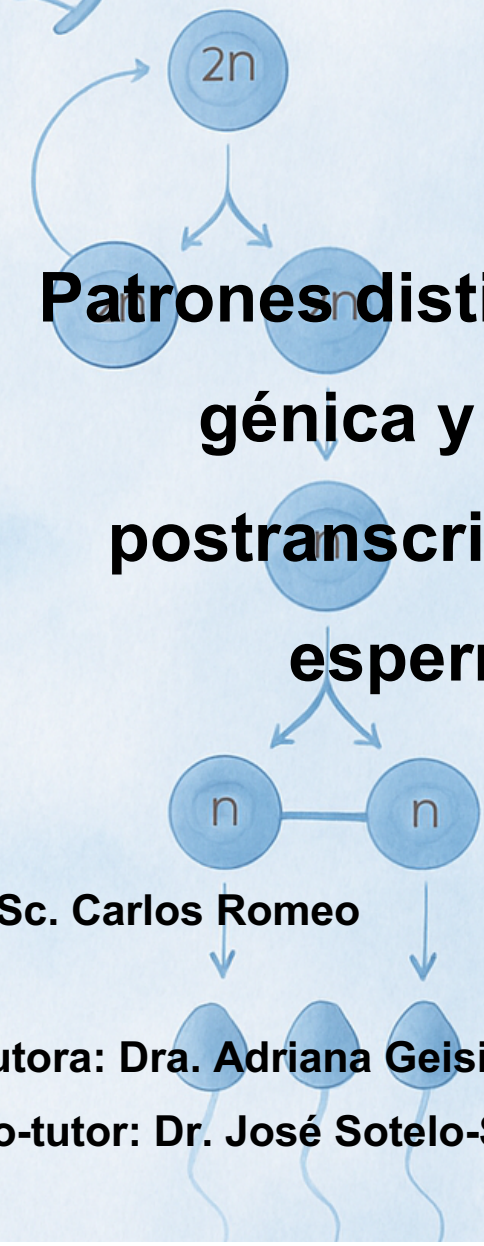
FACULTAD DE CIENCIAS
UDELAR fcien.edu.uy

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

PEDECIBA



Patrones distintivos de expresión génica y procesamiento postranscripcional durante la espermatogénesis

MSc. Carlos Romeo

Tutora: Dra. Adriana Geisinger

Co-tutor: Dr. José Sotelo-Silveira

mec MINISTERIO DE EDUCACIÓN Y CULTURA

IIBCE INSTITUTO DE INVESTIGACIONES BIOLÓGICAS CLEMENTE ESTABLE

cap comisión académica de posgrado universidad de la república

AGRADECIMIENTOS

A lo largo del camino recorrido durante la realización de esta tesis, muchas personas han contribuido de manera directa o indirecta con su apoyo, conocimientos y acompañamiento. A todas ellas, mi sincero agradecimiento.

En primer lugar, deseo agradecer especialmente a mis tutores, la Dra. Adriana Geisinger y el Dr. José Sotelo-Silveira, por su guía y confianza en mi trabajo. Su acompañamiento, tanto en lo académico como en la organización del proyecto, fue fundamental para que este trabajo llegara a buen puerto.

Agradezco también a los compañeros del área de genómica, por su predisposición a compartir conocimientos y su ayuda práctica en la interpretación y ejecución de análisis bioinformáticos. Gracias por enseñarme a utilizar herramientas, scripts y entornos que resultaron esenciales para el desarrollo de esta tesis.

De igual forma, agradezco a todos los integrantes del Departamento de Biología Molecular por el soporte brindado en los ensayos de biología molecular e inmunocitoquímica, así como por su disponibilidad para resolver dudas y aportar desde su experiencia.

Quiero agradecer a la Comisión Académica de Posgrado (CAP) por la beca de finalización de posgrado, de la cual fui beneficiario, y que me permitió dedicar el tiempo y la concentración necesarios para concluir este trabajo.

A mis amigos, por estar presentes en los momentos en que el ánimo flaquea y por ofrecer siempre un espacio de contención y desconexión tan necesario.

Finalmente, a mi familia, por su apoyo incondicional, paciencia y compañía a lo largo de todo este proceso.

A todos, muchas gracias.

RESUMEN

La espermatogénesis es un proceso tremendamente complejo, e involucra la división meiótica (durante la cual ocurren el apareamiento de los cromosomas homólogos y la recombinación o *crossing over*) y la diferenciación terminal hacia espermatozoides, o espermiogénesis. A nivel transcriptómico, durante la espermatogénesis se expresan el mayor número de genes tejido-específicos de entre todos los tejidos y procesos estudiados, el mayor número de ARNs no codificantes largos (lncRNAs), y una de las más elevadas tasas de *splicing* alternativo. A pesar de los avances en las tecnologías de secuenciación masiva y anotación genómica, persiste una fracción significativa del transcriptoma espermatogénico que no ha sido caracterizada, lo que sugiere la existencia de genes y transcriptos con funciones potencialmente relevantes en la biología reproductiva.

El presente trabajo se centró en la anotación de transcriptos y genes no anotados en el testículo del ratón adulto, un modelo ampliamente utilizado para estudiar la espermatogénesis. Se aplicaron enfoques bioinformáticos específicos para analizar datos de RNA-seq provenientes de cuatro poblaciones celulares testiculares representantes de etapas clave de la espermatogénesis en altísimo grado de pureza, obtenidas por citometría de flujo. Los datos fueron combinados con datos proteómicos para validar la expresión de proteínas derivadas de transcriptos no anotados.

A través del análisis se identificaron un total de 33.002 transcriptos no anotados, de los cuales 13.471 correspondieron a lncRNAs no reportados; es posible que, al menos una fracción de ellos, posea funciones regulatorias durante la espermatogénesis. Por otro lado, 2.794 transcriptos presentaron alto potencial codificante, lográndose confirmar 1.949 mediante análisis proteómico. Entre éstos, encontramos 22 genes no anotados que codifican 36 proteínas no reportadas previamente, así como 1.913 variantes de *splicing* codificantes para “nuevas” isoformas proteicas.

Los estudios de RT-PCR e inmunohistoquímica permitieron validar experimentalmente la expresión de variantes codificantes seleccionadas, como una isoforma no reportada de MSH5, que es una proteína esencial para la reparación de roturas de doble hebra y recombinación homóloga durante la meiosis. La localización

citoplasmática y patrón de expresión sugieren para esta isoforma, una función diferente de la de su contraparte canónica.

Un hallazgo importante es que las etapas de la profase meiótica temprana son particularmente ricas en transcriptos no anotados. Durante estas etapas tempranas, tienen lugar eventos únicos y fundamentales de la meiosis como el alineamiento y apareamiento de los cromosomas homólogos, cuyos fundamentos moleculares aún permanecen en gran parte desconocidos. La identificación de esta gran cantidad de transcriptos no anotados podría representar un paso significativo hacia el avance en la comprensión de estos procesos esenciales y de su regulación, proveyendo una fuente enorme de material para futuros estudios.

En conclusión, este estudio amplía el conocimiento del transcriptoma testicular, identificando variantes de *splicing* y genes no anotados que podrían desempeñar funciones críticas en la espermatogénesis. Estos hallazgos no sólo contribuyen a una mejor comprensión de la biología reproductiva, sino que también abren nuevas perspectivas para investigar mecanismos subyacentes a la infertilidad masculina y otros desórdenes reproductivos.

LISTA DE ABREVIATURAS

2C	Población celular con contenido 2C de ADN (espermatogonias y células somáticas)
A3SS	Sitio de *splicing* alternativo 3′ (*Alternative 3′ Splice Site*)
A5SS	Sitio de *splicing* alternativo 5′ (*Alternative 5′ Splice Site*)
D	Diploteno
eP	Paquiteno temprano (*early Pachytene*)
FDR	Tasa de descubrimiento falso (*False Discovery Rate*)
FPKM	Fragmentos por kilobase de transcripto por millón de lecturas
Ga	Genes anotados
Gn	Genes no anotados
GRCm38 / GRCm39	Ensamblajes de referencia del genoma del ratón
GTF	Formato de transferencia de genes (*Gene Transfer Format*)
HTSeq-counts	Software para conteo de lecturas por gen en RNA-seq
KEGG	Enciclopedia de genes y genomas de Kyoto (*Kyoto Encyclopedia of Genes and Genomes*)
L	Leptoteno
log2 FC	Logaritmo base 2 del cambio de expresión (*Fold Change*)
lncRNA	ARN largo no codificante (*long non-coding RNA*)
lP	Paquiteno tardío (*late Pachytene*)
LZ	Leptoteno/cigoteno (*Leptotene/Zygotene*: profase meiótica temprana)
mP	Paquiteno medio (*mid Pachytene*)
MIWI	Proteína PIWI-like 1, clave en células germinales
MXE	Exones mutuamente excluyentes (*Mutually Exclusive Exons*)
ORF	Marco abierto de lectura (*Open Reading Frame*)

2C	Población celular con contenido 2C de ADN (espermatogonias y células somáticas)
PS	Paquiteno (*Pachytene Spermatocytes*: espermatocitos en paquiteno)
PSI	Porcentaje de inclusión de exón (*Percent Spliced In*)
RI	Retención de intrón (*Retained Intron*)
RNA-seq	Secuenciación de ARN
rMATS	Software de análisis de splicing alternativo (*replicate Multivariate Analysis of Transcripts Spliginc*)
RS	Espermátidas redondas (*Round Spermatids*)
scRNA-seq	Secuenciación de ARNs de célula única (*Single-cell RNA-seq*)
SE	Exclusión o salto de exón (*Skipped Exon*)
Ta	Transcriptos anotados
Tn	Transcriptos no anotados
TPM	Transcriptos por millón
Z	Cigoteno

1. INTRODUCCIÓN

1.1. Espermatogénesis:

La espermatogénesis, proceso esencial en las especies con reproducción sexuada, representa la formación y maduración integral de los espermatozoides, células altamente diferenciadas y especializadas. Dicho proceso tiene lugar en los testículos. Durante este fenómeno biológico, las células germinales experimentan una diferenciación progresiva a partir de células precursoras, adquiriendo movilidad y la capacidad para fecundar un óvulo a medida que avanzan en su maduración. Además de su función primordial en la producción de espermatozoides, la espermatogénesis desempeña un papel esencial en la síntesis de hormonas esteroides como la testosterona, elementos fundamentales para el desarrollo y mantenimiento de las características sexuales secundarias en los individuos masculinos.

La espermatogénesis se inicia durante la pubertad y persiste a lo largo de toda la vida del individuo. Es importante resaltar que, aunque implica la producción continua de espermatozoides hasta la edad adulta, en el hombre el desarrollo completo de un espermatozoide requiere aproximadamente 74 días [1] y en el ratón, unos 28 días (Figura 1). Este proceso, altamente regulado, está influenciado por intrincados mecanismos hormonales y celulares. Por ejemplo, la hormona folículo estimulante (FSH) es necesaria para el inicio de la espermatogénesis, mientras que la testosterona desempeña un papel fundamental en el desarrollo y la función de los espermatozoides [2].

El proceso de espermatogénesis implica divisiones mitóticas de las espermatogonias (células precursoras de las células meióticas), diferenciación de las espermatogonias en espermatocitos, divisiones meióticas de los espermatocitos para producir espermátidas, y diferenciación de las espermátidas en espermatozoides [1].

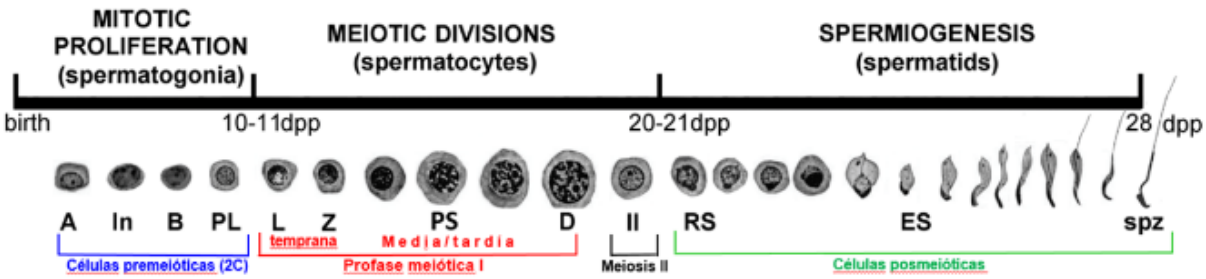


Figura 1. Espermatogénesis en el ratón. Las etapas emblemáticas se representan sobre la línea de tiempo, y su momento de aparición se expresa en días posparto (dpp). Células premeióticas (2C): espermatogonias A, intermedias (In), B y preleptoténicas (PL). Profase meiótica I temprana (LZ): leptoteno (L), cigoteno (Z). Profase meiótica I media/tardía: paquiteno (PS), diploteno (D). Espermatocitos II: II. Posmeiosis: espermátidas redondas (RS), elongadas (ES), espermatozoides (spz) (Figura tomada y modificada de Romeo *et al.,* 2024).

Una compleja serie de eventos moleculares coordina la espermatogénesis, que comprende la ejecución sincronizada de tres programas de expresión génica, correspondientes a los eventos mencionados más arriba. El primero de estos programas conduce a la proliferación de células madre espermatogoniales, las cuales se dividen varias veces mediante mitosis para dar lugar finalmente a los espermatocitos primarios que ingresarán en meiosis (segundo programa de expresión génica). Finalmente, el tercer programa corresponde a la espermiogénesis, durante la cual las espermátidas redondas, resultantes de la segunda división meiótica, se diferencian en espermatozoides [3,4] (Figura1).

Cualquier desviación de estos procesos puede desencadenar trastornos y, como resultado, conducir a la infertilidad. La investigación dedicada a la espermatogénesis es esencial para comprender a fondo los complejos procesos de producción y maduración de los espermatozoides, desempeñando un papel fundamental en la formulación de nuevas estrategias terapéuticas dirigidas a abordar la problemática de la infertilidad.

1.1.1. Complejidad celular

En el testículo, se encuentra una amplia diversidad de tipos celulares. Por un lado, se encuentran las células precursoras de los espermatozoides, que pasan por varios estadios de maduración hasta convertirse en espermatozoides maduros. Por otro, existe una compleja interacción con células circundantes que proporcionan soporte, facilitando esta especialización celular.

En los mamíferos, los túbulos seminíferos son la unidad funcional de los testículos y ocupan alrededor de dos tercios del volumen total del órgano. Estos túbulos están formados por una membrana basal, células de Sertoli (que proporcionan soporte estructural y funcional) y células germinales en diversas etapas de maduración [5]. Rodeando los túbulos seminíferos se encuentran las células mioides peritubulares, que brindan soporte al túbulo para facilitar el transporte de los espermatozoides en desarrollo [5].

En particular, las células de Sertoli son cruciales para el desarrollo exitoso de la espermatogenésis, debido a que regulan variados procesos [6]: poseen ramificaciones citoplasmáticas que envuelven y protegen a las células germinales, actuando como nodrizas y creando un microambiente en donde se puede desarrollar la espermatogénesis, actúan como macrófagos eliminando los cuerpos residuales de las espermátidas [5], y forman la barrera hematotesticular, que cumple una función clave en la protección inmunológica de las células de la línea germinal [7].

Por otra parte, los túbulos seminíferos están inmersos en un estroma que posee una diversidad de tipos celulares somáticos diferentes, entre los que destacan las células de Leydig. Los tipos celulares somáticos del testículo, en conjunto, desempeñan un papel crucial en la regulación parácrina y autócrina, así como en la producción de hormonas como la testosterona (producida por las células de Leydig), estrógenos, factor 3 tipo insulina (*Insulin-like factor 3*) y oxitocina, junto con la hormona antimülleriana, entre otras [5]. Un ejemplo de la regulación endócrina de la espermatogénesis es la estimulación tanto por parte de la hormona folículo estimulante (FSH) como de la hormona luteinizante (LH), las cuales actúan a través de la testosterona producida por las células de Leydig. Sin embargo, las células de la

línea germinal no poseen receptores ni para LH ni para testosterona. Tanto estas señales hormonales, como las de FSH en la vida adulta, son transmitidas mediante las células de Sertoli y las células peritubulares, a través de la producción de señales específicas, lo que habla, a las claras, del importante diálogo cruzado existente entre los distintos tipos celulares [6,8].

El entorno particular que se produce dentro de los túbulos seminíferos es tan complejo que, al día de hoy, no se ha logrado replicar con éxito [9], lo cual imposibilita la realización de estudios *in vitro*. Esto representa una importante limitación para efectuar estudios de expresión génica y funcionales sobre la espermatogénesis, ya que requiere que dichos estudios sean efectuados *in vivo*.

1.1.2. Mitosis en el marco de la espermatogénesis

La población de espermatogonias en los mamíferos se origina a partir de las células germinales conocidas como gonocitos. Estas células primero experimentan un período de multiplicación en la zona más externa de los túbulos seminíferos y luego migran hacia el interior a medida que se van diferenciando, donde se dividen para formar, en el ratón, espermatogonias de tipo A. Estas células van pasando por estadios intermedios, hasta culminar en las de tipo B [10]. Finalmente, las espermatogonias de tipo B se transforman en espermatocitos pre-leptoténicos e ingresan en el proceso de meiosis [10].

1.1.3. Meiosis

La meiosis es un proceso especializado de división celular que ocurre en las células precursoras de las células germinales (ovocitos y espermatocitos). Este proceso es crucial para la formación de los gametos (óvulos y espermatozoides).

La meiosis consta de dos divisiones celulares consecutivas: meiosis I y meiosis II. Durante la meiosis I, las células germinales experimentan una división para dar lugar a dos células hijas, cada una con la mitad de los cromosomas de la célula original. Esta división, conocida como división reduccional, es necesaria para reducir el

número de cromosomas en las células reproductoras y mantener la constancia del número cromosómico de generación en generación, disminuyendo el contenido de ADN de 4C (número cromosómico 2N), a 2C (1N). El proceso de meiosis I se encuentra subdividido en: profase I, la fase más larga y compleja, subdividida en cinco etapas, y en la cual profundizaremos más adelante; metafase I, durante la que los pares de cromosomas homólogos unidos se alinean en la placa metafásica de la célula y los microtúbulos se unen a los centrómeros de cada cromosoma homólogo, estableciendo la orientación para la siguiente fase; anafase I, en la que los cromosomas homólogos (cada uno compuesto por dos cromátidas hermanas) se separan y se mueven hacia polos opuestos de la célula, reduciéndose así el número de cromosomas a la mitad y asegurando que cada célula hija reciba sólo un cromosoma de cada par homólogo; telofase I, durante la cual los cromosomas alcanzan los polos opuestos y finalmente la citocinesis, en la que el citoplasma se divide, formándose dos células hijas haploides. Cada célula contiene un conjunto único de cromosomas recombinados, aún en forma de cromátidas hermanas unidas.

En la meiosis II, las células hijas se dividen nuevamente para producir en total cuatro células hijas, cada una con la mitad de los cromosomas de la célula original. Cada célula hija recibe una copia de cada cromosoma, reduciendo así el contenido de ADN a 1C en las espermátidas. Durante la fecundación, al unirse el espermatozoide con el ovocito, se restablece el número cromosómico de la especie [11,12] (Figura 2).

Dentro de la meiosis, se produce un proceso conocido como entrecruzamiento o "*crossing-over*", que implica el intercambio de segmentos de cromosomas homólogos durante la meiosis I. Este proceso es responsable de la variabilidad genética en la progenie, permitiendo que los individuos de una especie se diferencien genéticamente entre sí [13].

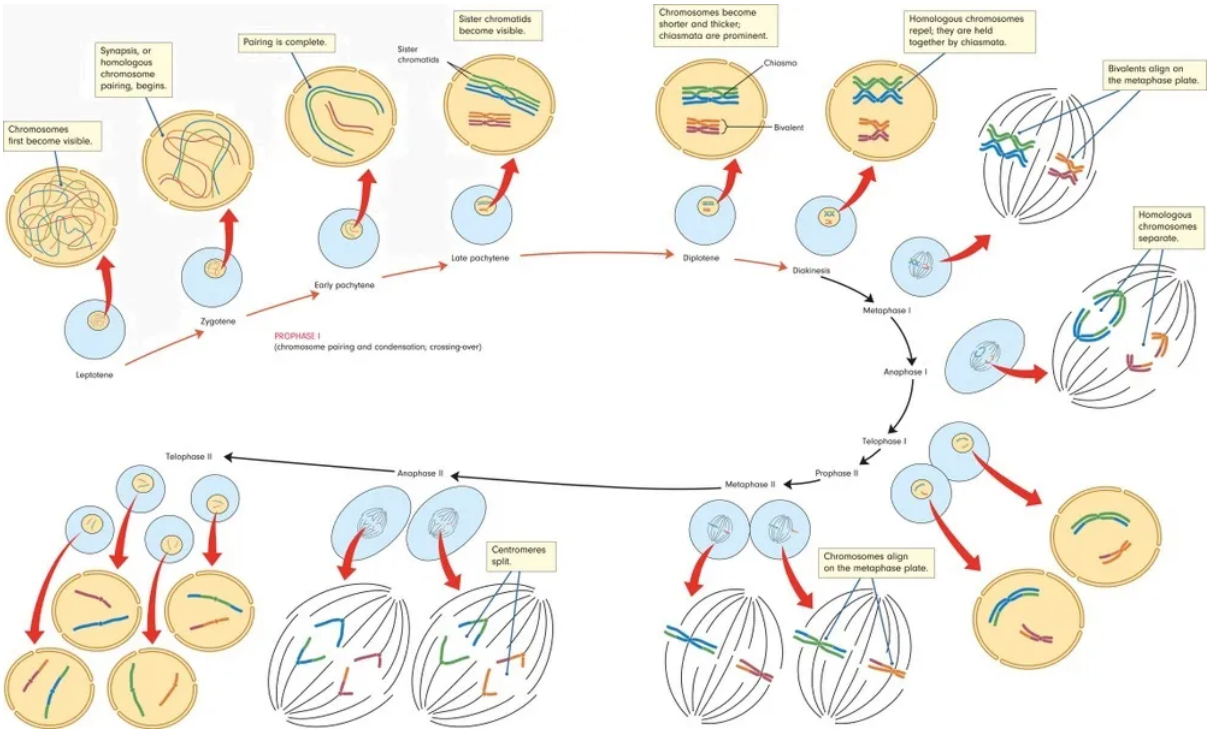


Figura 2. **Etapas de la meiosis I**. La meiosis I se divide en las siguientes fases: **profase I**: Fase más larga, subdividida en cinco etapas **(leptoteno, cigoteno, paquiteno, diploteno, diacinesis); metafase I**: Los cromosomas homólogos unidos se alinean en la placa metafásica de la célula; **anafase I**: Los cromosomas homólogos se separan y se mueven hacia polos opuestos de la célula; **telofase I y citocinesis**: Los cromosomas alcanzan los polos opuestos, y el citoplasma se divide (Figura tomada de https://www.elgencurioso.com/diccionario/profase-i/).

1.1.4. Profase de la Meiosis I

La profase I asume una importancia particular en esta investigación, y por ese motivo será abordada en detalle. Un componente crucial que define a la profase I es el complejo sinaptonémico (CS), una macroestructura proteica que facilita el alineamiento entre cromosomas homólogos. Esta estructura se compone de dos elementos laterales (ELs), un elemento central (EC) y filamentos transversos que conectan los ELs con el EC, otorgando al CS una apariencia escaleriforme distintiva [14–16]. La profase I, notablemente extensa, se caracteriza por eventos cruciales y se

subdivide en etapas distintas: leptoteno, cigoteno, paquiteno, diploteno y diacinesis, cada una marcada por acontecimientos significativos en el proceso de la meiosis (Figura 2).

Durante la etapa de leptoteno los cromosomas experimentan una condensación, manifestándose como filamentos delgados y alargados. En este período cada cromosoma consiste en dos cromátidas hermanas, réplicas idénticas del ADN previamente duplicado durante la interfase. Se forma la estructura de los elementos axiales (EAs), que posteriormente darán origen a los ELs del CS. Además, los telómeros se vinculan a la envoltura nuclear [17] y se desplazan sobre ella, adoptando una disposición característica denominada "*bouquet*" [18]. Este arreglo posiblemente facilite la aproximación de los cromosomas homólogos.

En el estadio de cigoteno comienza el proceso crucial de apareamiento de los cromosomas homólogos, donde se observa la coexistencia de EAs con segmentos ya ensamblados del CS [19]. La formación adecuada del CS es imprescindible para una segregación apropiada de los cromosomas homólogos durante la anafase I [20]. Durante las fases de leptoteno/cigoteno, los nódulos tempranos de recombinación, marcadores que indican los sitios relacionados con el intercambio de hebras [21,22], se hacen visibles.

El término paquiteno, de origen griego, que significa "hilos gruesos", se refiere a los cromosomas homólogos apareados que, tras su duplicación, forman tétradas. En este estadio, los CSs se encuentran completamente ensamblados de un extremo cromosómico al otro, permitiendo la ocurrencia de la recombinación (entrecruzamiento) entre cromosomas homólogos [11]. A medida que la profase avanza hacia el paquiteno, los nódulos de recombinación tardía, menos numerosos que los nódulos tempranos, se hacen evidentes, señalando los sitios donde realmente ocurre la recombinación genética [21,22].

En la etapa de diploteno, coincidente con el desensamblaje del CS, los cromosomas homólogos permanecen conectados sólo por los quiasmas, que representan vestigios de los puntos de recombinación [11,21,22]. Por último, la diacinesis representa la

migración de los cromosomas hacia la periferia del núcleo, la desaparición de los quiasmas, la formación del huso y la disolución de la envoltura nuclear (Figura 2).

El apareamiento y la recombinación de los cromosomas homólogos son fundamentales para garantizar la correcta segregación de los cromosomas a las células hijas [23]. Por otra parte, en la meiosis existen puntos de control, conocidos como "*checkpoints*", que supervisan todo el proceso, siendo uno de los más importantes el punto de control de la salida del paquiteno. Las alteraciones en el apareamiento o la recombinación de los cromosomas homólogos activan este punto de control, lo que provoca un arresto meiótico que puede resultar en patologías, incluyendo infertilidad [24,25].

Es importante señalar que la duración de cada una de las distintas etapas de la profase es muy diferente. En tanto el leptoteno y el cigoteno son etapas muy breves, el paquiteno abarca la mayor parte de la profase meiótica, y esto es así en todas las especies estudiadas. Por ejemplo, en el ratón, en tanto el leptoteno y cigoteno duran, en conjunto, poco más de un día, el paquiteno tiene una duración aproximada de 7 días (ver Figura 1).

1.1.5. Espermiogénesis

La espermiogénesis, que prosigue a la meiosis, es un proceso que culmina en la formación de espermatozoides. Los espermatozoides son células muy pequeñas y únicas, compuestas por una cabeza que contiene el material genético en su núcleo, y una cola que les proporciona la capacidad de moverse. Los espermatozoides, una vez producidos, son almacenados en el epidídimo, en donde maduran, hasta su eyaculación [26,27].

Una vez iniciada la espermiogénesis, se producen cambios citológicos profundos que transforman paulatinamente a las células resultantes de la meiosis II (espermátidas redondas) en espermatozoides. Destacan la aparición de un flagelo, cambios en la composición proteica y la ultraestructura de las mitocondrias, que terminan congregándose alrededor de la pieza media del flagelo [28], así como la modificación y reubicación del aparato de Golgi para formar el acrosoma [5] (Figura 3A), y el descarte

de gran parte del citoplasma, que es fagocitado por las células de Sertoli [6]. Además, el núcleo experimenta una enorme reducción de tamaño y su remodelación hacia una estructura condensada característica de los espermatozoides. Esta gran reducción del tamaño nuclear es posible gracias a una enorme compactación de la cromatina, lo que favorece la hidrodinámica y protege al material genético durante el viaje que emprenden los espermatozoides en busca del óvulo, pero es incompatible con la transcripción [29].

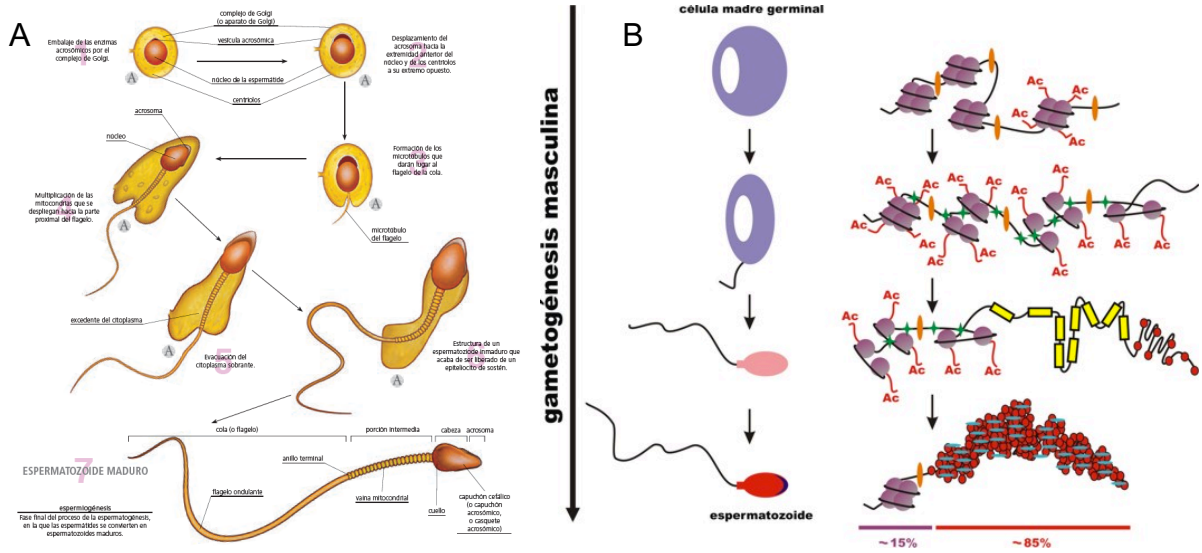


Figura 3. **Proceso de espermiogénesis**. Este proceso incluye varios cambios morfológicos y estructurales fundamentales. **A)** Cambios estructurales: Formación del acrosoma; condensación del núcleo; desarrollo del flagelo; eliminación del citoplasma; maduración final. Al completarse la formación del acrosoma, la condensación nuclear y el desarrollo del flagelo, el espermatozoide maduro presenta una estructura adaptada para la movilidad y la fertilización (Imagen tomada de https://gravida.com/blog/como-se-forman-los-espermatozoides/). **B)** Remodelación de la cromatina: A medida que las células germinales masculinas se diferencian desde células madre germinales hasta espermatozoides maduros, experimentan una transformación drástica en la estructura de su cromatina. Se observan cambios en las histonas, que se acetilan aumentando la expresión de genes que serán usados en etapas posteriores. A continuación, se da una sustitución gradual de histonas por proteínas de transición (en amarillo) que ayudan en la condensación de la cromatina. Finalmente, en las etapas más avanzadas, las histonas y proteínas de transición son reemplazadas casi completamente por protaminas (en rojo), lo que permite un empaquetamiento mucho más denso de la cromatina en los espermatozoides maduros.

Como parte de la remodelación nuclear, hacia el final de la espermiogénesis se genera un reemplazo secuencial de las histonas que empaquetan el ADN genómico, primero por proteínas de transición y luego por protaminas, proteínas exclusivas de la espermiogénesis [30,31] (Figura 3B). De esta forma, el ADN es empaquetado en una estructura superenrollada llamada toroide, aumentando su protección y generando, como consecuencia, el silenciamiento de la transcripción [5,30]. Debido a este silenciamiento, se requiere una transcripción temprana de ARNs cuyos productos serán necesarios en las espermátidas elongadas y espermatozoides, que deben producirse y almacenarse antes de su uso, lo que, a su vez, conlleva un complejo sistema de regulación traduccional. Un ejemplo de ello son las propias protaminas: en los ratones, los ARNm de las protaminas se sintetizan al menos siete días antes de su traducción y se almacenan en mRNPs (ribunucleoproteínas de unión a ARNm); la regulación precisa de su traducción, mediada por proteínas específicas, es clave, ya que alteraciones en el momento de la traducción de las protaminas suelen producir esterilidad [31,32]. Un mecanismo importante de regulación traduccional en estas células, entre otros, se ha relacionado con la regulación de la longitud de la cola poli(A) de ciertos ARNm [30]. Asimismo, se ha sugerido que el cuerpo cromatoide ("*chromatoid body*"), un gránulo distintivo de las espermátidas [33,34], desempeña un papel crucial en la regulación del ARN en las etapas posmeióticas y en la determinación de la fertilidad masculina [34–36], ya que el mismo alberga proteínas de unión a ARN o implicadas en vías de procesamiento del mismo, junto con ARNm y ARNs regulatorios como piRNAs, miRNAs [33] y ARNs no codificantes largos (lncRNAs), aunque estos últimos aún han sido escasamente caracterizados [37].

1.2. ARNs no codificantes, y procesamiento alternativo

1.2.1. Generalidades del ARN no codificante

Alrededor de dos tercios del genoma humano experimentan una transcripción generalizada, mientras que sólo aproximadamente el 2% se traduce en proteínas [38]. Los transcriptos que no resultan en la síntesis de proteínas son denominados ARNs no codificantes ("*noncoding RNAs*": ncRNAs), los cuales pueden clasificarse en ARNs

no codificantes cortos ("*small noncoding RNAs*": sncRNAs) y ARNs no codificantes largos ("*long noncoding RNAs*": lncRNAs). Entre los sncRNAs se encuentran los que interactúan con la proteína Piwi (piRNAs), que en los mamíferos están expresados casi exclusivamente en la línea germinal masculina [39], y los micro ARNs (miRNAs) que, si bien se expresan en diversos tejidos, desempeñan un papel importante durante la espermatogénesis [5].

Dado que los sncRNAs en la espermatogénesis han sido bastante estudiados, los mismos no se abordaron en este trabajo, razón por la cual no profundizaremos en su descripción. Por el contrario, sí nos dedicamos a la caracterización de los lncRNAs, de los cuales el conocimiento es aún bastante más escaso.

1.2.2. ARNs no codificantes largos (lncRNAs)

Los lncRNAs han recibido menos atención que los sncRNAs en la investigación, y sólo en los últimos años han empezado a ser estudiados más a fondo debido a la creciente comprensión de su relevancia funcional. Por oposición a sncRNAs, los lncRNAs se definen como aquellos ARNs no codificantes con una longitud superior a 200 pares de bases (pb). Este límite de 200 pb surgió debido a consideraciones experimentales, ya que los protocolos iniciales de purificación y secuenciación de ARN excluían a los ARNs no codificantes pequeños, lo que llevó a categorizar como lncRNAs a aquellos ARNs no codificantes con una longitud mayor [40–43].

Sin embargo, la premisa de que los lncRNAs no generan proteínas funcionales se ha cuestionado recientemente. Estudios de *ribosome-profiling* han demostrado que algunos lncRNAs tienen potencial para ser traducidos, y se ha observado la presencia de ribosomas asociados a ellos [44,45]. Se estima que aproximadamente el 40% de los lncRNAs en las células humanas podría codificar péptidos de más de 10 aminoácidos (aa), y alrededor de un tercio de los lncRNAs en células madre espermatogénicas de ratón podría contener al menos un marco abierto de lectura (ORF) [44,46]. Por lo tanto, los lncRNAs también se han definido como moléculas con un potencial codificante menor a 100 aa [46,47]. Se ha demostrado que algunos de estos pequeños péptidos tienen funciones específicas: por ejemplo, el péptido DWORF en ratones contrarresta los efectos de los inhibidores de la ATPasa Ca2+ del retículo sarcoplásmico (SERCA),

mientras que el péptido Spar regula negativamente la actividad de la proteína diana de la rapamicina mTORC1 [44]. Por lo tanto, algunos investigadores sugieren considerar a los ncRNAs como aquellos transcriptos que es "poco probable" que codifiquen proteínas funcionales [46,48]. Además, un mismo ARN puede contener o generar un transcripto codificante de proteínas y uno no codificante, lo que añade otra capa de complejidad a la definición [49].

Muchas características de los lncRNAs son consistentes en todos los vertebrados en los que se han estudiado hasta ahora: suelen ser relativamente cortos (en comparación con los ARNm), con un bajo número de exones (aunque estos exones suelen ser más largos que los de los ARNm), tener una baja conservación de secuencia, ser poco abundantes, y mostrar una expresión altamente restringida en el espacio y el tiempo [45,48,50–55]. Se cree que su bajo contenido de GC puede contribuir, en parte, a sus niveles generalmente menores de expresión en comparación con los ARNm [56]. Además, los lncRNAs suelen experimentar procesamiento alternativo por corte y empalme (*splicing*), aunque este proceso parece ser menos eficiente que en los ARNm [48,57]. En términos de patrones de expresión, la gran mayoría de los lncRNAs exhibe una expresión específica de tejido, con un gran número de ellos siendo específicos del testículo, como veremos más adelante [50,58,59].

Un tipo particular de lncRNAs en el genoma de los mamíferos proviene de elementos repetidos como los retrotransposones, que constituyen entre el 30% y el 50% del genoma. Estos elementos muestran actividad transcripcional, exhiben especificidad tisular, y están asociados a transcriptos codificantes de proteínas. Esto sugiere que podrían estar involucrados en la regulación de la expresión génica, ya sea como reguladores en *cis*, o como precursores de sncRNAs de doble hebra [40,60,61]. Por ejemplo, se ha demostrado que, en el humano y el ratón, los elementos transponibles LINE1 son esenciales para inducir la condensación de la cromatina, y se ha identificado el ARN LINE1 como un lncRNA (“*lncRNA-like*”) que participa en la renovación de las células madre embrionarias y en el desarrollo preimplantatorio del embrión [61,62]. Por otro lado, los pseudogenes también pueden actuar como lncRNAs al ser transcriptos [40].

Como hemos mencionado, los lncRNAs funcionales no muestran una alta conservación de secuencia primaria, lo que sugiere que evolucionan más rápidamente que los ARNs codificantes para proteínas, que están, en general, más altamente conservados a lo largo de la evolución de los vertebrados [48,63]. Por ejemplo, sólo el 72% de los lncRNAs intergénicos humanos se expresan también en los monos macacos, en comparación con un nivel de conservación del 98% en todos los primates para los genes codificantes de proteínas [64]. Es notable que, en un trabajo, se encontró que un grupo de lncRNAs conservados, expresados en células madre embrionarias humanas, se procesaba de manera diferente que sus homólogos en ratones, lo que resulta en una localización subcelular y una función diferentes en ambas especies. Dentro de este grupo, se identificaron 122 lncRNAs con secuencia conservada y 229 con posición conservada (sinténicos) entre ambas especies de mamíferos, que eran principalmente retenidos en el núcleo en ratones, pero procesados y exportados preferentemente al citoplasma en humanos [65].

Debido a su baja conservación en secuencia y bajos niveles de expresión, se cree que los lncRNAs pueden actuar en forma colectiva, o, incluso, que el simple evento de su transcripción podría tener un efecto funcional. Por ejemplo, en el ratón, el lncRNA *Airn* actúa como antisentido al promotor del receptor de insulina *Igf2r* y su transcripción es crucial para el silenciamiento del receptor, más allá de la función del propio transcripto [48,66]. De todos modos, a pesar de la falta de presiones de selección en su secuencia primaria, los lncRNAs pueden estar sujetos a restricciones en la estructura génica, los promotores y los patrones de *splicing* [67]. Además, aunque los lncRNAs tienden a no conservar toda su secuencia genética, es posible encontrar fragmentos cortos con secuencia conservada, especialmente en el extremo 5' [43,61,68]. Asimismo, algunos lncRNAs pueden contener elementos conservados como sitios de unión a factores de transcripción, motivos de *splicing* y señales de localización nuclear [57,61,69].

Con respecto a su modo de acción, los lncRNAs tienen la capacidad de unirse a ácidos nucleicos, interactuar directamente con proteínas, o facilitar la unión de proteínas a sus objetivos en el ADN o ARN. A diferencia de los sncRNAs, los lncRNAs pueden plegarse y formar estructuras secundarias que faciliten estas interacciones ARN-proteínas o ARN-ADN [48,70]. En términos generales, pueden actuar de diversas

formas: como competidores/inhibidores, activadores/reclutadores, precursores o potenciadores [49,59,71]. Como competidores, pueden unirse a proteínas de unión al ADN y evitar su acción, o pueden evitar la unión de miRNAs a ARNm, evitando su degradación posterior (ejemplos: *Xist*, *PANDA*, *RMST*). Como activadores, pueden reclutar modificadores de la cromatina y activarlos (como es el caso de *HOTAIR* o *Evf2*). Como precursores, pueden generar sncRNAs a través de su procesamiento por ARNasas (por ejemplo, *H19*, *HongrES2*). Como potenciadores, pueden interactuar en el contacto promotor-potenciador e inhibir la transcripción de genes codificantes [49,71]. Además, los promotores de lncRNAs también pueden actuar como potenciadores de genes cercanos [72].

1.2.3. ***Splicing*** **alternativo**

El procesamiento ("*splicing"*) alternativo contribuye de manera significativa a la complejidad de los transcriptomas en los organismos eucariotas multicelulares. Esta complejidad se refleja principalmente a nivel de las proteínas, y desempeña un papel fundamental en la fisiología y patología de estos organismos. Esta idea no sólo se basa en análisis genómicos completos, sino también en estudios detallados que han revelado que las modificaciones globales y específicas en el *splicing* regulan una amplia gama de procesos, incluida la diferenciación celular específica de tejidos y especies [73].

El *splicing* alternativo del ARN es un mecanismo que permite la producción de diferentes variantes de proteínas a partir de un solo gen. Este proceso ocurre cuando distintas combinaciones de exones son unidas durante el procesamiento del ARNm, generando así diversas isoformas del ARN [74].

Los ARNs y polipéptidos generados a partir de un solo gen mediante *splicing* alternativo suelen mostrar similitudes, pero no son idénticos, presentando regiones conservadas y divergentes que resultan en variaciones sutiles o radicales, tanto a nivel de ARNm como de proteínas [73]. Es importante destacar que el proceso de *splicing* no se limita a los genes que codifican ARNm, sino que también ocurre en ARNs no codificantes. Además, en el caso de los ARNm, el primer AUG, que es el

primer codón de un marco de lectura abierto traducible, no siempre coincide con el inicio del primer exón, pudiendo estar ubicado más adelante dentro del primer o incluso del segundo exón. Esta situación también se aplica a codones de inicio distintos de AUG, los cuales han demostrado ser más frecuentes de lo esperado [73]. Estos hallazgos subrayan la complejidad del proceso de *splicing*, y sugieren que la afirmación simplificada de que los exones codifican proteínas y los intrones no, es demasiado generalizada.

El proceso de *splicing* de los pre-ARNm transcritos por la ARN polimerasa II (Pol II) en los eucariotas es llevado a cabo por el espliceosoma, un complejo que se ensambla a lo largo de cada intrón del pre-ARNm utilizando pequeñas ribonucleoproteínas nucleares y un conjunto de proteínas auxiliares. Los componentes del espliceosoma reconocen y se unen a secuencias consenso ubicadas en los extremos 5' y 3' de cada intrón (el sitio de empalme 5' y el sitio de empalme 3') y catalizan dos reacciones de transesterificación consecutivas, que resultan en la eliminación del intrón y la unión covalente de los exones adyacentes (Figura 4A). Los sitios de empalme considerados "fuertes" (es decir, aquellos que se asemejan más a la secuencia consenso) son reconocidos de manera más eficiente por los componentes del empalme, lo que generalmente los coloca en posición de liderazgo en el proceso de *splicing* [73,75]. Cuando un sitio de empalme es considerado "débil" o subóptimo, es común que ocurra un *splicing* alternativo, donde el sitio débil se utiliza parcialmente en las reacciones de eliminación de intrones y unión de exones, lo que resulta en la generación de diferentes proporciones de variantes del ARNm maduro.

La elección de los sitios de empalme no sólo está determinada por sus secuencias intrínsecas, ya sean óptimas o subóptimas, que actúan en *cis*, sino también por la influencia de proteínas que actúan en *trans*, conocidas como factores de empalme. Estos factores reconocen y se unen a secuencias específicas presentes en exones o intrones, conocidas como potenciadores y silenciadores de empalme exónicos y potenciadores y silenciadores de empalme intrónicos. Adicionalmente, el *splicing* alternativo no sólo está regulado por la interacción de estos factores de empalme en *trans* con secuencias reguladoras en *cis* presentes en el pre-ARNm, sino que también

está vinculado a la transcripción por parte de la Pol II. De hecho, los factores de empalme son reclutados de manera co-transcripcional en el pre- ARNm [73,75].

Muchas proteínas reguladoras de unión al ARN funcionan de manera específica para células, tejidos o condiciones particulares, y tienen la capacidad de coordinar la regulación de "redes" funcionalmente coherentes de exones e intrones [76,77]. Por lo tanto, nuestra comprensión de los repertorios de variantes de *splicing* detectadas, así como de otras formas de variación en la transcripción en diversas condiciones celulares, y en el contexto de la fisiología normal y de la enfermedad, sigue aumentando de manera significativa [78].

El gen *CALCA* ofrece un excelente ejemplo de cómo el *splicing* alternativo y la poliadenilación alternativa pueden dar lugar a la expresión diferencial de proteínas a partir de un solo gen. Este gen de vertebrados se expresa en las neuronas del hipotálamo y en las células parafoliculares, también conocidas como células C, de la glándula tiroides, y codifica dos proteínas distintas: la calcitonina y el péptido 1 relacionado con el gen de la calcitonina (CGRP1). A través de una combinación de *splicing* alternativo y procesos de escisión y poliadenilación alternativos, se producen dos variantes de ARNm: una para CGRP1 en las neuronas, y otra para calcitonina en las células C de la tiroides. A pesar de que ambos ARNm comparten segmentos de secuencia codificante, generan dos proteínas maduras, con secuencias de aminoácidos completamente diferentes. Por lo tanto, aunque comparten el mismo gen, calcitonina y CGRP1 son proteínas distintas en términos de secuencia de aminoácidos, estructura y función. Esto justifica por qué la proteína neuronal se ha denominado "péptido relacionado con el gen de la calcitonina" en lugar de "péptido relacionado con la calcitonina" [73].

Existen varios mecanismos que pueden dar lugar a las variantes de *splicing* del ARN. Algunos ejemplos incluyen (Figura 4B):

a. Exclusión o salto (“*skipping*”) de exones: Este mecanismo implica la eliminación de un exón específico durante el proceso de *splicing*, dando lugar a una variante de transcripto en la que el exón no está presente [74,79].

b. Exones mutuamente excluyentes: En este mecanismo, dos exones son alternativos y excluyentes, es decir, sólo uno de los dos es incluido en el transcripto final, pero nunca ambos simultáneamente [74,79].
c. Uso alternativo de sitios de inicio y/o finalización de *splicing*: En este caso, el uso de un sitio de *splicing* alternativo en el extremo 5' o 3' del exón, acorta o alarga la región del exón incluida en el transcripto final [79,80].
d. Retención de intrón: Un intrón que normalmente sería eliminado se retiene en el transcripto final, generando una variante que incluye esta secuencia intrónica [79].

En todos los casos mencionados, el resultado puede ser la producción de proteínas diferentes. Estas variantes de *splicing* pueden ocasionar diferencias en la estructura, función y nivel de expresión de las proteínas resultantes, y pueden desempeñar un papel crucial en la variabilidad fenotípica y en la patogénesis de enfermedades genéticas [81]. Más aún: la regulación del *splicing* alternativo es tan relevante como la regulación de la transcripción en la configuración de los fenotipos y la fisiología celular, al menos en los organismos eucariotas multicelulares [73].

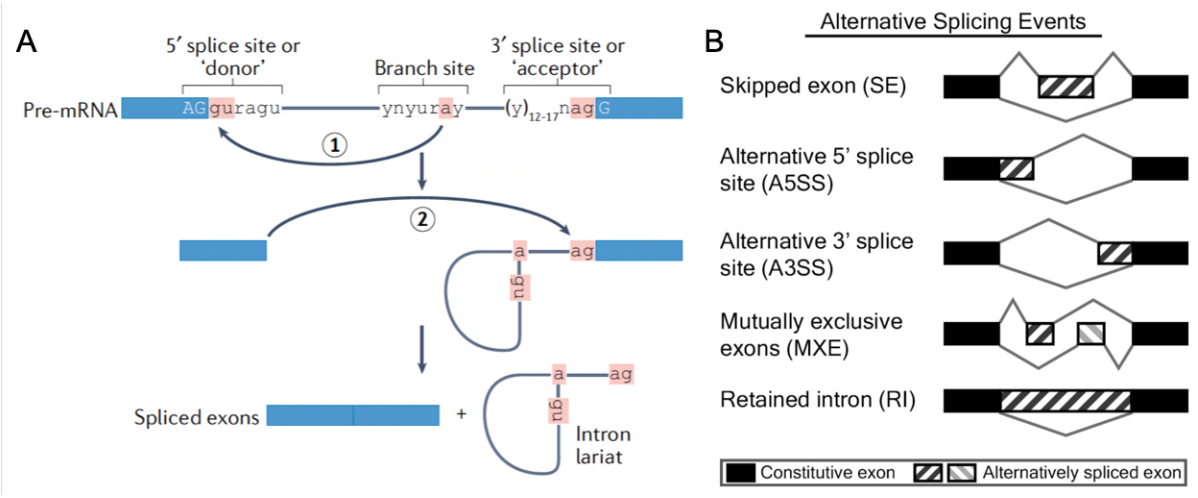


Figura 4. Proceso y eventos de *splicing* alternativo. A) Representación esquemática de la reacción de corte y empalme (*splicing*). La reacción ocurre en dos pasos, mostrando un segmento de pre-ARNm etiquetado con las secuencias consenso del sitio de corte y empalme "donador" (sitio de corte 5′), "aceptor" (sitio de corte 3′), y del sitio de ramificación (Figura tomada y modificada de Marasco & Kornblihtt, 2023). **B)** Esquema de los eventos de *splicing* alternativo. Se representan los principales tipos de eventos de *splicing* alternativo. **Skipped exon (SE)**:

exclusión o salto de exón**; Alternative 5' splice site (A5SS)**: utilización de un sitio de *splicing* alternativo en el extremo 5' del exón; **Alternative 3' splice site (A3SS)**: uso de un sitio de *splicing* alternativo en el extremo 3' del exón; **Mutually exclusive exons (MXE)**: exones mutuamente excluyentes; **Retained intron (RI)**: retención de intrón. Los exones constitutivos se muestran en negro, mientras que los exones o intrones sujetos a *splicing* alternativo se indican en blanco con líneas diagonales (Imagen tomada de https://rnaseq-mats.sourceforge.io/).

1.3. Estudios de expresión génica de la espermatogénesis de los mamíferos

Los estudios moleculares de la espermatogénesis suelen combinar diversas técnicas para obtener una comprensión completa de los procesos celulares y moleculares que regulan el desarrollo de los espermatozoides, así como para identificar posibles alteraciones asociadas con la infertilidad.

Uno de los enfoques comunes en estos estudios es el análisis de la expresión génica, para identificar los cambios asociados con la maduración de los espermatozoides y problemas relacionados con la espermatogénesis. Varios laboratorios han llevado a cabo estudios masivos de expresión génica durante la espermatogénesis de los mamíferos, inicialmente mediante el uso de microarreglos [82–86] y, en la última década, empleando métodos de secuenciación masiva [58,87–91]. Los modelos más empleados han sido ratón y humano [91].

Además de los mecanismos de expresión génica y procesamiento postranscripcional, los factores epigenéticos juegan un papel crucial en la espermatogénesis. Aunque no nos ocuparemos de ellos, es importante mencionarlos, aunque sea muy brevemente, debido a su relevancia y a la complejidad que añaden al estudio de este proceso. Un evento esencial en la espermatogénesis es la reprogramación epigenética, que incluye una desmetilación generalizada del ADN, seguida por un proceso de metilación de *novo*. Esta reprogramación ocurre tanto durante el desarrollo de las gónadas como en la espermatogénesis, estableciendo un patrón de hipometilación del ADN en la línea germinal masculina [41]. Se ha demostrado que alteraciones en

estos procesos epigenéticos, como la hipermetilación del ADN en ciertas secuencias, pueden estar relacionadas con una calidad deficiente de los espermatozoides y problemas de fertilidad [92]. Sin embargo, la relación exacta entre la metilación del ADN y la fertilidad sigue siendo un tema de discusión [5]. Además, dado que algunas modificaciones epigenéticas pueden heredarse, ha surgido preocupación sobre la posible relación entre las técnicas de reproducción asistida y ciertas patologías asociadas a la impronta genómica, como los síndromes de Prader-Willi, Beckwith-Wiedemann y Angelman [5].

1.3.1. Metodologías de purificación de células espermatogénicas para estudios de expresión génica

A pesar de los avances en el cultivo de células de la línea germinal masculina, hasta la fecha lo máximo que se ha logrado es producir células parecidas a espermátidas funcionales capaces de fecundar un óvulo en un estudio [93]. Sin embargo, estas células no poseen todas las características de las espermátidas, y no pueden diferenciarse en espermatozoides [93,94], así como tampoco presentan puntos característicos de control de la espermatogénesis [95]. Estas limitaciones subrayan la necesidad de trabajar con el organismo entero para comprender completamente el proceso.

A la actual incapacidad para cultivar eficazmente las células que componen la línea germinal masculina y reproducir de manera fiable los procesos biológicos, lo que obliga a trabajar *in vivo*, se añade la complejidad de la coexistencia en el testículo de los mamíferos de más de 30 tipos celulares diferentes, incluyendo células somáticas y germinales. Esta heterogeneidad celular se ve acentuada por la disparidad en la representación de los distintos tipos celulares, con algunos, como las espermatogonias o las células en profase meiótica temprana, teniendo una presencia mucho menor en comparación con otras, como los espermatocitos paquiténicos y las espermátidas [96]. Esta disparidad plantea desafíos significativos para el aislamiento y estudio de cada tipo celular de manera individual.

Con el fin de evaluar fluctuaciones en los niveles de expresión génica a lo largo de la espermatogénesis, así como de identificar genes de expresión específica de las

distintas etapas mencionadas más arriba, algunos de los estudios han comparado ARNs totales del testículo de individuos prepuberales de distintas edades [82,87,97], lo que se basa en atribuir los transcriptos encontrados en una edad y no en otra, a los distintos tipos celulares que van apareciendo a medida que progresa la primera onda espermatogénica. Sin embargo, esta aproximación no permite discriminar con exactitud qué tipo celular es el que está expresando un determinado transcripto. Además, fracasa en la identificación de transcriptos pertenecientes a los tipos celulares pobremente representados como los de profase meiótica temprana, que quedan diluidos frente a los de los tipos celulares más abundantes [98]. Adicionalmente, esta estrategia no toma en consideración las complejas interacciones celulares dentro del testículo (por ejemplo, entre células de línea germinal y células de Sertoli), a consecuencia de las cuales ciertos tipos celulares pueden cambiar sus patrones de expresión al entrar en contacto con los nuevos tipos celulares que van apareciendo [99,100]. Como un intento de superar estas limitaciones, algunos estudios han combinado el uso de testículos enteros de individuos de edades crecientes con un abordaje computacional de deconvolución, con el objetivo de convertir los perfiles de expresión temporal en perfiles de expresión específicos de tipo celular [88,101], aunque, por supuesto, el nivel de confiabilidad de los resultados es limitado.

A fin de poder realizar estudios moleculares en poblaciones celulares testiculares específicas, otros abordajes se han basado en el desarrollo de diversos métodos de purificación celular a lo largo del tiempo. Uno de estos métodos es el proceso de sedimentación conocido como "STA-PUT", que implica la disociación celular mediante enzimas seguida de una sedimentación basada en el tamaño celular en un gradiente de seroalbúmina bovina (BSA) [102]. Por ejemplo, Bao y colaboradores utilizaron el método STA-PUT para obtener poblaciones celulares a partir de testículos de ratones de diferentes edades, seguido de un análisis de microarreglos de secuencias de lncRNAs [103].

Un método alternativo de purificación celular, la elutriación, se lleva a cabo en un rotor de centrífuga especial, con cámara única lateral [104]. Es similar al STA-PUT, en el sentido de que también se basa en la separación de partículas según su velocidad de sedimentación. En la investigación de Soumillon y sus colegas [58], se empleó la elutriación para obtener fracciones de espermatocitos paquiténicos y espermátidas

redondas de ratón, con una pureza estimada del 70% y el 90%, respectivamente. La elutriación logra separaciones equivalentes a las obtenidas con el método STA-PUT en menos tiempo y con un rendiemiento de mayor número de células [104]. Sin embargo, en ambos procedimientos las poblaciones aisladas consisten principalmente en células paquiténicas y espermátidas redondas (por ejemplo, no se puede obtener células en profase meiótica temprana), y con un enriquecimiento moderado [105].

Más recientemente, ha surgido una metodología innovadora para la purificación de células mediante citometría de flujo, que ha ganado prominencia como la tecnología predominante en la obtención de fracciones de células testiculares enriquecidas o completamente puras, con niveles de pureza superiores al 95% [106]. Este enfoque, en el cual nuestro grupo ha tenido un rol protagónico, ha revolucionado la obtención de poblaciones celulares en diversas etapas de la espermatogénesis, incluyendo las etapas más tempranas de la meiosis, es decir, el leptoteno y el cigoteno [105,107]. La separación e identificación de estas poblaciones celulares se basa en una variedad de parámetros como el tamaño y la morfología celular, la complejidad interna, el contenido de ADN (2C, 4C o C), y la compactación de la cromatina. Para la marcación del ADN se utilizan tinciones con colorantes fluorescentes, siendo el *Hoechst 33342* el más común, a pesar de su desventaja de requerir una excitación en una longitud de onda del ultravioleta, lo que plantea preocupaciones sobre posibles daños en los ácidos nucleicos [108].

En nuestro laboratorio, se ha desarrollado un método innovador para la separación de células en diversos estadios de la espermatogénesis de roedores utilizando citometría de flujo, con un nivel de pureza excepcionalmente alto. Este enfoque comienza con una disgregación mecánica del tejido testicular, eliminando la necesidad de usar enzimas y contribuyendo a la preservación de las macromoléculas. Luego, las células en suspensión se tiñen con el colorante vital *Vybrant DyeCycle Green* (VDG), el cual se excita en el espectro visible, evitando así la exposición a la luz ultravioleta. Finalmente, las poblaciones celulares se separan y clasifican mediante citometría de flujo, logrando niveles de pureza que alcanzan un 98-99% [108,109]. Este método ha surgido como un logro significativo en nuestro grupo de trabajo

[98], permitiéndonos llevar a cabo investigaciones precisas en el campo molecular de la espermatogénesis.

1.3.2. Análisis transcriptomicos de la espermatogénesis

El desarrollo de las actuales tecnologías "ómicas" de alto rendimiento ha llevado al reconocimiento de la gran complejidad del espermatozoide, que transporta miles de ARNs y proteínas [3].

La complejidad a nivel celular en el tejido testicular se refleja igualmente en su perfil transcripcional. De hecho, se ha comprobado que los testículos destacan por su inmensa diversidad y complejidad transcriptómica, en comparación con otros tejidos: Expresan la mayor cantidad de genes tejido-específicos [110,111], junto con una abrumadora abundancia de lncRNAs [50,51,58,59,64,112,113]. Además, albergan una variada gama de sncRNAs, como piRNAs y miRNAs [114–119].

A pesar de las limitaciones metodológicas señaladas más arriba, los estudios transcriptómicos han significado un importante aporte que contribuyó a revelar al testículo de los mamíferos como un sistema sumamente interesante para los análisis de expresión génica diferencial:

a. Es el tejido con mayor número de genes expresados en forma exclusiva, y en forma diferencial, de todos los tejidos estudiados [58,110,111,120,121] (Figura 5A).

b. Posee una cantidad abrumadora de proteínas de unión a ARN [122].

c. Como hemos mencionado, alberga una variada gama de sncRNAs, como piRNAs y miRNAs [114–119]. De hecho, (en los mamíferos) los piRNAs se expresan de forma casi exclusiva en el testículo [39].

d. La abundancia de lncRNAs es enorme en comparación con otros tejidos [50,51,59,64,112,113]. En ese sentido, según los estudios de RNAseq, el número predicho de lncRNAs para diferentes especies de vertebrados sería mucho mayor en el testículo que en cualquier otro órgano o tejido [58] (ver Figura 5A).

e. La expresión génica es regulada en gran proporción a nivel postranscripcional, habiéndose desarrollado en el testículo algunos mecanismos originales de regulación postranscripcional como los mencionados más arriba, entre otros [123].

f. Posee la mayor tasa de procesamiento alternativo y cantidad de isoformas tejido-específicas (Figura 5B) de entre todos los tejidos estudiados, en conjunto con el cerebro [121,124–126], al menos en ratón y en humano.

De manera interesante, se ha observado que la complejidad del transcriptoma testicular en el ratón se origina principalmente en los espermatocitos primarios, y en particular, en las espermátidas redondas [58].

1.3.3. lncRNAs en la espermatogénesis

Durante la espermatogénesis se ha observado la existencia de lncRNAs altamente representados, específicos de cada etapa (87), de los cuales se conoce muy poco o nada de su funcionalidad, si bien se piensa que buena parte de los mismos estaría involucrada en funciones regulatorias [71].

Como se ha muestra en la Figura 5A, un estudio realizado por Soumillon y colaboradores [58] analizó el transcriptoma de múltiples órganos en diversas especies de mamíferos y aves, revelando que el testículo es el órgano con la transcripción más extensa del genoma. Este análisis mostró que el testículo expresa una cantidad significativamente mayor de lncRNAs en comparación con otros tejidos y órganos, destacando que los espermatocitos y las espermátidas son los tipos celulares con mayor expresión de estos ARN no codificantes [92]. Se ha confirmado que, durante el desarrollo del testículo, en roedores, los patrones de expresión de lncRNAs se conservan evolutivamente, sobre todo en regiones promotoras, lo que sugiere una importancia funcional clave para estas moléculas en la reproducción [92,127].

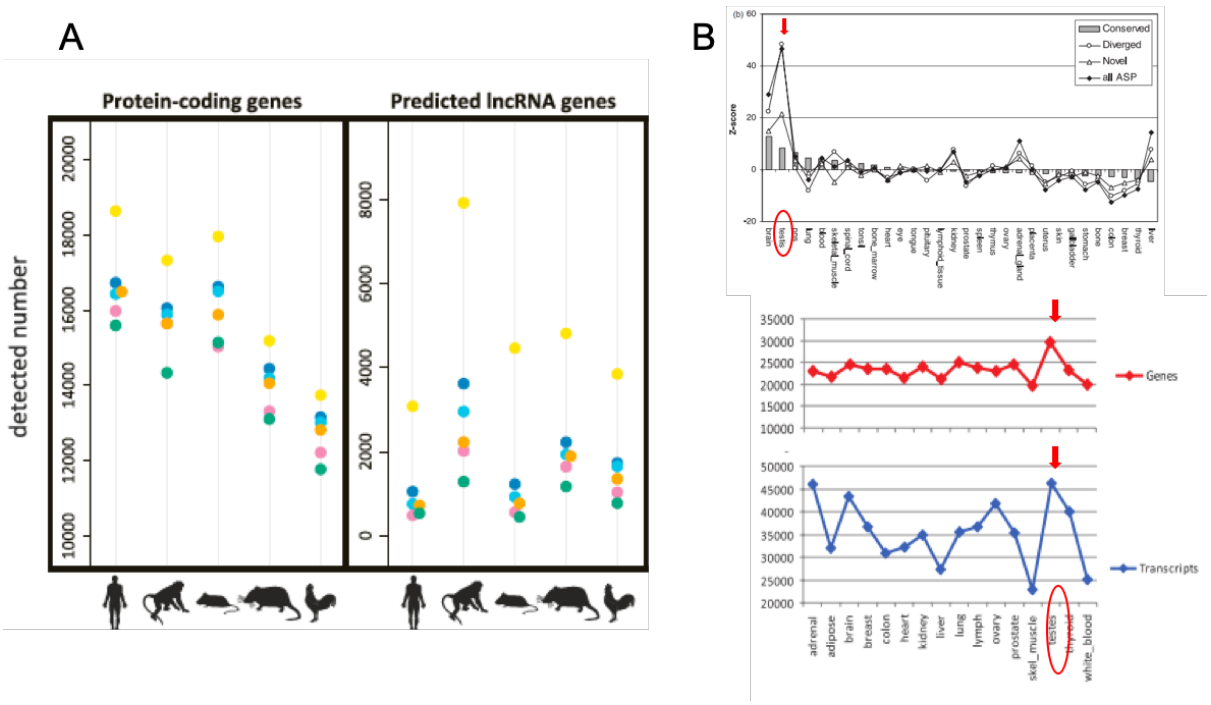


Figura 5. El testículo es el tejido con mayor número de genes y variantes de *splicing* expresadas. A) Número de genes codificantes de proteínas y genes de lncRNAs detectados en diferentes especies. El gráfico de la izquierda muestra el número de genes codificantes de proteínas, mientras que el gráfico de la derecha representa el número de genes de lncRNA predichos. Cada punto de color representa un tejido diferente, y los íconos en la parte inferior identifican las especies incluidas en el análisis, que abarcan humanos, monos, roedores y gallos (imagen tomada de Soumillon *et al.*, 2013). **B)** Distribución tisular del *splicing* alternativo en tejidos humanos. Se mide la sobre-representación de genes que contienen *splicing* alternativo en un tejido específico, en relación con los niveles de transcritos de fondo. La imagen superior corresponde a un estudio de ESTs ("*expressed sequence tags*"; Kan *et al.*, 2005), mientras que la imagen inferior corresponde a un estudio de RNAseq (Florea *et al.*, 2013).

Diversos lncRNAs han sido caracterizados y vinculados a funciones críticas durante la espermatogénesis. Por ejemplo, se ha observado que el silenciamiento del lncRNA *H19* provoca una reducción en el número de células dentro de los túbulos seminíferos de bovinos. Este lncRNA regula la expresión del receptor IGF-1R, esencial para la supervivencia de las células de Sertoli y espermatogénicas, influyendo directamente en la capacidad reproductiva masculina [128]. Estudios posteriores han revelado que la alteración de *H19* puede estar relacionada con la infertilidad en humanos, debido a la disfunción en la proliferación y maduración de las células germinales [92].

Otro ejemplo relevante es el del lncRNA *HongrES2*, identificado en las células testiculares de la rata. Este lncRNA es procesado en un miRNA y desempeña un papel esencial en la capacitación y maduración de los espermatozoides. Investigaciones recientes sugieren que este lncRNA está involucrado en la regulación epigenética durante la espermatogénesis, lo que lo convierte en un blanco potencial para terapias de tratamiento de la infertilidad masculina [127].

Asimismo, el lncRNA *Neat1* ha sido identificado como un regulador de la organización nuclear en los espermatocitos, al formar dominios conocidos como *paraspeckles* que modifican la transcripción del ARNm. En el testículo de la rata, *Neat1* desempeña un rol fundamental en la regulación de la expresión génica durante la espermatogénesis, lo que subraya su importancia en la fertilidad masculina [129].

1.3.4. *Splicing* alternativo en la espermatogénesis

En el contexto de la espermatogénesis, se ha demostrado que el *splicing* alternativo desempeña un papel crucial en la regulación de la expresión génica y en la maduración de los espermatozoides.

Como hemos mencionado más arriba, junto con el cerebro, los testículos se caracterizan por presentar la tasa más elevada de *splicing* alternativo, tanto en el humano como en el ratón, lo que da origen a numerosas isoformas de ARN y variantes de proteínas específicas de los testículos, sujetas a regulación temporal [80,130]. En consonancia con esto, como también hemos señalado, los testículos expresan una notable cantidad de proteínas de unión a ARN específicas y rigurosamente reguladas, incluyendo variantes de *splicing* únicas o expresadas de manera diferencial [121,130–133].

Se ha observado que varios genes de expresión específica de la espermatogénesis exhiben *splicing* alternativo, incluyendo genes involucrados en la regulación de la meiosis, la movilidad espermática y la fecundación [134]. Por ejemplo, el gen *PIWIL1*, que desempeña un papel crucial en la maduración de los espermatozoides de los bovinos, presenta diversas variantes de *splicing* que están relacionadas con la capacidad de fecundación [135].

De modo interesante, se han identificado variantes de procesamiento testículo-específicas que codifican isoformas con localización y función diferentes de las de sus isoformas somáticas. Sólo por citar un ejemplo, se conoce una isoforma específica del testículo de la poli(A) polimerasa (*TPAP*) que se localiza en el citoplasma, en contraste con la localización nuclear de la isoforma somática [123]. En ese sentido, se ha planteado que las células espermatogénicas poseen la capacidad de reclutar proteínas preexistentes para desempeñar nuevos roles, relacionados con la función del espermatozoide [121,136].

Además, se ha relacionado la alteración del *splicing* con trastornos testiculares patológicos [80,121,130,133,137–139]. Se ha observado que los patrones de *splicing* pueden estar alterados en casos de infertilidad, especialmente en situaciones de disfunción en la espermatogénesis, lo que sugiere su relevancia en la regulación de la espermatogénesis normal y en la etiología de la infertilidad [80].

Un dato curioso que ha surgido, se relaciona con la retención de intrones en la espermatogénesis. Si bien ésta existe en otros tejidos, juega en éstos un rol completamente diferente: en la mayoría de los tejidos es un mecanismo para regular a la baja la expresión, dado que al ser los ARNm exportados al citoplasma con un intrón retenido, se convierten en blancos para el mecanismo de degradación de transcriptos con codones *Stop* prematuros, o NMD [140]. Sin embargo, se ha reportado que en la espermatogénesis la retención de intrones jugaría un rol en la regulación temporal de la expresión de mensajeros. En las células espermatogénicas, y especialmente en las meióticas, ARNs con un intrón no procesado quedarían retenidos por largos períodos en el núcleo, para su posterior procesamiento y exportación tardías al citoplasma, en el momento en que se requiere la síntesis de la proteína [141,142].

Sorprendentemente, el procesamiento alternativo a lo largo de la espermatogénesis ha sido muy escasamente estudiado. Si bien de los varios estudios que han analizado la diversidad transcriptómica testicular a lo largo de la progresión de la espermatogénesis, algunos incluyeron una identificación y/o caracterización preliminar del procesamiento alternativo [58,87,90,143–145], sólo un número muy limitado de estudios ha abordado un análisis más detallado [88,134,146]. Más aún, dichos estudios

se basaron principalmente en enfoques de deconvolución computacional [88], o en conjuntos de datos ya disponibles en repositorios [134].

En síntesis, el *splicing* alternativo representa un mecanismo crucial en la regulación de la expresión génica y la maduración de los espermatozoides, y su perturbación puede estar asociada con trastornos en la espermatogénesis e infertilidad. No obstante, a pesar de que los estudios moleculares dirigidos a identificar y caracterizar las variantes de *splicing* en genes específicos de la espermatogénesis son esenciales para una mejor comprensión de estos procesos y para el desarrollo de posibles enfoques terapéuticos para la infertilidad, este proceso ha sido muy pobremente estudiado a lo largo de la espermatogénesis.

1.4. Antecedentes directos

A partir de la metodología desarrollada por nuestro grupo para la separación de estadios celulares de distintas etapas clave de la espermatogénesis, se generaron transcriptomas a lo largo de la espermatogénesis del ratón (*Mus musculus*) utilizando cuatro poblaciones celulares altamente purificadas, correspondientes a 4 etapas: espermatogonias (células premeióticas) y células somáticas testiculares (ambos tipos celulares son 2C en contenido de ADN), profase meiótica temprana (leptoteno-cigoteno: LZ), profase meiótica media (espermatocitos en paquiteno: PS; "*pachytene spermatocytes*"), y células posmeióticas (espermátidas redondas: RS; "*round spermatids*"). Estos estudios demostraron, entre otras cosas, que la incidencia de la regulación postranscripcional en la espermatogénesis es mucho mayor aún de lo que se conocía previamente [89]. Sin embargo, las genotecas generadas no eran específicas de hebra, lo que no permitía conocer la dirección original de la transcripción. En consecuencia, estas genotecas no resultaban adecuadas para el estudio de ARNs no codificantes, así como tampoco eran ideales para el ensamblaje de transcriptos nuevos. Debido a que muchos lncRNAs se transcriben a partir de hebras antisentido, e, incluso, solapando con genes de transcriptos codificantes [147], al emplear genotecas no direccionales resulta imposible saber si un transcripto

proviene de una hebra sentido o anti sentido, y se corre el riesgo de generar una quimera con los solapantes.

Se procedió entonces a generar nuevamente los transcriptomas de las cuatro poblaciones testiculares de ratón mencionadas, en altísimo grado de pureza y con tres réplicas biológicas para cada población (12 transcriptomas), pero esta vez se produjeron genotecas específicas de hebra. El estudio de los transcriptomas obtenidos permitió realizar un análisis de transcriptos no codificantes, poniendo en evidencia la existencia de lncRNAs de expresión diferencial de cada una de las etapas analizadas. Este estudio reveló que la gran mayoría de los lncRNAs diferenciales identificados correspondía a las células posmeióticas [148]. No obstante, es importante destacar que dicho análisis se basó únicamente en los transcriptos anotados en las bases de datos.

Para el presente trabajo de tesis nos propusimos reanalizar los datos crudos disponibles de las genotecas direccionales realizadas por nuestro grupo [148], para ensamblar transcriptos y genes nuevos, y así estudiar en mayor profundidad el universo aún no dilucidado de la espermatogénesis.

Es así que se presenta un análisis integral de la expresión génica en la espermatogénesis del ratón. Particularmente, se hizo énfasis en la identificación de genes codificantes “nuevos”, variantes de *splicing* y lncRNAs. Dado que el ratón es el modelo por excelencia para los estudios de genómica y transcriptómica en mamíferos, los resultados representan un importante aporte al conocimiento de la diversidad transcriptómica durante la espermatogénesis de los mamíferos. Más aún, teniendo en cuenta que el testículo se ha revelado como el tejido más complejo en términos de expresión génica de entre todos los tejidos estudiados, este trabajo constituye una aproximación al estudio de la complejidad real del transcriptoma en los mamíferos.

2. HIPÓTESIS Y OBJETIVOS

2.1. Hipótesis de trabajo

Dado que el testículo es un tejido de una inmensa complejidad, cuya totalidad de funciones y mecanismos regulatorios específicos aún no han sido completamente esclarecidos, proponemos que:

a. Existe una gran cantidad de variantes de *splicing* expresadas de manera específica o diferencial en las distintas etapas de la espermatogénesis del ratón, una parte considerable las cuales no ha sido identificada hasta el momento. Estas variantes podrían estar desempeñando funciones cruciales y hasta ahora no descritas;

b. Existen genes involucrados en la espermatogénesis que aún no han sido siquiera anotados.

En definitiva, mediante la confirmación de estas hipótesis, pretendemos demostrar que la complejidad transcriptómica durante la espermatogénesis es aún mayor que lo reportado hasta ahora.

2.2. Objetivo general

Contribuir a la comprensión a nivel molecular de la espermatogénesis de los mamíferos, en particular en relación a los patrones de expresión génica diferencial durante la misma.

2.2.1. Objetivos específicos

a. Analizar las especies de ARN presentes en diversas etapas representativas de la espermatogénesis del ratón.

b. Identificar los ARNs hasta el momento sin anotación.
c. Analizar el potencial codificante de los transcriptos hasta el momento no anotados, e identificar isoformas potencialmente relevantes para la espermatogénesis.
d. Identificar ARNs no codificantes largos (lncRNAs) no anotados, específicos de las diferentes etapas de la espermatogénesis.
e. Analizar comparativamente los transcriptomas de las diferentes etapas de la espermatogénesis.
f. Confirmar por técnicas alternativas resultados obtenidos en los puntos previos.

3. MATERIALES Y MÉTODOS

3.1. Análisis transcriptómicos

3.1.1. Datos crudos

Los datos crudos utilizados en esta tesis fueron obtenidos de bibliotecas de RNAseq específicas de hebra, generadas a partir de poblaciones celulares puras del testículo de ratones (*Mus musculus*) *CD1-Swiss*.

Para la confección de las bibliotecas se seleccionaron etapas clave a lo largo de la espermatogénesis del ratón, y se obtuvieron poblaciones celulares puras correspondientes a las distintas etapas de interés mediante clasificación por citometría de flujo. Las poblaciones celulares analizadas fueron: una población con contenido de ADN 2C, consistente en espermatogonias y células somáticas testiculares (le llamaremos población 2C); una población de células en profase meiótica temprana, es decir, espermatocitos en leptoteno y cigoteno (población LZ); una población de espermatocitos en profase meiótica media, es decir, paquiteno (población PS, del inglés: "*pachytene spermatocytes*"); y, por último, una población de células posmeióticas, específicamente espermátidas redondas (población RS, del inglés: "*round spermatids*"). Es importante señalar que la población 2C se componía de una mezcla heterogénea de tipos celulares (distintos tipos de espermatogonias, diferentes tipos celulares somáticos), y fue usada como control de los transcriptos expresados en células somáticas y pre-meióticas. Dicha población fue obtenida de un *pool* de hasta cinco individuos de 12-14 días posparto (dpp). La selección de animales de esta edad para la obtención de esta población fue intencional, para excluir la posibilidad de que la misma pudiera contener espermatocitos secundarios, que también son 2C pero aún no están presentes a esa edad temprana. Las poblaciones celulares LZ y PS se obtuvieron a partir de animales de 15 a 19 dpp, edad en la que estos tipos celulares están más representados en el total de células testiculares. Finalmente, la población RS se obtuvo de animales de 22 a 24 dpp, edad en la que ya hay espermátidas redondas, pero aún no se detectan espermatozoides.

Las librerías fueron generadas mediante el kit *Ovation RNA-Seq 1-16 for Mouse* (NuGEN, CA EE. UU.), específico para la obtención de librerías específicas de hebra, y las mismas se secuenciaron en la empresa Fasteris (Suiza) en una plataforma Illumina Hi-Seq 4000 durante el trabajo de Trovero y colaboradores [148]. En total, se utilizaron 12 bibliotecas que correspondían a las cuatro poblaciones celulares, cada una con tres réplicas biológicas, y los datos crudos fueron almacenados en el repositorio *Sequence Read Archive* (*SRA*) del NCBI (https://www.ncbi.nlm.nih.gov/sra), con número de acceso PRJNA548952 [148].

3.1.2. Datos crudos de otros trabajos, para estudios de reproducibilidad

Conjuntamente con los datos antes mencionados, se emplearon datos crudos de una publicación de Chen y colaboradores [90], descargados del repositorio de datos de expresión génica *Gene Expression Omnibus* (GEO) de NCBI (http://www.ncbi.nlm.nih.gov/gen/) con el ID de acceso: GSE107644. Los mismos fueron obtenidos mediante secuenciación de ARN de célula única (*single-cell RNA-seq*), y generados con base en el método *Smart-seq2*, con algunas modificaciones [90].

Por otra parte, los datos crudos de da Cruz y colaboradores [89] se descargaron del repositorio *SRA*, con número de acceso PRJNA317251. Estos datos habían sido generados por nuestro grupo mediante una amplificación lineal de ARNs utilizando el sistema *Ovation RNA-Seq v2* (NuGEN, San Carlos, CA), a partir de poblaciones celulares de alta pureza de distintas etapas espermatogénicas clasificadas por citometría de flujo. Las librerías fueron construidas y secuenciadas en Macrogen (Seúl, Corea).

Los datos crudos de estos dos trabajos se utilizaron únicamente para efectuar comparaciones con los nuestros, según se detalla más abajo, y no fueron incluidos en los análisis generales por no proceder de bibliotecas direccionales, entre otras diferencias.

3.1.3. Procesamiento de datos

Un esquema general del flujo de trabajo (*pipeline*) bioinformático seguido, se representa en la Figura 6.

El análisis de los datos se centró en moléculas ≥ 200 pb, lo cual incluye los ARNm y los lncRNAs. No fueron incluidos los ARNs de menor tamaño, ya que ni el método de extracción de ARN, ni el método construcción de bibliotecas, eran adecuados para ello.

Para el recorte por calidad de secuencia, eliminación de códigos de barra y adaptadores, fue empleado *TrimGalore* [149]. Se trabajó con lecturas pareadas mediante la opción *"--paired"*, reteniendo todas aquellas en las que, al menos, una de las lecturas del par, superara todos los controles de calidad. En estos casos, las lecturas no pareadas se conservaron mediante la opción *"--retain_unpaired"*. La línea utilizada fue:

"trim_galore --paired $lib'_1.fastq' $lib'_2.fastq' --retain_unpaired"

Las lecturas que pasaron el control de calidad (FastQC, Q > 28) se mapearon con *HISAT2* (http://daehwankimlab.github.io/hisat2/) (Figura 6), utilizando los parámetros dta (ensamblaje de transcriptos corriente abajo). Realizamos un alineamiento guiado por el genoma, utilizando tanto lecturas pareadas como no pareadas para cada población celular (de forma de emplear el máximo de lecturas posibles). Se utilizó el genoma de ratón anotado en *Ensembl* (versión Grcm38.92) como genoma de referencia. La línea empleada fue:

"hisat2 --dta -x /media/Disco5/carlos/referencia/referencia/grcm38/genome -1 160930_SNK268_B_L007_JCB-14_R1_val_1.fq.gz -2 160930_SNK268_B_L007_JCB-14_R2_val_2.fq.gz -U 160930_SNK268_B_L007_JCB-14_R1_unpaired_1.fq.gz,160930_SNK268_B_L007_JCB-14_R2_unpaired_2.fq.gz -S 14.sam"

Utilizamos *Strawberry* [150] para ensamblar nuevos transcriptos guiado por genoma, empleando un mínimo de 10 lecturas como soporte por sitio de empalme y por exón (ver Figura 6). Para generar una referencia única de nuestros ensamblajes, usamos *StringTie* con la opción de fusión [151]. Se empleó la siguiente línea:

“strawberry -o ./strawberry_newas/$lib -g /media/Disco7/carlos/referencia/HISAT2_index_primary_as/Mus_musculus.GRCm38.92.gtf -p2 --fr $lib'.bam'”

Por último, para evaluar si los datos generados eran reales o no, se utilizaron diferentes umbrales de profundidad. Se observó que los resultados no cambiaron sustancialmente entre diferentes coberturas por profundidad (Figura 7), por lo cual nos decidimos a emplear una cobertura mínima de 10x, la cual consideramos suficientemente sólida para proceder con nuestros análisis.

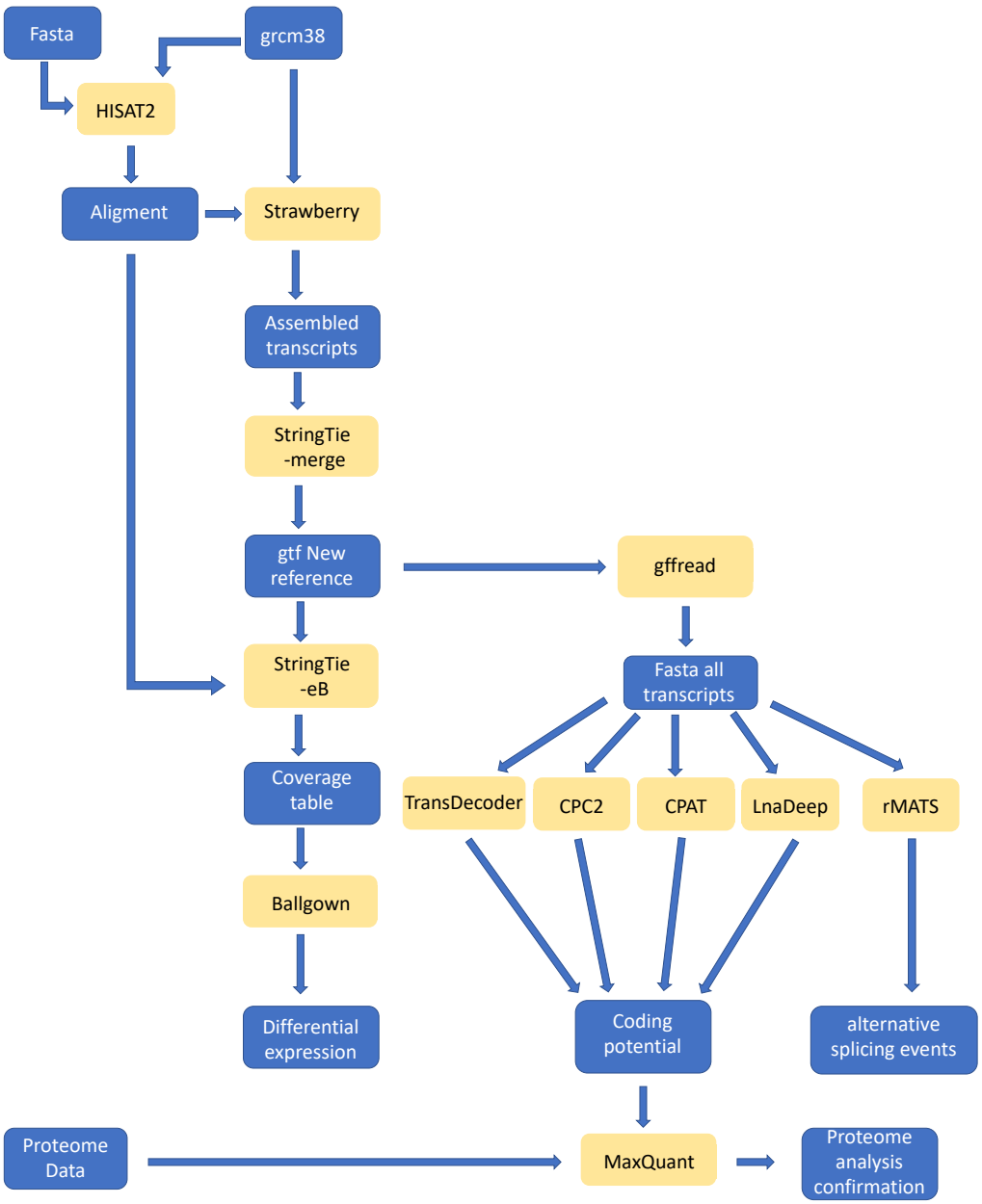


Figura 6. ***Pipeline* bioinformático**. El flujo de trabajo comenzó con los datos de secuencias en formato FASTA, que se alinearon con el genoma de referencia (GRCm38) utilizando el software ***HISAT2***. A partir del alineamiento, los transcriptos se ensamblaron mediante ***Strawberry*** y luego se combinaron en una referencia unificada utilizando ***StringTie-merge***. La nueva referencia en formato GTF fue procesada para generar una tabla de cobertura con ***StringTie-eB***, la cual fue utilizada por ***Ballgown*** para analizar la expresión diferencial de los transcriptos. En paralelo, se extrajeron todos los transcriptos en formato FASTA mediante **gffread**, para predecir el potencial codificante utilizando múltiples herramientas: **TransDecoder**, **CPC2**, **CPAT**, y **LnaDeep**. Para la identificación de eventos de *splicing* alternativo, se utilizó **rMATS**. El resultado del análisis del potencial codificante se validó contra los datos proteómicos mediante **MaxQuant**, y se realizó la confirmación de eventos de *splicing* alternativo.

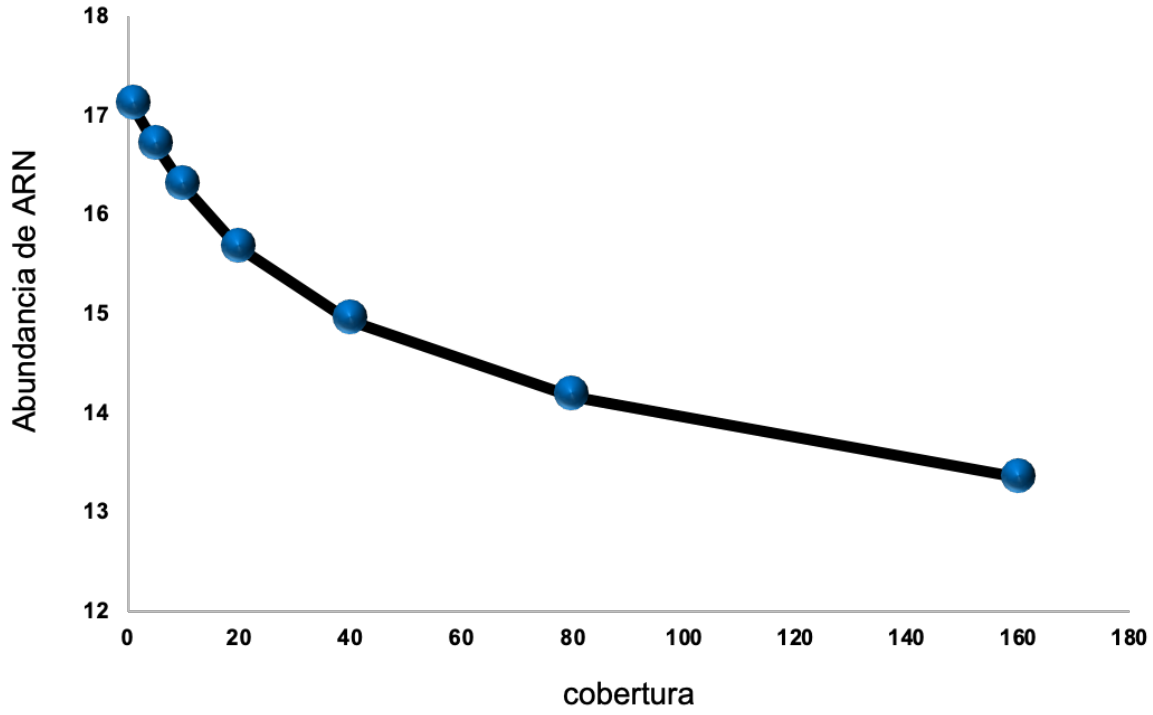


Figura 7. Gráfico semilogarítmico de los transcriptos identificados *vs.* la cobertura, para siete diferentes umbrales de abundancia de transcriptos. El eje de ordenadas (abundancia de ARN) indica la escala logarítmica (log_2) del número de transcriptos.

3.1.4. Matriz de correlación

Se construyó una matriz de correlación en *R Bioconductor* (http://www. R-project.org), calculando el coeficiente de correlación de Pearson entre la expresión en FPKM (*Fragments Per Kilobase per Million mapped*) de cada transcripto en cada una de las 12 muestras, con el fin de apreciar la fuerza de las correlaciones entre nuestras réplicas.

3.1.5. Saturación de transcriptos

Estimamos la saturación de los transcriptos mediante la rarefacción de las lecturas, con el objetivo de determinar si se alcanzaba la saturación en las cuatro poblaciones celulares y descartar posibles artefactos. Para este propósito, se realizaron recuentos con *FeatureCounts* [152] utilizando datos de este estudio, y comparándolos con los de da Cruz y colaboradores [89]. Se emplearon las siguientes condiciones: -O asigna las

lecturas a todas sus metafunciones superpuestas; -s0 indica lecturas sin orientación específica; -t especifica el tipo de característica(s) en una anotación GTF; y -g indica el tipo de atributo en la anotación GTF, con la referencia que generamos previamente. Posteriormente, en R, utilizamos la función "*estimate saturation*" de la biblioteca *RNAseQC* [153], que permite realizar un corte por profundidad y así observar cómo ocurre la detección de transcritos en función del número de lecturas. La línea utilizada en este caso fue:

"featureCounts -O -s0 -t transcript -g transcript_id -p -T 8 -a /media/Disco5/Disco7_Azul/carlos/map_HISAT2/map_HISAT2_primary/strawberry_newas_10/straw_primary_merge_10.gtf -o count_straw10/count_straw10_Irene SRR34376992C.bam SRR3437700LZ.bam SRR3437701PS.bam SRR3437702RS.bam"

3.1.6. Análisis de componentes principales (PCA)

Se compararon nuestras listas de RNAseq con las de otro estudio en el que se analizaron diferentes etapas espermatogénicas a nivel de scRNA-seq [90], empleando sus datos crudos. Los datos sin procesar de ese estudio fueron mapeados con la misma metodología que utilizamos para mapear nuestros propios datos. A continuación, se realizó el conteo de nuestros datos y los de scRNA-seq, con nuestro ensamblaje como referencia, utilizando *HTseq-counts* [154]. Las listas generadas fueron normalizadas con el paquete *limma* para R [155]. Se generó un análisis de componentes principales (PCA) utilizando *Seurat*, que emplea los valores de CPM (*Counts Per Million*) normalizados, y en escala logarítmica como entrada [156].

3.1.7. Detección de variantes de *splicing*, y análisis del potencial codificante

En la evaluación de las variantes de *splicing*, se empleó el software *rMATS* (http://rnaseq-mats.sourceforge.net/) con sus configuraciones por defecto (ver Figura 6). Para la determinación del número de transcriptos por gen, para transcriptos codificantes y no codificantes, los representamos normalizados como el porcentaje del total de transcriptos en cada categoría. Se realizó una prueba T para calcular los

valores estadísticos entre los tipos de células utilizando sus réplicas. Por otro lado, para la visualización de la estructura y expresión de transcriptos en el análisis individual de un gen, se aplicó la función *PlotTranscripts*, según lo descrito por Pertea y colaboradores [151].

El archivo GTF de referencia, que contenía nuestro ensamblaje, se transformó en un archivo FASTA mediante el uso de *gffread*, según la metodología propuesta por Pertea & Pertea en 2020 [157]. Este archivo FASTA se utilizó como entrada para diversos paquetes de software (ver Figura 6), con el propósito de clasificar los nuevos transcriptos dentro de la categoría de codificantes, o de la de no codificantes. Con este objetivo, se implementó, simultáneamente, el uso de cuatro paquetes de *software* distintos: i) *TransDecoder*, el cual se basa en *GeneMarkS-T*, y ofrece una alta precisión en la identificación de regiones codificantes y en la predicción de sitios de inicio de traducción [158]; ii) *CPC2*, un programa de rápido procesamiento, con una alta precisión especialmente para transcriptos largos no codificantes, siendo además el modelo de *CPC2* neutral con respecto a la especie [159]; iii) *LncADeep*, que predice las proteínas con las que interactúa un lncRNA utilizando redes neuronales profundas, basándose en información tanto de secuencia como de estructura, e integra el análisis de enriquecimiento de vías *KEGG* y *Reactome*, así como la detección de módulos funcionales, con las proteínas predichas como interactuantes [160]; y iv) *CPAT*, que utiliza un modelo de regresión logística basado en cuatro características de la secuencia, que son el tamaño del marco abierto de lectura, la cobertura del marco de lectura abierto, estadística *TESTCODE* de *Fickett*, y sesgo en el uso de hexámeros [161]. Los cuatro paquetes de software se utilizaron con sus ajustes predeterminados. En los análisis subsiguientes, únicamente se consideró la intersección de los resultados obtenidos con los cuatro paquetes de software, con el fin de asegurar una mayor confiabilidad de los resultados obtenidos en relación al potencial codificante de cada uno de los transcriptos en estudio.

3.1.8. Diagramas de Venn

Se generaron diagramas de Venn mediante el empleo del software *Bioinformatics & Evolutionary Genomics*, accesible en http://bioinformatics.psb.ugent.be/webtools/Venn/.

3.1.9. Anotación primaria, y predicción estructural de proteínas putativas

Con la intención de evaluar la funcionalidad de los genes no anotados, llevamos a cabo una anotación primaria mediante *Trinotate* [162], haciendo uso de todos los métodos y bases de datos disponibles, sin importar la especie (*BLASTX, SWISSPROT, RNAMMER, prot_id, BLASTP, Pfam, SignalP, TMHMM, eggNOG, KEGG, Gene Ontology BLAST, Gene Ontology Pfam*). Posteriormente, para comparar las distintas categorías de transcriptos entre sí y entre las distintas poblaciones celulares, generamos un diagrama de caja (*boxplot*) mediante *ggplot* [163]. En este análisis, que incluyó comparaciones estadísticas de las medias, empleamos *ggpubr* [164].

En el caso de las proteínas predichas para las que no se obtuvo una anotación primaria a través del análisis masivo, realizamos posteriormente una búsqueda manual mediante *BLASTP*.

En cuanto a la modelización de las proteínas predichas, recurrimos a *Swiss-Model* (https://swissmodel.expasy.org/interactive), y para llevar a cabo un análisis de los dominios proteicos putativos, empleamos *Pfam* (http://pfam-legacy.xfam.org).

3.1.10. Análisis de expresión diferencial

Se realizó un análisis de expresión diferencial entre las cuatro poblaciones celulares testiculares, mediante comparaciones por pares en orden cronológico, a lo largo de la evolución de la primera onda espermatogénica (LZ *vs* 2C; PS *vs* LZ; RS *vs* PS), y utilizando el software *Ballgown* [151]. Se aplicó un criterio de cambio de expresión (*fold change*: FC) log2 (FC) ≥2 o ≤-2, junto con un valor de $q < 0.05$, para filtrar los genes diferencialmente expresados (DE). Asimismo, se implementó un filtro adicional de cobertura mínima de 10X.

3.2. Ensayos de mesada

3.2.1. Animales de laboratorio

Los animales empleados durante la presente tesis fueron ratones CD1 *Swiss* de 30 dpp y 60 dpp, adquiridos del bioterio del instituto de Higiene de la Universidad de la República (UdelaR), Uruguay. La eutanasia de los ratones se llevó a cabo por dislocación cervical en estricto acuerdo con las normas del Comité de Ética en el Uso de Animales (CEUA) del Instituto de Investigaciones Biológicas Clemente Estable (IIBCE), bajo las regulaciones establecidas en la Ley Nacional de Experimentación Animal N°18611. El protocolo de experimentación fue aprobado por la Comisión Nacional de Experimentación Animal del Instituto de Investigaciones Biológicas Clemente Estable (protocolo 001/02/2012; código: 008/11; http://www.cnea.org.uy/index.php/instituciones/registro/10).

3.2.2. Confirmación experimental de variantes de *splicing* por RT-PCR

Para confirmar la existencia de las variantes de empalme seleccionadas, diseñamos cebadores específicos para PCR con el objetivo de amplificar, tanto los transcriptos recién identificados, como aquellos ya anotados. Los cebadores diseñados se detallan en el Anexo, Tabla S1.

El diseño de los cebadores se realizó empleando *Primer Blast* (https://www.ncbi.nlm.nih.gov/tools/primer-blast/), software que también se empleó para evaluar la especificidad de los mismos. Posteriormente se evaluaron las condiciones fisicoquímicas predichas para cada par de cebadores, utilizando el programa *OligoAnalyzer 3.1* (https://www.idtdna.com/calc/analyzer).

Se obtuvieron fracciones celulares altamente puras de cada una de las poblaciones analizadas (2C, LZ, PS y RS), a partir de individuos de las mismas edades que los empleados para la confección de las librerías de secuenciación. Las fracciones celulares se purificaron mediante citometría de flujo, utilizando las técnicas de clasificación celular desarrolladas en nuestro laboratorio y previamente descritas [148].

Brevemente, se prepararon suspensiones celulares por disgregación mecánica mediante un dispositivo *Medimachine* (Becton Dickinson, EE. UU.), y las mismas se tiñeron con el colorante vital *Vybrant DyeCycle Green* (VDG; Invitrogen, Life Technologies, Carlsbad, CA, EE. UU.) durante 1 hora, incubándolas a 35°C con leve agitación a 80 revoluciones por minuto (rpm), según lo descrito en publicaciones previas del grupo [109,148]. La clasificación se llevó a cabo en un citómetro de flujo *MoFlo Astrios EQ* (Beckman Coulter) en modo de purificación, el cual permite maximizar la pureza. Se obtuvieron triplicados de cada una de las poblaciones, con 3.000 células por muestra.

Las fracciones celulares clasificadas se utilizaron para la realización de RT-qPCR confirmativa mediante el kit *Power SYBR Green Cells-to-Ct* (Ambion-Life Technologies, CA EE. UU.), siguiendo las instrucciones del fabricante, y utilizando cebadores aleatorios para la síntesis de ADNc de cadena simple. La reacción de transcripción reversa se llevó a cabo en un volumen final de 40 µL según las recomendaciones del fabricante. Se emplearon 2 µL de ADNc en cada reacción de PCR en un volumen final de 20 µL, según las indicaciones del kit. Las reacciones de transcripción reversa y PCR se realizaron utilizando un sistema de detección de PCR en tiempo real *CFX96 Touch* (BioRad, Hercules, CA, EE. UU.), con triplicados biológicos.

Si bien la confirmación de las variantes de empalme podría haberse realizado por RT-PCR a tiempo final, optamos por utilizar este kit de RT-qPCR debido a su alta sensibilidad, considerando el bajo número de células clasificadas (este kit es ideal para amplificar ARNs a partir de cantidades pequeñas de células). Las reacciones de PCR luego se sometieron a análisis en geles de agarosa convencionales para la visualización de las variantes de *splicing* (similar a un PCR a tiempo final), y se tiñeron con *GelRed* al 10% (Biotium, Fremont, CA, EE. UU.) según los procedimientos habituales.

3.2.3. Diseño y producción de anticuerpos

Para la síntesis de péptidos y generación de anticuerpos específicos contra la probable isoforma proteica no anotada de MSH5, se recurrió al servicio de *GenScript*

(Piscataway, NJ, EE. UU.; https://www.genscript.com/?src=google&utm_source=google&utm_medium=cpc&utm_campaign=GenScript_new&jiraid=12194&gad_source=1&gclid=CjwKCAjwufq2BhAmEiwAnZqw8rCZEYtsXL_jiH0W0bE2qL_KqeY4HhstwBe9sdHOU8wiexVVcFSojhoCs2YQAvD_BwE).

Con ese objetivo, se diseñaron péptidos contra la región amino-terminal de la misma, que no poseía homología con las variantes ya anotadas.

Seleccionamos dos péptidos (SFSRPGDPVDVPAAC y CSGIGPCLPPGRPVG), los cuales se inocularon como antígenos en conejos para generar los anticuerpos primarios. Una vez generados los anticuerpos, los mismos se purificaron por afinidad. Los detalles de la generación y purificación de los anticuerpos se encuentran disponibles dentro de los servicios ofrecidos por *GenScript*, en el link indicado más arriba.

3.2.4. Inmunofluorescencia

Testículos de ratones de 30 dpp y 60 dpp fueron desprovistos de su túnica albugínea, y seccionados en fragmentos de aproximadamente 3 mm de diámetro. Los mismos fueron fijados por inmersión en buffer PHEM (60 mM PIPES, 2 mM HEPES, 10 mM EGTA, 2 mM MgCl2, pH 7,4) con paraformaldehído al 4%, toda la noche a 4°C. Luego se procedió a infiltrar el tejido con sacarosa para crio-protegerlo, primero en PHEM al 15% durante 1:30 horas a 4°C, y luego en solución PHEM con sacarosa al 30%, toda la noche a 4°C. Posteriormente, las muestras fueron infiltradas en medio de inclusión puro (*Jung Tissue Freezing Medium*, Leica Microsystems Nussloch GmbH, Alemania). El tejido se colocó en un molde en medio de inclusión nuevo, y se congeló a -20°C. Se realizaron criosecciones de 10 µm de espesor en crióstato (SLEE, medical GmbH, Alemania). Las criosecciones, recogidas sobre portaobjetos pretratados con sustancias adherentes (poli-L-Lisina, Invitrogen), se mantuvieron a -20°C hasta el momento de realizar la inmunofluorescencia.

Las criosecciones fueron permeabilizadas en PHEM con 0,1% de Tween 20 durante 10 minutos a temperatura ambiente. Luego, se bloquearon los grupos aldehído y

cetona libres con PHEM conteniendo 1% de borohidruro de sodio (213462, *Sigma Aldrich,* Saint Louis, EE. UU.), durante 30 minutos a temperatura ambiente. Posteriormente, se bloquearon las uniones inespecíficas con buffer de incubación (PHEM con 0,1% de BSA y 50 mM de glicina) conteniendo 5% de suero normal de cabra durante 30 minutos a temperatura ambiente. Los cortes fueron incubados durante 16 horas a 4°C con los anticuerpos específicos. El exceso del anticuerpo primario se lavó en PHEM con 0,1% de Tween 20 y 50 mM glicina, seguido de una incubación de 1 hora a temperatura ambiente con PBS 1X conteniendo el anticuerpo secundario (anti-conejo conjugado a Alexa 488; A-11034, Invitrogen), diluido 1:1000, y con DAPI 1 µg/mL. El exceso de anticuerpo secundario fue removido por sucesivos lavados en buffer PHEM con 0,1% de Tween 20 y 50 mM glicina. Se eliminó el exceso de humedad, y se realizó el montaje definitivo con *Prolong Gold Antifade* (P36961, Invitrogen).

Las imágenes de microscopía láser confocal se obtuvieron utilizando un microscopio *Zeiss LSM 800* (Carl Zeiss Microscopy, Alemania), equipado con un objetivo de inmersión en aceite de 63X y una apertura numérica de 1.4. Para la excitación, se emplearon láseres de 405 nm y 488 nm. Las fotografías fueron capturadas con una cámara digital en color *Axiocam* 506 (Carl Zeiss Microscopy), con una resolución de 1024 × 1024 píxeles, a partir de secciones individuales o consecutivas en el eje z con distancia variable (*stacks*). El procesamiento inicial de las imágenes se realizó con el *software Zen 2.3* (Carl Zeiss). Posteriormente, se procedió al análisis de las imágenes con el *software Fiji ImageJ* [165], con el cual se observaron en más detalle, se generaron imágenes 3D y proyecciones.

3.2.5. SDS-PAGE y Western Blot

Los testículos de ratones macho de 30 dpp se disecaron, se les extrajo la túnica albugínea, y se disgregaron mecánicamente en buffer de Laemmli 4X (Tris base pH 6,8 278 mM, SDS 4,4%, glicerol 44,4%, 2-mercapto-ethanol 10%, azul de bromofenol 0,02%), por pipeteado. Se procesaron fragmentos igualados en peso (~0,02 g – 0,03 g), y se homogeneizaron en igual relación masa/volumen.

Los lisados de proteínas, en una cantidad correspondiente a 7,5 x 10^5 células aproximadamente, en buffer de Laemmli, fueron cargados en geles de SDS-PAGE al 12%. Los geles se corrieron a 110 V hasta el borde del gel apilador, y luego a 100 V hasta que el estándar de 25 kDa llegara a la parte inferior del gel. El control estricto de la longitud de la corrida fue para mejorar la separación de las proteínas, ya que se esperaban unos pesos moleculares de alrededor de 75 kDa para la proteína MSH5 canónica y 39 kDa para la MSH5 no anotada.

Los geles de proteínas se transfirieron a membranas de nitrocelulosa *Amersham Protran* (GE10600002, Sigma-Aldrich) mediante transferencia líquida, por el método convencional (Matsudaira, 1987) a una tensión de 100 V durante 45 minutos empleando buffer Towbin (Tris 0,025 M, glicina 0,192 M, metanol 20%) como buffer de transferencia. Se utilizó para la transferencia el sistema *Mini Trans-Blot Electrophoretic Transfer Cell* (BioRad). Luego de la transferencia, las membranas se tiñeron durante 5 minutos en solución de Rojo Ponceau S (Sigma-Aldrich) al 0,1% en ácido acético para verificar la calidad de la misma.

Las membranas se incubaron a 4°C en TBST (Tris pH 7,4 100 Mm, NaCl 1,5 M, Tween 20 1%) con leche 10% durante 16 horas, para bloquear. Al día siguiente, las membranas en TBST/leche se incubaron a temperatura ambiente durante 1 hora con agitación. Después de realizar dos lavados breves con TBST, se procedió a incubar las membranas con el anticuerpo primario detallado previamente, diluido en TBST a una dilución 1:2.000. Las incubaciones se efectuaron durante 1:30 horas a temperatura ambiente con agitación. Se realizaron dos lavados de 15 minutos cada uno con TBST, y se incubó con un anticuerpo secundario anti-conejo acoplado a peroxidasa de rábano (34085 de Pierce, *Thermo Fisher Scientific*, con una dilución 1:10.000) en TBST/leche durante 45 minutos. Después de realizar dos lavados de 15 minutos con TBST, se reveló mediante el kit *SuperSignal West Pico* (*Thermo Fisher Scientific*), dejando actuar la solución reveladora durante 5 minutos. Finalmente, las imágenes se visualizaron en el equipo *Ibright 1500FL* (*Thermo Fisher Scientific*), mediante la opción para quimioluminiscencia.

3.3. Análisis proteómicos

Para estos estudios, se emplearon datos crudos de proteómica *shotgun* de testículo de ratón obtenidos por Giansanti y colaboradores [166], los cuales fueron descargados del repositorio de la *Proteomics Identifications Database* (PRIDE) de EMBL-EBI, con el ID de acceso PXD030983 (https://www.ebi.ac.uk/pride/archive/projects/PXD030983).

Los archivos descargados fueron los siguientes:
01938_GG2_P020143_S00_R10_R1.raw
01938_GG1_P020143_S00_R07_R1.raw
01938_GF4_P020143_S00_R27_R1.raw
01938_GF3_P020143_S00_R22_R1.raw
01938_GF1_P020143_S00_R06_R1.raw
01938_GF2_P020143_S00_R11_R1.raw

Los datos crudos de proteómica fueron analizados utilizando el software *MaxQuant* (https://www.maxquant.org/). Las muestras se cargaron individualmente como muestras independientes dentro de un mismo grupo experimental. Se emplearon los parámetros por defecto del programa, en donde se incluyen FDR menor a 0,01, cuatro corridas por muestra para mejorar el análisis y quedarnos con los resultados más confiables, al menos un péptido por proteína, y se tomó en cuenta que las muestras habían sido previamente digeridas con tripsina.

Para el análisis se empleó una referencia personalizada, la que fue generada durante la anotación primaria utilizando *Trinotate* [162]. Dentro de la misma se incluyen todas las proteínas codificantes anotadas, así como también las traducciones *in silico* de los transcriptos con alto potencial codificante. Con esta referencia se generaron péptidos trípticos *in silico*, los cuales fueron empleados para el “macheo” con el resultado obtenido por espectrometría de masas descargado previamente, de modo de realizar la identificación de los péptidos.

4. RESULTADOS

Con el objeto de contribuir a la comprensión a nivel molecular del complejo proceso de espermatogénesis en los mamíferos, en trabajos de investigación previos del grupo hemos analizado los transcriptomas codificantes de proteínas y de ARNs no codificantes largos (lncRNAs) durante la progresión del proceso de la espermatogénesis del ratón, modelo mamífero por excelencia. Este análisis se benefició de la metodología desarrollada y estandarizada por el grupo para el aislamiento de poblaciones celulares de diversas etapas de la espermatogénesis, mediante citometría de flujo [108,109,167,168]. Como hemos mencionado anteriormente, una gran ventaja de este desarrollo es la obtención de las diversas poblaciones celulares con altísimo nivel de pureza, lo que, dada la sensibilidad de las metodologías de secuenciación masiva, resulta esencial para garantizar la confiabilidad de los datos obtenidos, eliminando la posibilidad de asignación errónea de transcriptos a otros estadios por posible contaminación con otros tipos celulares. Otra gran ventaja es la inclusión, de forma innovadora, de una fracción de células altamente purificadas de leptoteno-cigoteno (LZ) [89,148]. La incorporación de esta fracción ha permitido una caracterización más minuciosa de la evolución de la profase meiótica y, particularmente, la asignación de transcriptos a la profase temprana, una fase de corta duración que hasta ahora había sido muy poco explorada a nivel molecular, principalmente por la dificultad de obtener estos tipos celulares en forma aislada. Sin embargo, estos análisis previos del grupo se centraron exclusivamente en genes anotados y, además, no abordaron el estudio de isoformas ni variantes de *splicing*.

Para el presente trabajo de tesis, nos basamos en una serie de premisas:

1. Existen reportes que indican que el testículo es el tejido que expresa mayor número de genes [58,112], y el tejido, o uno de los dos tejidos (considerando cerebro), con mayores niveles de *splicing alternativo* [80,124,130];

2. Existe muy escasa información a nivel masivo sobre la expresión diferencial de variantes de *splicing* e isoformas proteicas a lo largo de la espermatogénesis;

3. El testículo es el tejido con mayor expresión de lncRNAs, pero la identificación y caracterización de los mismos, y especialmente a lo largo de las distintas etapas de la espermatogénesis, es aún muy preliminar [58,59,112];

4. A pesar de que la profase meiótica temprana es una fase de muy corta duración, con poca representación en el total de células del testículo y, a causa de ello, su estudio ha sido comparativamente poco abordado, es muy relevante en el desarrollo de la reproducción sexuada, dados los eventos únicos y fundamentales que tienen lugar durante la misma.

En consecuencia, y con el propósito de profundizar el conocimiento y obtener una visión más integral de la complejidad y funcionalidad de la expresión génica durante la espermatogénesis, en este trabajo hemos analizado en detalle los datos brutos provenientes de trabajos anteriores de nuestro grupo, pero con un enfoque diferente. En particular, nos centramos en la identificación de genes expresados no anotados previamente, especies de ARN específicas, y proteínas putativas de las distintas etapas, hasta ahora no reportadas. Además, hemos analizado el *splicing* alternativo y sus variaciones a lo largo de la espermatogénesis, incorporando así una capa adicional de información sobre la diversidad de isoformas génicas. En este sentido, nuestra contribución busca enriquecer la comprensión global de los procesos que subyacen a la complejidad biológica de la espermatogénesis.

4.1. Diversidad transcriptómica y sus variaciones a lo largo de la espermatogénesis

4.1.1. Confiabilidad y reproducibilidad de nuestros datos

Iniciamos nuestros análisis realizando una matriz de correlación que reveló una alta reproducibilidad entre las réplicas biológicas, como se ilustra en la Figura 8. Además, comparamos nuestros datos con los de un estudio previo de referencia en la temática, que empleó secuenciación de ARN de células únicas para explorar 20 subtipos distintos de células espermatogénicas [90]. Es importante señalar que, además de las diferencias en la metodología de secuenciación masiva, otra diferencia importante entre dicho trabajo y el nuestro, es que en el mencionado trabajo se logró la obtención de todos esos tipos celulares mediante la combinación de marcaje transgénico con una sincronización artificial del ciclo del epitelio seminífero a través del uso simultáneo de un compuesto químico (WIN 18.446) y ácido retinoico. Nosotros, por el contrario, hemos trabajado en condiciones más naturales y, por ende, consideramos que nuestra asignación de transcriptos a las distintas etapas debería tener menos afectación por manipulación externa. Deseamos señalar, asimismo, que más allá de la "limitación" señalada, la elección de los datos de dicho trabajo para comparar con los nuestros se debió a que es el único reporte de scRNA-seq en el que se trabajó con poblaciones celulares puras de las distintas etapas de la espermatogénesis. A pesar de las diferencias, observamos una notable correlación entre ambos estudios, como se muestra en el análisis de componentes principales (Figura 9).

Una vez realizadas estas comparaciones, pudimos ver que estábamos en condiciones de proseguir con los siguientes análisis y asumir que los resultados son reales, y no un artefacto creado por el método de análisis o los datos empleados.

Es importante destacar que nuestros datos representan un conjunto profundamente detallado de lecturas con una sólida reproducibilidad, una característica que los hace útiles para una gran variedad de análisis. Esto es especialmente relevante para la identificación y caracterización de transcriptos de baja expresión. El nivel de profundidad y consistencia en los datos fortalece la robustez de nuestras

observaciones, respaldando la confiabilidad de los resultados obtenidos durante los análisis subsecuentes.

	2C			LZ			PS			RS		
2C	1,00	0,91	0,90	0,64	0,63	0,64	0,52	0,51	0,50	0,44	0,37	0,36
	0,91	1,00	0,90	0,66	0,65	0,66	0,54	0,54	0,52	0,45	0,39	0,38
	0,90	0,90	1,00	0,64	0,64	0,64	0,51	0,51	0,50	0,44	0,37	0,36
LZ	0,64	0,66	0,64	1,00	0,88	0,88	0,79	0,78	0,78	0,60	0,54	0,53
	0,63	0,65	0,64	0,88	1,00	0,89	0,77	0,77	0,77	0,60	0,54	0,54
	0,64	0,66	0,64	0,88	0,89	1,00	0,75	0,75	0,74	0,62	0,57	0,57
PS	0,52	0,54	0,51	0,79	0,77	0,75	1,00	0,94	0,92	0,69	0,67	0,67
	0,51	0,54	0,51	0,78	0,77	0,75	0,94	1,00	0,93	0,69	0,68	0,67
	0,50	0,52	0,50	0,78	0,77	0,74	0,92	0,93	1,00	0,68	0,67	0,66
RS	0,44	0,45	0,44	0,60	0,60	0,62	0,69	0,69	0,68	1,00	0,88	0,88
	0,37	0,39	0,37	0,54	0,54	0,57	0,67	0,68	0,67	0,88	1,00	0,94
	0,36	0,38	0,36	0,53	0,54	0,57	0,67	0,67	0,66	0,88	0,94	1,00

Figura 8. **Matriz de correlación** de expresión de RNA-seq entre las cuatro poblaciones celulares empleadas en el presente estudio, cada una con tres réplicas biológicas. La matriz muestra los coeficientes de correlación de Pearson entre muestras, con valores altos (en rojo) que indican una alta similitud en los perfiles de expresión génica. Se observa una excelente correlación entre las réplicas de cada población celular, lo que respalda la consistencia y reproducibilidad de los datos de expresión dentro de cada grupo.

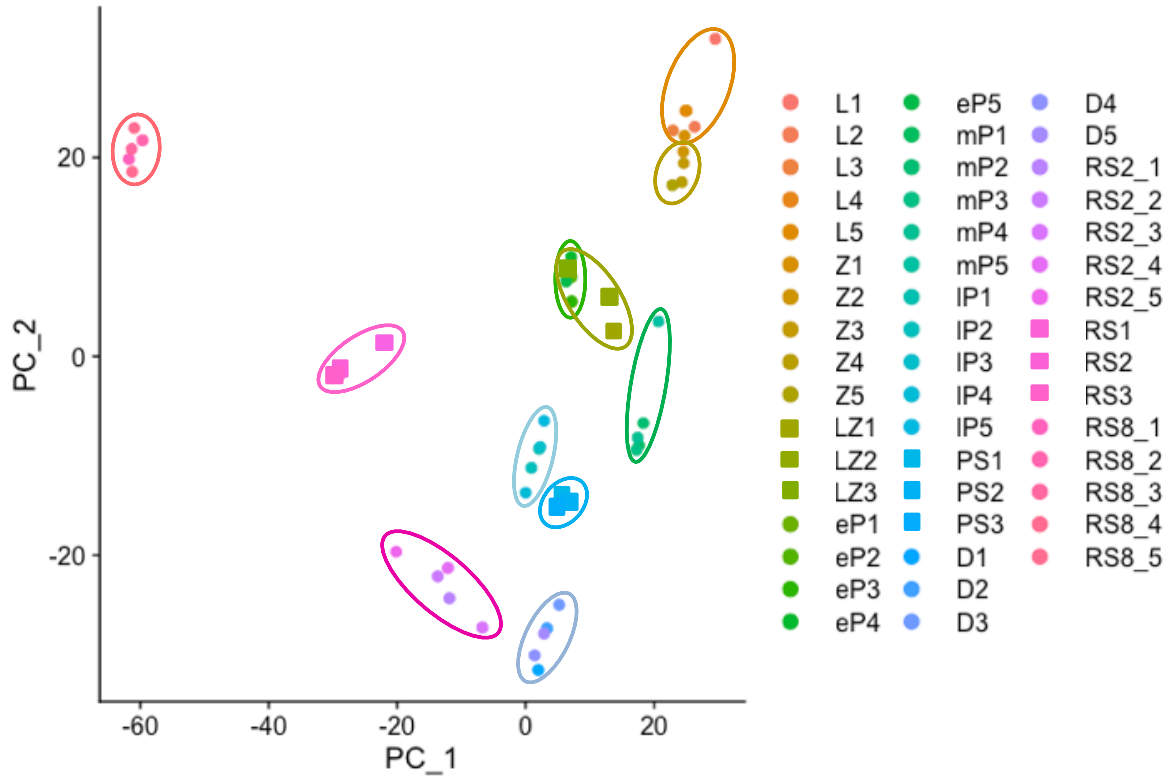


Figura 9. Análisis de componentes principales (PCA) comparando nuestros datos de RNA-seq con un estudio de secuenciación de ARN de células únicas (sc-RNAseq), de 20 subtipos celulares diferentes de la espermatogénesis (Chen *et al.,* 2018). Las poblaciones celulares de nuestro estudio están representadas con cuadrados, mientras que las poblaciones del estudio de sc-RNAseq están indicadas con círculos. Cabe destacar que la correlación es muy buena, considerando que hubo varias diferencias entre las condiciones de ambos experimentos. Es importante mencionar que los datos de nuestra población de células 2C no se incluyeron en la comparación, ya que, además de espermatogonias, contienen células somáticas testiculares que no fueron incluidas en el estudio de Chen *et al*. L: leptoteno; Z: cigoteno; LZ: lepto/cigoteno; eP: paquiteno temprano; mP: paquiteno medio; lP: paquiteno tardío; PS: espermatocitos en paquiteno; D: diploteno; RS: espermátidas redondas.

4.1.2. Identificación de transcriptos sin anotación

Para los análisis subsiguientes, aplicamos rigurosos criterios estableciendo umbrales específicos. Estos criterios incluyeron una cobertura mínima de 10X, un soporte mínimo de 10 lecturas por sitio de empalme, y un mínimo de 10 lecturas por soporte de exón. Bajo estas condiciones selectivas, identificamos un total de 37.793 genes expresados en los testículos que superaron todos los filtros aplicados, los cuales se pueden apreciar en la Tabla S2. De este conjunto, 21.156 genes (56%) ya estaban previamente anotados en las bases de datos, mientras que 16.637 genes (44%) no se encontraban anotados, como se ilustra en la Figura 10A. Estos 37.793 genes generaron 81.139 transcriptos distintos, de los cuales 48.137 (59%) ya contaban con anotaciones, mientras que 33.002 (41%) correspondían a transcriptos previamente no informados (Figura 10B).

Luego, dirigimos nuestra atención hacia la caracterización masiva de los 33.002 transcriptos no anotados. Dentro de este grupo, identificamos 14.667 transcriptos (44%) como variantes de *splicing* previamente no reveladas de genes conocidos, mientras que 18.335 transcriptos (56%) correspondieron a transcriptos emergentes de regiones no anotadas del genoma (Figura 10C y Figura 11).

Este análisis resalta la existencia de un número importante de genes y variantes de *splicing* expresados en el testículo aún por descubrir, evidenciando la complejidad y la diversidad subyacente a la expresión génica en este contexto biológico.

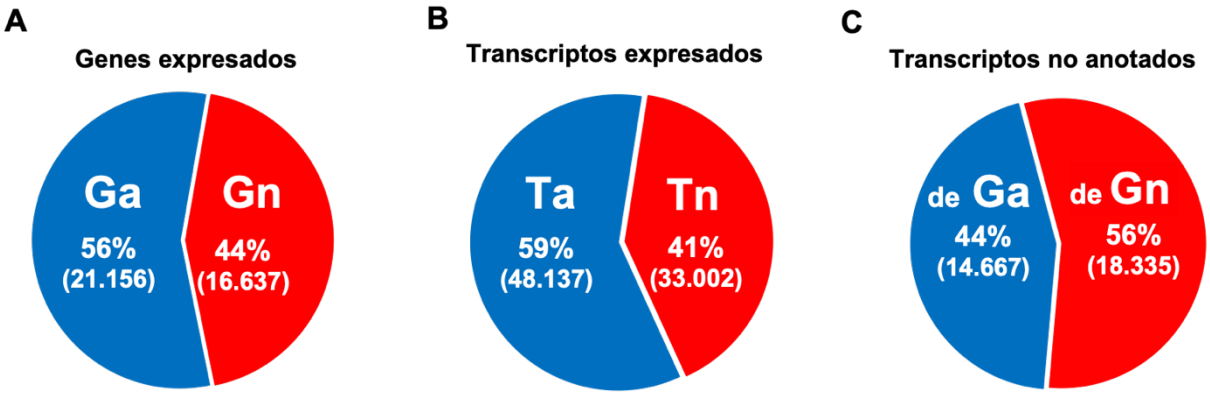


Figura 10. Proporción de genes y transcriptos expresados en las cuatro poblaciones celulares espermatogénicas. **A)** Gráfico circular que muestra la distribución de genes anotados (Ga: azul) y genes no anotados (Gn: rojo) expresados en estas poblaciones celulares. **B)** Gráfico circular de transcriptos anotados (Ta: azul) y transcriptos no anotados (Tn: rojo). **C)** Gráfico circular que ilustra el origen de los transcriptos no anotados en nuestras listas, destacando si se trata de variantes de *splicing* de genes ya anotados (de Ga: azul) o de transcriptos derivados de genes no anotados (de Gn: rojo).

4.1.3. Clasificación de los transcriptos sin anotación

Continuamos con un análisis del potencial codificante de los transcriptos no anotados. Durante este análisis, empleamos cuatro programas diferentes de análisis de potencial codificante y conservamos únicamente los resultados que eran consistentes entre sí, es decir, aquellos para los cuales los cuatro programas coincidían en su predicción de capacidad codificante o no codificante. Los que tuvieron clasificaciones discordantes en al menos uno de los programas, no fueron considerados en los subsiguientes análisis. La concordancia de los cuatro programas identificó 13.471 transcriptos como no codificantes (Figura 12A) y 2.794 como codificantes (Figura 12B). La Tabla S2 presenta en detalle los resultados individuales obtenidos con cada uno de los programas utilizados, proporcionando una visión más precisa de las clasificaciones realizadas.

Por lo tanto, del grupo de transcriptos sin anotación con los que proseguiremos el análisis, la mayoría resultó ser no codificante, una observación previsible dado que el

genoma codificante ha sido más exhaustivamente caracterizado en comparación con el no codificante.

Es relevante destacar que nuestro criterio, al ser muy riguroso, excluyó aproximadamente la mitad de los transcriptos (16.737; ver Figura 11). No obstante, por otro lado, nos permitió trabajar con un conjunto altamente confiable en términos de su potencial codificante. Dicho esto, cabe aclarar que para algunos análisis puntuales que se mostrarán más adelante, en los que el potencial codificante no era relevante, se empleó el conjunto completo de los 33.002 transcriptos no anotados, con el fin de tener un panorama más amplio del comportamiento de esos transcriptos.

El análisis del potencial codificante de los transcriptos arrojó diferencias en la cantidad de secuencias clasificadas como codificantes, según el programa utilizado. En orden decreciente, *TransDecoder* identificó la mayor cantidad de transcriptos con potencial codificante, seguido por *CPC2*, *LncDeep* y, finalmente, *CPAT*, que predijo la menor cantidad. Estas diferencias pueden atribuirse a la metodología específica de cada herramienta. *TransDecoder* basa su clasificación en la identificación de ORFs y la presencia de dominios proteicos conservados, lo que, en este caso, permitió detectar un mayor número de transcriptos codificantes. *CPC2* y *LncDeep*, al emplear modelos de aprendizaje automático basados en características estadísticas de la secuencia, presentan mayor flexibilidad, aunque con criterios distintos de clasificación. Por otro lado, *CPAT*, también basado en aprendizaje automático, mostró mayor restricción en la identificación de transcriptos codificantes, posiblemente debido a las características del modelo de referencia utilizado y su sensibilidad a la estructura de la secuencia. Estas discrepancias resaltan la importancia de considerar múltiples enfoques al evaluar el potencial codificante de transcriptos en estudios transcriptómicos.

De entre los 13.471 transcriptos sin potencial codificante no anotados, la abrumadora mayoría (12.297, es decir, 91%) correspondió a genes (específicamente, lncRNAs) no reportados, mientras que menos del 10% (1.174) representaron variantes de *splicing* no reportadas de genes no codificantes ya anotados (Figura 12C; ver también Figura 11).

Por otra parte, el número identificado de transcriptos espermatogénicos no anotados con un alto potencial codificante de proteínas, 2.794, es un número nada despreciable. A diferencia de lo ocurrido con los transcriptos no codificantes, la mayoría de ellos (2.571) se asoció con nuevas variantes de *splicing* de genes ya anotados. Esto era esperable, ya que es de suponer que la mayor parte de los genes codificantes para proteínas del genoma del ratón ya han sido anotados. Por otro lado, algo menos del 10% de los transcriptos "novedosos" presumiblemente codificantes (223), resultó corresponder a transcriptos de genes no anotados (Figura 12D; ver también Figura 11). Sorprendentemente, estos 223 transcriptos provenían de 191 genes aún no anotados en el genoma del ratón (Tabla S3). Estos resultados señalan la existencia de un número sustancial de genes presuntamente codificantes de proteínas en el genoma del ratón, que se expresan en células espermatogénicas, y que hasta ahora no se habían identificado.

En suma, hasta aquí hemos identificado 33.002 transcriptos no anotados, de los cuales para 13.471 se pudo determinar que no son codificantes, y para 2.794 se determinó su alto potencial codificante de proteínas. En tanto la mayoría de los transcriptos no codificantes provienen de genes “nuevos” (no anotados), la mayoría de los transcriptos codificantes corresponden a variantes de *splicing* no reportadas, originadas a partir de genes ya conocidos. No obstante ello, identificamos 191 genes codificantes, expresados en las células espermatogénicas, que no estaban aún anotados en el genoma del ratón.

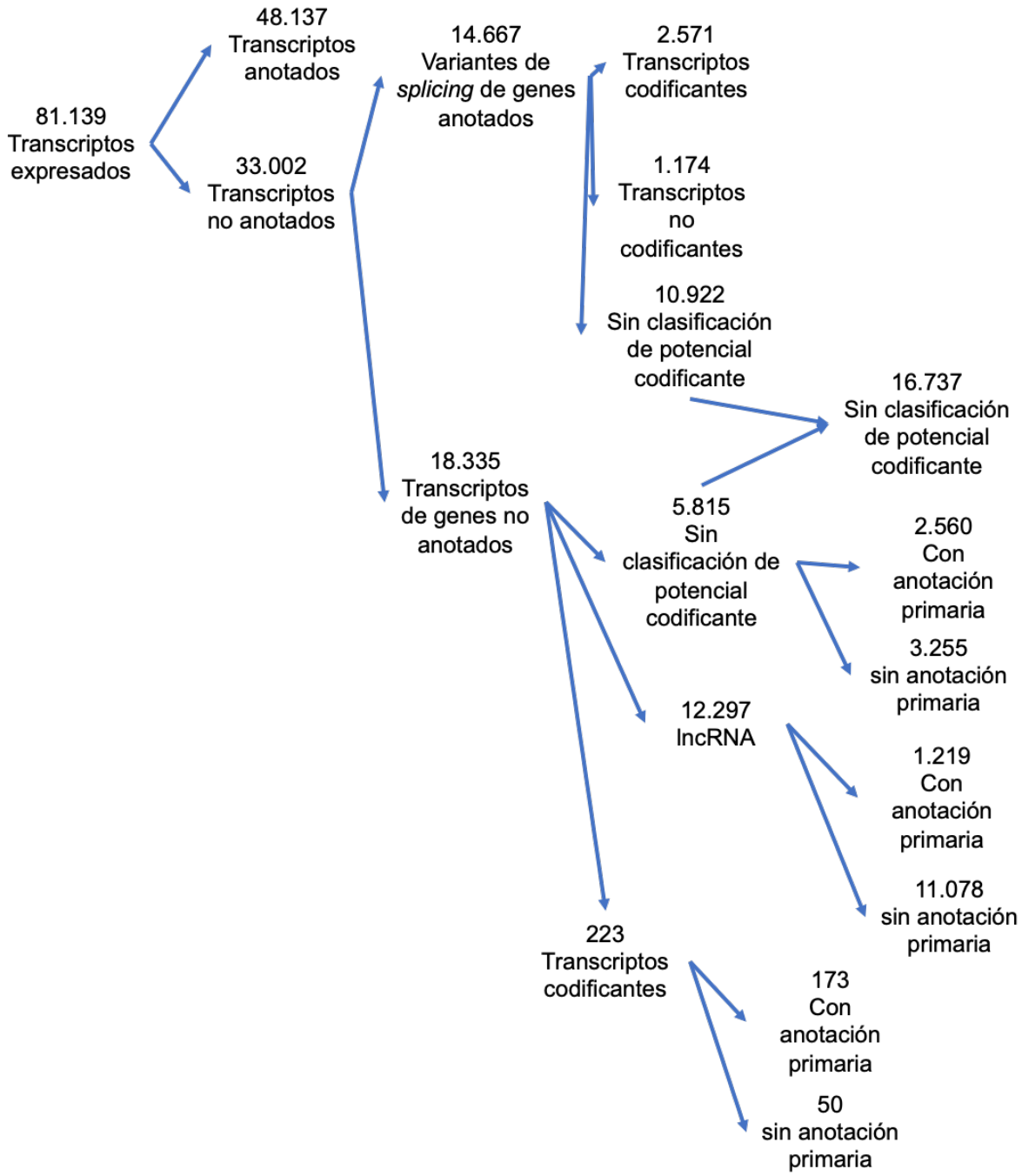


Figura 11. Esquema de distribución de transcriptos en las diferentes categorías empleadas a lo largo de las secciones 4.1.2 a 4.1.4 de la tesis. La distribución comienza en todos los transcriptos expresados en nuestras listas, y se va bifurcando a medida que se van generando las diferentes categorías a lo largo de los análisis.

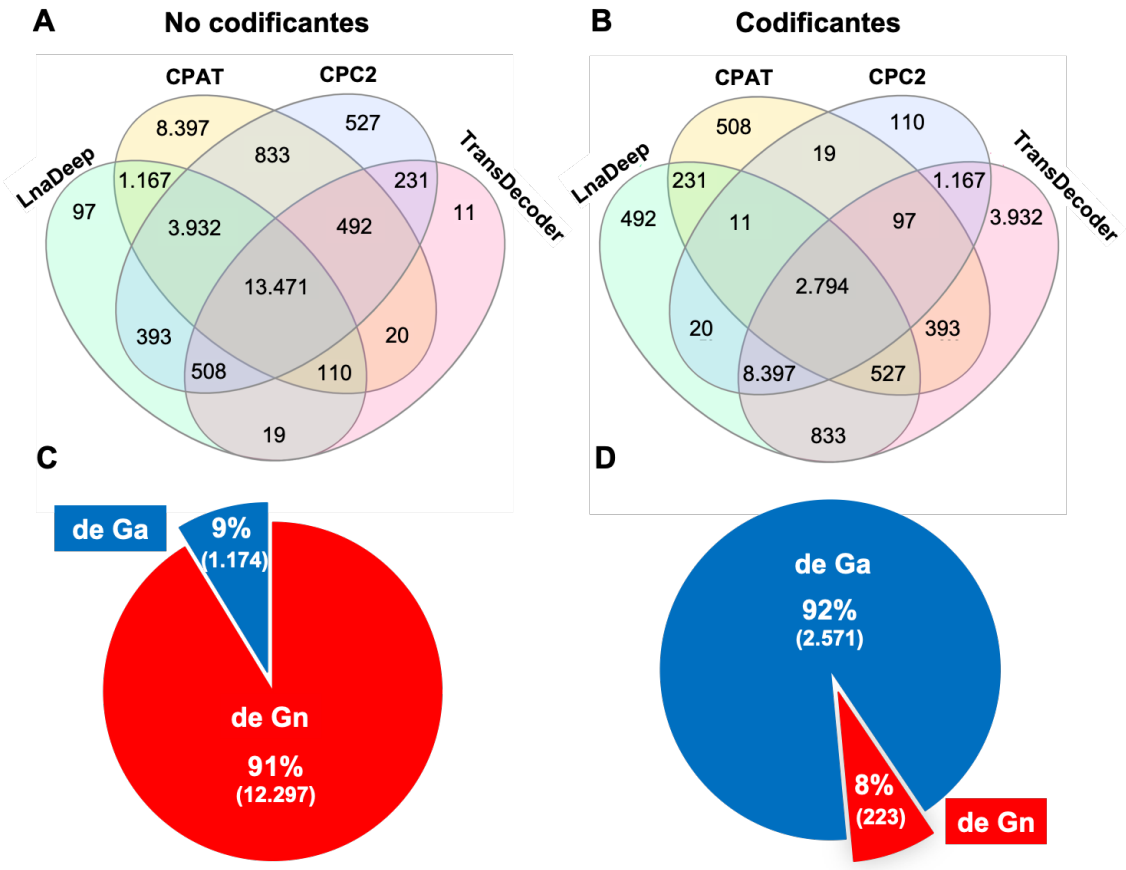


Figura 12. Potencial codificante de los transcriptos no anotados. A, B) Diagramas de Venn que muestran el análisis del potencial codificante de los transcriptos no anotados, evaluado mediante cuatro programas diferentes. **C, D)** Gráficos circulares de los transcriptos no codificantes y potencialmente codificantes que fueron identificados de forma coincidente como tales por los cuatro programas, clasificados en variantes de *splicing* no descritas de genes previamente anotados (de Ga: azul) y transcriptos de genes no anotados (de Gn: rojo). A, C: transcritos no codificantes; B, D: transcriptos codificantes.

4.1.4. Anotación primaria de transcriptos

Centrándonos en los 18.335 transcriptos provenientes de regiones no anotadas del genoma (ver Figuras 10C y 11), nos preguntamos hasta qué punto los mismos estarían relacionados a nivel de secuencia con genes ya anotados. Para ello, realizamos una anotación primaria con *Trinotate*, complementada con búsqueda manual en *BLASTP*. Como resultado, observamos que 3.952 de estos transcriptos presentaron similitud con genes ya anotados principalmente del ratón, rata o humano

(Figura 13A y Tabla S4). La anotación se basa en la suposición de que éstos podrían ser genes no identificados en el ratón, pero homólogos de genes ya anotados en esas otras especies o emparentados con éstos, o incluso corresponder a transcriptos de genes emparentados con otros ya anotados en el propio genoma del ratón (por ejemplo, pertenecientes a una misma familia).

Dentro de estos 3.952 transcriptos, 1.219 (producto de 1.153 genes) habían sido previamente clasificados como no codificantes mediante el uso de los cuatro programas de potencial codificante, y 173 transcriptos (producto de 162 genes) fueron clasificados como codificantes (Figura 13B), en tanto el resto no pasó el estricto criterio para determinación de potencial codificante (ver también Figura 11).

Por otro lado, de los restantes 14.383 transcriptos (que provenían de 13.194 genes) que no presentaron ninguna similitud con secuencias genómicas ya reportadas en las bases de datos, 11.128 habían logrado pasar el criterio para su clasificación con los cuatro programas de potencial codificante. De entre ellos, 11.078 (provenientes de 10.484 genes) habían sido clasificados como no codificantes, en tanto únicamente 50 transcriptos (derivados de 29 genes) fueron identificados como codificantes (Figura 13C, y Figura 11). Esto sugiere que la gran mayoría de los transcriptos no anotados que no encuentran siquiera similitud alguna con otros genes ya anotados en el propio genoma del ratón, ni en genomas de otros mamíferos, seguramente corresponda a lncRNAs. Más aún, esto podría obedecer en parte a la baja conservación de lncRNAs entre especies, y, en parte, a la pobre anotación de lncRNAs existente hasta la fecha.

A continuación, enfocamos nuestra atención en los 173 transcriptos (procedentes de 162 genes) con alto potencial codificante no anotados para los cuales pudimos encontrar cierta similitud con genes ya reportados, y llevamos a cabo análisis funcionales basados en dichas similitudes encontradas en ratón y otras especies. Identificamos similitudes significativas para algunas de las supuestas proteínas codificadas, destacándose conexiones con proteínas ribosómicas, proteínas con dedos de zinc, y con la proteína 92 asociada a cilios y flagelos (Tabla S3). Además, notamos que un subconjunto considerable, abarcando más de la mitad de estos genes, correspondía a productos de virus integrados y retroposones.

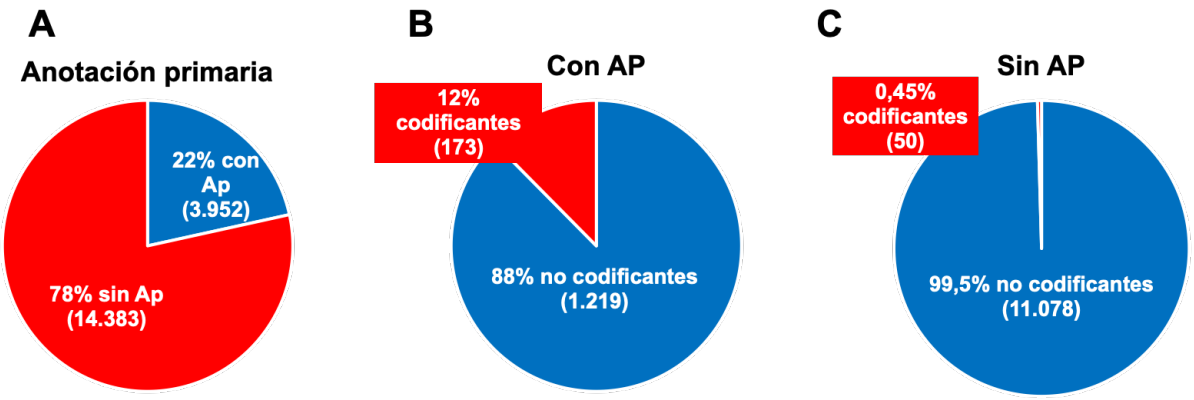


Figura 13. Anotación primaria de transcriptos de regiones no anotadas. A) De los 18.335 transcriptos analizados, el 22% presentó una anotación primaria (Ap), mientras que el 78% no mostró similitud con genes previamente anotados. **B**) Dentro de los 3.952 transcriptos con Ap, el 88% fueron clasificados como no codificantes, mientras que el 12% correspondió a transcriptos codificantes. **C**) De los 14.383 transcriptos sin Ap, casi el 100% de los transcriptos clasificados fueron no codificantes, con sólo 50 transcriptos presentando un alto potencial codificante.

Finalmente, es importante recordar que los 173 transcriptos mencionados, clasificados como codificantes, constituyen un subgrupo de los 223 transcriptos producto de 191 genes no anotados, con alto potencial codificante, mencionados en la sección 4.1.3. Se desprende, entonces, que para 29 de estos 191 genes presuntamente codificantes para proteínas (191 menos 162) y sus 50 transcriptos codificantes (223 menos 173) mostrados en la Figura 13C, no pudimos asociar ninguna función probable conocida en lo que respecta a la anotación primaria, ni siquiera por similitud con genes reportados en otras especies. En análisis posteriores profundizaremos de forma manual en el estudio de estos genes.

4.1.5. Expresión de los transcriptos no anotados a lo largo de las diferentes etapas de la espermatogénesis

Como siguiente paso, decidimos analizar la expresión de los transcriptos recién identificados, según su distribución en las diferentes poblaciones celulares. Inicialmente, comparamos los niveles de expresión de los transcriptos no anotados

con aquellos previamente anotados en la base de datos Ensembl, para cada una de las cuatro poblaciones celulares. De forma general, se observó que los niveles medios de expresión de los transcriptos no anotados, ya fueran codificantes o no codificantes, resultaron ser inferiores a los de los ya anotados, una tendencia que se mantuvo constante en todas las poblaciones celulares (Figura 14). Este hallazgo sugiere que al menos buena parte de los transcriptos no anotados podrían haber pasado desapercibidos en estudios anteriores, precisamente debido a su menor expresión.

Por otra parte, los transcriptos no codificantes presentaron, en general, niveles de expresión más bajos en comparación con los codificantes en todas las poblaciones celulares (ver Figura 14). Esto concuerda con investigaciones previas que han señalado que los lncRNAs tienden a exhibir niveles menores de expresión, en comparación con los ARNm [51,64].

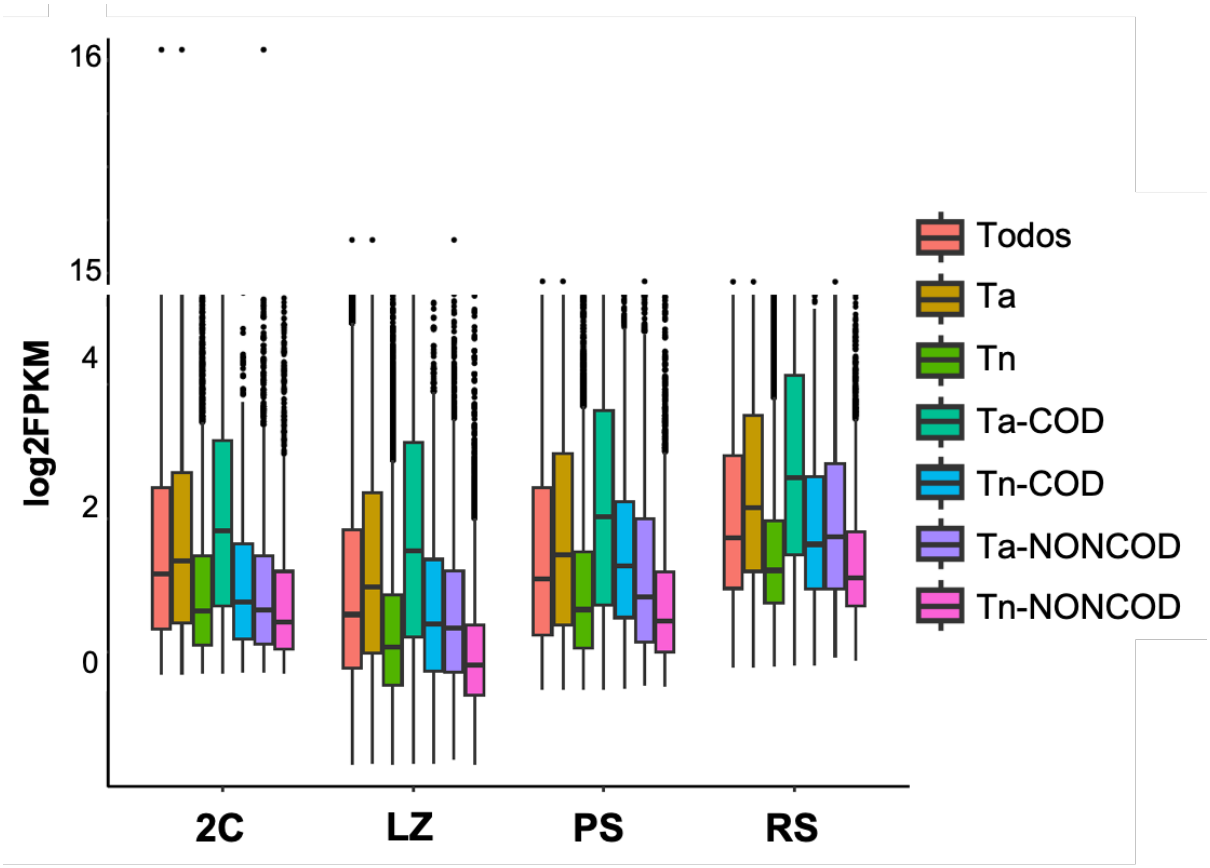


Figura 14. Distribución de expresión de los transcriptos en las cuatro poblaciones celulares testiculares. Gráfico de caja que muestra los niveles de expresión (log_2 FPKM) de todos los transcriptos detectados. Las categorías incluyen: *Todos*, que representa todos los

transcriptos detectados; *Ta*, que son los transcriptos anotados; *Tn*, los transcriptos no anotados; *Ta-COD*, transcritos anotados codificantes; *Tn-COD*, transcriptos no anotados codificantes; *Ta-NONCOD*, transcriptos anotados no codificantes; y *Tn-NONCOD*, transcriptos no anotados no codificantes. Este análisis permite observar la variación en la expresión entre tipos de transcriptos, y distinguir la contribución relativa de los transcriptos anotados y no anotados en cada población celular testicular.

En cuanto al número de transcriptos en cada población celular, la mayor contribución de los no anotados se observó en la etapa LZ, tanto para los codificantes como para los no codificantes (Figura 15A). Llamativamente, el 55% de los transcriptos no anotados expresados en LZ resultó ser exclusivo de esta etapa (Figura 15B, y Figuras 16A y B). Adicionalmente, al examinar específicamente los genes codificantes no anotados, se observó que 159 de los 191 identificados se expresaron en LZ, siendo casi la mitad de ellos (92 genes) exclusivos de esta etapa (ver Tabla S3).

El hallazgo de un mayor número de transcriptos no anotados pertenecientes a LZ, nos planteó la interrogante de si este fenómeno estaría reflejando un mayor número general de transcriptos expresados en LZ, o si podría ser simplemente consecuencia del hecho de que la mayor parte de los transcriptos de LZ aún están sin anotar (en ese caso, si se sumaran los transcriptos anotados a los no anotados, debería compensarse). Sin embargo, un análisis de distribución de la totalidad de los 81.139 transcriptos identificados, tanto anotados como no anotados, ratificó que LZ era el estadio con mayor número de transcriptos expresados, tanto en total, como de manera específica (Figura 16C).

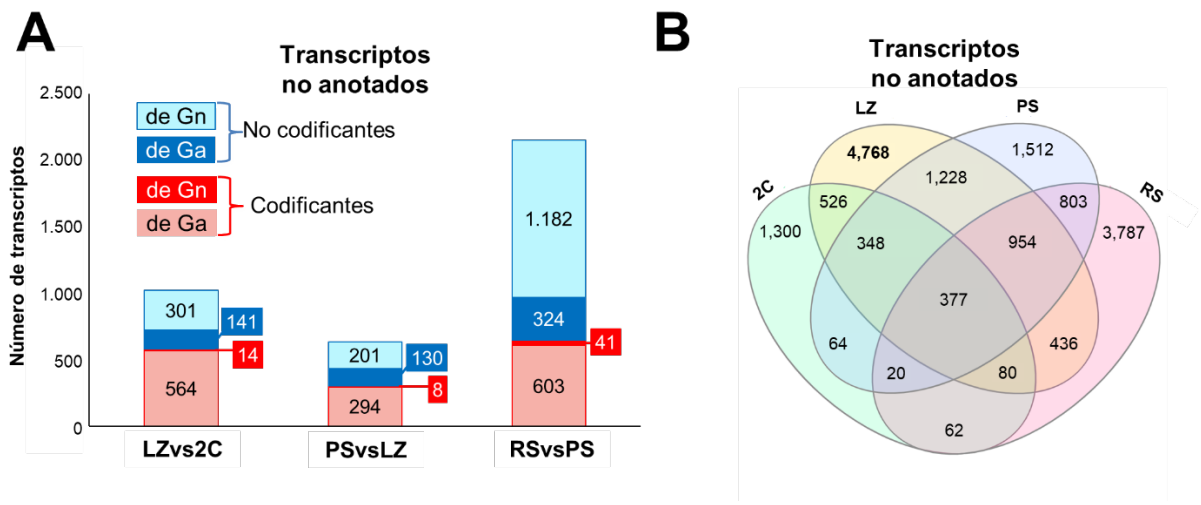
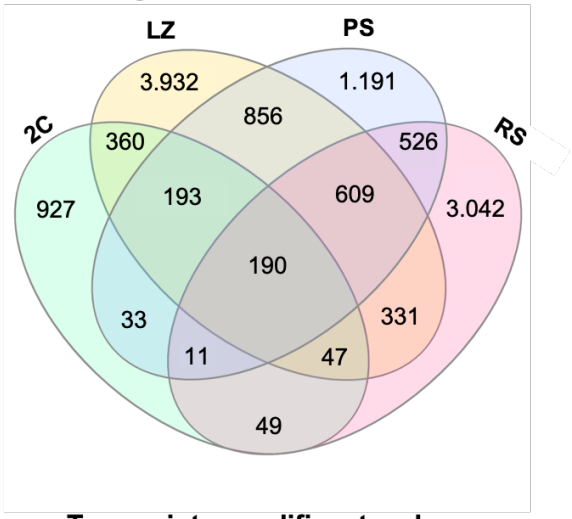


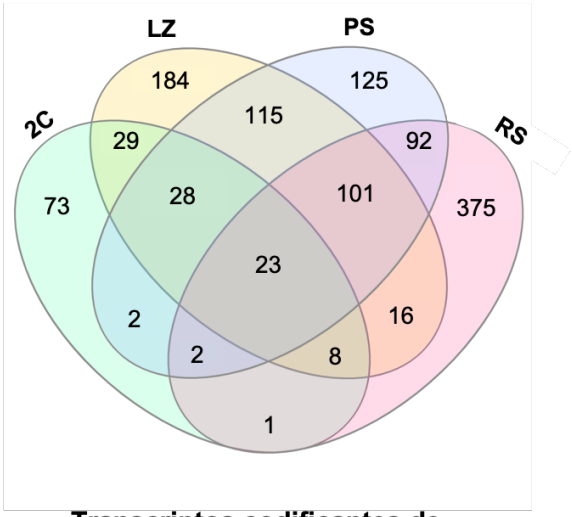
Figura 15. Distribución y clasificación de transcriptos no anotados en las diferentes poblaciones celulares testiculares. A) Transcriptos no anotados que fueron identificados de manera coincidente como tales por los cuatro programas utilizados para el análisis de potencial codificante (y representados en la Figura 11), distribuidos según su expresión en cada una de las cuatro poblaciones celulares testiculares. Los transcriptos se clasifican en codificantes o no codificantes, y en transcriptos de genes no anotados (de Gn) o variantes de *splicing* de genes ya anotados (de Ga). Es importante destacar que muchos transcriptos pueden estar expresados en más de una etapa, y por eso el número total supera los 16.265. **B)** Diagrama de Venn que indica la distribución de los transcriptos representados en el panel A, entre las cuatro poblaciones celulares testiculares.

A

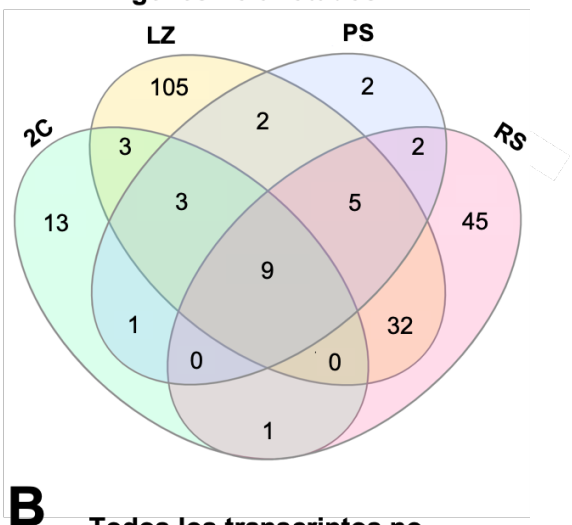
Transcriptos no codificantes de genes no anotados



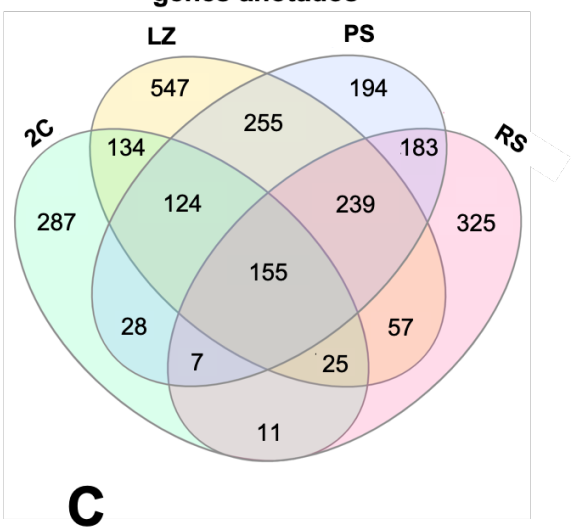
Transcriptos no codificantes de genes anotados



Transcriptos codificantes de genes no anotados

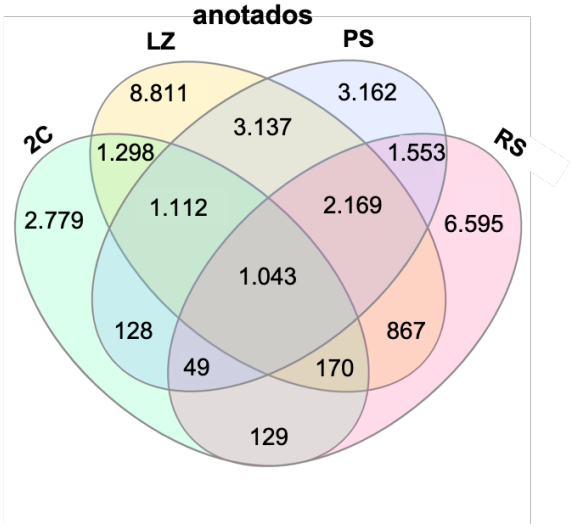


Transcriptos codificantes de genes anotados



B

Todos los transcriptos no anotados



C

Todos los transcriptos

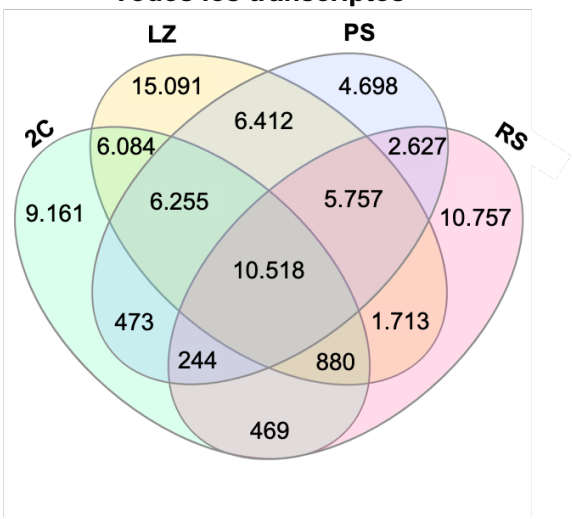


Figura 16. Distribución de los transcriptos en las cuatro poblaciones celulares testiculares. **A)** Representación de los transcriptos no anotados que fueron identificados coincidentemente como codificantes o no codificantes mediante los cuatro programas utilizados en el análisis de potencial codificante, representados en la Figura 12, y distribuidos de acuerdo a las diferentes categorías [es decir, codificantes o no codificantes; variantes de *splicing* (no anotadas) de genes ya anotados, o transcriptos de genes no anotados]. **B)** Representación de los 33.002 nuevos transcriptos identificados (antes del análisis de su potencial codificante), mostrando 6.708 transcriptos expresados en 2C; 18.607 en LZ; 12.353 en PS; y 12.575 en RS. **C)** Representación de todos los transcriptos detectados en nuestras listas (tanto los anotados como los no anotados).

Con el fin de corroborar estos resultados, y evaluar que no fueran un artefacto, realizamos un análisis de saturación de transcriptos. Dicho análisis incluyó datos del presente estudio y de uno anterior de nuestro grupo, en el que se analizó el transcriptoma de las mismas poblaciones celulares, pero empleando librerías de secuenciación no direccionales [89]. El análisis demostró que todas las poblaciones celulares alcanzaron la saturación (Figura 17A). Además, los histogramas de expresión de transcriptos entre las cuatro poblaciones presentaron una distribución similar (Figura 17B), validando así los resultados obtenidos. En resumen, estos análisis confirman que los resultados no son artefactos técnicos ni analíticos, sino reflejos genuinos de la biología subyacente.

Nuestros estudios también muestran que los transcriptos de LZ exhibieron, en general, los niveles de expresión más bajos para todos los tipos de transcriptos analizados (codificantes y no codificantes, anotados y no anotados), mientras que los transcriptos de RS presentaron los niveles de expresión más altos (ver Figura 14). Por consiguiente, se puede decir que LZ alberga un mayor número de transcriptos expresados, aunque éstos tienden a expresarse en niveles comparativamente menores.

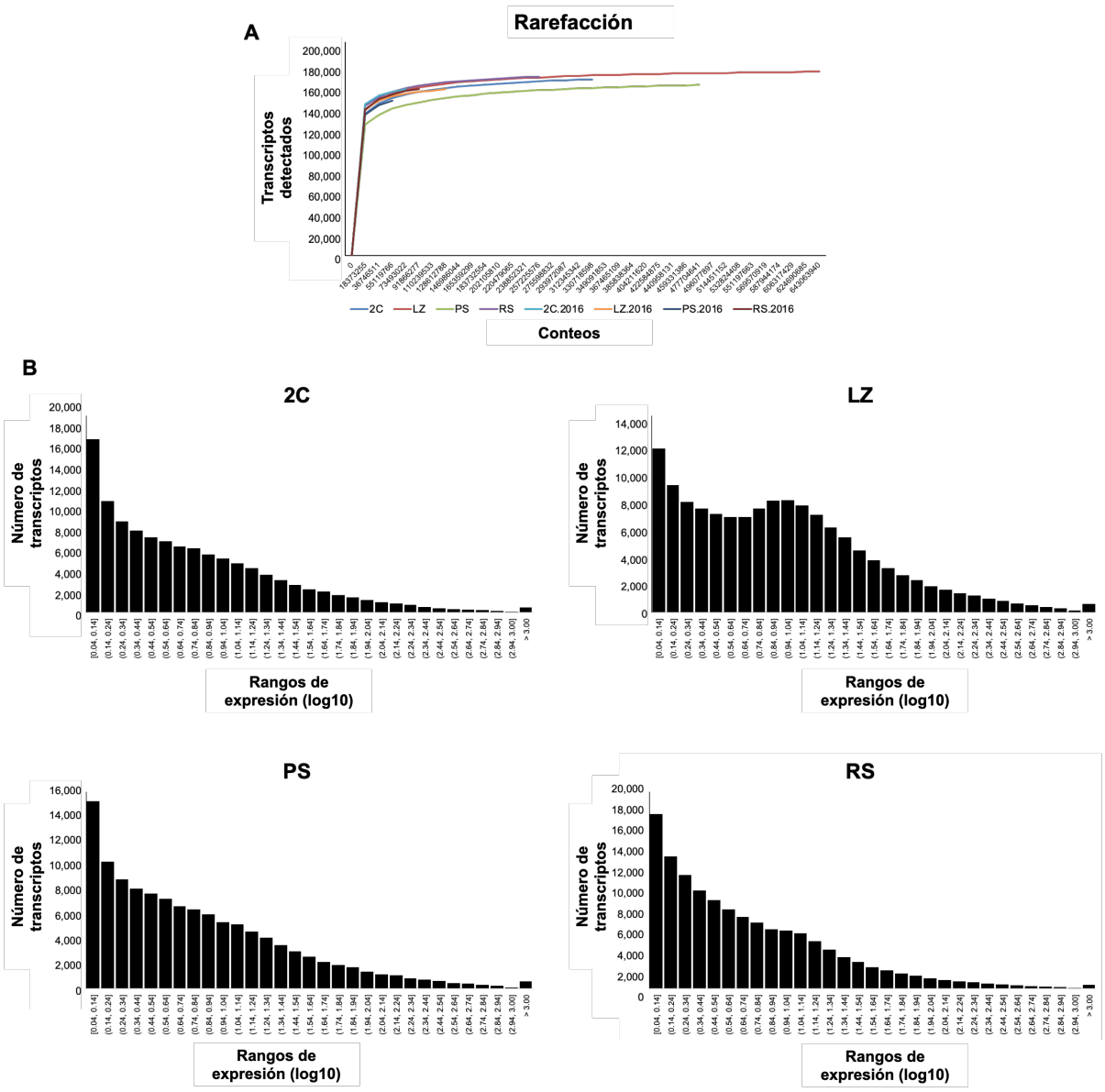


Figura 17. Saturación y distribución de expresión en las cuatro poblaciones celulares. **A)** Análisis de rarefacción basado en los transcriptos detectados a diferentes profundidades de secuenciación, incluyendo datos de da Cruz *et al*., 2016. **B)** Análisis de distribución de la expresión a través de un histograma. Para facilitar la interpretación gráfica, los valores correspondientes al percentil de expresión más bajo fueron excluidos (2C: 85.263; LZ: 68.740; PS: 84.947; RS: 76.905).

4.1.6. Expresión diferencial a lo largo de las distintas etapas de la espermatogénesis

Posteriormente, se procedió a un análisis de la expresión diferencial de los transcriptos recién identificados, realizando comparaciones pareadas a lo largo del progreso de la onda espermatogénica (log2 FC ≥ |2|, FDR < 0,05).

Observamos un mayor número de transcriptos no anotados expresados diferencialmente que cumplían con nuestros criterios en la transición de PS a RS (Figura 18), y esto es especialmente notorio para los transcriptos no codificantes. Este resultado sugiere que la transición de la profase meiótica a la espermiogénesis implica tanto la regulación negativa, como predominantemente, la regulación positiva de un gran número de genes y variantes de *splicing*, y, especialmente, de ARNs regulatorios. Por otra parte, si bien a primera vista este resultado podría parecer contradictorio con los obtenidos anteriormente, que mostraron mayor cantidad de transcriptos expresados en LZ, esta observación refleja que, como hemos visto, aunque LZ exhibe el mayor número de transcriptos exclusivos no anotados, muchos de ellos se expresan en niveles bajos y, por lo tanto, no cumplen con nuestro estricto criterio para la definición de expresión diferencial.

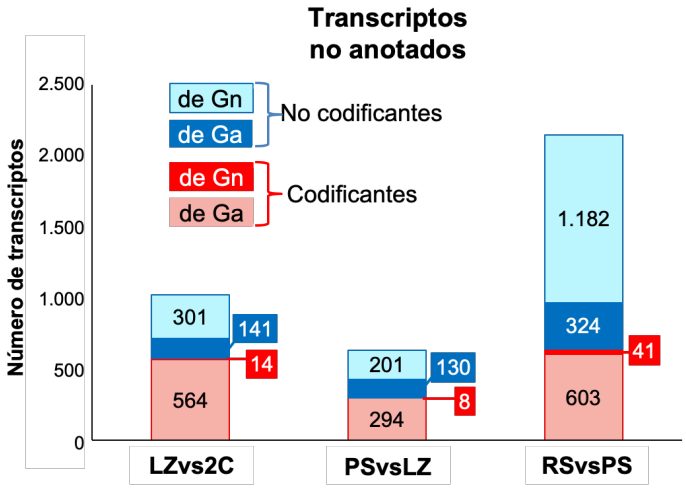


Figura 18. Comparación de la expresión diferencial (DE) de transcriptos codificantes y no codificantes entre pares de muestras de las cuatro poblaciones celulares testiculares en orden cronológico de aparición durante la onda espermatogénica. Las barras muestran el número de transcriptos detectados provenientes de genes diferencialmente expresados entre

las poblaciones de células específicas: LZ *vs* 2C, PS *vs* LZ, y RS *vs* PS. Los transcriptos se clasifican en cuatro categorías: transcriptos de genes no anotados (de Gn) y variantes de *splicing* de genes ya anotados (de Ga), tanto codificantes como no codificantes.

4.1.7. Caracterización de los tipos de *splicing* alternativo a lo largo de la espermatogénesis

A continuación, nos propusimos investigar qué tipos de *splicing alternativo* eran más predominantes a lo largo de la espermatogénesis, y si se producían cambios en ese sentido durante el proceso, es decir, si determinados tipos de *splicing* alternativo eran característicos de ciertas etapas específicas. Con ese fin, procedimos a caracterizar las variantes de *splicing* identificadas en nuestras listas, y utilizamos para ello tanto los transcriptos con anotación como los sin anotación. El análisis de los diferentes tipos de *splicing* alternativo en las cuatro poblaciones de células testiculares lo efectuamos mediante el software *rMATS*, utilizando las diferentes categorías de *splicing* alternativo definidas por el software: salto de exón (SE), sitio de empalme 5' alternativo (A5SS), sitio de empalme 3' alternativo (A3SS), exones excluyentes mutuamente (MXE), y retención de intrón (RI).

No observamos un enriquecimiento significativo en algún tipo de *splicing* alternativo en alguna etapa específica, al comparar los eventos entre las etapas estudiadas. Por el contrario, se destacó que SE fue el tipo de *splicing* más representado en las cuatro etapas, seguido de RI. Les siguieron A3SS, A5SS y MXE en ese orden, respectivamente, en las cuatro poblaciones de células testiculares (Figura 19).

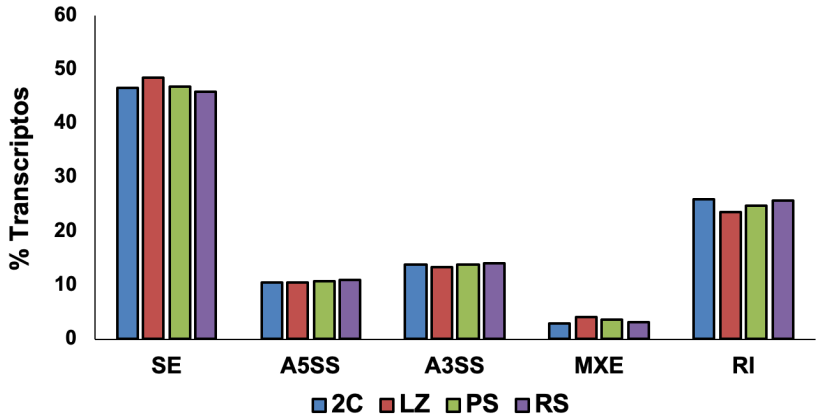


Figura 19. Análisis de los diferentes tipos de *splicing* alternativo (AS) a lo largo de la espermatogénesis. Gráfico de barras que muestra la distribución porcentual de los diferentes tipos de AS en las cuatro poblaciones celulares testiculares. SE: exón omitido; A5SS: sitio de empalme alternativo 5´; A3SS: sitio de empalme alternativo 3´; MXE: exones mutuamente excluyentes; RI: intrón retenido.

4.1.8. Cantidad de variantes de *splicing* en los genes expresados durante la espermatogénesis

Posteriormente, analizamos la cantidad de variantes de *splicing* por gen para los genes expresados durante la espermatogénesis. Para ello, calculamos el porcentaje de transcriptos en cada categoría y determinamos la proporción de genes que presentaban entre 1 y 10 transcriptos. No consideramos genes con más de 10 transcriptos, ya que estos casos eran excepcionales y representaban una minoría dentro del conjunto de datos.

Encontramos que, mientras que alrededor del 60% de los genes codificantes expresan sólo un transcripto por cada gen, entre el 85% y el 95% de los genes no codificantes presentan un solo transcripto (Figura 20, $p < 10^{-10}$). Asimismo, el número de genes con dos o más isoformas de *splicing* fue mayor para los genes codificantes que para los no codificantes (Figura 20, $p < 0,01$). Esto indica claramente que los genes codificantes suelen presentan mayor número de variantes de *splicing*. Por el contrario, los genes no codificantes, en general, presentaron un menor número de transcriptos por gen.

Es importante señalar que, básicamente, tanto los genes anotados como los no anotados se comportaron de manera similar en este aspecto, y esta afirmación es válida tanto para los transcriptos codificantes como para los no codificantes (ver Figura 20). Esto constituye una nueva evidencia de la confiabilidad de nuestros datos, ya que no hay razón para sospechar que los transcriptos anotados y los no anotados debieran comportarse de manera diferente.

Por otra parte, aunque aquí mostramos los resultados agrupados de las cuatro poblaciones de células testiculares, deseamos hacer notar que también hicimos el análisis del número de transcriptos por gen para las cuatro poblaciones celulares por separado. Sin embargo, no presentamos aquí dicho análisis, dado que el mismo no mostró diferencias significativas entre las distintas poblaciones.

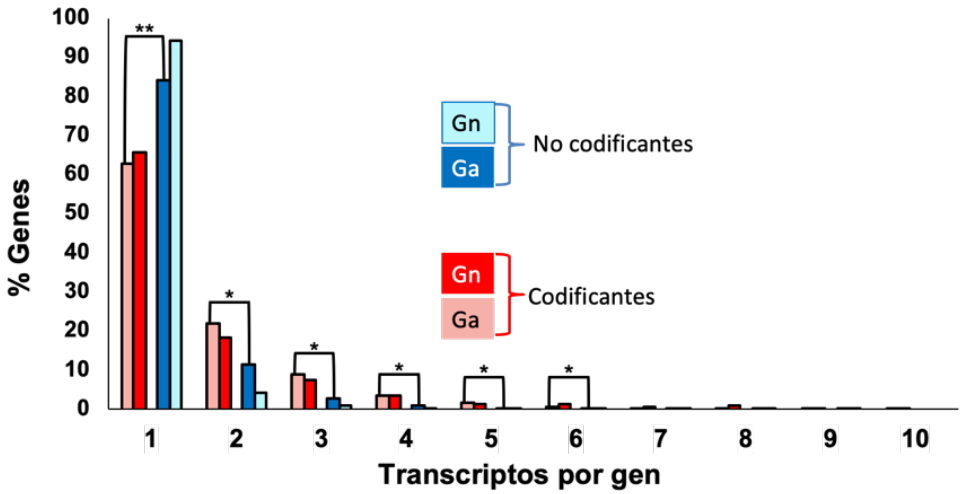


Figura 20. Análisis del número de variantes de *splicing* por gen para los genes expresados (codificantes y no codificantes) en las cuatro poblaciones celulares. Los datos están presentados como porcentaje del total. Sólo se consideraron genes con entre 1 y 10 variantes de *splicing* expresadas. Gn: genes no anotados; Ga: genes anotados. **$p < 10^{-10}$; *$p < 0.01$.

4.2. Estudios confirmatorios de transcriptos y proteínas no anotadas

4.2.1. Análisis en profundidad y confirmación de variantes de *splicing* representativas, con alto potencial codificante

Como se mencionó previamente, los niveles de expresión de los transcriptos no anotados fueron, en general, más bajos que los de los anotados (ver Figura 14). A pesar de esto, es importante destacar que algunos de los transcriptos "nuevos" identificados exhibieron niveles de expresión notablemente altos, y algunas isoformas de *splicing* alternativo no anotadas se expresaron a niveles mucho más elevados que sus isoformas ya anotadas. Suponemos que, al menos algunas de las variantes de *splicing* no anotadas que hemos identificado, con un alto potencial codificante, han de dirigir la síntesis de isoformas de proteínas específicas de los testículos, que hasta ahora han pasado desapercibidas. Podemos especular que, al menos algunas de ellas, podrían tener funciones novedosas e importantes para el desarrollo testicular y/o la espermatogénesis.

Con el fin de confirmar nuestros hallazgos, y a la vez contribuir al conocimiento de la espermatogénesis y los posibles productos involucrados en el desarrollo del proceso, seleccionamos siete ejemplos de estas variantes de *splicing* no anotadas para confirmar su descubrimiento mediante RT-PCR. Los criterios que empleamos para su selección, fueron los siguientes:

a. Que correspondieran a genes codificantes anotados que tuvieran gran cantidad de variantes de *splicing* no anotadas, expresadas en nuestras listas;

b. Que al menos una de las variantes de *splicing* no anotadas codificara una isoforma proteica putativa;

c. Que la supuesta "nueva" isoforma proteica fuera significativamente diferente (por ejemplo, con diferentes dominios proteicos) de la/s proteína/s anotada/s para ese gen;

d. Que la isoforma putativa no anotada exhibiera un nivel de expresión relativamente alto en, al menos, una de las etapas espermatogénicas analizadas; y

e. Que el gen anotado poseyera una función descrita interesante (por ejemplo, relacionada con la espermatogénesis y/o función testicular), o, en su defecto, que presentara algún rasgo específico que nos resultara particularmente interesante.

Uno de los genes seleccionados para el análisis fue *Msh5* (*MutS Homologue 5*), un gen de expresión diferencial del testículo (ver Anexo, Figura S1A) que codifica una proteína específica de la profase meiótica esencial para la progresión de la meiosis, e involucrada en la reparación de roturas del ADN ("*mismatch repair*") y en la recombinación meiótica [169]. En el ratón, *Msh5* muestra una regulación positiva en la etapa de LZ, y da lugar a la síntesis de una proteína de 833 aminoácidos. Identificamos quince variantes de *splicing* sin anotaciones para este gen (Tabla S5) y seleccionamos una de ellas para su confirmación, la cual también muestra una regulación positiva en LZ y exhibe una expresión considerablemente mayor que la de la canónica (Figura 21A,a). La variante de transcripto seleccionada, que se genera a través de un sitio de inicio alternativo y una combinación de todos los mecanismos de *splicing* alternativo descritos anteriormente (es decir, SE, A5SS, A3SS, MXE, RI), codifica una proteína putativa más corta, de 362 residuos. Esta proteína contendría una región carboxilo-terminal idéntica a la de la proteína canónica, pero una región amino-terminal completamente diferente (Figura 21B).

Otro gen en el que profundizamos fue *BC051142*, un gen cuya función aún no está definida, pero que, según la base de datos del *NCBI*, es altamente específico de los testículos (Figura S1B). Su homólogo humano, la proteína *Testis Expressed Basic Protein 1* (TSBP1), se ha vinculado con el hipogonadismo (https://www.genecards.org/cgi-bin/carddisp.pl?gene=TSBP1&keywords=BC051142). En nuestras listas de datos, *BC051142* se destacó por tener uno de los mayores números de variantes de *splicing*.

A pesar de que en Ensembl figuran 13 isoformas en total, de las cuales 8 son clasificadas como codificantes, nuestro análisis reveló un total de 25 isoformas de ARN detectadas a lo largo de la espermatogénesis (cuando utilizamos parámetros menos restrictivos, este número aumentó a 103 variantes de *splicing*), con al menos nueve isoformas adicionales codificantes de proteínas no anotadas para este gen. Ninguna de estas isoformas se detectó en la población 2C, y la expresión de todas ellas comienza en LZ, aumentando a medida que progresa la espermatogénesis (Tabla S5), lo que indica que la proteína y todas sus variantes serían específicas de la espermatogénesis. Para confirmar la existencia de estas isoformas, seleccionamos dos variantes codificantes de proteínas putativas, no anotadas (Figura 21A,b).

Seleccionamos también *Agbl5* (*ATP/GTP Binding Protein Like 5*) un gen altamente y diferencialmente expresado en los testículos (Figura S1C), que codifica una metalocarboxipeptidasa involucrada en la deglutaminación de la tubulina, esencial para la formación de espermatozoides funcionales. Se ha demostrado que AGBL5 (también conocida como CCP5) es crucial para mantener la integridad de los flagelos espermáticos, y otras funciones vinculadas a los microtúbulos durante la espermatogénesis [170,171]. Varios estudios han informado sobre múltiples variantes de *splicing*, y al menos una de ellas parece tener propiedades distintas [170]. Nosotros hemos identificado numerosas variantes de *splicing* adicionales, no anotadas, que se expresan a diferentes niveles a lo largo de las etapas de la espermatogénesis (Tabla S5). Específicamente, seleccionamos para su confirmación una variante altamente expresada que alcanza su pico de expresión en RS, y codifica una proteína putativa de 412 aminoácidos. Esto diferiría de la isoforma canónica, que tiene su mayor nivel de expresión en PS, y produce un producto proteico de 846 residuos de longitud (Figura 21A,c).

Otro candidato, *Larp1* (*La-Related Protein 1*), codifica una proteína de unión a ARN que regula la traducción y la estabilidad de los ARNm de proteínas ribosomales y factores de traducción aguas abajo del complejo TORC1 [172,173]. Los niveles de expresión de *Larp1* son significativamente mayores en los testículos, en comparación con otros tejidos (Figura S1D), sugiriendo un papel potencialmente importante en la espermatogénesis y/o en la regulación de la función testicular. Seleccionamos una variante no anotada de *Larp1* que codifica una isoforma putativa no reportada de 760

aminoácidos, distinta de la proteína canónica, de 1.072 residuos. Esta nueva variante se expresa a niveles muy superiores a los de la canónica, y principalmente en PS, donde los niveles de expresión de esta nueva isoforma son 15 veces mayores que los de la variante canónica (Figura 21A,d).

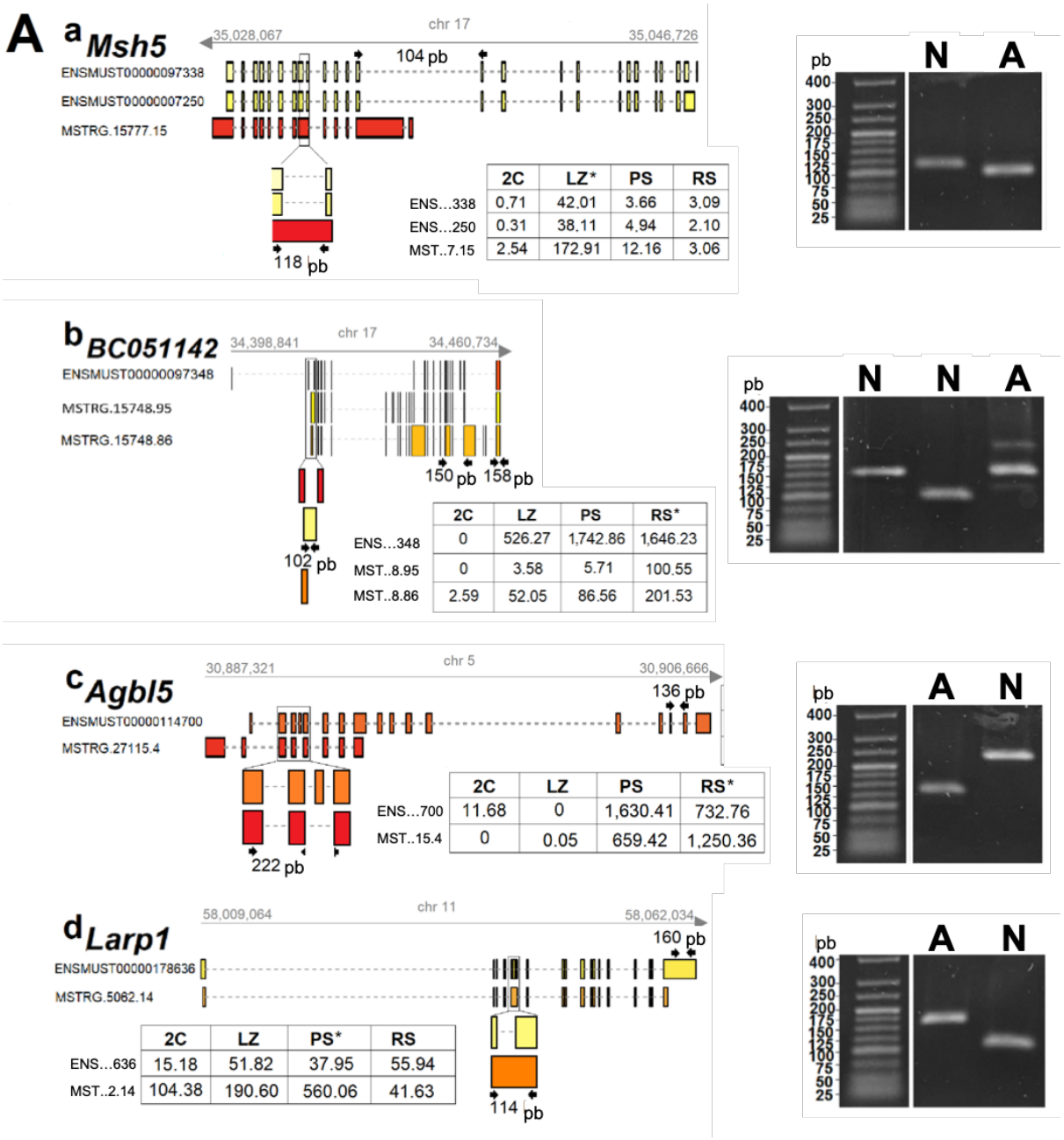
También estudiamos *Stk31* (*Serine-Threonine Kinase 31*), un gen con expresión específica del testículo (Figura S1E) que codifica una serín-treonín quinasa con un dominio Tudor [dominio estructural que generalmente se une a histonas metiladas, especialmente en argininas o lisinas, y participa en la regulación de la expresión génica y en procesos como la reparación del ADN [174,175]]. STK31 se localiza preferentemente en los gránulos germinales de los espermatocitos y en el acrosoma de las espermátidas, e interactúa con la proteína MIWI [176]. MIWI es una proteína clave en la regulación de la reprogramación epigenética durante el desarrollo de las células germinales masculinas, y juega un papel esencial en el silenciamiento de transposones y en la maduración de sncRNAs [177]. Se ha demostrado que STK31 es un "*cancer/testis antigen*" [familia de proteínas que están comúnmente presentes en células normales del testículo y en diversos tipos de células tumorales, pero rara vez en otros tejidos normales [178]] altamente expresado en varios tipos de cáncer, lo que sugiere su potencial como biomarcador y objetivo terapéutico en oncología [179–181]. Nosotros seleccionamos para su confirmación una variante de *splicing* más corta, pero más altamente expresada que la variante canónica a lo largo de las diferentes etapas de la espermatogénesis (Figura 21A,e). Esta nueva isoforma daría origen a una proteína putativa carente del dominio kinasa presente en la proteína canónica (Figura 21B), por lo cual seguramente no podría actuar como serín-treonín kinasa, lo que nos hace pensar que debería estar desempeñando una función diferente.

Otro gen estudiado fue *Baz1a* (*Bromodomain Adjacent to Zinc Finger Domain 1*a), un gen que exhibe una alta expresión en los testículos [182] (Figura S1F) y cuya expresión varía dinámicamente durante la espermatogénesis [183]. Este gen codifica una subunidad esencial de un complejo remodelador de cromatina dependiente de ATP, que se ha revelado como esencial para la expresión adecuada de genes espermatogénicos y para la fertilidad en ratones [182–184]. La investigación sobre *Baz1a* subraya su importancia en el mantenimiento de la integridad genética y funcional durante la producción de esperma, destacando su papel crucial en los mecanismos

epigenéticos que regulan la diferenciación celular y la fertilidad [183]. Para su confirmación, seleccionamos una variante de *splicing* expresada a niveles mucho más elevados que la canónica a lo largo de la espermatogénesis, y cuyos niveles son especialmente altos en PS (Figura 21A,f).

Por último, elegimos *Rsph1* (*Radial Spoke Head Component 1*). Este gen es de expresión específica de los testículos [185] (Figura S1G), y codifica un componente esencial de la cabeza de las espículas radiales presentes en el flagelo del espermatozoide [186]. Mutaciones en este gen resultan en discinesia ciliar primaria, lo que provoca problemas de motilidad en cilios y en el flagelo de los espermatozoides, y se han asociado con problemas de fertilidad humana [187]. Este gen se seleccionó como ejemplo de la generación de nuevas variantes mediante exclusión de exones, para lo cual elegimos una nueva variante presuntamente codificante, que excluye un exón presente en medio de la variante canónica (Figura 21A,g).

Logramos confirmar de forma exitosa la existencia de todas las variantes de empalme seleccionadas (Figura 21A,a-g), lo que representa un hito significativo en la validación de nuestros hallazgos, subrayando la consistencia y la precisión de nuestros datos en la identificación de isoformas espermatogénicas no anotadas.


A
a Msh5
chr 17
35,028,067
35,046,726
104 pb
ENSMUST00000097338
ENSMUST00000007250
MSTRG.15777.15
118 pb
2C LZ* PS RS
ENS...338 0.71 42.01 3.66 3.09
ENS...250 0.31 38.11 4.94 2.10
MST..7.15 2.54 172.91 12.16 3.06
pb
400
300
250
200
175
150
125
100
75
50
25
N A
b BC051142
chr 17
34,398,841
34,460,734
ENSMUST00000097348
MSTRG.15748.95
MSTRG.15748.86
150 pb
158 pb
102 pb
2C LZ PS RS*
ENS...348 0 526.27 1,742.86 1,646.23
MST..8.95 0 3.58 5.71 100.55
MST..8.86 2.59 52.05 86.56 201.53
pb
400
300
250
200
175
150
125
100
75
50
25
N N A
c Agbl5
chr 5
30,887,321
30,906,666
136 pb
ENSMUST00000114700
MSTRG.27115.4
222 pb
2C LZ PS RS*
ENS...700 11.68 0 1,630.41 732.76
MST..15.4 0 0.05 659.42 1,250.36
pb
400
300
250
200
175
150
125
100
75
50
25
A N
d Larp1
chr 11
58,009,064
58,062,034
160 pb
ENSMUST00000178636
MSTRG.5062.14
114 pb
2C LZ PS* RS
ENS...636 15.18 51.82 37.95 55.94
MST..2.14 104.38 190.60 560.06 41.63
pb
400
300
250
200
175
150
125
100
75
50
25
A N

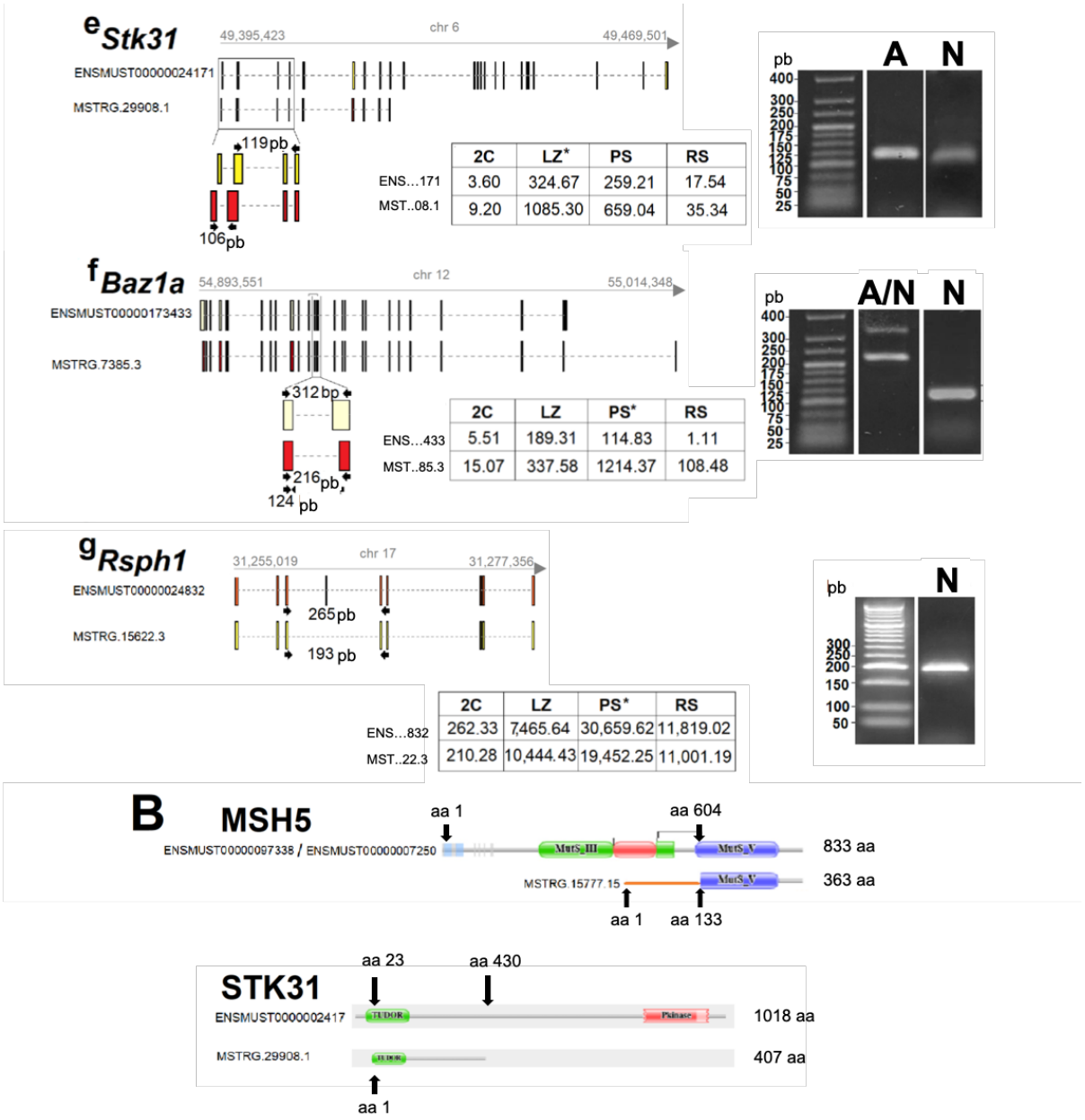


Figura 21. Confirmación de las diferentes variantes de *splicing* con alto potencial codificante seleccionadas. A) Representación esquemática de las variantes de *splicing* (anotadas y "nuevas"), junto con imágenes de geles de agarosa que muestran su amplificación por RT-PCR. Las anotaciones de Ensembl se presentan a la izquierda con el prefijo "ENSMUST" seguido del número correspondiente a ese transcripto, mientras que las isoformas no anotadas están etiquetadas como "MSTRG". Se muestran los pares de cebadores diseñados para la amplificación de las isoformas anotadas y no anotadas (indicados con flechas) junto con los tamaños esperados de los productos de PCR, en pares de bases (pb). La flecha gris sobre cada diagrama indica la dirección de transcripción, y se incluyen la ubicación genómica y el número de cromosoma. En cada caso se proporciona una tabla con la cobertura de los transcriptos anotados y no anotados en las cuatro

poblaciones celulares. Los asteriscos en la tabla indican la población celular en la que la isoforma no anotada tuvo el mayor nivel de expresión. En los geles de agarosa, **A** representa las variantes de empalme anotadas y **N** las no anotadas. *(a)* Variantes de *splicing* de ***Msh5*** que codifican la proteína canónica de 833 aminoácidos (amarillo) y una variante no anotada que codifica una isoforma de 362 residuos (rojo). *(b)* Variante más expresada de ***BC051142*** (rojo) y dos variantes no anotadas con diferentes patrones de expresión durante la espermatogénesis (una diferencial de la espermiogénesis, y otra, que incrementa progresivamente sus niveles de expresión desde la meiosis temprana hasta la espermiogénesis; amarillo y naranja, respectivamente). En el carril correspondiente a la variante anotada se observan bandas adicionales tenues, posiblemente debidas a isoformas débilmente expresadas (dado el alto número de isoformas detectadas para este gen, no fue posible diseñar cebadores que amplificaran exclusivamente una variante). *(c)* Transcripto canónico de ***Agbl5*** (naranja), que codifica una proteína de 846 residuos, y una variante no anotada seleccionada (rojo), que codifica una isoforma más corta, de 412 aminoácidos. *(d)* Variante no anotada de ***Larp1*** (naranja), que codifica una isoforma con mayores niveles de expresión y menor tamaño que la canónica (amarillo). *(e)* Variante no anotada seleccionada de ***Stk31*** (rojo), que codifica una isoforma algo más corta y con niveles de expresión bastante mayores, en comparación con la variante canónica (amarillo). La amplificación relativamente débil de la variante no anotada se debe a dificultades en el diseño de cebadores adecuados, ya que la corta región exclusiva de la variante más corta no permitía el diseño de un buen par de cebadores. *(f)* Representación de una isoforma anotada de ***Baz1a*** que codifica la proteína canónica (amarillo claro) y una variante no anotada (rojo) con niveles superiores de expresión. La amplificación fue realizada con un set de cebadores que simultáneamente amplifica ambas variantes (312 pb para el producto de amplificación de la anotada, y 216 pb para el de la no anotada). Además, a la derecha se presenta una banda correspondiente a un producto de amplificación realizado con un juego adicional de cebadores, que únicamente reconoce la isoforma no anotada (producto de 124 pb). *(g)* Variante de ARN anotada que codifica la proteína canónica ***Rsph1*** (rojo), y una nueva isoforma codificante generada por exclusión de exones (amarillo). Es de destacar que, aunque el par de iniciadores estaba diseñado para amplificar ambas variantes, no logramos amplificar la banda correspondiente a la variante anotada, posiblemente por competencia con la no anotada, de menor tamaño. **B)** Diagramas representativos de dos proteínas canónicas, y las proteínas putativas codificadas por las variantes no anotadas identificadas en este trabajo. Se toman estas dos a modo de ejemplo, para mostrar las diferencias entre ellas. **MSH5**: La línea naranja en la isoforma “nueva” representa los primeros 133 aminoácidos, que determinan una región amino-terminal completamente diferente de la de la proteína canónica. **STK31**: Ambas isoformas presentan

un dominio Tudor, pero la nueva variante predicha carece del dominio de serín/treonín quinasa, esencial en la isoforma canónica para su función.

4.2.2. Estudio de un caso particular: MSH5

De entre todas las variantes de *splicing* no anotadas confirmadas, codificantes para isoformas putativas de proteínas con importante función espermatogénica, el caso de *Msh5* nos interesó particularmente, por tratarse de un gen de altísima relevancia para la recombinación meiótica y el progreso de la meiosis, y cuya proteína canónica ha sido detectada en los núcleos meióticos en asociación con los sitios de recombinación (ver Anexo, Figura S2).

Con el objeto de caracterizar la nueva isoforma proteica cuyo ARNm confirmamos mediante RT-PCR, diseñamos un péptido contra la región exclusiva de esta variante proteica putativa de MSH5, el que fue empleado para la generación de anticuerpos específicos contra dicha variante.

En primer lugar, emprendimos la caracterización de esta nueva isoforma mediante Western blot empleando el anticuerpo generado. Al realizar dicho análisis, observamos una intrigante discrepancia: si bien esperábamos observar una banda de aproximadamente 38,8 kDa (recuérdese que la proteína predicha tendría 362 aminoácidos), la banda detectada presentó un peso molecular aparente mayor al esperado, situándose alrededor de los 60 kDa (Figura 22A). Es importante mencionar que la señal detectada fue sumamente específica y reproducible. Esta discrepancia entre el peso molecular aparente esperado y el observado, nos hace sospechar la existencia de modificaciones postraduccionales en esta proteína. De todos modos, este resultado estaría demostrando la existencia de la isoforma proteica en estudio, no sólo a nivel de transcripto, sino también de proteína. Alternativamente, no podemos descartar que estemos detectando una isoforma aún no caracterizada de MSH5, si bien esta posibilidad parece menos probable ya que debería tratarse de una isoforma con mayor expresión que la que hemos encontrado, pero que no habríamos identificado en nuestros análisis transcriptómicos.

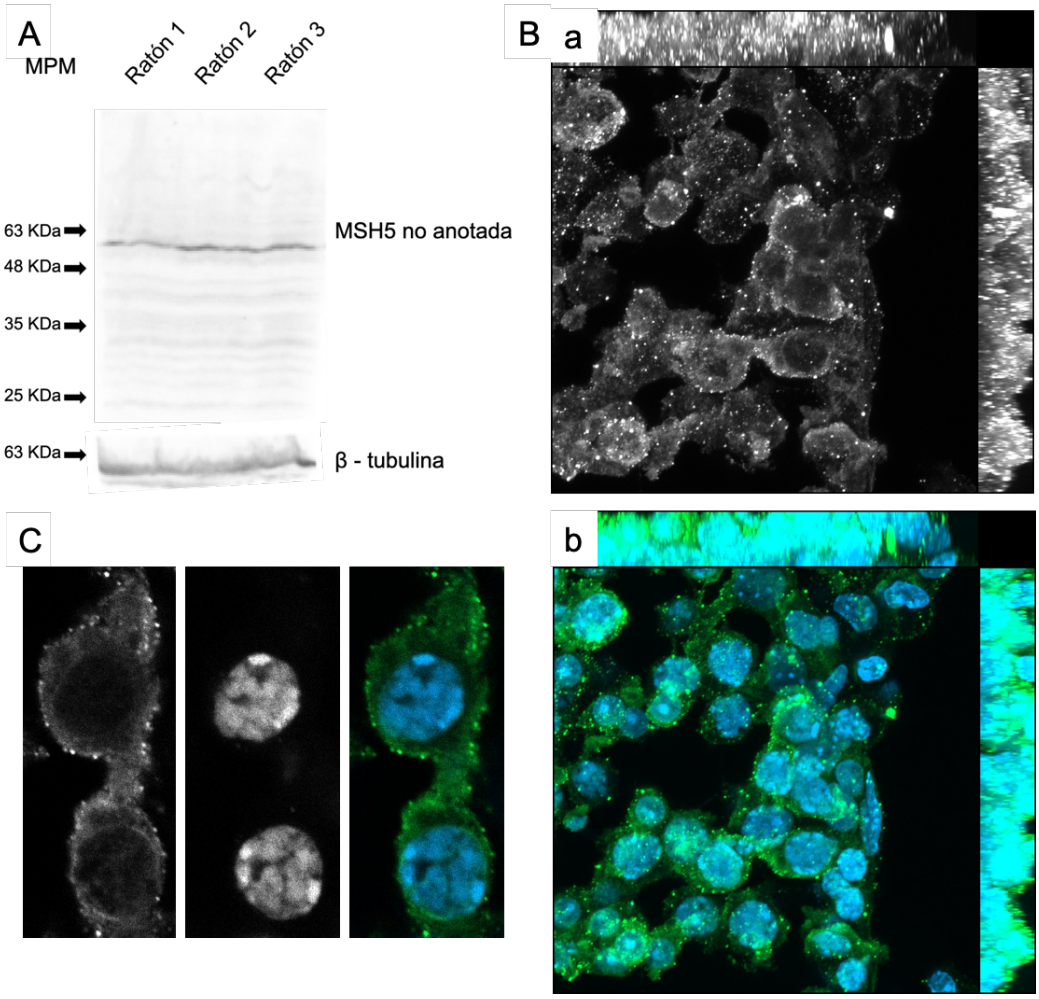


Figura 22. Localización de la isoforma no anotada de MSH5, mediante anticuerpo específico para esta variante. A) Western blot. Se aprecia una banda específica alrededor de los 60 kDa. **B)** Localización celular de la variante no anotada de MSH5 en criosecciones de testículo de ratón. *(a)* Vista panorámica de un montaje en Z-*stack* de criosección con marcaje para la variante no anotada de MSH5. En los bordes de la imagen se presentan proyecciones de los ejes X e Y, que muestran la distribución espacial de la señal a lo largo de la sección. (*b*) Misma imagen que en (*a*), pero mostrando la variante no anotada de MSH5 en verde, y los núcleos teñidos con DAPI (azul). **C)** Detalle a mayor aumento de dos células en planos unidos, con tres paneles que muestran la señal de la variante no anotada de MSH5 (izquierda), y la co-tinción con DAPI (derecha).

La variante proteica no anotada que identificamos, poseería una secuencia aminoacídica completamente diferente de la proteína MSH5 reportada en los primeros 133 aminoácidos, en tanto la región comprendida desde el aminoácido 134 hasta el final, sería idéntica en ambas isoformas. En consecuencia, esta nueva

proteína contendría un dominio MutS V parcial (ver Figura 21B), careciendo de parte del dominio ATPasa, que se ha reportado como crucial en la recombinación meiótica, facilitando la unión a estructuras de ADN y su interacción con la proteína meiótica MSH4 para regular los entrecruzamientos cromosómicos [188]. Adicionalmente, esta isoforma no reportada carecería del dominio MutS III, cuya función específica no se encuentra definida al día de hoy. Hipotetizamos, por lo tanto, que esta nueva variante debería estar desempeñando una función diferente de la de la proteína canónica, la que podría, o no, estar vinculada a la recombinación meiótica.

Con el fin de contribuir al conocimiento de esta nueva isoforma, empleamos el anticuerpo generado en ensayos de inmunofluorescencia. Los mismos revelaron una señal específica, que parece estar distribuida en el citoplasma y principalmente sobre el borde celular (Figura 22B), en tanto la señal no parece localizarse en el núcleo. La marca parece ser más intensa en los espermatocitos en profase temprana, coincidiendo con las etapas de mayor expresión a nivel del ARNm (ver Figura 21A,a). Si bien no hemos realizado cuantificaciones de los niveles de expresión de la proteína en cada uno de los tipos celulares, ni a nivel subcelular, que nos permitan delimitar con exactitud su localización y distribución, nos pareció que ello era irrelevante, dado que resulta evidente que el patrón de expresión de esta isoforma proteica es completamente diferente del de la proteína MSH5 canónica, que es lo que nos interesaba averiguar. Esta distinta localización y distribución sugieren una función de la variante de MSH5 muy diferente de la de la canónica; es importante señalar que, si participara en la reparación de rupturas de doble hebra y en la recombinación, sería esperable que se encontrara localizada dentro del núcleo, asociada a los complejos sinaptonémicos.

4.2.3. Confirmación masiva de “nuevas” proteínas generadas a partir de variantes de *splicing* y genes no anotados, mediante proteómica

Como hemos mencionado previamente, dentro de la sección 4.1.2., los transcriptos sin anotación identificados fueron 33.002. De éstos, unos 14.667 resultaron ser variantes de *splicing* de genes ya anotados, mientras que 18.335 fueron transcriptos

de genes hasta el momento no anotados. El análisis de potencial codificante seleccionó 2.794 transcriptos sin anotar con alto potencial codificante (ver Figura 11).

Si bien hemos profundizado en el estudio de algunos casos particulares de variantes de *splicing* potencialmente codificantes en las secciones previas, nos propusimos extender la confirmación por un abordaje complementario, a la vez que hacer una confirmación de la existencia de, al menos, algunas de estas variantes de *splicing* hasta el momento no anotadas, de alto potencial codificante, mediante un método más masivo. Concretamente, elegimos efectuar la confirmación de la existencia de al menos algunas de las proteínas, lo cual nos lleva a un paso más adelante, confirmando a su vez la existencia del producto proteico de los transcriptos identificados.

Con ese propósito, nos apoyamos en una investigación reciente, publicada durante el desarrollo de nuestro trabajo, en la que se llevó a cabo un ensamblaje *de novo* de transcriptos de ratón y se verificaron los transcriptos codificantes mediante el análisis de datos de proteómica [189]. Por un lado, realizamos ensamblaje de transcriptos de ratón, creando una base de datos de referencia que contenía todas las proteínas putativas generadas a partir de los transcriptos con alto potencial codificante que identificamos, incluyendo las posibles 2.794 proteínas no anotadas. Para llevar a cabo la confirmación de la existencia de las proteínas producto de los transcriptos que clasificamos como codificantes, empleamos datos crudos de proteómica generados mediante *shotgun*, del proyecto PXD030983, correspondientes a muestras de testículo completo de ratón [166]. Estos datos crudos fueron descargados y mapeados contra la referencia que creamos, previamente mencionada.

Como resultado del mapeo, pudimos observar un total de 16.825 proteínas expresadas. Al realizar la intersección con los 2.794 transcriptos “nuevos” que habíamos definido como probablemente codificantes, logramos confirmar la existencia de péptidos provenientes de 1.949 proteínas (Figura 23A), lo que equivale a la validación de un 70% del total. Esto da un respaldo importante a nuestros resultados.

De las 1.949 proteínas no anotadas validadas, 1.913 correspondieron a isoformas no anotadas, provenientes de 951 genes conocidos, en tanto 36 provinieron de 22 de los 191 genes no anotados hasta el momento, con alto potencial codificante (Figura 23B). Cabe destacar que, de estas 36 proteínas, para 18 obtuvimos información en la anotación primaria (por ejemplo, por similitud parcial con otras ya anotadas, de ratón o de otras especies), mientras que, para las otras 18, no obtuvimos ninguna información durante el proceso de anotación primaria (Tabla S6).

Con el objeto de profundizar en nuestras identificaciones, realizamos un análisis BLASTP de forma individual a partir de las secuencias aminoacídicas predichas de las 36 proteínas para las que fuimos capaces de encontrar péptidos provenientes de ellas (Tabla S6). Dado que este análisis individual fue realizado con varios meses de diferencia con la anotación primaria, nos permitió determinar si, en el plazo de tiempo transcurrido, habrían surgido anotaciones nuevas para las proteínas mencionadas. Observamos que las proteínas para las cuales habíamos obtenido una anotación primaria coincidían con los resultados del BLASTP, lo que reforzó la validez de nuestras anotaciones iniciales. Una excepción notable fue un gen cuya anotación en las bases de datos fue actualizada después de nuestra anotación primaria. Este caso correspondía a las proteínas MSTRG.38471.6 y MSTRG.38471.8, que originalmente habían sido identificadas como producto de un gen expresado en cloroplasto (https://www.uniprot.org/uniprotkb/P12222/entry), y fueron recientemente identificadas como proteína ácida nuclear de células germinales del ratón (https://www.ncbi.nlm.nih.gov/protein/NP_001369163.1?report=genbank&log$=prottop&blast_rank=1&RID=71FFSHCF016), concordando con nuestro hallazgo.

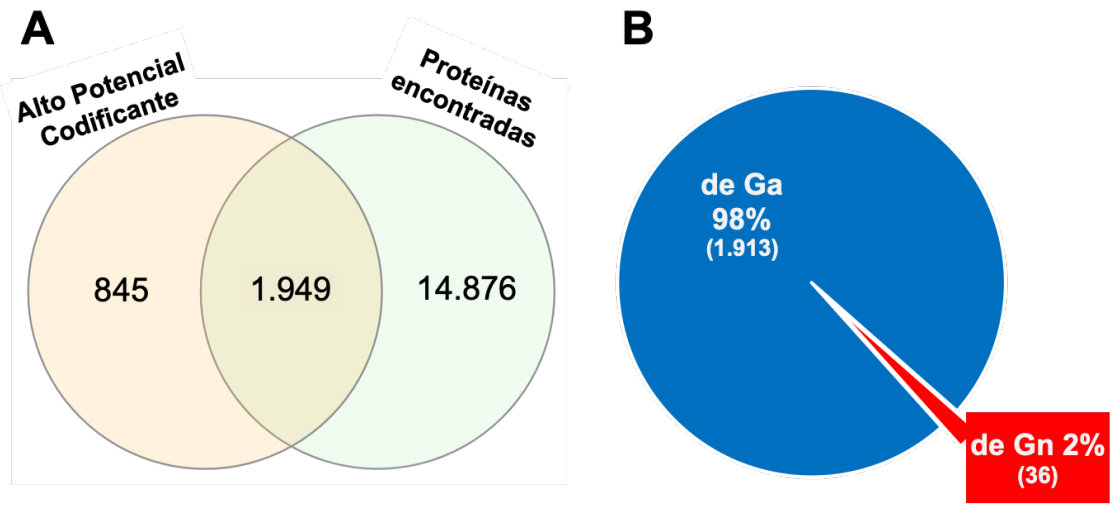


Figura 23. Validación de proteínas provenientes de transcriptos putativamente codificantes identificados en nuestro trabajo. A) Diagrama de Venn que representa la intersección entre proteínas identificadas mediante proteómica *shotgun* a partir del mapeo de los datos crudos de Giansanti *et al*., 2022 (16.825), y los transcriptos no anotados identificados en nuestro trabajo y previamente clasificados como con alto potencial codificante (2.794). De éstos, logramos identificar péptidos específicos correspondientes a 1.949 proteínas provenientes de los transcriptos de alto potencial codificante, mientras que 845 de estos transcriptos no pudieron ser confirmados a nivel de proteína. **B)** Gráfico circular que muestra la distribución de las 1.949 proteínas confirmadas a partir de los datos de proteómica, en función de su estatus de anotación. De ellas, el 98% (1.913) corresponde a nuevas proteínas generadas a partir de variantes de *splicing* codificantes de genes ya anotados, mientras que un 2% (36 proteínas) pertenecen a transcriptos provenientes de 22 genes completamente "nuevos", los cuales no habían sido previamente reportados en bases de datos existentes.

4.2.4. Proteínas putativas sin anotación primaria

Para las 18 proteínas para las que no se obtuvo ninguna anotación primaria, empleamos el mismo mecanismo de búsqueda individual que el utilizado para las que sí poseían anotación primaria, mediante BLASTP. De estas proteínas putativas analizadas, este estudio nos proporcionó coincidencias para las 18 proteínas nuevas. Dos de estas proteínas, MSTRG.38471.9 y MSTRG.38471.10, se asemejaron a la

suppressor protein SRP40-like de *Mus cairoli*, con una identidad superior al 88% y 86% respectivamente.

Por otro lado, los transcriptos correspondientes a las restantes 16 probables proteínas "matchearon" todos ellos contra una misma región del genoma, localizada sobre el cromosoma X. Se trata de una región descrita como de alta complejidad en lo que respecta a su secuencia nucleotídica, con zonas de alto contenido en GC, y, a su vez, con muchos repetidos, y que presenta alto interés precisamente por su complejidad y particularidades. En la versión GRCm38 del genoma del ratón, que fue la que empleamos como genoma de referencia durante este trabajo, en esa región del genoma, de aproximadamente de 2,5 Mb, había anotados únicamente tres genes, *Srsx,* codificante para una proteína de secreción rica en serinas, *Astx2,* presuntamente codificante para una proteína de la cual aún se desconoce la función, y *Gm17142*, que produciría un lncRNA (Figura 24, línea superior).

Llamativamente, observamos que, recientemente, en la última versión del genoma del ratón (GRCm39), hubo un cambio en la anotación de esta zona del cromosoma X, en la cual se aprecia un reordenamiento de la secuencia, reduciéndose el tamaño de la región a 260 Kb. Dicho cambio se basa en un trabajo publicado [190], en el cual se propone la existencia, en esa compleja región, de un gen atípico, al que se denominó *LaidX* (*Large amplified intrinsically disordered protein gene on the X chromosome*). Según el mencionado trabajo, el gen propuesto sería transcripto en un enorme ARNm de expresión específica de la espermatogénesis, con una longitud de unas 25 Kb, y que consistiría en un único exón con potencial codificante, que se traduciría en una única proteína predicha de más de 8.000 aa (Figura 24, línea inferior). Como su nombre lo indica, esta enorme proteína teórica no presentaría dominios caracterizables, sino que tendría una estructura intrínsecamente desordenada, y una función por ahora no definida.

Al ver este cambio en la anotación, procedimos a mapear nuestros datos sobre el nuevo ensamblado del genoma de referencia (GRCm39), con el fin de confirmar que nuestras lecturas mapearan contra esta región. En efecto, pudimos apreciar que esta zona, de poco más de 25 Kb, se encuentra cubierta por completo de lecturas, lo cual reafirma que se transcribiría por completo, de forma continua, generando un gran

transcripto único sin intrones en la región codificante (Figura 24, líneas media e inferior).

Por otro lado, aunque cuando se describió el gen *LaidX*, se lo predijo como potencialmente codificante [190], en el trabajo mencionado no se tenía evidencia de su producto proteico, por lo cual, la existencia de una proteína de más de 8.000 aa generada a partir de un único exón, podría resultar dudosa. En ese sentido, nosotros fuimos capaces de identificar péptidos provenientes de diferentes regiones de la presunta proteína (los 16 péptidos identificados corresponderían a la proteína LAIDX), siendo ésta la primera evidencia empírica de la existencia de una proteína producto de este gen.

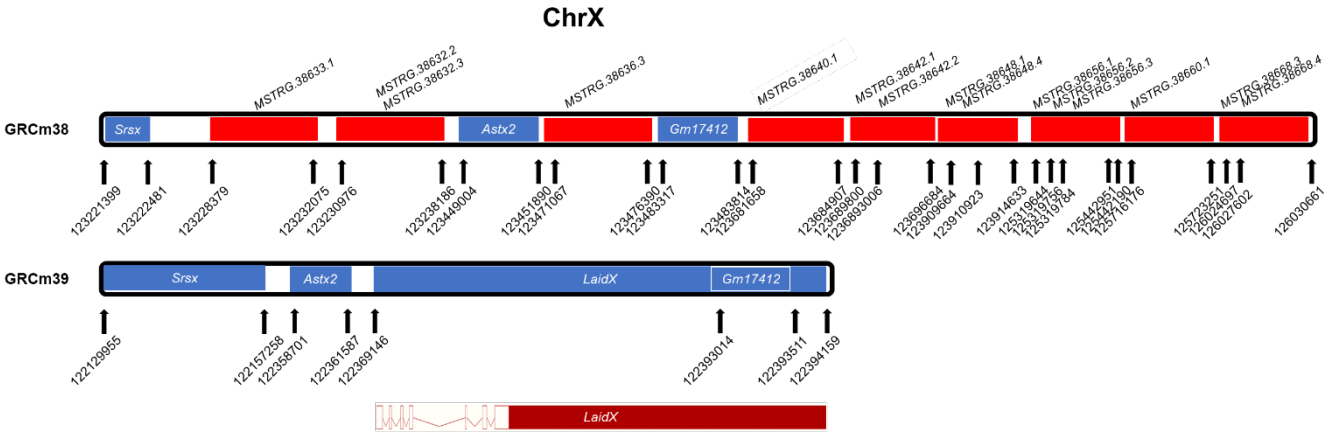


Figura 24. Representación esquemática de un cambio estructural en la anotación de una región del cromosoma X del ratón. Se puede ver como cambian la posición de los genes *Srsx* (cuya longitud cambia considerablemente), *Astx2* y *Gm17412*, y se anota el gen *LaidX*, del cual podemos ver la anotación presente en ENSEMBL en la parte inferior de la figura. A su vez, podemos apreciar la posición de nuestros genes no anotados en el cromosoma X en la versión GRCm38 del genoma (línea superior), y su posición relativa a los genes anotados en la última versión.

5. DISCUSIÓN

Se ha documentado ampliamente la complejidad del transcriptoma testicular, con informes que resaltan su heterogeneidad [58,136] y contribución significativa del *splicing* alternativo a dicha complejidad [80,121,130,133,146]. Además, es sabido que la adecuada ejecución del *splicing* alternativo específico de cada una de las etapas, es crucial para el éxito de la espermatogénesis [80,121,130,133,137–139,191]. Sin embargo, debido a la gran heterogeneidad celular en los testículos, es probable que cuando los estudios transcriptómicos se realizan a partir de testículo entero, o de poblaciones celulares pobremente enriquecidas en cierto tipo celular, muchas isoformas de ARN específicas de ese tipo celular caigan por debajo de los límites de detección. Por otro lado, aunque el scRNA-seq ha contribuido a mejorar la comprensión de la espermatogénesis [91], es importante tener en cuenta que, en general, las bibliotecas de scRNA-seq tienen una profundidad menor que las obtenidas de *pools* de células, limitando la detección y siendo un mal abordaje para el ensamblado de transcriptos. Además, para el caso concreto de la espermatogénesis, los estudios de scRNA-seq presentan la limitación adicional de que, dado el enorme desfasaje existente entre transcripción y traducción (muchos ARNm pueden almacenarse hasta semanas previo a su traducción; ver sección 1.1.5), por lo general los análisis de scRNA-seq no resultan informativos acerca de cuál es el tipo celular que está produciendo determinados transcriptos, ya que no realizan una clasificación previa de las células, sino una clasificación a partir de los datos de secuenciación [98].

En nuestro estudio, al utilizar poblaciones de células espermatogénicas específicas de estadio altamente puras, junto con bibliotecas de secuenciación de buena profundidad, hemos logrado identificar una gran cantidad de genes y transcriptos generados mediante *splicing* alternativo que no habían sido anotados previamente. En consecuencia, si bien el hecho de la gran diversidad transcriptómica del testículo era conocido, este trabajo demuestra que la misma es mayor de lo que se había informado anteriormente, y que aún queda mucho por reportar.

5.1. Los análisis de RNAseq revelan numerosos transcriptos no anotados en la profase meiótica temprana.

La población de células en etapas de LZ (es decir, profase meiótica temprana) se destacó por presentar la mayoría de las variantes de *splicing* no anotadas hasta el momento. Este fenómeno podría atribuirse, en gran parte, al hecho de que, como hemos mencionado, los espermatocitos en profase meiótica temprana rara vez se han incluido en estudios transcriptómicos tendientes al estudio de la expresión génica a lo largo de la espermatogénesis. Ha conspirado para ello la dificultad para obtener estas células como poblaciones celulares aisladas [89,148] o incluso enriquecidas, dada la breve duración del L y Z, lo que hace que su representación en la totalidad del testículo sea muy baja [89,148]. A su escaso número, se suma la dificultad de que estas células no presentan características particulares que favorezcan su enriquecimiento mediante estrategias de sedimentación celular, a diferencia, por ejemplo, de los espermatocitos paquiténicos, que son más fáciles de aislar por sedimentación o centrifugación en gradientes de albúmina, dado su volumen notablemente mayor que el de otros tipos celulares [104]. Como resultado de lo antedicho, los estudios de transcriptómica en poblaciones aisladas, en general, se han basado en el enriquecimiento en tipos celulares pertenecientes a otros estadios espermatogénicos más accesibles y con mayor representación en la totalidad de células, como la profase meiótica media/tardía y la espermiogénesis.

Otro factor importante que seguramente haya dificultado la detección previa de muchos transcriptos de LZ se relaciona con nuestros hallazgos, que indican niveles generales de expresión menores en esta etapa, en comparación con otras poblaciones celulares testiculares. Esto estaría en concordancia con antiguos estudios preliminares de las décadas del ´60 y ´70 que, a partir de la incorporación de precursores tritiados, habían sugerido la existencia de bajos niveles generales de transcripción en la profase meiótica temprana [192–194]. Naturalmente, la escasez de espermatocitos en LZ, sumada a los niveles comparativamente bajos de expresión de muchos de sus transcriptos, seguramente ha hecho que buena parte de los transcriptos específicos de la profase meiótica temprana hayan quedado diluidos

entre los de tipos celulares más abundantes, en los transcriptomas de testículo completo.

Todo esto contribuye a explicar por qué nosotros, empleando poblaciones celulares de altísima pureza obtenidas por citometría de flujo, hemos detectado un número tan importante de variantes de *splicing* no anotadas, expresadas principalmente en LZ. Más aún, de los 191 genes presuntamente codificantes que no estaban anotados en el genoma del ratón, hemos identificado la expresión de 159 de ellos en LZ, y casi la mitad son exclusivos de esta etapa. Esta observación sugiere que muchos de estos genes pueden haber pasado desapercibidos hasta ahora precisamente debido a su codificación de productos específicos de la profase meiótica temprana. Suena razonable pensar que un fenómeno similar podría ocurrir con tipos celulares escasos en otros tejidos heterogéneos, donde es probable que un gran número de transcriptos específicos de tipos celulares poco representados permanezca aún sin detectar.

Más allá de la abundancia de transcriptos no anotados en la etapa de LZ, esta fase también exhibió la mayor cantidad de transcriptos expresados en general, tanto anotados como no anotados. Nuestros resultados coinciden con un estudio de scRNA-seq que sugiere que las primeras etapas de la espermatogénesis expresan una mayor diversidad de genes, mientras que las etapas posteriores concentran una gran proporción de sus transcriptos en un conjunto más limitado de genes [195].

Proponemos que esta amplia expresión de genes e isoformas en LZ podría ser necesaria para el correcto desarrollo de los eventos tan únicos y exclusivos que tienen lugar durante la profase meiótica temprana. Es importante destacar que la base molecular de estos eventos sigue siendo en gran medida desconocida: aún no comprendemos completamente el papel de la formación del *bouquet* en el alineamiento cromosómico, ni cómo se reconocen los cromosomas homólogos entre sí. En este contexto, la identificación de todos estos genes y variantes de *splicing* sin anotar, tanto codificantes como no codificantes, podría representar un paso significativo hacia el avance en la comprensión de estos procesos esenciales y de su regulación, proveyendo una fuente enorme de material para futuros estudios.

5.2. Existe una gran cantidad de lncRNAs espermatogénicos aún no anotados.

El análisis del potencial codificante de los transcriptos no anotados reveló que la mayoría de ellos son no codificantes, como se muestra en la Figura 12. Este hallazgo es coherente con la noción de que la investigación sobre los lncRNAs es relativamente más nueva en comparación con la de los genes codificantes, lo que sugiere que apenas hemos comenzado a explorar la diversidad de lncRNAs existente, y que un vasto número de lncRNAs probablemente aún permanece sin descubrir.

Además, en un estudio previo donde intentamos realizar un análisis de conservación entre los lncRNAs espermatogénicos de ratón y humano, observamos que, para varios lncRNAs de una especie, había secuencias de ADN homólogas en la otra, pero no se había anotado un lncRNA correspondiente [148]. Si bien esto podría reflejar diferencias entre especies, lo que es consistente con la menor conservación de los patrones de expresión de los lncRNAs en comparación con la de los genes codificantes [64], también es indicativo, al menos en parte, de la parcialidad de la anotación de los lncRNAs.

El amplio número de lncRNAs espermatogénicos no anotados que hemos identificado, se suma a la considerable cantidad de lncRNAs ya anotados en células germinales masculinas, en comparación con cualquier otro tejido y tipo celular estudiado [50,51,58,64,112,113]. Esta abundancia puede atribuirse, en parte, a la cromatina más relajada de las células meióticas y posmeióticas, pero también, en parte, a los elevados niveles de regulación postranscripcional presentes en estas células (ver siguiente sección).

Un rasgo distintivo que hemos observado en los genes no codificantes, es que tienden a tener un menor número de transcriptos por gen en comparación con los codificantes de proteínas, lo que es indicativo de que suelen presentar menor cantidad de variantes de *splicing*. Este hallazgo estaría en línea con informes previos que indican que el *splicing* de los lncRNAs es menos eficiente que el de los ARNm [51,196]. Además, está en consonancia con nuestros resultados anteriores y los de otros grupos, que

muestran que los lncRNAs tienden a ser más cortos y a tener menos exones que los ARNm [50,51,148]. En suma, el nuevo hallago de que también presentan menor número de variantes de *splicing*, contribuye a la idea de que los lncRNAs son, en general, menos complejos que los ARNm.

5.3. La abundancia de transcriptos y variantes de *splicing* no anotados subraya la alta complejidad transcriptómica de las células meióticas y posmeióticas.

Las poblaciones de células meióticas y posmeióticas exhibieron un mayor número de transcriptos sin anotar, en comparación con la población de células 2C [58]. Este número comparativamente menor en la población 2C podría ser un reflejo de la complejidad intrínseca de los transcriptomas espermatogénicos, ya que, como hemos mencionado repetidamente, es sabido que las células meióticas y posmeióticas tienen transcriptomas extremadamente complejos [58].

Se ha sugerido que la complejidad de las células meióticas y posmeióticas podría ser una consecuencia del estado permisivo de su cromatina, resultado de la remodelación extensiva de la misma que ocurre durante estas etapas, debida al reemplazo de las histonas [58]. El reemplazo de histonas durante la espermatogénesis es un proceso en el que éstas son progresivamente sustituidas por protaminas, facilitando una mayor compactación del ADN en los espermatozoides maduros. Durante las primeras etapas, la cromatina se vuelve más laxa debido al intercambio temporal de histonas canónicas por variantes especializadas, permitiendo la remodelación epigenética y la transcripción de genes esenciales para la diferenciación celular [31]. En este contexto, algunos investigadores han sugerido que, al menos parte de esa enorme cantidad de transcriptos producidos en estas etapas, se debería a transcripción “promiscua” [58]. De ser así, podemos especular que parte del gran número de transcriptos no anotados identificados en estas células, representen transcripción promiscua. En relación con esto, un estudio reciente (que salió publicado en paralelo con el nuestro)

también identificó una abundancia de transcriptos no anotados en células germinales masculinas de ratón [189].

Más allá de si parte de la extensiva transcripción que tiene lugar durante la espermatogénesis es promiscua, también es cierto que la compleja diversidad transcriptómica observada en las células meióticas y, en particular, en las posmeióticas, se considera, en buena medida, parte de una estrategia adaptativa para regular la síntesis de proteínas en las espermátidas elongadas que, como hemos mencionado, son transcripcionalmente inertes (ver sección 1.1.5). Esta necesidad de disponer de todos los transcriptos para ser traducidos en momentos específicos de forma muy estrictamente regulada, ha impulsado el desarrollo de sofisticados mecanismos de regulación postranscripcional, destinados a cumplir con los estrictos requisitos de regulación, algunos de los cuales son exclusivos de los espermatocitos y espermátidas redondas [136,141,195,197]. Estos mecanismos de regulación postranscripcional probablemente requieran una abundancia de ARNs reguladores. De hecho, aunque se reconoce que una proporción significativa de los lncRNAs espermatogénicos puede carecer de función biológica, estudios recientes han destacado la importancia de varios de ellos en la regulación de la espermatogénesis y la preservación de la fertilidad [98,198–208].

Con respecto al estudio de los tipos de *splicing* predominantes, el análisis de nuestros datos de RNAseq reveló que el tipo más prevalente fue el salto de exón (SE), seguido de la retención de intrones (RI), no registrándose diferencias significativas, en este sentido, entre las cuatro poblaciones celulares estudiadas. El hallazgo de que los tipos de *splicing* más utilizados son SE y RI, concuerda con los resultados presentados en un estudio de reanálisis de datos de un repositorio público (cabe destacar que dicho análisis no incluyó la profase meiótica temprana) [134]. Asimismo, nuestros resultados están en línea con los de otro estudio [141], en el que se observó que la RI es uno de los patrones de *splicing* alternativo más comunes durante la diferenciación transmeiótica de las células germinales masculinas. Es interesante resaltar que en este último estudio se encontró una regulación positiva de los eventos de RI en los espermatocitos en comparación con las espermátidas, sugiriendo que la RI podría servir como un mecanismo de retención nuclear de transcriptos durante la meiosis, para su posterior traducción en las células germinales posmeióticas

transcripcionalmente inactivas [141]. Aunque nosotros no encontramos diferencias significativas en cuanto a la magnitud de la RI entre las cuatro poblaciones celulares estudiadas, es importante tener en cuenta que los resultados de ese trabajo y los nuestros no son directamente comparables, ya que nuestro análisis se centró en la prevalencia de las diversas categorías de *splicing* alternativo en diferentes etapas de las células espermatogénicas, y no en los eventos de *splicing* regulados diferencialmente.

En suma, la abundancia de transcriptos expresados resalta la diversidad de los eventos de expresión génica que tienen lugar durante la espermatogénesis, evidenciando la enorme complejidad de los procesos celulares involucrados en la formación de los espermatozoides. La presencia de un gran número de transcriptos no anotados y variantes de *splicing* aún no caracterizadas, sugiere que la comprensión completa de la regulación génica en estas etapas cruciales de la espermatogénesis, aún está muy lejos.

5.4. La caracterización de los patrones de *splicing* alternativo revela la existencia de variantes de interés, previamente desconocidas.

Un resultado importante de nuestro trabajo, es la identificación de numerosas variantes de *splicing* no anotadas de genes con funciones altamente relevantes en la espermatogénesis. Estas variantes presentaron elevados niveles de expresión, incluso, en ocasiones, notablemente superiores a los niveles de expresión de las variantes anotadas. En muchos casos, es probable que estas variantes hayan pasado desapercibidas debido a su expresión específica en etapas que a menudo están subrepresentadas en los estudios transcriptómicos, como es el caso de LZ. Lo más interesante es que algunos de estos transcriptos recién identificados, con un alto potencial codificante, podrían estar dando lugar a isoformas proteicas específicas de testículo, hasta ahora no conocidas. Podemos hipotetizar que, al menos algunas de estas nuevas isoformas, desempeñarían funciones distintas en la espermatogénesis, lo que es fuertemente apoyado por el hecho de que varias de ellas carecerían de

dominios clave para la ejecución de la función reportada para su contraparte canónica, o presentarían algunos dominios diferentes.

Un punto crucial, que avala la relevancia de nuestros hallazgos, fue nuestra capacidad para detectar estas interesantes variantes de *splicing* mediante abordajes alternativos al RNAseq, como RT-PCR. Deliberadamente, todas las variantes que seleccionamos fueron cuidadosamente elegidas, representando ejemplos de posibles isoformas proteicas no previamente descritas, cuya expresión varía a lo largo de las distintas etapas de la espermatogénesis, y cuyas proteínas canónicas, en la mayoría de los casos, son conocidas por desempeñar funciones esenciales en este proceso.

Un ejemplo ilustrativo, es la isoforma no anotada que identificamos para el producto del gen *Msh5*. MSH5 es una proteína reparadora de daño del ADN específica de la meiosis, crucial para la recombinación homóloga [209] y la progresión meiótica [169]. Dado que la nueva isoforma, altamente expresada en LZ, presentaría un dominio ATPasa incompleto, necesario para la reparación de roturas de doble cadena [188], hipotetizamos *a priori* que esta variante no anotada podría desempeñar una función diferente de la de la canónica durante la profase meiótica. En apoyo a esta idea, los resultados de inmunofluorescencia obtenidos con el anticuerpo que generamos, revelaron una distribución de la señal diferente de la reportada para la proteína canónica, no localizada en relación con los complejos sinaptonémicos y los puntos de recombinación, sino con una ubicación predominante en el citoplasma. Curiosamente, aunque en los ensayos de Western blot detectamos una banda específica para esta isoforma (que presenta un peso molecular distinto del de la proteína canónica), su peso molecular aparente no coincidió con el peso molecular esperado, sino que fue mayor. Sin embargo, este hallazgo coincide con los resultados obtenidos con un anticuerpo comercial anti-MSH5 dirigido contra una región carboxilo-terminal de la proteína canónica, región que comparten tanto la canónica como la isoforma no anotada (https://www.sigmaaldrich.com/UY/es/product/sigma/hpa062688). Ese resultado, validado por el *Protein Atlas* (https://www.proteinatlas.org/ENSG00000204410-MSH5), muestra una distribución celular similar a la observada en nuestros experimentos de inmunofluorescencia, y una banda en los ensayos de Western blot del mismo tamaño que la obtenida en nuestros estudios. El mayor peso molecular aparente de esta isoforma proteica nos

ha llevado a proponer que, posiblemente, la misma sea blanco de modificaciones postraduccionales. En esa misma dirección, no fuimos capaces de detectar la nueva isoforma de MSH5 mediante el análisis de proteómica, lo que nos refuerza la idea de que la misma podría ser blanco de modificaciones postraduccionales que alteraran su peso molecular. Las modificaciones postraduccionales hacen más complejo el análisis proteómico, ya que, en un principio, este análisis se basa en la comparación de masas de péptidos [210]. A su vez, los resultados de inmunofluorescencia concuerdan con lo reportado por Lahaye y colaboradores, quienes observaron la presencia de proteína MSH5 fuera del núcleo [211]. Un interesante desafío para el futuro es una mayor exploración de la función de esta nueva isoforma proteica en relación a la profase meiótica.

Otro ejemplo altamente llamativo es el del gen *BC051142*, un gen que, a pesar de ser uno de los que presentó mayor cantidad de variantes de *splicing*, codificar numerosas isoformas proteicas putativas específicas de las células germinales masculinas en nuestras listas, y haber sido asociado con el hipogonadismo en humanos, su función aún es completamente desconocida (https://www.genecards.org/cgi-bin/carddisp.pl?gene=TSBP1&keywords=BC051142). Este caso ilustra la considerable variabilidad existente a lo largo de la espermatogénesis, y la vasta cantidad de incógnitas que aún persisten en este campo de estudio.

El éxito en la validación de las nuevas isoformas para los siete genes seleccionados respalda la robustez de nuestro enfoque metodológico, y sugiere que nuestras observaciones tienen una base sólida y confiable. Para todos estos genes previamente mencionados, las probables isoformas proteicas recientemente identificadas diferirían significativamente de las canónicas. MSH5 es un ejemplo claro. STK31 es otro buen ejemplo de esto: mientras que la proteína conocida tiene un dominio Tudor y un dominio proteín-quinasa esencial para su función, la variante predicha carecería de este último, lo que sugiere que debería desempeñar un papel diferente. Estas variantes podrían, eventualmente, tener implicancias importantes en la regulación de la función celular, y podrían estar involucradas en distintas rutas biológicas, aún por descubrir.

Adicionalmente a la validación por RT-PCR, recurrimos al análisis proteómico como método de validación masiva. Dicho análisis confirmó la existencia de 1.949 proteínas derivadas de transcriptos previamente no anotados identificadas en nuestro estudio. Este hallazgo, a la vez que da un fuerte soporte a nuestros datos, subraya la notable complejidad del proteoma testicular, permitiendo ampliar nuestra comprensión sobre la diversidad proteica en este tejido. Por otra parte, esta lista confirmada de nuevas isoformas proteicas constituye una valiosa base de datos para futuros estudios. Como hemos señalado, las proteínas generadas por estas variantes de *splicing* podrían desempeñar funciones biológicas novedosas y, por ende, su caracterización puede abrir nuevas vías para entender mejor la biología del testículo y de los mecanismos moleculares subyacentes a la espermatogénesis. Las variantes de *splicing*, como mencionamos previamente, han sido reconocidas como mecanismos clave para generar diversidad funcional en las proteínas, y su relevancia en el testículo es particularmente significativa, dado que este órgano requiere una regulación precisa y multifacética. Además, la existencia de estas variantes no sólo sugiere posibles nuevas funciones, sino también la adaptación evolutiva del tejido testicular para optimizar procesos complejos, lo que resaltaría la importancia de este tipo de análisis para descubrir aspectos no explorados de la biología reproductiva.

5.5. Consideraciones sobre la anotación y caracterización de transcriptos codificantes hasta ahora no anotados.

La anotación primaria de las variantes de *splicing* de genes previamente identificados, presenta limitaciones cuando se trata de identificar de forma masiva los cambios funcionales potenciales en estas variantes aún no descritas. Esto se debe a que la anotación primaria se basa principalmente en la similitud de secuencias con genes conocidos, lo que generalmente asigna funciones preexistentes sin evaluar a fondo el impacto individual de cada variante [212]. Como resultado, muchas de estas variantes pueden estar subestimadas en términos de su relevancia funcional. En particular, la anotación funcional insuficiente puede ser un desafío, especialmente cuando se consideran los efectos que pueden tener estas variantes en procesos biológicos críticos, como la regulación del ciclo celular y la respuesta al estrés celular [213]. Para

el caso particular de la espermatogénesis, debemos recordar que se cuenta con una composición específica del *spliceosoma*, lo que potencia la generación de nuevas variantes de ARNm, que constituyen la base para la producción de la gran cantidad de diferentes isoformas proteicas, esenciales para el desarrollo y funcionalidad de los espermatozoides [214].

Hemos señalado varias veces a lo largo de este texto que el *splicing* alternativo no sólo permite diversificar las funciones proteicas, sino que también puede generar isoformas con nuevas funciones o, por el contrario, provocar la pérdida de dominios clave que alteran la funcionalidad de la proteína resultante [74,127]. Este fenómeno es especialmente evidente en genes involucrados en la señalización celular y el control del ciclo celular, donde la pérdida o adquisición de dominios funcionales puede influir de manera significativa en procesos esenciales como la apoptosis y la diferenciación celular [127]. Adicionalmente, se ha demostrado que muchos eventos de empalme alternativo en un organismo pueden no mostrar una ventaja selectiva inmediata, pero la existencia del empalme alternativo en sí misma tiene un valor adaptativo. Esto se debe a que proporciona una fuente de variabilidad de ARNm que puede ser aprovechada evolutivamente para generar variabilidad funcional relevante, sin la necesidad de eliminar funciones génicas existentes o generar nuevos genes [73].

Además, estudios como los de Tress y colaboradores [215] han demostrado que sólo una fracción de las variantes de *splicing* genera isoformas con relevancia biológica significativa, lo que subraya la importancia de realizar un análisis detallado de cada variante para evaluar su funcionalidad. De hecho, en nuestros estudios hemos llevado a cabo este tipo de análisis en casos específicos, observando si la variante generaba la pérdida o ganancia de dominios funcionales clave en la proteína. Este enfoque nos ha permitido entender mejor las implicancias biológicas de estas variantes, destacando la necesidad de ir más allá de la anotación primaria para descubrir nuevas funciones, especialmente en tejidos complejos como el testículo, donde la diversidad proteica es particularmente elevada.

Respecto a lo que se refiere a los genes hasta ahora no anotados, la información proporcionada por anotación primaria resulta más relevante, ya que nos podría dar un indicio de su función o, al menos, nos puede decir a qué se asemejan. Dentro de

estos genes no anotados, pusimos foco en 22 genes probablemente codificantes, para los cuales pudimos identificar péptidos provenientes de ellos mediante el análisis proteómico. Al realizar una búsqueda y análisis detallado de estos genes, encontramos que algunos de estos genes fueron anotados en el transcurso de esta investigación, constituyendo una validación adicional de nuestros análisis y resultados.

Uno de los resultados más sorprendentes de este estudio se aprecia en la expresión del gen *LaidX*, cuya existencia ha sido recientemente anotada en el genoma del ratón. Este gen, con un único transcripto de aproximadamente 25 Kb sin intrones, fue clasificado como codificante [190]. La región en la que se encuentra este gen constituye una zona muy compleja del cromosoma X, que históricamente ha representado un desafío para los métodos de secuenciación. Dicha región presenta múltiples elementos repetidos, y su anotación fue posible gracias a una corrección estructural que se efectuó en la versión GRCm39 del genoma del ratón. Este hallazgo es consistente con estudios previos que destacan cómo los avances en la tecnología de secuenciación, y la mejora de los ensamblajes genómicos, permiten la identificación de genes previamente ocultos, o mal anotados [216].

Al realizar un mapeo de nuestros datos sobre la nueva versión del genoma, observamos una cobertura continua en la región correspondiente al gen *LaidX*, lo cual coincide con los hallazgos reportados recientemente por Arlt y colaboradores [190]. Aún más interesante, es que nuestros análisis de proteómica revelaron la presencia de péptidos que corresponden a diferentes regiones de la proteína predicha de *LaidX*, lo que refuerza la hipótesis de la existencia de esta proteína a nivel biológico, constituyendo la primera evidencia empírica, más allá de la predicción teórica. Este hallazgo es especialmente relevante, ya que los genes sin intrones y los elementos repetidos, como los identificados en *LaidX*, han sido históricamente difíciles de caracterizar y, sin embargo, podrían desempeñar funciones cruciales aún no comprendidas completamente [217].

El descubrimiento de *LaidX* añade evidencia al creciente reconocimiento de que los genomas contienen muchas más regiones codificantes potenciales de lo que se pensaba previamente, especialmente en regiones repetitivas u "oscuras" del genoma

[218]. Este tipo de hallazgos abre nuevas líneas de investigación sobre la funcionalidad de genes previamente no caracterizados y la importancia de refinar nuestras herramientas bioinformáticas, para seguir descubriendo estas regiones exóticas pero funcionales.

Hay que mencionar que, además de la contribución que hemos hecho a identificar nuevos genes y transcriptos, y de profundizar en posibles anotaciones y funciones, hemos dejado por fuera 16.265 transcriptos, no por considerar que no fueran confiables, sino por presentar algún tipo de conflicto entre los programas de potencial codificante empleados para clasificarlos. Esto resalta aún más la complejidad y desconocimiento que tenemos hoy día, en la biología en general, a un nivel tal que, aún en un organismo modelo tan estudiado como el ratón, todavía nos queda mucho por profundizar para poder identificar con certeza los intrincados mecanismos moleculares necesarios para que se den, de forma correcta, muchos procesos biológicos.

6. CONCLUSIONES Y PERSPECTIVAS

En este estudio se abordó la caracterización de transcriptos no anotados en la espermatogénesis, utilizando análisis de RNAseq en poblaciones celulares altamente purificadas. Como conclusiones principales,

- La combinación de diferentes herramientas bioinformáticas permitió mejorar la anotación de transcriptos en el genoma del ratón, contribuyendo al conocimiento sobre la expresión génica durante la espermatogénesis, y generando una valiosa base de datos sobre transcriptos codificantes, variantes de *splicing* y lncRNAs expresados a lo largo de las distintas etapas del proceso.

- Se identificaron 33.002 transcriptos no anotados expresados en el testículo del ratón, de los cuales 18.607 se expresaron en la profase meiótica temprana (LZ), siendo 4.768 exclusivos de estas etapas y evidenciando que, a pesar de su corta duración y baja representación celular, dichas etapas son particularmente ricas transcripcionalmente, a diferencia de lo que se creía en el pasado.

- La gran mayoría de los transcriptos no anotados identificados corresponde a lncRNAs (13.471), lo que subraya la abundancia de éstos durante la espermatogénesis, y el menor conocimiento de los mismos respecto a los genes codificantes.

- De los transcriptos no anotados clasificados como codificantes (2.794), la mayoría resultaron ser variantes de *splicing* de genes ya anotados (2.571), lo que remarca la importancia del mecanismo de *splicing* alternativo en el testículo.

- El análisis proteómico permitió validar el 70 % de los transcriptos “nuevos” con alto potencial codificante identificados (1.949 de 2.794), confirmando de esta forma la existencia de un importante número de proteínas expresadas durante la espermatogénesis, hasta ahora no anotadas.

- Se logró mapear y ratificar la existencia del transcripto de *LaidX*, un gen sin intrones localizado en una región compleja del cromosoma X, detectando péptidos, por primera vez, en múltiples fragmentos de la proteína predicha, lo que constituye la primera evidencia empírica de su existencia y abre la puerta a su caracterización funcional. A la vez, aporta evidencia en la dirección de que regiones complejas u "oscuras" del genoma pueden, sin embargo, codificar proteínas atípicas, y en la necesidad de profundizar el estudio de esas regiones exóticas.

-Los resultados obtenidos confirman la presencia de una gran cantidad de transcriptos no anotados previamente, tanto codificantes como no codificantes, que potencialmente podrían cumplir funciones en la regulación de la meiosis y la espermiogénesis. En particular, la identificación de posibles nuevos genes codificantes y no codificantes resalta la complejidad del paisaje transcriptómico en la gametogénesis masculina, y abre nuevas preguntas sobre sus posibles funciones.

-La identificación de transcriptos sin homología conocida en bases de datos sugiere la existencia de elementos genéticos específicos de la espermatogénesis en mamíferos, cuya función y regulación aún no han sido descritas. La exploración de estos elementos podría ofrecer nuevas perspectivas sobre la evolución y especialización de los mecanismos moleculares que controlan la producción de gametos.

-En conjunto, estos resultados revelan que la complejidad del transcriptoma y proteoma testicular son aún bastante mayores de lo reconocido, aportando miles de transcriptos y numerosas isoformas proteicas hasta ahora desconocidas. Nuestros hallazgos plantean la posibilidad de que algunos de estos transcriptos sean elementos clave en la regulación de la fertilidad, lo que podría tener implicancias significativas, tanto en la comprensión de la biología básica de la reproducción, como en el desarrollo de posibles biomarcadores o estrategias terapéuticas para tratar casos de infertilidad masculina.

Este trabajo abre diversas perspectivas:

- A partir de los hallazgos realizados, futuras investigaciones podrán enfocarse en la validación funcional de los transcriptos identificados (o, al menos, algunos de ellos seleccionados), mediante experimentos de edición genética (como CRISPR-Cas9) y estudios de expresión en diferentes etapas del desarrollo espermatogénico. Estos estudios permitirán determinar si los transcriptos identificados y/o (para el caso de los codificantes) sus productos proteicos, desempeñan funciones esenciales en la regulación de la meiosis, la diferenciación celular y la maduración de los espermatozoides. Por ejemplo, se podrían estudiar las isoformas proteicas producto de genes esenciales para la espermatogénesis que hemos validado por RT-PCR, de modo de caracterizar su función (tal es el caso de la isoforma de MSH5, cuya caracterización hemos iniciado).

- Además, sería interesante analizar la conservación de estos transcriptos en otras especies, para determinar su posible papel evolutivo. Comparaciones entre mamíferos podrían revelar si estos genes desempeñan funciones más generales en la reproducción, o si son específicos del ratón.

- La integración de técnicas de secuenciación de tercera generación, como la secuenciación de largo alcance (PacBio o Nanopore), podrá permitir mejorar la caracterización estructural de estos transcriptos y definir con mayor precisión sus isoformas y regiones reguladoras. Complementariamente, estudios de interacción proteína-ARN y perfiles epigenéticos podrían contribuir a esclarecer los mecanismos mediante los cuales estos transcriptos regulan la espermatogénesis.

- También se podrían realizar estudios de mecanismos de regulación del *splicing*, analizando la expresión y actividad de factores de *splicing* específicos (por ejemplo, hnRNPs, proteínas SR) en cada etapa celular mediante Western-blot y ensayos de unión ARN-proteína, y/o emplear ensayos de reporteros de *splicing in vivo* para medir cómo alteraciones en factores clave afectan la inclusión/exclusión de exones.

Finalmente, deseamos mencionar que una significativa parte de este trabajo (exceptuando los análisis proteómicos, la caracterización de la isoforma de MSH5, y

algunos de los estudios de transcriptómica), fueron publicados en el trabajo: Romeo C, Trovero MF, Radío S, Smirsich P, Rodríguez-Casuriaga R, Rodríguez-Casuriaga R, Geisinger A, Sotelo-Silveira JR (2024). "*Uncovering a multitude of stage-specific splice variants and putative protein isoforms generated along mouse spermatogenesis*", *BMC Genomics* 25:295; doi: 10.1186/s12864-024-10170-z

Adicionalmente, este trabajo, en formato póster, recibió el Primer Premio en las Jornadas de la Sociedad de Bioquímica y Biología Molecular (JSSBM, Uruguay), en octubre 2023.

7. ANEXO: MATERIAL SUPLEMENTARIO

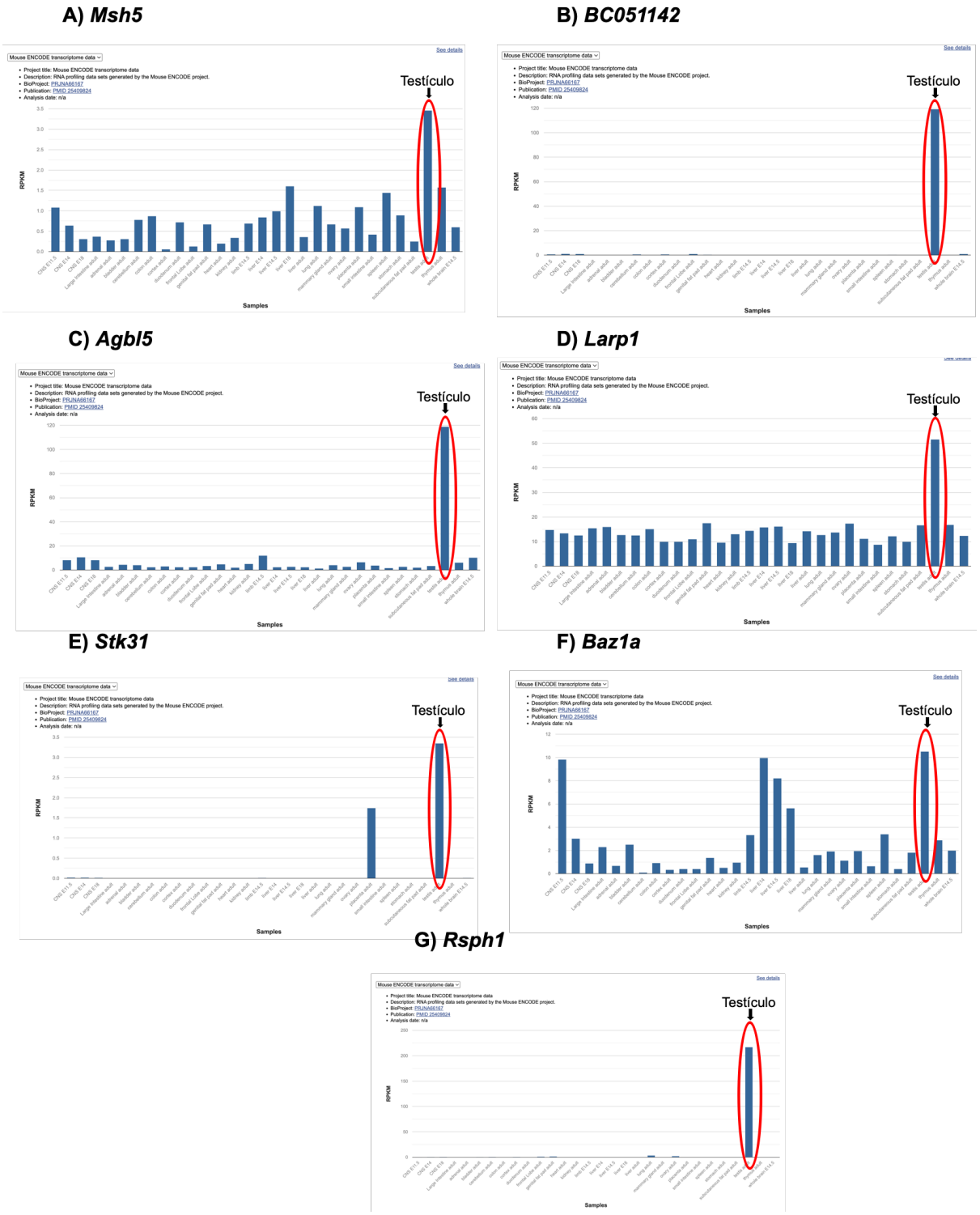


Figura Suplementaria S1. Niveles de expresión en diferentes tejidos, de los genes seleccionados para confirmar por RT-PCR. Se aprecian los niveles de expresión medidos en RPKM en los diferentes tejidos del ratón. Los datos fueron tomados de ENCODE, del proyecto PRJNA66167.

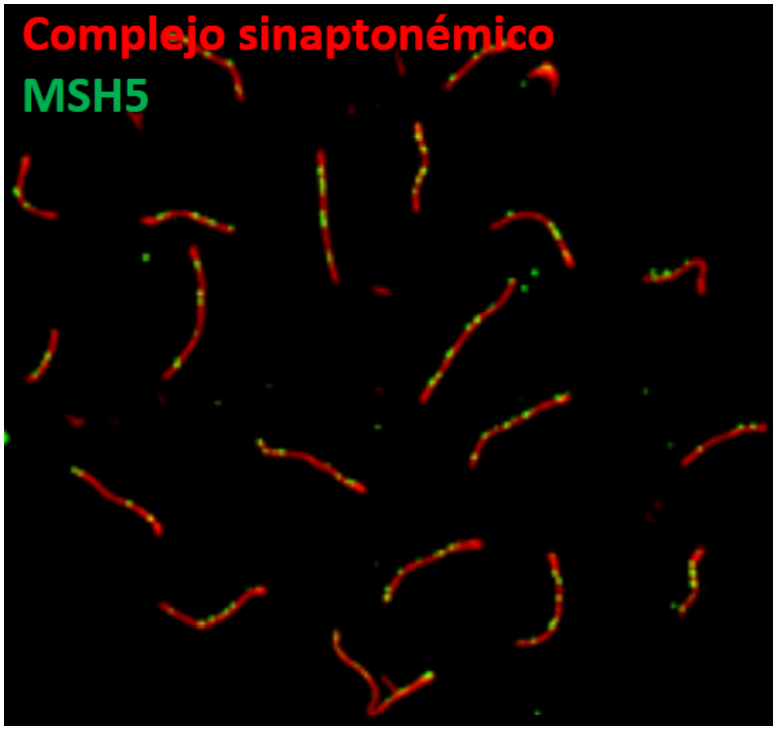


Figura Suplementaria S2. Localización de la proteína MSH5 canónica sobre los cromosomas meióticos. En rojo se aprecia la marca de SYCP3, un marcador de los complejos sinaptonémicos, y en verde, MSH5 canónica. (Imagen tomada y modificada de Milano y colaboradores, 2019).

Tabla Suplementaria S1. *Primers* empleados para los análisis de RT-PCR. Se presentan los nombres de los genes contra los cuales se realizaron las confirmaciones de variantes de *splicing*, los nombres específicos de cada par de *primers* utilizados, las secuencias de los mismos, y el tamaño esperado de los amplicones (en pares de bases).

Gen	Transcripto	Secuencia 5'->3' primer	Largo de Fragmento
Agbl5	ENSMUST00000114700.9	CTGCCAGAGGAGAGTTTCCG TCTCTCTGGTCTGGGGGATG	136
	MSTRG.27115.4	CCTGACTGTGCTGAAACGGA TGTCTCTGTCATCTCAAAGGTG	222
Baz1a	ENSMUST00000173433.8 and MSTRG.7385.3	ACTGCAATCTTCCAGGCGAT CGTAGCGTCGAATCCTCCTC	216-312
	MSTRG.7385.3	GTGAAGGCCCACTGTGTGAA AAACTGCAGCCTTTGGTGTC	124
BC051142	ENSMUST00000097348.10	GCTGCTTACCCCAGGAAACA CAACAGCCCGTCTCCATCAT	158
	MSTRG.15748.86	GTGAAAGCTGCACCCCTGTA GCAGGGTGGACTCAGAGAAT	150
	MSTRG.15748.95	AGATCCAGCTTTCTCTCTGGAA GAGGAAACTAGGTAGGAGAGCA	102
Larp1	ENSMUST00000178636.2	TCCTGTGACTCACTTGTCGC AGAGGGCTGAGTGAGGAGAG	160
	MSTRG.5062.14	CTCCGGGATGTCCAAGGGAG ACTTTCGGTAGCCAAACTGGT	114
Msh5	ENSMUST00000174556.8	ACGCATGAAGTTGTCCCACA GAGTGAGGCTCTCCTAGGCT	104
	MSTRG.15777.15	CAGATCTTTCCAGGCACCTCTC CAAGGTCCAATACCCGAGTCA	118

Rsph1	ENSMUST00000024832.9 y MSTRG.15622.3	ACAGGTACCAGGGCAAGTTC CCAGTAAGTCCCTGCGGTTC	193-265
Stk31	ENSMUST00000024171.14	GCTGTGGCGCAAAGTGTAAG TGGGCCCAAAATGTTACTGC	119
	MSTRG.29908.1	CTGCGCTCCCTCCGCTAT ATCCATTTGGACAAGTCCTGGAA	106

Tabla Suplementaria S2. Expresión y anotación de los transcriptos detectados. ENSMUST significa transcriptos anotados por Ensembl, mientras que MSTRG designa transcriptos no anotados. La tabla se accede en el siguiente link:

https://pmc.ncbi.nlm.nih.gov/articles/instance/10953240/bin/12864_2024_10170_MOESM8_ESM.xlsx

Tabla Suplementaria S3. Expresión y anotación de los 223 transcriptos recientemente identificados con alto potencial de codificación, que corresponden a 191 genes no anotados. La tabla se accede en el siguiente link:

https://pmc.ncbi.nlm.nih.gov/articles/instance/10953240/bin/12864_2024_10170_MOESM9_ESM.xlsx

Tabla suplementaria S4. Lista de los 3.952 transcriptos para los que se encontró una anotación primaria. Link de acceso:

https://docs.google.com/spreadsheets/d/1P90RBAspgTEUVGIUjkRfH_F3y53AWfqN/edit?usp=sharing&ouid=102469611408794891700&rtpof=true&sd=true

Tabla Suplementaria S5. Selección de variantes de *splicing* para confirmación molecular. Link de acceso:

https://docs.google.com/spreadsheets/d/1fWy99lVJNlLwwMPKhc08rPyWlwqnTEr0/edit?usp=sharing&ouid=102469611408794891700&rtpof=true&sd=true

Tabla Suplementaria S6. Proteínas confirmadas de genes no anotados. El link de acceso es:

https://docs.google.com/spreadsheets/d/1VyCzMbtj11NRosy_Dmgv0y7s6zYRiVbx/edit?usp=sharing&ouid=102469611408794891700&rtpof=true&sd=true

8. REFERENCIAS BIBLIOGRÁFICAS

1. Griswold MD. Spermatogenesis: The Commitment to Meiosis. *Physiol Rev*. 2016;96(1):1. doi:10.1152/PHYSREV.00013.2015
2. Sofikitis N, Giotitsas N, Tsounapi P, Baltogiannis D, Giannakis D, Pardalidis N. Hormonal regulation of spermatogenesis and spermiogenesis. *J Steroid Biochem Mol Biol*. 2008;109(3-5):323-330. doi:10.1016/J.JSBMB.2008.03.004
3. Du L, Chen W, Cheng Z, et al. Novel Gene Regulation in Normal and Abnormal Spermatogenesis. *Cells*. 2021;10(3):1-16. doi:10.3390/CELLS10030666
4. Erickson RP. Post-meiotic gene expression. *Trends in Genetics*. 1990;6(8):264-268. doi:10.1016/0168-9525(90)90209-O
5. Neto FTL, Bach PV, Najari BB, Li PS, Goldstein M. Spermatogenesis in humans and its affecting factors. *Semin Cell Dev Biol*. 2016;59:10-26. doi:10.1016/J.SEMCDB.2016.04.009
6. De Kretser DM, Loveland KL, Meinhardt A, Simorangkir D, Wreford N. Spermatogenesis. *Hum Reprod*. 1998;13 Suppl 1(SUPPL. 1):1-8. doi:10.1093/HUMREP/13.SUPPL_1.1
7. Wu S, Cheng CY. Blood-Testis Barrier. *Encyclopedia of Molecular Pharmacology*. Published online 2021:330-335. doi:10.1007/978-3-030-57401-7_10017
8. Recchia K, Jorge AS, Pessôa LV de F, et al. Actions and Roles of FSH in Germinative Cells. *International Journal of Molecular Sciences 2021, Vol 22, Page 10110*. 2021;22(18):10110. doi:10.3390/IJMS221810110
9. Richer G, Baert Y, Goossens E. In-vitro spermatogenesis through testis modelling: Toward the generation of testicular organoids. *Andrology*. 2020;8(4):879-891. doi:10.1111/ANDR.12741
10. Cannarella R, Condorelli RA, Mongioì LM, La Vignera S, Calogero AE. Molecular Biology of Spermatogenesis: Novel Targets of Apparently Idiopathic Male Infertility. *Int J Mol Sci*. 2020;21(5). doi:10.3390/IJMS21051728
11. Bolcun-Filas E, Handel MA. Meiosis: the chromosomal foundation of reproduction. *Biol Reprod*. 2018;99(1):112-126. doi:10.1093/BIOLRE/IOY021
12. Ishiguro K ichiro, Matsuura K, Tani N, et al. MEIOSIN Directs the Switch from Mitosis to Meiosis in Mammalian Germ Cells. *Dev Cell*. 2020;52(4):429-

445.e10. doi:10.1016/J.DEVCEL.2020.01.010/ATTACHMENT/52C7788A-4C0C-4DD8-832D-437EFAF068BB/MMC7.XLSX

13. Zickler D, Kleckner N. Recombination, Pairing, and Synapsis of Homologs during Meiosis. *Cold Spring Harbor Laboratory Press*. 2015;7(6):1-26. doi:10.1101/cshperspect.a016626
14. Fraune J, Schramm S, Alsheimer M, Benavente R. The mammalian synaptonemal complex: Protein components, assembly and role in meiotic recombination. *Exp Cell Res*. 2012;318(12):1340-1346. doi:10.1016/j.yexcr.2012.02.018
15. Page SL, Hawley RS. The genetics and molecular biology of the synaptonemal complex. *Annu Rev Cell Dev Biol*. 2004;20:525-558. doi:10.1146/ANNUREV.CELLBIO.19.111301.155141
16. Schücker K, Holm T, Franke C, Sauer M, Benavente R. Elucidation of synaptonemal complex organization by super-resolution imaging with isotropic resolution. *Proc Natl Acad Sci U S A*. 2015;112(7):2029-2033. doi:10.1073/PNAS.1414814112/VIDEO-1
17. Esponda P, Giménez-Martín G. The attachment of the synaptonemal complex to the nuclear envelope - An ultrastructural and cytochemical analysis. *Chromosoma*. 1972;38(4):405-417. doi:10.1007/BF00320159/METRICS
18. Scherthan H. Telomere attachment and clustering during meiosis. *Cellular and Molecular Life Sciences*. 2007;64(2):117-124. doi:10.1007/S00018-006-6463-2/METRICS
19. Alberts B, Johnson A, Lewis J, et al. *Molecular Biology of the Cell, Sixth Edition*.; 2014.
20. Zickler D, Kleckner N. The leptotene-zygotene transition of meiosis. *Annu Rev Genet*. 1998;32:619-697. doi:10.1146/ANNUREV.GENET.32.1.619
21. Moens PB, Marcon E, Shore JS, Kochakpour N, Spyropoulos B. Initiation and resolution of interhomolog connections: crossover and non-crossover sites along mouse synaptonemal complexes. *J Cell Sci*. 2007;120(Pt 6):1017-1027. doi:10.1242/JCS.03394
22. Handel MA, Schimenti JC. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nat Rev Genet*. 2010;11(2):124-136. doi:10.1038/NRG2723

23. Gerton JL, Hawley RS. Homologous chromosome interactions in meiosis: diversity amidst conservation. *Nature Reviews Genetics 2005 6:6*. 2005;6(6):477-487. doi:10.1038/nrg1614
24. Cohen PE, Pollack SE, Pollard JW. Genetic analysis of chromosome pairing, recombination, and cell cycle control during first meiotic prophase in mammals. *Endocr Rev*. 2006;27(4):398-426. doi:10.1210/ER.2005-0017
25. Roeder GS, Bailis JM. The pachytene checkpoint. *Trends Genet*. 2000;16(9):395-403. doi:10.1016/S0168-9525(00)02080-1
26. Teves ME, Roldan ERS. Spermbauplan and function and underlying processes of sperm formation and selection. *Physiol Rev*. 2022;102(1):7-60. doi:10.1152/PHYSREV.00009.2020/ASSET/IMAGES/LARGE/PHYSREV.00009.2020_F010.JPEG
27. Durairajanayagam D, Rengan AK, Sharma RK, Agarwal A. Sperm Biology from Production to Ejaculation. *Unexplained Infertility: Pathophysiology, Evaluation and Treatment*. Published online January 1, 2015:29-42. doi:10.1007/978-1-4939-2140-9_5
28. Eddy EM, O'Brien DA. Gene expression during mammalian meiosis. *Curr Top Dev Biol*. 1998;37(3):141-200.
29. Gan H, Cai T, Lin X, et al. Integrative proteomic and transcriptomic analyses reveal multiple post-transcriptional regulatory mechanisms of mouse spermatogenesis. *Mol Cell Proteomics*. 2013;12(5):1144-1157. doi:10.1074/MCP.M112.020123
30. Elliott D. Pathways of post-transcriptional gene regulation in mammalian germ cell development. *Cytogenet Genome Res*. 2003;103(3-4):210-216. doi:10.1159/000076806
31. Wang T, Gao H, Li W, Liu C. Essential Role of Histone Replacement and Modifications in Male Fertility. *Front Genet*. 2019;10:470883. doi:10.3389/FGENE.2019.00962/BIBTEX
32. Zhong J, Peters AH, Lee K, Braun RE. A double-stranded RNA binding protein required for activation of repressed messages in mammalian germ cells. *Nat Genet*. 1999;22(2):171-174. doi:10.1038/9684
33. Meikar O, Da Ros M, Korhonen H, Kotaja N. Chromatoid body and small RNAs in male germ cells. *Reproduction*. 2011;142(2):195-209. doi:10.1530/REP-11-0057

34. Anbazhagan R, Kavarthapu R, Dufau ML. Chromatoid Bodies in the Regulation of Spermatogenesis: Novel Role of GRTH. *Cells 2022, Vol 11, Page 613*. 2022;11(4):613. doi:10.3390/CELLS11040613
35. Parvinen M. The chromatoid body in spermatogenesis. *Int J Androl*. 2005;28(4):189-201. doi:10.1111/j.1365-2605.2005.00542.x
36. Kotaja N, Sassone-Corsi P. The chromatoid body: a germ-cell- specific RNA-processing centre. *Nat Rev Mol Cell Biol*. 2007;8(January):85-90.
37. Meikar O, Vagin V V., Chalmel F, et al. An atlas of chromatoid body components. *Rna*. 2014;20(4):483-495. doi:10.1261/rna.043729.113
38. Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. *Nature 2012 489:7414*. 2012;489(7414):101-108. doi:10.1038/nature11233
39. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature 2006 442:7099*. 2006;442(7099):199-202. doi:10.1038/nature04917
40. Atkinson SR, Marguerat S, Bähler J. Exploring long non-coding RNAs through sequencing. *Semin Cell Dev Biol*. 2012;23(2):200-205. doi:10.1016/J.SEMCDB.2011.12.003
41. Gao N, Li Y, Li J, et al. Long Non-Coding RNAs: The Regulatory Mechanisms, Research Strategies, and Future Directions in Cancers. *Front Oncol*. 2020;10:598817. doi:10.3389/FONC.2020.598817/BIBTEX
42. Kapranov P, Cheng J, Dike S, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*. 2007;316(5830):1484-1488. doi:10.1126/SCIENCE.1138341
43. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet*. 2009;10(3):155-159. doi:10.1038/NRG2521
44. Li LJ, Leng RX, Fan YG, Pan HF, Ye DQ. Translation of noncoding RNAs: Focus on lncRNAs, pri-miRNAs, and circRNAs. *Exp Cell Res*. 2017;361(1):1-8. doi:10.1016/J.YEXCR.2017.10.010
45. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics 2015 17:1*. 2015;17(1):47-62. doi:10.1038/nrg.2015.10
46. Xing J, Liu H, Jiang W, Wang L. LncRNA-Encoded Peptide: Functions and Predicting Methods. *Front Oncol*. 2021;10:622294 doi:10.3389/FONC.2020.622294/BIBTEX

47. Cao J. The functional role of long non-coding RNAs and epigenetics. *Biol Proced Online*. 2014;16(1):11. doi:10.1186/1480-9222-16-11
48. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell*. 2013;154(1):26. doi:10.1016/J.CELL.2013.06.020
49. Wu T, Du Y. LncRNAs: From Basic Research to Medical Application. *Int J Biol Sci*. 2017;13(3):295-307. doi:10.7150/IJBS.16968
50. Cabili M, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. 2011;25(18):1915-1927. doi:10.1101/GAD.17446611
51. Derrien T, Johnson R, Bussotti G, 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. doi:10.1101/GR.132159.111
52. Nakagawa S, Kageyama Y. Nuclear lncRNAs as epigenetic regulators-beyond skepticism. *Biochim Biophys Acta*. 2014;1839(3):215-222. doi:10.1016/J.BBAGRM.2013.10.009
53. Pauli A, Valen E, Lin MF, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*. 2012;22(3):577-591. doi:10.1101/GR.133009.111
54. Ran M, Chen B, Li Z, et al. Systematic Identification of Long Noncoding RNAs in Immature and Mature Porcine Testes. *Biol Reprod*. 2016;94(4). doi:10.1095/BIOLREPROD.115.136911
55. Sun J, Lin Y, Wu J. Long Non-Coding RNA Expression Profiling of Mouse Testis during Postnatal Development. *PLoS One*. 2013;8(10):e75750. doi:10.1371/JOURNAL.PONE.0075750
56. Niazi F, Valadkhan S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA*. 2012;18(4):825-843. doi:10.1261/RNA.029520.111
57. Melé M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res*. 2017;27(1):27-37. doi:10.1101/GR.214205.116
58. Soumillon M, Necsulea A, Weier M, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep*. 2013;3(6):2179-2190. doi:10.1016/J.CELREP.2013.05.031

59. Trovero MF, Geisinger A. ARNs no codificantes largos en patología testicular. *An Facultad Med*. 2019;6(1):10-27. doi:10.25184/anfamed2019v6n1a8
60. Faulkner GJ, Carninci P. Altruistic functions for selfish DNA. *Cell Cycle*. 2009;8(18):2895-2900. doi:10.4161/CC.8.18.9536
61. Mangiavacchi A, Liu P, Della Valle F, Orlando V. New insights into the functional role of retrotransposon dynamics in mammalian somatic cells. *Cellular and Molecular Life Sciences 2021 78:13*. 2021;78(13):5245-5256. doi:10.1007/S00018-021-03851-5
62. Honson DD, Macfarlan TS. A lncRNA-like Role for LINE1s in Development. *Dev Cell*. 2018;46(2):132. doi:10.1016/J.DEVCEL.2018.06.022
63. Frith MC, Bailey TL, Kasukawa T, et al. Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol*. 2006;3(1):40-48. doi:10.4161/RNA.3.1.2789
64. Necsulea A, Soumillon M, Warnefors M, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*. 2014;505(7485):635-640. doi:10.1038/NATURE12943
65. Sharma H, Carninci P. The Secret Life of lncRNAs: Conserved, yet Not Conserved. *Cell*. 2020;181(3):512-514. doi:10.1016/J.CELL.2020.04.012
66. Latos PA, Pauler FM, Koerner M V., et al. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science*. 2012;338(6113):1469-1472. doi:10.1126/SCIENCE.1228110
67. Nitsche A, Stadler PF. Evolutionary clues in lncRNAs. *Wiley Interdiscip Rev RNA*. 2017;8(1). doi:10.1002/WRNA.1376
68. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*. 2015;11(7):1110-1122. doi:10.1016/J.CELREP.2015.04.023
69. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet*. 2016;17(10):601-614. doi:10.1038/NRG.2016.85
70. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet*. 2014;15(1):7-21. doi:10.1038/NRG3606

71. Luk ACS, Chan WY, Rennert OM, Lee TL. Long noncoding RNAs in spermatogenesis: Insights from recent high-throughput transcriptome studies. *Reproduction*. 2014;147(5):131-141. doi:10.1530/REP-13-0594
72. Engreitz JM, Haines JE, Perez EM, et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*. 2016;539(7629):452-455. doi:10.1038/NATURE20149
73. Marasco LE, Kornblihtt AR. The physiology of alternative splicing. *Nat Rev Mol Cell Biol*. 2023;24(4):242-254. doi:10.1038/S41580-022-00545-Z
74. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol*. 2017;18(7):437-451. doi:10.1038/NRM.2017.27
75. Wilkinson ME, Charenton C, Nagai K. RNA Splicing by the Spliceosome. *Annu Rev Biochem*. 2020;89(Volume 89, 2020):359-388. doi:10.1146/ANNUREV-BIOCHEM-091719-064225/CITE/REFWORKS
76. Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet*. 2010;11(1):75-87. doi:10.1038/NRG2673
77. Braunschweig U, Gueroussov S, Plocik AM, Graveley BR, Blencowe BJ. Dynamic integration of splicing within gene regulatory pathways. *Cell*. 2013;152(6):1252-1269. doi:10.1016/J.CELL.2013.02.034
78. Ule J, Blencowe BJ. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol Cell*. 2019;76(2):329-345. doi:10.1016/J.MOLCEL.2019.09.017
79. Wang Y, Xie Z, Kutschera E, Adams JI, Kadash-Edmondson KE, Xing Y. rMATS-turbo: an efficient and flexible computational tool for alternative splicing analysis of large-scale RNA-seq data. *Nature Protocols 2024 19:4*. 2024;19(4):1083-1104. doi:10.1038/s41596-023-00944-2
80. Song H, Wang L, Chen D, Li F. The Function of Pre-mRNA Alternative Splicing in Mammal Spermatogenesis. *Int J Biol Sci*. 2020;16(1):38-48. doi:10.7150/IJBS.34422
81. Bonnal SC, López-Oreja I, Valcárcel J. Roles and mechanisms of alternative splicing in cancer — implications for care. *Nat Rev Clin Oncol*. 2020;17(8):457-474. doi:10.1038/S41571-020-0350-X

82. Shima JE, McLean DJ, McCarrey JR, Griswold MD. The Murine Testicular Transcriptome: Characterizing Gene Expression in the Testis During the Progression of Spermatogenesis1. *Biol Reprod*. 2004;71(1):319-330. doi:10.1095/biolreprod.103.026880
83. Rossi P, Dolci S, Sette C, et al. Analysis of the gene expression profile of mouse male meiotic germ cells. *Gene Expression Patterns*. 2004;4(3):267-281. doi:10.1016/j.modgep.2003.11.003
84. Huang SY, Tam MF, Hsu YT, et al. Developmental changes of heat-shock proteins in porcine testis by a proteomic analysis. *Theriogenology*. 2005;64(9):1940-1955. doi:10.1016/j.theriogenology.2005.04.024
85. Ellis PJI, Furlong RA, Conner SJ, et al. Coordinated transcriptional regulation patterns associated with infertility phenotypes in men. *J Med Genet*. 2007;44(8):498-508. doi:10.1136/jmg.2007.049650
86. Feig C, Kirchhoff C, Ivell R, Naether O, Schulze W, Spiess AN. A new paradigm for profiling testicular gene expression during normal and disturbed human spermatogenesis. *Mol Hum Reprod*. 2007;13(1):33-43. doi:10.1093/molehr/gal097
87. Laiho A, Kotaja N, Gyenesei A, Sironen A. Transcriptome profiling of the murine testis during the first wave of spermatogenesis. *PLoS One*. 2013;8(4). doi:10.1371/JOURNAL.PONE.0061558
88. Margolin G, Khil PP, Kim J, Bellani MA, Camerini-Otero RD. Integrated transcriptome analysis of mouse spermatogenesis. *BMC Genomics*. 2014;15(1). doi:10.1186/1471-2164-15-39
89. da Cruz I, Rodríguez-Casuriaga R, Santiñaque FF, et al. Transcriptome analysis of highly purified mouse spermatogenic cell populations: gene expression signatures switch from meiotic-to postmeiotic-related processes at pachytene stage. *BMC Genomics*. 2016;17(1). doi:10.1186/S12864-016-2618-1
90. Chen Y, Zheng Y, Gao Y, et al. Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res*. 2018;28(9):879-896. doi:10.1038/S41422-018-0074-Y
91. Rabbani M, Zheng X, Manske GL, et al. Decoding the Spermatogenesis Program: New Insights from Transcriptomic Analyses. *Annu Rev Genet*. 2022;56:339-368. doi:10.1146/ANNUREV-GENET-080320-040045

92. Chowdhary A, Satagopam V, Schneider R. Long Non-coding RNAs: Mechanisms, Experimental, and Computational Approaches in Identification, Characterization, and Their Biomarker Potential in Cancer. *Front Genet*. 2021;12:649619. doi:10.3389/FGENE.2021.649619/BIBTEX
93. Zhou Q, Wang M, Yuan Y, et al. Complete Meiosis from Embryonic Stem Cell-Derived Germ Cells in Vitro. *Cell Stem Cell*. 2016;18(3):330-340. doi:10.1016/J.STEM.2016.01.017
94. Lei Q, Lai X, Eliveld J, Chuva de Sousa Lopes SM, van Pelt AMM, Hamer G. In Vitro Meiosis of Male Germline Stem Cells. *Stem Cell Reports*. 2020;15(5):1140-1153. doi:10.1016/J.STEMCR.2020.10.006
95. Lei Q, van Pelt AMM, Hamer G. In vitro spermatogenesis: Why meiotic checkpoints matter. *Curr Top Dev Biol*. 2023;151:345-369. doi:10.1016/BS.CTDB.2022.04.009
96. Bellve AR, Cavicchia JC, Millette CF, O'Brien DA, Bhatnagar YM, Dym M. Spermatogenic cells of the prepuberal mouse. Isolation and morphological characterization. *Journal of Cell Biology*. 1977;74(1):68-85. doi:10.1083/jcb.74.1.68
97. Iguchi N, Tobias JW, Hecht NB. Expression profiling reveals meiotic male germ cell mRNAs that are translationally up- and down-regulated. *Proc Natl Acad Sci U S A*. 2006;103(20):7712-7717. doi:10.1073/pnas.0510999103
98. Geisinger A, Rodríguez-Casuriaga R, Benavente R. Transcriptomics of Meiosis in the Male Mouse. *Front Cell Dev Biol*. 2021;9. doi:10.3389/FCELL.2021.626020
99. Guo B, Zhu X, Li X, Yuan CF. The Roles of LncRNAs in Osteogenesis, Adipogenesis and Osteoporosis. *Curr Pharm Des*. 2021;27(1):91-104. doi:10.2174/1381612826666200707130246
100. Tan K, Song HW, Wilkinson MF. Single-cell RNAseq analysis of testicular germ and somatic cell development during the perinatal period. *Development*. 2020;147(3). doi:10.1242/DEV.183251
101. Ball RL, Fujiwara Y, Sun F, et al. Regulatory complexity revealed by integrated cytological and RNA-seq analyses of meiotic substages in mouse spermatocytes. *BMC Genomics*. 2016;17(1):1-17. doi:10.1186/s12864-016-2865-1

102. Bellvé AR. Purification, culture, and fractionation of spermatogenic cells. *Methods Enzymol*. 1993;225(C):84-113. doi:10.1016/0076-6879(93)25009-Q
103. Bao J, Wu J, Schuster AS, Hennig GW, Yan W. Expression Profiling Reveals Developmentally Regulated lncRNA Repertoire in the Mouse Male Germline1. *Biol Reprod*. 2013;89(5):1-12. doi:10.1095/biolreprod.113.113308
104. Meistrich ML. *Separation of Spermatogenic Cells and Nuclei from Rodent Testes.* Vol 15.; 1977.
105. Getun I V., Torres B, Bois PRJ. Flow Cytometry Purification of Mouse Meiotic Cells. *J Vis Exp*. 2011;50(50). doi:10.3791/2602
106. Geisinger A, Rodríguez-Casuriaga R. Flow cytometry for gene expression studies in mammalian spermatogenesis. *Cytogenet Genome Res*. 2010;128(1-3):46-56. doi:10.1159/000291489
107. Gaysinskaya V, Bortvin A. Flow cytometry of murine spermatocytes. *Curr Protoc Cytom*. 2015;72:7.44.1-7.44.24. doi:10.1002/0471142956.CY0744S72
108. Rodríguez-Casuriaga R, Santiñaque FF, Folle GA, Souza E, López-Carro B, Geisinger A. Rapid preparation of rodent testicular cell suspensions and spermatogenic stages purification by flow cytometry using a novel blue-laser-excitable vital dye. *MethodsX*. 2014;1:239-243. doi:10.1016/J.MEX.2014.10.002
109. Geisinger A, Rodríguez-Casuriaga R. Flow Cytometry for the Isolation and Characterization of Rodent Meiocytes. *Methods Mol Biol*. 2017;1471:217-230. doi:10.1007/978-1-4939-6340-9_11
110. Melé M, Ferreira PG, Reverter F, et al. Human genomics. The human transcriptome across tissues and individuals. *Science*. 2015;348(6235):660-665. doi:10.1126/SCIENCE.AAA0355
111. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347(6220). doi:10.1126/SCIENCE.1260419
112. Darbellay F, Necsulea A. Comparative Transcriptomics Analyses across Species, Organs, and Developmental Stages Reveal Functionally Constrained lncRNAs. *Mol Biol Evol*. 2020;37(1):240-259. doi:10.1093/MOLBEV/MSZ212
113. Hong SH, Kwon JT, Kim J, et al. Profiling of testis-specific long noncoding RNAs in mice. *BMC Genomics*. 2018;19(1). doi:10.1186/S12864-018-4931-3

114. Bortvin A. PIWI-interacting RNAs (piRNAs) - a mouse testis perspective. *Biochemistry (Mosc)*. 2013;78(6):592-602. doi:10.1134/S0006297913060059
115. de Mateo S, Sassone-Corsi P. Regulation of spermatogenesis by small non-coding RNAs: role of the germ granule. *Semin Cell Dev Biol*. 2014;29:84-92. doi:10.1016/J.SEMCDB.2014.04.021
116. He C, Wang K, Gao Y, et al. Roles of Noncoding RNA in Reproduction. *Front Genet*. 2021;12. doi:10.3389/FGENE.2021.777510
117. Hilz S, Modzelewski AJ, Cohen PE, Grimson A. The roles of microRNAs and siRNAs in mammalian spermatogenesis. *Development*. 2016;143(17):3061-3073. doi:10.1242/DEV.136721
118. Kotaja N. MicroRNAs and spermatogenesis. *Fertil Steril*. 2014;101(6):1552-1562. doi:10.1016/J.FERTNSTERT.2014.04.025
119. Yadav RP, Kotaja N. Small RNAs in spermatogenesis. *Mol Cell Endocrinol*. 2014;382(1):498-508. doi:10.1016/J.MCE.2013.04.015
120. Wrobel G, Primig M. Mammalian male germ cells are fertile ground for expression profiling of sexual reproduction. *Reproduction*. 2005;129(1):1-7. doi:10.1530/rep.1.00408
121. Naro C, Cesari E, Sette C. Splicing regulation in brain and testis: common themes for highly specialized organs. *Cell Cycle*. 2021;20(5-6):480-489. doi:10.1080/15384101.2021.1889187
122. Geisinger A. *Spermatogenesis in Mammals: A Very Peculiar Cell Differentiation Process*. Nova; 2008.
123. Kleene KC. Patterns, mechanisms, and functions of translation regulation in mammalian spermatogenic cells. *Cytogenet Genome Res*. 2003;103(3-4):217-224. doi:10.1159/000076807
124. Florea L, Song L, Salzberg SL. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Res*. 2013;2(0):188. doi:10.12688/f1000research.2-188.v2
125. Mazin P V., Khaitovich P, Cardoso-Moreira M, Kaessmann H. Alternative splicing during mammalian organ development. *Nat Genet*. 2021;53(6):925-934. doi:10.1038/S41588-021-00851-W
126. Kan Z, Garrett-Engele PW, Johnson JM, Castle JC. Evolutionarily conserved and diverged alternative splicing events show different expression and

functional profiles. *Nucleic Acids Res*. 2005;33(17):5659-5666. doi:10.1093/NAR/GKI834

127. Aliperti V, Skonieczna J, Cerase A. Long Non-Coding RNA (lncRNA) Roles in Cell Biology, Neurodevelopment and Neurological Disorders. *Non-Coding RNA 2021, Vol 7, Page 36*. 2021;7(2):36. doi:10.3390/NCRNA7020036
128. Sahlu BW, Zhao S, Wang X, et al. Long noncoding RNAs: New insights in modulating mammalian spermatogenesis. *J Anim Sci Biotechnol*. 2020;11(1):1-12. doi:10.1186/S40104-019-0424-8/TABLES/5
129. Liu KS, Li TP, Ton H, Mao XD, Chen YJ. Advances of Long Noncoding RNAs-mediated Regulation in Reproduction. *Chin Med J (Engl)*. 2018;131(2):226. doi:10.4103/0366-6999.222337
130. Legrand JMD, Hobbs RM. RNA processing in the male germline: Mechanisms and implications for fertility. *Semin Cell Dev Biol*. 2018;79:80-91. doi:10.1016/J.SEMCDB.2017.10.006
131. Grosso AR, Gomes AQ, Barbosa-Morais NL, et al. Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res*. 2008;36(15):4823-4832. doi:10.1093/NAR/GKN463
132. de la Grange P, Gratadou L, Delord M, Dutertre M, Auboeuf D. Splicing factor and exon profiling across human tissues. *Nucleic Acids Res*. 2010;38(9):2825-2838. doi:10.1093/NAR/GKQ008
133. Wu D, Khan FA, Huo L, Sun F, Huang C. Alternative splicing and MicroRNA: epigenetic mystique in male reproduction. *RNA Biol*. 2022;19(1):162-175. doi:10.1080/15476286.2021.2024033
134. Li Q, Li T, Xiao X, et al. Specific expression and alternative splicing of mouse genes during spermatogenesis. *Mol Omics*. 2020;16(3):258-267. doi:10.1039/C9MO00163H
135. Russell SJ, Stalker L, Gilchrist G, et al. Identification of PIWIL1 Isoforms and Their Expression in Bovine Testes, Oocytes, and Early Embryos 1. *Biol Reprod*. 2016;94(4):75-76. doi:10.1095/biolreprod.115.136721
136. Kleene KC. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mech Dev*. 2001;106(1-2):3-23. doi:10.1016/S0925-4773(01)00413-0

137. Bao J, Tang C, Li J, et al. RAN-binding protein 9 is involved in alternative splicing and is critical for male germ cell development and male fertility. *PLoS Genet*. 2014;10(12). doi:10.1371/JOURNAL.PGEN.1004825
138. Iwamori N, Tominaga K, Sato T, et al. MRG15 is required for pre-mRNA splicing and spermatogenesis. *Proc Natl Acad Sci U S A*. 2016;113(37):E5408-E5415. doi:10.1073/PNAS.1611995113
139. Hannigan MM, Zagore LL, Licatalosi DD. Ptbp2 Controls an Alternative Splicing Network Required for Cell Communication during Spermatogenesis. *Cell Rep*. 2017;19(12):2598-2612. doi:10.1016/J.CELREP.2017.05.089
140. He F, Jacobson A. Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. *Annu Rev Genet*. 2015;49:339. doi:10.1146/ANNUREV-GENET-112414-054639
141. Naro C, Jolly A, Di Persio S, et al. An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation. *Dev Cell*. 2017;41(1):82-93.e4. doi:10.1016/J.DEVCEL.2017.03.003
142. Paronetto MP, Sette C. Role of RNA-binding proteins in mammalian spermatogenesis. *Int J Androl*. 2010;33(1):2-12. doi:10.1111/j.1365-2605.2009.00959.x
143. Chalmel F, Lardenois A, Evrard B, et al. High-resolution profiling of novel transcribed regions during rat spermatogenesis. *Biol Reprod*. 2014;91(1). doi:10.1095/BIOLREPROD.114.118166
144. Zuo H, Zhang J, Zhang L, et al. Transcriptomic Variation during Spermiogenesis in Mouse Germ Cells. *PLoS One*. 2016;11(11). doi:10.1371/JOURNAL.PONE.0164874
145. Rolland AD, Evrard B, Darde TA, et al. RNA profiling of human testicular cells identifies syntenic lncRNAs associated with spermatogenesis. *Hum Reprod*. 2019;34(7):1278-1290. doi:10.1093/HUMREP/DEZ063
146. Schmid R, Grellscheid SN, Ehrmann I, et al. The splicing landscape is globally reprogrammed during male meiosis. *Nucleic Acids Res*. 2013;41(22):10170-10184. doi:10.1093/NAR/GKT811
147. Ilott NE, Ponting CP. Predicting long non-coding RNAs using RNA sequencing. *Methods*. 2013;63(1):50-59. doi:10.1016/j.ymeth.2013.03.019

148. Trovero MF, Rodríguez-Casuriaga R, Romeo C, et al. Revealing stage-specific expression patterns of long noncoding RNAs along mouse spermatogenesis. *RNA Biol*. 2020;17(3):350-365. doi:10.1080/15476286.2019.1700332
149. Lindgreen S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes*. 2012;5. doi:10.1186/1756-0500-5-337
150. Liu R, Dickerson J. Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-Seq. *PLoS Comput Biol*. 2017;13(11). doi:10.1371/JOURNAL.PCBI.1005851
151. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 2016;11(9):1650-1667. doi:10.1038/NPROT.2016.095
152. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930. doi:10.1093/BIOINFORMATICS/BTT656
153. Deluca DS, Levin JZ, Sivachenko A, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*. 2012;28(11):1530-1532. doi:10.1093/BIOINFORMATICS/BTS196
154. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-169. doi:10.1093/BIOINFORMATICS/BTU638
155. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/NAR/GKV007
156. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29. doi:10.1016/J.CELL.2021.04.048
157. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. *F1000Res*. 2020;9. doi:10.12688/F1000RESEARCH.23297.2/DOI
158. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res*. 2015;43(12). doi:10.1093/NAR/GKV227
159. Kang YJ, Yang DC, Kong L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res*. 2017;45(W1):W12-W16. doi:10.1093/NAR/GKX428

160. Yang C, Yang L, Zhou M, et al. LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*. 2018;34(22):3825-3834. doi:10.1093/BIOINFORMATICS/BTY428
161. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 2013;41(6). doi:10.1093/NAR/GKT006
162. Bryant DM, Johnson K, DiTommaso T, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Rep*. 2017;18(3):762-776. doi:10.1016/J.CELREP.2016.12.063
163. Wickham H. ggpolt2 Elegant Graphics for Data Analysis. *Use R! series*. Published online 2016:211.
164. Kassambara A. “ggplot2” Based Publication Ready Plots [R package ggpubr version 0.6.0]. Published online February 10, 2023. Accessed August 24, 2023. https://CRAN.R-project.org/package=ggpubr
165. Schindelin J, Arganda-Carreras I, Frise E, et al. Fiji: an open-source platform for biological-image analysis. *Nature Methods 2012 9:7*. 2012;9(7):676-682. doi:10.1038/nmeth.2019
166. Giansanti P, Samaras P, Bian Y, et al. Mass spectrometry-based draft of the mouse proteome. *Nature Methods 2022 19:7*. 2022;19(7):803-811. doi:10.1038/s41592-022-01526-y
167. Rodríguez-Casuriaga R, Folle GA, Santiñaque F, López-Carro B, Geisinger A. Simple and Efficient Technique for the Preparation of Testicular Cell Suspensions. *Journal of Visualized Experiments*. 2013;(78):1-7. doi:10.3791/50102
168. Rodríguez-Casuriaga R, Geisinger A, Santiñaque FF, López-Carro B, Folle GA. High-purity flow sorting of early meiocytes based on DNA analysis of guinea pig spermatogenic cells. *Cytometry Part A*. 2011;79 A(8):625-634. doi:10.1002/cyto.a.21067
169. Edelmann W, Cohen PE, Kneitz B, et al. Mammalian MutS homologue 5 is required for chromosome pairing in meiosis. *Nat Genet*. 1999;21(1):123-127. doi:10.1038/5075
170. Wu HY, Wei P, Morgan JI. Role of Cytosolic Carboxypeptidase 5 in Neuronal Survival and Spermatogenesis. *Sci Rep*. 2017;7. doi:10.1038/SREP41428

171. Giordano T, Gadadhar S, Bodakuntla S, et al. Loss of the deglutamylase CCP5 perturbs multiple steps of spermatogenesis and leads to male infertility. *J Cell Sci*. 2019;132(3). doi:10.1242/JCS.226951
172. Fonseca BD, Lahr RM, Damgaard CK, Alain T, Berman AJ. LARP1 on TOP of ribosome production. *Wiley Interdiscip Rev RNA*. 2018;9(5). doi:10.1002/WRNA.1480
173. Berman AJ, Thoreen CC, Dedeic Z, Chettle J, Roux PP, Blagden SP. Controversies around the function of LARP1. *RNA Biol*. 2021;18(2):207-217. doi:10.1080/15476286.2020.1733787
174. Uguen M, Liu T, James LI, Frye S V. Tudor-Containing Methyl-Lysine and Methyl-Arginine Reader Proteins: Disease Implications and Chemical Tool Development. *ACS Chem Biol*. 2025;20(1). doi:10.1021/ACSCHEMBIO.4C00661
175. Arkov AL, Wang JYS, Ramos A, Lehmann R. The role of Tudor domains in germline development and polar granule architecture. *Development*. 2006;133(20):4053-4062. doi:10.1242/DEV.02572
176. Bao J, Wang L, Lei J, et al. STK31(TDRD8) is dynamically regulated throughout mouse spermatogenesis and interacts with MIWI protein. *Histochem Cell Biol*. 2012;137(3):377-389. doi:10.1007/S00418-011-0897-9
177. Vourekas A, Zheng Q, Alexiou P, et al. Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. *Nature Structural & Molecular Biology 2012 19:8*. 2012;19(8):773-781. doi:10.1038/nsmb.2347
178. Fratta E, Coral S, Covre A, et al. The biology of cancer testis antigens: Putative function, regulation and therapeutic potential. *Mol Oncol*. 2011;5(2):164. doi:10.1016/J.MOLONC.2011.02.001
179. Zhong L, Liu J, Hu Y, et al. STK31 as novel biomarker of metastatic potential and tumorigenicity of colorectal cancer. *Oncotarget*. 2017;8(15):24354-24361. doi:10.18632/ONCOTARGET.15396
180. Bae DH, Kim HJ, Yoon BH, et al. STK31 upregulation is associated with chromatin remodeling in gastric cancer and induction of tumorigenicity in a xenograft mouse model. *Oncol Rep*. 2021;45(4). doi:10.3892/OR.2021.7993

181. Xiong J, Xing S, Dong Z, et al. STK31 regulates the proliferation and cell cycle of lung cancer cells via the Wnt/β-catenin pathway and feedback regulation by c-myc. *Oncol Rep*. 2020;43(2):395-404. doi:10.3892/OR.2019.7441
182. Dowdle JA, Mehta M, Kass EM, et al. Mouse BAZ1A (ACF1) is dispensable for double-strand break repair but is essential for averting improper gene expression during spermatogenesis. *PLoS Genet*. 2013;9(11). doi:10.1371/JOURNAL.PGEN.1003945
183. Yadav RP, Leskinen S, Ma L, Mäkelä JA, Kotaja N. Chromatin remodelers HELLS, WDHD1 and BAZ1A are dynamically expressed during mouse spermatogenesis. *Reproduction*. 2022;165(1):49-63. doi:10.1530/REP-22-0240
184. Moritz L, Hammoud SS. The Art of Packaging the Sperm Genome: Molecular and Structural Basis of the Histone-To-Protamine Exchange. *Front Endocrinol (Lausanne)*. 2022;13:895502. doi:10.3389/FENDO.2022.895502/BIBTEX
185. Tsuchida J, Nishina Y, Wakabayashi N, Nozaki M, Sakai Y, Nishimune Y. Molecular cloning and characterization of meichroacidin (male meiotic metaphase chromosome-associated acidic protein). *Dev Biol*. 1998;197(1):67-76. doi:10.1006/DBIO.1998.8885
186. Zheng W, Li F, Ding Z, et al. Distinct architecture and composition of mouse axonemal radial spoke head revealed by cryo-EM. *Proc Natl Acad Sci U S A*. 2021;118(4). doi:10.1073/PNAS.2021180118/-/DCSUPPLEMENTAL
187. Kott E, Legendre M, Copin B, et al. Loss-of-function mutations in RSPH1 cause primary ciliary dyskinesia with central-complex and radial-spoke defects. *Am J Hum Genet*. 2013;93(3):561-570. doi:10.1016/J.AJHG.2013.07.013
188. Milano CR, Kim Holloway J, Zhang Y, et al. Mutation of the ATPase Domain of MutS Homolog-5 (MSH5) Reveals a Requirement for a Functional MutSγ Complex for All Crossovers in Mammalian Meiosis. *G3 (Bethesda)*. 2019;9(6):1839-1850. doi:10.1534/G3.119.400074
189. Gill ME, Rohmer A, Erkek-Ozhan S, et al. De novo transcriptome assembly of mouse male germ cells reveals novel genes, stage-specific bidirectional promoter activity, and noncoding RNA expression. *Genome Res*. 2023;33(12):2060. doi:10.1101/GR.278060.123
190. Arlt MF, Brogley MA, Stark-Dykema ER, Hu YC, Mueller JL. Genomic Structure, Evolutionary Origins, and Reproductive Function of a Large Amplified

Intrinsically Disordered Protein-Coding Gene on the X Chromosome (Laidx) in Mice. Published online 2020. doi:10.1534/g3.120.401221

191. Liu W, Wang F, Xu Q, et al. BCAS2 is involved in alternative mRNA splicing in spermatogonia and the transition to meiosis. *Nat Commun*. 2017;8. doi:10.1038/NCOMMS14182
192. MONESI V. RIBONUCLEIC ACID SYNTHESIS DURING MITOSIS AND MEIOSIS IN THE MOUSE TESTIS. *J Cell Biol*. 1964;22(3):521-532. doi:10.1083/JCB.22.3.521
193. Kierszenbaum AL, Tres LL. Nucleolar and perichromosomal RNA synthesis during meiotic prophase in the mouse testis. *J Cell Biol*. 1974;60(1):39-53. doi:10.1083/JCB.60.1.39
194. Page J, De La Fuente R, Manterola M, et al. Inactivation or non-reactivation: what accounts better for the silence of sex chromosomes during mammalian male meiosis? *Chromosoma*. 2012;121(3):307-326. doi:10.1007/S00412-012-0364-Y
195. Green CD, Ma Q, Manske GL, et al. A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq. *Dev Cell*. 2018;46(5):651-667.e10. doi:10.1016/J.DEVCEL.2018.07.025
196. Tilgner H, Knowles DG, Johnson R, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res*. 2012;22(9):1616-1625. doi:10.1101/GR.134445.111
197. Licatalosi DD. Roles of RNA-binding Proteins and Post-transcriptional Regulation in Driving Male Germ Cell Development in the Mouse. *Adv Exp Med Biol*. 2016;907:123-151. doi:10.1007/978-3-319-29073-7_6
198. Anguera MC, Ma W, Clift D, Namekawa S, Kelleher RJ, Lee JT. Tsx produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. *PLoS Genet*. 2011;7(9). doi:10.1371/JOURNAL.PGEN.1002248
199. Joshi M, Rajender S. Long non-coding RNAs (lncRNAs) in spermatogenesis and male infertility. *Reprod Biol Endocrinol*. 2020;18(1). doi:10.1186/S12958-020-00660-6
200. Kataruka S, Akhade VS, Kayyar B, Rao MRS. Mrhl Long Noncoding RNA Mediates Meiotic Commitment of Mouse Spermatogonial Cells by Regulating Sox8 Expression. *Mol Cell Biol*. 2017;37(14). doi:10.1128/MCB.00632-16

201. Li K, Xu J, Luo Y, et al. Panoramic transcriptome analysis and functional screening of long noncoding RNAs in mouse spermatogenesis. *Genome Res*. 2021;31(1):13-26. doi:10.1101/GR.264333.120/-/DC1
202. Li L, Wang M, Wang M, et al. A long non-coding RNA interacts with Gfra1 and maintains survival of mouse spermatogonial stem cells. *Cell Death Dis*. 2016;7(3). doi:10.1038/CDDIS.2016.24
203. Li W, Ning JZ, Cheng F, et al. MALAT1 Promotes Cell Apoptosis and Suppresses Cell Proliferation in Testicular Ischemia-Reperfusion Injury by Sponging MiR-214 to Modulate TRPV4 Expression. *Cell Physiol Biochem*. 2018;46(2):802-814. doi:10.1159/000488738
204. Liu W, Zhao Y, Liu X, et al. A Novel Meiosis-Related lncRNA, Rbakdn, Contributes to Spermatogenesis by Stabilizing Ptbp2. *Front Genet*. 2021;12. doi:10.3389/FGENE.2021.752495
205. Lü M, Tian H, Cao YX, et al. Downregulation of miR-320a/383-sponge-like long non-coding RNA NLC1-C (narcolepsy candidate-region 1 genes) is associated with male infertility and promotes testicular embryonal carcinoma cell proliferation. *Cell Death Dis*. 2015;6(11). doi:10.1038/CDDIS.2015.267
206. Nakajima R, Sato T, Ogawa T, Okano H, Noce T. A noncoding RNA containing a SINE-B1 motif associates with meiotic metaphase chromatin and has an indispensable function during spermatogenesis. *PLoS One*. 2017;12(6). doi:10.1371/JOURNAL.PONE.0179585
207. Ni MJ, Hu ZH, Liu Q, et al. Identification and characterization of a novel non-coding RNA involved in sperm maturation. *PLoS One*. 2011;6(10). doi:10.1371/JOURNAL.PONE.0026053
208. Zhu Q, Sun J, An C, et al. Mechanism of LncRNA Gm2044 in germ cell development. Published online 2024. doi:10.3389/fcell.2024.1410914
209. Harfe BD, Jinks-Robertson S. DNA mismatch repair and genetic instability. *Annu Rev Genet*. 2000;34:359-399. doi:10.1146/ANNUREV.GENET.34.1.359
210. Miller RM, Smith LM. Overview and Considerations in Bottom-Up Proteomics. *Analyst*. 2023;148(3):475. doi:10.1039/D2AN01246D
211. Lahaye F, Lespinasse F, Staccini P, Palin L, Paquis-Flucklinger V, Santucci-Darmanin S. hMSH5 is a nucleocytoplasmic shuttling protein whose stability depends on its subcellular localization. *Nucleic Acids Res*. 2010;38(11):3655-3671. doi:10.1093/NAR/GKQ098

212. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics 2012 13:5*. 2012;13(5):329-342. doi:10.1038/nrg3174
213. Climente-González H, Porta-Pardo E, Godzik A, Eyras E. The Functional Impact of Alternative Splicing in Cancer. *Cell Rep*. 2017;20(9):2215-2226. doi:10.1016/J.CELREP.2017.08.012/ATTACHMENT/43258BFD-26EF-4B74-81A1-E5745A37F5E2/MMC8.PDF
214. Wang S, Cai Y, Li T, et al. CWF19L2 is Essential for Male Fertility and Spermatogenesis by Regulating Alternative Splicing. *Advanced Science*. 2024;11(31):2403866. doi:10.1002/ADVS.202403866
215. Tress ML, Abascal F, Valencia A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem Sci*. 2017;42(2):98-110. doi:10.1016/J.TIBS.2016.08.008
216. Schneider VA, Graves-Lindsay T, Howe K, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*. 2017;27(5):849-864. doi:10.1101/GR.213611.116
217. Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol*. 2008;4(11). doi:10.1371/JOURNAL.PCBI.1000176
218. Tollis M, Ferris E, Campbell MS, et al. Elephant Genomes Reveal Accelerated Evolution in Mechanisms Underlying Disease Defenses. *Mol Biol Evol*. 2021;38(9):3606-3620. doi:10.1093/MOLBEV/MSAB127

BMC Genomics

RESEARCH

Open Access

Uncovering a multitude of stage-specific splice variants and putative protein isoforms generated along mouse spermatogenesis

Carlos Romeo-Cardeillac[1,2], María Fernanda Trovero[1,5], Santiago Radío[2], Pablo Smircich[2], Rosana Rodríguez-Casuriaga[1], Adriana Geisinger[1,3*†] and José Sotelo-Silveira[2,4*†]

Abstract

Background Mammalian testis is a highly complex and heterogeneous tissue. This complexity, which mostly derives from spermatogenic cells, is reflected at the transcriptional level, with the largest number of tissue-specific genes and long noncoding RNAs (lncRNAs) compared to other tissues, and one of the highest rates of alternative splicing. Although it is known that adequate alternative-splicing patterns and stage-specific isoforms are critical for successful spermatogenesis, so far only a very limited number of reports have addressed a detailed study of alternative splicing and isoforms along the different spermatogenic stages.

Results In the present work, using highly purified stage-specific testicular cell populations, we detected 33,002 transcripts expressed throughout mouse spermatogenesis not annotated so far. These include both splice variants of already annotated genes, and of hitherto unannotated genes. Using conservative criteria, we uncovered 13,471 spermatogenic lncRNAs, which reflects the still incomplete annotation of lncRNAs. A distinctive feature of lncRNAs was their lower number of splice variants compared to protein-coding ones, adding to the conclusion that lncRNAs are, in general, less complex than mRNAs. Besides, we identified 2,794 unannotated transcripts with high coding potential (including some arising from yet unannotated genes), many of which encode unnoticed putative testis-specific proteins. Some of the most interesting coding splice variants were chosen, and validated through RT-PCR. Remarkably, the largest number of stage-specific unannotated transcripts are expressed during early meiotic prophase stages, whose study has been scarcely addressed in former transcriptomic analyses.

Conclusions We detected a high number of yet unannotated genes and alternatively spliced transcripts along mouse spermatogenesis, hence showing that the transcriptomic diversity of the testis is considerably higher than previously reported. This is especially prominent for specific, underrepresented stages such as those of early meiotic prophase, and its unveiling may constitute a step towards the understanding of their key events.

†Adriana Geisinger and José Sotelo-Silveira contributed equally to this work.

*Correspondence:
Adriana Geisinger
adriana.geisinger@gmail.com; ageisinger@fcien.edu.uy
José Sotelo-Silveira
jsotelosilveira@iibce.edu.uy; sotelojos@gmail.com

Full list of author information is available at the end of the article

Keywords Spermatogenesis, Transcriptome, Alternative splicing, lncRNAs, Testis

Background

Spermatogenesis can be defined as the execution of three consecutive yet overlapping processes that take place in the male gonad, inside the seminiferous tubules. The first process is the mitotic proliferation and differentiation of spermatogonia (meiotic precursor cells), which go through different stages until they become primary spermatocytes and enter meiosis. The second phase is the meiotic divisions, during which spermatocytes (I and II, corresponding to the first and second meiotic divisions, respectively) halve their DNA content, resulting in haploid spermatids. Recombination of homologous chromosomes, which involves a meiotic-specific protein structure, the synaptonemal complex, is a hallmark of meiotic prophase I. The third phase, spermiogenesis, is the differentiation of round spermatids (i.e. the outcome of meiosis II) into sperm (Fig. 1). Along the latter, spermatids undergo dramatic changes, namely: the acquisition of a flagellum; nuclear elongation; loss of most cytoplasm; acrosome formation; reorganization of mitochondria; and the sequential replacement of most histones first by transition proteins and then by protamines, with the consequence of chromatin compaction and massive transcriptional silencing during late spermiogenic stages [1, 2].

Besides germline cells at their various differentiation stages, different somatic cell types coexist within the mammalian testes: Sertoli cells, which support and nourish the germline cells inside the seminiferous tubules; peritubular myoid cells; and different types of interstitial cells, including testosterone-producing Leydig cells, fibroblasts, macrophages, endothelial cells, innate lymphoid cells, and mesenchymal cells [3]. In total, the testis is composed of over 30 different cell types, which makes it an extremely complex and heterogeneous tissue.

Testicular tissue and cell complexity are also reflected at the transcriptional level. It has been shown that, in different mammalian species, the testes exhibit the highest transcriptomic complexity and diversity compared to other tissues, expressing the largest number of tissue-specific genes [4–6] and an overwhelming majority of long noncoding RNAs (lncRNAs) [6–11], as well as a panoply of short noncoding RNAs (piRNAs, miRNAs) [12–17]. Moreover, together with the brain, the testes have been reported to present the highest rate of alternative splicing (AS) [6, 18–21], which generates a huge number of testis-specific, temporally regulated RNA isoforms and protein variants [22, 23]. In accordance with this, the testis expresses a very large number of specific and strictly-regulated RNA-binding proteins [24–26], including many unique or differentially expressed (DE) splicing factors [20, 22, 27–29]. Furthermore, splicing defects have been associated with testicular pathologies [20, 22, 23, 29–32]. Interestingly, the complexity of the testicular transcriptome has been reported to mostly derive from primary spermatocytes and, particularly, round spermatids [6].

A number of studies have analyzed testicular transcriptomic diversity along spermatogenesis progression, and some of them included the identification and/or preliminary characterization of AS in mouse [6, 33–38], rat [39], and human [40]. However, only a very limited number of studies have addressed a more detailed analysis [34, 35, 38]. Moreover, they were mostly based on computational deconvolution approaches [35] or available data sets [38].

We have previously profiled the transcriptomic fluctuations along mouse spermatogenesis, both for coding transcripts [41] and for lncRNAs [42]. The input was highly purified stage-specific spermatogenic cell populations by flow-cytometry [43–45], thus constituting a solid basis for generating highly reliable information. Of particular interest, our analyses included purified early

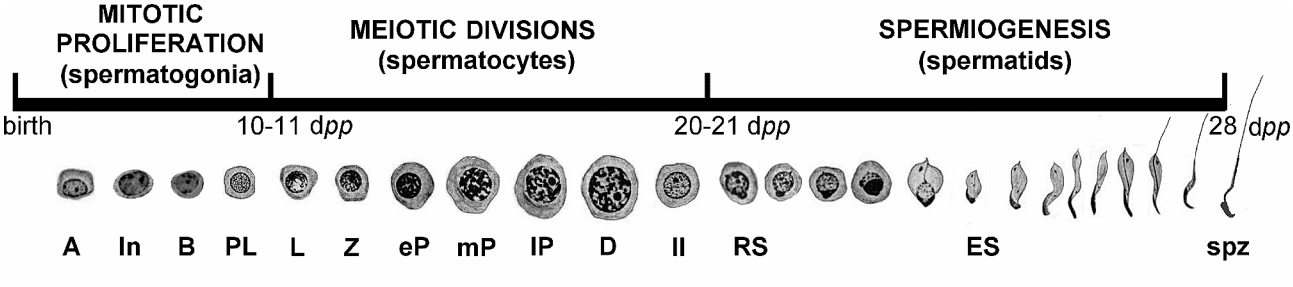


Fig. 1 Schematic representation of the timing of spermatogenesis in the mouse. The three main phases of the process are shown. Emblematic stages are graphically represented under the timeline, and their postnatal timing of appearance is expressed as days *postpartum* (d*pp*). Cell types represented correspond to type A, intermediate (In), type B and preleptotenic (PL) spermatogonia; leptotenic (L), zygotenic (Z), early (eP), medium (mP) and late (lP) pachytenic primary spermatocytes, and diplotenic ones (D); secondary spermatocytes (II); round (RS) and elongating (ES) spermatids, and spermatozoa (spz). Adapted from reference 94, under the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/)

meiotic prophase cell populations, which have been very rarely included in transcriptomic studies. Here, we used our highly confident data to provide a comprehensive study of the transcriptomic diversity along spermatogenesis. Due to the purity of the populations, added to the depth of the libraries, we have been able to detect genes and isoforms that are lowly expressed and/or specific to scarce cell types, which had not been detected so far.

Overall, our results identify a high number of unannotated transcripts and splice variants, both coding and noncoding, which helps contribute to the understanding of testis complexity and functionality. This is particularly conspicuous for short, poorly studied stages, such as early meiotic prophase. Therefore, even for a genome as well characterized as that of the mouse, when it comes to specific stages of spermatogenesis, there is still much transcriptomic diversity to be described, including undisclosed stage-specific protein isoforms.

Results

In previous reports, we have profiled the protein-coding and lncRNAs transcriptomes along mouse spermatogenesis, using isolated cell populations at different spermatogenic stages, and including a highly pure leptotene-zygotene (LZ) fraction [41, 42]. The latter allowed us to analyze early meiotic prophase, which is a scarce, short-lived stage, and therefore had been very poorly characterized at the molecular level. However, in those analyses we only studied annotated genes. Moreover, expressed genes were accounted for, but not splice variants. Here, we used our highly reliable raw data to identify unannotated expressed genes, stage-specific RNA species and unreported putative proteins, as well as to analyze AS and its variations along spermatogenesis in order to have a more complete idea of its real extension (see complete pipeline in Fig. 2).

A correlation matrix showed high reproducibility between biological replicas (Supplementary Figure S1). Besides, contrasting our data with those from another study, namely a single-cell RNA sequencing (scRNA-seq) of 20 different spermatogenic cell subtypes [37], rendered a good correlation despite the different methodologies employed in both studies (Supplementary Figure S2). Overall, our data is a very deep set of reads with robust reproducibility, and therefore it is useful to characterize even lowly expressed transcripts.

Identification of unannotated coding and noncoding transcripts

We applied strict cut-offs for downstream analyses (e.g. 10X coverage as minimum; 10 reads as minimum support per splice site; 10 reads as minimum per exon support). Under the selected conditions, we identified 37,793 testis-expressed genes that passed all the filters, of which 21,156 (56%) were already annotated in databases, and 16,637 (44%) were unannotated genes (Fig. 3A, and Supplementary Figures S3A and S3B). These 37,793 genes gave rise to 81,139 different transcripts (Supplementary Table S1). Of these transcripts, 48,137 (59%) were already annotated, while 33,002 (41%) were unreported transcripts (Fig. 3B, and Supplementary Figures S3A and S3C).

We then focused on the characterization of the 33,002 unannotated transcripts. Among them, we identified 14,667 (44%) as undisclosed splice variants of already known genes, while 18,335 (56%) corresponded to transcripts arising from regions of the genome for which there are no annotated transcripts (Fig. 3C, and Supplementary Figures S3A and S3D). This shows that there is still a very high number of testis-expressed genes and splice variants to be unveiled.

Next, we analyzed the coding potential of the unannotated transcripts. For this purpose, we used four different software programs and only kept the results found in common among them (i.e., those transcripts for which all four programs coincide that they are, or are not, coding). The coincidence of the four programs identified 13,471 transcripts as noncoding (Fig. 4A), and 2,794 as coding (Fig. 4B). Therefore, most of the "novel" transcripts are noncoding. This is as expected since the coding genome has been much more characterized than the noncoding one. We note that our established criterion, which is very restrictive, excluded over half of the transcripts (e.g. if only three of the four programs coincided), but in turn allowed us to keep working with a highly reliable subset of transcripts in terms of their high or low coding potential.

Of the unannotated noncoding transcripts, the vast majority (12,297 transcripts, i.e. 91%) corresponded to unreported genes, namely lncRNAs, while less than 10% (1,174 transcripts) were undisclosed splice variants of already annotated noncoding genes (Fig. 4C).

Concerning the unannotated spermatogenic transcripts with highly reliable protein-coding potential, their identified number (2,794) is not negligible at all. Contrarily to the noncoding transcripts, the majority among them (2,571) corresponded to novel splice variants of already annotated genes, while less than 10% of the "novel" putatively coding transcripts, namely 223, were transcripts of unannotated genes (Fig. 4D, and Supplementary Table S2). Surprisingly, these 223 transcripts come from 191 yet unannotated genes. These results indicate that in the mouse genome there is still a significant number of putative protein-coding genes and AS coding isoforms that are expressed in spermatogenic cells, which have remained undetected so far.

We then focused on these 191 unannotated genes with high coding potential, and conducted functional analysis

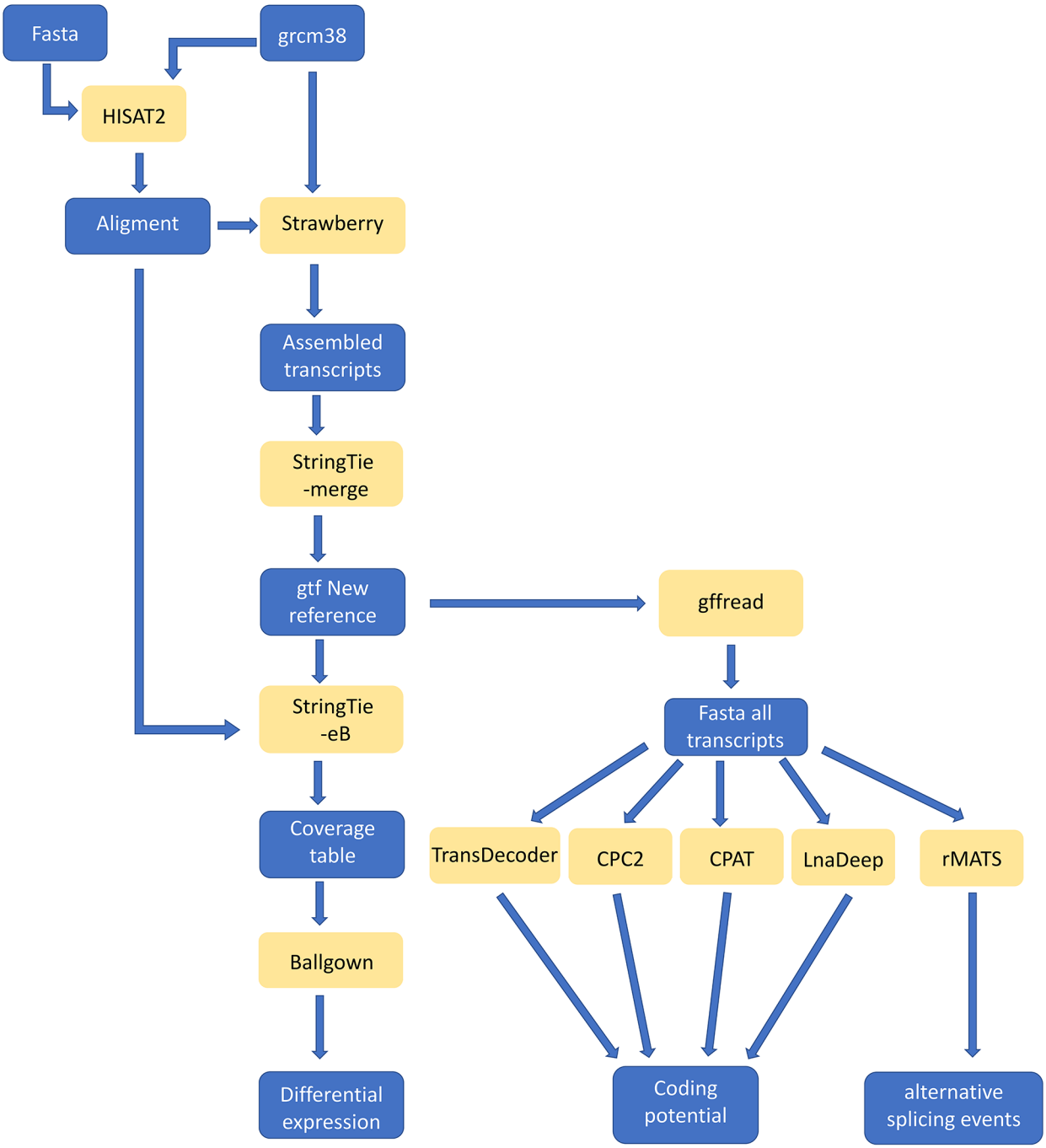


Fig. 2 Flow chart of the followed bioinformatics pipeline. The data files are represented in blue, while the different employed software is represented in yellow

based on similarities with annotated genes from mouse and other species. Some of the similarities for encoded putative proteins were with ribosomal proteins, zinc finger proteins, and with putative Cilia and Flagella-Associated Protein 92 (MSTRG.30402.2; BlastX match to human FLJ43738). Besides, a relatively large subset (over half of the genes) corresponded to the products of retroposons and, to a lesser extent, integrated viruses. Finally, for 29 of these 191 genes (corresponding to 49 transcripts), no known probable function was associated (see Supplementary Table S2).

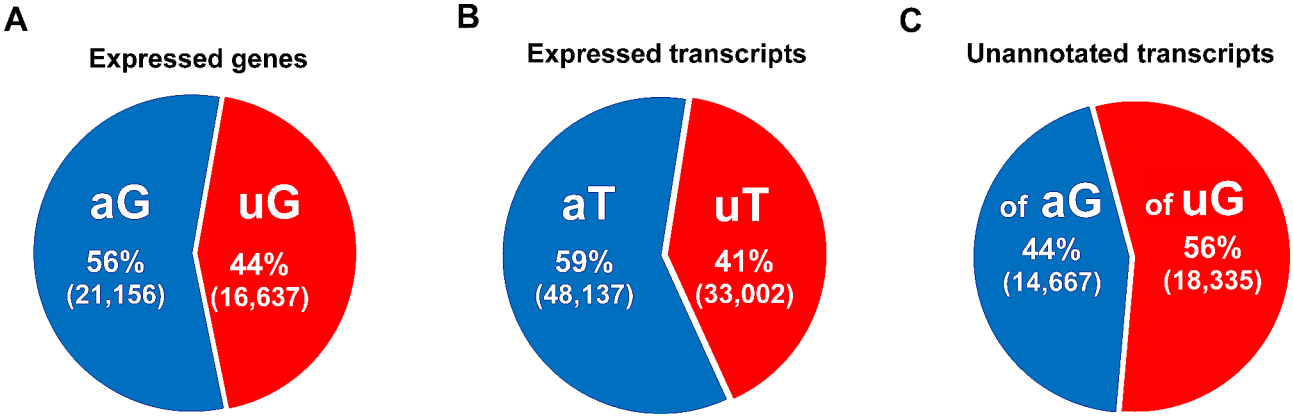


Fig. 3 Genes and transcripts expressed in our lists. (**A**) Pie chart of annotated genes (aG: blue) and unannotated genes (uG: red) expressed in the four spermatogenic cell populations that passed all the filters. (**B**) Pie chart of annotated transcripts (aT: blue) and unannotated transcripts (uT: red) expressed in the four spermatogenic cell populations. (**C**) Pie chart showing the origin of the unannotated transcripts in our lists, either undisclosed splice variants of already annotated genes (of aG: blue), or transcripts arising from unannotated genes (of uG: red)

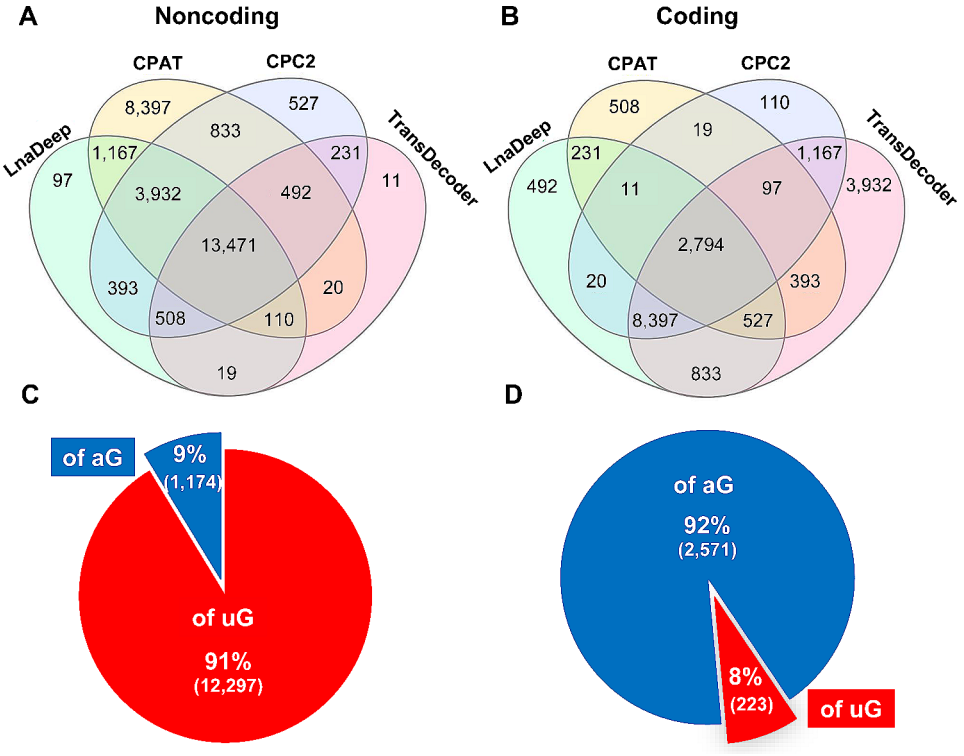


Fig. 4 Coding potential of the unannotated transcripts. (**A**, **B**) Venn diagrams showing the analysis of the coding potential for the unannotated transcripts through four different software programs. (**C**, **D**) Pie charts of the unannotated noncoding and putative protein-coding transcripts that were coincidentally identified as such with the four programs, and classified into undisclosed splice variants of already annotated genes (of aG: blue) and transcripts of unannotated genes (of uG: red). **A**, **C**: noncoding transcripts; **B**, **D**: coding transcripts

Expression of the newly identified transcripts along the different spermatogenic stages

As a next step, we analyzed the expression of the newly identified transcripts distributed by cell population. In the first place, we compared the expression levels of the unannotated transcripts with those previously annotated in Ensembl database, for each of the four cell populations. Overall, the median expression levels of the unannotated transcripts - both coding and noncoding - were lower than those of the annotated ones, and this turned out to be valid for all the cell populations (Fig. 5A). This may help explain why these transcripts had not been detected so far. Additionally, noncoding transcripts showed lower expression levels than coding ones for all cell populations (see Fig. 5A), which is in agreement with previous reports

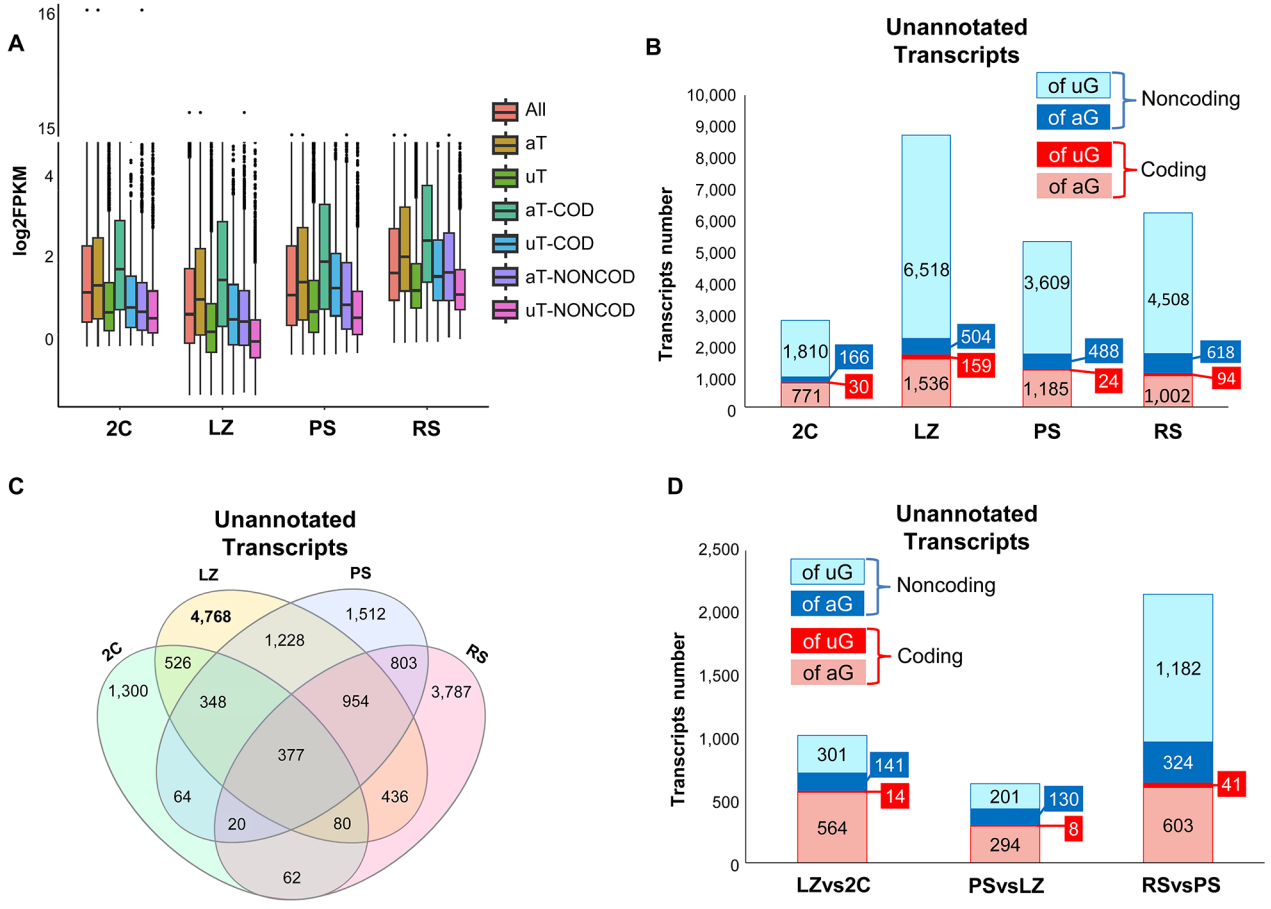


Fig. 5 Distribution of the transcripts in the four testicular cell populations. (**A**) Box plot of expression levels (log2FPKM) of all detected transcripts. All: all detected transcripts; aT: annotated transcripts; uT: unannotated transcripts; aT-COD: annotated coding transcripts; uT-COD: unannotated coding transcripts; aT-NONCOD: annotated noncoding transcripts; uT-NONCOD: unannotated noncoding transcripts. (**B**) Unannotated transcripts that were coincidentally identified as such with the four programs for coding potential analysis (and depicted in Fig. 4), distributed according to their expression in each of the four testicular cell populations. Transcripts are categorized into coding or noncoding, and transcripts of unannotated genes (of uG) or splice variants of already annotated genes (of aG). Note that many transcripts may be expressed in more than one stage. (**C**) Venn diagram indicating the distribution of the transcripts represented in B, in the four testicular cell populations. (**D**) DE coding and noncoding transcripts between pairwise sample comparisons of the four populations in chronological order. Of uG and of aG denote the same as in B

that indicate that the expression levels of lncRNAs are, in general, lower than those of mRNAs [9, 11].

Concerning the number of transcripts in each of the cell populations, interestingly, the largest contribution of the unannotated transcripts was on behalf of the LZ stage, both for noncoding and for coding transcripts (Fig. 5B). Furthermore, 55% of the unannotated LZ-expressed transcripts were exclusive of LZ (Fig. 5C, and Supplementary Figure S4A; see also Supplementary Figure S4B). In this regard, when we particularly looked at the unannotated coding genes, strikingly, 159 out of the 191 identified were expressed in LZ, and almost half (92 genes) were exclusive of LZ (see Supplementary Table S2). This led us to ask whether this would be the reflect of a greater number of transcripts expressed in LZ in general. Indeed, when we compared the total number of transcripts (both annotated and unannotated together) between the four testicular cell populations, LZ presented the highest number (Supplementary Figure S4C). Transcript saturation analysis including the data from the present study as well as from a previous one [41] showed that all the cell populations reached saturation (Supplementary Figure S5A). Moreover, the transcript expression histograms among all the four populations presented a similar distribution (Supplementary Figure S5B), thus helping validate the results. Altogether, these analyses confirm that the results are not an artifact of either the technique or the conducted analysis.

On the other hand, LZ was, in general, the stage with the lowest expression levels for all types of transcripts (both coding and noncoding, either annotated or not), while round spermatids (RS) transcripts exhibited the

highest overall expression levels (see Fig. 5A). Thus, LZ expresses the largest number of stage-specific transcripts, although these are, in general, expressed at comparatively lower levels.

Next, we analyzed the differential expression of the newly identified transcripts among pairwise comparisons along the progress of the spermatogenic wave (log2 FC $\geq |2|$, FDR < 0.05). We observed the highest number of DE unannotated transcripts that passed our established criteria at the pachytene spermatocytes (PS) - to - RS transition (Fig. 5D), and this is especially so for noncoding transcripts. This indicates that the transition from meiotic prophase to spermiogenesis involves the differential expression of a high number of genes and splice variants. Besides, this result is also reflecting the fact that, although as stated above, LZ expresses the highest number of exclusive unannotated transcripts, many of them are expressed at low levels, and therefore they do not pass our strict criterion for the definition of differential expression.

Characterization of spermatogenic-specific AS

We proceeded to further characterize the identified splice variants in our lists (both annotated and unannotated). In first place, we analyzed the different AS types in the four testicular cell populations by means of rMATS, and using the different AS categories defined by the software, i.e.: skipping exon (SE), alternative 5´ splice site (A5SS), alternative 3´ splice site (A3SS), mutually exclusive exons (MXE), and retained intron (RI). There were no significant enrichments when comparing AS events among the analyzed stages. On the other hand, SE and RI were the most abundant AS types, followed by A3SS, A5SS and MXE, in that order, in the four testicular cell populations (Fig. 6A).

Then, we studied the number of splice variants per gene. Clearly, we found that most splicing isoforms are generated by coding genes. Contrarily, noncoding genes, in general, have a lower number of transcripts per gene (Fig. 6B). Indeed, while about 60% of the coding genes express only one transcript, between 85% and 95% of the noncoding genes present only one transcript ($p < 10^{-10}$). Likewise, the number of genes with two or more splicing isoforms was higher for coding than for noncoding genes ($p < 0{,}01$). Basically, both annotated and unannotated genes behaved similarly in this regard, and this statement is valid both for coding and for noncoding transcripts (see Fig. 6B). This shows the reliability of our data, as there is no reason to suspect that the annotated and unannotated transcripts should behave differently. Furthermore, we note that although we are here showing the grouped results of the four testicular cell populations, there were no significant differences when the four populations were analyzed separately (not shown).

In-depth analysis of representative newly identified putative protein-coding splice variants

As stated above, the expression levels of the unannotated transcripts were, in general, lower than those of annotated ones (see Fig. 5A). Notwithstanding this, it is worth mentioning that some of the newly identified transcripts presented very high expression levels, with some AS isoforms being much more highly expressed than the already annotated ones (see Supplementary Table S1, and examples mentioned below).

We chose seven examples of these unannotated splice variants to confirm the discovery through RT-PCR (Fig. 7A), using the following criteria: (i) Annotated coding genes that would have a high number of expressed splice variants in our lists; (ii) That at least one of the splice variants would code for an unannotated putative protein isoform; (iii) That the putative novel protein isoform would be significantly different (e.g. with different protein domains) from the annotated one(s); (iv) That the putative novel isoform would exhibit a relatively high expression level in at least one of the analyzed

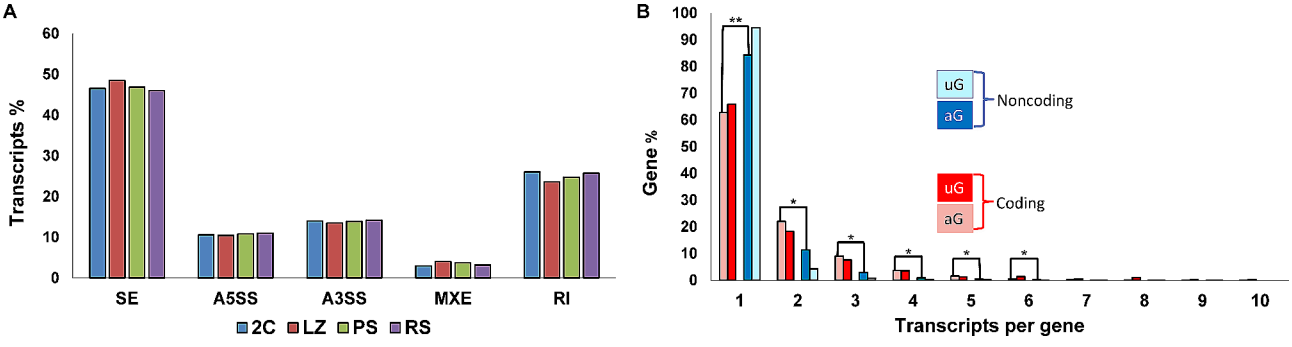


Fig. 6 Analysis of spermatogenic-specific alternative splicing (AS). (**A**) Bar graph representing the distribution of different AS types (percentage) along the four testicular cell populations. SE: skipping exon; A5SS: alternative 5´ splice site; A3SS: alternative 3´ splice site; MXE: mutually exclusive exons; RI: retained intron. (**B**) Classification of the expressed genes (coding and noncoding), according to their number of splice variants in our lists. The data are presented as percentage of the total. Only genes with 1 to 10 expressed splice variants were considered. uG: unannotated genes; aG: annotated genes. ** $p < 10^{-10}$; * $p < 0.01$

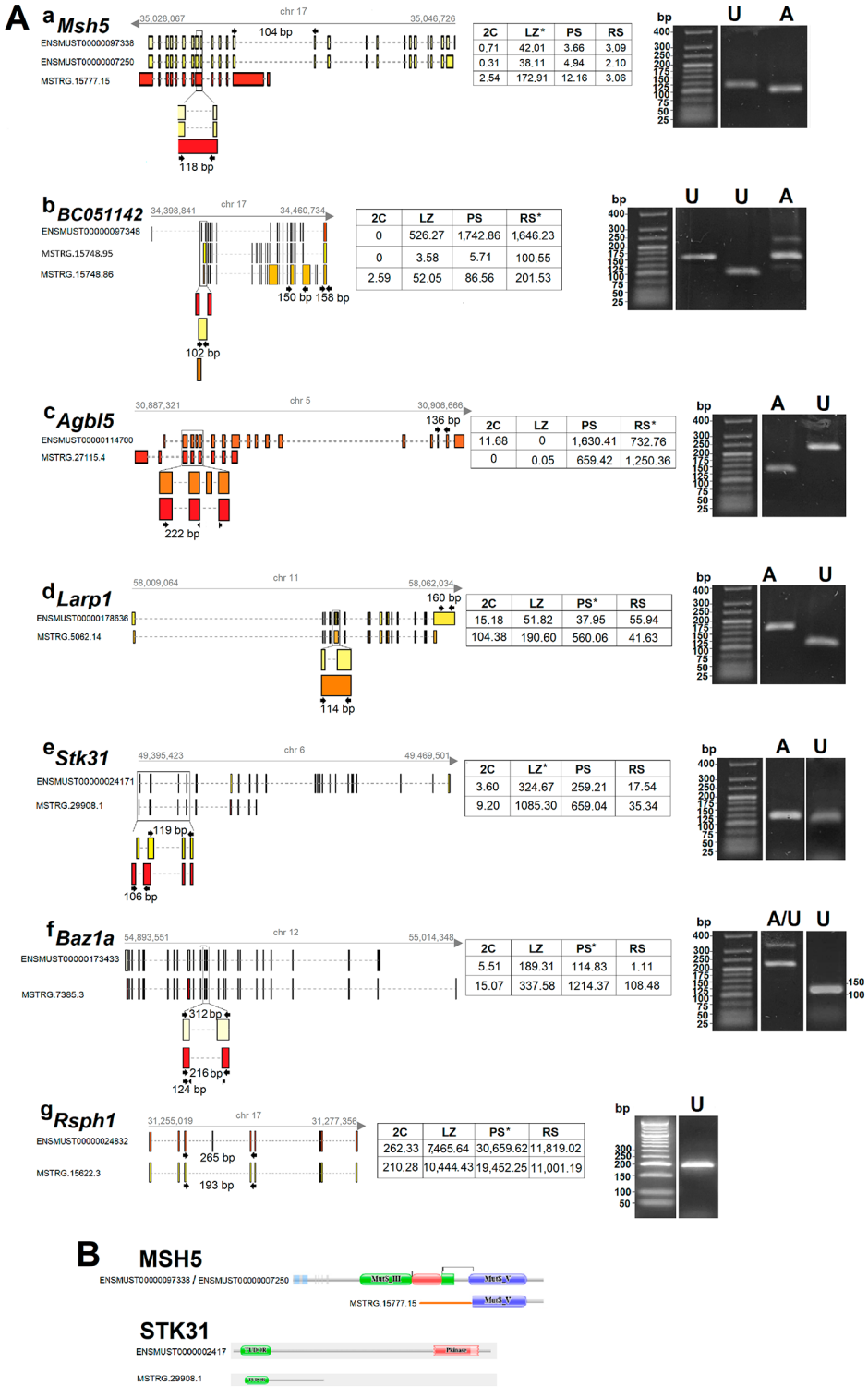


Fig. 7 (See legend on next page.)

spermatogenic stages; and (v) That the annotated gene would have an interesting described function (e.g. testis-related), or would present a specific trait that we considered particularly interesting.

One of the selected genes was *MutS homologue 5* (*Msh5*), which is upregulated in LZ, and in mouse directs the synthesis of an 833 amino acids protein (Fig. 7A,a). We have identified fifteen unannotated splice variants

(See figure on previous page.)

Fig. 7 RT-PCR confirms the expression of different examples of selected putatively protein-coding splice variants. (**A**) Schematic representations of the splice variants (annotated and newly identified), and agarose gels showing their RT-PCR amplification. Ensembl annotations are depicted on the left as "ENSMUST" followed by the corresponding Ensembl numberings. Unannotated transcript isoforms are depicted with the label "MSTRG". The designed primer sets for the amplification of either the annotated or the unannotated isoforms are shown for each case (arrows), together with the expected PCR product sizes (bp). The gray arrow above each diagram indicates transcription direction. Genomic location, as well as chromosome number, are indicated in each case. Whenever necessary, magnified insets are shown below each representation for better visualization of the amplified regions. A table with the coverage of the annotated and unannotated transcripts in the four cell populations is included in each case. The asterisks indicate the cell population in which the unannotated isoform was most highly expressed. In the agarose gels, A stands for the annotated splice variants, and U for the unannotated ones. Gels have been cropped for the sake of clarity (original agarose gels are presented as Supplementary Figure S7). (**a**) *Msh5* splice variants that encode the canonical 833 amino acids protein (yellow), and an unannotated splice variant encoding a putative 362 residues isoform (red). (**b**) *BC051142* most highly expressed annotated variant (red), and two unannotated putatively coding variants with a differential expression pattern along spermatogenesis (one is mostly differential of spermiogenesis, while the other progressively increases from early meiotic prophase to spermiogenesis; yellow and orange, respectively). In the lane corresponding to the annotated variant, two additional faint bands can be observed, most probably corresponding to the amplification of a couple of weakly expressed isoforms (due to the extremely high number of isoforms detected for this gene, it was not possible to design primer combinations to exclusively recognize only one variant). (**c**) *Agbl5* canonical transcript (orange), which encodes an 846 residues protein, and a selected unannotated variant (red) encoding a putative much shorter isoform of 412 amino acids. (**d**) The chosen *Larp1* unannotated splice variant (orange) encodes a putative not reported protein isoform of 760 amino acids, unlike the canonical one (yellow), whose encoded protein is 1,072 residues long. As shown, in PS the expression levels of the new variant are fifteen-fold higher than those of the canonical one. (**e**) The unannotated *Stk31* isoform we chose for confirmation (red) encodes a shorter variant that is much more highly expressed than the canonical one (yellow). The comparatively poorer amplification of the unannotated variant is due to the fact that the region did not allow the design of a good pair of primers. (**f**) Representation of an annotated *Baz1a* isoform (light yellow), and the unannotated splice variant (red), which is much more highly expressed all along spermatogenesis, upregulates in PS, and directs the synthesis of a shorter protein. In this case, amplification was performed with a primer set that simultaneously amplifies a region of both the annotated (312 bp) and unannotated (216 bp) variants. The annotated isoform is poorly amplified, presumptively due to its competition with the newly identified one, which is expressed at much higher levels (see table). Besides, a band corresponding to the amplification product with a primer set that only recognizes the unannotated isoform is shown to the right. (**g**) *Rsph1* was chosen as an example of a novel coding isoform generated through exon-skipping (yellow, while the canonical isoform is represented in red). Although the primer set was intended to amplify the annotated variant as well, yielding a larger, 265 bp band, the latter was not detected most probably because of its competition with the unannotated isoform. **B)** Representative schematic diagrams of two of the canonical and unannotated putative protein isoforms, to exemplify the differences between them. MSH5: The orange line in the "novel" isoform represents the first 133 amino acids, which are completely different from those of the canonical protein. STK31: While both isoforms present a Tudor domain, the predicted variant would lack the protein-kinase domain, which is essential for its function as a serin-threonine kinase in the canonical isoform

for this gene (see Supplementary Table S1), and chose for confirmation one of them, which is also upregulated in LZ but much more highly expressed than the canonical one (see Fig. 7A,a). The selected transcript variant, which is generated through an alternative start site and a combination of all the above described AS mechanisms (i.e. SE, A5SS, A3SS, MXE, RI), encodes a putative shorter, 362 residues protein, containing an identical carboxyl-terminal region to that of the canonical protein, but a completely different amino-terminal region (see Fig. 7B).

We also chose *BC051142*, which ranked among the genes with the highest number of splice variants in our lists, as we detected 25 RNA isoforms expressed along spermatogenesis (when we used slightly less restrictive parameters, the number of expressed RNA isoforms for this gene raised to 103 splice variants). While there are eight putatively coding isoforms annotated in Ensembl, our analysis unveils the existence of at least nine additional unannotated protein-coding isoforms for this gene. None of the isoforms was detected in the 2 C cell population (i.e. somatic testicular cells and spermatogonia), and the expression of all of them starts in LZ, raising along spermatogenesis progress (Supplementary Table S1). In particular, we selected two unannotated putative protein-coding isoforms (Fig. 7A,b), for confirming their existence.

We also chose *ATP/GTP Binding Protein Like 5* (*Agbl5*), for which we have found several unannotated coding splice variants that are expressed at different levels along spermatogenic stages (see Supplementary Table S1). In particular, we selected for confirmation a very highly expressed variant that attains its expression peak in RS and encodes a putative 415 amino acids protein, unlike the canonical isoform whose highest expression level is in PS, and whose protein product is 846 residues long (Fig. 7A,c).

Additionally, we chose *La-Related Protein 1* (*Larp1*), *Serine-Threonine Kinase 31* (*Stk31*), *Bromodomain Adjacent to Zinc Finger Domain 1a* (*Baz1a*), and *Radial Spoke Head Component 1* (*Rsph1*). For all these genes, we have selected for confirmation unannotated highly expressed splice variants (Fig. 7A,d-g) that encode putative proteins that significantly differ from the annotated ones. In most cases, these novel variants are much more highly expressed than the canonical ones (see Fig. 7A,d-f). At least for some of them, their protein products would lack key domains (Fig. 7B), suggesting that these putative isoforms would accomplish different roles than the canonical ones.

We have been able to confirm the existence of all the chosen splice variants (see Fig. 7A,a-g), which further validates the results from our lists and shows the high

reliability of our data concerning the identification of unannotated spermatogenic isoforms.

Discussion

The RNAseq analysis of highly pure stage-specific spermatogenic cell populations reveals a high number of undisclosed transcripts in early meiotic prophase

Different reports have indicated that the testis has a particularly complex transcriptome [2, 6], with AS significantly contributing to its complexity [20, 22, 23, 29, 34]. Moreover, it is known that proper stage-specific AS is critical for successful spermatogenesis [20, 22, 23, 29–32, 46]. However, due to the heterogeneous composition of the testis, most likely an important number of cell type-specific RNA isoforms fall below the detection limits when whole testes or poorly purified cell types, are employed for transcriptome studies. Moreover, despite scRNA-seq allows to study the transcripts of individual cells, which has recently helped improve the understanding of spermatogenesis [47], it is important to take into account that scRNA-seq libraries are lower in depth than those for bulk sequencing, which does not allow the detection and assembly of low expression transcripts. Here, the use of highly purified stage-specific spermatogenic cell populations, added to the depth of the sequencing libraries, allowed us to detect a high number of yet unannotated genes and AS transcripts, hence showing that the transcriptomic diversity of the testis is considerably higher than previously reported.

The LZ cell population showed the majority of unannotated splice variants. This can be partly explained by our finding that they present lower overall expression levels compared to those of the other testicular cell populations, which would be in agreement with early reports that suggested the existence of low global transcription levels in mouse testes during early meiotic prophase [48–50].

Another important factor that surely contributed to hamper the previous detection of many LZ transcripts, is that these stages are very short and difficult to obtain as isolated cell populations, and therefore they have been rarely used in transcriptomic studies in comparison to other spermatogenic stages such as medium/late meiotic prophase and spermiogenesis [41, 42]. Besides, it is reasonable to think that, due to the scarceness of these cell types, specific transcripts of them may have become diluted among those of the most abundant cell types in whole testes transcriptomes. Remarkably, 159 out of the 191 newly identified putatively coding genes are expressed in LZ spermatocytes, and almost half of them are exclusive of LZ; we can reason that they may have gone unnoticed so far precisely because they encode LZ-specific products. Of note, surely something similar happens with scanty cell types in other heterogeneous tissues, where a high number of specific transcripts must still be undetected.

Beyond the fact that LZ stages presented the largest number of unannotated transcripts among all the analyzed cell populations, they also showed the highest number of transcripts considering both those annotated and unannotated together. In fact, our results are in line with a scRNA-seq study that has suggested that early spermatogenic stages express a higher number of genes, while later stages tend to concentrate a higher fraction of their transcripts on a narrower set of genes [3]. We propose that this high number of LZ-genes and isoforms could be required to accomplish the unique events that take place during early meiotic prophase. Noteworthy, the molecular groundwork of such events is largely unknown: we still do not really understand the role of bouquet formation, neither how homologous chromosomes recognize each other. In this scenario, the identification of all these unannotated genes and splice variants (both coding and noncoding), may represent a step forward toward the understanding of these essential processes and how they are regulated.

A large amount of still unannotated spermatogenic lncRNAs

The analysis of the coding potential of the unannotated transcripts, indicated that the highest number of them are noncoding (see Fig. 4). This makes sense as research regarding lncRNAs is much more recent than that of coding genes, and indicates that, when it comes to lncRNAs, we have only seen the tip of the iceberg, and there is still a high number of them to be discovered.

In relation to this, in a previous study, while attempting to conduct conservation analysis between spermatogenic lncRNAs of mouse and human, we found that for several lncRNAs from one species there were homologous DNA sequences in the other, but a cognate lncRNA was not annotated [42]. Although certainly this may be evidencing species-specific expression differences - which agrees with the fact that the expression patterns of lncRNAs are less conserved than those of coding genes [11] - this result may be also reflecting, at least in part, the incompleteness of the annotation of lncRNAs.

We have detected most of the DE lncRNA transcripts at the transition from meiosis (PS) to spermiogenesis (RS). This agrees with our previous observation that most of the differential expression of lncRNA genes along spermatogenesis takes place in spermiogenesis [42], and extends this result to splice variants.

The high numbers of unannotated spermatogenic lncRNAs we have identified, which add to the much higher amount of already annotated lncRNAs in male germ cells than in any other analyzed tissues and cell types [6–11], may be partly interpreted as a consequence

of the relaxed chromatin of meiotic and post-meiotic cells, but also for the high levels of post-transcriptional regulation that are present in these cells (see next section).

A particular characteristic we found for noncoding genes was a lower number of transcripts per gene in comparison to protein-coding ones, thus indicating that noncoding genes tend to have less AS isoforms. The latter is in consonance with some earlier reports that indicated that the splicing of lncRNAs was less efficient than that of mRNAs [9, 51]. Besides, this is also in line with our results and those of other groups, which showed that lncRNAs tend to be shorter and have less exons than mRNAs [7, 9, 42], adding to the conclusion that lncRNAs are, in general, less complex than mRNAs.

The number of unannotated transcripts and splice variants reinforces the concept of the high transcriptomic complexity of meiotic and post-meiotic cells

The meiotic and post-meiotic cell populations presented a higher number of unannotated transcripts compared to the 2 C cell population. This comparatively low number of the 2 C population may be reflecting the already known fact that meiotic and post-meiotic cells have extremely complex transcriptomes [6].

The widespread transcriptome complexity of male meiotic and post-meiotic cells has been proposed to be a consequence of their permissive chromatin state, which in turn results from the extensive chromatin remodeling that, due to histone replacement, takes place during these stages [6]. In this regard, we can speculate that at least part of the high number of unannotated transcripts that we found in meiotic and post-meiotic cells represents promiscuous transcription. In connection with this, while this manuscript was under review, a paper by Peters and collaborators [52] also showed a high number of novel unannotated transcripts in mouse male germ cells. Remarkably, the authors analyzed whether the expression of the high number of discovered transcripts could be influenced by repetitive elements in a cell type-specific manner, and found no evidence supporting that hypothesis.

On the other hand, the extensive transcriptome diversity of meiotic and especially of post-meiotic cells is also viewed as part of a strategy to regulate protein synthesis in the transcriptionally inert elongating and elongated spermatids. The need to have all the transcripts available to be translated in a timely fashion led to the development of diverse post-transcriptional regulatory mechanisms - some of which are unique to spermatocytes and RS - to accomplish the strict regulation requirements [1, 2, 25, 53]. In turn, these post-transcriptional regulation mechanisms most probably require a high amount of regulatory RNAs. In fact, although a large proportion of the spermatogenic lncRNAs are probably nonfunctional, at least for some of them, their importance for the regulation of spermatogenesis progression and fertility preservation, is being demonstrated [2, 54–63].

In summary, our results indicate that the transcriptomic complexity of spermatogenic cells is even higher than previously reported, and reinforces the concept that AS is particularly prominent for meiosis and spermiogenesis.

Characterization of AS patterns reveals previously unknown interesting splice variants

The analysis of our RNAseq data showed SE to be the most abundant AS type, followed by RI, for the four cell populations. This is in agreement with the results shown by Li et al. in a reanalysis study of repository-available data (of mention, early meiotic prophase was not included in that study) [38]. Our results also agree with those of Naro et al. [53]., who found RI as one of the most represented regulated AS patterns in the trans-meiotic differentiation of male germ cells. Noteworthy, they observed that RI events were upregulated in spermatocytes compared to spermatids, suggesting that intron retention represents a modality of nuclear retention of transcripts in meiosis, for their timely translation in inactive post-meiotic germ cells [53]. Although we did not detect significant differences regarding RI between the four cell populations, it must be noted that these results are not comparable, as we only analyzed the prevalence of the diverse AS categories in the different spermatogenic cell populations, but not differentially regulated splicing events.

We also detected some unannotated splice variants with much higher expression levels than the annotated ones. In many cases, they may have gone unnoticed because they are highly expressed in a specific stage, which is often poorly represented (i.e., LZ). More important, for the newly identified AS transcripts with high coding potential, despite the limitation that the confirmation of the existence of their protein products is pending, most probably at least part of them encode unnoticed testis-specific protein isoforms. We can hypothesize that, at least some of them, have "novel" testis-specific functions. A key aspect in understanding the physiological validity of the discovery of interesting unannotated splice variants is that we were able to detect them using an alternative approach to RNA-seq, i.e. RT-PCR. Remarkably, they all represent examples of previously undisclosed, putative protein-coding isoforms that are DE along spermatogenic stages, and whose canonical proteins, in most cases, are known to play essential roles in spermatogenesis. In some cases, the putative unannotated protein isoforms would lack important domains.

An interesting example of the above is the undisclosed isoform we detected for *Msh5*. MSH5 is a meiotic-specific mismatch repair protein involved in homologous recombination [64] that has proved to be essential for meiotic progression [65]. The novel isoform, whose transcript is abundant in LZ, would have an incomplete ATPase domain that is required for double strand breaks repair [66], thus suggesting that this unannotated isoform could be accomplishing a different role during meiotic prophase.

Another, curious, example is *BC051142*,a gene that according to NCBI database is highly testis-specific (see https://www.ncbi.nlm.nih.gov/gene/?term=BC051142), and whose human homolog, *Testis Expressed Basic Protein 1* (*TSBP1*), has been associated with hypogonadism (https://www.genecards.org/cgi-bin/carddisp.pl?gene=TSBP1&keywords=BC051142). However, despite it encodes a high number of spermatogenic-specific different putative protein isoforms, its function is still unknown. Therefore, it constitutes an excellent example to illustrate the enormous variability that exists throughout spermatogenesis, and all that remains to be unveiled.

Concerning *Agbl5*, it is a highly testis-biased gene (https://www.ncbi.nlm.nih.gov/gene/?term=agbl5+mus+musculus) that encodes a metallocarboxypeptidase involved in tubulin deglutamylation, which is essential for the formation of functional sperm. It has been shown that AGBL5 (also known as CCP5) is necessary for the integrity of sperm flagella and for other microtubule-based functions during spermatogenesis [67, 68]. Although various splice variants have been reported, at least one of them even with apparently distinct properties [67], according to our findings several other unannotated coding splice variants expressed along spermatogenesis would exist.

Larp1 encodes an RNA-binding protein that regulates the translation and stability of mRNAs for ribosomal proteins and translation factors downstream of TORC1 complex [69, 70], and is most highly expressed in the testis compared to other tissues (https://www.ncbi.nlm.nih.gov/gene/73158). *Stk31* is a testis-biased gene (https://www.ncbi.nlm.nih.gov/gene/77485) that encodes a serine-threonine kinase with a Tudor domain, which is preferentially localized in germinal granules of spermatocytes and acrosomal cap of spermatids, interacting with MIWI protein [71]. Besides, it has been shown to be a cancer/testis antigen highly expressed in several types of cancers [72–74]. *Baz1a* is highly [75] and dynamically expressed during spermatogenesis [76], and encodes a defining subunit of an ATP-dependent chromatin remodeler complex essential for proper spermatogenic gene expression and fertility in mouse [75]. *Rsph1*, whose expression is testis-restricted [77] (see https://www.ncbi.nlm.nih.gov/gene/22092), directs the synthesis of a component of radial spokes head of cilia and sperm flagella [78], and mutations in this gene have been related to fertility problems in humans, resulting in primary ciliary dyskinesia and motility alterations of cilia and sperm [79]. For all these genes, the newly identified putative protein isoforms would differ significantly from the canonical ones. STK31 is an example of this: while the known protein has a Tudor domain and a protein-kinase domain that is essential for its function as a serin-threonine kinase, the predicted variant would lack the latter, thus suggesting that it should play a different role.

Conclusions

In this work, we generated a great amount of highly reliable information about gene expression along spermatogenesis, from pure flow sorted stage-specific mouse spermatogenic cell populations. Our results reveal a high number of yet unannotated spermatogenic lncRNAs, undisclosed splice variants of coding genes, and even some unannotated protein-coding genes. At least part of the newly identified splice variants encodes putative isoforms of important spermatogenic proteins. Besides, we have delved into the characterization of spermatogenic alternative splicing. Importantly, the largest number of spermatogenic stage-specific unannotated transcripts and splice variants are expressed during early meiotic prophase, a stage that has been scarcely studied in former transcriptomic analyses. We propose that these may be related to the unique and complex processes that take place during these stages.

Overall, this study shows that testicular transcriptomic diversity is considerably higher than previously reported. A general conclusion we can draw is that not only a great deal of existing variability in terms of spermatogenic non-coding RNAs and stage-specific protein variants is still to be revealed, but we do not even know the exact number of coding genes yet, even in a model as studied as the mouse.

Methods

Raw data

The raw data employed in this study came from stranded RNAseq libraries of testis-specific cell populations representative of landmark stages along mouse spermatogenesis, obtained through flow sorting [42] (SRA repository access number PRJNA548952). The cell populations were: 2 C (a heterogeneous population with 2 C DNA content, consisting of spermatogonia and testicular somatic cells); LZ (leptotene and zygotene spermatocytes); PS (pachytene spermatocytes); and RS (round spermatids), totaling 12 libraries, i.e. four different cell populations, with three biological replicates each. As previously stated [42], the 2 C cell population was obtained from a testicular cell suspension of a pool of up to five individuals of 12–14

days *postpartum* (d*pp*), which excludes the possibility that this population contains spermatocytes II; LZ and PS cell populations were classified from 15 to 19 d*pp* animals, and RS from 22 to 24 d*pp* animals.

General data processing

Neither the RNA extraction method nor the library type focused on small RNAs, and therefore the analysis was centered on mRNAs and lncRNAs. Moreover, only molecules ≥ 200 bp were considered in this study, and every genome unit that generated transcripts above that size, was considered a gene.

Low-quality reads (Q < 20) and adapter sequences were trimmed using TrimGalore [80]. Reads that passed quality control were mapped with HISAT2 (http://daehwankimlab.github.io/hisat2/), employing dta (downstream-transcriptome-assembly) parameters. We performed genome-guided alignment, using both paired and unpaired reads for each cell population, and discarding reads with multimapping. *Mus musculus* Ensembl database (Grcm38.92 release) was used as reference genome.

We used Strawberry [81] to assemble new transcripts under the guidance of genome alignment, employing 10 reads as minimum support per splice site, and per exon. Besides, during the set-up we used different depth cut-offs and found that the results did not substantially change. We therefore chose to work with a minimum of 10X coverage, as it turned out to be a strong support (Supplementary Figure S6). In order to generate a unique reference of our assemblies, we employed StringTie, with merge option [82].

A correlation matrix was constructed in R bioconductor (http://www.R-project.org), calculating Pearson's correlation coefficient between FPKM expression of every transcript in each of the 12 samples, to appreciate the strength of the correlations between ourreplicates.

We analyzed transcripts discovery saturation throughout rarefaction curves at different read depths, with the aim of checking if we reached saturation in the 4 cell populations, and to rule out artifacts. For this purpose, we carried out counts with FeatureCounts [83] using the data from this paper, and compared them to those of da Cruz et al. [41]. (SRA repository access number PRJNA317251). The following conditions were used: -O assigns reads to all their overlapping meta-features; -S0 indicates unstranded reads; -t specifies feature type(s) in a GTF annotation; and -g states for attribute type in GTF annotation, with the reference that we previously generated. Subsequently, in R, we employed the function "estimate saturation" from the RNAseQC library [84], which allows cutting by depth and thus seeing how transcript detection occurs, based on the number of reads.

Data comparison with single cell RNAseq studies

For comparison of our RNA lists with those from another report in which different spermatogenic stages were studied at scRNA-seq level [37], we downloaded the raw data from NCBI's Gene Expression Omnibus (GEO) data repository (http://www.ncbi.nlm.nih.gov/gen/) with the accession ID: GSE107644. We mapped the raw data from that study with the same pipeline used for our own data, then performed the counts of our data and those of single seq with our assembly employing HTseq-counts [85], and the generated lists were normalized with limma package for R [86], using the function removeBatchEffect. A Principal Component Analysis (PCA) was generated by means of Seurat (that uses normalized log CPM [Counts Per Million] values as input) [87].

Detection of splice variants, analysis of coding potential, annotation, and structural prediction of putative proteins

The generated reference GTF file containing our assembly was converted to a FASTA file by means of gffread [88]. We used this FASTA file as input for the different employed software packages, to categorize the new transcripts into coding or noncoding. For this categorization, we used four different software packages in parallel: TransDecoder [89], CPC2 [90], LncADeep [91], and CPAT [92], all of them with their default parameters. For further analysis, we proceeded with the intersection of the four software packages.

Venn diagrams were constructed using free Bioinformatics & Evolutionary Genomics software (http://bioinformatics.psb.ugent.be/webtools/Venn/).

We used rMATS software (http://rnaseq-mats.sourceforge.net/) with its default parameters, for the analysis of the different types of AS patterns. For the determination of the number of transcripts per gene for coding and noncoding transcripts, we plotted them normalized as the percentage of total transcripts in each category. T-test was conducted to calculate statistical values between the cell types using their replicates. We used PlotTranscripts function [82] to see the transcript structure and expression for single gene analysis.

With the aim of assessing the functionality of the unannotated genes, we conducted a primary annotation by means of Trinotate [93], using all the software´s available methods and databases (BLASTX using SWISSPROT, RNAMMER, prot_id, BLASTP, Pfam, SignalP, TMHMM, eggNOG, KEGG, Gene Ontology BLAST, Gene Ontology Pfam). Modeling of predicted proteins was conducted through Swiss-Model (https://swissmodel.expasy.org/interactive), and an analysis of putative protein domains was performed with Pfam (http://pfam-legacy.xfam.org).

Differential gene expression analysis

Differential gene expression between the four testicular cell populations was obtained employing StringTie -e (quantification function) -B (output option for Ballgown analysis), --fr (stranded library fr-secondstrand), and using our assembly as a reference to generate the counts and FPKM.

Pairwise comparisons were made in chronological order of appearance along the first spermatogenic wave (LZ vs. 2 C; PS vs. LZ; RS vs. PS), by means of Ballgown software [82]. A log2 fold change (FC) ≥2 or ≤-2, and q value<0.05 was used to filter the DE genes. We also filtered by a minimum of 10X coverage.

All followed bioinformatics protocols are illustrated in Fig. 2.

Animals and Ethics statement

Animal procedures were performed following the recommendations of the Uruguayan National Commission of Animal Experimentation (CNEA, http://www.cnea.org.uy), approved experimental protocol 001/02/2012 (code: 008/11). Male CD-1 Swiss mice (*Mus musculus*) were obtained from the animal facility at Instituto de Higiene of Facultad de Medicina (UdelaR, Montevideo, Uruguay). Animals were euthanized by cervical dislocation, in accordance with the National Law of Animal Experimentation 18,611 (Uruguay). Immediately after euthanasia testes were dissected and tunica albuginea was removed, before proceeding to the preparation of testicular cell suspensions for sorting and RT-PCR.

Confirmative RT-PCR

For the confirmation of the selected splice variants, we designed specific primers to amplify, either the newly identified transcripts or the annotated ones. Especially designed primers are listed in Supplementary Table S3.

Cell fractions containing 3,000 cells each from 2 C, LZ, PS, and RS populations were sorted as previously described [42]. Briefly, cell suspensions were prepared and stained with Vybrant DyeCycle Green (VDG; Invitrogen, Life Technologies, Carlsbad, CA), as instructed [45]. The sorting was conducted in a MoFlo Astrios EQ (Beckman Coulter) in Purify mode (with 1–2 drops). The sorted cell fractions were used for confirmative RT-PCR by means of the Power SYBR Green Cells-to-Ct Kit (Ambion-Life Technologies) essentially as instructed, using random primers for first strand cDNA synthesis. We used 2 µL cDNA in 20 µL final volume PCR reaction following the instructions of the Cells-to-Ct Kit, and employing a CFX96 Touch Real-Time PCR Detection System 1 (BioRad, Hercules, CA), with three biological replicas each. Although RT-qPCR was not mandatory for the confirmation of splice variants, we chose to use this kit for its high sensitivity, given the low input of sorted cells. The PCR reactions were run in conventional agarose gels and stained with GelRed (Biotium, Fremont, CA, USA).

Abbreviations

aG	Annotated genes
AS	Alternative splicing
aT	Annotated transcripts
A3SS	Alternative 3′ splice site
A5SS	Alternative 5′ splice site
d*pp*	Days *postpartum*
DE	Differentially expressed
lncRNAs	Long noncoding RNAs
LZ	Leptotene-zygotene
MXE	Mutually exclusive exons
PS	Pachytene spermatocytes
RI	Retained intron
RS	Round spermatids
scRNA-seq	Single-cell RNA sequencing
SE	Skipping exon
uG	Unannotated genes
uT	Unannotated transcripts

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10170-z.

Supplementary Figure S1.Correlation matrix for the 4 cell populations with 3 biological replicas each.

Supplementary Figure S2.Principal component analysis (PCA) comparing our RNAseq data with those of a scRNA-seq of 20 different spermatogenic cell subtypes [37]. The cell populations from our study are represented as squares, while those from the single-cell study are depicted as circles. Notably, the correlation is very good taking into consideration that many conditions in both experiments were different. As an example, in this single-cell study the spermatogenic process was manipulated through a combination of transgenic labeling and artificial synchronization of the cycle of the seminiferous epithelium, and therefore a slight shift in the time of appearance of some transcripts cannot be ruled out. Of mention, the data from our 2C cell population was not included for comparison, as besides spermatogonia it contains somatic testicular cells, which were not included in the single-cell study.L: leptotene; Z: zygotene; LZ: lepto/zygotene; eP: early pachytene; mP: medium pachytene; lP: late pachytene; PS: pachytene spermatocytes; D: diplotene; RS: round spermatids; RS2_1-5: early round spermatids, steps 1-2; RS8_1-5: late round spermatids, steps 7-8.

Supplementary Figure S3.Genes and transcripts expressed in our lists.A) Flow chart representing the process of categorizing the genes expressed in the four testicular cell populations, and the expressed transcripts generated from them. The categories are, in each case, annotated or unannotated, and, for the unannotated transcripts, high or low coding potential. The number of genes or transcripts in each category is indicated. It is important to recall that the number of categorized transcripts according to coding potential is only a subset, as we only kept the intersection of the four used software programs. The individual result of each program is shown at the bottom of the figure.**B-D)**Number of expressed genes and transcripts arising from them, discriminated by the four testicular cell populations. **B)** Pie chart of annotated genes (aG: blue) and unannotated genes (uG: red) expressed in each of the four cell populations that passed all the filters. **C)** Pie chart of annotated transcripts (aT: blue) and unannotated transcripts (uT: red) expressed in each of the four spermatogenic cell populations. **D)** Pie chart showing the origin of the unannotated transcripts in our lists for each of the four cell populations, either undisclosed splice variants of already annotated genes (of aG: blue), or transcripts arising from unannotated genes (of uG: red). Note that the unannotated genes and transcripts are more stage-specific than the annotated ones. As a consequence, the different cell populations share a higher number of annotated expressed genes/transcripts compared to

the unannotated ones. Due to the transcripts in common, this is visualized as a higher proportion of annotated genes and transcripts when they are separately analyzed by cell population.

Supplementary Figure S4. Transcript distribution in the four testicular cell populations. A) Representation of the unannotated transcripts that were coincidentally identified as coding or noncoding with the four programs for coding potential analysis and depicted in Figure 5C, but distributed according to the different categories (*i.e.* coding or noncoding; splice variants of already annotated genes or transcripts of unannotated genes). **B)** Representation of all the 33,002 newly identified transcripts (previous to their filtration for coding potential), and showing 6,708 transcripts as expressed in 2C; 18,607 in LZ; 12,353 in PS; and 12,575 in RS. **C)** Representation of all the detected transcripts in our lists (both annotated and unannotated).

Supplementary Figure S5. Saturation and expression distribution in the four cell populations.A) Rarefaction analysis in the studied samples, including data of da Cruz *et al*., 2016 [41]. **B**) Histogram distribution analysis of expression in the four testicular cell populations. The values of the lowest expression range (corresponding to 2C: 85,263 transcripts; LZ: 68,740; PS: 84,947; and RS: 76,905), were excluded from the graph to have a clearer representation.

Supplementary Figure S6.Semi-logarithmic plot of identified transcripts *vs* coverage for 7 different transcript abundance cut-offs. The ordinate axis (RNA abundance) indicates the logarithmic scale (log2) of transcripts number.

Supplementary Figure S7: Original agarose gels from Figure 7. The cropped regions are demarcated by red squares.

Supplementary Table S1: Expression and annotation of detected transcripts. ENSMUST stands for Ensembl-annotated transcripts, while MSTRG designates unannotated transcripts.

Supplementary Table S2: Expression and annotation of the 223 newly identified transcripts with high coding potential, that correspond to 191 unannotated genes.

Supplementary Table S3: List of the PCR primers used in this study.

Acknowledgements
The authors want to thank MSc Federico Santiñaque from the flow cytometry core at IIBCE (SECIF-IIBCE) for his valuable help concerning sorting of spermatogenic cell populations.

Author contributions
CRediT author statement: Carlos Romeo: Formal analysis, Investigation, visualization, validation, Writing - Original Draft, Writing - Review & Editing. María Fernanda Trovero: Formal analysis. Santiago Radio: Formal analysis. Pablo Smircich: Formal analysis, Supervision. Rosana Rodriguez-Casuriaga: Validation, Supervision.Adriana Geisinger: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Project administration, Funding acquisition, Supervision, Resources. Jose Sotelo: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Funding acquisition, Supervision, Resources.

Funding
This work was supported by Comisión Sectorial de Investigación Científica (CSIC), UdelaR (Uruguay) under an I + D Groups grant to AG and Ricardo Benavente, and Agencia Nacional de Investigación e Innovación (ANII, Uruguay), under grant FCE_1_2021_1_166510 to AG. CR was awarded with a short PhD scholarship from Comisión Académica de Posgrado (CAP), UdelaR.

Data availability
The datasets used and analysed during the current study are available in the SRA repository, with access number PRJNA548952, (https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA548952).

Declarations

Ethics approval
Animal procedures were performed following the recommendations of the Uruguayan National Commission of Animal Experimentation (CNEA), approved experimental protocol 001/02/2012 (code: 008/11; http://www.cnea.org.uy/index.php/instituciones/registro/10). Animals were euthanized by cervical dislocation, in accordance with the National Law of Animal Experimentation 18,611 (Uruguay).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Laboratory of Molecular Biology of Reproduction, Department of Molecular Biology, Instituto de Investigaciones Biológicas Clemente Estable (IIBCE), 11,600 Montevideo, Uruguay
[2]Department of Genomics, IIBCE, 11,600 Montevideo, Uruguay
[3]Biochemistry-Molecular Biology, Facultad de Ciencias, Universidad de la República (UdelaR), 11,400 Montevideo, Uruguay
[4]Department of Cell and Molecular Biology, Facultad de Ciencias, UdelaR, 11,400 Montevideo, Uruguay
[5]Present address: Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Received: 2 October 2023 / Accepted: 28 February 2024
Published online: 20 March 2024

References

1. Kleene KC. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech Dev. 2001;106(1–2):3–23.
2. Geisinger A, Rodríguez-Casuriaga R, Benavente R. Transcriptomics of meiosis in the male mouse. Front Cell Dev Biol. 2021;9.
3. Green CD, Ma Q, Manske GL, Shami AN, Zheng X, Marini S, et al. A Comprehensive Roadmap of Murine Spermatogenesis defined by single-cell RNA-Seq. Dev Cell. 2018;46(5):651–67e10.
4. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348(6235):660–5.
5. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347:6220.
6. Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep. 2013;3(6):2179–90.
7. Cabili M, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25(18):1915–27.
8. Darbellay F, Necsulea A. Comparative transcriptomics analyses across species, organs, and Developmental stages Reveal functionally constrained lncRNAs. Mol Biol Evol. 2020;37(1):240–59.
9. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, 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.
10. Hong SH, Kwon JT, Kim J, Jeong J, Kim J, Lee S et al. Profiling of testis-specific long noncoding RNAs in mice. BMC Genomics. 2018;19(1).
11. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505(7485):635–40.
12. Bortvin A. PIWI-interacting RNAs (piRNAs) - a mouse testis perspective. Biochem (Mosc). 2013;78(6):592–602.
13. de Mateo S, Sassone-Corsi P. Regulation of spermatogenesis by small non-coding RNAs: role of the germ granule. Semin Cell Dev Biol. 2014;29:84–92.
14. Kotaja N. MicroRNAs and spermatogenesis. Fertil Steril. 2014;101(6):1552–62.
15. Yadav RP, Kotaja N. Small RNAs in spermatogenesis. Mol Cell Endocrinol. 2014;382(1):498–508.

16. Hilz S, Modzelewski AJ, Cohen PE, Grimson A. The roles of microRNAs and siRNAs in mammalian spermatogenesis. Development. 2016;143(17):3061–73.
17. He C, Wang K, Gao Y, Wang C, Li L, Liao Y et al. Roles of noncoding RNA in Reproduction. Front Genet. 2021;12.
18. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. Genome Biol. 2004;5(10).
19. Kan Z, Garrett-Engele PW, Johnson JM, Castle JC. Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles. Nucleic Acids Res. 2005;33(17):5659–66.
20. Naro C, Cesari E, Sette C. Splicing regulation in brain and testis: common themes for highly specialized organs. Cell Cycle. 2021;20(5–6):480–9.
21. Mazin PV, Khaitovich P, Cardoso-Moreira M, Kaessmann H. Alternative splicing during mammalian organ development. Nat Genet. 2021;53(6):925–34.
22. Legrand JMD, Hobbs RM. RNA processing in the male germline: mechanisms and implications for fertility. Semin Cell Dev Biol. 2018;79:80–91.
23. Song H, Wang L, Chen D, Li F. The function of Pre-mRNA Alternative Splicing in Mammal Spermatogenesis. Int J Biol Sci. 2020;16(1):38–48.
24. Idler RK, Yan W. Control of messenger RNA fate by RNA-binding proteins: an emphasis on mammalian spermatogenesis. J Androl. 2012;33(3):309–37.
25. Licatalosi DD. Roles of RNA-binding proteins and post-transcriptional regulation in driving male germ cell development in the mouse. Adv Exp Med Biol. 2016;907:123–51.
26. MacDonald CC. Tissue-specific mechanisms of alternative polyadenylation: Testis, brain, and beyond (2018 update). Wiley Interdiscip Rev RNA. 2019;10(4).
27. Grosso AR, Gomes AQ, Barbosa-Morais NL, Caldeira S, Thorne NP, Grech G, et al. Tissue-specific splicing factor gene expression signatures. Nucleic Acids Res. 2008;36(15):4823–32.
28. de la Grange P, Gratadou L, Delord M, Dutertre M, Auboeuf D. Splicing factor and exon profiling across human tissues. Nucleic Acids Res. 2010;38(9):2825–38.
29. Wu D, Khan FA, Huo L, Sun F, Huang C. Alternative splicing and MicroRNA: epigenetic mystique in male reproduction. RNA Biol. 2022;19(1):162–75.
30. Bao J, Tang C, Li J, Zhang Y, Bhetwal BP, Zheng H et al. RAN-binding protein 9 is involved in alternative splicing and is critical for male germ cell development and male fertility. PLoS Genet. 2014;10(12).
31. Iwamori N, Tominaga K, Sato T, Riehle K, Iwamori T, Ohkawa Y, et al. MRG15 is required for pre-mRNA splicing and spermatogenesis. Proc Natl Acad Sci U S A. 2016;113(37):E5408–15.
32. Hannigan MM, Zagore LL, Licatalosi DD. Ptbp2 controls an alternative splicing network required for cell communication during spermatogenesis. Cell Rep. 2017;19(12):2598–612.
33. Laiho A, Kotaja N, Gyenesei A, Sironen A. Transcriptome profiling of the murine testis during the first wave of spermatogenesis. PLoS ONE. 2013;8(4).
34. Schmid R, Grellscheid SN, Ehrmann I, Dalgliesh C, Danilenko M, Paronetto MP, et al. The splicing landscape is globally reprogrammed during male meiosis. Nucleic Acids Res. 2013;41(22):10170–84.
35. Margolin G, Khil PP, Kim J, Bellani MA, Camerini-Otero RD. Integrated transcriptome analysis of mouse spermatogenesis. BMC Genomics. 2014;15(1).
36. Zuo H, Zhang J, Zhang L, Ren X, Chen X, Hao H et al. Transcriptomic variation during spermiogenesis in mouse germ cells. PLoS ONE. 2016;11(11).
37. Chen Y, Zheng Y, Gao Y, Lin Z, Yang S, Wang T, et al. Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. Cell Res. 2018;28(9):879–96.
38. Li Q, Li T, Xiao X, Ahmad DW, Zhang N, Li H, et al. Specific expression and alternative splicing of mouse genes during spermatogenesis. Mol Omics. 2020;16(3):258–67.
39. Chalmel F, Lardenois A, Evrard B, Rolland AD, Sallou O, Dumargne MC et al. High-resolution profiling of novel transcribed regions during rat spermatogenesis. Biol Reprod. 2014;91(1).
40. Rolland AD, Evrard B, Darde TA, Le Beguec C, Le Bras Y, Bensalah K, et al. RNA profiling of human testicular cells identifies syntenic lncRNAs associated with spermatogenesis. Hum Reprod. 2019;34(7):1278–90.
41. da Cruz I, Rodríguez-Casuriaga R, Santiñaque FF, Farías J, Curti G, Capoano CA et al. Transcriptome analysis of highly purified mouse spermatogenic cell populations: gene expression signatures switch from meiotic-to postmeiotic-related processes at pachytene stage. BMC Genomics. 2016;17(1).
42. Trovero MF, Rodríguez-Casuriaga R, Romeo C, Santiñaque FF, François M, Folle GA, et al. Revealing stage-specific expression patterns of long noncoding RNAs along mouse spermatogenesis. RNA Biol. 2020;17(3):350–65.
43. Rodríguez-Casuriaga R, Folle GA, Santiñaque F, López-Carro B, Geisinger A. Simple and efficient technique for the preparation of testicular cell suspensions. J Visualized Experiments. 2013;(78):1–7.
44. Rodríguez-Casuriaga R, Santiñaque FF, Folle GA, Souza E, López-Carro B, Geisinger A. Rapid preparation of rodent testicular cell suspensions and spermatogenic stages purification by flow cytometry using a novel blue-laser-excitable vital dye. MethodsX. 2014;1:239–43.
45. Geisinger A, Rodríguez-Casuriaga R. Flow cytometry for the isolation and characterization of rodent meiocytes. Methods Mol Biol. 2017;1471:217–30.
46. Liu W, Wang F, Xu Q, Shi J, Zhang X, Lu X et al. BCAS2 is involved in alternative mRNA splicing in spermatogonia and the transition to meiosis. Nat Commun. 2017;8.
47. Rabbani M, Zheng X, Manske GL, Vargo A, Shami AN, Li JZ, et al. Decoding the spermatogenesis program: new insights from transcriptomic analyses. Annu Rev Genet. 2022;56:339–68.
48. Monesi V. Ribonucleic acid synthesis during mitosis and meiosis in the mouse testis. J Cell Biol. 1964;22(3):521–32.
49. Kierszenbaum AL, Tres LL. Nucleolar and perichromosomal RNA synthesis during meiotic prophase in the mouse testis. J Cell Biol. 1974;60(1):39–53.
50. Page J, De La Fuente R, Manterola M, Parra MT, Viera A, Berríos S, et al. Inactivation or non-reactivation: what accounts better for the silence of sex chromosomes during mammalian male meiosis? Chromosoma. 2012;121(3):307–26.
51. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. 2012;22(9):1616–25.
52. Gill ME, Rohmer A, Erkek-Ozhan S, Liang CY, Chun S, Ozonov EA, Peters AHFM. De novo transcriptome assembly of mouse male germ cells reveals novel genes, stage-specific bidirectional promoter activity, and noncoding RNA expression. Genome Res. 2023;33(12):2060–78. https://doi.org/10.1101/gr.278060.123
53. Naro C, Jolly A, Di Persio S, Bielli P, Setterblad N, Alberdi AJ, et al. An orchestrated intron retention program in meiosis controls timely usage of transcripts during germ cell differentiation. Dev Cell. 2017;41(1):82–93e4.
54. Anguera MC, Ma W, Clift D, Namekawa S, Kelleher RJ, Lee JT. Tsx produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. PLoS Genet. 2011;7(9).
55. Ni MJ, Hu ZH, Liu Q, Liu MF, Lu MH, Zhang JS et al. Identification and characterization of a novel non-coding RNA involved in sperm maturation. PLoS ONE. 2011;6(10).
56. Lü M, Tian H, Cao YX, He X, Chen L, Song X et al. Downregulation of miR-320a/383-sponge-like long non-coding RNA NLC1-C (narcolepsy candidate-region 1 genes) is associated with male infertility and promotes testicular embryonal carcinoma cell proliferation. Cell Death Dis. 2015;6(11).
57. Li L, Wang M, Wang M, Wu X, Geng L, Xue Y et al. A long non-coding RNA interacts with Gfra1 and maintains survival of mouse spermatogonial stem cells. Cell Death Dis 2016;7(3).
58. Kataruka S, Akhade VS, Kayyar B, Rao MRS. Mrhl Long noncoding RNA mediates meiotic commitment of mouse spermatogonial cells by regulating Sox8 expression. Mol Cell Biol. 2017;37(14).
59. Nakajima R, Sato T, Ogawa T, Okano H, Noce T. A noncoding RNA containing a SINE-B1 motif associates with meiotic metaphase chromatin and has an indispensable function during spermatogenesis. PLoS ONE. 2017;12(6).
60. Li W, Ning JZ, Cheng F, Yu WM, Rao T, Ruan Y, et al. MALAT1 promotes cell apoptosis and suppresses cell proliferation in testicular ischemia-reperfusion injury by sponging MiR-214 to modulate TRPV4 expression. Cell Physiol Biochem. 2018;46(2):802–14.
61. Joshi M, Rajender S. Long non-coding RNAs (lncRNAs) in spermatogenesis and male infertility. Reprod Biol Endocrinol. 2020;18(1).
62. Li K, Xu J, Luo Y, Zou D, Han R, Zhong S, et al. Panoramic transcriptome analysis and functional screening of long noncoding RNAs in mouse spermatogenesis. Genome Res. 2021;31(1):13–26.
63. Liu W, Zhao Y, Liu X, Zhang X, Ding J, Li Y et al. A novel meiosis-related lncRNA, Rbakdn, contributes to spermatogenesis by stabilizing Ptbp2. Front Genet. 2021;12.
64. Harfe BD, Jinks-Robertson S. DNA mismatch repair and genetic instability. Annu Rev Genet. 2000;34:359–99.
65. Edelmann W, Cohen PE, Kneitz B, Winand N, Lia M, Heyer J, et al. Mammalian MutS homologue 5 is required for chromosome pairing in meiosis. Nat Genet. 1999;21(1):123–7.

66. Milano CR, Kim Holloway J, Zhang Y, Jin B, Smith C, Bergman A, et al. Mutation of the ATPase domain of MutS Homolog-5 (MSH5) reveals a requirement for a functional MutSγ complex for all crossovers in mammalian meiosis. G3. (Bethesda). 2019;9(6):1839–50.
67. Wu HY, Wei P, Morgan JI. Role of Cytosolic Carboxypeptidase 5 in neuronal survival and spermatogenesis. Sci Rep. 2017;7.
68. Giordano T, Gadadhar S, Bodakuntla S, Straub J, Leboucher S, Martinez G et al. Loss of the deglutamylase CCP5 perturbs multiple steps of spermatogenesis and leads to male infertility. J Cell Sci. 2019;132(3).
69. Fonseca BD, Lahr RM, Damgaard CK, Alain T, Berman AJ. LARP1 on TOP of ribosome production. Wiley Interdiscip Rev RNA. 2018;9(5).
70. Berman AJ, Thoreen CC, Dedeic Z, Chettle J, Roux PP, Blagden SP. Controversies around the function of LARP1. RNA Biol. 2021;18(2):207–17.
71. Bao J, Wang L, Lei J, Hu Y, Liu Y, Shen H, et al. STK31(TDRD8) is dynamically regulated throughout mouse spermatogenesis and interacts with MIWI protein. Histochem Cell Biol. 2012;137(3):377–89.
72. Zhong L, Liu J, Hu Y, Wang W, Xu F, Xu W, et al. STK31 as novel biomarker of metastatic potential and tumorigenicity of colorectal cancer. Oncotarget. 2017;8(15):24354–61.
73. Xiong J, Xing S, Dong Z, Niu L, Xu Q, Liu P, et al. STK31 regulates the proliferation and cell cycle of lung cancer cells via the Wnt/βcatenin pathway and feedback regulation by cmyc. Oncol Rep. 2020;43(2):395–404.
74. Bae DH, Kim HJ, Yoon BH, Park JL, Kim M, Kim SK et al. STK31 upregulation is associated with chromatin remodeling in gastric cancer and induction of tumorigenicity in a xenograft mouse model. Oncol Rep. 2021;45(4).
75. Dowdle JA, Mehta M, Kass EM, Vuong BQ, Inagaki A, Egli D et al. Mouse BAZ1A (ACF1) is dispensable for double-strand break repair but is essential for averting improper gene expression during spermatogenesis. PLoS Genet. 2013;9(11).
76. Yadav RP, Leskinen S, Ma L, Mäkelä JA, Kotaja N. Chromatin remodelers HELLS, WDHD1 and BAZ1A are dynamically expressed during mouse spermatogenesis. Reproduction. 2022;165(1):49–63.
77. Tsuchida J, Nishina Y, Wakabayashi N, Nozaki M, Sakai Y, Nishimune Y. Molecular cloning and characterization of meichroacidin (male meiotic metaphase chromosome-associated acidic protein). Dev Biol. 1998;197(1):67–76.
78. Zheng W, Li F, Ding Z, Liu H, Zhu L, Xu C et al. Distinct architecture and composition of mouse axonemal radial spoke head revealed by cryo-EM. Proc Natl Acad Sci U S A. 2021;118(4).
79. Kott E, Legendre M, Copin B, Papon JF, Dastot-Le Moal F, Montantin G, et al. Loss-of-function mutations in RSPH1 cause primary ciliary dyskinesia with central-complex and radial-spoke defects. Am J Hum Genet. 2013;93(3):561–70.
80. Lindgreen S. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes. 2012;5.
81. Liu R, Dickerson J, Strawberry. Fast and accurate genome-guided transcript reconstruction and quantification from RNA-Seq. PLoS Comput Biol. 2017;13(11).
82. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11(9):1650–67.
83. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
84. Deluca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28(11):1530–2.
85. Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
86. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
87. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–3587e29.
88. Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. F1000Res. 2020;9.
89. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. 2015;43(12).
90. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017;45(W1):W12–6.
91. Yang C, Yang L, Zhou M, Xie H, Zhang C, Wang MD, et al. LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. Bioinformatics. 2018;34(22):3825–34.
92. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41(6).
93. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A tissue-mapped Axolotl De Novo Transcriptome enables identification of limb regeneration factors. Cell Rep. 2017;18(3):762–76.
94. Rodríguez-Casuriaga R, Geisinger A. Contributions of Flow Cytometry to the Molecular Study of Spermatogenesis in mammals. Int J Mol Sci. 2021;22(3):1151. https://doi.org/10.3390/ijms22031151

Publisher's Note