
UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Fisiología, Biología Molecular y Celular

# Estudio de señales en cis y trans en la determinación de los patrones de *splicing* alternativo

Tesis a presentar para optar al título de
**Doctor de la Universidad de Buenos Aires, área Ciencias Biológicas**

Lic. Maximiliano Sebastián Beckel

Director de Tesis: **Dr. Ariel Chernomoretz**
Codirector: **Dr. Marcelo Yanovsky**
Consejero de estudios: **Dr. Alejandro Colman-Lerner**

Lugar de Trabajo:
**Laboratorio de biología de sistemas integrativa. Instituto de Investigaciones Bioquímicas de Buenos Aires (IIBBA), Fundación Instituto Leloir (FIL)**

Buenos Aires, 2022

# Estudio de señales en cis y trans en la determinación de los patrones de *splicing* alternativo


## Resumen

El *splicing* se encuentra presente a lo largo de las distintas especies de organismos eucariotas, representando una pieza clave en la inmensa mayoría de los procesos de regulación. Para que el mismo pueda llevarse a cabo, debe darse el reconocimiento de una serie de señales. En particular, los sitios donores y aceptores de *splicing* definen los límites del intrón y su reconocimiento constituye uno de los primeros pasos dentro del ciclo de *splicing*. A lo largo del genoma, estos sitios presentan una cierta variabilidad en su secuencia, la cual se encuentra relacionada con la posibilidad de regular su reconocimiento y establecer diversos patrones de *splicing* alternativo. Por otro lado, este reconocimiento se encuentra influido por la acción de una gran variedad de factores en trans que son reclutados al lugar donde se está llevando a cabo el *splicing*.

En la primera parte de esta tesis, buscamos analizar la variabilidad de secuencia que presentan los sitios donores a lo largo de diversas especies eucariotas. Para esto construimos un modelo estadístico de máxima entropía para determinar patrones de correlación no triviales entre los distintos pares de posiciones que constituyen los sitios donores de *splicing*, en busca de determinar cuáles de éstos son comunes entre las 30 especies analizadas, y cuáles son característicos de solo algunos grupos. Si bien el proceso de *splicing* conserva un gran número de elementos en común en los organismos eucariotas, en los últimos años múltiples estudios han destacado la posible relevancia funcional de pequeñas diferencias entre las especies. Así logramos establecer patrones de correlación característicos en los sitios donores de *splicing* que distinguen a las especies vegetales de los metazoos y los hongos.

En la segunda parte, nos abocamos al estudio del efecto sobre el *splicing* de PRMT5, un factor que, mediante la metilación de proteínas, participa de la regulación de múltiples procesos moleculares. Se ha propuesto que la acción de PRMT5 podría favorecer el reconocimiento de sitios donores débiles. Con el objetivo de determinar en qué medida el efecto que tiene PRMT5 sobre los patrones de splicing se encuentra influído por señales en cis, se realizaron experimentos de secuenciación de ARN (RNA-Seq) en plantas de *Arabidopsis thaliana* mutantes de PRMT5, pertenecientes a dos ecotipos distintos: Columbia (Col-0) y Landsberg erecta (Ler). De esta manera se buscó analizar el efecto que tiene la mutación ante la variabilidad genética que presentan estos ecotipos. Para poder discriminar los cambios relacionados con variaciones de las secuencias en cis, se analizó también híbridos F1 Col-0 X Ler y Ler X Col-0. Mediante estos análisis se llegó a la conclusión de que una parte importante de los patrones de *splicing* afectados por la mutación de PRMT5 dependen de señales en cis, teniendo una particular relevancia la fortaleza del sitio donor.



**Palabras clave:** *splicing* alternativo, PRMT5, *Arabidopsis thaliana*, bioinformática, genómica.

# Study of cis and trans signals in the determination of alternative splicing patterns


## Abstract

Splicing is present throughout eukaryotic organisms, representing a key part in a large number of regulatory processes. To be carried out, there must be recognition of a series of signals present both within and around the intron. In particular, the splicing donor and acceptor sites define the boundaries of the intron and are critical for its proper recognition. Throughout the genome, these sites present a certain variability in their sequence, which is related to the possibility of regulating their recognition and establishing different patterns of alternative splicing. On the other hand, this recognition is influenced by the action of a wide variety of factors in trans that are recruited.

In the first part of this work, we analyzed the sequence variability of donor sites across various eukaryotic species. We built a maximum entropy statistical model to determine non-trivial correlation patterns between the different pairs of positions that constitute the splicing donor sites, seeking to determine which of these are common among the 30 species analyzed, and which are characteristic of just some groups. Although the splicing process preserves a large number of elements in common between the different species, in recent years multiple studies have highlighted the possible functional relevance of small interspecific differences. Using this approach, we were able to establish characteristic correlation patterns in splice donor sites that distinguish plant species from metazoans and fungi.

In the second part of this thesis, we study the effect of PRMT5 on splicing . PRMT5 is a Protein arginine methyltransferase that participates in the regulation of multiple molecular processes, such as chromatin modification and alternative splicing. It has been proposed that the action of PRMT5 could favor the recognition of weak donor sites. In order to determine to what extent the effect PRMT5 has on splicing patterns is influenced by cis signals, sequencing experiments were performed (RNA-Seq) in PRMT5 wildtype and mutant *Arabidopsis thaliana* plants belonging to two different ecotypes: Columbia (Col-0) and Landsberg erecta (Ler). In this way, we sought to analyze the effect that the mutation has on the genetic variability that these ecotypes have. In order to discriminate the changes related to variations in the cis sequences, F1 Col-0 X Ler and Ler X Col-0 hybrids were also analyzed. Through these analyses, it was concluded that an important part of the splicing patterns affected by the PRMT5 mutation depend on cis signals, having a particular relevance the strength of the donor site.



**Key words:** alternative splicing, PRMT5, *Arabidopsis thaliana*, bioinformatics, genomics.

*«Oh tiempo tus pirámides»*
*La biblioteca de Babel*
**Jorge L. Borges**

# Índice general

# Índice de figuras

# Índice de cuadros

# Agradecimientos

Por más largos que se hagan los años que lleva realizar un doctorado, no dejan de estar constituido de momentos. Y detrás de esos momentos estuvieron muchas personas a las cuales quiero agradecer:

A mi director, Ariel Chernomoretz, por permitirme ser "el biólogo" del grupo. Seguramente nunca más vaya a dibujar tantas veces el esquema de un gen eucariota como en estos años, al contarles qué es el *splicing* a mis compañeros físicos.

A mi co-director, Marcelo Yanovsky, por la pasión que muestra al intentar entender cada arista y cada detalle de los datos.

A toda la unidad de Bioinformática del Instituto Leloir, tanto a los que están como a los que estuvieron, con los que compartí gran parte de esos momentos que hoy llamo doctorado. Tantas charlas, comidas, fiestas, partidos de fútbol, congresos pasaron durante estos años que no alcanzarían las páginas de esta tesis para relatarlas.

Especialmente quiero mencionar a dos personas. Bruno Kaufman, sin él los primeros capítulos de esta tesis no existirían. Y Andrés Rabinovich, a quien le debo gran parte de las cosas que logré aprender en estos años.

A mi familia, mis tías y primos. Especialmente a mi hermana, Mariela, por estar siempre para todo lo que necesite y apoyarme(soportarme) como solo la familia puede hacerlo.

A mis amigos: Thor, Mati, Galle, Cata, Moha, Panther, Marian, Mirko, Lean; por llenar esos otros momentos, los que no están relacionados con el doctorado, desde la época de Tecno.

A Juan Lenardi, mi hermano de diferente familia. Tu apoyo en estos años ha sido inmenso y espero que el futuro nos encuentre en alguna de nuestras charlas infinitas.

Este recorrido podría seguir por otros nombres que, a pesar de no figurar en el papel, no dejan de estar hoy conmigo.

# Dedicatoria

Mi interés por la biología nació el día que me pregunté por qué yo no tenía los ojos verdes de mi mamá, Silvia.

Ella es la razón por la que hoy puedo decir que soy biólogo, pero no solo por eso. Incluso en los momentos más difíciles, siempre se aseguró de que no dejara de estudiar y me apoyó en cada paso que di.

Por eso, a pesar de que hoy ya no esté conmigo en este paso, le dedico esta tesis y todo lo que pueda lograr de aquí en más en esta profesión.

# Contribuciones

Los resultados expuestos en los capítulos 3 y 4 de esta tesis forman parte de un artículo que está siendo evaluado para su publicación.

A partir del trabajo que detallo en el capítulo 5, se está escribiendo un manuscrito para su próximo envío a una revista y realizando los experimentos de validación de los eventos de *splicing* destacados por el análisis de los datos de RNA-Seq.

Además de los resultados relatados a lo largo de esta tesis, durante mi doctorado participé de dos colaboraciones con otros grupos de investigación:

- Con el Laboratorio de Fisiología de Proteínas del Departamento de Química Biológica, FCEN, UBA, a cargo del Dr. Diego Ferreiro; se realizó un estudio en el que se buscó caracterizar los genes que codifican para proteínas que presentan motivos repetitivos en sus secuencias, denominados Ankirinas. Los resultados que se obtuvieron ya han sido plasmados en un manuscrito que espera su próximo envío a una revista.

- Con el Laboratorio de Neurobiología Molecular del IBioBA - CONICET - MPSP, a cargo del Dr. Damián Refojo; se realizó el análisis de datos procedentes de secuenciación masiva de ARN de cultivos primarios de neuronas, en los que se buscó determinar los efectos que a nivel transcriptómico tienen distintos protocolos experimentales utilizados en experimentos de neurobiología. En este momento se están realizando los últimos experimentos necesarios para la posterior escritura del manuscrito correspondiente.

# Abreviaturas más frecuentes

**3'ss:** del inglés *3' splice site*

**5'ss:** del inglés *5' splice site*

**5'Alt, 3'Alt:** dadores/aceptores 5'/3' alternativos

**ADN:** ácido desoxirribonucleico

**ARN:** ácido ribonucleico

**ARNm:** ácido ribonucleico maduro, mensajero

**BAM:** del inglés Binary Alignment Map

**BPS:** del inglés *branching point site*

**Col-0:** Columbia (ecotipo de Arabidopsis thaliana)

**FDR:** del inglés false discovery rate

**Ler:** Landsber (ecotipo de Arabidopsis thaliana)

**PIR:** del inglés *percent intron retention*

**PSI:** del inglés *percent inclusion* o *percent spliced in*

**PRMT:** metil-transferasa de residuos arginina en proteínas

**p-valor:** valor de probabilidad

**RI:** retención de intrón

**RNA-Seq:** secuenciación masiva del ARN

**SE:** salteo de exon

**snRNAs:** del inglés *small nuclear ribonucleic acid*

**snRNP:** del inglés *small nuclear ribonucleoproteins*

**SREs:** del inglés *splicing regulatory elements*

# Capítulo 1

# Introducción

## 1.1. Proceso de splicing

Los procesos moleculares que median entre la transcripción de los ARN mensajeros (ARNm) y su disponibilidad para su uso en la traducción de proteínas, son enormemente complejos, particularmente en los organismos eucariotas. Un conjunto de estos procesos esta involucrado en lo que se denomina como la maduración del precursor del ARN mensajero (pre-ARNm) que será fundamental para que pueda ser exportado del núcleo al citoplasma y alcance los ribosomas, en los cuales se producirá la traducción. Entre estos mecanismos podemos mencionar el agregado de una GTP (guanosina-5'-trifosfato) al extremo 5' del ARNm (capuchón o *cap* en inglés) y hacia el extremo 3' el corte del transcripto y el agregado de una cola de poli-adeninas. Mientras que, por otro lado, a lo largo del transcripto se produce el corte y eliminación de segmentos internos que no serán parte del transcripto maduro. Este proceso se denomina *splicing* y será el eje central de esta tesis.

El *splicing* es un mecanismo que esta íntimamente ligado a la estructura que tienen los genes eucariotas. Éstos se caracterizan por el hecho de que las regiones que formarán parte del transcripto maduro, exones, se encuentran interrumpidas por segmentos que son eliminados como parte del proceso de maduración, y que se denominan intrones. El descubrimiento de los intrones y, por ende, del *splicing* se produjo en forma simultanea por dos trabajos independientes en 1977[12,29], en los que, mediante la hibridación de un gen y el ARNm codi-

ficado por el mismo, se pudo observar por microscopia electrónica (MI) que había segmentos del gen que no estaban presentes en el transcripto maduro (Fig. 1.1). Desde entonces se ha producido una revolución en nuestra comprensión de los mecanismos de regulación genética y las bases moleculares detrás de la evolución de los organismos eucariotas.

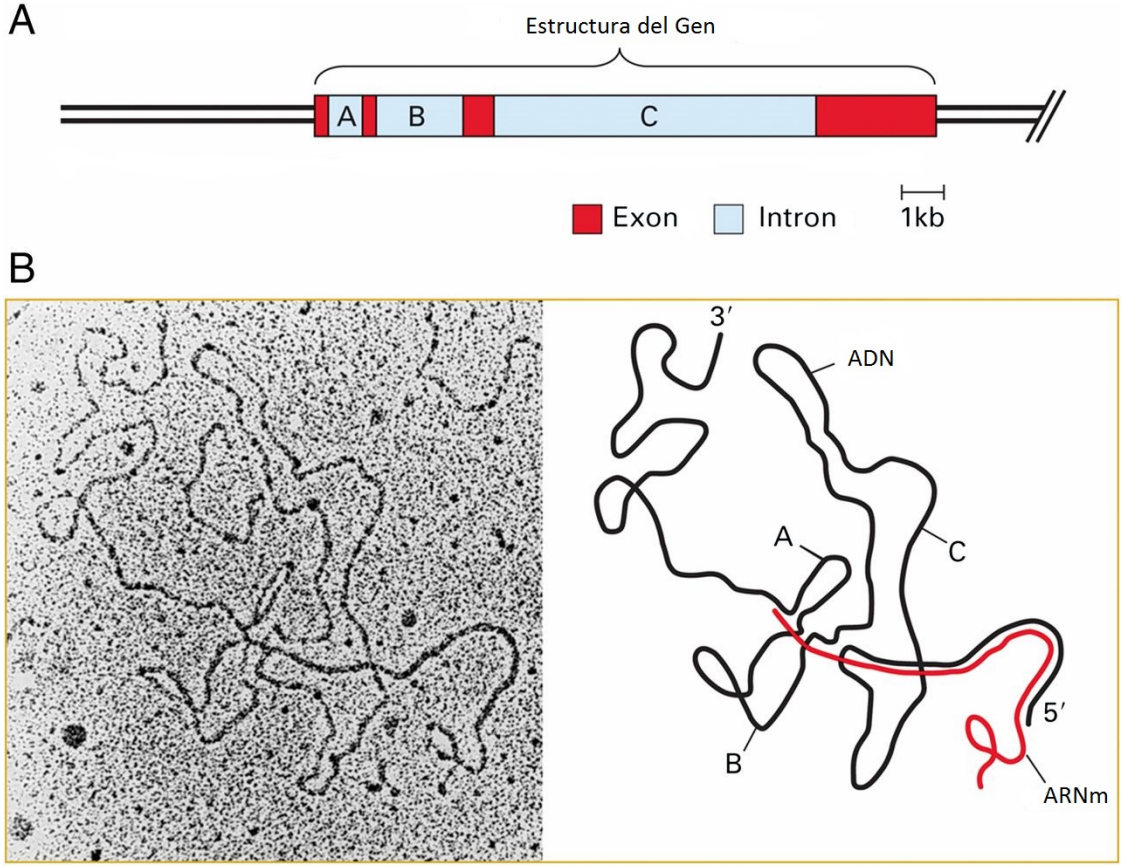



**Figura 1.1**: **Descubrimiento del *splicing*.** *(A) Diagrama de la estructura del gen utilizado en el experimento, en rojo se observa la posición de los exones y en celeste la de los intrones (A, B y C). (B) Imagen de MI del híbrido entre el ARNm maduro y la hebra codificante del gen. (C) Interpretación del resultado, las zonas en las que ARNm (en rojo) es paralelo al ADN (en negro) corresponde a regiones de hibridación, mientras que los segmentos no hibridados A, B y C dan cuenta de la existencia de los intrones. Tomado de Berka et al. 2016*[13]*, adaptado de Berget et al. 1977*[12]*.*

La ventaja evolutiva que genera la presencia de los intrones en los genes eucariotas puede resultar, en un primer momento, difícil de comprender, debido a que su transcripción involucra un alto costo energético para las células, a pesar de que los intrones serán finalmente

eliminados sin involucrarse en la creación de proteínas. Esto se vuelve más paradójico aún si pensamos que la presencia de intrones se extiende a lo largo casi todos los genomas eucariotas (con la excepción de los nucleomorfos[92]) y cuya densidad se vuelve particularmente alta en los organismos más complejos, como las plantas y los animales. Solo por dar un ejemplo, en el caso del *Homo sapiens*, las secuencias intrónicas constituyen un 25 % del total del genoma, siendo entre 4 y 5 veces más de lo que representan los exones[77].

En gran medida esto se explica por el fenoméno de *splicing* alternativo, en el cual el corte y empalme de exones producido durante el *splicing* puede llevarse a cabo a partir de diferentes lugares dentro de un mismo gen. Las secuencias que marcan el límite entre el exon y el intrón son de vital importancia para que se lleve a cabo el *splicing*. En los extremos 5' y 3' del intrón tenemos los sitios donor y aceptor de *splicing*, respectivamente. Ambos sitios presentan una determinada secuencia consenso que permitirá el reconocimiento de los mismos por la maquinaria encargada de llevar a cabo el *splicing*. En los sitios que se alejen de estas secuencias consenso, este reconocimiento dependerá de determinados mecanismos regulatorios que harán que los mismos puedan o no ser reconocidos. Los detalles de la forma en la que estos sitios son reconocidos serán desarrollados con mayor profundidad en los siguientes apartados.

Dependiendo de qué sitios de *splicing* sean reconocidos se van a originar distintos ARN mensajeros a partir de un único gen. Y en el caso de los genes que codifican proteínas, esto puede derivar también en un aumento de los polipéptidos que pueden codificarse a partir de un conjunto dado de genes. El *splicing* alternativo sería el principal mecanismo para aumentar la diversidad proteica, superando a otros como la iniciación de la transcripción alternativa o los sitios de corte y poliadenilación alternativos[56,119].

De esta forma, el *splicing* alternativo representa una capa de regulación que se integra a los demás mecanismos regulatorios presentes tanto a nivel transcripcional y post-transcripcional, para determinar la calidad y la cantidad de las proteínas presentes en un determinado tejido, etapa embrionaria o contexto fisiológico.

La relación entre estas capas regulatorias se hace patente en el hecho de que gran parte del *splicing* se lleva a cabo de manera co-transcripcional[14,36,62], principalmente en los intrones que se encuentran más cerca del extremo 5' del transcripto[82]. Se han presentado dos modelos, no excluyentes entre sí, para explicar la relación que se establece entre la transcripción y el *splicing*[19,88]. Por un lado, el modelo kinético relaciona la velocidad de elongación de la ARN polimerasa II (ARN PolII) con el reconocimiento de los sitios de splicing. Si la velocidad es alta más sitios estarán disponibles a la hora de llevarse a cabo el *splicing*, lo que favorecería el reconocimiento de sitios consenso; mientras que en el caso opuesto, la baja velocidad de elongación podría ayudar a que los sitios de *splicing* más alejados del consenso puedan ser reconocidos ya que no tendrían otros sitios con quienes competir a la hora de producirse el *splicing*. Por otro lado, el modelo de reclutamiento hace énfasis en las interacciones que se dan entre la ARN PolII y distintos factores relacionados con el *splicing*. Muchas de estas interacciones se dan mediante el dominio C terminal de la misma (CTS), la cual está expuesta a un gran número de modificaciones post-traduccionales.

Estos modelos se complejizan aún más si tenemos en cuenta las múltiples interacciones que mantiene la maquinaria molecular relacionada con el *splicing* y la estructura de la cromatina y sus remodeladores. Sin embargo, antes de seguir desarrollando los mecanismos de regulación del splicing, debemos concentrarnos en cuáles son los actores encargados de llevar a cabo el proceso del *splicing*, conformando en conjunto lo que se denomina spliceosoma, y los diversos pasos en los que el mismo se produce.

### 1.1.1. Ciclo de splicing

Hasta el momento se han identificado tres clases diferentes de intrones. Los pertenecientes al Grupo I y II han sido encontrados en las secuencias de ADN de algunas especies de bacterias y de organelas[116], hallándose incluso algunos intrones del Grupo I en los ARN ribosomales de ciertos protistas y hongos[122,125]. Lo que tienen en común los intrones de ambos grupos es que, mediante diferentes mecanismos, logran producir por su propia cuenta

el proceso de *splicing* sin que requieran de la actividad enzimática de una proteína. Sin embargo, el *splicing* de los intrones que pertenecen al Grupo III, que constituyen la inmensa mayoría de los intrones presentes en los genomas nucleares de los organismos eucariotas, necesita para llevarse a cabo la participación de una compleja maquinaria molecular llamada spliceosoma.

Durante los diferentes pasos que conlleva el proceso de *splicing* se pone en juego la participación de 5 ARN pequeños nucleares (snRNA, por sus siglas en inglés) y cientos de proteínas, que conforman distintos complejos moleculares que serán los encargados de llevar a cabo las reacciones químicas necesarias para el corte y empalme de los exones. En cada ciclo de *splicing* (Fig. 1.2), diferentes componentes del spliceosoma realizan el reconocimiento de tres señales conservadas en los intrones: el sitio donor de *splicing* en el extremo 5' del intrón (5'ss, por 5' *splicing site*), la secuencia del punto de ramificación (BPS, *branching point sequence*) y el sitio aceptor de *splicing* en el extremo 3' del intrón (3'ss). Por otra parte, los 5 snRNA mencionados anteriormente forman 5 complejos ribonúcleoproteicos: U1, U2, U4, U5, y U6 snRNP (del inglés, *small nuclear ribonucleoproteins*). Pasemos ahora a realizar un breve recorrido de las diferentes etapas y pasos que conforman un ciclo de *splicing* para entender el rol que adquieren cada uno de estos componentes.

El ensamblado del spliceosoma comienza por el reconocimiento del 5'ss por parte de U1, el mismo se realiza por la complementariedad de bases que se establece entre éste y la porción 5' del snRNA de U1 snRNP[124]. A esto se suma la unión de las proteínas SF1 y U2AF a los sitios BP y 3'SS, respectivamente, para formar el primero de los complejos del ciclo: el complejo E. Tanto las SF1 como U2AF son desplazadas por la U2 snRNP, cuyo snRNA forma hibridiza con el BPS, en la formación de lo que se denomina como pre-spliceosoma o complejo A. En un siguiente paso, se asocia el U4/U6.U5 tri-snRNP en el cual se da una unión por hibridación de los snRNA's de U4 y U6, mientras que U5 establece uniones proteína-proteína con los demás complejos. Con la incorporación de los 5 snRNP's se establece el armado del complejo pre-B, el cual sufrirá una serie de reestructuraciones que llevarán a la

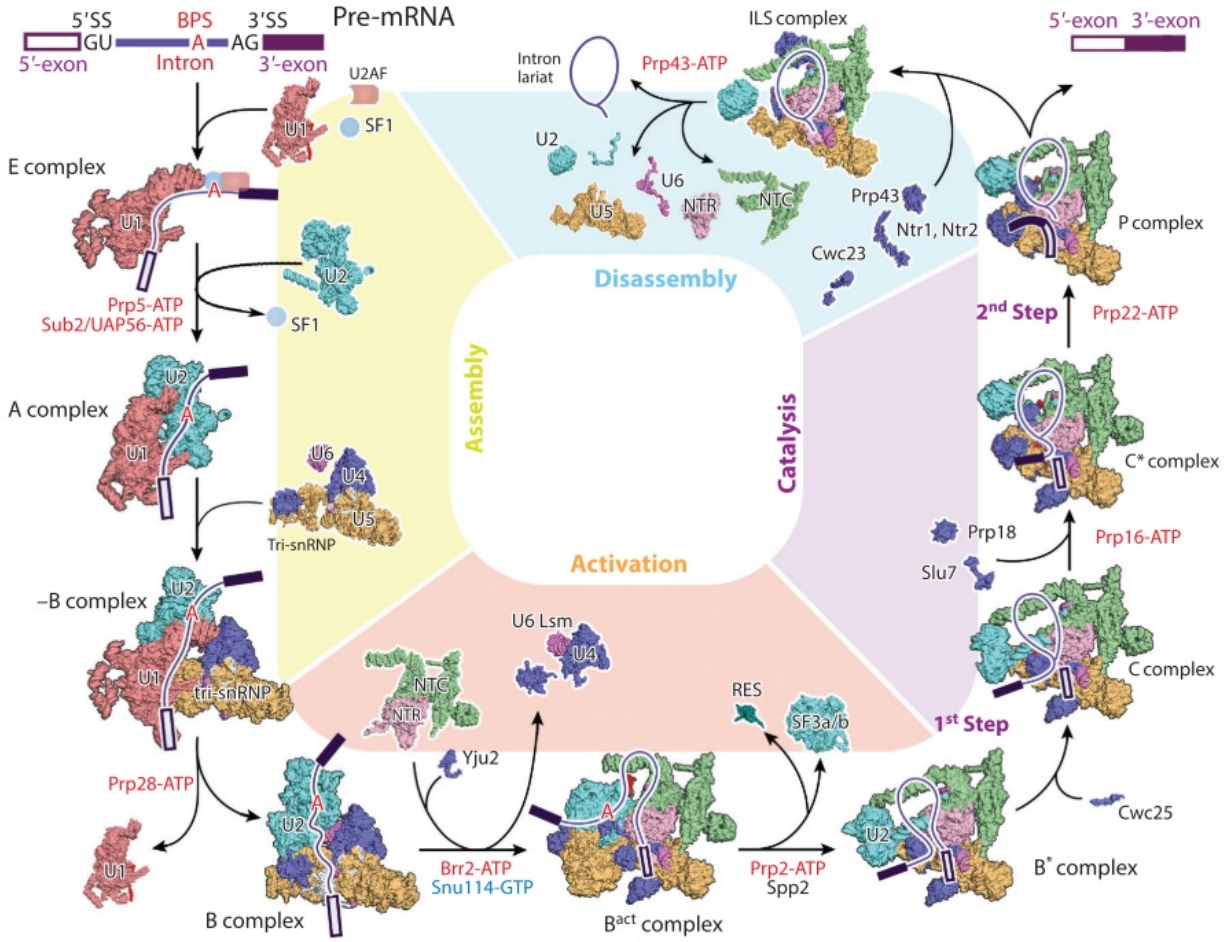



**Figura 1.2**: **Ciclo de splicing**. *Tomado de Wan et al. 2020*[165]

.

activación del sitio catalítico donde tendrán lugar las reacciones químicas necesarias para el corte y empalme de los exones.Hasta este momento el 5'ss se mantuvo hibridado con U1. La primera de las reestructuraciones involucradas en la activación del sitio catalítico implica el desplazamiento de U1 para que el 5'ss pueda pasar a interactuar con U6. Este paso es facilitado por la acción de la ATPasa/helicasa Prp28 y da lugar a la conformación del complejo B, el cual aún no cuenta con la activación del complejo catalítico[16].

Para que se lleve a cabo dicha activación, el U4 snRNP que se mantenía unido con U6 mediante la hibridación de su snRNA debe ser desplazado por los complejos NTC (del inglés, *NineTeen Complex*) y NTR (NTC-related complex). De esta manera se permite la

reorganización de la porción del U6 snRNA que será clave en la reacción de *splicing*, la ISL (del inglés, *Intramolecular Stem Loop*). En segundo lugar, una serie de reorganizaciones del complejo mediadas por la acción de ATPasas/helicasas ubica el BPS en el sitio activo, en la proximidad del híbrido 5'ss/U6 snRNA[164,169]. Así se alcanza el complejo B* con un sitio catalítico activo.

La reacción de *splicing* ocurre en dos pasos: (1) el grupo 2'hidroxilo de una adenosina presente en el BPS ataca el grupo fosfodiester en el 5'ss, produciendo el corte del exón presente en 5' y la formación de una estructura intermedia en la que el intrón en forma de lazo, con el primer nucleótido del intrón unido covalentemente a la A del BPS, sigue unido al exón en 3'. Como resultado del primer paso, tenemos la estructura en forma de lazo en el centro del sitio catalítico, imposibilitando la entrada del 3'ss para proseguir al segundo paso catalítico[164]. (2) El complejo es nuevamente remodelado mediante la acción de Prp16 (pasando de complejo C a C*), permitiendo el posicionamiento del 3'ss en el sitio activo[47,175]. De esta manera se posibilita la reacción de ligamiento de los dos exones, en el que el grupo 3' hidroxilo del exon 5' ataca el grupo fosfodiester del sitio 3', ligando estos exones y permitiendo la liberación del intrón en forma de lazo.

Para completar el ciclo de *splicing*, lo único que resta es liberar los exones ligados de la maquinaria de *splicing*, cuya conformación en este punto se denomina complejo P. Esto se logra mediante la acción de ATPasas como Prp22 y Prp16[176] que liberan, por un lado, los exones del ARNm ya ligados y, por el otro, el intrón en forma de lazo unido a varios componentes del spliceosoma (complejo ILS, por *Intron Lariat Spliceosome*). Finalmente el complejo ILS se desarmará completamente mediante la acción de Prp43, dejando todos los componentes del spliceosoma listos para un nuevo ciclo de *splicing*.

### 1.1.2. Señales en cis

#### Sitios de splicing

Como se vio en la sección anterior, el reconocimiento de las señales presentes en las secuencias que delimitan el intrón y del BP resulta esencial para llevar adelante el proceso de *splicing*. El sitio 5'ss abarca principalmente 9 posiciones en la interfaz exón-intrón: las 3 primeras corresponden a las últimas posiciones del exón río arriba y las siguientes 6 a las primeras del intrón. La señal más importante es el dinucleótido GT en las primeras dos posiciones intrónicas, presente en la enorme mayoría de los 5'ss procesados por el spliceosoma mayor. De manera similar, el 3'ss marca el final del intrón con el dinucleótido AG, también altamente conservado, que es antecedido por una secuencia variable de citosinas y timinas conocida como tracto de pirimidinas. En cuanto a la secuencia relacionada con el BPS, lo más importante a destacar es la presencia de una adenosina que es clave en la primera reacción química en el proceso de *splicing*. En los seres humanos está presente en el 92 % de los BPS[54].

Todas estas señales presentan una secuencia consenso que se mantiene prácticamente invariable a lo largo de las distintas especies de eucariotas, aunque puede observarse diferente grado de conservación dependiendo de la especie considerada (Fig. 1.3). Por ejemplo, en *S. cerevisiae* la secuencia del 5'ss presenta una muy baja variabilidad en su parte intronica, mientras que en el caso de *A. thaliana* ocurre lo contrario. Una situación análoga se presenta para el BPS, el cual presenta una alta conservación en *P. chrysosporium* mientras que en otras las especies se muestra más variable.

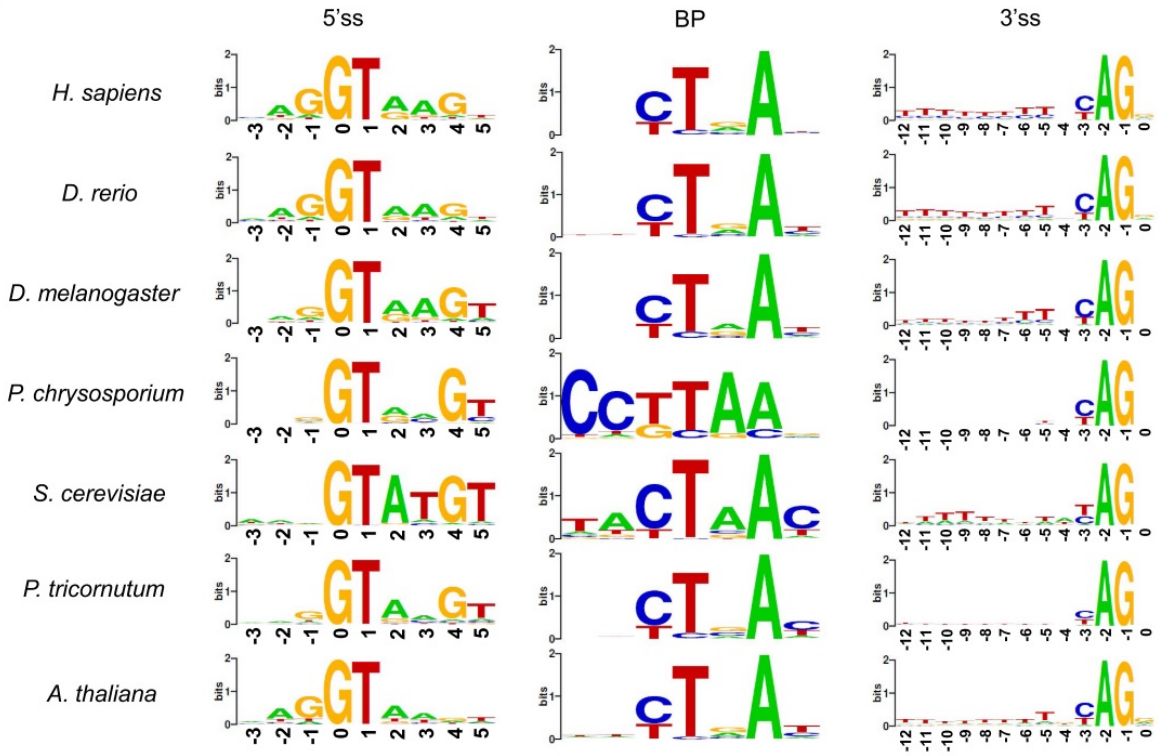



**Figura 1.3**: **Contenido de Información (CI) de los sitios de *splicing* de diferentes especies.** *5'ss: sitio donor de splicing; 3'ss: sitio aceptor de splicing; BP: sitio de ramificación. Tomado de Iwata & Gotoh 2011*[74]

.

El reconocimiento de estas señales adquiere relevancia en distintos puntos del ciclo de *splicing*. Mientras que el 5'ss y el BPS son importantes en el ensamblaje inicial del spliceosoma y en la primer reacción de *splicing* en la que se forma la estructura intrónica en forma de lazo, el reconocimiento del 3'ss es crucial para el segundo paso catalítico, que involucra el ligamiento de los dos exones. En particular, el reconocimiento del sitio 5'ss pasa por varios rearreglos a lo largo del ciclo (Fig. 1.4). Durante el ensamblado del spliceosoma, desde el complejo E hasta el complejo pre-B, el reconocimiento del 5'ss se da por hibridación con una secuencia conservada del extremo 5' del U1 snRNA. El duplex formado por los dos ácidos nucleicos es estabilizado por cuatro componentes del snRNP, Luc7, Yhc1/U1-C, SmD3 y SmB. Estas interacciones se dan principalmente sobre los grupos fosfatos de las cadenas de los ácidos nucléicos. De esta manera, la interacción se da, por un lado, de manera secuencia

específica, mediante la formación del dúplex 5'ss/U1 snRNA, y, por el otro lado, de forma inespecífica por unión al esqueleto de fosfatos de las cadenas de ácidos nucleicos. Durante la activación del spliceosoma (formación del complejo B), la acción de Prp28 disocia a snRNA U1 del 5'ss permitiendo la aproximación de la secuencia ACAGA del snRNA U6 que finalmente llevará a la formación de un duplex 5'ss/U6 en el complejo B*. Con estos rearreglos una zona del snRNA de U5 que antes formaba un duplex con el snRNA de U6, llamada *loop I*, es liberada y forma una unión con las posiciones exónicas del 5'ss; la cual se mantendrá hasta la finalización de la etapa catalítica del ciclo.

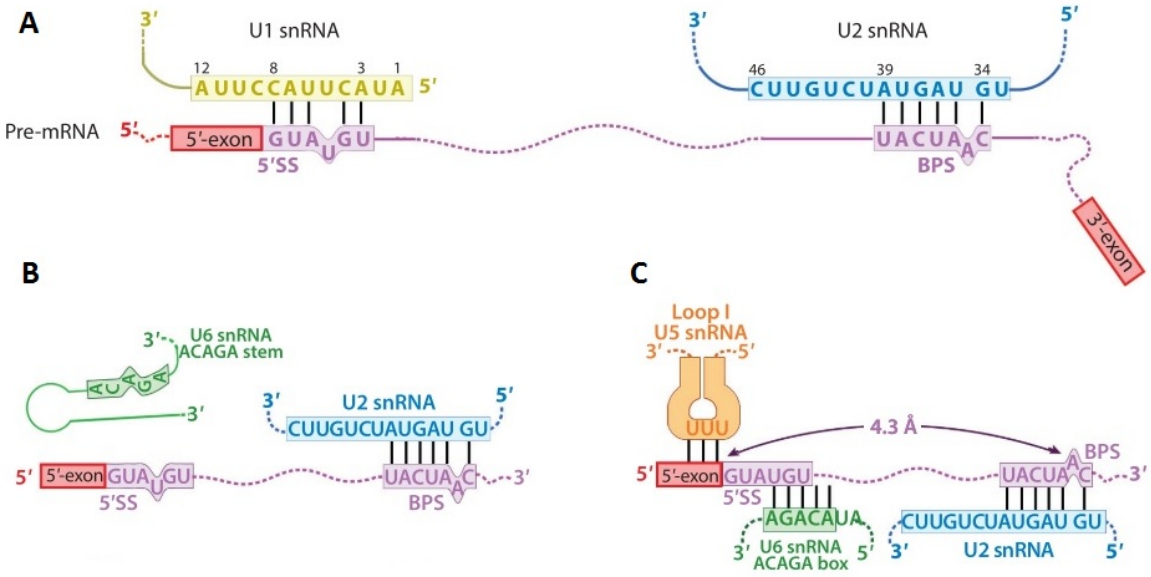



**Figura 1.4**: **Reconocimiento del 5'ss y BPS durante el ciclo de splicing.** *Adaptado de Wan et al. 2020*[165]

.

Además del proceso de *splicing* tal como fue descripto en las secciones anteriores, existe en varias especies eucariotas una variante del mismo llevada a cabo por lo que se denomina como spliceosoma menor. Éste presenta una serie de diferencias en cuanto a los componentes moleculares que lo conforman. Las snRNPs U1, U2, U4 y U6 son reemplazas aquí por U11, U12, U4atac y U6atac, respectivamente. Sin embargo, solamente 7 proteínas resultan ser específicas de este spliceosoma y están presentes en los complejos U11 y U12 que, a diferencia

del spliceosoma mayor, están presentes en el núcleo como un di-snRNP U11/U12. En el ciclo de *splicing* no se observan grandes cambios respecto al realizado por el spliceosoma mayor, aunque varios estudios marcan que en este caso el proceso se realiza de manera más ineficiente, lo que podría tener un alto impacto en la regulación de la expresión de los transcriptos cuyo procesamiento depende de este spliceosoma. Los intrones procesados por este spliceosoma, denominados intrones U12, representan solamente el 0.5 % del total de intrones presentes en un genoma. Al principio se pensó que lo que identificaba a estos intrones era que estaban delimitados por los di-nucleótidos AT-AC, a diferencia del resto de los intrones donde predomina GT-AG. Sin embargo, luego se mostró que los intrones procesados por el spliceosoma menor no necesariamente tenían que tener esos extremos y que lo que en verdad diferenciaba a sus señales de *splicing* era un mayor grado de conservación en los 5'SS y en el BPS[162] (Fig. 1.5). Al mismo tiempo que los intrones con extremos AT-AC también podían ser target del spliceosoma mayor.

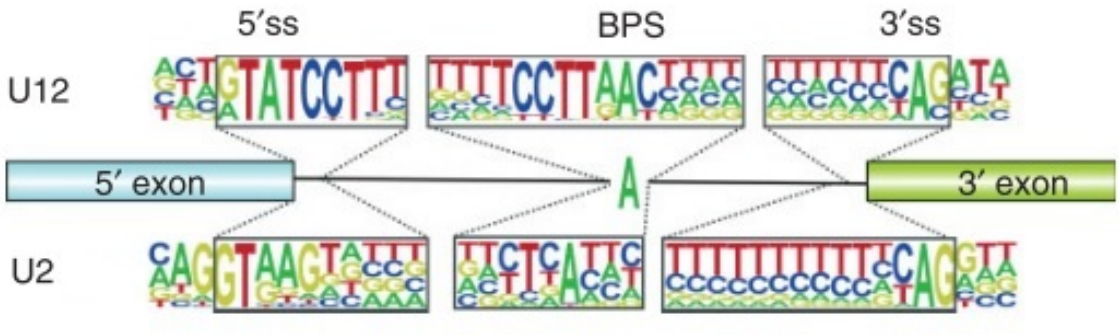



**Figura 1.5**: **Secuencias consenso para las señales de *splicing*** *intrones procesados por el spliceosoma menor, U12, y mayor, U2. Modificado de Turunen et al. 2013*[162].

### Elementos regulatorios en cis

Además de las señales de *splicing* vistas hasta ahora, se encontraron otras señales presentes en los transcriptos que, en ciertos casos, son importantes para el reconocimiento de los sitios de *splicing* y la determinación de los patrones de *splicing* alternativos. Estas secuencias se denominan SRE (del inglés *splicing regulatory element*) y pueden promover - en cuyo caso se les dice *enhancers* - o inhibir - *silencers* - el proceso de splicing. Dependiendo en qué parte se encuentran estas secuencias podemos tener: ESE (*exon splicing enhancer*), ESS (*exon splicing silencers*), ISE (*intrón splicing enhancer*) o ISS (*intrón splicing silencers*) (Fig. 1.6). Estas secuencias son reconocidas por proteínas reguladoras del *splicing* que establecen interacciones con diferentes componentes del spliceosoma favoreciendo o desfavoreciendo el desarrollo del ciclo de *splicing*.

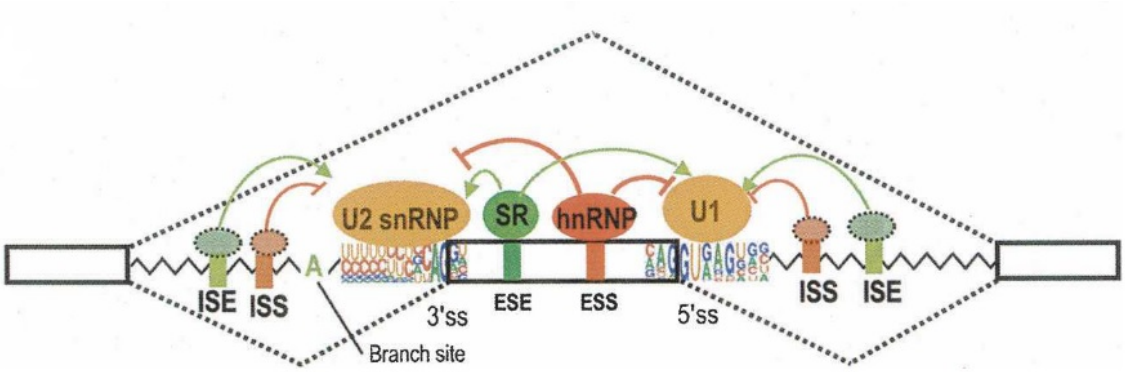



**Figura 1.6**: *Esquema de las señales en cis involucradas en la regulación del splicing. Tomado de Wang et al. 2008*[167].

Los motivos que pueden conformar los SRE son muy diversos. Un ejemplo bien caracterizado de ISE son las repeticiones de G ($G_n$; $n \geq 3$) que si se encuentran en las cercanías de los 5'ss y 3'ss pueden favorecer su reconocimiento. Otro ejemplo son las repeticiones CA dentro de las regiones intrónicas o variaciones de la secuencia UGCAUG, la cual constituye un importante ISE en determinados tejidos, como cerebro y músculo.

La función de las SRE suele tener un efecto aditivo por lo que al aumentar el número de copias de las mismas aumenta también su efecto sobre el splicing. Esto puede ser debido

tanto por un incremento en la afinidad por el factor de *splicing* como por un aumento del número de factores que son reclutados a la zona. Por otro lado, su efecto también se ve alterado dependiendo de la ubicación que el SRE tenga dentro del pre-mARN. Mientras que los $G_n$ funcionan como *enhancers* del *splicing* cuando están ubicados en los intrones, también pueden actuar como *silencers* si se encuentran dentro del exón[43]. Otra posible diferencia puede darse en algunos casos si el efecto del factor de *splicing* depende de la distancia que el mismo tenga respecto a los sitios de *splicing*. Todas estas variables provocan que el efecto que puede tener la presencia de un SRE sea dependiente del contexto en el que el mismo se encuentra[51].

### 1.1.3. Efectos regulatorios en trans

#### Factores de splicing

Las proteínas relacionadas con el reconocimiento de los SRE provienen principalmente de dos familias de proteínas: las SR y las hnRNP. Ambas familias contienen un gran número de representantes en las distintas especies eucariotas, siendo piezas claves tanto de la expresión génica como en el procesamiento del ARN. En el Cuadro 1.1, se muestra algunos de los factores de *splicing* y los motivos que reconocen deducidos a partir de la técnica CLIP (del inglés *crosslinking and immunoprecipitation*).

Las proteínas SR se caracterizan por la presencia de un dominio hacia su extremo C-terminal enriquecido en los aminoácidos serina (S) y arginina (R) que, por este motivo, se denomina dominio RS; mientras que en su extremo N-terminal se encuentra un dominio de reconocimiento de ARN (dominio RRM). El dominio RS participa de un gran número de posibles interacciones proteína-proteína y proteína-ARN, mediante las cuales produce el reclutamiento de los componentes del spliceosoma a los sitios de *splicing*, favoreciendo así su reconocimiento. De esta forma, las proteínas SR suelen cumplir una función de activación del *splicing*. Sin embargo, de la misma forma que se vio con los SRE, su efecto depende del contexto. Si la unión de una proteína SR se da dentro de la región exónica (mediante

| Factor de *splicing* | Motivo |
|---|---|
| Proteínas SR | |
| SRSF1 | GGAGGA |
| SRSF2 | SSNG |
| SRSF3 | Rico en Pirimidinas |
| SRSF4 | Rico en Purinas |
| Proteínas hnRNP | |
| hnRNP A1 | GNNAGN |
| hnRNP A2/B1 | GGUAGU |
| hnRNP C | Rico en Pirimidinas |
| hnRNP F | Rico en GU |
| hnRNP H1 | $G_n$ |
| hnRNP L | Rico en CA |
| hnRNP M | GGUGG |
| hnRNP U | Rico en GU |
| Otras | |
| CUGBP1 | UGUU |
| MBNL1/L2 | YGCY |
| NOVA | YCAY |
| RBM4 | CGG o CTAACG |
| RBM5 | UCAUC o UGUAA |

**Cuadro 1.1**: *Factores de splicing y sus motivos de reconocimiento identificados mediante la técnica de CLIP (del inglés crosslinking and immunoprecipitation). En los motivos: Y representa una C o U; S representa una C o G; y N cualquier nucleótido. Tomado de Fu et al. 2014*[51].

motivos ESE), esto conduce favorecer el splicing, mientras que si se da dentro del intrón conduce a su supresión[145]. Pero también se ha visto que su efecto depende de su interacción con otras proteínas de unión a ARN (RBP, por sus siglas en inglés), pudiéndose establecer un efecto sinérgico entre las mismas[65].

Las proteínas hnRNP, cuyo nombre deriva de ribonucleoproteínas heterogeneas nucleares (en inglés *heterologeneous nuclear ribonucleoproteins*), son caracterizadas comúnmente como reguladores negativos del splicing. De la misma forma que ya fue discutido para el caso de las proteínas SR, este efecto depende del contexto en el cual se esté dando. La represión del *splicing* por parte de las hnRNPs no parece interferir con el reconocimiento inicial de los

sitios de *splicing* pero lleva a la formación de complejos no productivos del spliceosoma que impiden completar el ciclo de splicing[43], aunque los detalles de estos mecanismos no están completamente resueltos.

### Efectos de la cromatina

Como ya vimos al comienzo de esta Introducción, el hecho de que gran parte del *splicing* se realice de manera co-transcripcional, lo inserta en una compleja red de regulaciones que lo vincula con la transcripción pero también con los múltiples mecanismos regulatorios que involucran modificaciones epigenéticas, como la estructura y remodelación de la cromatina.

La estructura exón-intrón de los genes eucariotas, se ve correlacionada con una diferenciación de las marcas epigenéticas: en las regiones exónicas hay un mayor nivel de ocupación de nucleosomas, de metilación del ADN y de la presencia de ciertas marcas de histonas, como H3K36me3 y H3K27me2. Especialmente se ha encontrado una mayor densidad de nucleosomas en exones que se encuentran flanqueados por intrones largos y que presentan sitios de *splicing* débiles, lo que conduce a pensar que éstos serían importantes para el reconocimiento de los exones por parte de la maquinaria de *splicing*[151,158]. Algo similar ocurre con la metilación del ADN, se ha observado que los exones alternativos, que pueden o no ser incluidos en el transcripto final mediante *splicing* alternativo, presentan un nivel menor de metilación que los exones constitutivos[57]. Finalmente, las marcas epigenéticas de las histonas también pueden cumplir un rol importante a la hora de definir los patrones de splicing. Por ejemplo, la proteína remodeladora de la cromatina Chd1 que reconoce la metilación de la H3K4, se encuentra involucrada en el reclutamiento de diversos factores involucrados en el *splicing*. Esto conduce a que frente a un *knockdown* de Chd1, disminuya la asociación de los componentes de snRNP U2 a la cromatina y la eficiencia del procesamiento del ARN se vea comprometida[147].

Sin embargo, la regulación que estos procesos realizan es recíproca ya que el *splicing* también ejerce un efecto importante tanto en la transcripción como en la estructura de la cromatina. En varios trabajos se ha observado que la presencia de intrones tiene un efecto

positivo sobre la expresión génica[28,120] y que el *splicing* favorece tanto la iniciación como la elongación de la transcripción[35,99]. Por otro lado, el *splicing* también se vio relacionado con la presencia de la tri-metilación de la H3K36, la cual disminuye notoriamente si el *splicing* es inhibido o en el caso de genes sin intrones[15].

## 1.2. *Splicing* alternativo y sus efectos regulatorios

El *splicing* alternativo (SA) que, como vimos anteriormente, permite la obtención de más de un transcripto maduro a partir de una secuencia génica, puede ser clasificado en diferentes tipos dependiendo de los sitios de *splicing* que son reconocidos (Fig. 1.7). En primer lugar, encontramos el 5' y 3' alternativos donde se da la presencia de más de un sitio, 5'ss o 3'ss respectivamente, que se encuentran en competencia, es decir, que puede ser reconocido uno u otro bajo determinadas circunstancias. En estos casos se da una alteración en el largo del exón que será parte del transcripto maduro. Por otro lado, podemos tener el caso de un exón que no sea reconocido por la maquinaria de *splicing*, con lo cual será excluido del ARNm. Los exones salteados (SE) pueden ser uno o más y puede darse el caso de exones contiguos que sean mutuamente excluyentes, es decir, que en un transcripto dado solo pueda estar uno u otro pero nunca los dos. Finalmente, cuando ningún sitio de *splicing* que flanquea a un determinado intrón es reconocido por el spliceosoma lo que se produce es la retención del intrón (RI), lo que implica que la secuencia del mismo se conservará en el transcripto luego de su procesamiento. Un caso especial de retención de intrones lo constituyen los exitrones, que son intrones que se encuentran en la región interna de exones que codifican para proteínas[107].

La proporción con la que estos tipos de *splicing* alternativo ocurren en las diferentes especies de eucariotas es bastante variable. Mientras que en los animales el tipo de SA más común es el salteo de exones, llegando en el caso de los humanos alrededor del 40 % del total de eventos; en las especies vegetales este tipo de SA no alcanza el 10 %, siendo ampliamente superado en abundancia por los eventos de retención de intrones, que cuentan

con una frecuencia de entre 28 y 64 %[24,42,79,104]. En humanos solo el 5 % de los eventos de *splicing* alternativo corresponden a un IR[81,129], aunque éste puede llegar a ser importante ciertas enfermedades[75,78,172] y tejidos[10,18].

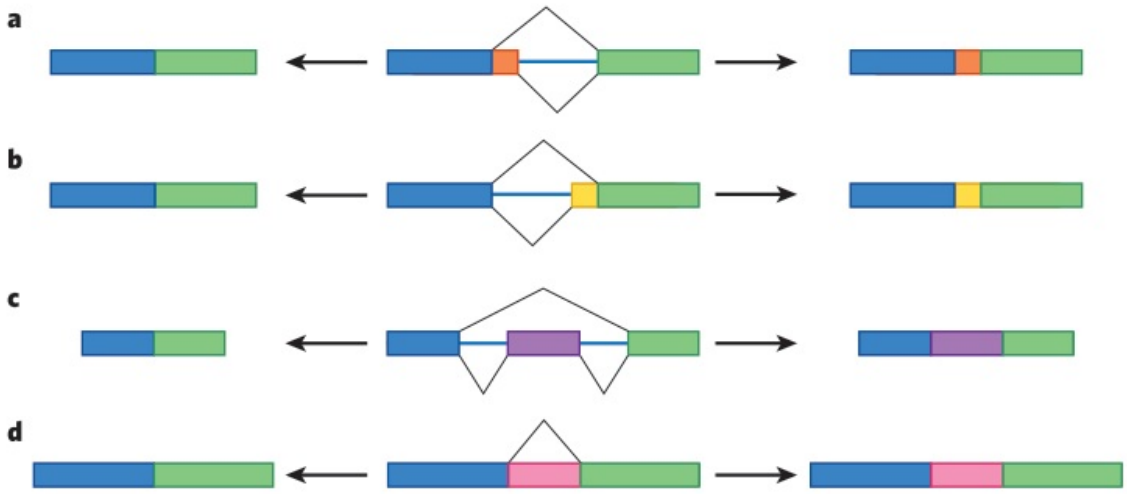



**Figura 1.7**: **Clasificación de eventos de *splicing*.** *(a) 5' alternativo; (b) 3'alternativo; (c) Salteo de exón; y (d) Retención de intron. En la región central se muestra el pre-ARNm, hacia la izquierda y derecha se muestra los transcriptos maduros con y sin la región alternativo, respectivamente. Tomado de Nilsen & Graveley 2010*[118].

Los diferentes tipos de *splicing* alternativo también están relacionados con diferentes consecuencias que tiene este proceso en el flujo de información genética.Una de las principales funciones del SA que se había mencionado es la de aumentar el número de polipéptidos que pueden producirse a partir de un dado conjunto de genes, esto conlleva un aumento en la complejidad de los organismos, que se ve reflejado en la diversificación de los tejidos y de los procesos de regulación, sin la necesidad de aumentar proporcionalmente el tamaño de su genoma. Las variaciones introducidas por el *splicing* alternativo pueden traducirse a nivel proteico de múltiples maneras. Si la inclusión o exclusión de la región exónica alternativa, sea ésta por salteo de exones o por 5'ss o 3'ss alternativo, mantiene el marco de lectura del resto de la proteína se puede generar una variante proteica que tenga o no un determinado dominio funcional o que el mismo se encuentre interrumpido. Por otro lado, si el evento de

SA modifica el marco de lectura del transcripto puede generar proteínas que se encuentren truncadas. En todos estos casos se puede dar no solo una diferenciación funcional de las proteínas resultantes, sino también de las posibles interacciones proteína-proteína, proteína-ADN o proteína-ARN que éstas puede llevar a cabo, reconfigurando las redes de regulación celular.

Los experimentos de transcriptómica, ligados con tecnología de secuenciación masiva como RNAseq, han mostrado que la gran mayoría de los genes multi-exónicos de los organismos más complejos tienen más de una isoforma, alcanzando el 70 % en especies vegetales[24], mientras que en el caso de los seres humanos humanos casi todos los genes que contienen intrones presentan más de un transcripto bajo ciertas condiciones[113,121,167]. Si bien existen ejemplos bien documentados de las consecuencias que a nivel proteico tiene el *splicing* alternativo, aún es una incógnita cuánto de los eventos de SA que se ven a nivel transcriptómico contribuye efectivamente en un aumento de la diversidad del proteoma. Algunos autores afirman que esta contribución puede ser menor a lo que podría esperarse a partir de los resultados de transcriptómica. Estudios de proteómica han detectado que en la mayoría de casos, cada gen tiene una única isoforma proteica principal[1]. Los exones alternativos que sí estan bien documentados a nivel proteicos suelen tener un efecto menor en la estructura de la proteína y estar conservados a lo largo de distintas especies. En cambio, la mayoría de los exones alternativos detectados en transcriptómica presentan una baja presión de selección y una gran acumulación de variaciones de alto impacto[161]. Parte del desacuerdo entre los datos surgidos de estudios de transcriptómica y los obtenidos mediante proteómica, pueden deberse a varios factores, entre los cuales podemos mencionar las limitaciones que pueden darse a la hora de detectar isoformas proteicas minoritarias cuya expresión se daba bajo un contexto muy determinado o en tejidos específicos. Esto se suma al hecho de que gran parte de los polipéptidos surgidos por SA que pasan ciertos criterios de calidad puedan ser eliminados por la célula antes de que puedan ser detectados. Otros estudios que intentan integrar métodos de transcriptómica y proteómica han obtenido mejores acuerdos entre los

mismos, concluyendo que efectivamente el SA tiene un rol importante en la composición y diversidad del proteoma[100,170].

Más allá de estas discusiones, las consecuencias del *splicing* alternativo no se limitan solamente al aumento del proteoma. En la sección anterior, al discutir la influencia mutua que tienen el *splicing*, la estructura de la cromatina y la transcripción, tuvimos un buen ejemplo de esto. Sin embargo, la regulación que ejerce el SA sobre la expresión genética se da de múltiples maneras. La retención de intrones juega un rol fundamental en muchos de estos casos. Los transcriptos maduros que contienen un intrón retenido pueden tener diferentes destinos dependiendo de la ubicación que este intrón tiene dentro de su secuencia y de las señales presentes en el mismo. En primer lugar, la retención de un intrón puede derivar en el secuestro de los transcriptos en el núcleo, lo que puede ocasionar su degradación pero también puede llevar a su acumulación y posterior liberación frente a determinadas señales. Esto puede llegar a ser un mecanismo de respuesta rápida frente a estímulos estresantes como se ha observado en plantas. En segundo lugar, los eventos de IR han sido relacionados con el mecanismo de degradación de proteínas denominado NMD (del inglés *Non-sense Mediated Decay*) mediante la inserción de un codón de terminación prematuro en la secuencia (denominado PTC, del inglés *Premature Termination Codon*). También los RI pueden afectar la traducción mediante la inserción en el extremo 5'UTR (región no traducida del extremo 5' del ARNm) de un uORF (marco de lectura abierta en la zona UTR), o aumentar la eficiencia de traducción mediante lo que se denomina como IME (del inglés *Intron-Mediated Enhancement*). Finalmente, y como ya vimos al hablar del caso de los exitrones, la retención de un intrón también puede generar una variante protéica. Incluso no en todos los casos en los que el intrón retenido presente un PTC éste conducirá a NMD, sino que también puede derivar en la producción de proteínas truncadas, con efectos funcionales importantes.

Lo dicho hasta aquí sirve para dar cuenta del papel central que cumple el proceso de *splicing* en los diversos procesos de regulación de la expresión génica y en el establecimiento y diversidad de las redes de interacción proteica. Si tomamos en cuenta que el mismo está

presente en prácticamente todos los organismos eucariotas, nos lleva a la pregunta por la manera en la que el *splicing* evolucionó y el grado de conservación que mantiene en los diferentes grupos taxonómicos.

## 1.3. Evolución del proceso de *splicing* y su diversidad en eucariotas

El origen evolutivo de los intrones es aún hoy tema de debate entre los especialistas, surgiendo una gran cantidad de hipótesis que lo pueden ubicar en la etapa de la vida previa al uso del ADN como soporte de la información genética (llamada como teoría del ïntrón primero", *intrón first* en inglés) o posterior a la incorporación del ADN en el flujo de información pero previo al surgimiento de los eucariotas (teoría del intrón temprano, *intrón early* en inglés) o que los intrones son una innovación surgida a partir de la aparición de los eucariotas (teoría del intrón tardío, *intrón late* en inglés); además de una serie de teorías que ubican el origen de los intrones en algún punto intermedio entre los momentos ya mencionados. Si bien cada una de estas teorías tiene evidencia que la apoya, cualquier opción que establezca el origen de los intrones antes del surgimiento de los eucariotas debería postular que los mismos se perdieron tanto en bacterias como en los distintos linajes de archaeas, lo que resulta poco probable[163].

La gran cantidad de similitudes que hay entre los intrones procesados por el spliceosoma y los intrones auto-catalíticos del grupo II lleva a pensar que los primeros derivan de estos últimos. Estas similitudes no se limitan solamente a los detalles del mecanismos mediante el cual ambos tipos de intrones llevan a cabo el *splicing*, sino también, en comparaciones tanto funcionales como estructurales entre los distintos dominios presentes en los intrones del grupo II y los snRNA que forman parte del spliceosoma. La principal hipotesis del surgimiento de los intrones spliceosomales propone que estos surgieron a partir de intrones del grupo II de proteobacterias simbiontes (que luego darían lugar a las mitocondrias) de LECA (último ancestro común eucariota), que invadieron los genes nucleares y, eventualmente, perdieron

su capacidad auto-catalítica y, por ende, necesitaron para su procesamiento de factores en trans, dando lugar al surgimiento del spliceosoma. Más allá de esta hipótesis, no necesariamente todos los intrones actuales derivan de este proceso, sino que se han formulado varios mecanismos endógenos mediante los cuales se puede adquirir nuevos intrones. Entre ellos, podemos mencionar la acción de transposones, duplicación interna de los genes o el proceso de "intronización" de secuencias traducibles, relacionado con la ventaja evolutiva que supone la eliminación de PTCs en las secuencias que codifican proteínas. A estas observaciones hay que sumar la evidencia aportada por el estudio de las secuencias de los sitios de *splicing*. Sverdlov[155] ha identificado que intrones "antiguos" presentan un mayor grado de conservación en la parte intrónica de los sitios de *splicing*, mientras que en los intrones "jóvenes" ocurre lo contrario, es la parte exónica la que presenta una mayor conservación. Esto ha sido interpretado como evidencia de que a lo largo de la evolución hubo una migración de la señal desde la parte exónica hacia la parte intrónica de la misma. Lo cual es compatible con la existencia de "proto-sitios de splicing" a partir de los cuales los intrones pueden ser insertados en una secuencia.

Estrechamente vinculado con el surgimiento de los intrones spliceosomales, se encuentra también la pregunta por el origen del spliceosoma. Estudios filogenéticos indican que ya el último ancestro común de todos los eucariotas tenía un spliceosoma complejo, compuesto por los 5 snRNA y al menos 80 proteínas[30]. Por otro lado, hay evidencia de que el spliceosoma menor también se originó muy temprano en la evolución de los eucariotas[134]. Esto indicaría que el spliceosoma de LECA sería similar, tanto en composición como en función, con el que presentan hoy en día los distintos linajes eucariotas. Más allá de esta similitud, a lo largo de la evolución se han dado tanto procesos de complejización como de simplificación del spliceosoma. Mientras que en plantas y animales se registra una gran expansión de los componentes del spliceosoma y de los factores que regulan el *splicing*, en algunos linajes de protistas y hongos se han perdido algunas de las subunidades del spliceosoma o, incluso, éste se ha perdido por completo, como es el caso de algunas especies de microesporidias[5],

diplomonadas[8] y los nucleomorfos[92].

Como es de esperarse, esta simplificación o pérdida del spliceosoma correlaciona con una reducción en el número de intrones que presentan los genomas de estas especies. De manera inversa, muchos intentos de reconstruir las posibles características que presentó el genoma de LECA llegan a la conclusión de que contaba con una alta densidad de intrones. Comparable con la de los genomas modernos de plantas y animales, al mismo tiempo que supera la que presentan los protistas actuales[32,86,132]. Muchas variables relacionadas con el *splicing* parecen estar correlacionadas: organismos con genomas ricos en intrones suelen presentar también una gran cantidad de sitios de *splicing* débiles y una alta tasa de *splicing* alternativo[72,73]. Así especies con baja cantidad de intrones tienen a su vez sitios de *splicing* fuertes y una baja incidencia de SA, como es el caso de *Saccharomyces cerevisiae*. Estos indicios llevan a pensar que el *splicing* alternativo tiene un origen temprano, estando ya presente en el ancestro común de plantas, animales y hongos. Cumpliendo ya un rol importante en los organismos unicelulares que precedieron el surgimiento de la multicelularidad, donde el *splicing* alternativo también pudo haber tenido una importancia capital[141].

A pesar de que los principales pasos del ciclo de *splicing* son similares en los distintos eucariotas[30,80], algunas diferencias en cuanto a factores de transcripción y *splicing*, la arquitectura de los genes y divergencia en las secuencias utilizadas, sugieren que ciertas características del *splicing* y su regulación pueden ser específicas de ciertos linajes. Como consecuencia de estas diferencias, se ha observado que los intrones provenientes de animales no son apropiadamente removidos cuando se los transfecta a células vegetales[20,61,67]. De una forma similar, muchos intrones de mamíferos no pueden ser reconocidos por *Saccharomyces cerevisiae*[11,93]. A esto se suma el hecho de que el spliceosoma de varios linajes ha sido poco estudiado. Hasta hace poco tiempo no había sistemas *in vitro* para caracterizar el spliceosoma de las plantas, quedando aún sin conocerse muchos de los detalles del mecanismo de *splicing* en esos organismos[24].

Como ya se ha mencionado, las plantas y los animales difieren en cuanto a cuál es el

tipo de SA predominante: retención de intrones o salteo de exones, respectivamente. Varios factores pueden ayudar a explicar estas diferencias, entre los que se encuentra el largo promedio de los intrones. Mientras que en el caso de los intrones animales el largo promedio se encuentra en los 5kb, en las plantas los intrones son significativamente más pequeños, con un largo promedio de 160pb[74,106], lo que sugiere que el reconocimiento de los intrones y exones puede diferir en éstas últimas[129]. Por otro lado, estas diferencias relacionadas con el tipo de SA predominante pueden vincularse con diferencias en cuanto a qué función está cumpliendo el *splicing* en estos organismos. Al estar principalmente relacionados con los procesos de NMD y secuestro de transcriptos en el núcleo[48,79], los eventos de retención de intrones no representan un aporte a la diversidad proteica[60,66]. De esta manera, los eventos de RI podrían ser parte de una respuesta rápida de regulación de los niveles de ciertos transcriptos, particularmente durante el proceso de desarrollo de las plantas y las respuestas a estímulos del ambiente[76,129].

Muchos factores de *splicing* están involucrados en respuesta a estrés en plantas[89], incluso componentes del spliceosoma como SKIP[46], SAD1[34,174], LSm4[179], STA1[94] y RBM25[26]. Otro ejemplo importante son las proteínas SR[23,25,178] que presentan 19 miembros en el genoma de *Arabidopsis thaliana*, casi el doble de los que encontramos en humanos[79]. La presencia de múltiples parálogos se da en varias familias de reguladores de *splicing*[85,166]. Esto pudo deberse a duplicaciones del genoma[3,33] y está relacionado con el hecho de que muchas mutaciones en los componentes del spliceosoma no son letales como en otros linajes, aunque pueden tener efectos a nivel de desarrollo y fisiológico[152]. Por otro lado, cabe destacar que estas duplicaciones también pueden ser el puntapié para una progresiva diferenciación funcional de las proteínas parálogas.

Aunque las secuencias consenso de las principales señales de *splicing* son comparables en plantas y animales[128,146], éstas son susceptibles a cambios evolutivos. Iwata y Gotoh[74] analizaron cinco características relacionadas con el *splicing*: el contenido de información de los 5'ss, 3'ss y BPS, el largo del intrón y la composición nucleotídica dentro de los

introns en 61 especies de eucariotas. Encontraron que estas características reflejaban las relaciones filogenéticas entre las especies, especialmente los 5'ss y 3'ss. Trabajos recientes han destacado la significancia evolutiva que pueden llegar a tener las secuencias no consenso. Las combinaciones de sitios de *splicing* no consenso más comunes son GC-AG y AT-AC (considerando el dinucleótido con el que empieza y termina el intrón). La frecuencia con la que se dan las combinaciones de sitios de *splicing* decrece a medida que diverge más de la combinación canónica GT-AG[126]. Sin embargo, pueden observarse algunas diferencias entre las especies en cuanto a la prevalencia de algunas combinaciones de sitios no consenso. Por ejemplo, la combinación GA-AG parece estar particularmente representada en *Eurytemora affinis* y *Oikopleura dioica*[50]. A pesar de su baja abundancia, algunas combinaciones de sitios de *splicing* presentan una alta conservación a través de distintas especies vegetales, lo que podría ser un indicador de su relevancia funcional[126]. En consonancia con esto, algunos sitios de *splicing* no canónicos parecen ser más frecuentes ante condiciones de estrés[2,126]. Por otro lado, la frecuencia de la combinación GC-AG es el doble en plantas respecto a animales[126]. Los trabajos citados analizan las combinaciones de di-nucleótidos presentes en los sitios de *splicing*, las restantes posiciones suelen ser mucho más variables, lo que se refleja en un contenido de información más bajo[74]. Sin embargo, esto no quiere decir que no sean relevantes a nivel funcional. Por ejemplo, la disautonomía familiar está estrechamente vinculada con un patrón de *splicing* aberrante en el exón 20 del gen IKBKAP debido a un mal apareamiento de la posición +6 del 5'ss con el U1 snRNA. Se ha observado que un apareamiento correcto en la posición -1 puede favorecer la aparición de un patrón de *splicing* normal, mostrando la importancia e interdependencia de estas dos posiciones en este caso[22].

## 1.4. Estudio del *splicing* alternativo mediante tecnologías de secuenciación

Los grandes avances que desde el cambio de siglo han tenido las tecnologías de secuenciación permitieron el desarrollo de un sinfín de aplicaciones que profundizaron y, en gran

medida, redefinieron el estudio de los mecanismos de biología molecular. El surgimiento de la secuenciación masiva de ARN, denominada *RNA-Seq*, revolucionó el campo de la transcriptómica, ampliando enormemente las posibilidades de estudiar a nivel sistémico las diferencias tanto cualitativas como cuantitativas de los transcriptos presentes en un determinado organismo, tejido o, incluso al día de hoy, célula. Si bien en un principio su uso predominante fue el de determinar la expresión diferencial de genes (DGE, del ingles *Differential Gene Expression*), su utilidad se demostró en un gran abanico de problemas, como por ejemplo el análisis de los patrones de *splicing*. Sin embargo, el estudio de los patrones de *splicing* a partir de datos de secuenciación implica la resolución de algunos desafíos que son inherentes al procedimiento que conlleva el mismo.

El protocolo de un experimento típico de RNA-Seq involucra una serie de pasos, algunos propiamente experimentales y otros computacionales. En primer lugar, se debe realizar la extracción del ARN que, dependiendo del tipo de experimento, puede ser seguido por un enriquecimiento en ARNm mediado por el reconocimiento de las colas de poli-A. Luego se produce la síntesis de los ADNc (de ADN copia) a partir de los ARN extraídos, la fragmentación de los mismos y la ligación de adaptadores, conformando así una librería de secuencias. En la placa de secuenciación se da la formación de los *clusters* mediante la amplificación de los fragmentos, para luego pasar a la secuenciación. Utilizando la tecnología de secuenciación Illumina se obtienen finalmente entre 10 y 30 millones de lecturas (o *reads* en inglés). Los siguientes pasos son computacionales y resultan fundamentales para lograr una interpretación biológicamente significativa de los resultados de la secuenciación. Las lecturas deben ser alineadas o ensambladas a un transcriptoma, lo que permite la asignación de cada una a una determinada región de un gen/transcripto. Para la cuantificación de la expresión de los genes y/o transcriptos se debe previamente pasar por pasos de filtrado y normalización de las cuentas asignadas a cada uno de ellos para, finalmente, poder realizar la estimación de la expresión diferencial entre distintas muestras.

Las lecturas producidas mediante esta tecnología tienen una longitud corta, de entre

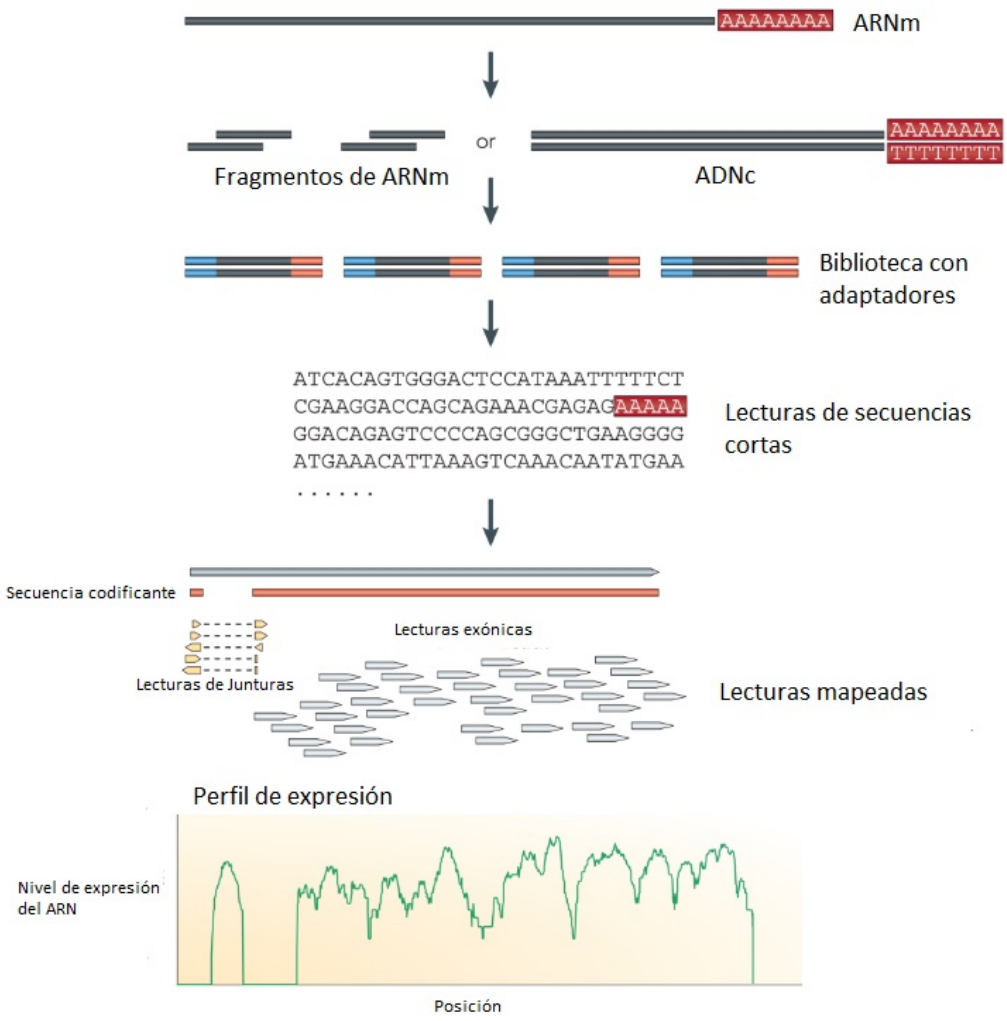



**Figura 1.8**: *Esquema de un experimento de RNA-Seq. Adaptado de Wang et al. 2008*[167].

150 y 200pb; esto representa un desafío a la hora de poder establecer de manera unívoca la cuantificación de los diversos transcriptos procedentes de un único gen por *splicing* alternativo. Si bien existen otras tecnologías que permiten una lectura mucho más larga, llegando a abarcar la longitud total de buena parte de los transcriptos, tanto por profundidad de secuenciación como por su bajo costo, las tecnologías de secuenciación asociadas a lecturas cortas son aún las más utilizadas[153].

Para el estudio del *splicing* es importante tener en cuenta que una lectura puede alinear de principio a fin con una determinada región del genoma, pero también, si la misma se

alinea sobre dos exones contiguos en un dado transcripto, vamos a ver que el alineamiento de esta lectura sobre el genoma va a estar interrumpido por la presencia del intrón (no presente en el ARNm maduro). A este tipo de lecturas se las denomina "junturas", y son de gran importancia para poder determinar variaciones en la cantidad relativa de las isoformas. Ante un evento de *splicing* alternativa, varias métricas que hacen uso de la información de junturas pueden utilizarse según el tipo de evento. Por ejemplo, en el caso de un salteo de exones se utiliza la métrica PSI (del inglés *Percent Spliced In*) que compara la cantidad de junturas que incluyen al exón y las que lo excluyen. Por otro lado, para el caso de la retención de intrones, tenemos el PIR (del inglés *Percent Intrón Retention*) que compara las junturas que delimitan un determinado intrón con la cantidad de lecturas que alinean sobre la secuencia del mismo. Entre diferentes muestras estas métricas pueden variar, lo que puede ser indicativo de variaciones en los patrones de *splicing*. Los métodos para cuantificar los eventos de *splicing* en una dada muestra pueden utilizar estas métricas, pero también, los conteos de lecturas de la región normalizado por millones de lecturas alineadas (del inglés *Fragments Per Kilobase of transcript per Million mapped reads* (FPKM) o *Reads Per Kilobase of transcript per Million mapped reads* (RPKM)). Los métodos que toman en consideración solamente las junturas tienen el problema que éstas representan solamente una pequeña parte del total de las lecturas generadas, por lo que no toman en cuenta una porción importante de la información generada por la secuenciación.

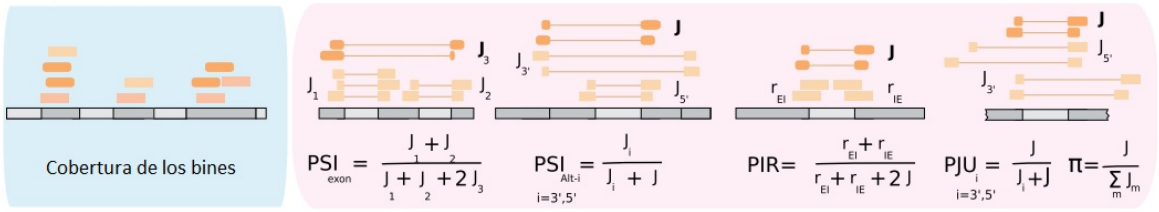



**Figura 1.9**: *Métricas para cuantificar eventos de splicing a partir del uso de junturas. Adaptado de Mancini et al. 2021*[103].

Los métodos computacionales que estiman la ocurrencia de *splicing* diferencial entre distintas condiciones pueden separarse entre: (1) los métodos basados en la reconstrucción

de isoformas (por ejemplo cuffdiff2[160] y DiffSplice[70]) y los métodos basados en conteos de regiones sub-génicas (por ejemplo DEXSeq[7], edgeR[131], rMats[145]). Por lo general, estos últimos presentan una mayor sensibilidad y precisión respecto a los primeros[110]. Por otro lado, los métodos basados en isoformas son incapaces de detectar eventos de *splicing* nuevos que no se correspondan con ninguna de los transcriptos anotados.

Como medio para superar las limitaciones antes mencionadas, recientemente en el *Laboratorio de biología de sistemas integrativa* se diseñó una metodología denominada *ASpli*[103] que permite integrar la información derivada del conteo de lecturas con la que proviene de las junturas. Este marco de trabajo, que será utilizado a lo largo de esta tesis, se centra en el análisis de bines, es decir, de regiones sub-génicas no superponibles que pueden ser pensados como " bloques" de secuencias que pueden ser incluidos o excluidos de un transcripto. Los bines pueden ser tanto exónicos como intronicos, anotados como no anotados.

# Capítulo 2

# Objetivos

## 2.1. Objetivo general

El objetivo general de esta tesis de doctorado es entender algunos de los aspectos que determinan el reconocimiento y posterior procesamiento de los intrones realizado por el proceso de splicing, poniendo énfasis en la relación entre la fortaleza de los 5'ss y la acción de un factor regulatorio como es PRMT5.

## 2.2. Objetivos específicos

- Elaborar de un modelo estadístico basado en el principio de máxima entropía de las secuencias de los 5'ss para entender la variabilidad que estas secuencias presentan a lo largo de un genoma e identificar los patrones de covariaciones entre las distintas posiciones de los 5'ss.

- Comparar, mediante el modelo generado, los 5'ss de diversas especies eucariotas para establecer qué patrones de covariación entre las posiciones de los 5'ss se encuentran conservados y cuáles son específicos para un dado linaje.

- Analizar datos de RNA-Seq procedentes de un diseño experimental de dos factores, en el que se compara el efecto de la mutación de PRMT5 en dos ecotipos de *Arabidopsis thaliana*, Col-0 y Ler.

- Analizar datos de RNA-Seq procedentes de plantas híbridas F1 entre los dos ecotipos antes mencionados y evaluar el efecto de la mutación de PRMT5 en las mismas.
- Determinar la relación entre la fortaleza de los 5'ss y el efecto que tiene PRMT5 sobre los patrones de splicing a nivel sistémico.

# Capítulo 3

# Modelo de máxima entropía para las secuencias 5'ss

## 3.1. Introducción

El reconocimiento de las señales presentes en los intrones es un requisito fundamental para llevar a cabo el proceso de *splicing*. Sin embargo, estas secuencias presentan una alta variabilidad a lo largo de los genomas de la mayoría de las especies eucariotas, la cual está estrechamente relacionada con la posibilidad de regular el reconocimiento de estos sitios según los diversos contextos celulares; dando lugar así al *splicing* alternativo. En este sentido, debe darse un fino equilibrio entre la variabilidad necesaria para los procesos regulatorios pero, al mismo tiempo, mantener una cierta cantidad de información que asegure la fidelidad del proceso de *splicing* y evite la potencial producción de transcriptos que alteren la funcionalidad celular.

Dada la importancia de los sitios de *splicing*, muchos estudios se han llevado a cabo para determinar cuáles son los patrones de estas secuencias que son más relevantes a nivel biológico. Los enfoques basados en teoría de la información han sido extensamente usados para resolver estos interrogantes. Ya a comienzos de la década de los '90, Schneider y Stephens utilizaron teoría de la información para analizar un total de 1800 sitios de *splicing*. Encontraron que había posiciones más variables que otras pero que el 80 % del contenido de información de los 5'ss se encontraba en su parte intrónica[154]. Otra investigación que pone

en relación la parte exónica e intrónica del 5'ss es el ya citado trabajo de Sverdlov *et al.* que propone una migración de la información del sitio de *splicing* desde su parte exónica a la intrónica[155]. Una investigación más reciente y que incorpora al análisis un mayor número de especies (61 en total), encontró que, si bien los 5'ss son comparables entre ellas, hay diferencias significativas en el contenido de información de algunas de las posiciones, en particular la +5; dejando entrever cierto grado de especificidad en los patrones encontrados. Las estadísticas que surgen de evaluar la frecuencia de cada uno de los cuatro nucleótidos en cada posición del 5'ss, como puede ser los logos consenso (como los mostrados en la Fig. 1.3) no logran capturar por completo la información estadística relevante en las secuencias. Para esto hay que tener en cuenta las correlaciones de orden mayor que se dan entre las posiciones de los 5'ss indicando, por ejemplo, si existe una covariación entre los nucleótidos de dos posiciones distintas. En este sentido, ya el trabajo de Stephens y Schneiner encontraron valores de información mutua significativos entre las siguientes posiciones del 5'ss de humanos: (-2,+4), (-1,+5) y (-2,+5)[154]. Nuevamente vemos que lo que ocurre en el lado exónico de la señal no parece ser independiente de lo que ocurre del lado intrónico. Las asociaciones entre las posiciones (-1, +5) y (-2, +5) también fueron reportadas por otros trabajos siguiendo diversas metodologías[22,157].

Las correlaciones entre dos posiciones de los 5'ss también fueron analizadas por Sahashi *et al.* encontrando que la no complementariedad entre la secuencia de los 5'ss y el U1 snRNA en ciertas posiciones podía ser compensada por una correcta complementariedad en otras[137]. En este mismo sentido, Denisov *et al.* realizó un análisis comparativo de los genomas de tres especies de mamíferos encontrando patrones de epistasis bien definidos entre los nucleótidos de las diferentes posiciones del 5'ss[38]. Los nucleótidos que resultan complementarios a la secuencia de U1 snRNA mantienen una epistasis positiva entre sí dentro de las regiones exónicas e intrónicas de los sitios. Sin embargo, lo contrario ocurre si se considera las relaciones entre posiciones de ambos lados del límite del exón: un nucleótido que tiende a fortalecer el sitio en la parte exónica parece desfavorecer la presencia de uno

que la fortalezca en la parte intrónica, y viceversa[38].

Con el objetivo de entender con mayor profundidad estos patrones presentes en los 5'ss, pero también con la intención de comprender la manera en la que éstos han ido variando a lo largo de la evolución de las especies eucariotas, construimos un modelo estadístico generativo a partir de las secuencias de los 5'ss anotadas en los genomas de distintas especies eucariotas, considerando un modelo de máxima entropía. En términos generales, el principio de máxima entropía puede ser entendido de la siguiente manera: si tenemos una cierta variable aleatoria $X$ con valores conocidos $x_1, x_2, x_3, ..., x_n$ pero cuyas probabilidades $p_1, p_2, p_3, ..., p_n$ son desconocidas, el principio de máxima entropía asigna probabilidades de forma tal que se maximice la entropía de la información de la variable $X$ bajo un cierto conjunto de restricciones $m$.

Los modelos basados en el principio de máxima entropía han sido ampliamente utilizados en diversos problemas biológicos, entre los que podemos mencionar: estudios sobre el nivel de actividad neuronal[53,133,143], ingeniería inversa de redes de regulación genética[95,130], estudios sobre la interacción entre factores de transcripción y el ADN[130,140], análisis de estructura de proteínas guiado por información evolutiva[44,114,148], etc. Detrás de la estrategia llevada a cabo por estos modelos se encuentra la idea de que las correlaciones de orden bajo, particularmente correlaciones de a pares, tienen una gran relevancia para entender la información biológica que subyace a los diversos procesos moleculares y fisiológicos.

## 3.2. Materiales y métodos

### Genomas analizados

Para el análisis de los 5'ss tuvimos en cuenta 30 especies eucariotas, entre las que incluimos los genomas de 5 especies de hongos, 8 de plantas y 17 de metazoos (Ver Tabla 3.1). Tanto la información de las secuencias genómicas - en formato FASTA - como la anotación de los genomas - en formato GTF o GFF - fue descargada de la base de datos *Ensembl*, https://ensemblgenomes.org/.

| Organismo | Código | Ensamblado | 5'ss (únicas) | 5'ss GT (únicas) |
|---|---|---|---|---|
| *Aspergillus nidulans* | ani | ASM1142v1 | 24563 (3954) | 24514 (3919) |
| *Coprinopsis cinerea okayama* | cci | CC3 | 61427 (4737) | 61325 (4670) |
| *Cryptococcus neoformans* | cne | cryp_neof_125_91_V1 | 36046 (2637) | 35220 (2435) |
| *Magnaporthe oryzae* | mor | MG8 | 22772 (3126) | 22567 (3039) |
| *Neurospora crassa* | ncr | NC12 | 16739 (2522) | 16533 (2423) |
| *Arabidospis thaliana* | ath | TAIR10 | 136036 (11286) | 130069 (6803) |
| *Hordeum vulgare* | hvu | IBSCv2 | 499846 (48879) | 413160 (10107) |
| *Medicago truncatula* | mtr | MedtrA17_4.0 | 156335 (11396) | 148511 (6964) |
| *Oryza sativa* | osa | IRGSP-1.0 | 130702 (12688) | 124646 (8418) |
| *Physcomitrium patens* | ppa | Phypa V3 | 194684 (8310) | 191147 (6991) |
| *Populus trichocarpa* | ptri | Pop_tri_v3 | 179803 (10168) | 176766 (9285) |
| *Solanum lycopersicum* | sly | SL3.0 | 141263 (12562) | 135225 (8400) |
| *Vitis vinifera* | vvi | 12X | 108097 (9614) | 105266 (7713) |
| *Anas platythynchos* | apl | CAU_duck1.0 | 152272 (9075) | 146465 (5405) |
| *Bos taurus* | bta | ARS_UCD1.2 | 203086 (11919) | 194432 (6624) |
| *Canis lupus familiaris* | clu | CanFam3.1 | 215912 (12437) | 206340 (5521) |
| *Danio rerio* | dre | GRCz11 | 276776 (11723) | 268477 (6720) |
| *Equus caballus* | eca | EquCab3.0 | 216007 (10367) | 208399 (4984) |
| *Gorilla gorilla* | ggo | gorGor4 | 212738 (14620) | 200262 (7697) |
| *Homo sapiens* | hsa | GRChv38 | 502197 (12129) | 488939 (6206) |
| *Monodelphis domestica* | mdo | ASM229v1 | 208511 (14090) | 196940 (6970) |
| *Mus musculus* | mmu | GRCm39 | 391997 (9026) | 384141 (5369) |
| *Ornythorhyncus anatinus* | oan | mOrnAna1.p.v1 | 187119 (13517) | 175225 (6562) |
| *Oryctolagus cuniculus* | ocu | OryCun2.0 | 147142 (12204) | 138124 (5217) |
| *Sarcophilus harrisii* | sha | mSarHar1.11 | 227688 (10104) | 220325 (5637) |
| *Salmo salar* | ssa | ICSASGv2 | 516921 (28445) | 455502 (14209) |
| *Sus scrofa* | ssc | Sscrofa11.1 | 260642 (21094) | 243064 (8300) |
| *Xenopus tropicalis* | xtr | Xenopus_tropicalis_v9.1 | 231618 (20862) | 205430 (10794) |
| *Caenorhabditis elegans* | cel | WBcel235 | 127661 (7625) | 124619 (5351) |
| *Drosophila melanogaster* | dme | BDGP6.32 | 63121 (4023) | 62451 (3780) |

**Cuadro 3.1**: **Genomas analizados.** *Para cada genoma analizado, se muestra el código del ensamblado utilizado, el número de sitios 5'ss totales extraídos de la anotación y de los 5'ss que presentan el di-nucleótido GT al inicio del intrón. Entre paréntesis se muestra el número de secuencias únicas.*

## Modelo estadístico

Para cada uno de los genomas, buscamos aproximar una función de distribución de la probabilidad conjunta, $P(\vec{S})$, asociada al conjunto de secuencias de los 5'ss observada. Cada elemento de este conjunto es una secuencia de 9 nucleótidos de largo (los tres primeros de la parte exónica y los 6 últimos de la parte intrónica) $\vec{S} = (s_{-3}, s_{-2}, s_{-1}, s_1, s_2, s_3, s_4, s_5, s_6)$

donde $s_i \in \{A, C, G, T\}$.

La distribución buscada debe ser compatible con las probabilidades marginales observadas para una dada posición o entre dos posiciones, $f_i(s_i)$ y $f_{ij}(s_i, s_j)$ respectivamente, tomando éstas como restricciones para encontrar la distribución mediante la maximización de la entropía. Bajo este marco, estimamos una función de densidad de probabilidad,$\hat{P}(\vec{S})$, que puede ser expresada como:

$$\hat{P}(\vec{S}) = \frac{1}{Z} e^{-E_d(\vec{S})} \tag{3.1}$$

con

$$E_d(\vec{S}) = -\sum_{i=1}^{9} h_i(s_i) - \sum_{i<j}^{9} J_{ij}(s_i, s_j) \tag{3.2}$$

siendo una medida de energía derivada a partir de los datos.. $Z$ es la función de partición que puede ser considerada aquí como una constante de normalización. $h_i(s_i)$ y $J_{ij}(s_i, s_j)$ son los parámetros de nuestro modelo que deben ser estimados. Éstos conforman las contribuciones de las estadísticas de una posición y de dos posiciones a la medida de energía de una dada secuencia. En total hay 36 parámetros de una posición (4 nucleótidos posibles para cada una de las 9 posiciones del 5'ss) y 576 parámetros de interacción entre dos posiciones (16 combinaciones de nucleótidos para los 36 pares de sitios posibles) que deben ser estimados

$$\sum_{\forall j \neq i} \sum_{s_j} \hat{P}(\vec{S}) = f_i(s_i) \tag{3.3}$$

$$\sum_{\forall k,l \neq i,j} \sum_{s_k, s_l} \hat{P}(\vec{S}) = f_{ij}(s_i, s_j) \tag{3.4}$$

De aquí en más por conveniencia se va a omitir la dependencia de las variables $s_i$ (por ejemplo, $J_{ij}(s_i, s_j) \equiv J_{ij}$)

## Procedimiento de ajuste

Para ajustar los 612 parámetros de nuestro modelo implementamos un esquema de descenso gradiente regularizado siguiendo lo hecho por Espada *et al.* (2015)[45]. Utilizando un

procedimiento de Monte Carlo, en cada iteración generamos un ensamble de 100.000 secuencias compatibles con la ecuación 3.1 y los parámetros del modelo obtenidos hasta esa iteración. Luego los parámetros ajustados son actualizados siguiendo los siguientes criterios:

$$h_i^{t+1} \leftarrow h_i^t - \epsilon_h[f_i - f_i^m]$$

si $J_{ij}^t = 0$

$$J_{ij}^{t+1} \leftarrow \begin{cases} 0 & \text{if } |\Delta| < \gamma \\ \epsilon_j[\Delta - sign(\Delta)] & \text{if } |\Delta| > \gamma \end{cases} \tag{3.5}$$

si $J_{ij}^t \neq 0$

$$J_{ij}^{t+1} \leftarrow \begin{cases} J_{ij}^t + \epsilon_j[\Delta - \gamma sign(J_{ij}^t)] & \text{if } \eta > 0 \\ 0 & \text{if } \eta < 0 \end{cases} \tag{3.6}$$

con $\Delta = f_{ij} - f_{ij}^{model}$ and $\eta = (J_{ij}^t + \epsilon_j[\Delta - \gamma sign(J_{ij}^t)]) * J_{ij}^t$.

El parámetro de regularización $\gamma$ evita que los parámetros de interacción que no sean suficientemente fuertes se aparten de cero.

Para cada genoma analizado, se estimó una familia de modelos de ajuste con diferentes valores de $\gamma$. A partir de la disminución secuencial de los valores de $\gamma$ se obtienen modelos con un aumento controlado de la complejidad. Con esta estrategia se puede identificar de manera precisa el mínimo conjunto de parámetros de interacción $J_{ij}$ necesarios para ajustar las frecuencias observadas de co-aparición de nucleótidos en los pares de posiciones del 5'ss. Este proceso se ilustra en la Fig. 3.1 para el caso de *Homo sapiens*.

En el panel izquierdo se muestra $\Delta = max[abs(f_{ij} - P_{ij})]$ (la máxima devianza absoluta entre las frecuencias de co-aparición observadas, $f_{ij}$, y estimadas, $P_{ij}$) en función del número de parámetros $J_{ij}$ distintos de cero. Esto se muestra para distintos valores de regularización $\gamma$, desde 0,01 hasta 0,05. Como puede observarse, una regularización más permisiva genera un modelo más complejo, es decir, con más parámetros distintos de cero; la cual se ajusta mejor a las frecuencias observadas. Para el valor de $\gamma = 0,45$ mostramos los valores de los parámetros $J_{ij}$ en función de la correspondiente correlación observada en las secuencias

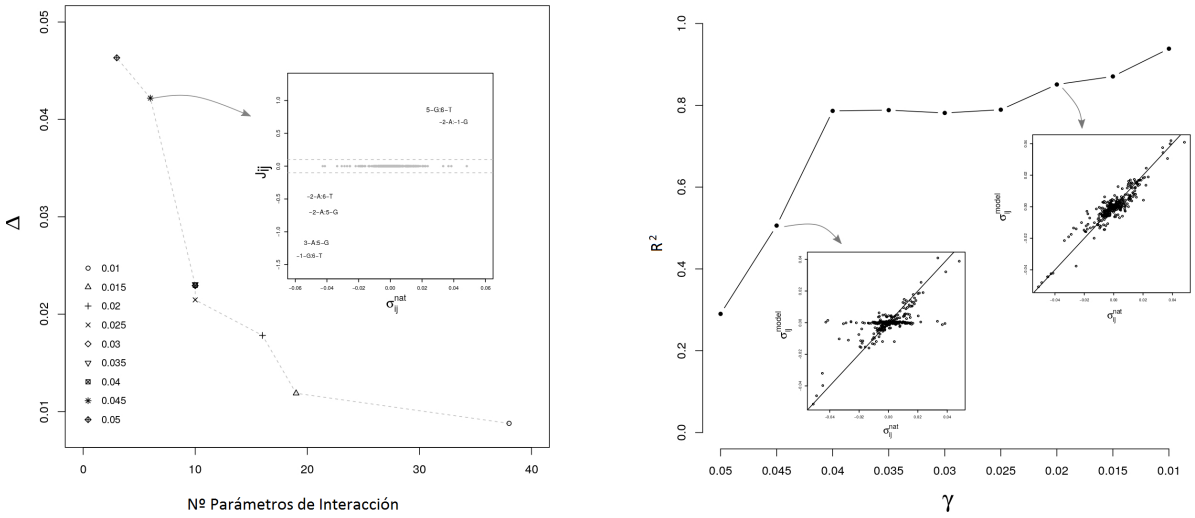



**Figura 3.1**: ***El rol de la regularización***. *Panel izquierdo: Desviación máxima absoluta entre las probabilidades entre pares de posiciones observadas y estimadas,*$\Delta = max[abs(f_{ij} - P_{ij})]$*, en función del número de parámetros* $J_{ij}$ *significativamente distintos a cero a distintos niveles de regularización (mostrados como diferentes símbolos en la figura,* $\gamma \in [0{,}01, 0{,}05]$*) para el caso de Homo sapiens. Gráfico interno: parámetros de interacción* $J_{ij}(s_i, s_j)$ *en función de las correlaciones entre pares de posiciones observadas* $\sigma_{i,j}^{nat}$ *para el modelo con* $\gamma = 0{,}045$*. Panel derecho: coeficientes de correlación* $r^2$ *entre correlaciones de pares de posiciones observadas y modeladas en función del parámetro de regularización* $\gamma$*. Gráficos internos: relación entre correlaciones de pares de posiciones modeladas,* $\sigma_{i,j}^{model}$*, y observadas,* $\sigma_{i,j}^{nat}$*, en* $\gamma = 0{,}045$ *and* $\gamma = 0{,}02$*.*

de los 5'ss, $\sigma_{ij} = f_{ij} - f_i f_j$. En este caso puede observarse que son 6 los parámetros de interacción reconocidos por este modelo que tienen la mayor asociación con la correlación de co-aparición que presentan los datos. En particular, puede verse que los modelos obtenidos con un nivel de regularización de $\gamma = 0,025$ experimentan una considerable reducción del valor de $\Delta$ con 10 parámetros $J_{ij}$ distintos a cero. Por otro lado, a partir del valor $0,015$ para $\gamma$, el valor de $\Delta$ se estabiliza.

A cada nivel de regularización los modelos ajustados pueden generar un ensamble de secuencias, con determinadas frecuencias para cada posición, $P_i$, y para los pares de posiciones, $P_{ij}$. En el panel derecho de la Fig. 3.1 se muestra el coeficiente de correlación $R^2$ de la regresión de los valores estimados mediante el modelo y los observados. Puede obser-

varse que con un nivel más bajo de regularización, es decir con la inclusión de un mayor número de parámetros, el modelo puede obtener cada vez mayores valores de correlación. Un comportamiento similar puede verse en todas las especies analizadas(ver Fig. A.1).

Para entender mejor el comportamiento de los parámetros ajustados, $\{h_i\}$ y $\{J_{ij}\}$, mostramos en la Fig. 3.2 un resumen de los resultados obtenidos en el modelo para humanos a un $\gamma = 0{,}025$.

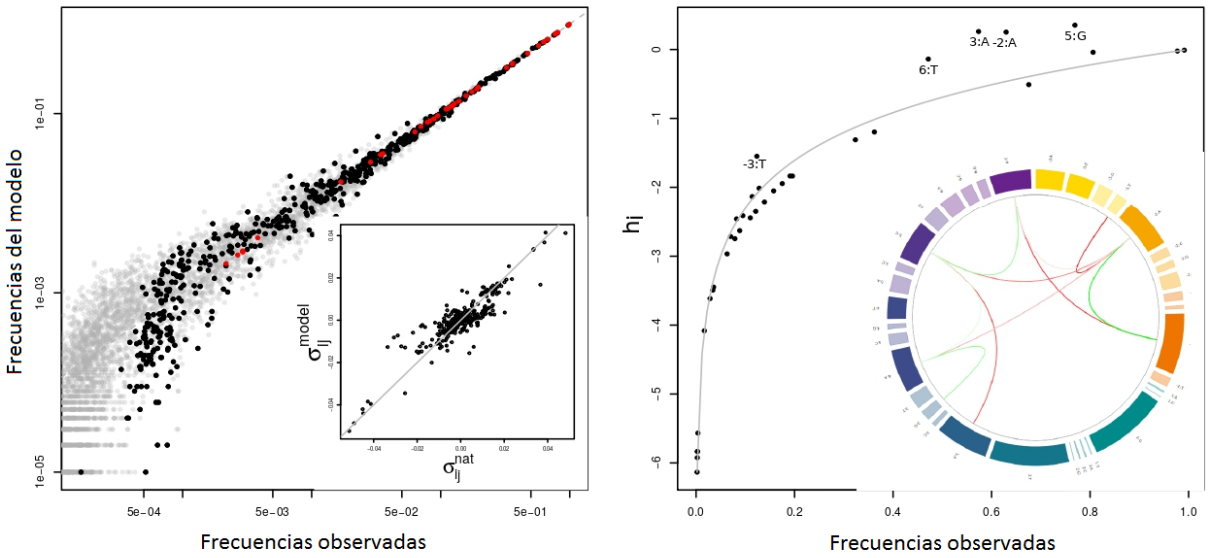



**Figura 3.2**: ***Frecuencias del modelo y parametros ajustados con*** $\gamma = 0{,}025$ . ***Panel izquierdo:*** *frecuencias estimadas de una posiciones, pares de posiciones y entre tres posiciones* $(p_i, p_{ij}, p_{ijk})$ *en función de las correspondientes frecuencias observadas* $(f_i, f_{ij}, f_{ijk})$ *son mostradas como puntos rojos, negros y grises, respectivamente. Gráfico interno: relación entre las correlaciones de pares de sitios estimadas,* $\sigma_{ij}^{model}$*, y observadas,* $\sigma_{ij}^{nat}$*.* ***Panel derecho:*** *valores ajustados para los* $h_i(s_i)$ *en función de las frecuencias observadas por posición,* $f_i(s_i)$*. La línea continua describe los valores esperados bajo un modelo en el que las posiciones sean independientes:*$h_i^{indep}(s_i) \sim \log(f_i(s_i))$*. Gráfico interno: Representación en forma de circos de las interacciones entre pares de posiciones. El anillo externo esta dividido en 36 bloques representando los 4 nucleótidos en cada una de las 9 posiciones del 5'ss. El área de cada bloque es proporcional a la probabilidad observada de cada nucleótido en cada posición,* $f_i(s_i)$*. Los colores cálidos corresponden a las posiciones intrónicas y los colores fríos para las exónicas. Interacciones positivas y negativas son representadas en color verde y rojo, respectivamente.*

En el panel de la izquierda de la Fig. 3.2 se muestra la relación entre los valores estimados

y observados para las frecuencias de una posición (puntos rojos) y entre pares de posiciones (puntos negros). En ambos casos se obtiene un $R^2 > 0{,}99$. En el panel interno se puede ver la relación entre las correlaciones modeladas y las observadas, obtenido al nivel de regularización elegido un $R^2 \sim 0{,}84$. A pesar de que nuestro procedimiento de ajuste solo toma en cuenta las frecuencias de una posición y de pares de posiciones, también puede reproducir las frecuencias de orden superior. En el gráfico se muestra la correlación entre valores estimados y observados de las probabilidades entre tres posiciones (puntos grises), obteniendose un $R^2 = 0{,}95$.

En el panel derecho se muestra los valores para el parámetro $h_i$ en función de las frecuencias observadas $f_i$. En una línea gris continua se muestra como referencia la estimación de los parámetros $h_i$ si no se tuviera en cuenta los patrones de interacción entre las posiciones, en tal caso $h_i \sim log(f_i)$. Se resaltan cinco casos en los que existe una desviación considerable de lo esperado bajo estas condiciones. También se muestra una representación en forma de *circos* que resume de manera gráfica gran parte de la información estadística relevante de las secuencias de 5'ss para una dada especie, humanos en este caso. El anillo está formado por 36 bloques que representan los 4 nucleótidos posibles, $\{A, C, G, T\}$, en las 9 posiciones del 5'ss. Los colores cálidos son asignados a las posiciones exónicas $\{-3, -2, -1\}$ y los colores fríos a las posiciones intrónicas $\{1, 2, 3, 4, 5, 6\}$. El ancho de cada bloque da cuenta de la frecuencia con la que un dado nucleótido es observado en una dada posición. Las lineas que unen diferentes pares de posiciones representan los parámetros $J_{ij}$ distintos de cero, los cuales pueden ser negativos (líneas rojas) o positivos (líneas verdes).

## 3.3. Resultados

### Escala de energía

De acuerdo con nuestro estimador de las probabilidades de una secuencia introducido en la ec. 3.1, la función de energía definida en ec. 3.2 representa una medida cuantitativa de la frecuencia en la que una dada secuencia puede ser encontrada a lo largo del genoma.

Mientras que secuencias de baja energía corresponden con los 5'ss más frecuentes, las de alta energía están asociadas a secuencias raras e infrecuentes.

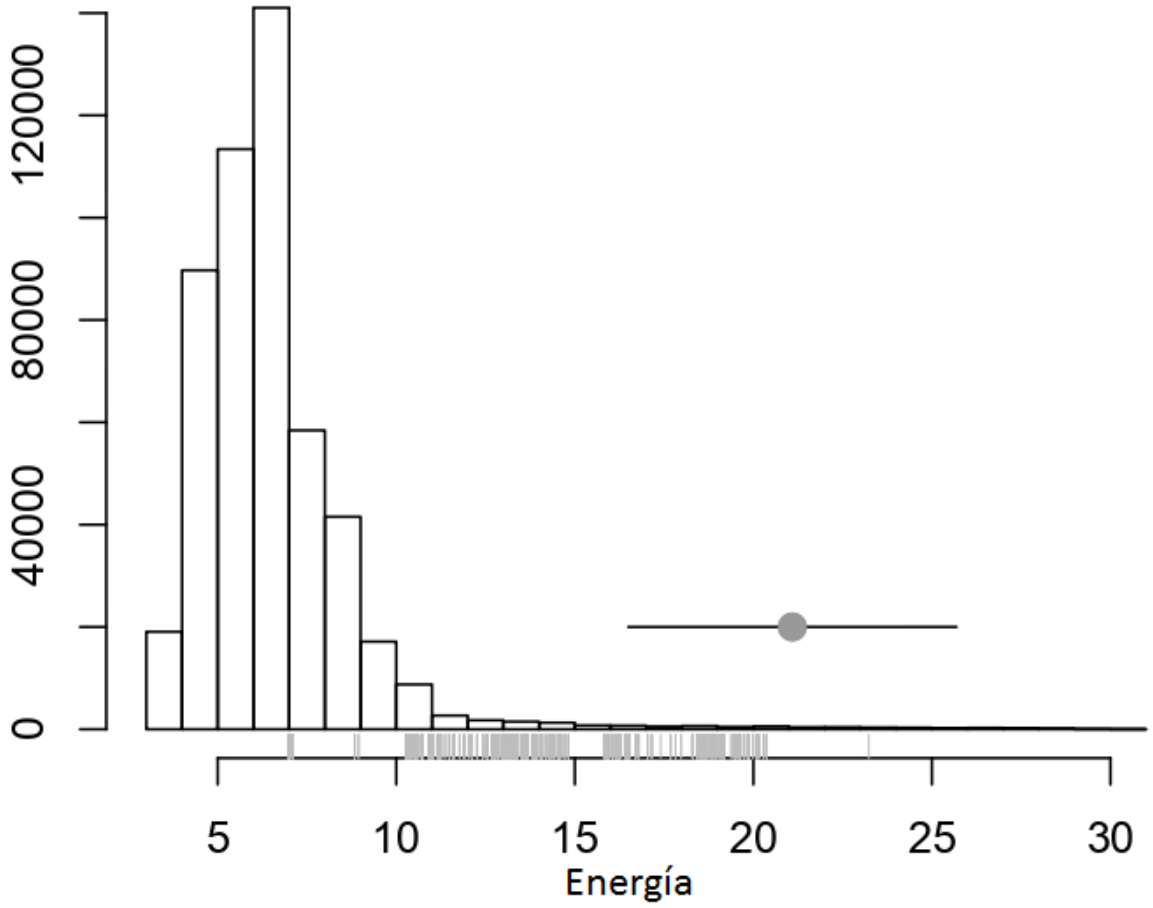



**Figura 3.3**: ***Energías de las secuencias de los 5'ss de Homo sapiens*** $(\gamma = 0{,}025)$***.*** *Distribución de energía para las 502.197 secuencais de los 5'ss extraídas del genoma humano. Las barras grises corresponden a las energías obtenidas por los 5'ss asociados a U12. El punto gris muestra el valor de energía promedio observado en un ensamble de 2.000 secuencias aleatorias, mientras que la linea negra indica el intervalo de energía* $\pm\sigma$ *de esta distribución nula.*

La figura 3.3 muestra la distribución de frecuencia de las energías de los 50.2197 5'ss analizados en el genoma humano. El 90 % de las secuencias presentan una energía con un valor en el rango $E_d \in [4{,}1, 9{,}7]$. La secuencia de mínima energía, con un valor de

$E_d(\vec{S}^*) = 3{,}50$, es $\vec{S}^* = \{C, A, G, G, T, A, A, G, T\}$. Merece destacarse que, al mismo tiempo, esta secuencia también es la que presenta una mayor complementariedad con el snRNA de U1. Por otro lado, esta secuencia es el mínimo global dentro del paisaje energético de secuencias. Un ensamble de 2.000 secuencias generadas de manera aleatoria fueron utilizadas como referencia para comparar con nuestra escala de energía. En la Fig. 3.3 se muestra un punto gris y una linea negra representando la media y desvío estándar de energías que obtuvieron estas secuencias aleatorias ($E_d = 21{,}1 \pm 4{,}6$). Esto marca los valores de energía que pueden tomar secuencias completamente desordenadas.

Como puede observarse, la distribución de energías presenta una asimetría con una cola hacia los valores de mayor energía. Una de las preguntas que nos podemos hacer es qué energías obtendrían los 5'ss asociados al reconocimiento por el spliceosoma menor. Tomamos 136 sitios de splicing que son procesados por este spliceosoma de la *Intron Annotation and Orthology Database*[115]. En la Fig. 3.3 puede observarse los valores de energía obtenidos para estas secuencias como barras grises. En la mayoría de los casos se observa que estas secuencias presentan una alta energía, por lo que podemos concluir que éstas presentan estadística distinta a la que predomina en las secuencias utilizadas en el modelo.

Por otro lado, la distribución de energías nos habla de una escala de variabilidad en las secuencias de los 5'ss del genoma que va desde la secuencia con complementariedad perfecta al snRNA de U1 hasta secuencias que presentan un bajo nivel de información. La gran mayoría de las secuencias presenta una energía media entre estos extremos, lo que nos lleva a pensar en una región de variabilidad energética en la que se favorece el reconocimiento de los sitios pero evitando el estado de máximo orden, permitiendo así un grado óptimo de regulación sobre el proceso de *splicing*.

Finalmente el hecho de que la secuencia de mayor complementariedad con el snRNA de U1 sea a su vez la secuencia de mínima energía nos llevó a preguntarnos por la correlación que hay entre estas dos medidas. En la Fig. 3.4 se muestra la energía libre de dimerización estimada entre las secuencias de los 5'ss la porción complementaria a éste del snRNA de

U1 en función de nuestra escala de energía. Como puede verse existe un cierto grado de correlación entre estas variables, en el que secuencias con una mayor complementariedad con snRNA de U1 tienden a tener energías menores.

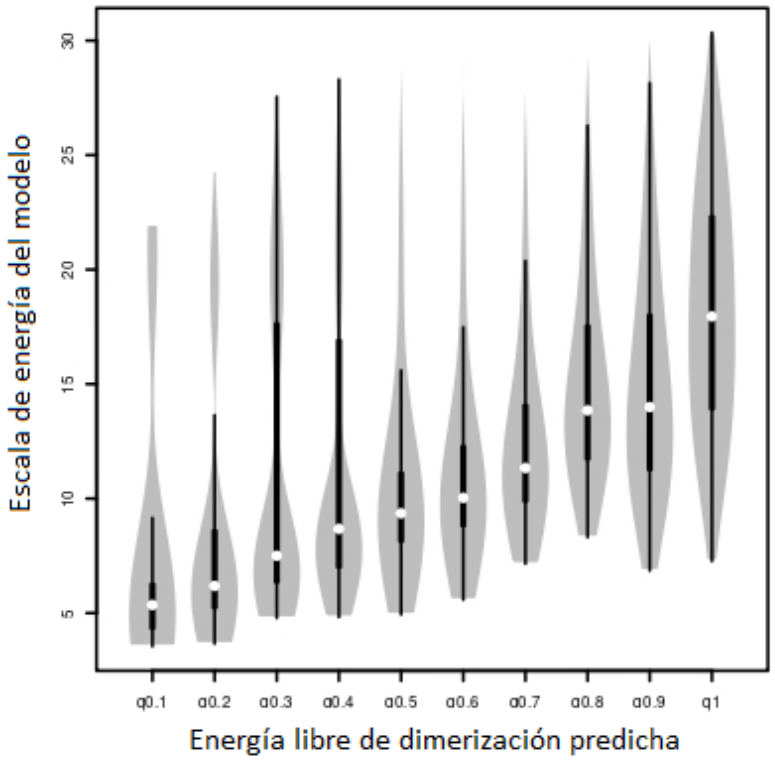



**Figura 3.4**: **Relación con la energía de unión entre los 5'ss y el snRNA de U1.** *Gráfico de cajas de los valores de energía estimados para los 502.197 5'ss extraídos de la anotación del genoma de humanos para decil de la energía libre de dimerización predicha entre el sitio y la porción complementaria a éste en el snRNA U1. Esta última fue estimada utilizando el programa RNAcofold del paquete de R ViennaRNA 2.0*[101]*, usando parámetros estándar.*

## Paisaje energético

Para caracterizar en mayor profundidad el paisaje de energía de las secuencias 5'ss, analizamos el modelo realizado en base al genoma humano con un $\gamma = 0{,}025$ considerando un enfoque de redes complejas. Representamos cada secuencia única como un nodo en la red, estableciendo un enlace solamente entre secuencias que sean distintas en una única

posición (como ejemplo se muestra en la Fig. 3.5 un sub-grafo de esta red). En teoría de la información, esto equivale a decir que conectamos entre sí secuencias que tienen una distancia de Hamming igual a 1. Esta medida da cuenta del número mínimo de sustituciones que debo realizar para cambiar de una dada cadena de caracteres a otra. De esta manera construimos una red dirigida, donde los enlaces apuntan hacia la secuencia de menor energía. Encontramos 243 secuencias que actuaban en nuestra red como mínimos locales, es decir, todas las secuencias con las que estaban conectadas tenían una energía mayor que éstas; por lo que el número de enlaces que salen "hacia afuera" de estos nodos es igual a 0. Para cada uno de estos nodos atractores pudimos estimar la componente de la red que se dirige hacia ellos en un número finito de pasos. Esta es una medida de la extensión que tiene la base de atracción de una dada secuencia atractora.

En la Figura 3.6 se muestra un gráfico de violín para la distribución de energías de las secuencias que pertenecen a bases de atracción de la red dirigida. Las primeras 15 muestran bases de atracción de secuencias que presentan el di-nucleótido GT al comienzo del intrón, luego se representa en una única distribución las energías pertenecientes a bases en las que este dinucleótido no está presente y, por último, de secuencias que no están enlazadas con la componente gigante de la red. El punto rojo marca la energía correspondiente a la secuencia atractora en cada una de las bases, cuya secuencia se incluye como etiqueta en el eje horizontal del gráfico. Los números en la parte superior del gráfico hacen referencia al número de secuencias presentes en cada una de las bases de atracción.

La primer distribución de la Fig. 3.6 corresponde a la mayor base de atracción, conteniendo 9.593 secuencias. Este número representa el 80 % del conjunto completo de secuencias de 5'ss que se encuentran presentes en el genoma de *Homo sapiens* y el 96 % del subconjunto de éstas que tienen el di-nucleótido GT al comienzo del intrón. Nuevamente vemos que el atractor asociado a esta base es la secuencia que presenta el mejor apareamiento con el snRNA de U1: $\vec{S}^* = \{C, A, G, G, T, A, A, G, T\}$, y la menor energía en nuestra escala: $E_d(\vec{S}^*) = 3{,}5$. El resto de las bases de atracción presentan mínimos locales de mucha mayor

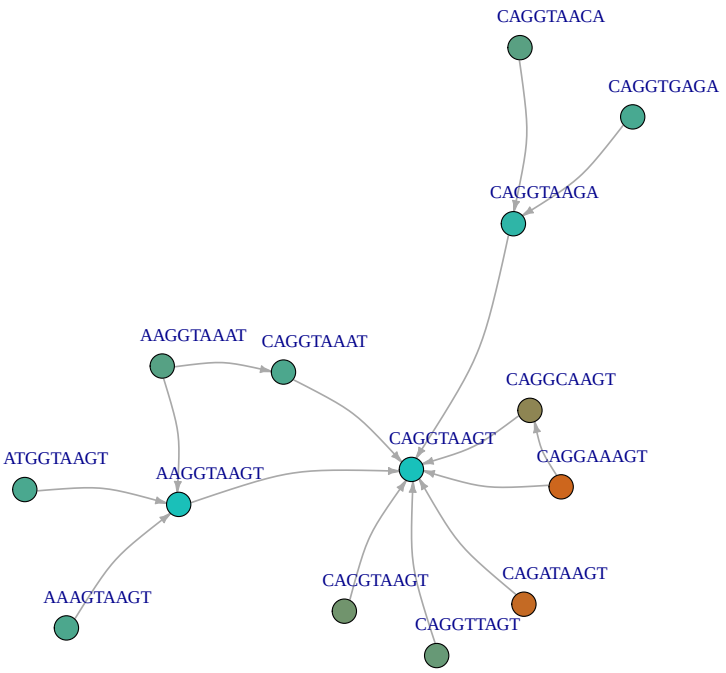



**Figura 3.5**: **Subgrafo de secuencias de 5'ss.** *Cada nodo del grafo corresponde una secuencia de 5'ss. Los colores de los mismos representan el valor de energía obtenido a partir del modelo realizado para Homo sapiens. Los enlaces se encuentran dirigidos hacia el nodo de menor energía, nótese que en este subgrafo la secuencia CAGGTAAGT actúa como una base de atracción hacia la cual se dirigen todos los enlaces.*

energía, de los cuales solamente 14 presentan como atractor una secuencia con un GT al comienzo del intrón. Estos resultados indican que nuestro sistema presenta un mínimo global bien definido, dentro de un amplio panorama energético.

## Robustez de los patrones de interacción

Para poder determinar qué tan robustos son los patrones de interacción inferidos a partir de nuestros modelos, consideramos para el caso de *Homo sapiens* diferentes subconjuntos de secuencias de 5'ss para construir los modelos: conjunto total de las secuencias 5'ss presentes

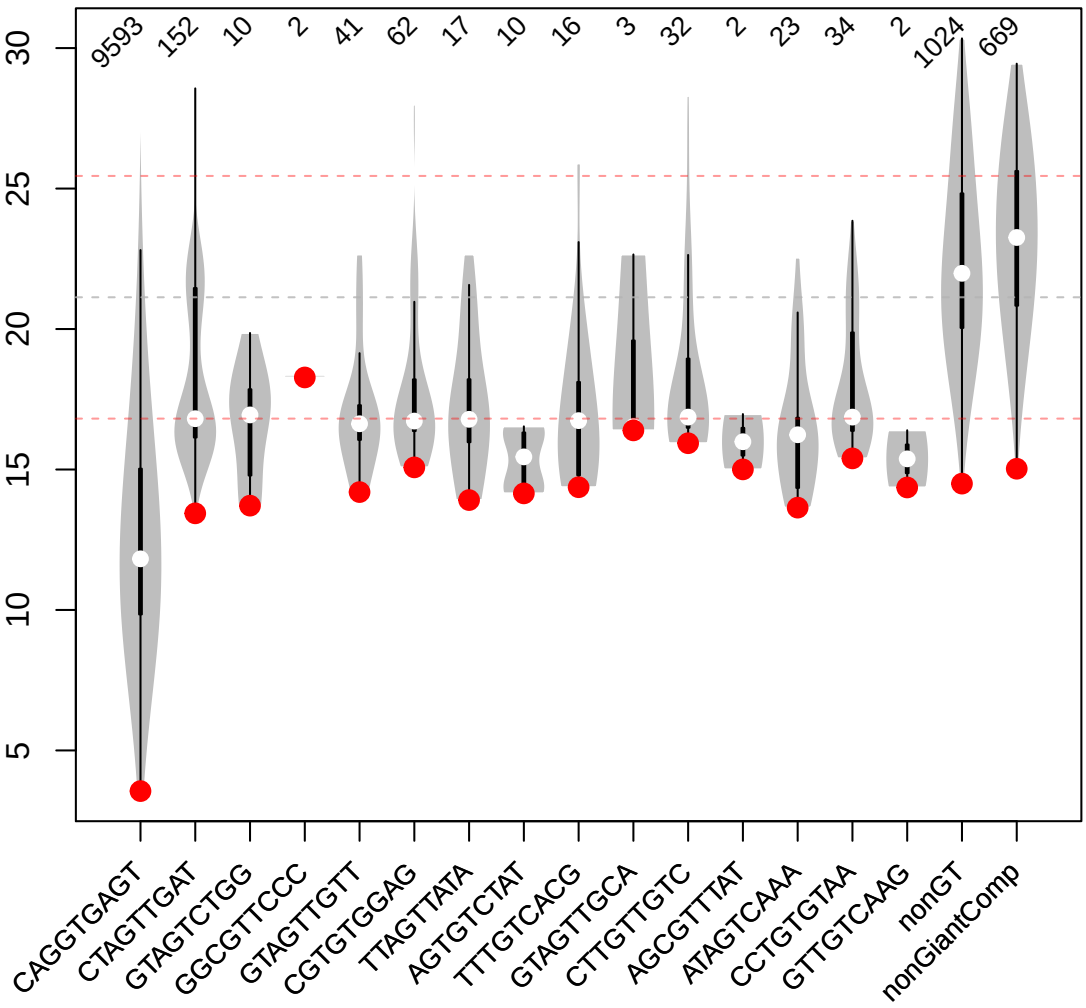



**Figura 3.6**: **Bases de atracción en la red dirigida de secuencias de 5'ss.** *Gráfico de violín para las distribuciones de energías de las secuencias que corresponden a distintas bases de atracción en la red dirigida de secuencias de 5'ss. Se muestran las 14 bases en las que el atractor corresponde a una secuencia con el dinucleótido canónico GT al comienzo del intrón. Unidas en una única distribución se muestran, por un lado, las secuencias que no presentan este dinucleótido (nonGT) y, por el otro, las que se encuentran separadas de la componente principal de la red (nonGiantComp). El punto blanco corresponde a la mediana de energía para cada distribución y el punto rojo al valor de energía del atractor en cada caso.*

en la anotación del genoma (502.497), un subconjunto que tome solamente las secuencias en las que esté presente el di-nucleótido GT al comienzo de los intrón (488.939) y dos

subconjuntos de 5'ss que tengan evidencia experimental de ser efectivamente reconocidos en el proceso de splicing.

Para obtener información sobre 5'ss soportados por datos de trsncriptómica utilizamos *RJunBase* http://www.rjunbase.org/. Esta base de datos integra información sobre junturas encontradas en 10.283 experimentos de RNA-Seq tanto de tejidos humanos normales como cancerosos presentes en las bases de datos *The Cancer Genome Atlas* (TCGA) y *Genotype-tissue Expression* (GTEx). Para tener en cuenta una determinada juntura tuvimos que: 1) sea una juntura lineal; 2) esté presente en la anotación del genoma; 3) corresponda a un gen que codifica proteínas; y 4) se encuentre expresada en tejido normal. Teniendo en cuenta estos criterios, obtuvimos 233.961 5'ss. También fue considerado un segundo subconjunto más restrictivo que tiene en cuenta solamente junturas cuya mediana de expresión en tejidos normales sea mayor a 5 cuentas.

Los cuatro paneles de la Fig.3.7 muestran los circos que se obtienen de los modelos ajustados a partir de los cuatro subconjuntos de 5'ss ya mencionados. Como puede apreciarse, no se observan cambios sustanciales en los patrones de interacción de los diferentes modelos, lo que indica la robustez de los mismos frente a los distintos conjuntos analizados.

## 3.4. Conclusión

El reconocimiento de los 5'ss durante el proceso de *splicing* es sumamente complejo por el gran número de factores que ejercen una influencia en el mismo. En primer lugar, hay que mencionar que, a excepción del di-nucleótido GT al comienzo del intrón, las secuencias de los 5'ss son altamente variables. En segundo lugar, este reconocimiento se da en distintas etapas del ciclo de *splicing* por parte de diferentes complejos que se unen de manera total o parcial al 5'ss. Por otro lado, no sólo la hibridación entre el sitio de *splicing* y los snRNA determinan el reconocimiento de los primeros, sino que también participan múltiples proteínas que ayudan a estabilizar la unión del complejo de forma inespecífica. Por último, hay que tener en cuenta que el *splicing* se da en un determinado contexto genómico y celular.

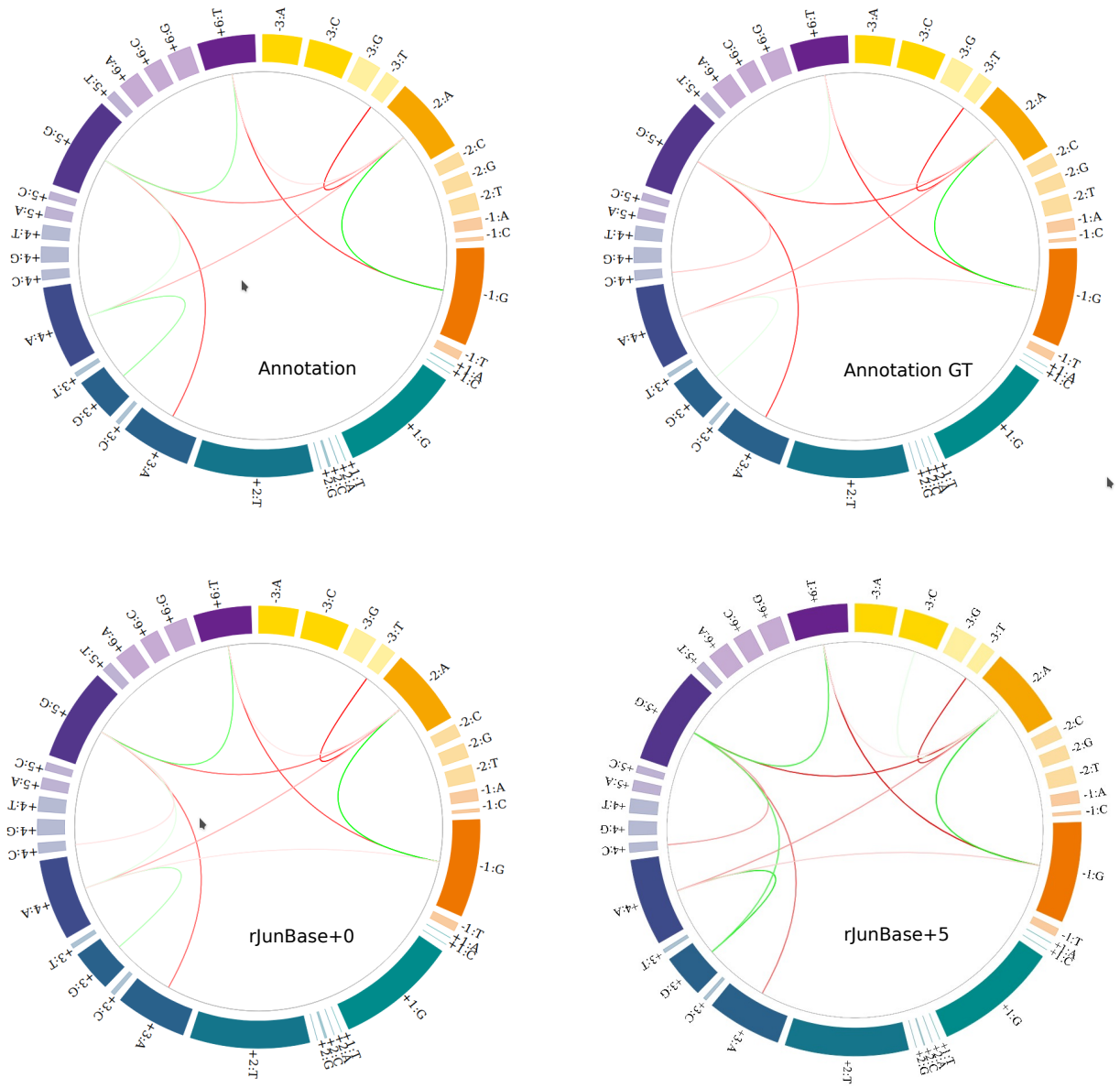



**Figura 3.7**: **Diagrama de circos para los patrones de interacción estimados para los 5'ss de *Homo Sapiens*.** *Los patrones de interacción para el conjunto completo de secuencias de 5'ss extraídas de la anotación del genoma (502.197) y del subconjunto de éstas que presentan GT al comienzo del intrón (488.939) se muestran en el panel superior izquierdo y derecho, respectivamente. Los patrones de interacción para los modelos generados a partir de 5'ss asociados a junturas en datos de transcriptómica de tejidos normales de acuerdo a la base de datos* RJunBase *(233.961) se muestran en el panel inferior izquierdo. En el panel inferior derecho se muestra el modelo generado solamente a partir de junturas que presentan al menos una mediana de expresión de 5 en esta base de datos (114.745).*

Muchos son los factores tanto en cis, como la estructura genética y la presencia de secuencias resaltadoras/silenciadoras, como en trans, factores de *splicing* y remodelaciones de la cromatina, que establecen distintas capas de regulación que determinan en cierto grado el nivel de reconocimiento de cada 5'ss en particular. En este trabajo buscamos identificar señales tanto conservadas como divergentes en las secuencias de los 5'ss de distintos organismos que reflejan, en cierto modo, la influencia que todos estos factores dejaron en la composición de estas secuencias.

Nuestra estrategia de maximización de la entropía nos permitió recuperar resultados obtenidos por trabajos previos y arrojar luz sobre nuevas regularidades presentes en los 5'ss. La escala de energía definida a partir de nuestros modelos, $E_d$ (Eq. 3.2), recupera la idea detrás de la medida *SD-Score* (definida como el logaritmo de la frecuencia de una secuencia 5'ss) introducida por Sahari y colaboradores para predecir el resultado del *splicing* en mini-genes diseñados artificialmente[137].

Vimos también que la $E_d$ correlaciona con la energía de dimerización estimada de los 5'ss con la porción del snRNA de U1 encargado de su reconocimiento (Figura **??**). A pesar del gran número de factores en cis y trans que influyen sobre el reconocimiento de los sitios de *splicing*, nuestra escala de energías parece reflejar ciertos aspectos funcionales de este reconocimiento. En este sentido, $E_d$ no solo nos habla de la frecuencia en la que una dada secuencia de 5'ss se encuentra en el genoma sino también se relaciona con el aspecto funcional de las mismas. La mayor parte de las secuencias que se observan en los genomas presentan valores intermedios de energía(Fig. 3.3), lo cual es consistente con lo observado por trabajos previos en los que se concluye que un apareamiento perfecto con la maquinaria de *splicing* podría ser evitado en favor de un aumento en la variabilidad de los sitios de *splicing*, siendo la formación de entre 5 y 6 pares de Watson-Crick una cantidad óptima para que se de el reconocimiento[22]. De esta forma, se alcanza un buen grado de fidelidad en el reconocimiento, al mismo tiempo que se permite la regulación del mismo ante distintos contextos celulares.

En nuestros modelos pudimos identificar patrones de interacción entre pares de posicio-

nes en las secuencias que realizan un aporte importante en la definición de la escala $E_d$. Por un lado, se observó que las interacciones entre nucleótidos consenso hacia el interior de la región exónica o intrónica del sitio tienden a ser positivas. Lo contrario ocurre en las interacciones entre posiciones que relacionan las regiones exónicas e intrónicas, en estos casos los nucleótidos consenso mantienen una relación negativa (Table 4.2). Estos resultados extienden observaciones previas realizadas sobre sitios de *splicing* de mamíferos, como humanos y ratón[22,38,137], y apoyan la idea de que se favorece la complementariedad del 5'ss o bien en la parte exónica o bien en la intrónica, pero no en ambas partes del sitio al mismo tiempo.

Las interacciones significativas entre pares de posiciones de los 5'ss dan cuenta de que éstas no pueden ser tomadas como independientes. Nuestro trabajo reporta que las probabilidades conjuntas entre los pares de posiciones presentan una información biológica relevante y están estrechamente relacionadas con las relaciones filogenéticas entre los organismos. Los 5'ss son reconocidos por el snRNP U1 en el complejo temprano del spliceosoma y por los complejos U5 y U6 snRNP en la etapa pre-catalítica del ciclo de *splicing*. En este contexto, relaciones de compensación entre las distintas posiciones del sitio podrían ser las principales responsables de mantener la fidelidad del proceso de *splicing* incluso cuando la variabilidad de cada posición en particular sea alta. Pequeñas diferencias en cuanto a la evolución de estos mecanismos de compensación pueden estar a la base de las diferencias que se observan en cuanto a los patrones de interacción entre posiciones en las diversas especies eucariotas.

# Capítulo 4

# Estudio comparativo de los 5'ss en plantas, hongos y metazoos

## 4.1. Materiales y métodos

### Time-tree

A partir de la lista de especies utilizadas para generar los modelos de 5'ss obtuvimos un *time-tree* en el que se encuentra representadas las relaciones evolutivas entre ellas (http://www.timetree.org/search/goto-timetree, último acceso 20 Diciembre de 2021), como puede verse en la Fig. 4.1. De las 30 especies analizadas, hubieron dos que no se encontraron en la base de datos TimeTree por lo que para los siguientes análisis fueron reemplazadas por especies evolutivamente muy cercanas. Estas dos especies fueron: *Magnaporthe oryzae* (mor), que fue reemplazada por *Pseudohalonectria lignicola* (ambas pertenecen a la familia *Magnaporthacease*), y *Coprinopsis cinerea* (cci), que fue reemplazada por *Coprinopsis lagopus* (siendo ambas del género *Coprinopsis*).

### Estimación de señales filogenéticas

Para estudiar la asociación y posible señal filogenética en los parámetros de interacción encontrados en los modelos de las distintas especies analizadas, que en total suman un todal de 41 interacciones distintas, utilizamos el proceso de aleatorización Maddison-Slatkin[102]. Este es un enfoque de *bootstraping* no paramétrico para generar una distribución de valores

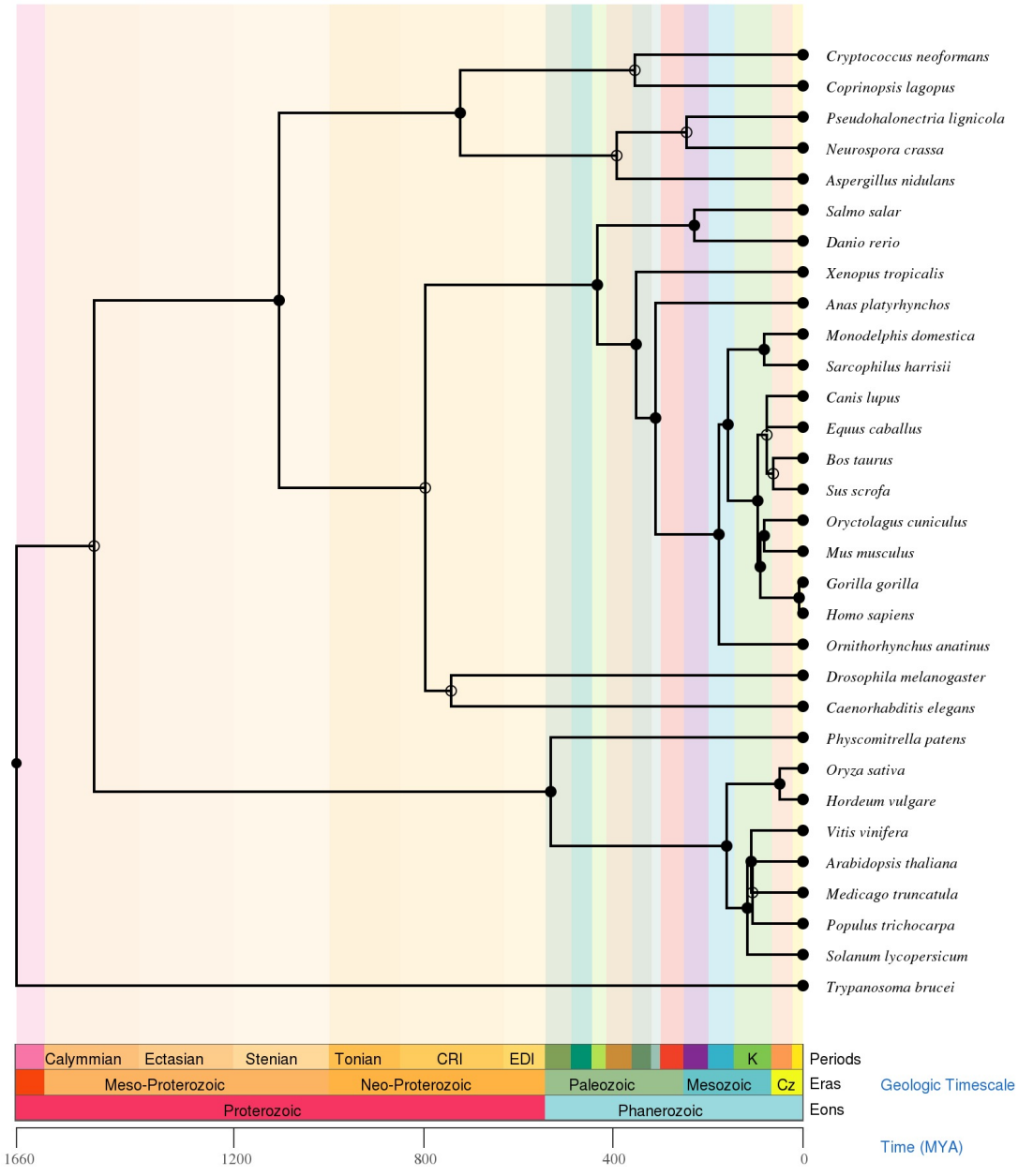



**Figura 4.1**: ***Time-tree***. *Para las especies analizadas, obtenido de* `http://www.timetree.org/search/goto-timetree`

esperados de una prueba estadística. En nuestro caso, consideramos un puntaje de parsimonia definido como el número de cambios de la variable binaria presencia/ausencia para un dado parámetro de interacción. Para cada uno de estos parámetros se le asignó de forma aleatoria un determinado estado de presencia/ausencia a cada especie 1.000 veces y se estimaron los valores de parsimonia (metodología de Sankoff[139]) utilizando la función *parsimony* presente en el paquete de R *phagorn*[142]. Se estimaron los p-valores para la fracción de eventos aleatorios con un valor de parsimonia mayor o igual al valor observado para el parámetro de interacción bajo análisis, corrigiendo por múltiple testeo mediante el método de Bonferroni.

## Dendrogramas inferidos de las probabilidades entre dos posiciones

Consideramos distancias euclideas entre las matrices triangulares de probabilidades $P_{ij}$ de los modelos de las genomas analizados. El dendrograma fue producido usando el método de agrupamiento por enlazamiento completo.

## Comparación de dendrogramas

Para la comparación de dendrogramas utilizamos las funciones que se encuentran implementadas en el paquete de R *dendextend*[52]. En particular, utilizamos un tipo de diagramas llamados *tanglegrams* que logran comparar de manera visual y cualitativa dos ordenamientos jerárquicos. Por otro lado, usamos la función *Bk-permutations* para llevar a cabo un análisis de *boostrap* (1.000 permutaciones) de los índices Fowlkes-Mallows para comparar particiones generadas a partir de cortar a diferentes niveles el dendrograma de interés[49]. Dadas dos particiones distintas de $n$ objetos, y siendo $m_{i,j}$ el número de elementos comunes entre el cluster i-ésimo y j-ésimo de las particiones, el índice Fowlkes-Mallows es definido como:

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}}$$

donde

$$T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{i,j}^2 - n \tag{4.1}$$

$$P_k = \sum_{i=1}^{k}(\sum_{j=1}^{k} m_{i,j})^2 - n \tag{4.2}$$

$$Q_k = \sum_{j=1}^{k}(\sum_{i=1}^{k} m_{i,j})^2 - n \tag{4.3}$$

$0 \leq B_k \leq 1$. Un valor más alto del índice de Fowlkes–Mallows indica un mayor grado de similitud entre los clusters de las dos particiones comparadas.

## 4.2. Resultados

### Patrones de interacción conservados

Nuestro procedimiento de ajuste permite identificar conjuntos de interacciones entre pares de posiciones en los 5'ss. Algunos de ellos involucran nucleótidos que son consenso en las posiciones que interactúan, por ejemplo la interacción que une al -1G y el +6T en *Homo sapiens* (Fig. 3.7). Otros ponen en relación un nucleótido consenso con otro que no lo es en las posiciones que interactúan, como la interacción entre -2A y -3T. En la Tabla 4.2 se muestra para los genomas analizados el promedio de la intensidad de las interacciones entre los diferentes casos que pueden darse: nucleótidos que son consenso (C) o no (NC), tanto en posiciones que son exónicas (E) como intrónicas (I). A pesar de algunas diferencias entre las distintas especies, algunos patrones parecen ser compartidos por la gran mayoría de los casos.

Interacciones positivas entre nucleótidos consensos de posiciones exónicas (EC-EC) o intrónicas (IC-IC) están presentes en casi todas las especies, lo que indicaría que el consenso parece reforzarse en cada lado de la secuencia del 5'ss. Sin embargo, también parece ser común a las distintas especies una interacción negativa entre nucleótidos consensos de posiciones que se encuentran en ambos lados del sitio (IC-EC). Esto nos esta indicando que

| Especies | IC-EC | IC-ENC | INC-EC | INC-ENC | IC-IC | IC-INC | INC-INC | EC-EC | EC-ENC | ENC-ENC |
|---|---|---|---|---|---|---|---|---|---|---|
| **cne** | -0.41 | 0 | 0 | 0 | -0.14 | -0.01 | 0 | 0.88 | 0.13 | -0.13 |
| **ani** | -0.5 | 0 | 0.03 | 0 | 0.21 | 0.02 | 0 | 0.82 | 0 | -0.15 |
| **ncr** | -0.68 | 0 | 0 | 0 | 0.26 | 0.03 | 0 | 0.68 | 0 | -0.07 |
| **mor** | -0.45 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0.87 | 0 | -0.15 |
| **cci** | -0.59 | 0 | 0.04 | 0 | 0.15 | 0.09 | 0 | 0.7 | 0 | -0.35 |
| **ath** | -0.39 | 0 | 0.05 | 0 | 0.37 | 0 | 0 | 0.83 | 0 | -0.11 |
| **hvu** | 0 | 0.09 | 0 | -0.06 | -0.13 | 0.92 | 0.09 | 0 | -0.16 | -0.29 |
| **mtr** | -0.3 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0.95 | 0 | 0 |
| **osa** | -0.24 | 0 | 0.01 | 0 | 0.23 | 0.05 | -0.04 | 0.94 | 0 | 0 |
| **ppa** | -0.77 | 0 | -0.01 | 0 | -0.02 | -0.01 | -0.03 | 0.64 | 0 | -0.08 |
| **ptri** | -0.42 | 0 | 0 | 0 | -0.02 | 0 | 0 | 0.91 | 0 | 0 |
| **sly** | -0.36 | 0 | -0.01 | 0 | 0.21 | 0 | 0 | 0.91 | 0 | -0.07 |
| **vvi** | -0.37 | 0 | 0 | 0 | -0.02 | 0 | 0 | 0.93 | 0 | -0.04 |
| **apl** | -0.4 | 0 | 0 | 0 | 0.29 | 0.01 | 0 | 0.87 | 0 | 0 |
| **bta** | -0.28 | 0 | 0 | 0 | 0.57 | 0 | 0 | 0.78 | 0 | 0 |
| **clu** | -0.3 | 0 | 0 | 0 | 0.52 | 0.03 | 0 | 0.79 | 0 | 0 |
| **dre** | -0.45 | 0 | 0 | 0 | 0.23 | 0 | -0.02 | 0.86 | 0 | 0 |
| **eca** | -0.37 | 0 | 0 | 0 | 0.48 | 0 | 0 | 0.8 | 0 | 0 |
| **ggo** | -0.18 | 0 | 0 | 0 | 0.59 | -0.01 | 0 | 0.78 | -0.1 | 0 |
| **hsa** | -0.48 | 0 | 0 | 0 | 0.14 | -0.01 | 0 | 0.86 | 0 | 0 |
| **mdo** | -0.22 | 0 | 0 | 0 | 0.67 | 0 | 0 | 0.69 | -0.15 | 0 |
| **mmu** | -0.51 | 0 | 0 | 0 | 0.08 | 0 | 0 | 0.86 | -0.06 | 0 |
| **oan** | -0.19 | 0 | 0 | 0 | 0.81 | -0.06 | 0.03 | 0.55 | 0 | 0 |
| **ocu** | -0.24 | 0 | 0 | 0 | 0.61 | 0 | 0 | 0.76 | 0 | 0 |
| **sha** | -0.33 | 0 | 0 | 0 | 0.6 | 0.01 | 0 | 0.73 | -0.05 | 0 |
| **ssa** | 0.14 | 0 | -0.06 | 0 | 0.74 | 0.01 | 0.02 | 0.63 | -0.16 | 0 |
| **ssc** | -0.14 | 0 | 0 | 0 | 0.7 | -0.04 | 0 | 0.7 | -0.07 | 0 |
| **xtr** | 0.11 | 0 | -0.04 | 0 | 0.56 | -0.02 | 0.03 | 0.81 | -0.13 | -0.04 |
| **dme** | -0.89 | 0 | 0 | 0 | -0.23 | -0.04 | 0 | 0.38 | 0 | -0.08 |
| **cel** | -0.9 | 0 | 0.05 | 0 | -0.21 | -0.01 | 0 | -0.35 | 0 | -0.13 |

**Cuadro 4.1**: **Patrones conservados** ($\gamma = 0{,}025$)**.** *La media de la interacción entre diferentes tipos de posiciones son mostradas para los organismos analizados. EC: posiciones con nucleótidos consenso en el exón, ENC: posiciones con nucleótidos no consenso en el exón, IC: posiciones con nucleótidos consenso en el intrón, y INC: posiciones con nucleótidos no consenso en el intrón.*

la presencia de la secuencia consenso tanto del lado intrónico como del exónico en un mismo sitio está estadísticamente desfavorecida. Este resultado se encuentra en consonancia con estudios previos que encontraron relaciones de epistasis negativa entre ambos lados del 5'ss en mamíferos[38,137]. Los mismos resultados se obtienen si se analiza el modelo con un mayor número de interacciones, con un $\gamma = 0{,}015$ (Tabla Supl.B.1).

## Diferencias entre las especies analizadas

A pesar de la alta conservación del mecanismo general del *splicing* a lo largo de los *eucariotas*, se ha encontrado diferencias entre las especies en cuanto al contenido de información observado en las posiciones del 5'ss[74,144]. Estas diferencias también se reflejan en nuestros datos, siendo la más destacable la disminución de contenido de información de la posición +5 en los vegetales y de la posición +6 tanto en vegetales como hongos 4.2.

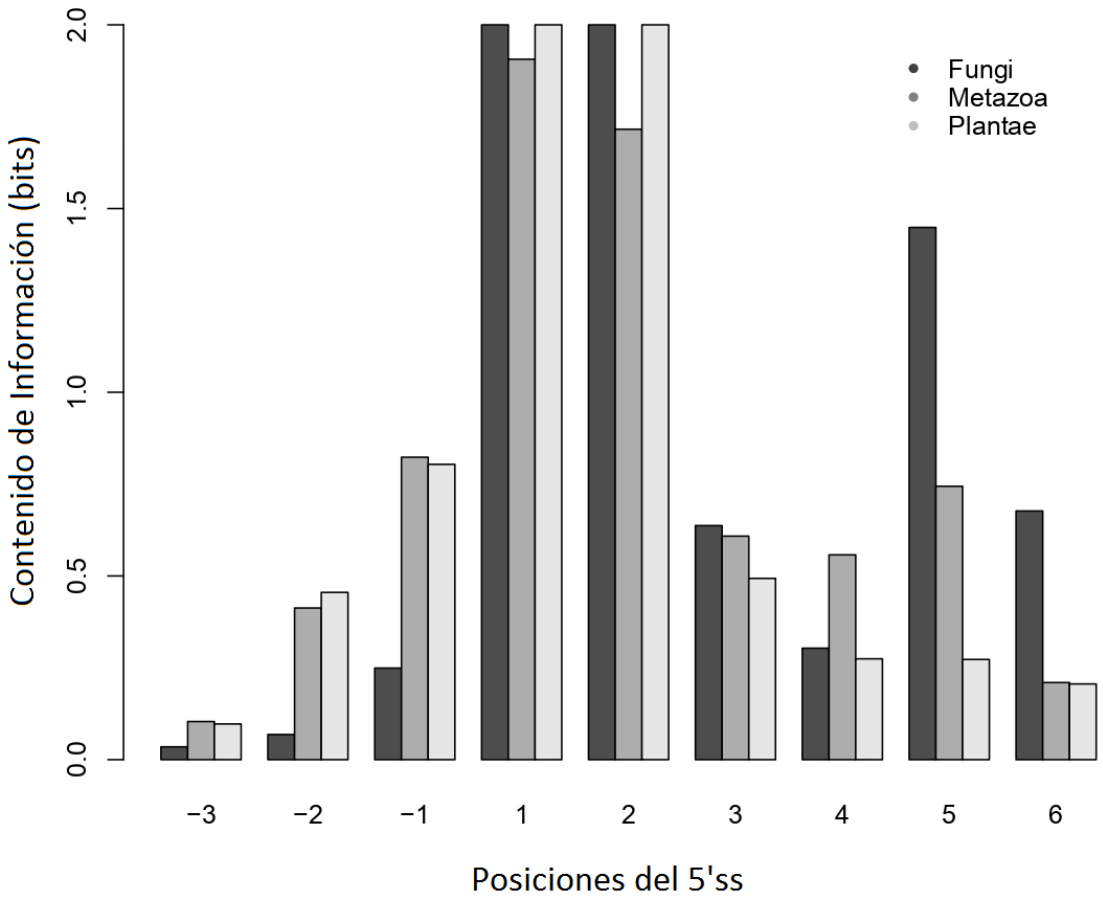



**Figura 4.2**: **Contenido de Información de los 5'ss** *Media del Contenido de Información de cada posición del 5'ss en Fungi, Metazoa y Plantae.*

De la misma manera, también las probabilidades marginales entre pares de posiciones de los 5'ss $P_{ij}$ presentan una variación entre las diferentes especies. En el panel izquierdo

de la Fig.4.3 mostramos un tanglegrama en el cual comparamos un dendrograma obtenido a partir de considerar las $P_{ij}$ para los modelos de $\gamma = 0{,}025$ y el TimeTree para las especies analizadas.

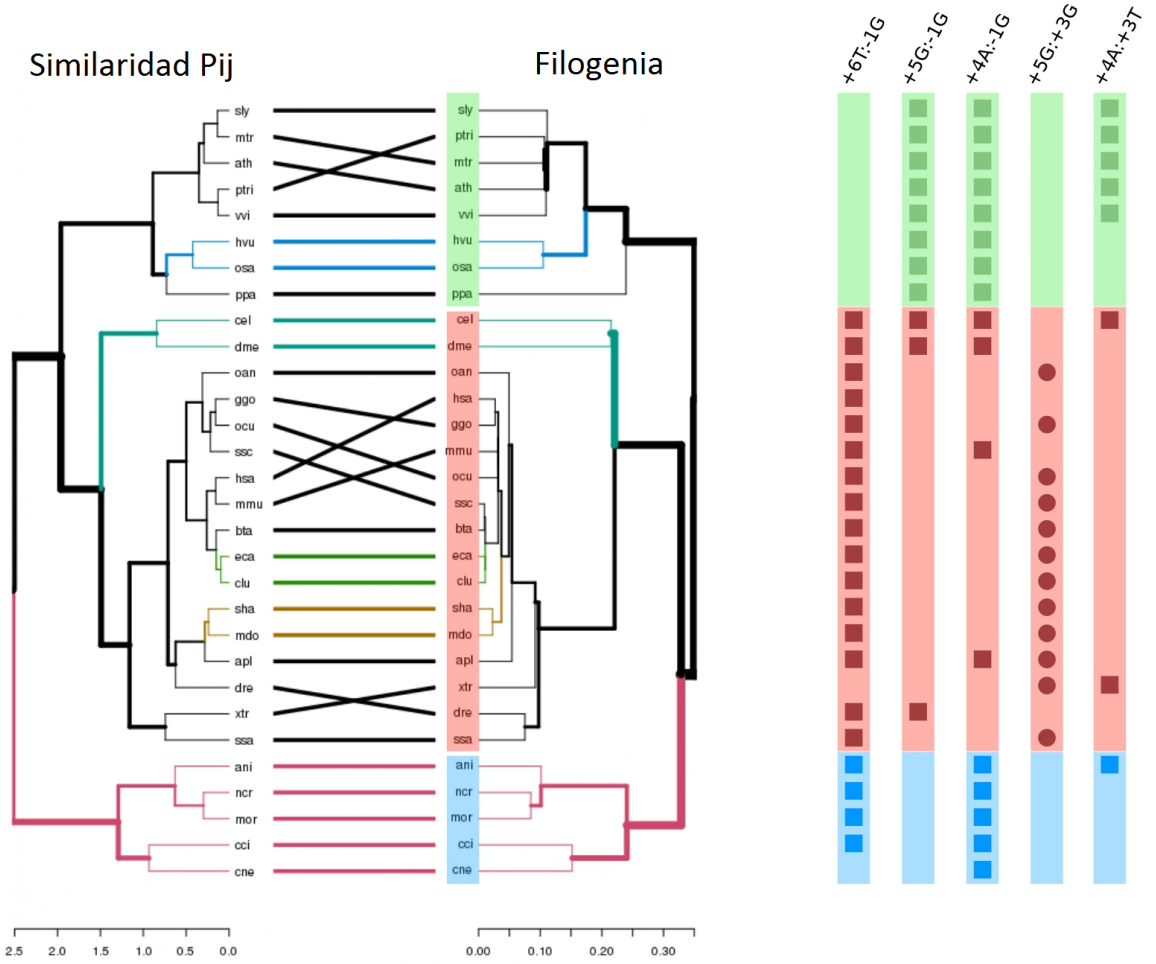



**Figura 4.3**: **Patrones de presencia/ausencia de los parámetros de interacción A**, *Tanglegrama que compara el dendrograma producido a partir de la similaridad en los $P_{ij}$ en las distintas especies y las relaciones filogenéticas que estas guardan entre sí.* **B**, *Matriz de Presencia/Ausencia para los parámetros de interacción que resultaron mostrar una señal filogenética significativa. Con cuadrados y círculos se representa si la interacción es negativa o positiva, respectivamente. Verde: plantas, Rojo: animales, Azul: hongos.*

Ambos ordenamientos presentan una gran concordancia. En particular, puede observarse claramente la separación entre los tres grupos principales: animales, plantas y hongos. En la comparación entre los dos ordenamientos se obtiene una correlación cofrenética de 0,9; y,

por otro lado, los índices de Fowlkes-Mallows resultaron ser significativos ($p_v < 10^{-4}$) para casi todo el rango de grupos de k-clusters ($2 < k < 29$) (Figura 4.4)

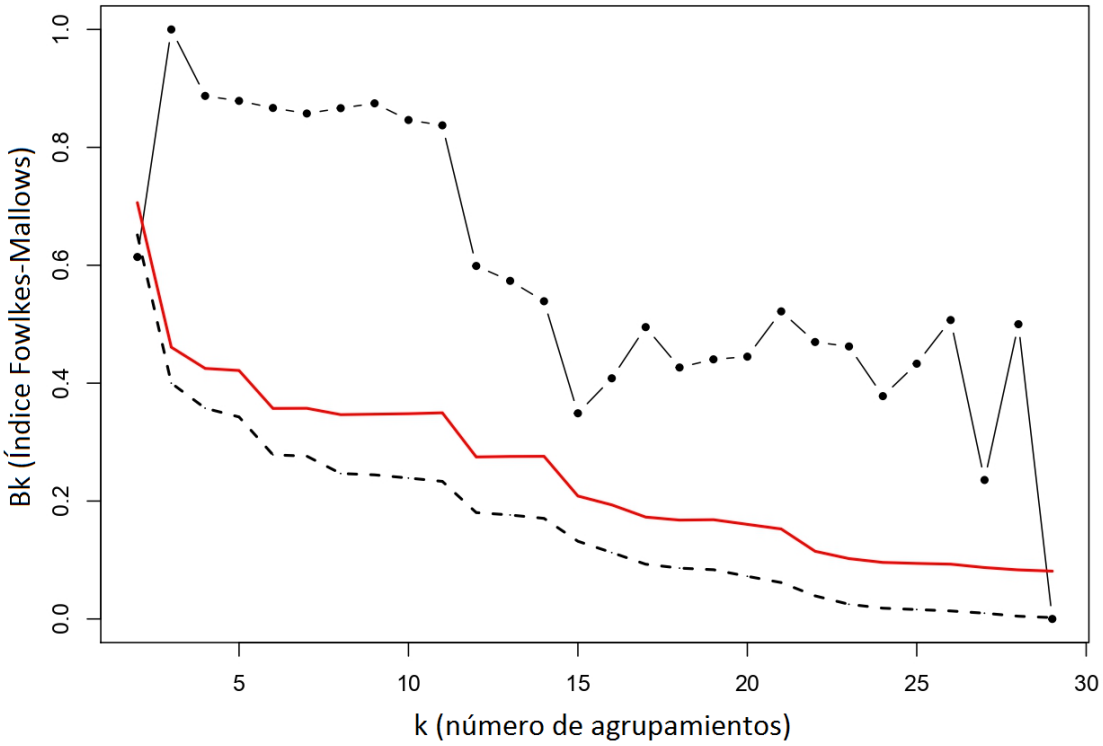



**Figura 4.4**: **Análisis de Fowlkes-Mallows.** *Particiones* K-cuts *se realizaron para los ordenamientos jerárquicos derivados de la filogenia y de los* $P_{ij}$. *Para cada valor de* $k$ *el índice de similaridad de agrupamiento Fowlkes-Mallows,* $B_k$, *fue estimado junto a 10.000 valores para el modelo nulo, obtenidos por rearreglo de las etiquetas. La línea negra punteada corresponde a valores de* $B_k$ *observados en función del número de agrupamientos* $k$. *Los valores esperados bajo la hipótesis nula de no relación entre los ordenamientos son mostrados como una línea negra discontinua. La línea roja continua representa la región de rechazo del 95 % basado en la distribución de valores de* $B_k$. *Todos los cálculos fueron realizados utilizando la función* Bk_plot *del paquete de R* dendextend [52].

## Señal filogenética de los parámetros de interacción

El panel A de la Figura 4.3 sugiere que las estadísticas relacionadas con las interacciones entre pares de sitios pueden llegar a estar vinculadas con las relaciones filogenéticas entre las especies analizadas. A continuación nos preguntamos si en los parámetros de interacción

obtenidos por nuestros modelos puede encontrarse una señal filogenética significativa.

Para cada uno de los 41 parámetros $J_{ij}$, identificados en al menos una especie, llevamos a cabo el procedimiento de Maddison-Slatkin para poner a prueba la presencia de señal filogenética. Encontramos cinco asociaciones estadísticamente significativas (ver Tabla 4.2). El número de pasos evolutivos observados para los caracteres analizados (es decir, para los parámetros) es inferido utilizando la parsimonia de Sankoff[?] y mostrado en la columna *SM.obs*. El mínimo, la mediana y el máximo número de pasos evolutivos detectados en 1.000 muestras de *bootstrapping*, y los p-valores corregidos por el método de Bonferroni son reportados en las columnas *SM.null* y *SM.pv*, respectivamente. Los patrones de presencia/ausencia de estas interacciones son mostrados en el Panel B de la Figura 4.3. Para favorecer la claridad, utilizamos una simplificación de la notación de los parámetros de interacción. Por ejemplo, *-1G:+6T* denota el parámetro $J_{-1,+6}(G,T)$.

Como podemos ver hay ciertas interacciones que se encuentran predominantemente en algunos de los grupos analizados pero no en otros. Las tres primeras interacciones mostradas en la Tabla 4.2 corresponden con interacciones negativas que ponen en relación el nucleótido consenso G en la posición exónica -1 (-1G) con las últimas tres posiciones de la parte intrónica del 5'ss. Mientras que la interacción -1G:+6T se encuentra en casi la totalidad de las especies analizadas de *Metazoa* y *Fungi* pero ausente en plantas, la interacción -1G:+5G se encuentra en todas las especies vegetales pero solo se encuentra en 3 especies de animales y ninguna de hongos. En cuanto a -1G:+4A se encuentra en la totalidad de las especies de plantas y hongos pero muy pocas especies animales. Cabe destacar que de las especies animales que presentan estas dos últimas interacciones, se encuentran las dos especies de invertebrados incluidas en este análisis, *C. elegans* y *D. melanogaster*.

Las últimas dos interacciones en la Tabla 4.2 corresponden a interacciones entre posiciones intrónicas:+3G:+5G y +3T:+4A. Ambas muestran un patrón de aparición mucho más complejo que los ejemplos anteriores. Mientras que la primera se encuentra presente en el 70 % de las especies animales analizadas, la segunda se encuentra solamente en las especies

| Parámetro | Plantas | Metazoos | Fungi | MS.obs | MS.null | MS.pvD |
|---|---|---|---|---|---|---|
| -1G:+6T | 0 | 16 | 4 | 3 | 3,9,10 | 4.1e-3 |
| -1G:+5G | 8 | 3 | 0 | 3 | 4,9,11 | 0 |
| -1G:+4A | 8 | 4 | 5 | 3 | 3,10,13 | 4.1e-3 |
| +3G:+5G | 0 | 12 | 0 | 4 | 4,10,12 | 4.1e-3 |
| +3T:+4A | 5 | 2 | 1 | 3 | 3,7,7 | 5.0e-2 |

**Cuadro 4.2**: **Señal filogenética de los parámetros de interacción**. *Los parámetros estadísticamente significativos según la prueba de Maddison-Slatkin son mostrados en las distintas filas. Para cada uno de estos parámetros se muestra en cuantas especies de plantas, animales y hongos fue detectado. MS.obs es el número de pasos evolutivos de Sankoff observados. En la sexta columna se muestra el mínimo, mediana y máximo de este valor para muestras obtenidas por* bootstraping. *Los p valores corregidos por el método de Bonferroni son reportados en la última columna.*

de plantas dicotiledóneas y en un par de especies de animales y una de hongos.

Estos resultados sugieren que hay señal filogenética en algunos de los parámetros de interacción de nuestros modelos, dando evidencia de los procesos de divergencia que han sufrido a lo largo de la evolución de los distintos clados de eucariotas y dejando abierta la pregunta sobre qué efectos a nivel funcional pueden tener estas diferencias.

## Modelos para los distintos agrupamientos de organismos

Con el objetivo de resaltar los patrones específicos de las interacciones entre pares de posiciones en los 5'ss de plantas, animales y hongos; realizamos un muestreo balanceado de 800.0000 secuencias en plantas y animales, y de 161.547 para hongos. A partir de los muestreos estimamos modelos de máxima entropía para cada uno de estos grupos.

La representación de los modelos resultantes puede verse en la Fig.4.5, mostrando los resultados para dos niveles de regularización: $\gamma = 0{,}015$ y $\gamma = 0{,}025$. Muchas de las interacciones ya discutidas previamente son recuperadas mediante estos modelos. Por ejemplo, las interacciones positivas entre nucleótidos consenso que se encuentran en y otro lado de la señal siguen presentes en los tres grupos. Una interacción negativa que resulta ser común a todos los casos es la -2A:-3T, que relaciona un nucleótido consenso con uno no consenso

dentro de la región exónica.

De la misma manera, también recuperamos las diferencias entre estos grupos que habíamos mostrado con el análisis de Maddison-Slatkin. Una fuerte interacción negativa entre -1G y +6T se puede observar tanto para el caso de animales como para el de hongos, la cual es reemplazada por la interacción -1G:+5G en el caso del modelo hecho para plantas. Por otro lado, la interacción -1G:+4A es detectada en los modelos de plantas y hongos pero ausente en el caso de animales, mientras que la interacción positiva +3G:+5G es solo encontrada en este último grupo. Estas diferencias en los patrones de interacción se mantienen incluso en los modelos más complejos en los que han sido relajados los niveles de regularización, $\gamma = 0{,}015$ (Fig 4.3-B). Por otro lado, estos patrones de interacción observados en estos modelos no cambian de manera significativa si se toman en cuenta solamente las secuencias de los 5'ss que presentan el GT canónico al comienzo del intrón(Fig. A.2).

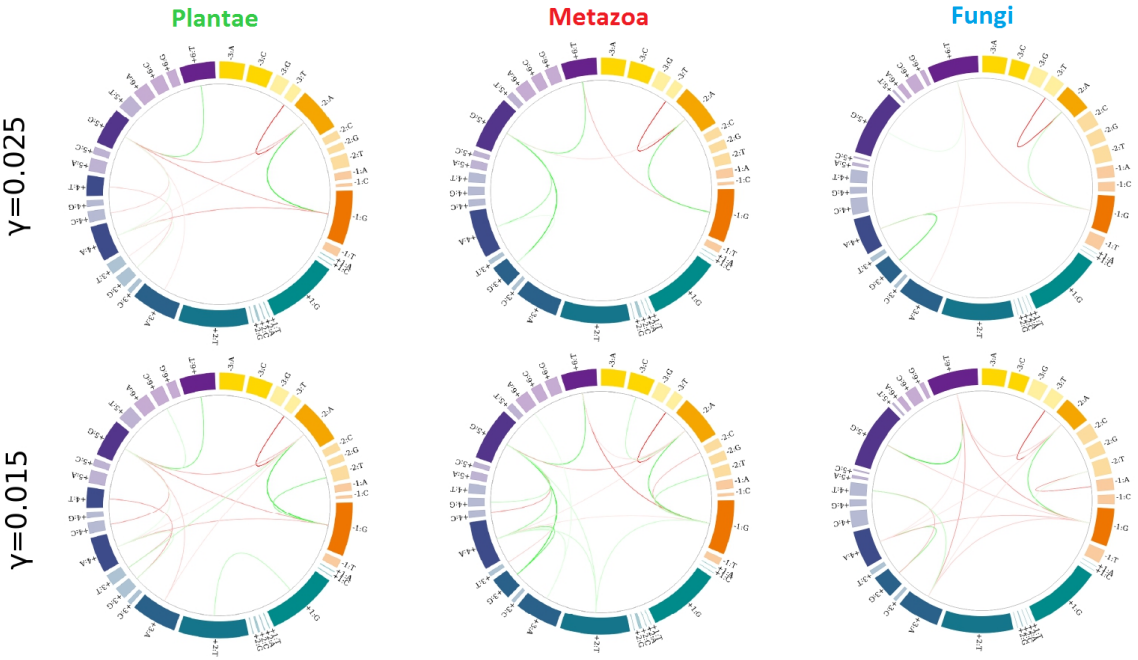



**Figura 4.5**: **Patrones de interacción para los modelos de plantas, animales y hongos**. *Diagrama de circos para los modelos construidos con* $\gamma = 0{,}025$ *(arriba) y* $\gamma = 0{,}015$ *(abajo).*

## 4.3. Conclusión

Nuestros resultados sugieren que los parámetros de interacción de nuestros modelos están asociados a las relaciones filogenéticas establecidas entre las especies analizadas. Estas asociaciones han resultado ser robustas en tanto que no se ven significativamente afectadas por el grado de regularización utilizado para construir los modelos y se ven tanto al analizar las especies por separado como al construir un modelo que reúne secuencias de las distintas especies de plantas, animales y hongos.

Varias de las interacciones detectadas en este trabajo ya han sido reportadas en trabajos previos analizando un numero reducido de especies, principalmente mamíferos como humano, ratón[22,38,137,154,157,177]. Por ejemplo, las interacciones positivas entre las posiciones intrónicas +4:+5 y +5:+6, y la interacción negativa -2:+5 son recuperadas por nuestro modelo y coinciden con los resultados de los trabajos citados. Para destacar es el hecho de que en nuestros modelos encontramos la interacción -1:+5 para las especies vegetales pero, a diferencia de las investigaciones previas, no la encontramos en los metazoos. Esta interacción involucra dos posiciones con alto contenido de información (4.2), especialmente para el caso de animales donde el IC para la posición +5 es bastante más alto que para las plantas. Esta interacción negativa es encontradas en trabajos como el de Yeo y Burge, quienes analizaron 12.700 introns de 1.821 transcriptos distintos[177], o el de Carmel y colaboradores, en el que se infiere una relación de compensación entre -1G:+5G a partir de un análisis comparativo de 8.869 exones homólogos entre humano y ratón[22]. Resulta interesante que en un estudio reciente en el que caracterizan funcionalmente el procesamiento de 32.768 5'ss únicos en tres contextos genéticos diferentes a partir de la técnica de secuenciación masiva en paralelo, encuentran que la interacción entre estas dos posiciones es positiva ya que si alguna de éstas se la muta colocando un nucleótido no consenso para esa posición, el reconocimiento de ese sitio se reduce en un 20 %[173]. Esto podría indicar que el efecto compensatorio que observa el resto de los trabajos podría no estar reflejándose completamente a nivel funcional. En nuestro caso, la interacción -1:+5 no fue relevada ni en el caso del genoma humano (con

50.2197 secuencias) ni en el ensamble de secuencias de animales (con 800.000 secuencias). Ya en las secuencias se observa una baja correlación entre la aparición de -1G y +5G en los datos, ya sea que se considere el conjunto total de secuencias como si se tenga en cuenta solamente las secuencias con el di-nucleótido canónico GT al inicio del intrón (ver A.3).

De acuerdo con nuestro modelo, la relación de compensación entre las posiciones exónicas y las últimas intrónicas se da por un entramado de relaciones entre pares de posiciones que, por un lado, involucra la estabilización de las secuencias consenso mediante las relaciones +5G:+6T y -1G:-2A y, por otro lado, la interacción negativa entre -2A:+5G y -1G:+6T(ver Fig. 4.5). Al menos para el caso de *Homo sapiens*, comprobamos que este entramado de relaciones se mantiene inalterado, incluso si consideramos solamente sitios que tienen evidencia transcripcional de ser reconocidos activamente por la maquinaria de *splicing* (Fig. 3.3). La relevancia de la posición +6, a pesar de su alta variabilidad, ya fue reportada en conección con las aberraciones del *splicing* que se dan en la disautonomía familiar[22]. Cambios en la posición -1 pueden llegar a rescatar el *splicing* aberrante del exón 20 del gen IKBKAP causado por un mal apareamiento del 5'ss con el U1 snRNA en la posición +6.

En plantas detectamos la interacción negativa -1G:+5G que parece reemplazar la interacción -1G:+6T observada en animales(see Fig. 4.5). Cabe destacar que, en el caso de las plantas, las posiciones +5 y +6 presentan un bajo contenido de información: 0.27 y 0.21 bits, respectivamente. Lo que puede sugerir que el entramado de relaciones entre posiciones puede estar jugando un rol clave para mantener la fidelidad en el reconocimiento de los sitios frente a la alta variabilidad que tienen las posiciones individuales de los mismos.

Varias diferencias han sido resaltadas entre los animales y las plantas en cuanto a los detalles relacionados al mecanismo de *splicing*, los más importantes se relacionan con el largo de los intrones y con la prevalencia de un tipo de *splicing* alternativo por sobre otro[24,129]. Nuestros resultados exponen diferencias significativas en los patrones de interacción entre las posiciones de los 5'ss. Las implicancias que estas diferencias pueden tener en cuanto a la historia evolutiva y el funcionamiento del *splicing* requiere de futuras investigaciones.

# Capítulo 5

# Efecto de la mutación de PRMT5 en distintas accesiones de *A. thaliana.*

## 5.1. Introducción

Las modificaciones post-traduccionales (MPT) se encuentran a la base de la gran mayoría de los mecanismos de transducción de señales, siendo una manera adecuada para producir cambios de manera rápida y, en muchos casos, reversible de los componentes que integran los diversos procesos celulares. Estos cambios cambian las propiedades bioquímicas de las proteínas que los sufren, afectando su estabilidad, su localización o las interacciones que pueden establecer. Entre todos los tipos de MPT que se conocen, la metilación de argininas se destaca por varios motivos. Ocurre en prácticamente todas las especies de organismos eucariotas; siendo conocidos sus efectos en la modificación de la estructura de la cromatina, la regulación de la expresión génica y del procesamiento del ARN, respuesta al daño del ADN, entre otros, afectando a una gran cantidad de procesos moleculares y fisiológicos, como la diferenciación celular y la respuesta ante estímulos externos.

Esta modificación es llevada a cabo por una familia de proteínas llamadas PRMT (del inglés *Protein Arginine Metil-Transferases*), la cual se divide en cuatro tipos distintos. El tipo I y II producen la misma monometilación, $\omega - N^G - monometilarginina$ (MMA), pero se diferencian en la dimetilación que producen: mientras que las PRMT que pertenecen al primer tipo producen una dimetilación asimétrica, $\omega - N^G, N'^G - monometilarginina$ (aDMA), las

del segundo tipo producen una dimetilación simétrica, $\omega - N^G, N^G - monometilarginina$ (sDMA). En cuanto a las PRMT del Tipo III, solamente producen la monometilación $\omega - N^G - monometilarginina$. Finalmente, la metilación producida por las del Tipo IV es la más infrecuente, encontrada solamente en *Saccharomyces cereviceae*, e involucra la formación de $\delta - N^G - monometilarginina$[9]. El Tipo I esta integrado por PRMT1, PRMT2, PRMT3, PRMT4/CARM1, PRMT6 y PRMT8; mientras que el Tipo II está representado por PRMT5 y PRMT9. La única metiltransferasa del Tipo III es PRMT7.

La metilación de argininas es una MPT muy común, en tejidos de mamíferos aproximadamente el 0,5 % de las argininas se encuentra metilada. La adición de grupos metilos a los residuos de arginina de una proteína permite su reconocimiento y unión con proteínas que presentan determinados dominios, como Tudor, PHD y WD40. Un tipo de blanco importante de las PRMTs son las histonas[4]. PRMT1 lleva a cabo la dimetilación asimétrica de la arginina 3 de la histona 4 (H4R3), mientras que CARM1 está asociada a la dimetilación H3R26 y H3R42. Todas estas marcas están relacionadas con activación transcripcional. Por otro lado, el efecto contrario tienen las marcas producidas por PRMT5 -H2AR3, H4R3, H3R8- y PRMT6 -H3R2 y H2AR29- las cuales tienen un efecto inhibitorio. Sin embargo, el efecto que las PRMT tienen sobre la transcripción no se limita a su acción directa sobre las histonas. Por ejemplo, PRMT5 metila el factor de elongación de la transcripción SPT5, modificando su interacción con ARN polimerasa II.

El procesamiento del ARN es otro de los procesos en los que las PRMTs ejercen un papel regulatorio muy importante. Muchas proteínas que se unen al ARN (RBPs, por sus siglas en inglés *RNA-binding proteins*) son blanco de mono y di-metilaciones, en especial las pertenecientes a las ribonucleoproteínas nucleares heterogéneas (hnRNP, por sus siglas en inglés). PRMT1 regula la localización y funciones de varias RBPs, como Sam68[31], hnRNP A2[117] o FUS[159]. Algo similar ocurre con PRMT4/CARM1, asociado a la metilación de los factores de splicing SAP49 y U1 snRNP C[27]. Entre las metiltransferesas de Tipo III, PRMT9 metila a SF3B2, uno de los componente del U2 snRNP, mientras que PRMT5

es reconocido como una pieza clave en la maduración de las ribonucleoproteínas pequeñas nucleares (snRNPs) y el mantenimiento de la fidelidad del proceso de splicing. PRMT5 metila a tres de las siete proteínas Sm (D1, D3 y B), siendo esta metilación importante para el reconocimiento de las mismas por parte del dominio Tudor de la proteína SMN (del inglés *survival motor neuron*), la cual promueve la maduración de las snRNPs[111,112]. Además, en *Arabidopsis thaliana* se ha mostrado que PRMT5 también metila a la proteína AtLSm4[37]. Los efectos de la ausencia o falta de función de PRMT5 sobre el *splicing* han sido relacionados con múltiples fenotipos. Un ejemplo importante de esto se ha observado en la regulación del *splicing* de la proteína MDM4, la cual es un represor clave de la vía de p53. La falta de dimetilación de las Sm en ausencia de PRMT5 lleva a la expresión de una isoforma corta de MDM4, la cual es degradada vía NMD, liberando así la vía de p53. Esto ha sido encontrado en varios tipos de cancer, incluyendo melanoma, neoplasias malignas hematológicas[39,58,84] y glioblastoma[136]. También se ha observado patrones de *splicing* aberrantes en transcriptos del gen MYC en células B de linfomas[84]. Por otro lado, los efectos en el splicing de la ausencia o inhibición de PRMT5 impide el reparado del ADN mediante recombinación homóloga, llevando al arresto del ciclo celular y apoptosis[64]. Por todo esto, PRMT5 ha tomado una gran relevancia clínica en el estudio y posible tratamiento de diferentes tipos de cáncer[63].

En plantas también han sido caracterizados los efectos de PRMT5, con la enorme ventaja de que en estos organismos, los mutantes que carecen de esta proteína son viables, a diferencia de lo que ocurre en los mamíferos. PRMT5 está involucrado en múltiples procesos, como el control del tiempo de floración, fotomorfogénesis, ritmos circadianos y respuesta a estrés salino; tanto por su efecto en la transcripción mediado por su acción epigenética como por su relación con el *splicing*. Análisis comparativos de los mutantes *prmt5* y *prmt4*, muestran grandes similitudes en cuanto a la alteración del tiempo de floración, defectos en las vías de señalización relacionadas con luz y la reducción en la tolerancia a sal. Estas similitudes se relacionan con efectos similares también en cuanto a los genes cuya expresión se ve afectada y de los que presentan patrones de *splicing* alterados[68]. En el caso de los eventos de *splicing*

alternativo asociados a la mutación de PRMT5, pero no a la de PRMT4, se observó un enriquecimiento de 5'ss débiles que se alejaban de la secuencia consenso. Esto indica que el efecto de PRMT5 sobre el *splicing* puede deberse, al menos en parte, a la estabilización de las uniones ARN-ARN que se establecen entre 5'ss débiles y la porción de snRNA de U1 que es complementaria a la secuencia consenso de 5'ss[68].

Como ya se ha destacado anteriormente, las señales en *cis* tienen un rol preponderante en el proceso de *splicing* y su regulación. Varios estudios indican que las diferencias que se encuentran en los patrones de *splicing* alternativo tanto a nivel intra como interespecífico se deben principalmente a variaciones de los elementos regulatorios en *cis*[10,87,109,138]. Por ejemplo, estudios sobre la divergencia del splicing en transcriptomas de cerebro de humanos y chimpancés determinaron que la evolución de las señales regulatorias en *cis* realizan la mayor contribución en el surgimiento de patrones de *splicing* especie-específicos[98]. Conclusiones similares alcanzan estudios que comparan diferencias intraespecíficas entre cepas de ratones[55] o entre accesiones de *Arabidopsis thaliana*[168].

Con el objetivo de entender la relación que existe entre la acción de PRMT5 y la secuencia que regulan en cis el proceso de *splicing*, en este estudio comparamos los efectos que tiene la mutación *prmt5* en dos accesiones diferentes de *Arabidopsis thaliana*: Columbia (Col-0) y Landsberg erecta (Ler). En este sentido, buscamos relacionar los efectos que la mutación tiene sobre la transcripción y los patrones de *splicing* con las diferencias que existen entre estas accesiones en cuanto a sus secuencias genómicas, poniendo especial énfasis en los polimorfismos de nucleótido único (SNP, del inglés *Single Nucleotide Polymorphism*) presentes en las secuencias de los 5'ss. Para esto se planteó un diseño experimental de dos factores en el cual se realizó un experimento de RNA-Seq de plantas salvajes y mutantes para PRMT5 de cada una de las accesiones antes mencionados, Col-0 y Ler (Figura 5.1). Por otro lado, para poder discriminar las diferencias en los patrones de *splicing* originadas por efectos en *cis* o en *trans*, se analizó mediante RNA-Seq los transcriptomas de plantas producto de las cruzas entre Col-0 y Ler-1, tanto salvajes como mutantes.

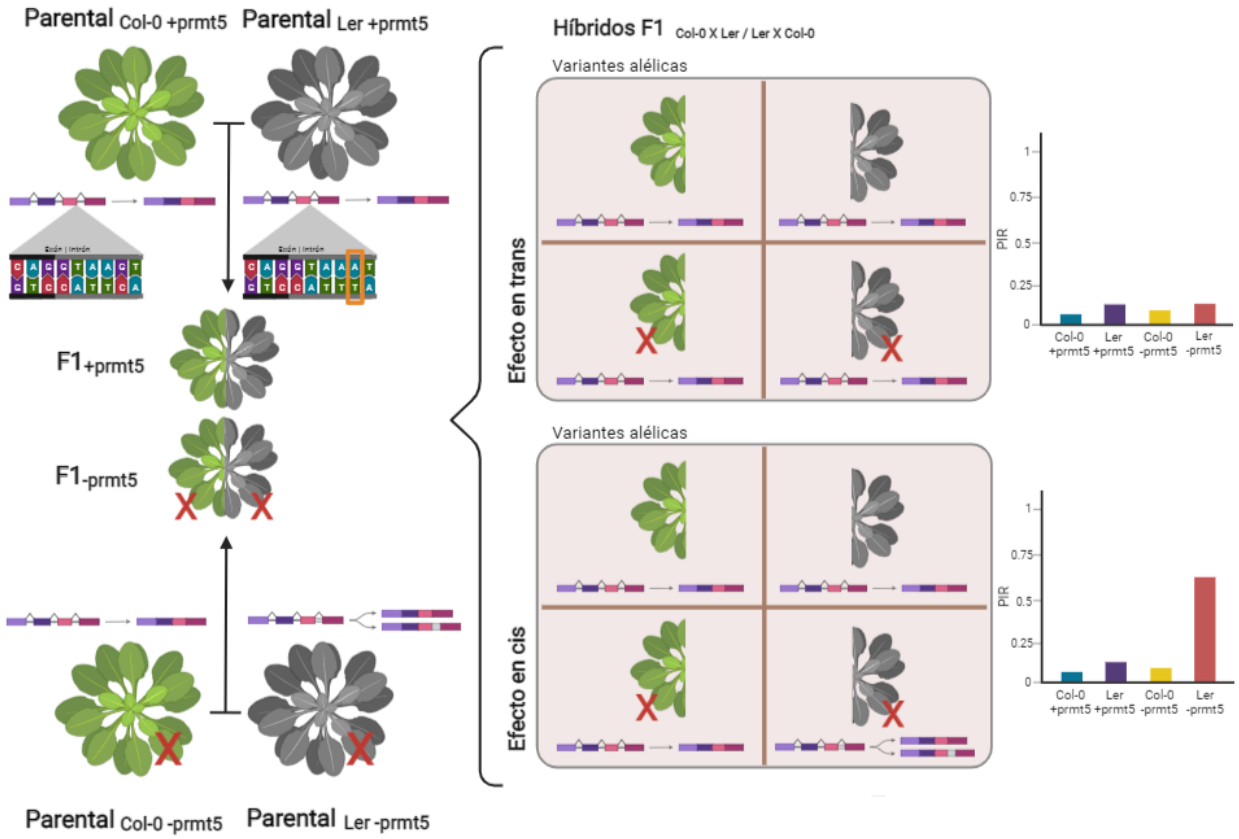



**Figura 5.1**: **Esquema general del experimento**. *A partir de plantas salvajes (+prmt5) y mutantes (-prmt5, cruz roja en el esquema) procedentes de dos accesiones distintas de A. thaliana, Columbia (Col-0, planta verde en el esquema) y Landsberg erecta (Ler, planta gris en el esquema), se produjo mediante cruzamiento plantas híbridas F1 Col-0 X Ler y Ler X Col-0 (plantas de ambos colores). Gracias a las diferencias en las secuencias genómicas presentes en las accesiones (SNP/Indels) podemos identificar en los experimentos realizados de RNAseq cuáles lecturas provienen de los alelos de cada uno de ellos. De esta manera podemos calcular la métrica PIR (Percentage Intron Retention) para cada alelo por separado y evaluar el efecto interacción en los eventos de splicing. En las plantas híbridas, los alelos procedentes de las accesiones se encuentran en un mismo ambiente celular, en particular, están expuestos a la misma cantidad y cualidad de reguladores de splicing. Si los eventos de interacción que se observan en las plantas parentales también se registran en las plantas híbridas es indicio de que el efecto se da por diferencias en cis, por ejemplo, debido a las variaciones de secuencia entre una accesión y otra. En cambio, si el efecto interacción desaparece en las plantas híbridas, podemos decir que las diferencias en cis no alcanzan para explicar el resultado visto en los parentales. Por lo que ese evento de interacción se deberá interpretar como producido por un efecto en trans.*

## 5.2. Materiales y métodos

### Crecimiento de plantas, extracción y secuenciación de ARN

Se sembraron semillas de los distintos genotipos en medio Murashige-Skoog conteniendo 0.8 % de agar. Las semillas fueron estratificadas por 4 días en oscuridad a 4 °C, y luego fueron cultivadas bajo luz blanca continua a 22 °C. Luego de 9 días, se cosecharon plántulas enteras de cada genotipo, y se extrajo el ARN total utilizando el *RNeasy Plant Mini Kit* de QIAGEN, siguiendo las instrucciones del fabricante. Luego, se prepararon bibliotecas de ADNc siguiendo las instrucciones del *TruSeq RNA Sample Preparation Guide* de Illumina. Brevemente, a partir de 3 $\mu$g de ARN total se purificó ARNm polyadenilado, el cual luego de ser fragmentado fue utilizado para la síntesis de ADNc utilizando transcriptasa reversa (SuperScript II; Invitrogen) y hexámeros aleatorios. Finalmente se agregaron adaptadores específicos para cada muestra y las bibliotecas de ADNc fueron secuenciadas en la plataforma Illumina GAIIx, obteniéndose secuencias de 100 pares de bases. Las secuencias obtenidas fueron analizadas con la versión 1.3 del *pipeline* Illumina, y filtradas por calidad utilizando procesos usuales de Illumina. Los archivos de secuencias resultantes fueron generados en el formato FASTQ.

### Introgresión de la mutación *prmt5-5*, obtenida en el background Col-0, al background Ler

La mutante *prmt5-5* obtenida en el background Columbia (Col-0) fue cruzada durante 7 generaciones con plantas salvajes de la accesión Landsberg erecta (Ler), verificando la eficacia de la introgresión luego de cada cruza.

### Obtención de híbridos F1 de plantas salvajes ( WT Col-0 x WT Ler) y de mutantes (*prmt5-5* Col x *prmt5-5* Ler)

Para obtener backgrounds híbridos de plantas salvajes (F1 Col-0 x Ler), así como del mutante *prmt5-5* (F1 *prmt5-5* Col x *prmt5-5* Ler), se realizaron decenas de cruzas de plantas

salvajes de la accesión Col-0 con plantas salvajes de la accesión Ler, así como de plantas mutantes *prmt5-5* en el *background* Col-0 con mutantes *prmt5-5* en el *background* Ler, y se utilizaron las semillas y plántulas resultantes de dichas cruzas para la caracterización de los transcriptomas de los distintos genotipos por secuenciación masiva (RNA-seq).

## Mapeo de las lecturas de RNA-Seq

Previo al paso de mapeo de las lecturas, se verificó la calidad de los archivos fastq utilizando *TrimGalore* (https://github.com/FelixKrueger/TrimGalore), mediante la detección y remoción de secuencias adaptadoras con *Curadapt*[108] y el control de calidad de las mismas mediante FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc).

Para el mapeo de las lecturas elegimos el alineador *STAR v2.7.9a* debido a su buen rendimiento y precisión en el alineamiento de lecturas provenientes de junturas[41]. Fueron consideradas solo las lecturas que lograron ser mapeadas a una única posición del genoma. Se utilizó el modo "2-pass" de *STAR* para mejorar la sensibilidad en la detección de nuevas junturas, no coincidentes con las coordenadas de las isoformas presentes en la anotación del genoma utilizada.

Para las lecturas que provienen de la accesión Col-0 se utilizó el genoma de referencia TAIR10. Para el caso de las lecturas provenientes de la accesión Ler, utilizamos una estrategia ya previamente descripta[171]. Creamos un genoma de referencia partiendo del genoma de Col-0 y utilizando la información de las variantes de nucleótido único (SNP) y de inserción/deleción (Indel) de Ler presentes en el sitio del Proyecto *1001 Genomes* (https://1001genomes.org/). La secuencia del "pseudo-genoma" es inferida a partir de reemplazar el alelo de referencia por el alelo alternativo de Ler, utilizando la función *pseudogeno* implementada en el software *GEAN*[149]. En cuanto a la anotación de las características genómicas, construimos archivos GFF (del inglés, *general feature format*) a partir de la proyección de las coordenadas genómicas presentes en TAIR10 en las coordenadas de la accesión Ler, mediante la función *liftgff* de *GEAN*.

En el caso de las lecturas provenientes de los híbridos F1, las mismas fueron alineadas utilizando tanto TAIR10 como el pseudo-genoma de Ler. Cada lectura fue identificada con un determinado parental según en cuál de estos dos alineamientos obtuvo la calidad más alta. Solo las lecturas que pudieron ser asignadas a alguno de los parentales de forma inequívoca fueron utilizadas para el análisis de *splicing* alternativo alélico.

Para el procesamiento de los archivos BAM creados en los alineamientos arriba detallados, se utilizó las herramientas que provee el *software samtools*[96]. Fueron consideradas solamente las lecturas que luego del alineamiento se encuentren debidamente pareadas y con una calidad de mapeo superior a 20.

## Análisis de expresión diferencial de genes y *splicing* alternativo

El conteo de la cantidad lecturas que se superponen con distintas *features* genómicas, ya sea genes como exones/intrones o junturas, fue realizado con el paquete de R *ASpli*[103]. Este mismo paquete fue utilizado tanto para la detección de los genes diferencialmente expresados (DEG) como para el uso diferencial de bines en las comparaciones entre plantas salvajes y mutantes *prmt5-5*. En estas comparaciones se tomó como criterio de significancia un *Fold Change*(FC) mayor a 1.5 y un *False Discovey Rate*(FDR) menor a 0.05.

En el análisis de *splicing* diferencial en la comparación entre las dos accesiones se utilizó un test exacto de Fisher para comparar los valores de PSI/PIR entre las dos accesiones o alelos de los híbridos. Los p-valores obtenidos fueron ajustados por múltiple testeo mediante el método de Benjamini–Hochberg (BH). Como criterio de significancia se tomó un q-value menor a 0.1 en las tres réplicas y un $|\Delta PIR/PSI|$ promedio mayor a 0.1.

Para estimar el efecto de interacción entre la accesión (Col-0 o Ler) y el genotipo (salvaje o mutante de *PRMT5*) se utilizó el método de Altman y Bland[6,55,109,168]. El *ratio* entre los valores de PSI/PIR de las plantas salvajes y mutantes fue comparado entre ambas accesiones, tanto para las plantas parentales como para el caso de los híbridos F1 Col-0 X Ler y Ler X Col-0. En cada una de estas comparaciones se calculó el error estándar de la diferencia de

los ratios entre las dos accesiones, el cual fue utilizado para la construcción de valores-z y los correspondientes p-valores. Estos últimos fueron ajustados mediante el método de BH, tomando como significativos aquellos que obtuvieron un q-value menor a 0.05.

## Análisis de enriquecimiento de *pathways* de *KEGG*

KEGG (*Kyoto Encyclopedia of Genes and Genomes*) es una base de datos con información sobre una gran diversidad de funciones y vías metabólicas o *pathways* presentes en los sistemas biológicos; conteniendo información sobre los diversos componentes y sus funciones biológicas, así como también una red de sus interacciones moleculares. Con el fin de entender el significado biológico de los genes que resultaron diferencialmente expresados (DEG) en la comparación entre plantas salvajes y mutantes para PRMT5, hicimos un análisis de enriquecimiento de *pathways* de KEGG, utilizando el paquete de R *clusterProfiler*. En este análisis lo que se busca es determinar si el conjunto de DEG encontrados están significativamente enriquecidos en algunos de los conjuntos de genes reunidos en los diferentes *pathways* que conforman KEGG. Se ajustaron los p-valores según el método de BH y se tomó como significativos aquellos que superaran un valor de 0.05. Para la representación de las vías metabólicas y los genes diferencialmente expresados en ellas se utilizó el paquete de R *pathways*.

## Análisis de motivos de RBPs y SNPs

Para la determinación de la presencia de motivos asociados a proteínas de unión a ARN (RBP, del inglés *RNA Binding Protein*) en los eventos de *splicing* relacionados con la mutación de *PRMT5*, se utilizó el servidor web *rMAPS2* [71,123]. Para la obtención y manipulación de las secuencias genómicas correspondientes a *A. thaliana* se usaron los paquetes de R: *GenomicFeatures*, *Biostrings*, *rtracklayer*. Para el análisis y manipulación de la información referida a las variaciones genómicas entre Col-0 y Ler se utilizó el paquete de R *VariantAnnotation*.

## 5.3. Resultados

### Variabilidad genómica entre las accesiones Col-0 y Ler relacionada con los sitios de *splicing*

Los SNPs presentes en la accesión Ler fueron obtenidas a partir del Proyecto *1001 Genomes*, contabilizándose un total de 568.741 SNP/Indels. A partir de la anotación del genoma de *A. thaliana* podemos explorar en qué regiones del genoma se encuentran estas variaciones de secuencia y entender su posible impacto funcional. En la Figura 5.2 podemos observar que existen cerca de 250.000 variaciones de secuencia en las regiones promotoras de los genes, mientras que las que se localizan dentro de las regiones codificantes de los mismos son 130.747 y dentro de intrones, 152.998. Para obtener los SNP/Indels que se ubican dentro de las regiones correspondientes a sitios de *splicing*, se definieron éstas de la siguiente forma: para los 5'ss se tuvo en cuenta las últimas tres posiciones exónicas (-3 a -1) y las primeras 6 intrónicas (1 a 6), mientras que para los 3'ss se consideraron las últimas 12 posiciones del intrón y la primera del exón siguiente. Así definido, se obtuvo que en Ler hay 3.615 SNPs en los 5'ss y 8.426 en los 3'ss.

En la Figura 5.3 puede observarse la secuencia logo para los 5'ss y 3'ss para los genes que resultaron expresados en la condición salvaje de Col-0 (de un total de 20.289 genes se obtuvieron 153.213 sitios de *splicing*). Por otro lado, se puede ver el porcentaje de SNPs que se ubican en cada una de las posiciones de los sitios. Por ejemplo, de los 3.615 SNPs que se encuentran en los 5'ss, el 11 % se ubica en la posición -3 mientras que en la posición 1 hay menos del 0.01 %. En ambos sitios de *splicing* se observa una menor cantidad de SNPs en las posiciones que están más conservadas. Esto se ve claramente en los 3'ss en donde la presencia de SNPs en las posiciones -2 y -1 es muy poco frecuente en relación al resto de las posiciones. Esto se relaciona con el posible impacto funcional que puede tener un cambio en la secuencia en las posiciones más conservadas, modificando la probabilidad de que el sitio pueda ser reconocido por la maquinaria de *splicing*.

Concentrándonos en los 3.615 SNPs presentes en los sitios donores de *splicing*, podemos

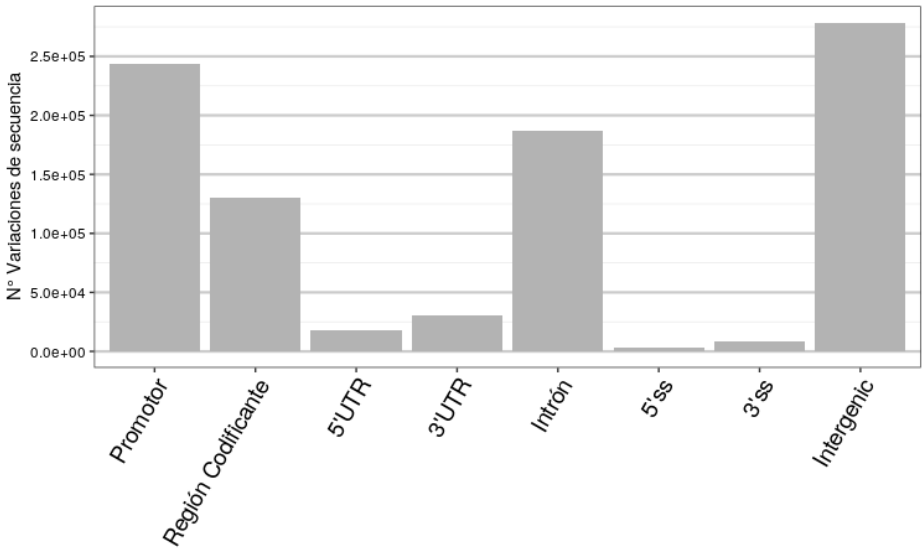



**Figura 5.2**: **Variaciones en las secuencias genómicas entre Col-0 y Ler**. *Cantidad de SNP/Indels cuyas coordenadas genómicas coinciden con regiones con distintas funciones anotadas: promotores, regiones codificantes, intrones, etc. La región promotora fue definida como el intervalo que abarca 1000 bp río arriba del comienzo de la transcripción y 100 bp río abajo de la misma.*

calcular el efecto que estos SNPs tienen en relación a la escala de energía presentada en el primer capítulo de esta tesis, utilizando el modelo correspondiente a *Arabidopsis thaliana*. Esta medida de energía de las secuencias de los 5'ss nos permite establecer una escala de variabilidad. Como podemos ver en la Figura 5.4, la gran mayoría de los SNPs presentes en sitios donores de *splicing* generan un cambio pequeño en nuestra escala de energías. Esto se relaciona con el hecho ya mencionado de que la mayoría de los SNPs se encuentran en posiciones más variables de los 5'ss, con lo cual su efecto funcional podría no ser tan abrupto. Sin embargo, también podemos notar que hay un conjunto de SNPs que producen cambios importantes en la escala de energía. En estos casos podríamos esperar que el correlato funcional de estos cambios sea mayor.

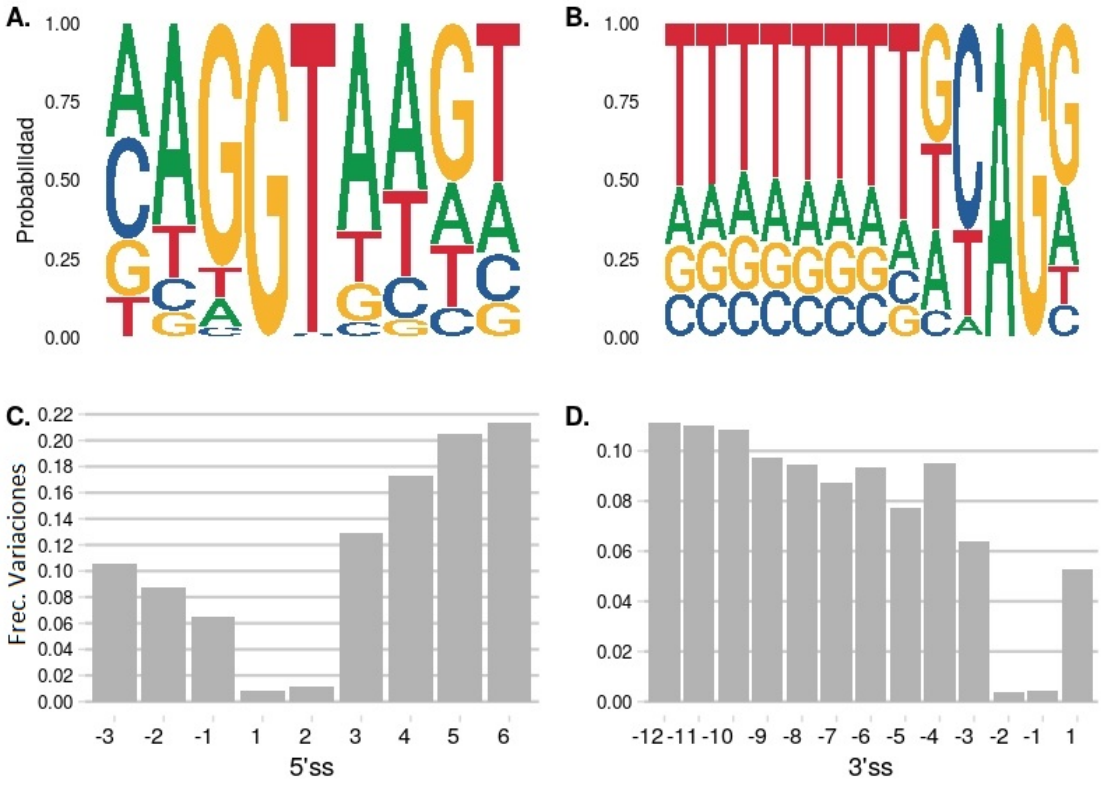



**Figura 5.3**: **SNPs en sitios de *splicing***. ***A, B*** *Secuencias Logo para los sitios 5'ss y 3'ss de Arabidopsis thaliana, respectivamente. Se consideran 153.213 secuencias pertenecientes a sitios de splicing relacionados con genes que se encuentran expresados en las muestras salvajes de Col-0.* ***C, D:*** *Número de SNPs entre Col-0 y Ler mapeados en los sitios de splicing, distinguiendo la posición dentro de los mismos en la que se encuentran. En total hay 3.615 y 8.426 SNPs en los 5ss y 3ss, respectivamente.*

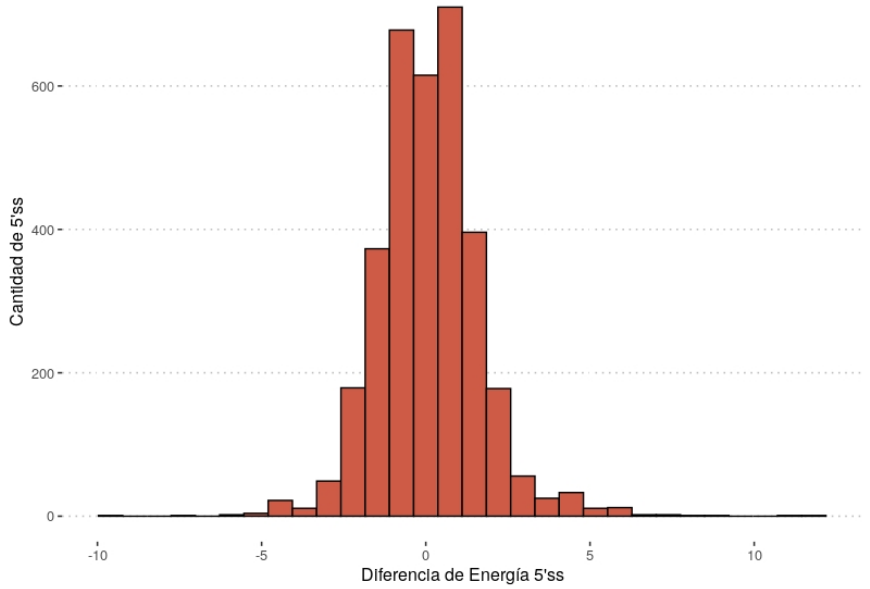



**Figura 5.4**: **Diferencia de energía de 5'ss con variaciones de secuencia entre Col-0 y Ler**. *Distribución de las diferencias de energías para los 5'ss en Col-0 y Ler con variaciones de secuencia entre estas accesiones, calculadas según el modelo estadístico desarrollado en la primera parte de esta tesis para A.thaliana. Para el caso de las secuencias de los 5'ss correspondientes a la accesión Ler, se reemplazó el alelo de referencia por el alternativo según la información extraída de 1001Genomes.*

## Efecto de la mutación de *PRMT5* en Col-0 y Ler

Plantas salvajes y mutantes *prmt5-5* de las accesiones Col-0 y Ler fueron recolectadas para realizar un experimento de RNA-Seq, utilizando tres réplicas biológicas para cada una de las condiciones. Las lecturas provenientes de plantas de la accesión Col-0 fueron mapeadas al genoma de referencia TAIR10, mientras que las pertenecientes a la accesión Ler fueron mapeadas contra un pseudogenoma generado a partir de la anotación de TAIR10 y la información de SNP/Indel de esta accesión, extraída de *1001 Genomes*. En promedio, aproximadamente 54 millones de lecturas fueron alineadas en una única posición del genoma (Apéndice B, Ref. B).

Para todos los análisis que realizamos, excluimos de los mismos todas las zonas del genoma en las que se den una superposición de elementos genómicos, como pueden ser transcriptos anti-sentido, para evitar posibles errores en la cuantificación de la expresión génica o en el uso de los bines. Para la determinación de los genes diferencialmente expresados (DEG, por sus siglas en inglés) se tomó dos criterios de significancia: FC (en inglés, *Fold Change*) mayor a 1.5 y FDR menor a 0.05. Considerando estos criterios se encontraron 1903 DEG en la comparación entre plantas salvajes y mutantes para *PRMT5* en la accesión Col-0, y 2508 DEG para el caso de la accesión Ler. Entre ambos conjuntos hay 975 genes en común. Este número representa el 51,2 % y 38,9 % de los DEG encontrados en Col-0 y Ler, respectivamente. En ambos casos, se registra un mayor número de genes cuya expresión aumenta en el mutante respecto al estado salvaje. Al mismo tiempo que resulta muy bajo el número de genes que muestran una tendencia opuesta entre ambas accesiones, es decir, que se encuentran sobre-expresados en las muestras mutantes en una accesión pero subexpresados en la otra(Fig. 5.5 A).

Para realizar un análisis de los procesos biológicos en los que están involucrados los genes cuya expresión se encuentra afectada por la mutación de *PRMT5*, se realizó un análisis de enriquecimiento de *pathways* de *KEGG*, mediante el paquete de R *clusterprofiler*. Se consideró como criterio de significancia un p-valor corregido por múltiple testeo mediante

el método de BH menor a 0.05. En ambas accesiones se encuentran un conjunto similar de *pathways* afectados por la mutación (Figura 5.5 B y C). Podemos destacar una aumento de la expresión en el mutante de genes relacionados con el ciclo de *splicing* y con la degradación del ARN. Así como también una disminución de la expresión de genes relacionados con la transducción de señales hormonales y el proceso de fotosíntesis, entre otros.

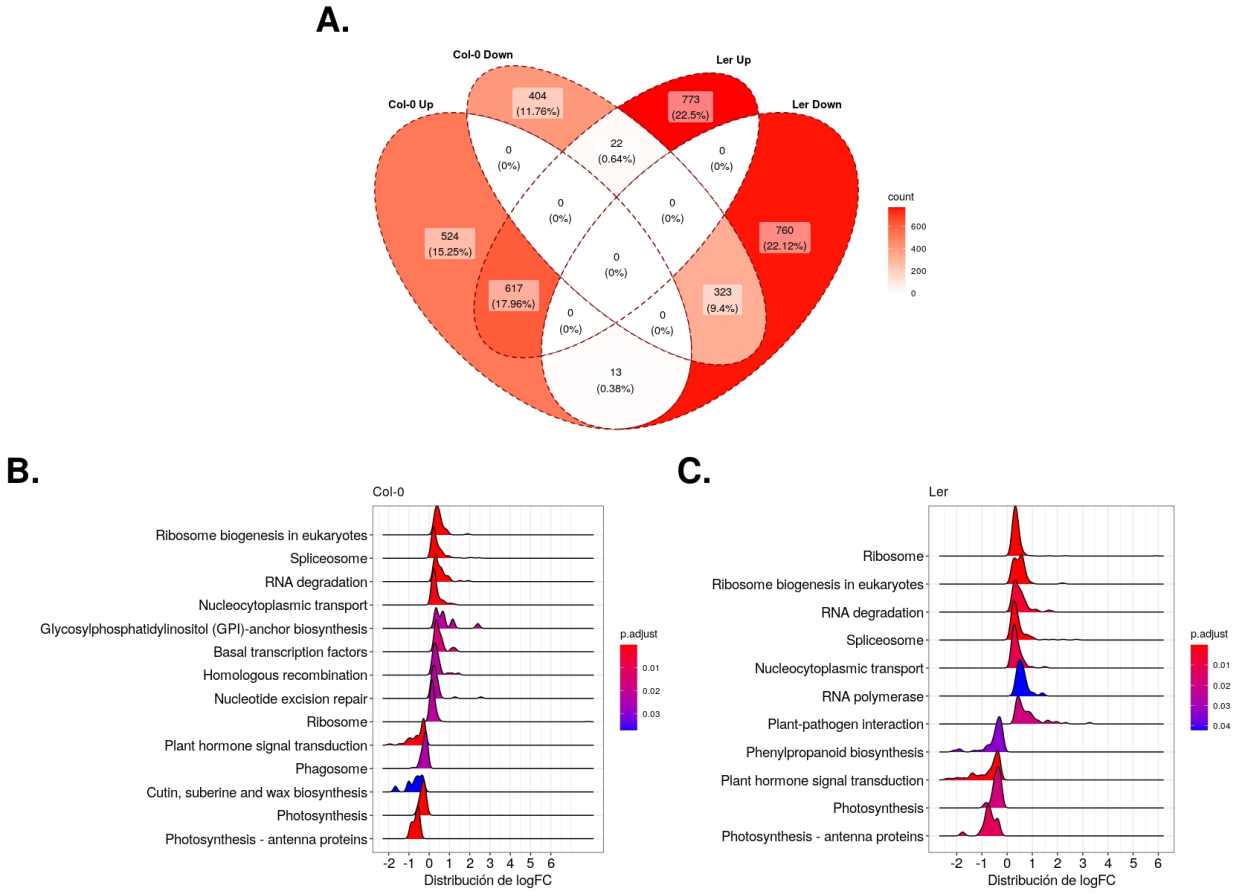



**Figura 5.5**: **Genes Diferencialmente Expresados (DEG).** *Comparación entre plantas salvajes y mutantes para PRMT5 en ambas accesiones, Col-0 y Ler.* **A.** *Diagrama de Venn de los DEG encontrados en el análisis. Se muestra los conjuntos de genes cuya expresión resultó significativamente mayor en el mutante (Col-0 Up y Ler Up) y en los que se encontró que la misma fue menor Col-0 Down y Ler Down).* **B.** *y* **C.** *Distribución de logFC de los genes cuyas vías metabólicas resultaron significativas en el análisis de enriquecimiento en KEGG para Col-0 y Ler, respectivamente.*

Si nos concentramos en la expresión de los genes que codifican para los principales componentes del spliceosoma, podemos observar que en ambas accesiones se da un aumento

de la expresión de estos genes en el mutante respecto al salvaje. Como puede verse en la Figura 5.6, los genes cuya expresión se encuentra afectada corresponden a componentes del spliceosoma que participan en las distintas etapas del ciclo de *splicing*. Este efecto de la mutación de *PRMT5* parece ser similar tanto en Col-0 como en Ler (ver Figura suplementaria A.4), siendo la correlación de los FC obtenidos para los genes que pertenecen a componentes del spliceosoma en Col-0 y Ler de 0,83 (calculada a partir de 80 genes pertenecientes a este *pathway* y expresados en ambas accesiones).

En cuanto a los cambios en los patrones de *splicing* alternativo que se dieron en los mutantes respecto a las muestras salvajes, se tomaron los siguientes criterios de significancia: (1) un FC mayor a 1.5 en la cobertura del bin involucrado en el evento de *splicing*; (2) FDR asociado a ese cambio menor a 0.05; y (3) $\Delta PIR/PSI$ entre las condiciones mayor a 0.1. De esta manera se logró identificar 1130 eventos de *splicing* diferenciales para la accesión Col-0, y 947 para el caso de Ler. Nuevamente encontramos una alta coincidencia en los cambios que se producen en ambas accesiones, siendo la intersección de 696 eventos; representando un 61.6 % y 73.5 % de los eventos que cambian en Col-0 y Ler, respectivamente (Fig. 5.7.A). El tipo de *splicing* alternativo que predomina ampliamente entre los eventos diferenciales observados es la retención de intrón (IR), representando un 92.1 % en Col-0 y del 94.7 % en Ler.

Como mencionamos en la Introducción, en trabajos previos[68] se observó que los sitios donores de *splicing* asociados a eventos diferenciales en mutantes de *PRMT5* tendían a ser sitios débiles, es decir, que sus secuencias se alejan de la secuencia consenso esperada para los 5'ss de *A.thaliana*. Para poder determinar si en estas muestras puede observarse la misma tendencia, se utilizó el modelo energético desarrollado en la primera parte de esta tesis para caracterizar los 5'ss. En modo de recordatorio, dentro del marco de los modelos de máxima entropía realizados para las secuencias de los 5'ss, una alta energía corresponde a secuencias que tienen una baja frecuencia a lo largo del genoma y, por lo tanto, a sitios que se alejan de la secuencia consenso. En la Figura 5.7.B podemos ver la distribución de energías

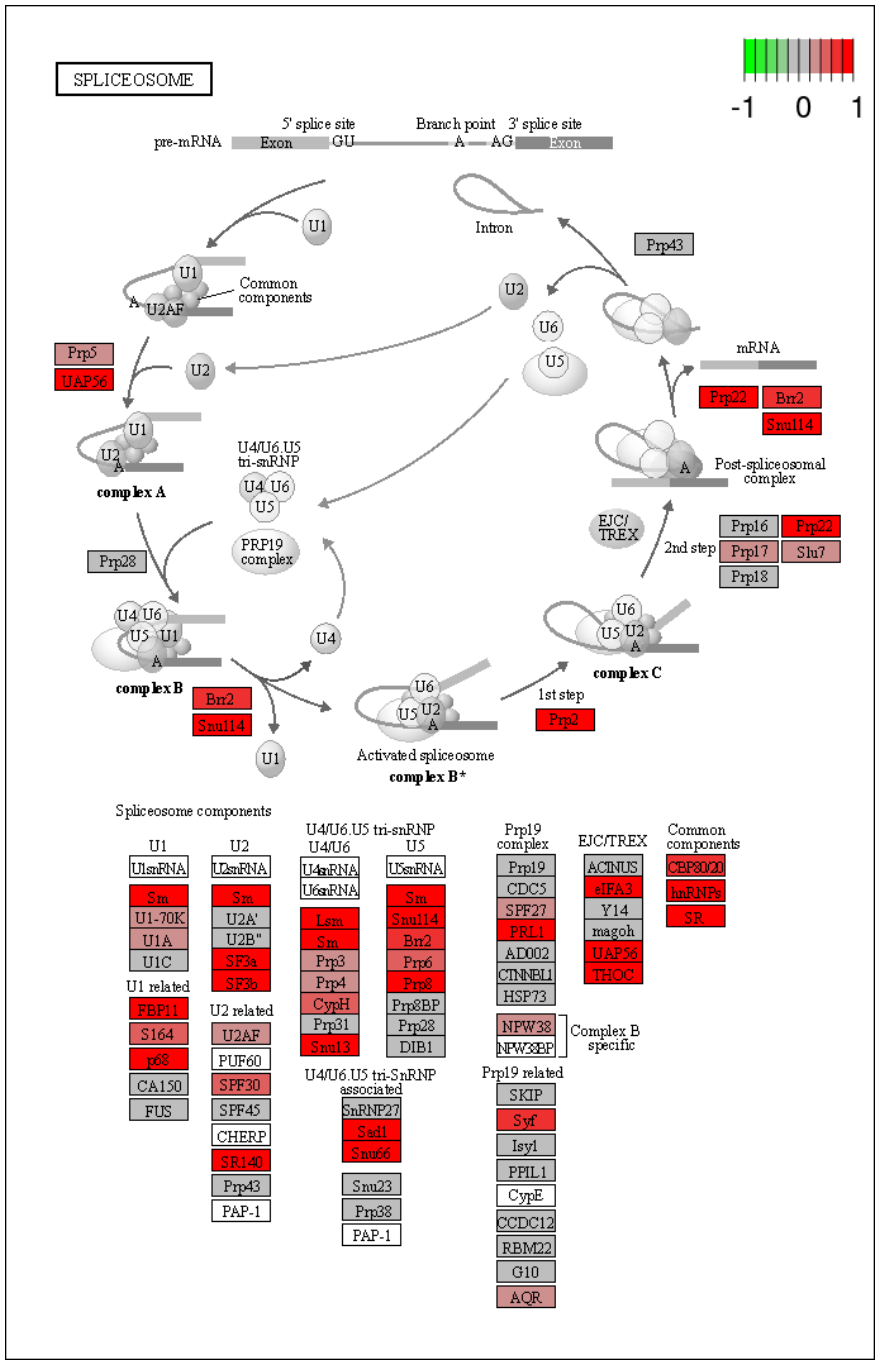



**Figura 5.6**: **DEG relacionados con el ciclo de *splicing* en Col-0**. *Análisis de enriquecimiento en pathways de KEGG para Col-0. Se consideró el FC obtenido a partir de la comparación entre plantas salvajes y mutantes para PRMT5 en la accesión Col-0. Los genes resaltados en rojo presentan un aumento de su expresión en esta última condición. Diagrama realizado a partir del paquete de R pathways.*

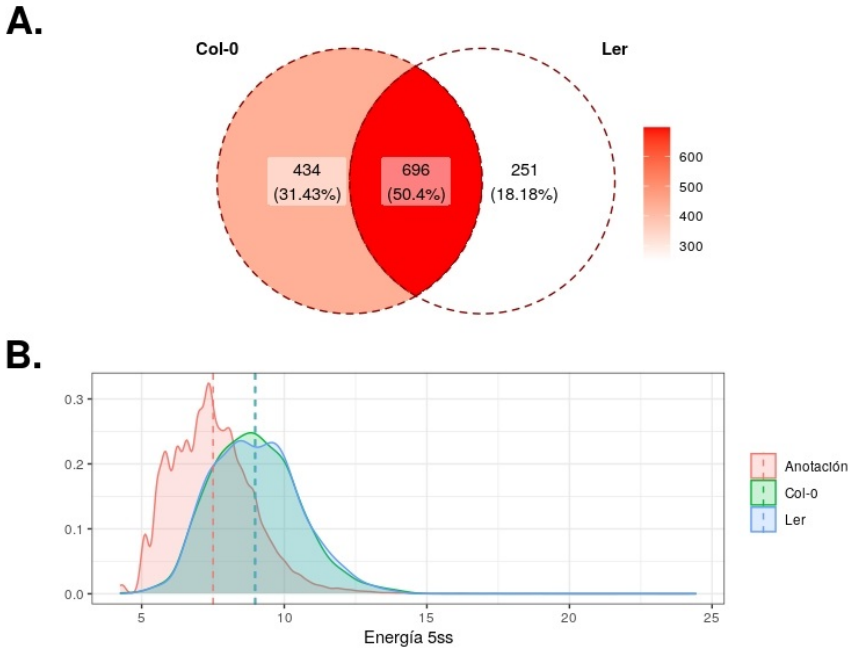



**Figura 5.7**: **Bines diferenciales entre plantas salvajes y mutantes**. **A.** *Diagrama de Venn entre los bines que sufrieron un cambio significativo entre las muestras salvajes y las mutantes prmt5-5 en Col-0 y Ler.* **B.** *Distribución de energías para los sitios donores de splicing (5'ss) de los eventos de retención de intrones en cada una de estas comparaciones. Anotación: 5'ss extraídos de la anotación del genoma y pertenecientes a genes expresados en al menos una condición; Col-0: 5'ss de eventos de retención de intrones en Col-0; Ler: ídem para el caso de la accesión Ler. Con líneas puntuadas se indica la media de las distribuciones.*

para los 5'ss anotados en el genoma de *A. thaliana* (TAIR10) y que pertenecen a genes que se encuentran expresados en al menos una de las condiciones evaluadas (criterio: mínimo de 10 lecturas que solapen las coordenadas genómicas del gen y una densidad de lecturas media mínima de 0.05). Al mismo tiempo, también podemos observar las distribuciones de energías para los 5'ss asociados a eventos de IR en las comparaciones entre muestras salvajes y mutantes, para la accesión Col-0 (verde) y Ler (azul).

Para ambas accesiones, las secuencias de los 5'ss de los bines diferenciales muestran un corrimiento de la distribución hacia energías más altas, indicando que éstos corresponden a sitios más alejados de la secuencia consenso. Por otro lado, si consideramos el $\Delta PIR$ obtenido para estos eventos de retención de intrones, encontramos que en 1028 de 1033 y 887 de 893 eventos hay una mayor retención en las muestras mutantes de *PRMT5* de Col-0 y Ler, respectivamente. Estos dos resultados llevan a concluir que en las plantas mutantes de ambas accesiones podrían existir dificultades en el reconocimiento de sitios donores de *splicing* débiles, lo que conlleva a la retención de los intrones asociados a esos sitios. Este tipo de *splicing* alternativo suele estar asociado a vías de regulación que afectan la expresión de los genes mediante mecanismos como NMD. Encontramos que alrededor del 25 % de los genes que encuentran modificados sus patrones de *splicing* en las muestras mutantes, también tienen diferencias significativas en los niveles de expresión.

## Diferencias en los patrones de *splicing* entre las accesiones Col-0 y Ler

Para analizar las diferencias entre Col-0 y Ler en cuanto a los patrones de *splicing* tanto en las plantas salvajes como mutantes para PRMT5, se realizó un Test Exacto de Fisher para determinar diferencias significativas en las métricas PIR/PSI; ajustando los p-valores mediante el método BH. Se tomó como significativos los eventos con un FDR $<0.05$ en las tres réplicas y un valor absoluto promedio de $\Delta PIR/PSI > 0.1$.

La cantidad de eventos diferenciales para el caso de las plantas salvajes es 686, mientras que para las plantas mutantes ese valor asciende a 1265. Como podemos ver en la Figura

5.8.A, esta diferencia se debe principalmente a un aumento en la cantidad de eventos de retención de intrones presente en las plantas mutantes. En suma, esto indica que existe una diferencia más pronunciada en los patrones de *splicing* entre las accesiones cuando PRMT5 no es funcional.

Tomando los eventos de retención de intrones,buscamos evaluar la posible influencia que tienen las diferencias de secuencia entre las accesiones en la determinación de los mismos. Para esto se calculó la densidad de SNP/Indels dentro de un rango genómico que comprende tanto el intrón retenido como los dos exones contiguos. En la Figura 5.8.B podemos comparar la densidad de variaciones de secuencia que muestran los eventos de IR entre plantas salvajes y mutantes, comparada con la densidad de SNP/Indels que muestran eventos de retención de intrón presentes en la anotación del genoma pero que no resultaron significativamente diferenciales en ninguna de las dos condiciones antes mencionadas. Tanto para el caso de las muestras salvajes como mutantes, los eventos de IR significativos tienen una mayor densidad de SNP/Indel en comparación a los eventos de IR no significativos, mostrando valores medios de 6.253 y 6.702 contra 3.5362 obtenido por éstos últimos.

Dada la importancia que tienen los sitios de *splicing*, a continuación nos concentramos en las variaciones de secuencia que se dan exclusivamente en ellos. Para el caso de los sitios donores (5'ss), se consideró SNPs que se ubiquen dentro de los últimas tres posiciones del exón anterior al intrón evaluado y las primeras 6 posiciones del mismo. Para el caso de los sitios aceptores (3'ss), se tomaron en cuenta las últimas 13 posiciones intrónicas y la primer posición del exón siguiente. Para ambos sitios se encontró que la frecuencia de SNPs es significativamente mayor para los sitios involucrados en una retención de intrón, ya sea en plantas salvajes o mutantes, respecto a la que muestran sitios asociados a eventos de IR no significativos (Figura 5.8.C). Resulta interesante destacar que para el caso de los sitios donores, se encontró una mayor frecuencia de SNPs en el caso de las muestras mutantes que en las salvajes. Esto es consistente con el hecho de que PRMT5 ayude al reconocimiento de los sitios 5'ss débiles, ya que implica que cambios en estas secuencias deberían tener un

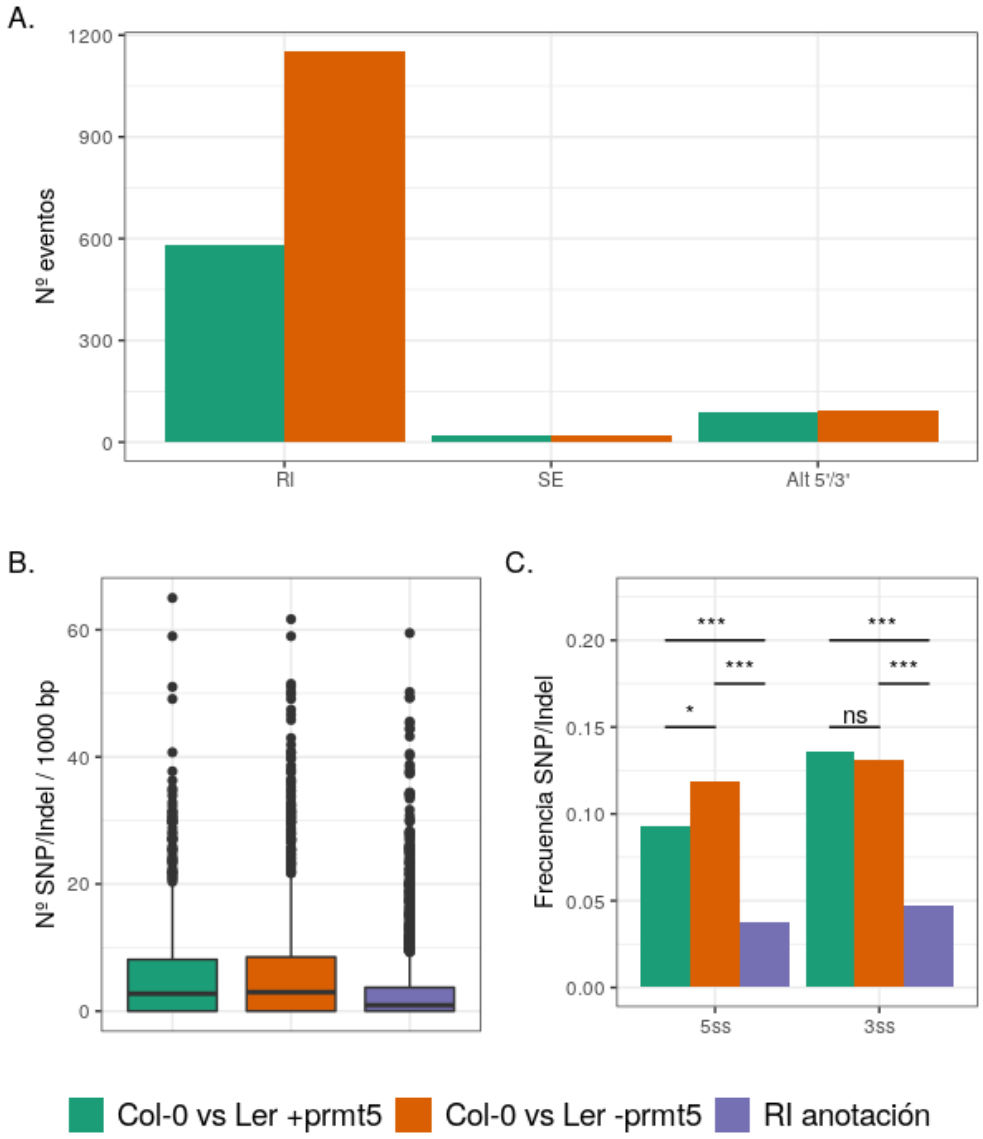



**Figura 5.8**: **Eventos de *splicing* diferenciales entre Col-0 y Ler.A.** *Cantidad de eventos de splicing diferenciales en la comparación entre las accesiones en plantas salvajes (wt, barras verdes) y mutantes (prmt5, barras naranjas) según el tipo de splicing alternativo (IR: retención de intrón, ES: salteo de exón, Alt 5'/3': extremo 5' o 3' alternativo).* **B.** *Densidad de variaciones en la secuencia entre Col-0 y Ler en torno a los eventos de retención de intrones en plantas salvajes (+prmt5), mutantes (-prmt5) y para eventos de IR anotados en el genoma de referencia pero que no dieron señal diferencial (RI anotación).* **C.** *Frecuencia de ocurrencia de SNP/Indels en los sitios donores (5ss) y aceptores (3ss) de splicing para eventos de retención de intrones en las plantas salvajes (+prmt5), mutantes (-prmt5) y en casos de retención de intrones anotados en el genoma de referencia (RI anotación). Test de Fisher para poner a prueba diferencias en las frecuencias observadas. Significancia: "***": p-valor <0.001; "**": p-valor <0.01; "*": p-valor <0.05, "ns": p-valor >0.05.*

mayor impacto funcional cuando PRMT5 está ausente.

Por otro lado, se evaluó el efecto de interacción entre ambas condiciones, es decir, se buscó eventos de *splicing* en los que el efecto de la mutación de *PRMT5* haya sido distinto en ambas accesiones. Para esto se realizó una estrategia similar a la utilizada en trabajos previos[55,109,168]. Se calculo el *ratio* entre las métricas PIR/PSI obtenidas para las muestras salvajes y mutantes en cada accesión, $PIR_{mutante}/PIR_{salvaje}$. Si tomamos como ejemplo un caso retención de intrón, un *ratio* mayor a 1 indica que en las plantas en las que *PRMT5* se encuentra mutado hay un mayor nivel de retención del intrón. Luego se calculo la diferencia entre los *ratios* obtenidos para cada accesión y, utilizando el error estándar de esta medida, se construyeron valores z y estimaron p-valores. Finalmente, se ajustaron los mismos por múltiple testeo, considerando como significativos los que obtuvieron un fdr $<0.05$.

En la Figura 5.9 podemos ver uno de estos eventos: el intrón 5 del gen At5g20220 no se encuentra afectado por la ausencia de PRTM5 en la accesión Col-0 y pero sí en la accesión Ler, en el que aumenta la retención del intrón. Si observamos la secuencia del 5'ss de este intrón encontramos que presenta un SNP en la posición 6. En Col-0 la secuencia es AAA|GTTAGT, mientras que en Ler es AAA|GTTAGC.

Se obtuvo un total de 90 eventos de *splicing* con un efecto interacción significativo. Si nos concentramos en las diferencias de secuencias asociadas a estos eventos, obtenemos que en 27 de los mismos hay al menos un SNP en la región correspondiente al sitio donor de *splicing*. Este valor representa un 30 % del total, un porcentaje que resulta ser más del doble que el obtenido anteriormente en las comparaciones entre plantas salvajes y mutantes de *PRMT5* (Figura 5.8.C). A partir de estos casos, nos preguntamos si el efecto que había tenido la mutación estaba relacionado con el cambio en la secuencia del sitio donor. Para esto calculamos la diferencia de energía que habia en la secuencia entre Col-0 y Ler, utilizando el modelo desarrollado en la primera parte de esta tesis. En la Figura 5.10 se muestra la relación entre la diferencia de los *ratios* $PIR_{mutante}/PIR_{salvaje}$ entre las accesiones y la diferencia de energías entre las secuencias de los 5'ss originada por la presencia de SNPs.

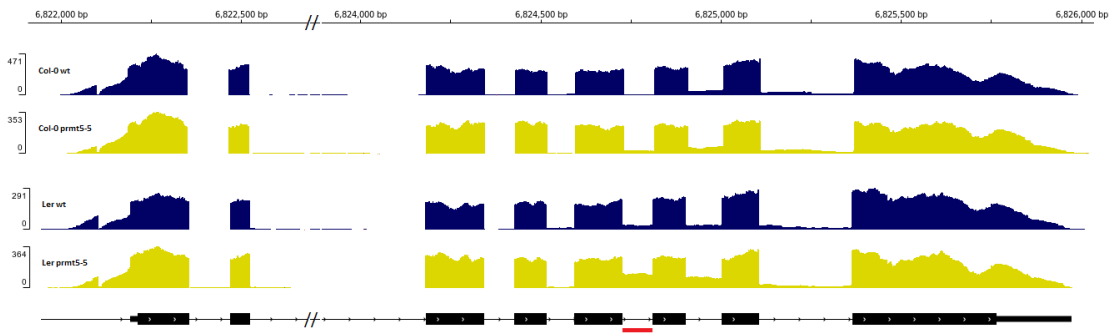



**Figura 5.9**: **Efecto interacción en At5g20220**. *Gráfico de cobertura en la región correspondiente al gen At5g20220 en las cuatro condiciones: Col-0 salvaje, Col-0 mutante de PRMT5, Ler salvaje y Ler mutante de PRMT5. Se indica con una línea roja el intrón que resultó significativo en la evaluación del efecto interacción entre estos factores.*

Los casos en los que la diferencia entre los *ratios* es positiva indica que la mutación de *PRMT5* produjo una mayor retención de intrón en Ler respecto a Col-0. Como podemos ver en la Figura 5.10, en la mayoría de los eventos esto esta asociado a que la secuencia de los 5'ss sean más débiles en Ler (lo que hace que la diferencia de energías también sea positiva). De una manera análoga, para los casos en los que el efecto de la mutación es mayor en Col-0 se da que en esa accesión se encuentra la variante de 5'ss más débil (diferencia de energías negativa). De esta forma encontramos que el efecto sobre los patrones de *splicing* de la mutación de *PRMT5* resulta ser mayor en la accesión con la variante de sitio donor más débil.

Finalmente, se investigó la influencia que diferentes factores regulatorios de *splicing*, RBP (del inglés, *RNA-binding proteins*), pueden tener sobre los cambios observados ante la mutación de *PRMT5*. Para esto se tomaron los eventos retención de intrones obtenidos en la comparación entre plantas salvajes y mutantes en Col-0 y se calculó el enriquecimiento de motivos de secuencias en una región que comprende tanto los intrones retenidos como los exones flanqueantes. Para esto se utilizó el *software* rMAPS2[71,123], el cual determina el enriquecimiento de motivos mediante el análisis de los patrones espaciales de los sitios de unión conocidos para más de 100 RBPs para los eventos de *splicing* evaluados, lo que llama

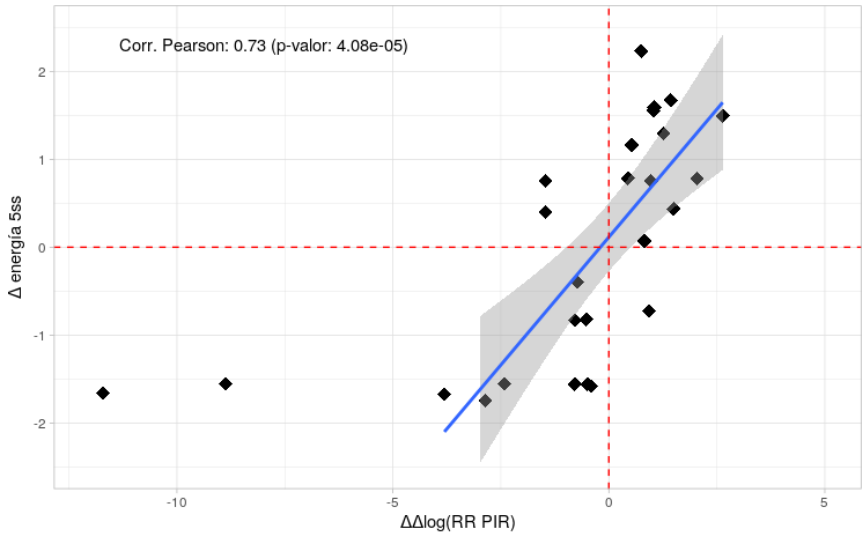



**Figura 5.10**: **5'ss con variación de secuencia en eventos de interacción entre genotipo y accesión**. *Gráfico de la relación que hay entre las diferencias entre las accesiones cuanto a las energías de los 5'ss producidas por la presencia de SNPs y el efecto de interacción genotipo x accesión. Para el cálculo de la correlación de Pearson fueron excluidos los dos eventos cuya diferencia en el efecto de interacción es menor a -5.*

"mapa de ARN´´ (*RNA-map*, en inglés). De esta forma, se encontró un total de 70 motivos asociados a RBPs conocidos en los eventos de retención de intrones analizados (Figura 5.11). Es importante destacar que la determinación de cuáles de estos motivos son efectivamente sitios de unión de RBPs en *A. thaliana* requiere de la realización de más experimentos. El presente análisis tiene un fin exploratorio.

Luego en los eventos IR significativos para el efecto de interacción, se analizó si existían SNP/Indels en las secuencias de los motivos hallados anteriormente. Sobre un total de 87 eventos, se encontraron SNPs sobre la secuencia de motivos en 39 de ellos(un 44.8 %). Estos motivos corresponden a 43 de los 70 encontramos en los eventos IR de la comparación entre plantas salvajes y mutantes para *PRMT5*. Como puede verse en el Cuadro 5.1, muchos de estos motivos tienen una gran coincidencia en su secuencia, siendo predominante los que presentan secuencias de poli-pirimidinas(Cuadro 5.1). En futuras investigaciones se deberá estudiar en mayor profundidad la posible relación entre estos RBPs y el mecanismo de acción de PRMT5 que estas investigaciones sugieren.

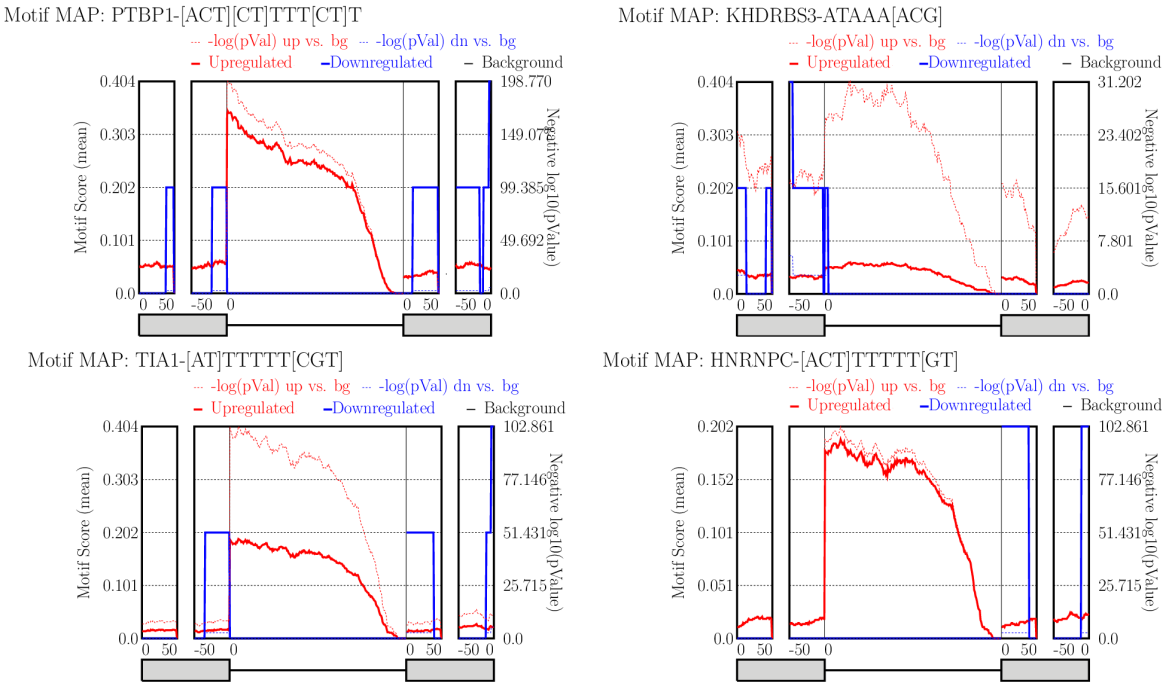



**Figura 5.11**: **Identificación de motivos relacionados con RBPs en los eventos IR asociados a la mutación de *PRMT5***. *Análisis realizado con el software rMAPS2. Se muestran los diagramas de enriquecimiento de motivos a lo largo de los eventos de IR relacionados con cuatro RBPs: PTB1, KHDRBS3, TIA1 y HNRNPC. En el eje de la izquierda se muestra el Motif Score, que da cuenta de cómo varía la densidad de motivos a lo largo del evento, y a la derecha se muestra el p-valor resultante de la comparación entre los eventos up o down regulados y los eventos tomados como control. Estos últimos corresponden a los eventos de retención de intrón anotados en el genoma pero cuya expresión no fue modificada por la mutación de PRMT5.*

| RBP | Motivo | N° eventos IR con SNPs |
|---|---|---|
| PTBP1 | [ACT][CT]TTT[CT]T | 10 |
| KHDRBS3 | ATAAA[ACG] | 6 |
| TIA1 | TTTTT[CGT][GT] | 6 |
| HNRNPC/L1 | [ACT]TTTTT[GT] | 6 |
| ZNF638 | [CGT]GTT[GC][GT]T | 5 |
| RBM3 | [AG]A[AGT]AC[GT]A | 5 |
| CPEB4 | TTTTTT | 5 |
| RALY | TTTTTT[CGT] | 5 |
| HuR | TTTTTT[GT] | 5 |
| RBM47 | GATGA[AT] | 4 |
| FXR2 | [AGT]GAC[AG][AG][AG] | 4 |
| KHDRBS2 | [AG]ATAAA[AC] | 4 |
| BRUNOL4 | [GT]GTGT[GT][GT] | 3 |
| RBMS3 | [ACT]ATATA | 3 |
| KHDRBS1 | TAAAA[ACG][ACG] | 3 |
| MATR3 | [AC]ATCTT[AG] | 3 |
| CPEB2 | C[ACT]TTTTT | 3 |
| A1CF | [AT]TAATT[AG] | 3 |
| SART1 | A[AG]AAAA[AC] | 2 |
| RBMS3 | [AC]TATA[GT][AC] | 2 |
| TUT1 | [AC][AG]ATACT | 2 |
| BRUNOL5 | TGTGT[GT][GT] | 2 |
| ZCRB1 | G[AG][ACT]TTAA | 2 |
| KHDRBS1 | ATAAAA[ACG] | 2 |
| TARDBP | GAATG[AGT] | 1 |
| PCBP1 | CC[AT][AT][ACT]CC | 1 |
| FXR1 | A[CT]GAC[AG] | 1 |
| ESRP1 | TGGTGG | 1 |
| SAMD4A | GC[GT]GG[ACT][AC] | 1 |
| RBM38 | [GT][GT]GTGT[GT] | 1 |
| 9G8 | [AT]GGAC[AG]A | 1 |
| FMR1 | [GT]GACA[AG]G | 1 |
| RBM45 | GACGA[AC][ACG] | 1 |
| U2AF2 | TTTTT[CT]C | 1 |
| RBFOX1 | [AT]GCATG[AC] | 1 |
| ZC3H10 | [GC][GC]AGCG[AC] | 1 |
| RBM8A | [AG][CT]GCGC[CGT] | 1 |
| MSI1 | TAGT[AT][AG]G | 1 |

**Cuadro 5.1**: **Motivos de RBPs con SNPs en eventos de IR de interacción**.

## Análisis de las variaciones de splicing en híbridos F1 y el efecto de la mutación de *PRMT5*

Las diferencias en los patrones de *splicing* que se observan entre las muestras pertenecientes a distintas accesiones pueden deberse a una combinación de diversos factores. Por un lado, tenemos el efecto que tienen diferencias de secuencia que se dan en las proximidades genómicas del evento de *splicing* evaluado, como pueden ser cambios de secuencia en sitios donores o aceptores, o secuencias que estimulan o silencian el *splicing*, como las ISE/ISS o ESE/ESS. A estos efectos los llamamos "efectos en cis". Por otro lado, los cambios también pueden darse por diferencias en cuanto a los factores regulatorios que afectan al *splicing*. Por ejemplo, debido a cambios en la expresión de algún factor clave o por cambios en las modificaciones post-traduccionales que sufre (como ya vimos, PRMT5 puede afectar ambos niveles al mismo tiempo). A esto lo llamamos "efectos en trans".

Con el objetivo de desacoplar estos dos tipos de efectos, se decidió realizar la comparación de los efectos de la mutación de *PRMT5* sobre los híbridos F1 entre las accesiones evaluados anteriormente, Col-0 y Ler. En el caso de los híbridos, el contexto en "trans" se ve mayormente igualado, ya que las variantes procedentes de ambas accesiones se encuentran ante los mismos factores regulatorios. En este sentido, si encontramos diferencias en el *splicing* entre las variantes alélicas de los híbridos, podemos inferir que las mismas se producen debido a los cambios en las secuencias en "cis", y no por diferencias en la abundancia o estado de los factores regulatorios.

Siguiendo el protocolo ya descripto en la sección de "Materiales y Métodos", se obtuvieron los eventos de *splicing* alélico diferenciales tanto para los híbridos F1 Col-0 X Ler como para Ler X Col-0, es decir, para los dos híbridos recíprocos. Luego se calculó el efecto interacción entre la mutación de PRMT5 y las variantes alélicas de la forma ya descripta más arriba. Para poder comparar con los resultados obtenidos para el caso de las plantas parentales, se volvió a realizar el análisis anterior pero a partir de muestras cuya cantidad de lecturas fue reducida hasta obtener una profundidad similar a la que presentan las muestras de los

híbridos (en inglés, *downsampling*). De esta manera, se comparan experimentos de RNA-Seq que tienen una profundidad de secuenciación comparable. Como podemos ver en la Figura 5.12, algunos eventos de interacción solo se dan en las plantas parentales y no en los híbridos (en la figura, At3g14660) indicando que al igualar el contexto en trans, las diferencias que se ven entre las accesiones desaparecen. Sin embargo, en los eventos que son compartidos (en la figura, At1g69620) las diferencias se mantienen tanto en los parentales como en los híbridos.

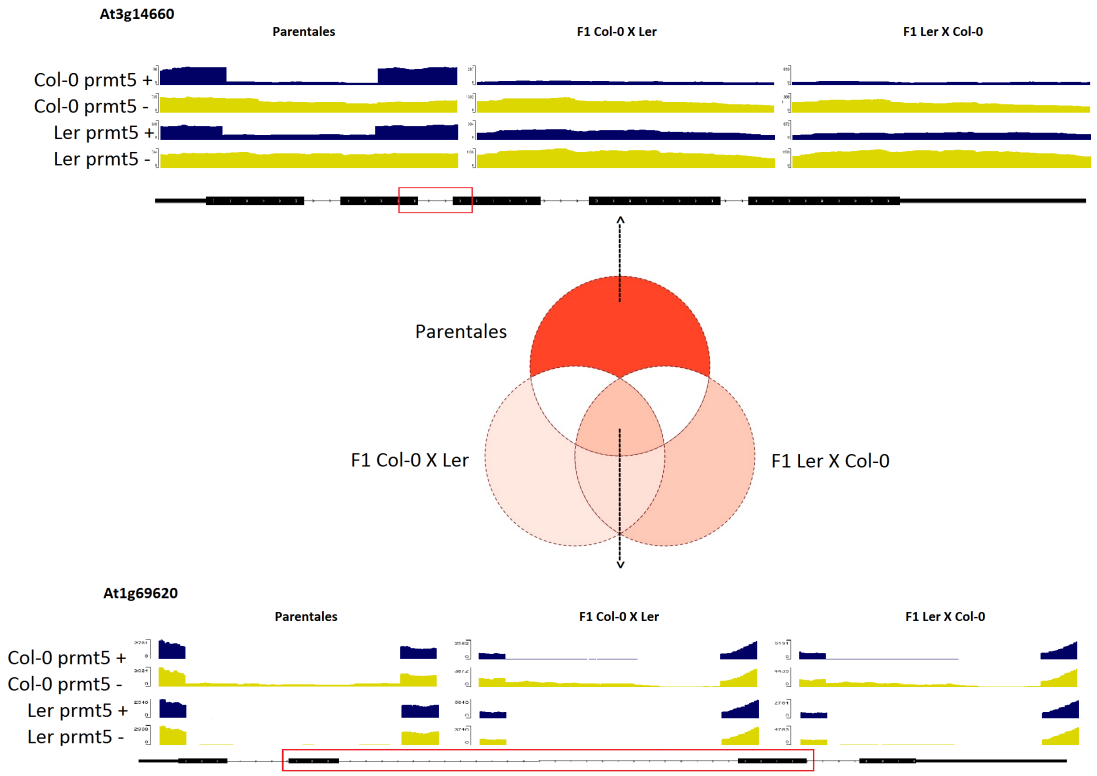



**Figura 5.12**: **Ejemplos de evento de interacción entre genotipo y accesión en los híbridos F1**. *Se muestra la cobertura de lecturas alrededor del evento analizado (marcado con un recuadro rojo en el modelo de la estructura del gen). En el intrón 2 del gen At3g14660 puede verse que las diferencias en las plantas parentales no se reproducen en los híbridos F1. Mientras que en el intrón 2 del gen At1g69620 se produce su retención solamente en la accesión Col-0 cuando PRMT5 se encuentra mutado, efecto que se da tanto en las plantas parentales como en los híbridos.*

En la Figura 5.13.A podemos ver un diagrama de Venn de los resultados obtenido para

el efecto interacción de los dos tipos de híbridos y los parentales. Tanto en el híbrido Col-0 X Ler como para Ler X Col-0 se obtiene un numero similar de eventos diferenciales: 35 y 36, respectivamente. Lo que resulta interesante de remarcar es que estos conjuntos no son idénticos. Los eventos compartidos por ambos híbridos son 22, mientras que el resto resultan significativos solo en uno de los dos. Si en la comparación incluimos a los eventos que resultaron significativo en la comparación entre las plantas parentales, encontramos que, de un total de 47 eventos, comparte con los híbridos aprox. un 42 % de los mismos. Como en el caso de los híbridos ambos alelos se encuentran en un mismo contexto trans, podemos decir que si un evento que aparece en los parentales también aparece en los híbridos es porque las diferencia que muestran en cuanto a su respuesta a la ausencia de PRMT5 se debe a las diferencias en cis que hay entre ambas accesiones. Por lo cual, podemos concluir que el efecto que produce PRMT5 sobre el *splicing* depende en gran medida del entorno local de los eventos de *splicing*.

Si analizamos la densidad de variaciones en las secuencias de los eventos de *splicing*(Figura 5.13.B), encontramos que tanto para ambos híbridos como para los eventos en la intersección entre los eventos parentales y éstos, se haya una mayor densidad de SNP/Indels que para los eventos que solamente resultan significativos para los parentales. Estos resultados apuntan a concluir que un porcentaje importante de los eventos que observamos como significativos en las plantas parentales se explican a partir de las diferencias en las secuencias en cis que se encuentran en las proximidades del evento de *splicing* diferencial y no por diferencias en la presencia/activación de factores en trans.

En cuanto a la presencia de variaciones de secuencia en motivos relacionados con RBPs en los eventos IR, realizamos un análisis similar al ya realizado para el efecto interacción de las muestras parentales (Cuadro 5.1). En los eventos analizados en las tres interacciones se ve una alta coincidencia en los motivos que se ven alterado por la presencia de SNPs/Indels (Cuadro 5.2). Esta coincidencia también se replica en el conjunto más restrictivo de los eventos que son compartidos tanto por los eventos significativos en los parentales y en al

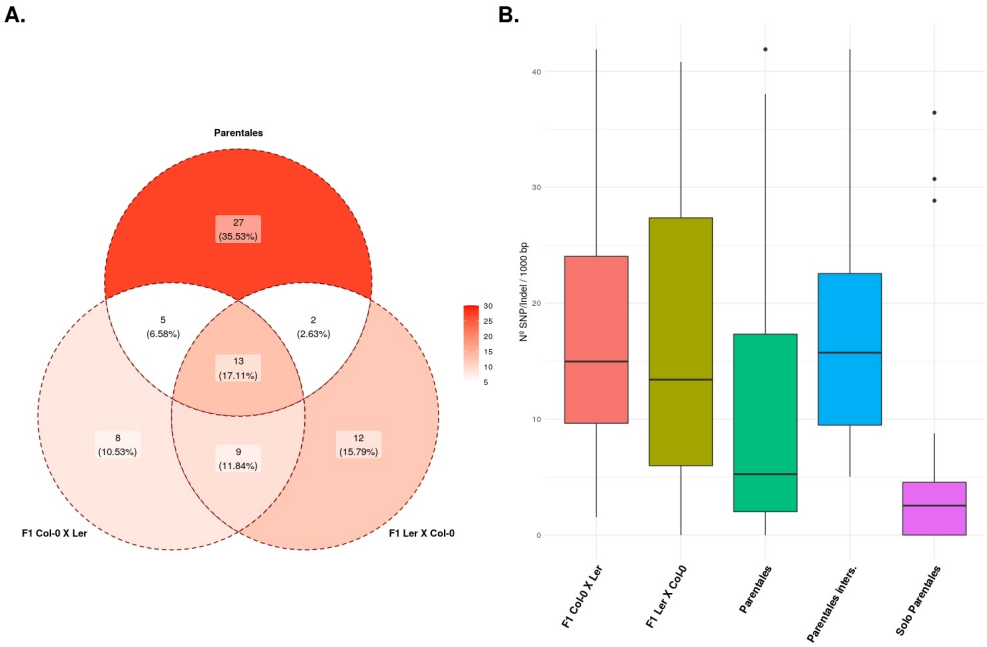
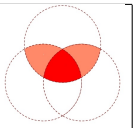



**Figura 5.13**: **Efecto interacción entre genotipo y accesión en los híbridos**. ***A.*** *Diagrama de Venn de los eventos de interacción significativos en los híbridos F1 y en las muestras sub-sampleadas de las plantas parentales.* ***B.*** *Densidad de variaciones de secuencia en torno a los eventos de interacción correspondientes a: los híbridos F1 Col-0 X Ler y Ler X Col-0, las plantas parentales, la intersección entre los conjuntos anteriores y los eventos que sólo dan significativos en las muestras parentales.*

menos uno de los híbridos.

| **RBP** | ***Parentales*** | **F1 Col-0 X Ler** | **F1 Ler X Col-0** |  |
|---|---|---|---|---|
| PTBP1 | 7 | 10 | 8 | 4 |
| ZNF638 | 5 | 6 | 4 | 4 |
| KHDRBS3 | 5 | 6 | 5 | 5 |
| RBM3 | 4 | 4 | 1 | 3 |
| TIA1 | 4 | 8 | 6 | 3 |
| SART3 | 4 | 4 | 3 | 3 |
| PABPC1 | 4 | 4 | 3 | 3 |
| HNRNPC/L1 | 4 | 7 | 6 | 3 |
| CPEB4 | 4 | 6 | 5 | 3 |
| RALY | 4 | 6 | 5 | 3 |
| BRUNOL4 | 4 | 3 | 3 | 2 |
| HuR | 4 | 5 | 5 | 3 |
| FXR2 | 3 | 4 | 2 | 3 |
| CPEB2 | 3 | 3 | 5 | 2 |
| RBM47 | 3 | 5 | 6 | 3 |
| MATR3 | 3 | 2 | 3 | 2 |
| KHDRBS2 | 3 | 4 | 3 | 3 |
| FXR1 | 3 | 3 | 3 | 3 |
| BRUNOL5 | 3 | 3 | 3 | 2 |

**Cuadro 5.2**: **Motivos de RBPs con SNPs en eventos de IR de interacción de híbridos F1**.*Se consideran los eventos de IR resultantes del efecto de interacción en las muestras parentales (con downsampling) y en los dos híbridos F1 recíprocos, Col-0 X Ler y Ler X Col-0. Las últimas tres columnas muestran los motivos con SNPs en los eventos que tienen en común las muestras parentales con al menos uno de los híbridos, y los eventos que aparecieron solamente en uno de los híbridos pero no en el recíproco.*

## 5.4. Conclusión

PRMT5 es un regulador clave en una enorme cantidad de procesos distintos que, a su vez, se encuentran interrelacionados, como son la transcripción, el procesamiento y transporte de los ARNm, el remodelamiento de la cromatina, etc. Esta diversidad de procesos hace de PRMT5 una pieza clave a la hora de entender cómo estos mecanismos de regulación dialogan entre sí; pero también, por otro lado, hace que entender de manera pormenorizada el mecanismo mediante el cual PRMT5 afecta los patrones generales de expresión y *splicing* sea un desafío. Si nos concentramos en los efectos que esta metil-transferasa tiene sobre el *splicing*, se ha visto que los eventos que se encuentran afectados por ella están asociados a sitios donores débiles, lo cual es explicado por la metilación que PRMT5 produce sobre las proteínas Sm[17]. Esta modificación post-traduccional produciría una estabilización de la unión de los sitios 5'ss que tienen una baja complementariedad con el snRNA de U1 con la maquinaria de *splicing*, favoreciendo así su reconocimiento. Sin embargo, la manera en la que el efecto de PRMT5 sobre el *splicing* depende de las señales presentes en cis no ha sido completamente estudiada. En este trabajo, mediante la comparación de los efectos de la mutación de esta metil-transferasa en dos accesiones de *Arabidopsis thaliana* pudimos determinar la influencia que tienen las variaciones en las secuencias genómicas relacionadas con el *splicing* sobre la acción de PRMT5.

Cuando analizamos los efectos globales que presenta la mutante en cada uno de los accesiones por separado, éstos no parecen diferir mucho entre sí. Tanto al analizar los genes diferencialmente expresados como los eventos de *splicing* que se vieron afectados, la coincidencia entre las accesiones es de aprox. 50 %. En cuanto al significado biológicos de los genes afectados en su expresión, se encuentran en ambas accesiones que éstos participan principalmente de vías metabólicas relacionadas con los ácidos nucleicos y con la transducción de señales. Estos resultados coinciden con investigaciones anteriores que encuentran que las plantas mutantes de PRMT5 presentan sobre-expresados genes relacionados con el procesamiento del ARN, especialmente los involucrados con el *splicing*, mientras que genes relacionados con la

respuesta a estrés y a luz se encuentran sub-expresados[68]. Resulta interesante observar el efecto que tiene la mutación de PRMT5 sobre la expresión de genes relacionados con el ciclo de *splicing*. Se sabe que esta metil-transferasa modifica post-traduccionalmente varios factores de *splicing*, especialmente las proteínas Sm y LSm4[37]. Esta acción de PRMT5 resulta importante para el funcionamiento y la fidelidad del proceso de *splicing*. Por otro lado, se observa que la mutación de *PRMT5* genera un aumento de la expresión de estos genes y de muchos otros componentes que intervienen en el procesamiento de los intrones. Esto implica una doble regulación por parte de PRMT5 del *splicing*: la metilación de proteínas como las Sm favorece la eficiencia del *splicing*, al mismo tiempo que a nivel transcripcional, la expresión de muchos de estos genes se encuentra controlada por esta misma metil-transferasa. Podemos llegar a entender esta doble acción como un posible efecto de compensación, en donde la perdida de eficiencia producida por la ausencia de la metilación de estos factores sea parcialmente compensada por el aumento de expresión de los distintos componentes que participan del proceso de *splicing*. De todas maneras, no podemos asegurar si los cambios en la expresión observados entre las plantas salvajes y mutantes de PRMT5 son producto de cambios en el estado de metilación de histonas o de otro tipo de proteínas blanco que regulen la transcripción.

En cuanto a los eventos de *splicing* diferenciales entre la condición salvaje y mutante pudimos confirmar, mediante el uso del modelo para los 5'ss desarrollado en la primera parte de esta tesis, lo observado por trabajos anteriores: los 5'ss asociados a eventos de retención de intrones producidos por la mutación de PRMT5 resultan ser sitios débiles. Sin embargo, esta explicación no parece ser suficiente para explicar los efectos que PRMT5 tiene sobre el *splicing*. Si bien el conjunto de los 5'ss afectados son estadísticamente más débiles, no todos ellos lo son y, aún más importante, no todos los 5'ss débiles son afectados. Para poder entender mejor la relación entre el efecto de PRMT5 y la fortaleza de los sitios de *splicing* comparamos lo que ocurre con estos mutantes en las accesiones.

Si nos concentramos en las diferencias que muestran Col-0 y Ler en cuanto a sus patrones

de *splicing* tanto en la condición salvaje como entre las mutantes, observamos que en ambos casos estos eventos están asociados a una mayor densidad de variaciones de secuencia que eventos que no resultaron diferenciales. Esto indica que estas diferencias de secuencia juegan un rol importante a la hora de explicar estas alteraciones en el *splicing*. Si bien tanto en la condición salvaje y mutante esta densidad de SNP/Indels es similar, observamos que el número de eventos de *splicing* diferenciales en esta última es mayor (principalmente debido a un aumento de eventos de retención de intrones). Cuando PRMT5 está ausente se producen un mayor número de diferencias en los patrones de *splicing* entre las accesiones, lo que nos lleva a pensar que las diferencias en las secuencias se traducen en diferencias funcionales en mayor medida de lo que ocurre en las plantas salvajes. Entre estos eventos diferenciales, encontramos que hay una mayor frecuencia de SNPs en los 5'ss en el caso de las plantas prmt5-5 respecto a lo observado en las plantas salvajes, lo que no ocurre para los sitios 3'ss. Todos estos resultados nos llevan a concluir que las plantas mutantes podrían ser más sensibles a variaciones en las secuencias genómicas. Estos resultados nos llevan a proponer que PRMT5 podría ser un factor importante a la hora de mantener la fidelidad del proceso de *splicing* frente a determinadas variaciones, como pueden ser cambios en las secuencias genómicas. En este sentido, PRMT5 actuaría a modo de "buffer"manteniendo la estabilidad del sistema frente a los cambios. Esto resulta particularmente interesante si tenemos en cuenta que, en animales, ya se ha caracterizado a PRMT5 como un modulador de la integridad genómica al afectar el *splicing* de varios genes relacionados con el reparado del ADN[83,156]. De esta manera, al favorecer la estabilidad de los patrones de *splicing* frente a cambios en las secuencias no sólo contribuye con la regulación de este proceso en particular, sino también, con la robustez de múltiples capas regulatorias.

El efecto interacción entre estos dos factores (genotipo y accesión), nos permite identificar aquellos eventos de *splicing* que cambiaron de manera distinta ante la mutación de PRMT5 en Col-0 respecto a Ler. Esto nos permite evaluar la dependencia de la respuesta a las diferencias entre las accesiones. Encontramos que un alto porcentaje de estos eventos

presentan algún SNP en la región que corresponde al 5'ss. Utilizando nuevamente nuestro modelo energético para evaluar la diferencia de fortaleza del 5'ss entre ambas accesiones, pudimos evaluar su relación con la diferencia en el efecto de la mutación y encontramos una alta correlación entre estas variables (superior a 0.7), lo que nos lleva a concluir que la secuencia de los 5'ss resulta de gran importancia en la determinación del efecto de PRMT5. Sin embargo, solo una pequeña proporción de los sitios 5'ss que presentan una variación de secuencia entre Col-0 y Ler están asociados con eventos de *splicing* con un efecto interacción significativo (recordemos que en total hay más de 3500 sitios con variaciones de secuencia). Esto nos llevó a extender nuestro análisis para investigar la posible influencia de otros factores de *splicing*, por lo que realizamos un análisis de motivos de secuencia relacionados con proteínas de unión a ARN (RBPs) conocidas, encontrando en los eventos de interacción una serie de posibles SRE que presentan variaciones de secuencia entre Col-0 y Ler. Estos motivos están relacionados con importantes reguladores del *splicing*, entre los cuales se destacan los pertenecientes a la familia de las hnRNPs, aunque la determinación precisa de cuáles son los factores de *splicing* que regulan estos eventos requerirá de futuras investigaciones.

De esta manera, podemos concluir que el efecto de PRMT5 no sólo se encuentra relacionado con la fortaleza de los sitios 5'ss, sino también, con otros SRE que pueden estar ejerciendo una influencia importante en la determinación de los eventos afectados. Sin embargo entre las accesiones existen otros tipos de diferencias que no se encuentren directamente relacionadas con el entorno local en el que se produce cada evento de *splicing* sino, por ejemplo, con la cantidad o estado de distintos reguladores. Por este motivo resulta de importancia el estudio del efecto de la mutación en PRMT5 en las plantas híbridas F1. En estas plantas las variantes alélicas de ambas accesiones están inmersas en un mismo contexto celular. Por lo que en este caso sí vamos a poder decir que las diferencias que encontramos las podemos relacionar directamente a diferencias locales presentes en el entorno del evento de *splicing* diferencial, como puede ser un cambios en las secuencias genómicas o variaciones en el estado de la cromatina.

En este sentido, encontramos que una alta proporción, 42 %, de los eventos que pueden encontrarse en las plantas parentales también se observa en el caso de los híbridos. Para poder comprender completamente este resultado es importante destacar algunas limitaciones en cuanto a la metodología utilizada. A partir de la secuenciación masiva del ARNm de los híbridos, se asignó cada lectura a uno de las dos accesiones según en cual genoma se obtuvo una calidad de mapeo más alta. Las lecturas que obtuvieron la misma calidad de mapeo para los genomas de ambas accesiones fueron descartadas del análisis. Esto provoca que las regiones del genoma en las que haya una baja densidad de SNP/Indels entre Col-0 y Ler presenten una gran disminución de su cobertura, lo que impide que puedan ser estadísticamente evaluadas. Al analizar los eventos que tuvieron un efecto interacción significativo en el caso de las plantas parentales pero que no obtuvieron los mismos resultados en los casos de los híbridos, efectivamente encontramos que son regiones del genoma que presentan una menor densidad de variaciones de secuencia. Estas consideraciones nos llevan a pensar que el porcentaje de coincidencia encontrado entre los eventos de los parentales y los eventos de los híbridos puede ser una subestimación de la coincidencia que pudiera encontrarse si la densidad de variaciones fuera homogénea a lo largo de los distintos genes. Por lo que la coincidencia real entre los eventos encontrados en los parentales y los híbridos sea posiblemente aún mayor al 42 %, reforzando más la idea de que el efecto de PRMT5 está mediado por las secuencias en cis que se encuentran en las cercanías de los eventos de *splicig*.

Otro aspecto interesante a destacar es la existencia de un conjunto de eventos de *splicing* que resultan significativos solamente en los híbridos pero no en las muestras parentales. Estos eventos podrían deberse a posibles factores en trans que enmascaran en los parentales los efectos en cis locales que presentan estos eventos. Por otro lado, otro resultado interesante es que la coincidencia entre los eventos que resultan significativos en los híbridos recíprocos no es completa. Una posible explicación es que cada alelo esté inmerso en un contexto en el que prime un conjunto distinto de marcas epigenéticas que influyan en los patrones de *splicing* de esos eventos y que estén relacionadas con el parental del cual provino el alelo.

A partir de todos estos resultados podemos intentar dar un marco explicativo que eche luz sobre los posibles mecanismos por los cuales PRMT5 ejerce su acción sobre los patrones de *splicing*. Algunas investigaciones recientes[76,105] han propuesto que las PRMTs intervienen específicamente en la fracción del *splicing* que se produce de forma post-transcripcional. Si bien en proporción es minoritario, de alrededor de un 20 % en cultivo de células humanas[59] y de 28 % en *A. thaliana*[76], este tipo *splicing* tendría un rol fisiológico importante. Por lo general, esta asociado a transcriptos que mantienen intrones sin remover que se encuentran retenidos en el núcleo hasta que alguna señal dispare una serie de procesos que permiten la finalización de la remoción de los intrones restantes y la exportación del transcripto hacia el citoplasma. Este mecanismo confiere una importante ventaja en situaciones de estrés ya que aumentaría la velocidad de respuesta. Por otro lado, la detención en el núcleo de los transcriptos con intrones retenidos los protege de la degradación mediada por el mecanismo de NMD. En el *splicing* post-transcripcional, suele estar asociado a unas organelas sin membranas llamadas "motas nucleares"(*nuclear speckles*, en inglés)[59,97]. Estas últimas son conocidas por actuar como lugares de reserva de factores de *splicing* y componentes del spliceosoma pero en donde también puede llevarse a cabo el *splicing*. En la regulación de la conformación de estas organelas participa una gran cantidad de modificaciones post-traduccionales, entre ellas, la metilación de los residuos arginina[69,127,135].

Teniendo en mente estos trabajos, podemos pensar la manera en la que los resultados obtenidos a lo largo de esta tesis pueden vincularse con ellos.Una de las características que suelen presentar los eventos de *splicing* que se procesan de manera post-transcripcional es que están asociados a sitios débiles , característica que también se da en los eventos afectados por PRMT5. Si el efecto sobre el *splicing* de esta metil-transferasa se da mayormente en la fracción de eventos que se procesan post-trascripcionalmente, entonces cualquier evento que disminuya las posibilidades de que un transcripto se pueda procesar de manera co-transcripcional también lo haría más dependiente de PRMT5 para su correcta maduración. De esta manera podrían explicarse las diferencias que se observan en cuanto a la dependencia

de un mismo evento de *splicing* a PRMT5 en una accesión y otra cuando hay una variación de secuencia en sus 5'ss. Pero por otro lado, también tenemos las variaciones que se dan en los motivos de secuencia relacionados con RBPs. Distintos factores de *splicing* pueden favorecer o impedir el procesamiento co-transcripcional, por lo que es esperable encontrar que también modificaciones en los SRE asociados a los mismos provoque diferencias en la sensibilidad de un dado evento de *splicing* a la ausencia de PRMT5. Una característica en común que tiene la mayoría de los motivos que encontramos que tienen alguna variación entre las accesiones en los eventos de interacción es que están caracterizados por secuencias donde predominan las pirimidinas, en particular poli-U. Muchos son los factores de *splicing* que pueden unirse a este tipo de secuencias, como por ejemplo, UBP1 que es un conocido regulador de la transcripción y de la maduración de los ARNm que colabora con el procesamiento de intrones sub-óptimos en plantas[90,91]. UBP1 es el ortólogo en plantas de TIA1[150], el cual tiene un rol importante en la regulación del *splicing*, estabilidad y traducción de ARNm en condiciones de estrés y durante el proceso de neurodesarrollo[21]. En animales se ha visto que tanto UBP1[150] como TIA1[40] están involucrados en la formación de gránulos de estrés en el citoplasma. Más estudios se necesitarán en el futuro para poder determinar de manera más precisa cuáles son los factores de *splicing* vinculados a la sensibilidad a PRMT5, especialmente para el caso de las plantas donde esto se encuentra aún menos estudiado. De la misma forma que más experimentos se requerirán para determinar la relación de PRMT5 con los *nuclear speckles* y la regulación del *splicing* post-transcripcional. Entendemos que el presente trabajo abre un campo de preguntas que posibilitará el delineamiento de futuras investigaciones que extenderán nuestro conocimientos sobre el rol clave que PRMT5 tiene en el fidelidad y coordinación de las distintas capas regulatorias tanto a nivel transcripcional como post-transcripcional.

# Capítulo 6

# Conclusiones generales

A lo largo de esta tesis nos hemos encargado de investigar distintos aspectos que del proceso de *splicing* y su regulación. En particular, nos interesó estudiar la relación entre sus determinantes en cis, como son las secuencias de los sitios de *splicing*, y los factores que intervienen en trans, como PRMT5.

A continuación se da un breve listado de las principales conclusiones a las que se llegó en los diferentes capítulos de esta tesis y los caminos que se abren para futuras investigaciones:

- Se logró elaborar un modelo estadístico basado en el principio de máxima entropía de las secuencias de los 5'ss para entender la variabilidad que estas secuencias presentan a lo largo de un genoma e identificar los patrones de co-variaciones entre las distintas posiciones de los 5'ss.

- En los modelos realizados para los distintos genomas analizados se encontró una serie de patrones en común: 1. dentro de las partes intrónica y exónica de los 5'ss, las secuencias consenso se ven reforzadas entre sí mediante interacciones positivas; 2. lo opuesto ocurre si consideramos las interacciones entre posiciones intrónicas y exónicas, entre las que se dan interacciones negativas; 3. las posiciones de los 5'ss que presentan un bajo contenido de información tienen, sin embargo, interacciones significativas con otras posiciones del sitio.

- Estos resultados se encuentran en coincidencia con la idea que plantea que los 5'ss

mantendrían una alto apareamiento de secuencia con snRNA de U1 solamente en una de sus dos porciones: o bien en la región intrónica o bien en la exónica, pero no en ambas a la vez. Esto se relaciona a la mayor posibilidad de regulación que permiten los sitios de *splicing* que presentan un nivel de complementariedad intermedio.

- En este contexto, podemos pensar que relaciones de compensación entre las distintas posiciones del sitio podrían ser importantes a la hora de mantener la fidelidad del proceso de *splicing* incluso cuando la variabilidad de cada posición individual sea alta.

- En total se analizaron 30 genomas de diversas especies eucariotas, tanto animales, vegetales como hongos. Esto permitió extender algunos de los resultados que trabajos previos habían observado en análisis más acotados, considerando solamente mamíferos. Algunos patrones de interacción entre las posiciones de los 5'ss resultaron ser comunes a todas las especies analizadas. Entre éstas podemos mencionar las interacciones positivas entre las posiciones +4:+5 y +5:+6, y las interacción negativa -2:+5.

- Sin embargo, otras interacciones resultaron ser específicas para ciertos grupos de organismos. La interacción -1:+5 solo fue encontradas en las especies vegetales, mientras que la interacción -1:+6 es propia de animales. La importancia de la posición 6, a pesar de su alta variabilidad, ha sido reportada en determinados contextos, como lo es en las alteraciones del *splicing* producidas en la disautonomía familiar.

- Nuestros resultados exponen diferencias significativas en los patrones de interacción entre las posiciones de los 5'ss. Las implicancias que estas diferencias pueden tener en cuanto a la historia evolutiva y el funcionamiento del *splicing* requiere de futuras investigaciones.

- Los experimentos de RNASeq realizados sobre plantas de *Arabidopsis thaliana* provenientes de dos accesiones distintas, nos permitió estudiar el vínculo que hay entre el efecto de PRTM5 y la fortaleza de los sitios de *splicing*. Concluyendo que la sensibili-

dad de un evento de *splicing* a la acción de PRMT5 está altamente influenciada por la secuencia de los sitios 5'ss.

- Por otro lado, en los eventos de *splicing* que se vieron afectados por la mutación de PRMT5 de manera distinta en las accesiones (eventos de interacción) también se observó variaciones de secuencia en distintos motivos que podrían ser sitios de unión a RBPs, lo que amplía el escenario antes centrado solamente en los 5'ss. Otros factores de *splicing* también podrían determinar la dependencia de un evento de *splicing* a PRMT5.

- El estudio de híbridos F1 entre las dos accesiones analizadas, nos permitió concluir que una importante proporción de los eventos alterados en las plantas parentales se debe efectivamente a cambios que se dan en el entorno local de estos eventos, y no a diferencias en la cantidad o estados de otros factores en trans.

- Los resultados expuestos hasta aquí nos llevan a plantear la hipótesis de que PRMT5 podría tener un rol importante en la estabilidad de los patrones de *splicing* ante la presencia de cambios, como pueden ser variaciones en las secuencias genómicas. Esto se condice con el aumento de diferencias entre las accesiones que se registra cuando PRMT5 está ausente.

- Estudios recientes relacionan el efecto de PRMT5 con el *splicing* post-traduccional, destacando su importancia para el procesamiento de los intrones que no pudieron removerse co-transcripcionalmente. En este contexto, podríamos entender los resultados obtenidos en esta tesis de la siguiente manera: todos los factores que modifiquen las probabilidades de que un intrón pueda ser procesado de manera co-transcripcional, harán que el mismo sea más o menos dependiente de la acción de PRMT5. Tanto la fortaleza de los 5'ss como la unión de factores de *splicing* podrían tener este efecto. Futuras investigaciones tendrán que determinar los detalles del mecanismo mediante el cual PRMT5 actúa y cuáles son los factores que se encuentran asociados.

# Bibliografía


[1] Federico Abascal, Iakes Ezkurdia, Juan Rodriguez-Rivas, Jose Manuel Rodriguez, Angela del Pozo, Jesús Vázquez, Alfonso Valencia, and Michael L. Tress. Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. *PLoS Computational Biology*, 11(6):1004325, jun 2015.

[2] Emmanuel L. Abebrese, Syed H. Ali, Zachary R. Arnold, Victoria M. Andrews, Katharine Armstrong, Lindsay Burns, Hannah R. Crowder, R. Thomas Day, Daniel G. Hsu, Katherine Jarrell, Grace Lee, Yi Luo, Daphine Mugayo, Zain Raza, and Kyle Friend. Identification of human short introns. *PLoS ONE*, 12(5):e0175393, may 2017.

[3] Keith L. Adams and Jonathan F. Wendel. Polyploidy and genome evolution in plants, 2005.

[4] Ayaz Ahmad and Xiaofeng Cao. Plant PRMTs Broaden the Scope of Arginine Methylation. *Journal of Genetics and Genomics*, 39(5):195–208, 2012.

[5] Donna E. Akiyoshi, Hilary G. Morrison, Shi Lei, Xiaochuan Feng, Quanshun Zhang, Nicolas Corradi, Harriet Mayanja, James K. Tumwine, Patrick J. Keeling, Louis M. Weiss, and Saul Tzipori. Genomic survey of the non-cultivatable opportunistic human pathogen, Enterocytozoon bieneusi. *PLoS Pathogens*, 5(1):1–10, 2009.

[6] Douglas G Altman and J Martin Bland. Interaction revisited : the difference between two estimates. *BMJ*, 326:219, 2003.

[7] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, 2012.

[8] Jan O. Andersson, Åsa M. Sjögren, David S. Horner, Colleen A. Murphy, Patricia L. Dyal, Staffan G. Svärd, John M. Logsdon, Mark A. Ragan, Robert P. Hirt, and Andrew J. Roger. A genomic survey of the fish parasite Spironucleus salmonicida indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics*, 8, 2007.

[9] François Bachand. Protein arginine methyltransferases: From unicellular eukaryotes to humans. *Eukaryotic Cell*, 6(6):889–898, 2007.

[10] Nuno L. Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Y. Xiong, Serge Gueroussov, Leo J. Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Çolak, Tae Hyung Kim, Christine M. Misquitta-Ali, Michael D. Wilson, Philip M. Kim, Duncan T. Odom, Brendan J. Frey, and Benjamin J. Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, dec 2012.

[11] Jean D. Beggs, Johan Van Den Berg, Albert Van Ooyen, and Charles Weissmann. Abnormal expression of chromosomal rabbit $\beta$-globin gene in Saccharomyces cerevisiae. *Nature*, 283(5750):835–840, 1980.

[12] S. M. Berget, C. Moore, and P. A. Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(8):3171–3175, 1977.

[13] Arnold J. Berk. Discovery of RNA splicing and genes in pieces. *Proceedings of the National Academy of Sciences of the United States of America*, 113(4):801–805, 2016.

[14] A. L. Beyer and Y. N. Osheim. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes development*, 2(6):754–765, 1988.

[15] Nicole I. Bieberstein, Fernando Carrillo Oesterreich, Korinna Straube, and Karla M. Neugebauer. First exon length controls active chromatin signatures and transcription. *Cell Reports*, 2(1):62–68, 2012.

[16] Carsten Boesler, Norbert Rigo, Maria M. Anokhina, Marcel J. Tauchert, Dmitry E. Agafonov, Berthold Kastner, Henning Urlaub, Ralf Ficner, Cindy L. Will, and Reinhard Lührmann. A spliceosome intermediate with loosely associated tri-snRNP accumulates in the absence of Prp28 ATPase activity. *Nature Communications*, 7(May), 2016.

[17] Hero Brahms, Lydie Meheus, Veronique De Brabandere, Utz Fischer, and Reinhard Lührmann. Symmetrical dimethylation of arginine residues in spliceosomal Sm protein B/B and the Sm-like protein LSm4, and their interaction with the SMN protein. *Rna*, 7(11):1531–1542, 2001.

[18] Ulrich Braunschweig, Nuno L. Barbosa-Morais, Qun Pan, Emil N. Nachman, Babak Alipanahi, Thomas Gonatopoulos-Pournatzis, Brendan Frey, Manuel Irimia, and Benjamin J. Blencowe. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, 24(11):1774–1786, nov 2014.

[19] Ulrich Braunschweig, Serge Gueroussov, Alex M. Plocik, Brenton R. Graveley, and Benjamin J. Blencowe. Dynamic integration of splicing within gene regulatory pathways. *Cell*, 152(6):1252–1269, 2013.

[20] J W S Brown, G Feix, and D Frendewey'. Accurate in vitro splicing of two pre-mRNA plant introns in a HeLa cell nuclear extract. Technical Report 11, 1986.

[21] Loryn P. Byres, Marat Mufteev, Kyoko E. Yuki, Wei Wei, Alina Piekna, Michael D. Wilson, Deivid C. Rodrigues, and James Ellis. Identification of TIA1 mRNA targets during human neuronal development. *Molecular Biology Reports*, 48(9):6349–6361, 2021.

[22] Ido Carmel, Saar Tal, Ida Vig, and Gil Ast. Comparative analysis detects dependencies among the 5 splice-site positions. *Rna*, 10(5):828–840, 2004.

[23] Raquel F. Carvalho, Dóra Szakonyi, Craig G. Simpson, Inês C.R. Barbosa, John W.S. Brown, Elena Baena-González, and Paula Duque. The arabidopsis SR45 splicing factor, a negative regulator of sugar signaling, modulates SNF1-related protein kinase 1 stability. *Plant Cell*, 28(8):1910–1925, aug 2016.

[24] Saurabh Chaudhary, Waqas Khokhar, Ibtissam Jabre, Anireddy S.N. Reddy, Lee J. Byrne, Cornelia M. Wilson, and Naeem H. Syed. Alternative splicing and protein diversity: Plants versus animals. *Frontiers in Plant Science*, 10(June):1–14, 2019.

[25] Tao Chen, Peng Cui, Hao Chen, Shahjahan Ali, Shoudong Zhang, and Liming Xiong. A KH-Domain RNA-Binding Protein Interacts with FIERY2/CTD Phosphatase-Like 1 and Splicing Factors and Is Important for Pre-mRNA Splicing in Arabidopsis. *PLoS Genetics*, 9(10):1003875, oct 2013.

[26] Chunhong Cheng, Zhijuan Wang, Bingjian Yuan, and Xia Li. RBM25 mediates abiotic responses in plants. *Frontiers in Plant Science*, 8:292, mar 2017.

[27] Donghang Cheng, Jocelyn Côté, Salam Shaaban, and Mark T. Bedford. The Arginine Methyltransferase CARM1 Regulates the Coupling of Transcription and mRNA Processing. *Molecular Cell*, 25(1):71–83, 2007.

[28] T Choi, M Huang, C Gorman, and R Jaenisch. A generic intron increases gene expression in transgenic mice. *Molecular and Cellular Biology*, 11(6):3070–3074, 1991.

[29] Louise T. Chow, Richard E. Gelinas, Thomas R. Broker, and Richard J. Roberts. An amazing sequence arrangement at the 5 ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, 1977.

[30] Lesley Collins and David Penny. Complex spliceosomal organization ancestral to extant eukaryotes. *Molecular Biology and Evolution*, 22(4):1053–1066, apr 2005.

[31] Jocelyn Cote, Francois-Michel Boisvert, Marie-Chloé Boulanger, Mark T. Bedford, and Stéphane Richard. Sam68 RNA Binding Protein Is an In Vivo Substrate for Protein Arginine N-Methyltransferase 1. *Molecular Biology of the Cell*, 14:274–287, 2003.

[32] Miklos Csuros, Igor B. Rogozin, and Eugene V. Koonin. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Computational Biology*, 7(9):1–9, 2011.

[33] Liying Cui, P. Kerr Wall, James H. Leebens-Mack, Bruce G. Lindsay, Douglas E. Soltis, Jeff J. Doyle, Pamela S. Soltis, John E. Carlson, Kathiravetpilla Arumuganathan, Abdelali Barakat, Victor A. Albert, Hong Ma, and Claude W. DePamphilis. Widespread genome duplications throughout the history of flowering plants. *Genome Research*, 16(6):738–749, jun 2006.

[34] Peng Cui, Shoudong Zhang, Feng Ding, Shahjahan Ali, and Liming Xiong. Dynamic regulation of genome-wide pre-mRNA splicing and stress tolerance by the Sm-like protein LSm5 in Arabidopsis. *Genome Biology*, 15(1):1–18, jan 2014.

[35] Christian Kroun Damgaard, Søren Kahns, Søren Lykke-Andersen, Anders Lade Nielsen, Torben Heick Jensen, and Jørgen Kjems. A 5 Splice Site Enhances the Recruitment of Basal Transcription Initiation Factors In Vivo. *Molecular Cell*, 29(2):271–278, 2008.

[36] Rita Das, Jiong Yu, Zuo Zhang, Melanie P. Gygi, Adrian R. Krainer, Steven P. Gygi, and Robin Reed. SR Proteins Function in Coupling RNAP II Transcription to Pre-mRNA Splicing. *Molecular Cell*, 26(6):867–881, 2007.

[37] Xian Deng, Lianfeng Gu, Chunyan Liu, Tiancong Lu, Falong Lu, Zhike Lu, Peng Cui, Yanxi Pei, Baichen Wang, Songnian Hu, and Xiaofeng Cao. Arginine methylation

mediated by the Arabidopsis homolog of PRMT5 is essential for proper pre-mRNA splicing. *PNAS*, 104(44):19114–19119, 2010.

[38] Stepan Denisov, Georgii Bazykin, Alexander Favorov, Andrey Mironov, and Mikhail Gelfand. Correlated Evolution of Nucleotide Positions within Splice Sites in Mammals. *PLoS ONE*, 10(12):e0144388, 2015.

[39] Michael Dewaele, Tommaso Tabaglio, Karen Willekens, Marco Bezzi, Shun Xie Teo, Diana H.P. Low, Cheryl M. Koh, Florian Rambow, Mark Fiers, Aljosja Rogiers, Enrico Radaelli, Muthafar Al-Haddawi, Soo Yong Tan, Els Hermans, Frederic Amant, Hualong Yan, Manikandan Lakshmanan, Ratnacaram Chandrahas Koumar, Soon Thye Lim, Frederick A. Derheimer, Robert M. Campbell, Zahid Bonday, Vinay Tergaonkar, Mark Shackleton, Christine Blattner, Jean Christophe Marine, and Ernesto Guccione. Antisense oligonucleotide-mediated MDM4 exon 6 skipping impairs tumor growth. *Journal of Clinical Investigation*, 126(1):68–84, 2016.

[40] Xiufang Ding, Siyu Gu, Song Xue, and Shi Zhong Luo. Disease-associated mutations affect TIA1 phase separation and aggregation in a proline-dependent manner. *Brain Research*, 1768(April):147589, 2021.

[41] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[42] Gabriele Drechsel, André Kahles, Anil K. Kesarwani, Eva Stauffer, Jonas Behr, Philipp Drewe, Gunnar Rätsch, and Andreas Wachter. Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *Plant Cell*, 25(10):3726–3742, oct 2013.

[43] Steffen Erkelenz, William F. Mueller, Melanie S. Evans, Anke Busch, Katrin Schöneweis, Klemens J. Hertel, and Heiner Schaal. Position-dependent splicing activation and

repression by SR and hnRNP proteins rely on common mechanisms. *Rna*, 19(1):96–102, 2013.

[44] Rocío Espada, R. Gonzalo Parra, Thierry Mora, Aleksandra M. Walczak, and Diego U. Ferreiro. Inferring repeat-protein energetics from evolutionary information. *PLoS Computational Biology*, 13(6):1–16, 2017.

[45] Rocío Espada, R. Gonzalo Parra, Manfred J. Sippl, Thierry Mora, Aleksandra M. Walczak, and Diego U. Ferreiro. Repeat proteins challenge the concept of structural domains. *Biochemical Society Transactions*, 43:844–849, 2015.

[46] Jinlin Feng, Jingjing Li, Zhaoxu Gao, Yaru Lu, Junya Yu, Qian Zheng, Shuning Yan, Wenjiao Zhang, Hang He, Ligeng Ma, and Zhengge Zhu. SKIP Confers Osmotic Tolerance during Salt Stress by Controlling Alternative Gene Splicing in Arabidopsis. *Molecular Plant*, 8(7):1038–1052, jul 2015.

[47] Sebastian M. Fica, Chris Oubridge, Wojciech P. Galej, Max E. Wilkinson, Xiao Chen Bai, Andrew J. Newman, and Kiyoshi Nagai. Structure of a spliceosome remodelled for exon ligation. *Nature*, 542(7641):377–380, 2017.

[48] Sergei A. Filichkin, Henry D. Priest, Scott A. Givan, Rongkun Shen, Douglas W. Bryant, Samuel E. Fox, Weng Keen Wong, and Todd C. Mockler. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Research*, 20(1):45–58, jan 2010.

[49] E. B. Fowlkes and C. L. Mallows. Comparing Two A Method for Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 2014.

[50] Katharina Frey and Boas Pucker. Animal, Fungi, and Plant Genome Sequences Harbor Di ff erent Non-Canonical Splice Sites. *cells*, 9(458), 2020.

[51] Xiang Dong Fu and Manuel Ares. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, 15(10):689–701, 2014.

[52] Tal Galili. dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, 31(22):3718–3720, 2015.

[53] Elad Ganmor, Ronen Segev, and Elad Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23):9679–9684, 2011.

[54] Kaiping Gao, Akio Masuda, Tohru Matsuura, and Kinji Ohno. Human branch point consensus sequence is yUnAy. *Nucleic Acids Research*, 36(7):2257–2267, 2008.

[55] Qingsong Gao, Wei Sun, Marlies Ballegeer, Claude Libert, and Wei Chen. Predominant contribution of cis- regulatory divergence in the evolution of mouse alternative splicing . *Molecular Systems Biology*, 11(7):816, 2015.

[56] Daniel Gautheret, Olivier Poirot, Fabrice Lopez, Stéphane Audic, and Jean Michel Claverie. Alternate polyadenylation in human mRNAs: A large-scale analysis by EST clustering. *Genome Research*, 8(5):524–530, 1998.

[57] Sahar Gelfman, Noa Cohen, Ahuvi Yearim, and Gil Ast. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome Research*, 23(5):789–799, 2013.

[58] Sarah V. Gerhart, Wendy A. Kellner, Christine Thompson, Melissa B. Pappalardi, Xi Ping Zhang, Rocio Montes De Oca, Elayne Penebre, Kenneth Duncan, Ann Boriack-Sjodin, Baochau Le, Christina Majer, Michael T. McCabe, Chris Carpenter, Neil Johnson, Ryan G. Kruger, and Olena Barbash. Activation of the p53-MDM4 regulatory axis defines the anti-tumour response to PRMT5 inhibition through its role in regulating cellular splicing. *Scientific Reports*, 8(1):1–15, 2018.

[59] Cyrille Girard, Cindy L. Will, Jianhe Peng, Evgeny M. Makarov, Berthold Kastner, Ira Lemm, Henning Urlaub, Klaus Hartmuth, and Reinhard Luhrmann. Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nature Communications*, 3, 2012.

[60] Janett Göhring, Jaroslaw Jacak, and Andrea Barta. Imaging of endogenous messenger RNA splice variants in living cells reveals nuclear retention of transcripts inaccessible to nonsense-mediated decay in Arabidopsis. *Plant Cell*, 26(2):754–764, mar 2014.

[61] Gregory J. Goodall and Witold Filipowicz. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell*, 58(3):473–483, aug 1989.

[62] Janina Görnemann, Kimberly M. Kotovic, Katja Hujer, and Karla M. Neugebauer. Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Molecular Cell*, 19(1):53–63, 2005.

[63] Ernesto Guccione and Stéphane Richard. The regulation, functions and clinical relevance of arginine methylation. *Nature Reviews Molecular Cell Biology*, 20(10):642–657, 2019.

[64] Pierre-Jacques Hamard, Gabriel E. Santiago, Fan Liu, Daniel L. Karl, Concepcion Martinez, Na Man, Adnan Mookhtiar, Stephanie Duffort, Sarah Greenblatt, Ramiro E. Verdun, and Stephen Nimer. PRMT5 regulates DNA repair by controlling the alternative splicing of key histone-modifying enzymes. *Cell Rep.*, 24(10):2643–2657, 2018.

[65] J. Han, J.-H. Ding, C. W. Byeon, J. H. Kim, K. J. Hertel, S. Jeong, and X.-D. Fu. SR Proteins Induce Alternative Exon Skipping through Their Activities on the Flanking Constitutive Exons. *Molecular and Cellular Biology*, 31(4):793–802, 2011.

[66] Lisa Hartmann, Theresa Wießner, and Andreas Wachter. Subcellular compartmentation of alternatively spliced transcripts defines SERINE/ARGININE-RICH PROTEIN30 expression. *Plant Physiology*, 176(4):2886–2903, apr 2018.

[67] Klaus Hartmuth and Andrea Barta. In vitro processing of a plant pre-mRNA in a hela cell nuclear extract. *Nucleic Acids Research*, 14(19):7513–7528, oct 1986.

[68] Carlos E. Hernando, Sabrina E. Sanchez, Estefanía Mancini, and Marcelo J. Yanovsky. Genome wide comparative analysis of the effects of PRMT5 and PRMT4/CARM1 arginine methyltransferases on the Arabidopsis thaliana transcriptome. *BMC Genomics*, 16(1):1–15, 2015.

[69] Mario Hofweber, Saskia Hutten, Benjamin Bourgeois, Emil Spreitzer, Annika Niedner-Boblenz, Martina Schifferer, Marc David Ruepp, Mikael Simons, Dierk Niessing, Tobias Madl, and Dorothee Dormann. Phase Separation of FUS Is Suppressed by Its Nuclear Import Receptor and Arginine Methylation. *Cell*, 173(3):706–719.e13, 2018.

[70] Yin Hu, Yan Huang, Ying Du, Christian F. Orellana, Darshan Singh, Amy R. Johnson, Anaïs Monroy, Pei Fen Kuan, Scott M. Hammond, Liza Makowski, Scott H. Randell, Derek Y. Chiang, D. Neil Hayes, Corbin Jones, Yufeng Liu, Jan F. Prins, and Jinze Liu. DiffSplice: The genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Research*, 41(2):1–18, 2013.

[71] Jae Y. Hwang, Sungbo Jung, Tae L. Kook, Eric C. Rouchka, Jinwoong Bok, and Juw W. Park. rMAPS2: An update of the RNA map analysis and plotting server for alternative splicing regulation. *Nucleic Acids Research*, 48(W1):W300–W306, 2021.

[72] Manuel Irimia, David Penny, and Scott W. Roy. Coevolution of genomic intron number and splice sites. *Trends in Genetics*, 23(7):321–325, jul 2007.

[73] Manuel Irimia, Jakob Lewin Rukov, David Penny, and Scott William Roy. Functional

and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evolutionary Biology*, 7(188), 2007.

[74] Hiroaki Iwata and Osamu Gotoh. Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics*, 12(1):45, 2011.

[75] Aishwarya G. Jacob and Christopher W.J. Smith. Intron retention as a component of regulated gene expression programs, sep 2017.

[76] Jinbu Jia, Yanping Long, Hong Zhang, Zhuowen Li, Zhijian Liu, Yan Zhao, Dongdong Lu, Xianhao Jin, Xian Deng, Rui Xia, Xiaofeng Cao, and Jixian Zhai. Post-transcriptional splicing of nascent RNA contributes to widespread intron retention in plants. *Nature Plants*, 6(7):780–788, jul 2020.

[77] Bong-Seok Jo and Sun Shim Choi. The Functional Benefits of Introns in Genomes. *Genomics Inform*, 13(4):112–118, 2015.

[78] Hyunchul Jung, Donghoon Lee, Jongkeun Lee, Donghyun Park, Yeon Jeong Kim, Woong Yang Park, Dongwan Hong, Peter J. Park, and Eunjung Lee. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nature Genetics*, 47(11):1242–1248, nov 2015.

[79] Maria Kalyna, Craig G. Simpson, Naeem H. Syed, Dominika Lewandowska, Yamile Marquez, Branislav Kusenda, Jacqueline Marshall, John Fuller, Linda Cardle, Jim McNicol, Huy Q. Dinh, Andrea Barta, and John W.S. Brown. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Research*, 40(6):2454–2469, mar 2012.

[80] Norbert F. Käufer and Judith Potashkin. Analysis of the splicing machinery in fission yeast: A comparison with budding yeast and mammals. *Nucleic Acids Research*, 28(16):3003–3010, aug 2000.

[81] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: Diversification, exon definition and function, may 2010.

[82] Yevgenia L. Khodor, Jerome S. Menet, Michael Tolan, and Michael Rosbash. Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. *Rna*, 18(12):2174–2186, 2012.

[83] Shinseog Kim, Ufuk Günesdogan, Jan J. Zylicz, Jamie A. Hackett, Delphine Cougot, Siqin Bao, Caroline Lee, Sabine Dietmann, George E. Allen, Roopsha Sengupta, and M. Azim Surani. PRMT5 Protects Genomic Integrity during Global DNA Demethylation in Primordial Germ Cells and Preimplantation Embryos. *Molecular Cell*, 56(4):564–579, 2014.

[84] Cheryl M. Koh, Marco Bezzi, Diana H.P. Low, Wei Xia Ang, Shun Xie Teo, Florence P.H. Gay, Muthafar Al-Haddawi, Soo Yong Tan, Motomi Osato, Arianna Sabò, Bruno Amati, Keng Boon Wee, and Ernesto Guccione. MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. *Nature*, 523(7558):96–100, 2015.

[85] Csaba Koncz, Femke DeJong, Nicolas Villacorta, Dóra Szakonyi, and Zsuzsa Koncz. The spliceosome-activating complex: Molecular mechanisms underlying the function of a pleiotropic regulator, jan 2012.

[86] Eugene V. Koonin, Miklos Csuros, and Igor B. Rogozin. Whence genes in pieces: Reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley Interdisciplinary Reviews: RNA*, 4(1):93–105, 2013.

[87] Eli Koren, Galit Lev-Maor, and Gil Ast. The emergence of alternative 3 and 5 splice site exons from constitutive exons. *PLoS Computational Biology*, 3(5):0895–0908, 2007.

[88] A.R. Kornblihtt. Coupling transcription and alternative. *Adv. Exp. Med. Biol.*, 623:175–189, 2007.

[89] Tom Laloum, Guiomar Martín, and Paula Duque. Alternative Splicing Control of Abiotic Stress Responses, feb 2018.

[90] Mark H. L. Lambermon, Yu Fu, Dominika A. Wieczorek Kirk, Marcel Dupasquier, Witold Filipowicz, and Zdravko J. Lorković. UBA1 and UBA2, Two Proteins That Interact with UBP1, a Multifunctional Effector of Pre-mRNA Maturation in Plants. *Molecular and Cellular Biology*, 22(12):4346–4357, 2002.

[91] Mark H.L. Lambermon, Gordon G. Simpson, Dominika A. Wieczorek Kirk, Maja Hemmings-Mieszczak, Ulrich Klahre, and Witold Filipowicz. UBP1, a novel hnRNP-like protein that functions at multiple steps of higher plant nuclear pre-mRNA maturation. *EMBO Journal*, 19(7):1638–1649, 2000.

[92] Christopher E. Lane, Krystal Van Den Heuvel, Catherine Kozera, Bruce A. Curtis, Byron J. Parsons, Sharen Bowman, and John M. Archibald. Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50):19908–19913, 2007.

[93] Christopher J. Langford and Dieter Gallwitz. Evidence for an intron-contained sequence required for the splicing of yeast RNA polymerase II transcripts. *Cell*, 33(2):519–527, 1983.

[94] Byeong Ha Lee, Avnish Kapoor, Jianhua Zhu, and Jian Kang Zhu. Stabilized1, a stress-upregulated nuclear protein, is required for pre-mRNA splicing, mRNA turnover, and stress tolerance in Arabidopsis. *Plant Cell*, 18(7):1736–1749, jul 2006.

[95] Timothy R Lezon, Jayanth R Banavar, Marek Cieplak, Amos Maritan, and Nina V Fedoroff. Using the principle of entropy maximization to infer genetic interaction

networks from gene expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 103(50):19033–19038, 2006.

[96] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[97] Susan E. Liao and Oded Regev. Splicing at the phase-separated nuclear speckle interface: A model. *Nucleic Acids Research*, 49(2):636–645, 2021.

[98] Lan Lin, Shihao Shen, Peng Jiang, Seiko Sato, Beverly L. Davidson, and Yi Xing. Evolution of alternative splicing in primate brain transcriptomes. *Human Molecular Genetics*, 19(15):2958–2973, 2010.

[99] Shengrong Lin, Gabriela Coutinho-Mansfield, Dong Wang, Shatakshi Pandit, and Xiang Dong Fu. The splicing factor SC35 has an active role in transcriptional elongation. *Nature Structural and Molecular Biology*, 15(8):819–826, 2008.

[100] Yansheng Liu, Mar Gonzàlez-Porta, Sergio Santos, Alvis Brazma, John C. Marioni, Ruedi Aebersold, Ashok R. Venkitaraman, and Vihandha O. Wickramasinghe. Impact of Alternative Splicing on the Human Proteome. *Cell Reports*, 20(5):1229–1241, aug 2017.

[101] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(26), 2011.

[102] W. P. Maddison and M. Slatkin. Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution*, 45(5):1184–1197, 1991.

[103] Estefania Mancini, Andres Rabinovich, Javier Iserte, Marcelo Yanovsky, and Ariel

Chernomoretz. ASpli: Integrative analysis of splicing landscapes through RNA-Seq assays. *Bioinformatics*, 37(17):2609–2616, 2021.

[104] Kranthi K. Mandadi and Karen Beth G. Scholthof. Genome-wide analysis of alternative splicing landscapes modulated during plant-virus interactions in brachypodium distachyon. *Plant Cell*, 27(1):71–85, feb 2015.

[105] Maxim I Maron, Alyssa D Casill, Varun Gupta, Jacob S Roth, Simone Sidoli, Charles C Query, Matthew J Gamble, and David Shechter. Type I and II PRMTs inversely regulate post-transcriptional intron detention through Sm and CHTOP methylation. *eLife*, 11:72867, 2022.

[106] Yamile Marquez, John W.S. Brown, Craig Simpson, Andrea Barta, and Maria Kalyna. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research*, 22(6):1184–1195, jun 2012.

[107] Yamile Marquez, Markus Höpfler, Zahra Ayatollahi, Andrea Barta, and Maria Kalyna. Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Research*, 25(7):995–1007, 2015.

[108] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

[109] C. Joel McManus, Joseph D. Coolon, Jodi Eipper-Mains, Patricia J. Wittkopp, and Brenton R. Graveley. Evolution of splicing regulatory networks in Drosophila. *Genome Research*, 24(5):786–796, 2014.

[110] Arfa Mehmood, Asta Laiho, Mikko S. Venälaïnen, Aidan J. McGlinchey, Ning Wang, and Laura L. Elo. Systematic evaluation of differential splicing tools for RNA-seq studies. *Briefings in Bioinformatics*, 21(6):2052–2065, 2020.

[111] Gunter Meister, Christian Eggert, Dirk Bühler, Hero Brahms, Christian Kambach, and Utz Fischer. Methylation of Sm proteins by a complex containing PRMT5 and the putative U snRNP assembly factor pICln. *Current Biology*, 11(24):1990–1994, 2001.

[112] Gunter Meister and Utz Fischer. Assisted RNP assembly: SMN and PRMT5 complexes cooperate in the formation of spliceosomal UsnRNPs. *The EMBO Journal*, 21(21):5853–5863, 2002.

[113] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B. Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, dec 2012.

[114] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49), 2011.

[115] Devlin C. Moyer, Graham E. Larue, Courtney E. Hershberger, Scott W. Roy, and Richard A. Padgett. Comprehensive database and evolutionary dynamics of U12-type introns. *Nucleic Acids Research*, 48(13):7066–7078, 2020.

[116] Jigeesha Mukhopadhyay and Georg Hausner. Organellar introns in fungi, algae, and plants. *Cells*, 10(8), 2021.

[117] Ralph C. Nichols, Xiao Wei Wang, Jie Tang, B. Jonell Hamilton, Frances A. High, Harvey R. Herschman, and William F.C. Rigby. The RGG domain in hnRNP A2 affects subcellular localization. *Experimental Cell Research*, 256(2):522–532, 2000.

[118] Timothy W Nilsen and Brenton R Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463, 2010.

[119] Dawn E. Ouelle, Frédérique Zindy, Richard A. Ashmun, and Charles J. Sherr. Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell*, 83(6):993–1000, 1995.

[120] Richard D. Palmiter, Eric P. Sandgren, Mary R. Avarbock, D. Diane Allen, and Ralph L. Brinster. Heterologous introns can enhance expression of transgenes in mice. *Proceedings of the National Academy of Sciences of the United States of America*, 88(2):478–482, 1991.

[121] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, dec 2008.

[122] Ioannis A. Papaioannou, Chrysoula D. Dimopoulou, and Milton A. Typas. Çryptic"group-I introns in the nuclear SSU-rRNA gene of Verticillium dahliae. *Current Genetics*, 60(3):135–148, 2014.

[123] Juw Won Park, Sungbo Jung, Eric C. Rouchka, Yu Ting Tseng, and Yi Xing. rMAPS: RNA map analysis and plotting server for alternative exon regulation. *Nucleic Acids Research*, 44(1):W333–W338, 2016.

[124] Clemens Plaschka, Pei Chun Lin, Clément Charenton, and Kiyoshi Nagai. Prespliceosome structure provides insights into spliceosome assembly and regulation. *Nature*, 559(7714):419–422, 2018.

[125] I. V. Poverennaya and M. A. Roytberg. Spliceosomal Introns: Features, Functions, and Evolution. *Biochemistry (Moscow)*, 85(7):725–734, 2020.

[126] Boas Pucker and Samuel F Brockington. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genomics*, 19(980), 2018.

[127] Seema Qamar, Guo Zhen Wang, Suzanne J. Randle, Francesco Simone Ruggeri, Juan A. Varela, Julie Qiaojin Lin, Emma C. Phillips, Akinori Miyashita, Declan Williams, Florian Ströhl, William Meadows, Rodylyn Ferry, Victoria J. Dardov, Gian G. Tartaglia, Lindsay A. Farrer, Gabriele S. Kaminski Schierle, Clemens F. Kaminski, Christine E. Holt, Paul E. Fraser, Gerold Schmitt-Ulms, David Klenerman, Tuomas Knowles, Michele Vendruscolo, and Peter St George-Hyslop. FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation-$\pi$ Interactions. *Cell*, 173(3):720–734.e15, 2018.

[128] Anireddy S.N. Reddy. Alternative splicing of pre-messenger RNAs in plants in the genomic era, 2007.

[129] Anireddy S.N. Reddy, Mark F. Rogers, Dale N. Richardson, Michael Hamilton, and Asa Ben-Hur. Deciphering the plant splicing code: Experimental and computational approaches for predicting alternative splicing and splicing regulatory elements, feb 2012.

[130] F. Remacle, Nataly Kravchenko-Balasha, Alexander Levitzki, and R. D. Levine. Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(22):10324–10329, 2010.

[131] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

[132] Igor B. Rogozin, Liran Carmel, Miklos Csuros, and Eugene V. Koonin. Origin and evolution of spliceosomal introns. *Biology Direct*, 7(1):1, 2012.

[133] Yasser Roudi, Erik Aurell, and John A. Hertz. Statistical physics of pairwise probability models. *Frontiers in Computational Neuroscience*, 3(NOV):1–15, 2009.

[134] Anthony G. Russell, J. Michael Charette, David F. Spencer, and Michael W. Gray. An early evolutionary origin for the minor spliceosome. *Nature*, 443(7113):863–866, 2006.

[135] Veronica Ryan, Gregory L. Dignon, Gul H. Zerze, Charlene V. Chabata, Rute Silva, Alexander E. Conicella, Joshua Amaya, Kathleen A. Burke, Jeetain Mittal, and Nicolas L. Fawzi. Mechanistic view of hnRNPA2 low complexity domain structure, interactions, and phase separation altered by disease mutation and arginine methylation. *Mol Cell.*, 69(3):465–479, 2018.

[136] Patty Sachamitr, Jolene C. Ho, Felipe E. Ciamponi, Wail Ba-Alawi, Fiona J. Coutinho, Paul Guilhamon, Michelle M. Kushida, Florence M.G. Cavalli, Lilian Lee, Naghmeh Rastegar, Victoria Vu, María Sánchez-Osuna, Jasmin Coulombe-Huntington, Evgeny Kanshin, Heather Whetstone, Mathieu Durand, Philippe Thibault, Kirsten Hart, Maria Mangos, Joseph Veyhl, Wenjun Chen, Nhat Tran, Bang Chi Duong, Ahmed M. Aman, Xinghui Che, Xiaoyang Lan, Owen Whitley, Olga Zaslaver, Dalia Barsyte-Lovejoy, Laura M. Richards, Ian Restall, Amy Caudy, Hannes L. Röst, Zahid Quyoom Bonday, Mark Bernstein, Sunit Das, Michael D. Cusimano, Julian Spears, Gary D. Bader, Trevor J. Pugh, Mike Tyers, Mathieu Lupien, Benjamin Haibe-Kains, H. Artee Luchman, Samuel Weiss, Katlin B. Massirer, Panagiotis Prinos, Cheryl H. Arrowsmith, and Peter B. Dirks. PRMT5 inhibition disrupts splicing and stemness in glioblastoma. *Nature Communications*, 12(1):1–17, 2021.

[137] Kentaro Sahashi, Akio Masuda, Tohru Matsuura, Jun Shinmi, Zhujun Zhang, Yasuhiro Takeshima, Masafumi Matsuo, Gen Sobue, and Kinji Ohno. In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5 splice sites. *Nucleic Acids Research*, 35(18):5995–6003, 2007.

[138] Noboru Jo Sakabe and Sandro José de Souza. Sequence features responsible for intron retention in human. *BMC Genomics*, 8:1–14, 2007.

[139] David Sankoff. Minimal Mutation Trees of Sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.

[140] Marc Santolini, Thierry Mora, and Vincent Hakim. A general pairwise interaction model provides an accurate description of in Vivo transcription factor binding sites. *PLoS ONE*, 9(6), 2014.

[141] Bernhard Schaefke, Wei Sun, Yi Sheng Li, Liang Fang, and Wei Chen. The evolution of posttranscriptional regulation. *Wiley Interdisciplinary Reviews: RNA*, 9(5):1–20, 2018.

[142] Klaus Peter Schliep. phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.

[143] Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[144] Schraga Schwartz, David Burstein, Genome Res, Supplemental Research Data, References Article, Genome Research, and Cold Spring Harbor Laboratory Press. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Research*, pages 88–103, 2008.

[145] Manli Shen and William Mattox. Activation and repression functions of an SR splicing regulator depend on exonic versus intronic-binding position. *Nucleic Acids Research*, 40(1):428–437, 2012.

[146] Craig G. Simpson, Graham Thow, Gillian P. Clark, S. Nikki Jennings, Jenny A. Watters, and John W.S. Brown. Mutational analysis of a plant branchpoint and polypyrimidine tract required for constitutive splicing of a mini-exon. *RNA*, 8(1):47–56, 2002.

[147] Robert J. Sims, Scott Millhouse, Chi Fu Chen, Brian A. Lewis, Hediye Erdjument-Bromage, Paul Tempst, James L. Manley, and Danny Reinberg. Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Postinitiation Factors and Pre-mRNA Splicing. *Molecular Cell*, 28(4):665–676, 2007.

[148] Michael Socolich, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, 2005.

[149] Baoxing Song, Qing Sang, Hai Wang, Huimin Pei, Xiang Chao Gan, and Fen Wang. Complement Genome Annotation Lift Over Using a Weighted Sequence Alignment Strategy. *Frontiers in Genetics*, 10(November):1–10, 2019.

[150] Reed Sorenson and Julia Bailey-Serres. Selective mRNA sequestration by OLIGOURIDYLATEBINDING PROTEIN 1 contributes to translational control during hypoxia in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 111(6):2373–2378, 2014.

[151] Noah Spies, Cydney B. Nielsen, Richard A. Padgett, and Christopher B. Burge. Biased Chromatin Signatures around Polyadenylation Sites and Exons. *Molecular Cell*, 36(2):245–254, 2009.

[152] Dorothee Staiger and John W.S. Brown. Alternative splicing at the intersection of biological timing, development, and stress responses, oct 2013.

[153] Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[154] R.Michael Stephens and Thomas Dana Schneider. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *Journal of Molecular Biology*, 228(4):1124–1136, dec 1992.

[155] Alexander V Sverdlov, Igor B Rogozin, Vladimir N Babenko, and Eugene V Koonin. Evidence of Splice Signal Migration from Exon to Intron during Intron Evolution. *Current Biology*, 13:2170–2174, 2003.

[156] Darren Qiancheng Tan, Ying Li, Chong Yang, Jia Li, Shi Hao Tan, Desmond Wai Loon Chin, Ayako Nakamura-Ishizu, Henry Yang, and Toshio Suda. PRMT5 Modulates Splicing for Genome Integrity and Preserves Proteostasis of Hematopoietic Stem Cells. *Cell Reports*, 26(9):2316–2328.e6, 2019.

[157] T. A. Thanaraj and A. J. Robinson. Prediction of exact boundaries of exons. *Briefings in bioinformatics*, 1(4):343–356, 2000.

[158] Hagen Tilgner, Christoforos Nikolaou, Sonja Althammer, Michael Sammeth, Miguel Beato, Juan Valcárcel, and Roderic Guigó. Nucleosome positioning as a determinant of exon recognition. *Nature Structural and Molecular Biology*, 16(9):996–1001, 2009.

[159] Miranda L. Tradewell, Zhenbao Yu, Michael Tibshirani, Marie Chloé Boulanger, Heather D. Durham, and Stéphane Richard. Arginine methylation by prmt1 regulates nuclear-cytoplasmic localization and toxicity of FUS/TLS harbouring ALS-linked mutations. *Human Molecular Genetics*, 21(1):136–149, 2012.

[160] Cole Trapnell, David G. Hendrickson, Martin Sauvageau, Loyal Goff, John L. Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1):46–53, 2013.

[161] Michael L. Tress, Federico Abascal, and Alfonso Valencia. Alternative Splicing May Not Be the Key to Proteome Complexity, feb 2017.

[162] Janne J. Turunen, Elina H. Niemelä, Bhupendra Verma, and Mikko J. Frilander. The significant other: Splicing by the minor spliceosome. *Wiley Interdisciplinary Reviews: RNA*, 4(1):61–76, 2013.

[163] Julian Vosseberg and Berend Snel. Domestication of self-splicing introns during eukaryogenesis: the rise of the complex spliceosomal machinery. *Biology Direct*, 12(30), 2017.

[164] Ruixue Wan, Rui Bai, Chuangye Yan, Jianlin Lei, and Yigong Shi. Structures of the Catalytically Activated Yeast Spliceosome Reveal the Mechanism of Branching. *Cell*, 177(2):339–351.e13, 2019.

[165] Ruixue Wan, Rui Bai, Xiechao Zhan, and Yigong Shi. How Is Precursor Messenger RNA Spliced by the Spliceosome? *Annual Review of Biochemistry*, 89:333–358, 2020.

[166] Bing Bing Wang and Volker Brendel. The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. *Genome biology*, 5(12), 2004.

[167] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, nov 2008.

[168] Xuncheng Wang, Mei Yang, Diqiu Ren, William Terzaghi, Xing Wang Deng, and Guangming He. Cis-regulated alternative splicing divergence and its potential contribution to environmental responses in Arabidopsis. *Plant Journal*, 97(3):555–570, 2019.

[169] Zbigniew Warkocki, Peter Odenwälder, Jana Schmitzová, Florian Platzmann, Holger Stark, Henning Urlaub, Ralf Ficner, Patrizia Fabrizio, and Reinhard Lührmann. Reconstitution of both steps of Saccharomyces cerevisiae splicing with purified spliceosomal components. *Nature Structural and Molecular Biology*, 16(12):1237–1243, 2009.

[170] Robert J. Weatheritt, Timothy Sterne-Weiler, and Benjamin J. Blencowe. The ribosome-engaged landscape of alternative splicing. *Nature Structural and Molecular Biology*, 23(12):1117–1123, dec 2016.

[171] Thomas M. Winkelmüller, Frederickson Entila, Shajahan Anver, Anna Piasecka, Baoxing Song, Eik Dahms, Hitoshi Sakakibara, Xiangchao Gan, Karolina Kułak, Aneta Sawikowska, Paweł Krajewski, Miltos Tsiantis, Ruben Garrido-Oter, Kenji Fukushima, Paul Schulze-Lefert, Stefan Laurent, Paweł Bednarek, and Kenichi Tsuda. Gene expression evolution in pattern-triggered immunity within Arabidopsis thaliana and across Brassicaceae species. *Plant Cell*, 33(6):1863–1887, 2021.

[172] Justin J.L. Wong, Amy Y.M. Au, William Ritchie, and John E.J. Rasko. Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. *BioEssays*, 38(1):41–49, jan 2016.

[173] Mandy S Wong, Justin B Kinney, and Adrian R Krainer. Quantitative Activity Profile and Context Dependence of All Human 5prime; Splice Sites. *Molecular Cell*, 71, 2018.

[174] Liming Xiong, Zhizhong Gong, Christopher D. Rock, Senthil Subramanian, Yan Guo, Wenying Xu, David Galbraith, and Jian Kang Zhu. Modulation of Abscisic Acid Signal Transduction and Biosynthesis by an Sm-like Protein in Arabidopsis. *Developmental Cell*, 1(6):771–781, dec 2001.

[175] Chuangye Yan, Ruixue Wan, Rui Bai, Gaoxingyu Huang, and Yigong Shi. Structure of a yeast step II catalytically activated spliceosome. *Science*, 355(6321), 2017.

[176] Chuangye Yan, Ruixue Wan, and Yigong Shi. Molecular mechanisms of pre-mRNA splicing through structural biology of the spliceosome. *Cold Spring Harbor Perspectives in Biology*, 11(1), 2019.

[177] Gene Yeo, Shawn Hoon, Byrappa Venkatesh, and Christopher B. Burge. Variation in sequence and organization of splicing regulatory elements in vertebrate genes.

*Proceedings of the National Academy of Sciences of the United States of America*, 101(44):15700–15705, 2004.

[178] Wentao Zhang, Bojing Du, Di Liu, and Xiaoting Qi. Splicing factor SR34b mutation reduces cadmium tolerance in Arabidopsis by regulating iron-regulated transporter 1 gene. *Biochemical and Biophysical Research Communications*, 455(3-4):312–317, dec 2014.

[179] Zhaoliang Zhang, Shupei Zhang, Ya Zhang, Xin Wang, Dan Li, Qiuling Li, Minghui Yue, Qun Li, Yu e. Zhang, Yunyuan Xu, Yongbiao Xue, Kang Chong, and Shilai Bao. Arabidopsis floral initiator SKB1 confers high salt tolerance by regulating transcription and pre-mRNA splicing through altering histone H4R3 and small nuclear ribonucleoprotein LSM4 methylation. *Plant Cell*, 23(1):396–411, jan 2011.

# Apéndice A

# Figuras suplementarias

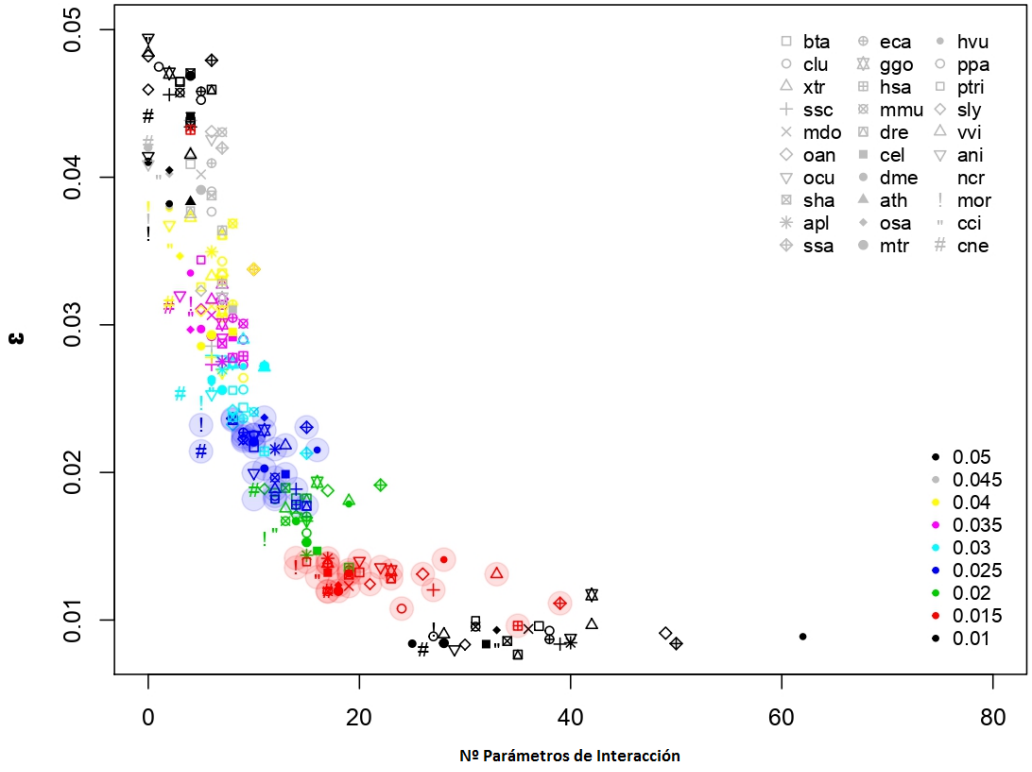



**Figura A.1**: **Convergencia de los modelos.** *Devianza absoluta máxima entre las probabilidades de co-ocurrencia entre dos posiciones observadas y estimadas,* $\Delta = max[abs(f_{ij} - P_{ij})]$*, en función de la cantidad de parámetros de interacción con valor distinto a cero, bajo distintos niveles de regularización (en distintos colores,* $\gamma \in [0{,}01, 0{,}05]$*). Los resultados para los distintos genomas analizados se muestran en la figura con distintos símbolos (ver legenda). Con sombreado azul y rojo se resaltan los resultados obtenidos con* $\gamma = 0{,}025$ *y* $0{,}015$*, respectivamente.*

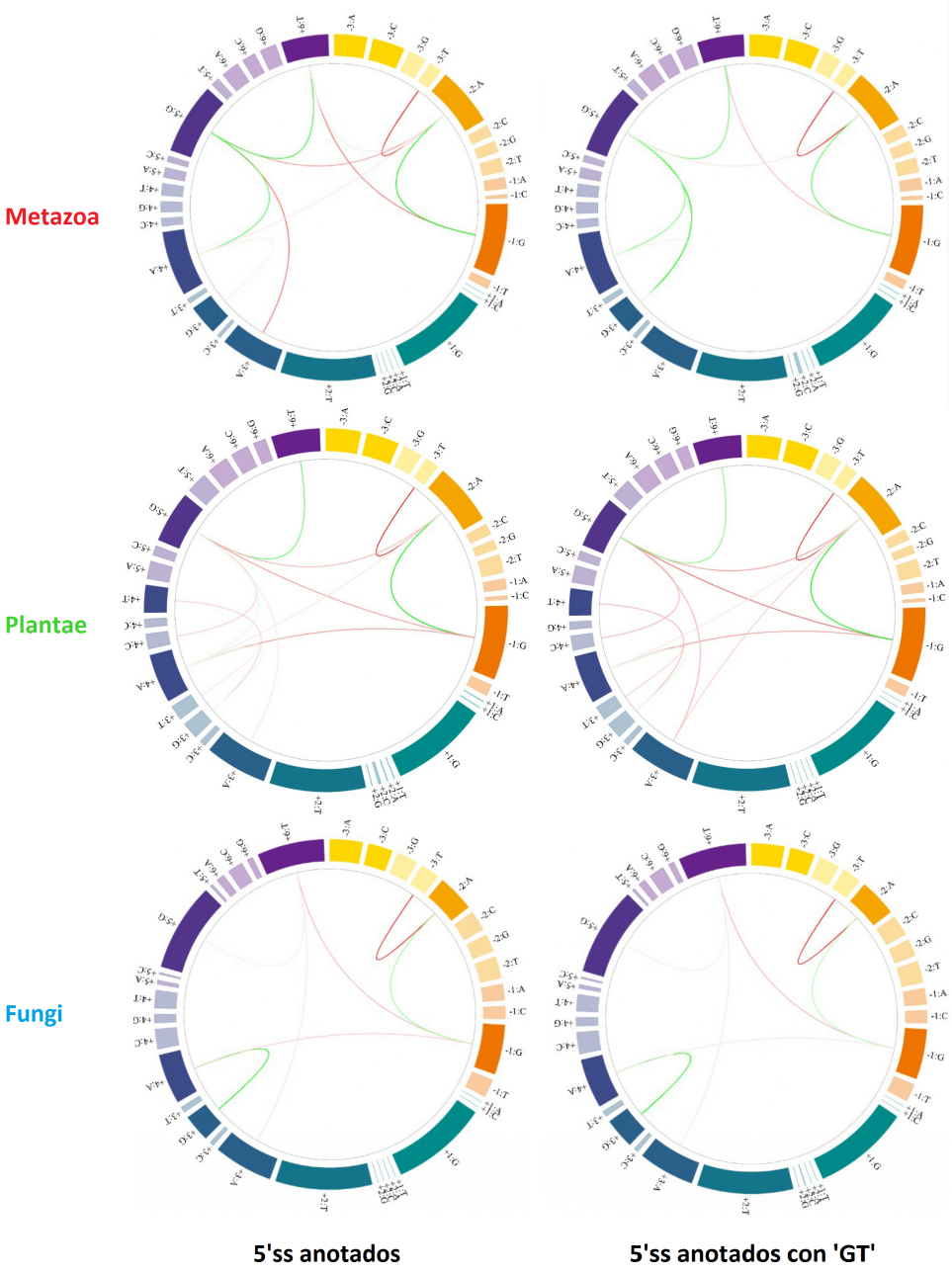



**Figura A.2**: **Diagramas de circos de los modelos para plantas, animales y hongos**. *Para los modelos construidos a partir de las secuencias de los 5'ss de animales, plantas y hongos; teniendo en cuenta todas las secuencias de las anotaciones de los genomas de las especies o solamente las que presentan el di-nucleótido consenso GT al comienzo del intrón.*

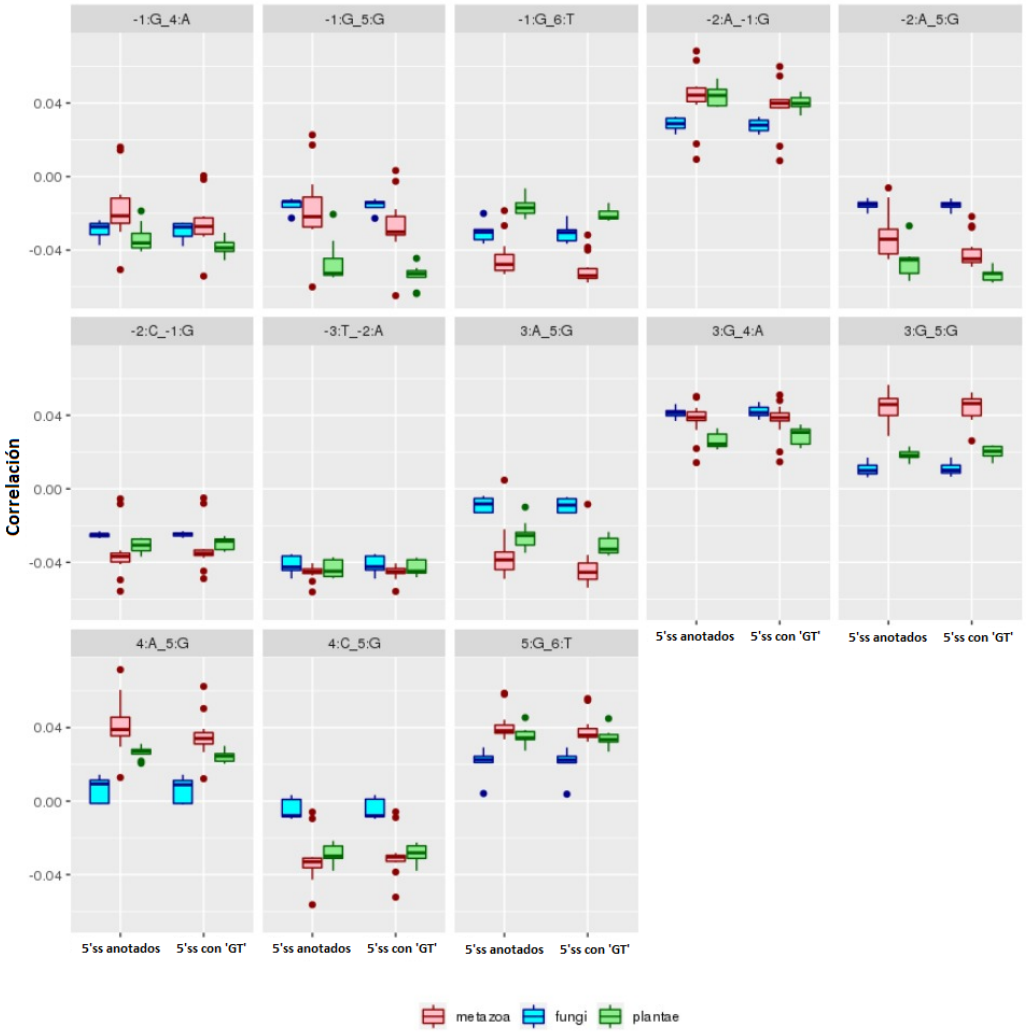



**Figura A.3**: **Correlaciones entre pares de nucleótidos**. *Cada panel muestra gráficos de cajas de la distribución de los valores de correlación obtenida para las especies de hongos (azul), animales (rojo) y plantas (verde) para un determinado por de nucleótidos, tanto en el conjunto entero de los 5'ss que se encuentran en el genoma como para el subconjunto de éstos que presentan una GT al inicio del intrón.*

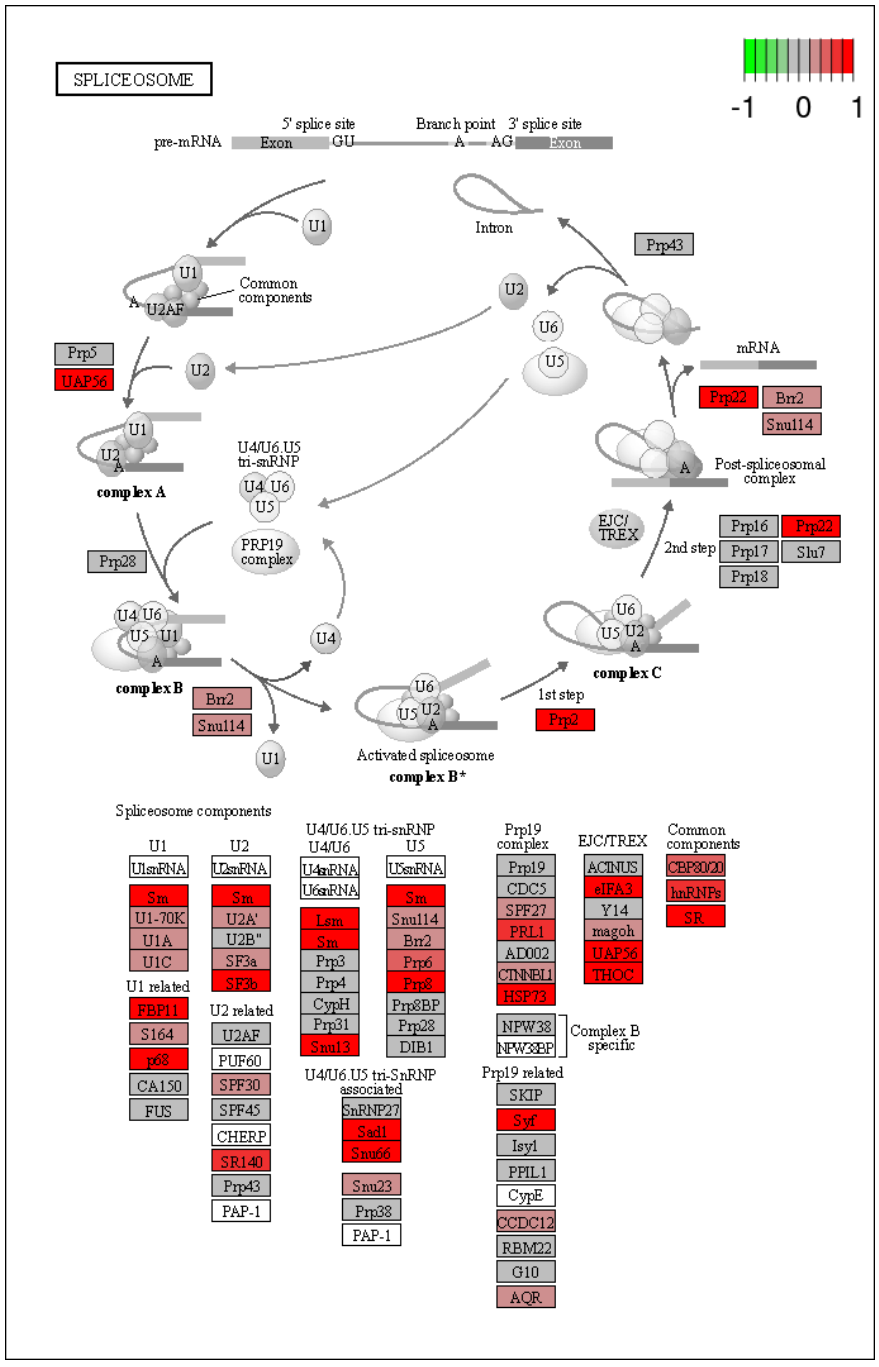



**Figura A.4**: **DEG relacionados con el ciclo de *splicing* en Ler**. *Análisis de enriquecimiento en pathways de KEGG para Ler. Se consideró el FC obtenido a partir de la comparación entre plantas salvajes y mutantes para PRMT5 en la accesión Ler. Los genes resaltados en rojo presentan un aumento de su expresión en esta última condición. Diagrama realizado a partir del paquete de R pathways.*

# Apéndice B

# Tablas suplementarias

| Especies | IC-EC | IC-ENC | INC-EC | INC-ENC | IC-IC | IC-INC | INC-INC | EC-EC | ENC-EC | ENC-ENC |
|---|---|---|---|---|---|---|---|---|---|---|
| **cne** | -0.08 | 0 | 0 | 0 | 0.34 | 0 | 0 | 0.92 | 0 | -0.17 |
| **ani** | -0.48 | 0 | 0 | 0 | 0.06 | 0 | 0 | 0.88 | 0 | 0 |
| **ncr** | -0.56 | 0 | 0 | 0 | 0.26 | 0 | 0 | 0.79 | 0 | 0 |
| **mor** | -0.29 | 0 | 0 | 0 | 0 | 0 | 0 | 0.96 | 0 | 0 |
| **cci** | -0.28 | 0 | 0 | 0 | 0.24 | 0.04 | 0 | 0.93 | 0 | 0 |
| **ath** | -0.37 | 0 | 0.02 | 0 | 0.19 | 0 | 0 | 0.91 | 0 | 0 |
| **hvu** | 0 | 0.13 | 0 | -0.08 | 0 | 0.94 | 0.19 | 0 | 0 | -0.23 |
| **mtr** | -0.56 | 0 | 0 | 0 | 0.07 | 0 | 0 | 0.82 | 0 | 0 |
| **osa** | -0.22 | 0 | 0 | 0 | 0.07 | 0.02 | -0.03 | 0.97 | 0 | 0 |
| **ppa** | -0.84 | 0 | 0 | 0 | 0.05 | 0 | -0.03 | 0.55 | 0 | 0 |
| **ptri** | -0.42 | 0 | 0 | 0 | 0.02 | 0 | 0 | 0.91 | 0 | 0 |
| **sly** | -0.33 | 0 | 0 | 0 | 0.23 | 0 | 0 | 0.92 | 0 | 0 |
| **vvi** | -0.34 | 0 | 0 | 0 | 0.01 | 0 | 0 | 0.94 | 0 | 0 |
| **apl** | -0.41 | 0 | 0 | 0 | 0.39 | 0 | 0 | 0.83 | 0 | 0 |
| **bta** | -0.27 | 0 | 0 | 0 | 0.59 | 0 | 0 | 0.76 | 0 | 0 |
| **clu** | -0.33 | 0 | 0 | 0 | 0.56 | 0 | 0 | 0.76 | 0 | 0 |
| **dre** | -0.34 | 0 | 0 | 0 | 0.24 | 0 | 0 | 0.91 | 0 | 0 |
| **eca** | -0.36 | 0 | 0 | 0 | 0.51 | 0 | 0 | 0.78 | 0 | 0 |
| **ggo** | -0.22 | 0 | 0 | 0 | 0.57 | 0 | 0 | 0.79 | 0 | 0 |
| **hsa** | -0.55 | 0 | 0 | 0 | 0.16 | 0 | 0 | 0.82 | 0 | 0 |
| **mdo** | -0.19 | 0 | 0 | 0 | 0.57 | 0 | 0 | 0.8 | 0 | 0 |
| **mmu** | -0.58 | 0 | 0 | 0 | 0.12 | 0 | 0 | 0.81 | 0 | 0 |
| **oan** | -0.23 | 0 | 0 | 0 | 0.61 | 0 | 0 | 0.76 | 0 | 0 |
| **ocu** | -0.27 | 0 | 0 | 0 | 0.59 | 0 | 0 | 0.76 | 0 | 0 |
| **sha** | -0.31 | 0 | 0 | 0 | 0.63 | 0 | 0 | 0.71 | 0 | 0 |
| **ssa** | 0.11 | 0 | -0.05 | 0 | 0.61 | 0 | 0 | 0.78 | 0 | 0 |
| **ssc** | -0.18 | 0 | 0 | 0 | 0.62 | 0 | 0 | 0.77 | 0 | 0 |
| **xtr** | 0.17 | 0 | -0.05 | 0 | 0.59 | 0 | 0 | 0.79 | 0 | 0 |
| **dme** | -0.97 | 0 | 0 | 0 | -0.26 | 0 | 0 | 0 | 0 | 0 |
| **cel** | -0.97 | 0 | 0.06 | 0 | -0.21 | 0 | 0 | 0 | 0 | -0.1 |

**Cuadro B.1**: **Patrones de interacción** ($\gamma = \mathbf{0{,}015}$) *Se muestra la media de intensidad de la interacción de diferentes tipos de posiciones. EC: posiciones con nucleótidos consenso en el exón, ENC: posiciones con nucleótidos no consenso en el exón, IC: posiciones con nucleótidos consenso en el intrón, y INC: posiciones con nucleótidos no consenso en el intrón.*

| **Muestra** | **Col-0 Tair10** | **Ler pseudogenoma** |
|---|---|---|
| Col-$0_{wt} - replica1$ | 66312170 | - |
| Col-$0_{wt} - replica2$ | 36566624 | - |
| Col-$0_{wt} - replica3$ | 71906834 | - |
| $\text{Ler}_{wt} - replica1$ | - | 55484532 |
| $\text{Ler}_{wt} - replica2$ | - | 54223928 |
| $\text{Ler}_{wt} - replica3$ | - | 37727040 |
| Col-$0_{prmt5} - replica1$ | 49259634 | - |
| Col-$0_{prmt5} - replica2$ | 39783546 | - |
| Col-$0_{prmt5} - replica3$ | 64347230 | - |
| $\text{Ler}_{prmt5} - replica1$ | - | 71333436 |
| $\text{Ler}_{prmt5} - replica2$ | - | 58802752 |
| $\text{Ler}_{prmt5} - replica3$ | - | 51625468 |
| Col-$0_{wt} X \text{Ler}_{wt} - replica1$ | 26446074(3061378) | 26385428(2893088) |
| Col-$0_{wt} X \text{Ler}_{wt} - replica2$ | 42057684(4901710) | 41988048(4648090) |
| Col-$0_{wt} X \text{Ler}_{wt} - replica3$ | 53473244(6205970) | 53389466(5917522) |
| $\text{Ler}_{wt} X Col - 0_{wt} - replica1$ | 32683604(3762578) | 32646732(3609772) |
| $\text{Ler}_{wt} X Col - 0_{wt} - replica2$ | 126701648(14900664) | 126385756(14013622) |
| $\text{Ler}_{wt} X Col - 0_{wt} - replica3$ | 87050630(10193638) | 86866886(9678214) |
| Col-$0_{prmt5} X \text{Ler}_{prmt5} - replica1$ | 55468462(6585926) | 55406802(6159686) |
| Col-$0_{prmt5} X \text{Ler}_{prmt5} - replica2$ | 43138138(5048696) | 43091660(4692680) |
| Col-$0_{prmt5} X \text{Ler}_{prmt5} - replica3$ | 70169254(8264284) | 70128358(7645704) |
| $\text{Ler}_{prmt5} X Col - 0_{prmt5} - replica1$ | 47475188(5657418) | 47415662(5325872) |
| $\text{Ler}_{prmt5} X Col - 0_{prmt5} - replica2$ | 74359472(8845602) | 74331938(8187574) |
| $\text{Ler}_{prmt5} X Col - 0_{prmt5} - replica3$ | 26041508(3112144) | 26021268(2903566) |

**Cuadro B.2**: **Lecturas mapeadas por muestra***: Número de lecturas mapeadas al genoma de referencia de Col-0, TAIR10, y al genoma de Ler construido a partir de la información de SNPs/Indels extraída de 1001 Genomes. En el caso de los híbridos, se muestra el total de lecturas mapeadas para cada genoma y, entre paréntesis, el número de lecturas asignadas de forma inequívoca a uno de los dos alelos.*