

MÁS ALLÁ DEL DOGMA CENTRAL DE LA BIOLOGÍA MOLECULAR: EL ARN NO CODIFICANTE Y EL QUE CODIFICA

Juan Carlos García-Soriano y Puri Fortes

Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra (UNAV). Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd). *Cancer Center* de la Clínica Universidad de Navarra (CCUN), Pamplona

<https://doi.org/10.18567/sebbmrev.226.202509.dc4>

En 1958 Francis Crick postuló el Dogma central de la Biología Molecular, una idea que Jacques Monod immortalizó diciendo: «lo que ocurre en las bacterias, ocurre en el elefante». Este dogma establece un flujo de información del ADN al ARN y de ahí a las proteínas (Figura 1). El ADN se transcribe para dar ARN mensajero (ARNm) utilizando un código similar de cuatro nucleótidos (nt) y éste se traduce a proteínas cambiando la información a 20 aminoácidos (aa). Este dogma ponía en el centro de la ecuación al ARNm e ignoraba a otros ARNs que no se traducen, y por ello son llamados «no codificantes». Estos incluyen, por ejemplo, el ARN de transferencia (ARNt), el ribosómico (ARNr) o los pequeños nucleares (ARNpn), que son necesarios para la maduración del ARNm o su traducción a proteínas. Más adelante, se descubrieron los micro ARNs (miARN). Habían pasado desapercibidos por su pequeño tamaño (algunos con tan solo 21 nt). Sin embargo, son numerosos, se expresan a altos niveles y resultaron ser extraordinariamente relevantes para controlar la expresión génica de los ARNm con los que interactúan. Algo parecido ocurrió con ARNs más largos no codificantes (ARNlnc).

Los ARNlnc se identificaron en estudios transcriptómicos, una poderosa tecnología que permite mapear todos los ARNs producidos en una célula. Esta técnica mostró que casi todo el ADN de la célula se transcribe en ARN. Para sorpresa de la comunidad científica se encontró que, de todo este ARN, el ARNm constituye tan solo un 2 %. El 90 % restante, es ARN no codificante que incluye, además de ARNt, ARNr, o miARN, a los ARNlnc. Estos últimos son los más numerosos. Los últimos análisis del genoma humano cuantifican casi 20.000 genes que



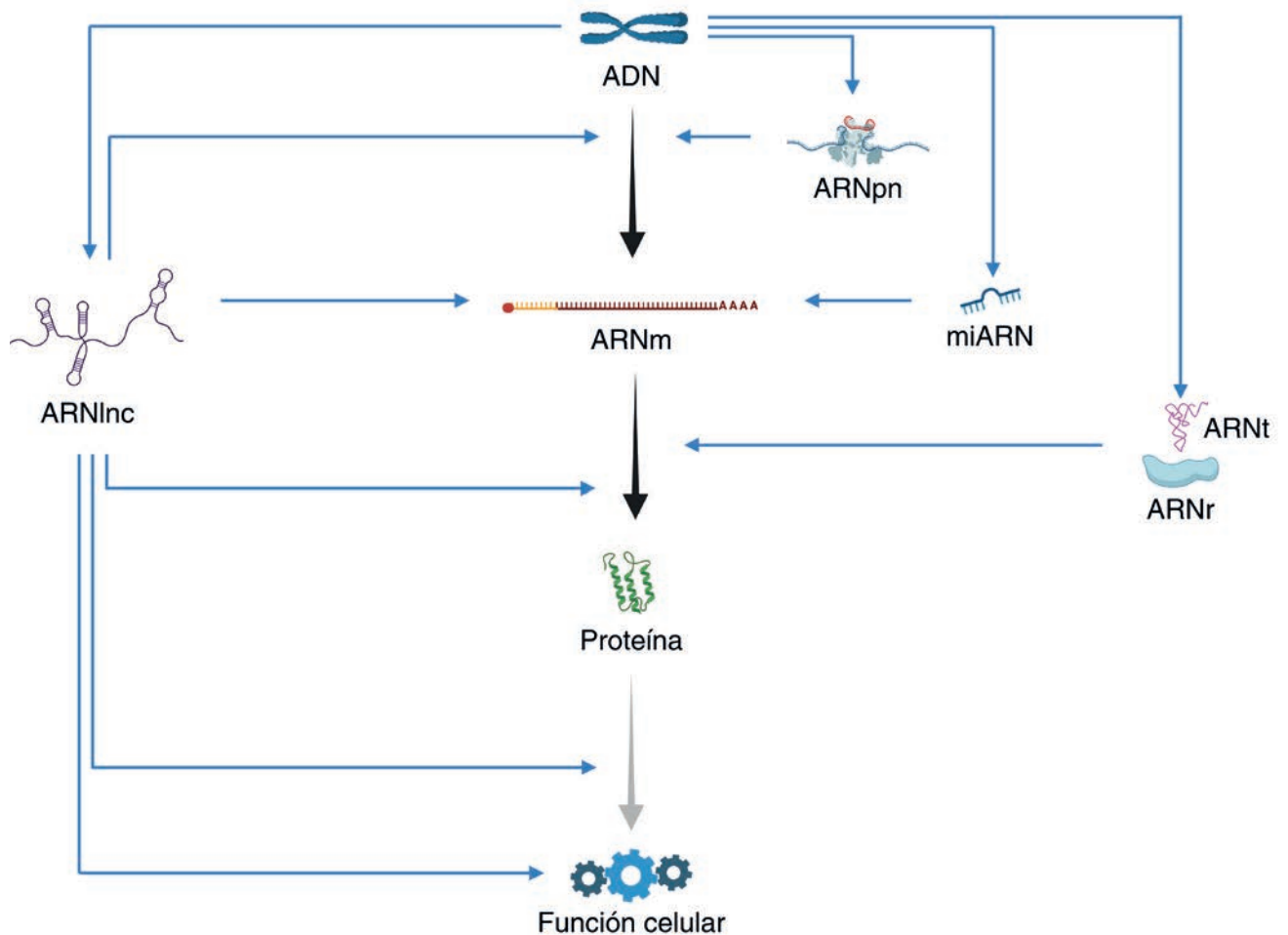


Figura 1

Dogma Central de la Biología Molecular con ARNs diversos. En el centro, con flechas negras, se muestra el flujo de información bajo la visión tradicional del Dogma: el ADN se transcribe para dar lugar a ARN mensajero (ARNm) y este se traduce a proteína. A la derecha de esta línea se muestran los ARN pequeños nucleares (ARNnpn; que participan en la maduración del ARN), los ARN de transferencia y ribosomales (ARNt y ARNr; que participan en la traducción) y los microARN (miARN; que disminuyen la estabilidad y la traducción de los ARNm con los que interaccionan). A la izquierda se muestran los ARN largos no codificantes (ARNlnc), que ejercen una gran variedad de funciones de regulación de la expresión génica. Tanto las proteínas como los ARNlnc pueden formar parte de maquinarias celulares no relacionadas directamente con la expresión génica.

codifican proteínas (se transcriben a ARNm) y unos 50.000 que producen ARNlnc. Esta cifra incluye también a los pseudogenes, antiguos genes codificantes que perdieron esa capacidad. Como la mayoría de los estudios hasta la fecha se habían enfocado en los genes codificantes (esos 20.000) y en las proteínas que estos producen, los análisis transcriptómicos indicaron que nos falta por estudiar lo que se produce en la mayor parte del genoma.

Enseguida aparecieron detractores a la idea de que esto fuese relevante. Los críticos argumentaban que los bajos niveles de producción de los ARNlnc y su escasa conservación entre especies los convertían en «basura» genética, sin funciones importantes. Esto de la conservación es fascinante. La mayoría de los genes que codifican proteínas, esos 20.000 en humanos, son muy parecidos a los de organismos tan diversos como gusanos o moscas, a pesar de las

evidentes diferencias en complejidad evolutiva. Sin embargo, los ARNlnc son muy distintos entre especies y su número aumenta con la complejidad evolutiva: hay menos en bacterias, más en moscas, muchos más en mamíferos, y los primates somos los que tenemos más genes de ARNlnc en nuestro genoma. Se postula que esta abundancia de ARNlnc en primates podría deberse, en parte, a una infección masiva de retrovirus en un ancestro común. Estos virus

insertaron sus genomas en el de nuestro antepasado, colonizándolo y dejando innumerables réplicas de genes virales en prácticamente la mitad de su genoma. Lo que a primera vista parecería perjudicial, pudo haber conferido una ventaja evolutiva tan única, que hoy somos los herederos de ese primate ancestral.

Nuestro genoma no tiene mucha «basura», sino, en gran medida, restos de virus que, modificados por la evolución, nos han dotado de miles de ARNinc diferentes. Una teoría evolutiva muy interesante, conocida como «*same brick, different building*» (mismo ladrillo, diferente edificio), que sugiere que las proteínas están muy conservadas porque actúan como los «ladrillos» de los «edificios» que formamos los seres vivos. Así como una caseta de perro y la catedral de León pueden construirse con ladrillos similares, la complejidad de un organismo no reside tanto en los «ladrillos» sino en el «arquitecto» o regulador que dicta su disposición específica en cada especie. Es tentador pensar que los ARNinc podrían ser los «arquitectos», los reguladores que controlan qué ladrillo (proteína) se produce en cada tejido y dónde debe colocarse. Esto ofrece una explicación elegante a la asombrosa diversidad biológica que se obtiene con un conjunto de componentes moleculares compartidos en gran parte.

Los resultados obtenidos durante los últimos años no son suficientes para validar el papel de los ARNinc como los «arquitectos» del genoma de esta compleja teoría evolutiva. Sin embargo, sí indican que los ARNinc estudiados no son «basura». La mayoría de los genes que producen ARNinc analizados responden a estímulos de una forma similar a los genes codificantes. Los ARNinc y los ARNm comparten procesos de síntesis y maduración, aunque

esta última parece que funciona un poco peor con los ARNinc. Sin embargo, en general, comparados con los ARNm, los ARNinc se localizan más preferentemente en el núcleo de la célula, son más específicos de tejido y de especie, y se producen a niveles mucho más bajos. Estas características están directamente relacionadas con la función de los ARNinc, que en muchos casos sirven para regular la expresión de genes concretos. Ciertos ARNinc actúan en *cis*, es decir, ejercen su función en el mismo sitio del genoma en el que se sintetizan, actuando sobre genes vecinos. Esta acción tan localizada explicaría una función con muy baja expresión. Otros funcionan en *trans*, como las proteínas, viajan por la célula hasta localizarse donde ejercen su función. Independientemente de actuar en *cis* o *trans*, los ARNinc han de plegarse en una estructura determinada por su secuencia, como las proteínas, y unirse a ADN, proteínas o a otros ARNs para formar complejos activos. Se han descrito ARNinc que contribuyen a modificar el genoma haciéndolo más o menos accesible a la maquinaria de transcripción, otros que modulan la transcripción de determinados genes directamente, o bien que alteran la maduración del ARNm, su transporte, su estabilidad o su traducción (Figura 2). En general, muchos de los ARNinc estudiados son excelentes reguladores de la expresión génica. Como son específicos de tejido, contribuyen a que, aunque todas las células compartan el mismo ADN, esta información se utilice de forma específica en distintos tejidos.

Además de funciones reguladoras, muchos ARNinc se unen a maquinarias celulares para modular su función. Ese es el caso de NIHCOLE, que se une a la maquinaria de reparación del ADN y aumenta su eficacia (Figura 3).

Sorprendentemente, NIHCOLE no se detecta en células sanas, pero sí a niveles elevados en tumores de distinto origen: cáncer de mama, de pulmón, de colon y el cáncer de hígado más prevalente (hepatocarcinoma). En este último caso, donde NIHCOLE se ha estudiado en más detalle, los pacientes con peor pronóstico tienen niveles más elevados de NIHCOLE. Esto se debe, probablemente, a que el crecimiento tumoral y los tratamientos de radio y quimioterapia producen roturas en el ADN que se reparan más eficazmente cuando las células tumorales tienen NIHCOLE. Las células con NIHCOLE deberían seleccionarse y proliferar más, y eso las vuelve vulnerables. En el laboratorio hemos demostrado que bloquear la producción de NIHCOLE en células tumorales induce su muerte. Al igual que NIHCOLE, numerosos ARNinc están implicados en el crecimiento tumoral o el desarrollo de otras enfermedades.

Otra función ejercida por los ARNinc es quizá la más inesperada. A pesar de haber sido bautizados como «no codificantes», algunos ARNinc se traducen para producir proteínas no canónicas, generalmente pequeñas, y por ello llamadas microproteínas (miP). Al igual que los miARN hace un par de décadas, las miP han pasado desapercibidas hasta hace poco. La dificultad para identificar estas miP reside en que la región de un ARN que se traduce, el marco de lectura abierto (ORF, por sus siglas en inglés *open reading frame*), es difícil de identificar, especialmente cuando es pequeño. Esto se debe a la complejidad de distinguir entre ORF verdaderos de los que no se traducen. Por ello, para diferenciar entre ORF genuinos y espurios, se estableció por consenso que las secuencias codificantes comienzan con los nt AUG (el codón iniciador), tienen más de 300 nt y están flanqueadas por

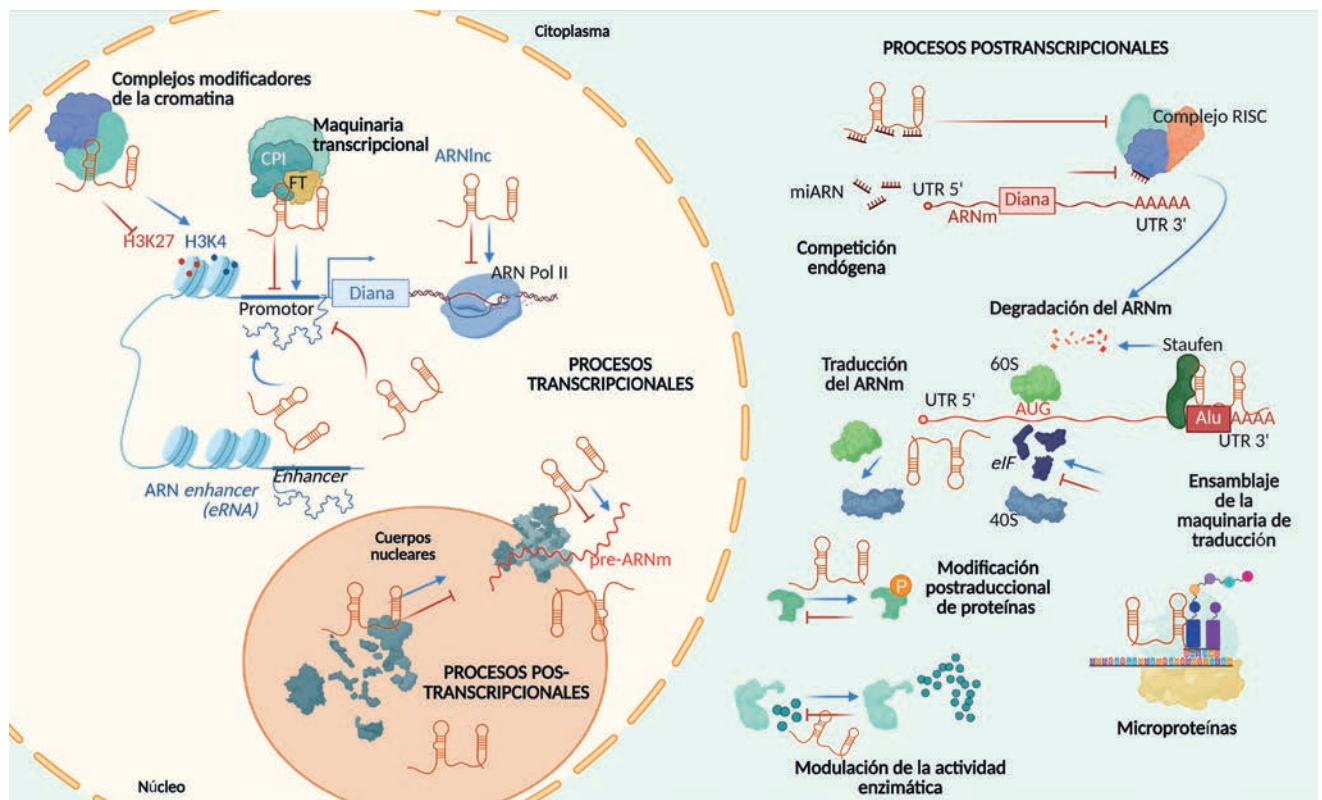


Figura 2

Funciones de los ARNlnc. A la izquierda se muestra un esquema del núcleo de la célula, donde los ARNlnc se esquematizan uniéndose a maquinarias implicadas en modificación de la cromatina, en transcripción y en procesos posttranscripcionales. A la derecha se esquematiza el citoplasma, donde ciertos ARNlnc están implicados en varios procesos posttranscripcionales relacionados con la estabilidad del ARNm (por ejemplo, compitiendo por miARN) y su traducción. Algunos ARNlnc pueden modular la modificación de proteínas o su actividad o pueden traducirse para producir microproteínas. Se indican procesos de activación (flechas azules) o inhibición (líneas rojas).

regiones no traducidas (UTR, por sus siglas en inglés *untranslated region*) 5' y 3' (Figura 4). Esto dibujó un gen modelo en el que cada ARNm producía una única proteína canónica. Sin embargo, esta visión ha sido desafiada por los datos obtenidos con el perfilado de ribosomas (RiboSeq), una tecnología novedosa que permite identificar fragmentos de ARN protegidos por los ribosomas durante la traducción. El RiboSeq ofrece una imagen precisa de toda la traducción celular con una resolución a nivel de nt. A la información de RiboSeq se suma la proteómica, que utiliza espectrometría de masas para identificar trozos de proteínas (péptidos) de extractos celulares o unidos

al complejo mayor de histocompatibilidad (CMH). Esta última aplicación, conocida como *ligandoma*, explota el hecho de que las proteínas celulares se fragmentan en péptidos (epítopos). Algunos de estos se unen al CMH, que los presenta al sistema inmune, controlando que la célula no exprese proteínas «extrañas» (como las virales). Si el sistema inmune detecta epítopos anómalos, desencadena la destrucción celular. Así, cualquier miP que se sintetice, incluso a bajos niveles, puede ser detectada en estudios de *ligandoma* si los epítopos resultantes se unen fuertemente al CMH.

Los análisis de los datos de RiboSeq y proteómica han revelado una complejidad inesperada.

Algunas bases de datos de células humanas incluyen más de 500.000 miP, mientras que otras, con mayor nivel de confianza, describen más de 30.000. Las miP se han clasificado, según su ORF, en aquellas derivadas de ARN circulares (circORF, otra variante de la diversidad del ARN celular), largos o cortos no codificantes (lncORF y sncORF, por sus siglas en inglés) y aquellas traducidas desde la UTR 5' o 3' de ARNm (5'ORF y 3'ORF) o incluso desde las secuencias codificantes canónicas, solapando con la ORF de la proteína principal, pero en una fase de lectura diferente (5' o 3'sORF o internas, intORF) (Figura 4). La mayoría de las miP son 5'ORF (47 %) y lncORF (21 %), seguidas de intORF (10 %). De

hecho, se estima que alrededor del 75 % de todos los ARNm no solo producen una proteína canónica, sino también una o más miP. Esto nos lleva a un nuevo paradigma donde cada ARN es capaz de generar múltiples proteínas.

Todo lo anterior modifica la imagen clásica de los genes codificantes. Primero, porque muchos ARNs codifican para más de una proteína, siendo los ARNlnc los más complejos (hasta 8 miP/ARNlnc en nuestros datos). Además, porque las secuencias codificantes pueden ser mucho más pequeñas de 300 nt, con miP de tan solo 3 nt (un solo aa). Asimismo, a diferencia de las proteínas canónicas, cuyo ORF siempre empieza con el codón AUG

(traducido a metionina), el de las miP puede iniciar en AUG o en una mutación puntual de esta secuencia, resultando en codones que codifican para leucina, valina, treonina, isoleucina, arginina o lisina. Quizá las miP no comiencen con estos aminoácidos, sino que, en esta traducción no canónica, se permitan apareamientos incorrectos por los que se traduzca una metionina a partir de un codón que no codifica para ella. Nuestros resultados muestran que esta iniciación no requiere una secuencia lineal concreta (lo que se conoce como secuencia Kozak), sino que parece depender de estructuras secundarias de ARN alrededor del codón de inicio. Por lo tanto, las miP que

comienzan en codones distintos de AUG provienen de regiones estructuradas. Quizás, si el ribosoma se «entretiene» con estas estructuras, su menor velocidad le permite aceptar pequeños errores de inicio. Por último, comparado con las proteínas canónicas, las miP están enriquecidas en aminoácidos hidrófobos, lo que podría activar su degradación.

Al igual que ocurrió con los ARNlnc, se ha sugerido que las miP podrían no ser relevantes. Esto se basa en que muchas podrían ser inestables, su tamaño pequeño dificulta que formen estructuras robustas y ordenadas, y no están conservadas. Sin embargo, el número de miP aumenta con la complejidad del desarrollo,

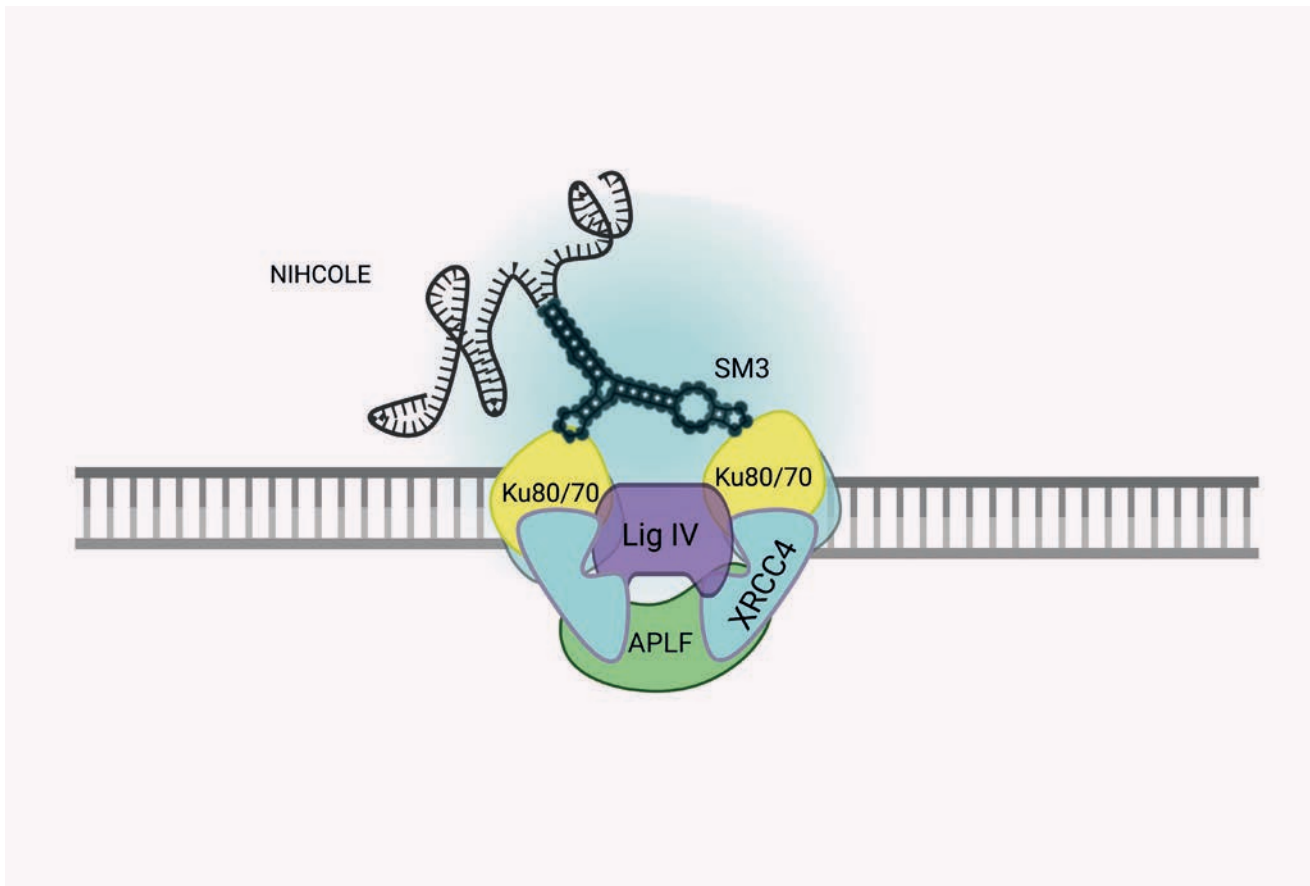


Figura 3

Mecanismo de acción de NIHCOLE. NIHCOLE tiene un dominio denominado SM3 que interacciona con las proteínas Ku (heterodímero de Ku80 y Ku70), que reconoce los extremos rotos del ADN. Este complejo se une a la maquinaria formada por APLF y XRCCA unido a la ligasa IV, que une los extremos del ADN roto. En presencia de NIHCOLE, aumenta la sinapsis con la que se unen los dos extremos del ADN y se mejora la eficacia de ligación.

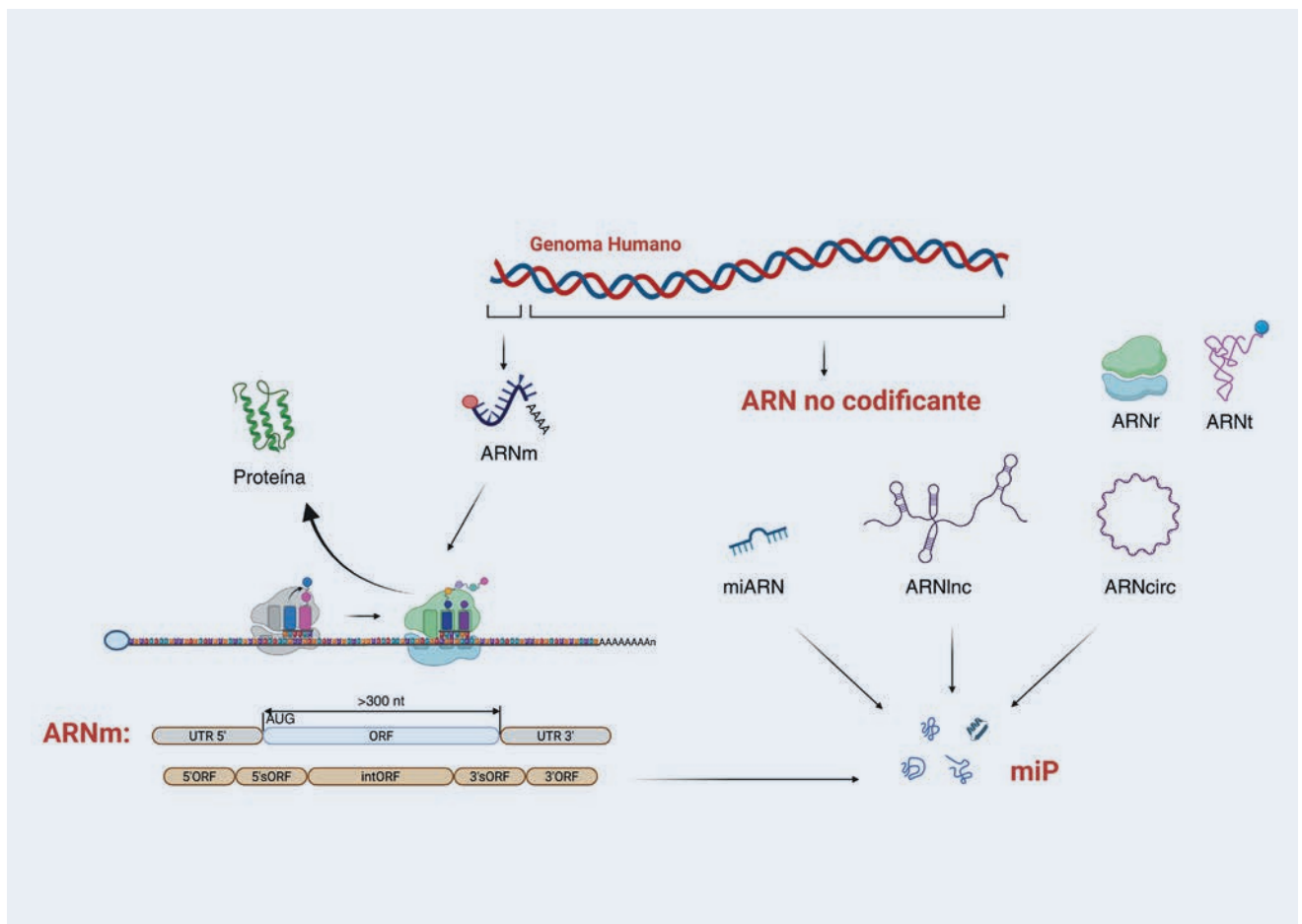


Figura 4

Esquema del genoma humano codificante y no codificante. La región del genoma humano catalogada clásicamente como codificante (que se transcribe a ARNm y da lugar a proteínas canónicas) es pequeña. La región no codificante es mayoritaria. A partir de ella se transcriben ARNr, ARNt, miARN, ARNlnc y ARNcirc. Sin embargo, algunos de los ARN considerados no codificantes tienen marcos de lectura abiertos (ORF, por sus siglas en inglés *open reading frame*) y pueden traducirse a proteínas generalmente pequeñas denominadas microproteínas (miP). Estos incluyen tanto los ARN no codificantes como las regiones no traducidas (UTR) de los ARNm. En este último caso se detectan proteínas no canónicas derivadas de UTR 5' y 3' (5' y 3' sORF o de regiones que comparten secuencia con el ORF de la proteína canónica, pero están en otra fase de lectura (5' o 3' sORF, cuando el ORF solapa con la zona 5' o 3' UTR y parcialmente con el ORF de la proteína canónica, o intORF cuando solapa completamente con éste).

siendo los humanos quienes poseen el mayor número, lo cual no es sorprendente ya que tenemos más ARNlnc capaces de producir las (se estima que un 14-22 % de ARNlnc producen miP). Por ello, se ha sugerido que muchas de estas miP podrían ser «basura», y representar la producción de proteínas muy inestables y sin función, debido a que evolutivamente son demasiado recientes. Sin embargo, ya se han descrito unas 200 miP funcionales.

Curiosamente, las miP también pueden actuar en *trans* o en *cis*,

de manera independiente de la secuencia. La función de estas últimas es el mero acto de traducción: la traducción de algunos 3'ORF puede mejorar la expresión del ORF canónico cercano, mientras que la de ciertos 5'ORF puede disminuirla. La mayoría de las miP estudiadas funcionan en *trans*, se localizan en músculos (incluido el corazón), y están unidas a membranas o mitocondrias. Afectan a: (i) la función mitocondrial y el metabolismo celular; (ii) la expresión génica (procesamiento, estabilidad, modificación y traducción del

ARN, y ensamblaje y estabilidad de proteínas); (iii) la señalización celular, proliferación y diferenciación; (iv) las funciones relacionadas con el sistema inmunitario; (v) la reparación del ADN y el estrés celular. Estudios preliminares estiman que un 10-19 % de las miP son importantes para la supervivencia celular, y que la modulación de los niveles del 72 % de las miP produce cambios celulares detectables.

Por último, las miP podrían ser clave en el desarrollo de terapias contra el cáncer. Muchos de los ARNlnc que se expresan en

cáncer no se detectan en tejido sano. Entre ellos, se han identificado varios que se traducen a miP. Estas miP, específicas de cáncer, aun siendo no funcionales o inestables, pueden degradarse dentro de la célula y ser presentadas al sistema inmune a través de las moléculas del CMH. Si las identificamos, se podrían desarrollar vacunas contra el cáncer con las que activar a las células del sistema inmune para que reconozcan a estas miP y, tras ello, eliminen las células tumorales.

De hecho, estos factores no conservados, tanto ARNlnc como miP, de evolución tan reciente, podrían ser especialmente relevantes en la carcinogénesis y en el crecimiento tumoral. Las células cancerosas se enfrentan a un enorme estrés por la falta de oxígeno y nutrientes, vigilancia del sistema inmune, o por tratamientos como la quimio o la radioterapia, que buscan eliminarlas.

Para sobrevivir, solo tienen una opción: evolucionar y adaptarse usando su genoma. En células altamente proliferativas, las mutaciones, reorganizaciones genómicas y alteraciones epigenéticas pueden generar heterogeneidad celular y contribuir a la selección de células en división resistentes al estrés. Estos cambios pueden afectar la parte conservada del genoma, permitiendo la selección de células que han activado distintos oncogenes. Esto conlleva un riesgo, ya que se podrían dañar genes esenciales. Sin embargo, los cambios en la parte no conservada deberían ser más favorables, más numerosos (dado que alrededor del 95 % del genoma no está conservado) y, por lo tanto, generar una mayor diversidad de la cual se pueden seleccionar células resistentes. Nuestra hipótesis es que, en el estresante microambiente tumoral, la parte no conservada del genoma podría

generar una gran cantidad de ARNlnc, y estos, a su vez, diversas miP. Algunos de estos factores podrían ser capaces de regular la expresión de genes que alivian el estrés o que forman estructuras que mejoran los mecanismos para resolverlo y/o impulsar la progresión tumoral. Este es el caso del ARNlnc NIHCOLE. Las células que expresen estos factores, nuevos oncogenes, tenderían a ser seleccionadas por su ventaja adaptativa. Estos factores no conservados a menudo se han ignorado. Sin embargo, creemos que podrían ser dianas terapéuticas ideales, ya que las células tumorales deberían ser dependientes de su expresión. Los ARNlnc y las miP, a pesar de que en ciertos ámbitos se consideran «basura» genética, podrían estar mostrándonos el camino hacia nuevas y efectivas terapias contra el cáncer y otras enfermedades.

Para leer más

Baena-Angulo C, Platero AI, Couso JP. Cis to Trans: small ORF functions emerging through evolution. *Trends in Genetics* 41 (2025) 119-131. DOI: 10.1016/j.tig.2024.10.012

Camarena ME, Theunissen P, Ruiz M, Ruiz-Orera J, Calvo-Serra B, Castelo R, Castro C, Sarobe P, Fortes P, Perera-Bel J, Albà MM. Microproteins encoded by noncanonical ORFs are a major source of tumor-specific antigens in a liver cancer patient meta-cohort. *Science Advances* 10 (2024) eadn3628. DOI: 10.1126/sciadv.adn3628

Chen LL, Kim VN. Small and Long non-coding RNAs: Past, present, and future. *Cell* 187 (2024) 6451-85. DOI: 10.1016/j.cell.2024.10.024

Ruiz-Orera J, Villanueva-Cañas JL, Albà MM. Evolution of new proteins from translated sORFs in long-non-coding RNAs. *Experimental Cell Research* 391 (2020) 111940. DOI: 10.1016/j.yexcr.2020.111940

Statello L, Guo C-J, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Review Molecular Cell Biology* 22 (2021) 96-118. <https://doi.org/10.1038/s41580-020-00315-9>

Unfried JP, Marín-Baquero M, Rivera-Calzada Á, Razquin N, [...], Llorca O, Lees-Miller SP, Fortes P. Long Noncoding RNA NIHCOLE Promotes Ligation Efficiency of DNA Double-Strand Breaks in Hepatocellular Carcinoma. *Cancer Research* 81 (2021) 4910-4925. DOI: 10.1158/0008-5472.CAN-21-0463