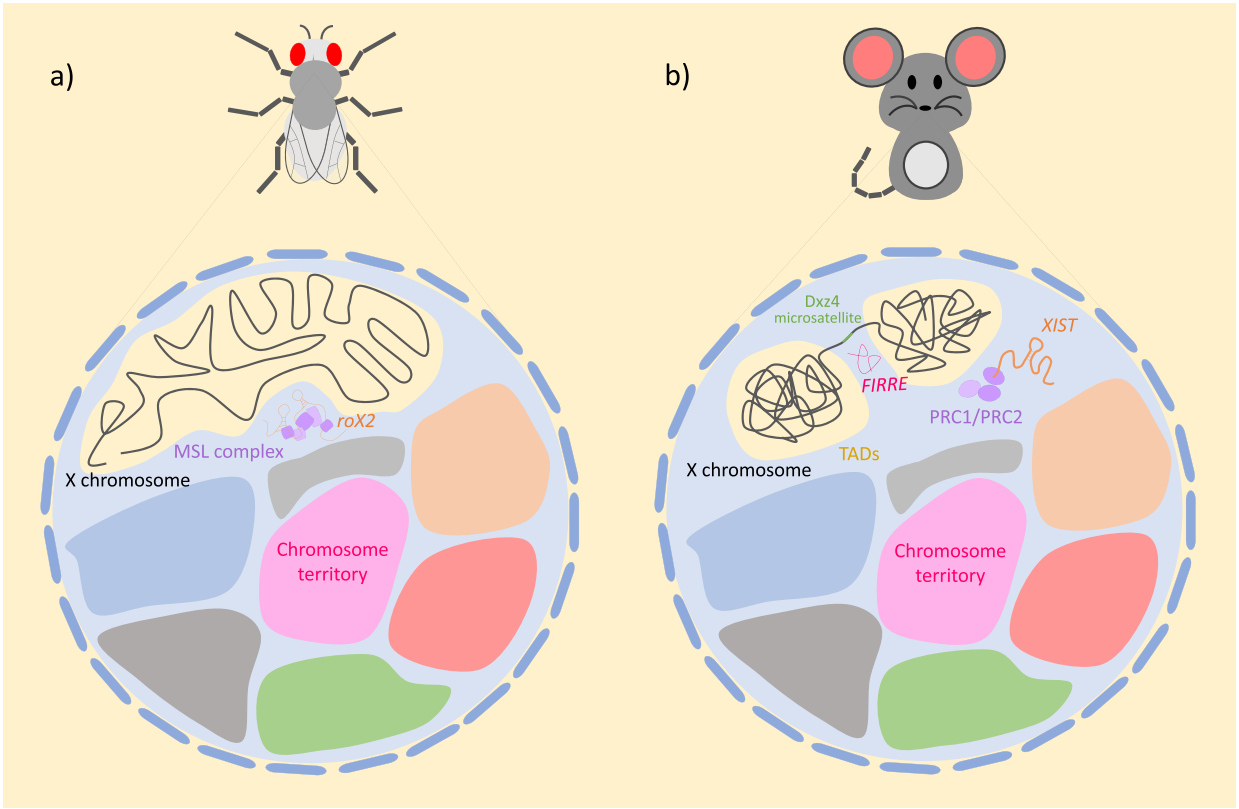


Fig. 2 Nuclear localization, domains, and genetic factors in X chromosome regulation. (a) The X chromosome in the fruit fly is located near the nuclear envelope, with a structure similar to other chromosomes. Overexpression of this chromosome is triggered by the joint activity of the lncRNA *ROX2* and the MSL complex. (b) The X chromosome in placental mammals is located near the nuclear envelope, structured into two super-domains (TADs) that are maintained by the *dxz4* microsatellite and the lncRNA *FIRRE*. Inactivation of the X chromosome is controlled by the lncRNA *XIST* that interacts with the Polycomb repressive complex (PRC). Chromosome territories are regions of the nucleus preferentially occupied by particular chromosomes in interphase.

lethal dosage compensation complex (MSLc) to promote the three-dimensional molding of the X chromosome so that its active chromatin regions interact with other insulating proteins (Jordan and Larschan, 2021). Lastly, in hermaphrodites (XX) of *Caenorhabditis elegans*, dosage compensation is active on both X chromosomes, where the expression levels are lowered by half (**Fig. 3**), rather than completely inhibiting one sex chromosome, to attain the expression levels shown by males (with a single X chromosome) (Strome *et al.*, 2014). The activity of the DC complex is important in this process (**Fig. 3**) (Strome *et al.*, 2014). In *C. elegans*, however, it is not yet known whether a lncRNA participates in regulating the X chromosomes (**Fig. 3**).

The reasons why in some species dosage compensation is achieved through overexpression and in others through repression of the X chromosome is still debated. It has been proposed that the strength of X and Y *cis*-regulatory elements may play an important role in the evolution of dosage compensation systems (Lenormand and Roze, 2022). Hence, it could be hypothesized that the presence of strong X *cis*-regulatory elements could lead to X overexpression whereas weak X *cis*-regulatory elements could be associated with X inactivation.

3D Structure of the X Chromosome

Chromatin-remodeling proteins are one of the main actors in the mechanisms that bring dosage equivalence between sex chromosomes. The lncRNAs indirectly or directly attract these proteins to modify or re-organize specific regions on the X chromosomes (Makki and Meller, 2021). The inactive X chromosome in mammals is organized differently from the autosomal and active X chromosomes. As mentioned, the inactive X chromosome shows two topologically associated macrodomains (TADs) that are stabilized near the periphery of the nuclear envelope (**Fig. 2**) (Fang *et al.*, 2019). In *C. elegans*, the chromatin architecture of both X chromosomes in hermaphrodites changes, and the sex chromosomes are relocated closer to the periphery, distinctly from autosomes, which strongly interact with the nuclear lamina (Makki and Meller, 2021). Finally, in the fruit fly, the topology of the X chromosome in males is stretched and opts a similar shape to that of the autosomes, which are composed of multiple TADs (**Fig. 2**) (Quinn and Chang, 2015; Schauer *et al.*, 2017).

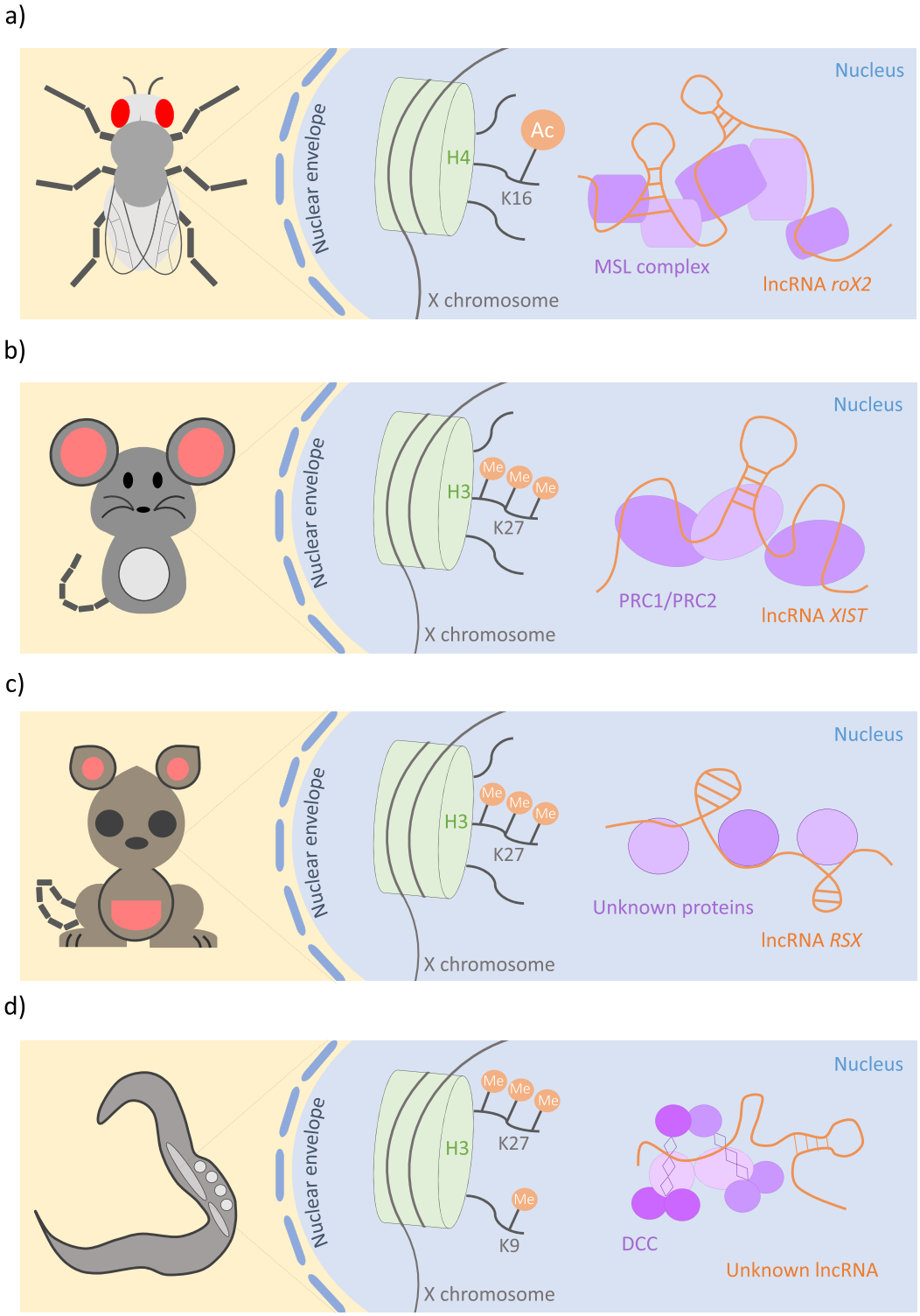


Fig. 3 Histone modifications following X inactivation or X overexpression. (a) In the fruit fly, the lncRNA *ROX2* interacts with the MSL complex to specifically acetylate the lysine 16 on histone 4 (a transcription activation mark), which initiates the overexpression of the X chromosome. (b) In placental mammals, the lncRNA *XIST* interacts with the Polycomb repressive complex (PRC) to specifically trimethylate the lysine 27 on histone 3 (a transcription repression mark), triggering the inactivation of the X chromosome. (c) In marsupials, the lncRNA *RSX* is hypothesized to interact with the Polycomb repressive complex (PRC) to trimethylate the lysine 27 on histone 3 (a transcription repression mark) and induce the inactivation of the X chromosome. (d) In *C. elegans*, the two X chromosomes in hermaphrodites are downregulated by the tri-methylation of lysine 27 on histone 3 and the methylation of lysine 9 on histone 3 (two transcription repression marks). This process is catalyzed by the DC complex and, potentially, by still unknown lncRNAs.

Infertility, Diseases, and Mosaicism

The X chromosome is involved in several diseases and infertility problems, particularly among males. For example, male infertility is commonly caused by defects of sperm. It has recently been revealed that the X chromosome of mammals is enriched in genes that are expressed during spermatogenesis (Vockel *et al.*, 2021). Since males only have one X chromosome, this makes them more likely to develop loss-of-function mutations affecting sperm production. It has also been shown that the X chromosome expression levels in individuals with Turner syndrome (having a single X chromosome) and Klinefelter syndrome (XXY) are more balanced, thus causing the individuals to have milder diseases compared to other trisomies (e.g., trisomy of chromosome 21) that have more serious consequences (Pereira and Doria, 2021). Also, aberrations in *XIST* expression and, in some cases, disruption of X-chromosome inactivation as a whole, are related to Alzheimer's disease (Chanda and Mukhopadhyay, 2020). Moreover, as we age, some cells in our bodies become mosaic, meaning they can lose either the male or female X chromosome or the male Y. However, recent studies showed that the frequency of loss of the male X chromosome in leukocytes is rare relative to the female X chromosome (Zhou *et al.*, 2021).

Conclusions

Ever since Susumu Ohno proposed that one X chromosome in females of placental mammals could be silenced (Ohno, 1967; Beutler, 1998), researchers have been fascinated by this process that involves both sex-specific and chromosome-specific genetic signals that regulate the epigenetic landscape and the expression output of entire chromosomes to restore cellular homeostasis between males and females. Similar processes have been investigated in marsupials, the fruit fly, and *C. elegans*. In all of these cases, chromatin-modifying proteins and lncRNAs are the major players in molecular mechanisms. Recent advances in genomics and molecular biology have revealed the structure of the sex chromosomes during the epigenetic changes, and have identified new genetic elements involved in the complex regulatory pathways that control gene expression levels of the X chromosome. Although many advances have been made in the field over the past few years, numerous aspects of the dosage compensation mechanisms and their relationship with human health and development remain unanswered. Also, probably many other species of animals with sex chromosomes have evolved interesting but still unknown dosage mechanisms to solve potential expression unbalances between males and females. Sex chromosomes are influenced by selection forces that lead to the evolution of sex-specific genes, the fixation of sex-antagonistic genes, and the emergence of sex-beneficial genes. The study of these genes has been important in understanding the regulation of chromosomal-wide gene expression levels and their role in health, sexual selection, and the evolution of the species. Finally, analyses of the structure of the X chromosome have opened a wide array of opportunities for the study of chromatin dynamics and epigenetic markers that will be the focus of future work for many decades to come.

Acknowledgments

This work was supported by grant from PAPIIT-UNAM (gran IN201920). Mariela Tenorio Perez is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 814927 from CONACYT.

References

Bachtrog, D. (2013) Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. Nat. Rev. Genet. 14, 113–124.
Bachtrog, D., Mank, J. E. and Peichel, C. L. *et al.* (2014) Sex determination: Why so many ways of doing it? PLOS Biol. 12, e1001899.
Balaton, B. P., Cotton, A. M. and Brown, C. J. (2015) Derivation of consensus inactivation status for X-linked genes from genome-wide studies. Biol. Sex Differ. 6, 35.
Balaton, B. P., Dixon-McDougall, T., Peeters, S. B. and Brown, C. J. (2018) The eXceptional nature of the X chromosome. Hum. Mol. Genet. 27, R242–R249.
Bellott, D. W., Hughes, J. F. and Skaletsky, H. *et al.* (2014) Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature 508, 494–499.
Berletch, J. B., Yang, F., Xu, J., Carrel, L. and Disteche, C. M. (2011) Genes that escape from X inactivation. Hum Genet 130, 237–245.
Beutler, E. (1998) Susumu Ohno: The father of X-inactivation. Cytogenet. Cell Genet. 80, 16–17.
Brockdorff, N. and Turner, B. M. (2015) Dosage compensation in mammals. Cold Spring Harb. Perspect. Biol. 7, a019406.
Chanda, K. and Mukhopadhyay, D. (2020) LncRNA Xist, X-chromosome instability and alzheimer's disease. Curr. Alzheimer Res. 17, 499–507.
Charlesworth, B. and Charlesworth, D. (2000) The degeneration of Y chromosomes. Philos. Trans. R. Soc. Lond. B Biol. Sci. 355, 1563–1572.
Conrad, T. and Akhtar, A. (2012) Dosage compensation in Drosophila melanogaster: Epigenetic fine-tuning of chromosome-wide transcription. Nat. Rev. Genet. 13, 123–134.
Cortez, D., Marin, R. and Toledo-Flores, D. *et al.* (2014) Origins and functional evolution of Y chromosomes across mammals. Nature 508, 488–493.
Ercan, S. (2015) Mechanisms of x chromosome dosage compensation. J. Genom. 3, 1–19.
Fang, H., Disteche, C. M. and Berletch, J. B. (2019) X inactivation and escape: Epigenetic and structural features. Front. Cell Dev. Biol. 7, 219.
Furman, B. L. S., Metzger, D. C. H. and Darolti, I. *et al.* (2020) Sex chromosome evolution: So many exceptions to the rules. Genome Biol. Evol. 12, 750–763.
Gelbart, M. E., Larschan, E., Peng, S., Park, P. J. and Kuroda, M. I. (2009) Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. Nat. Struct. Mol. Biol. 16, 825–832.
Jordan 3rd, W. and Larschan, E. (2021) The zinc finger protein CLAMP promotes long-range chromatin interactions that mediate dosage compensation of the Drosophila male X-chromosome. Epigenet. Chromatin 14, 29.
Julien, P., Brawand, D. and Soumillon, M. *et al.* (2012) Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. PLOS Biol. 10, e1001328.
Lenormand, T. and Roze, D. (2022) Y recombination arrest and degeneration in the absence of sexual dimorphism. Science 375, 663–666.

Li, Y., Zheng, M. and Lau, Y. F. (2014) The sex-determining factors SRY and SOX9 regulate similar target genes and promote testis cord formation during testicular differentiation. Cell Rep. 8, 723–733.
Makki, R. and Meller, V. H. (2021) When down is up: Heterochromatin, nuclear organization and X upregulation. Cells 10.
Mank, J. E. (2013) Sex chromosome dosage compensation: Definitely not for everyone. Trends Genet. 29, 677–683.
Ohno, S. (1967) Sex chromosomes and sex linked genes. Berlin, Heidelberg: Springer.
Patrat, C., Ouimette, J. F. and Rougeulle, C. (2020) X chromosome inactivation in human development. Development 147.
Payer, B. and Lee, J. T. (2008) X chromosome dosage compensation: how mammals keep the balance. Annu. Rev. Genet. 42, 733–772.
Pereira, G. and Doria, S. (2021) X-chromosome inactivation: Implications in human disease. J. Genet. 100.
Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. and Panning, B. (2002) Xist RNA and the mechanism of X chromosome inactivation. Annu. Rev. Genet. 36, 233–278.
Quinn, J. J. and Chang, H. Y. (2015) In situ dissection of RNA functional subunits by domain-specific chromatin isolation by RNA purification (dChIRP). Methods Mol. Biol. 1262, 199–213.
Schauer, T., Ghavi-Helm, Y. and Sexton, T. *et al.* (2017) Chromosome topology guides the drosophila dosage compensation complex for target gene activation. EMBO Rep. 18, 1854–1868. https://doi.org/10.15252/embr.201744292.
Sekido, R. and Lovell-Badge, R. (2008) Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer. Nature 453, 930–934.
Sinclair, A. H., Berta, P. and Palmer, M. S. *et al.* (1990) A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Nature 346, 240–244.
Skaletsky, H., Kuroda-Kawaguchi, T. and Minx, P. J. *et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423, 825–837.
Sprague, D., Waters, S. A. and Kirk, J. M. *et al.* (2019) Nonlinear sequence similarity between the Xist and Rsx long noncoding RNAs suggests shared functions of tandem repeat domains. RNA 25, 1004–1019.
Strome, S., Kelly, W. G., Ercan, S. and Lieb, J. D. (2014) Regulation of the X chromosomes in Caenorhabditis elegans. Cold Spring Harb. Perspect. Biol. 6.
Tukiainen, T., Villani, A. C. and Yen, A. *et al.* (2017) Landscape of X chromosome inactivation across human tissues. Nature 550, 244–248.
Vockel, M., Riera-Escamilla, A., Tuttelmann, F. and Krausz, C. (2021) The X chromosome and male infertility. Hum. Genet. 140, 203–215.
Zhou, W., Lin, S. H. and Khan, S. M. *et al.* (2021) Detectable chromosome X mosaicism in males is rarely tolerated in peripheral leukocytes. Sci. Rep. 11, 1193.

Regulación de los cromosomas sexuales por RNAs largos no codificantes

Mariela Tenorio y Diego Cortez
Centro de Ciencias Genómicas de la UNAM

Los cromosomas sexuales son los responsables de la determinación del sexo en múltiples especies, incluyendo al humano. A raíz de la evolución de esta función, estos cromosomas han tenido cambios masivos a tal punto de apagar o activar casi por completo su expresión según el tipo de células en las que están presentes (Mank, 2013).

Este fenómeno es entendible bajo la luz de la teoría del origen de los cromosomas sexuales propuesta por el genetista Hermann Müller (Bachtrog, 2013). La teoría de Müller propone que los cromosomas sexuales se originan a partir de un par de cromosomas autosomales. Por ejemplo, en los mamíferos, hace aproximadamente 180 millones de años, un cromosoma autosomal adquirió un gen capaz de activar la cascada de señalización que desarrolla el testículo (Bachtrog, 2013; Marais y Galtier, 2003). Este gen lo conocemos ahora como *SRY*. La aparición de este gen produjo que un autosoma se transformara en un cromosoma sexual específico de machos, el cromosoma Y; a la pareja de este cromosoma la llamamos X. Rápidamente la región alrededor del gen *SRY* fue aislada de la recombinación por una gran inversión cromosomal, lo que ocasionó que los cromosomas X y Y dejaran de recombinar y *SRY* se fijara en los machos de la población. La falta de recombinación del Y condujo a la acumulación de mutaciones, de secuencias repetidas y, posteriormente, a la pérdida masiva de material genético (Bachtrog, 2013; Gatler, 2014). Así, los machos se quedaron con un cromosoma X y un cromosoma Y degenerado, mientras que las hembras conservaron dos cromosomas X. La expresión de aquellos genes que anteriormente conservaba el Y se perdió, provocando un desbalance de expresión génica entre machos y hembras. Casi de manera simultánea a la degeneración del cromosoma Y, evolucionó un mecanismo que pudiera restablecer el balance de expresión génica entre los dos sexos (Ercan, 2015).

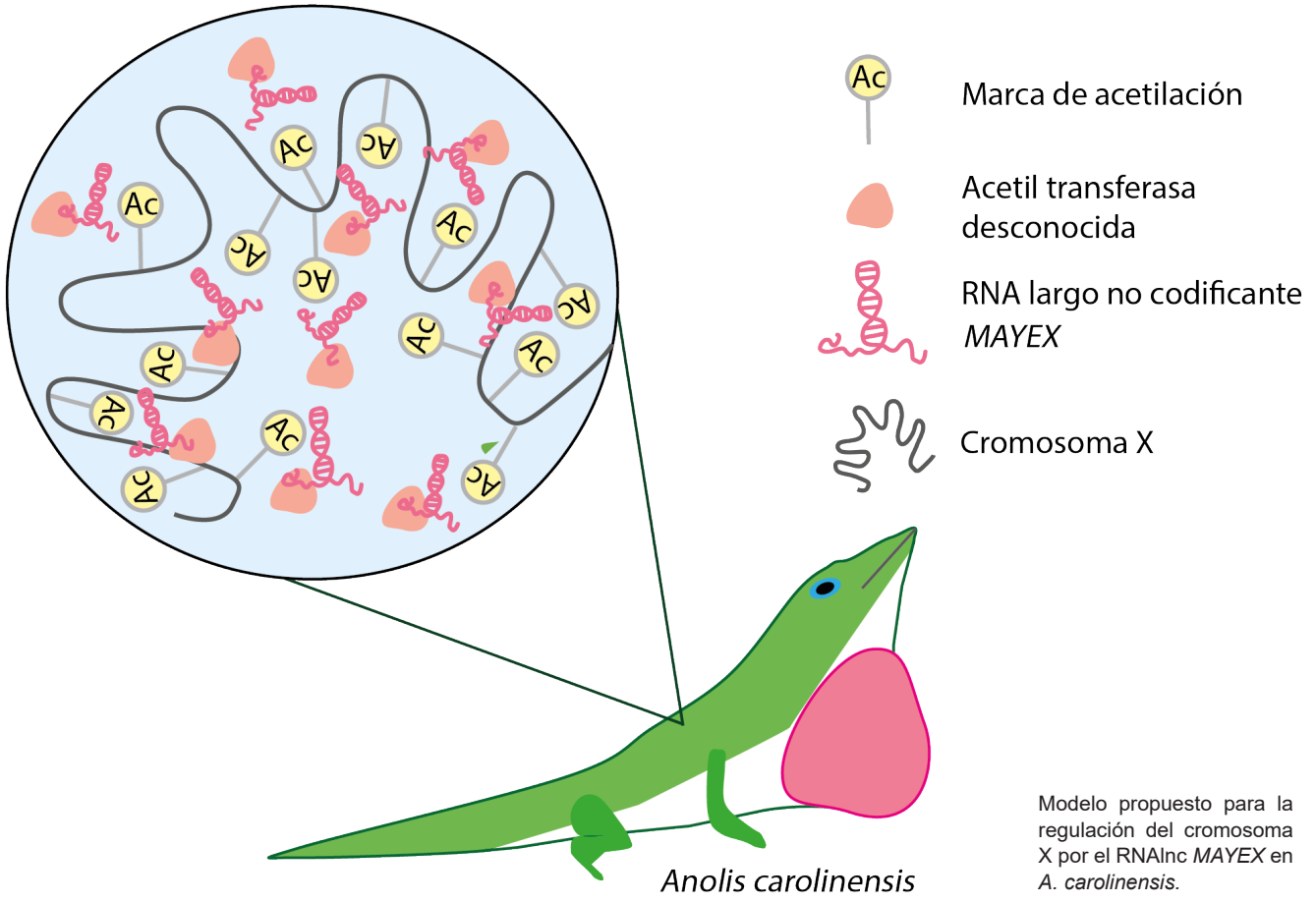
Los mecanismos de compensación de dosis génica ayudan a igualar la expresión de los cromosomas X en machos y hembras (Gatler, 2014). En estos procesos hay cambios en las marcas epigenéticas del cromosoma X que regulan sus niveles de expresión. Los mecanismos de compensación de dosis génica son variados y pueden ocurrir tanto en machos como en hembras. En mamíferos, por ejemplo, las hembras apagan casi por completo uno de sus cromosomas X a través de la metilación de las histonas, específicamente trimetilan la lisina 27 de la histona 3 (H3K27me3) (Payer y Lee, 2008). En cambio, en *Drosophila*, el cromosoma X de los machos aumenta sus niveles de expresión porque se hiper-acetila, específicamente se acetila la lisina 16 de la histona 4 (H4K16ac) (Conrad y Akhtar, 2012).

En los mecanismos que se conocen en mamíferos, marsupiales y en la mosca de la fruta, el sistema es orquestado por RNAs largos no codificantes (RNAlnc), que son transcritos que no se traducen a proteínas y miden más de 200 nucleótidos (Wang y Chang 2011). Los RNAlnc son capaces de responder a diversos estímulos, reclutan enzimas modificadoras de la cromatina hacia genes específicos y funcionan como andamios al formar complejos ribonucleoproteicos para actuar sobre las histonas de los cromosomas X (Wang y Chang 2011). Los RNAlnc que desencadenan la inactivación del cromosoma X de las hembras son *XIST* (mamíferos placentarios) y RSX (en marsupiales). Por otro lado, el RNAlnc que participa en la hiperacetilación del cromosoma X en machos de la mosca de la fruta *ROX2* (Quinn y Chang, 2015). Con nuestro trabajo hemos podido añadir un RNAlnc a la regulación de los cromosomas X. Hablamos de la regulación del cromosoma X de la lagartija verde, *Anolis carolinensis*.

Los cromosomas sexuales XY de *A. carolinensis* aparecieron aproximadamente hace 160 millones de años (Marin *et al*., 2017). Al igual que en mamíferos, el cromosoma Y de la lagartija verde se degeneró a tal grado que sólo conserva 7 genes de los 350 que tenía cuando era un autosoma (Marin *et al*., 2017). De forma similar a lo que ocurre en *Drosophila*, el cromosoma X en los machos de *A. carolinensis* presenta una hiper-acetilación (H4K16ac) del cromosoma X en machos; primer caso de regulación del X de los machos en vertebrados.

Durante el trabajo de tesis de doctorado de Mariela Tenorio, identificamos un RNAlnc que sólo está activo en el cromosoma X de los machos al que denominamos como *MAYEX* por "Male-specific long non-coding RNA AmplifYing the Expression of the X". *MAYEX* está fuertemente asociado con la maquinaria de acetilación y logra crear un dominio que permite que diferentes regiones del cromosoma X se plieguen hacia el locus de *MAYEX* y se sobreacetilen, logrando así la sobre-regulación del cromosoma X.

Aún quedan muchas preguntas sin respuesta, pues no conocemos las proteínas con las que interactúa *MAYEX*, ni los factores específicos de machos que pudieran estar regulando la expresión de *MAYEX* para que se active únicamente en machos.



Modelo propuesto para la regulación del cromosoma X por el RNAlnc *MAYEX* en *A. carolinensis*.

Bibliografía

1. Bachtrog D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature reviews. Genetics*, 14(2), 113–124. https://doi.org/10.1038/nrg3366
2. Conrad, T., Akhtar, A. (2012). Dosage compensation in *Drosophila melanogaster*: epigenetic fine-tuning of chromosome-wide transcription. *Nature reviews. Genetics*, 13(2), 123–134. https://doi.org/10.1038/nrg3124
3. Ercan S. (2015). Mechanisms of x chromosome dosage compensation. *Journal of genomics*, 3, 1–19. https://doi.org/10.7150/jgen.10404
4. Gartler S. M. (2014). A brief history of dosage compensation. *Journal of genetics*, 93(2), 591–595. https://doi.org/10.1007/s12041-014-0360-5
5. Mank J. E. (2013). Sex chromosome dosage compensation: definitely not for everyone. *Trends in genetics : TIG*, 29(12), 677–683. https://doi.org/10.1016/j.tig.2013.07.005
6. Marais, G., Galtier, N. (2003). Sex chromosomes: how X-Y recombination stops. *Current biology : CB*, 13(16), R641–R643. https://doi.org/10.1016/s0960-9822(03)00570-0
7. Marin, R., Cortez, D., Lamanna, F., *et al.* (2017). Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage. *Genome research*, 27(12), 1974–1987. https://doi.org/10.1101/gr.223727.117
8. Payer, B., Lee, J. T. (2008). X chromosome dosage compensation: how mammals keep the balance. *Annual review of genetics*, 42, 733–772. https://doi.org/10.1146/annurev.genet.42.110807.091711
9. Quinn, J. J., Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nature reviews. Genetics*, 17(1), 47–62. https://doi.org/10.1038/nrg.2015.10
10. Wang, K. C., y Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Molecular cell*, 43(6), 904–914. https://doi.org/10.1016/j.molcel.2011.08.018

Agradecimientos

Mariela Tenorio Pérez es alumna del Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) y recibió la beca 814927 del CONACYT.
Nuestro trabajo fue financiado por PAPIIT-UNAM no. IN201920.