Tesis Doctoral

Identificación y caracterización de eventos de splicing alternativo en Arabidopsis thaliana utilizando herramientas de transcriptómica de alto rendimiento

Mancini, Estefanía

2016-12-27

Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Identificación y caracterización de eventos de *splicing* alternativo en *Arabidopsis thaliana* utilizando herramientas de transcriptómica de alto rendimiento

Lic. Estefania Mancini

Tesis a presentar para optar al título de

Doctor en el área de Ciencias Biológicas

de la Universidad de Buenos Aires

Director de tesis: **Dr. Marcelo Yanovsky**
Director asistente: **Dr. Ariel Chernomoretz**
Consejero de estudios: **Dr. Pablo Cerdán**
Lugar de trabajo:
Laboratorio de genómica comparativa del desarrollo vegetal. Laboratorio de biología de sistemas integrativa. Instituto de Investigaciones Bioquímicas de Buenos Aires (IIBBA) Fundación Instituto Leloir (FIL).

Ciudad Autónoma de Buenos Aires, Argentina.

Fecha de defensa: 27 de diciembre de 2016

Índice general

Índice de figuras

Índice de cuadros

Identificación y caracterización de eventos de *splicing* alternativo en *Arabidopsis thaliana* utilizando herramientas de transcriptómica de alto rendimiento

Resumen

El mecanismo de *splicing* alternativo es un mecanismo de regulación post-transcipcional presente en los organismos eucariotas que amplía las capacidades regulatorias y funcionales de los genes mediante la generación de múltiples isoformas. Las consecuencias de este proceso son proteínas funcional y estructuralmente diferentes como así también productos no traducibles con funciones regulatorias en sí mismos. El advenimiento de las nuevas tecnologías de secuenciación masiva, incluyendo las que se utilizan para ARN (RNA-Seq) permiten estudiar cambios transcripcionales incluyendo splicing alternativo de una manera inusitada. Sin embargo, la descripción de los cambios en los patrones de *splicing* bajo diferentes condiciones no es trivial y ha dado lugar durante los últimos años al desarrollo de múltiples herramientas bioinformáticas.

En este trabajo de tesis, se describe **ASpli**, un paquete de R integrativo y fácil de usar que facilita el análisis de los cambios en el *splicing* alternativo tanto en eventos anotados como nuevos a partir de datos de RNA-Seq. Nuestra propuesta es combinar la información estadística del uso diferencial de exones, intrones y junturas junto con las métricas de inclusión (PSI) y retención de intrón (PIR) que se obtienen por el análisis de las junturas sobre cada región genómica en estudio. La utilización de este abordaje integral intenta cuantificar de manera más exacta la ocurrencia de eventos de splicing alternativo. El desarrollo del paquete fue fundamental para el análisis de la remodelación del transcriptoma ante numerosos tratamientos en la planta modelo *Arabidopsis thaliana* así como en otros organismos. Como parte de los resultados, se describen los análisis del efecto de un pulso de luz en plantas y sus consecuencias globales en la regulación del *splicing* alternativo.

Palabras clave: *splicing* alternativo, RNA-seq, bioinformática, *Arabidopsis thaliana*, genómica

Identification and characterization of alternative splicing events in *Arabidopsis thaliana* using high throughput sequencing

Abstract

Alternative splicing (AS) is a common mechanism of post-transcriptional gene regulation in eukaryotic organisms that expands the functional and regulatory diversity of a single gene by generating multiple mRNA isoforms that encode structurally and functionally distinct proteins. The development of novel high-throughput sequencing methods for RNA (RNA-Seq) has provided a powerful means of studying AS under multiple conditions and in a genome-wide manner. Despite the fact that a plethora of bioinformatic tools have been developed in the last few years, using RNA-Seq to study changes in AS under different experimental conditions is not trivial.

In this thesis, we describe **ASpli**, an integrative and user-friendly R package that facilitates the analysis of changes in both annotated and novel AS events. Our method combines statistical information from exon, intron, and splice junction differential usage, with information from splice junction reads to calculate differences in the percentage of exon inclusion **(PSI)** and intron retention **(PIR)**, which reliably reflect the magnitude of changes in the relative abundance of different annotated and novel AS events. This approach is intended to improve the analysis of RNA-Seq data for the quantification and discovery of AS events. The package has been intensively used for the analysis of alternative splicing in a genome-wide manner in *Arabidopsis thaliana* as well as other organisms. As part of the results, we present the analysis of the accute effects of light on alternative splicing in light-grown plants.

Key words: alternative splicing, RNA-Seq, bioinformatics, *Arabidopsis thaliana*, genomics

Abreviaturas más frecuentes

- ADN: ácido desoxiribonucleico
- ARN: ácido ribonucleico
- ARNm: ácido ribonucleico mensajero
- pre ARNm: ácido ribonucleico inmaduro
- poliA: poliadenilación
- RNA-Seq: secuenciación masiva del ARN
- ADNc: ADN complementario
- EST: del inglés *expressed sequence tag*
- RT-PCR: del inglés *Reverse Transcription Polymerase Chain Reaction*
- SE: salteo de exon
- RI: retención de intrón
- 5'Alt, 3'Alt: dadores/aceptores 5'/3' alternativos
- PSI: del inglés *percent inclusion* o *percent spliced in*
- PIR: del inglés *percent intron retention*
- *snRNPs*: del inglés *small nuclear ribonucleoproteins*
- *snRNAs*: del inglés *small nuclear ribonucleic acid*
- miARNs: micro ácido ribonucleicos
- BAM: del inglés *Binary Alignment Map*
- FDR: del inglés *false discovery rate*
- p-valor: valor de probabilidad
- GO: del inglés *Gene onthology*
- *hnRNPs*: del inglés *heterogeneous nuclear ribonucleoproteins*
- *SREs*: del inglés *splicing regulatory elements*

Agradecimientos

A lo largo de esta tesis me han acompañado muchas personas al punto que recordarlas y homenajearlas con esta pequeña dedicatoria realmente me acongoja. Especialmente quiero agradecer a las siguientes personas que de una u otra manera me sostuvieron para que no extrañe tanto el aire puro y las distancias cortas, entre tantas otras cosas.

A **Marcelo Yanovsky**, por su inconmensurable entrega, dedicación, paciencia y ejemplo en lo cotidiano. Consciente de que este trabajo no refleja exactamente todo lo que hicimos juntos, tranquila porque queda reflejado en múltiples artículos y colaboraciones que surgieron a partir de la tesis e ilusionada que sea un vínculo que dure para siempre. Gracias Marce por ser infinita fuente de inspiración para todos.

A **Ariel Chernomoretz**, por su ejemplo de disfrutar de hacer ciencia sin fines de lucro.

A **Cristina Marino**, porque su presencia en el laboratorio me dio lugar a creer que se puede hacer buena ciencia siendo mujer y madre y porque siempre está dispuesta a dar un consejo, una mano y organizar una salida.

A **Martín Vázquez**, porque cuando me abrió las puertas de la plataforma de Genómica y Bioinformática, me abrió los ojos a hacer ciencia desde otro lado, desde la innovación, la proactividad, los datos y la tecnología aplicada sin perder la pasión. Y porque gracias a él también conocí el Instituto Leloir y la facultad de Ciencias Exactas de la cual hoy egreso.

A **Andrea Gamarnik**, porque con la excusa de hablar con ella por un proyecto de secuenciación, vine a Buenos Aires, conocí Leloir y entendí que había preguntas de biología interesantísimas para hacerle a *las secuencias*. Y luego, por esas cosas de la vida, *otras secuencias* llegaron a nuestras manos. Gracias Andrea por depositar tu confianza en nosotros sin dudarlo ni un segundo. Y por ser un ejemplo para todos para que haya más mujeres científicas.

A **Fernán Agüero**, porque me acompañó como miembro del comité de seguimiento y luego

me dio lugar como docente en la Universidad de San Martín, a la cual le tomé un afecto especial. Gracias Fernán por tus consejos profesionales y personales siempre atinados y Gracias Javier y Emilio, compañeros docentes de la materia por los mates y los ratos y acomodarse los horarios para que pueda dar clases sin perder congresos y viajes.

A **Luis Esteban** y **Esteban Serra**, que desde los orígenes de mi vida profesional me han acompañado. Siempre dispuestos a tomar un café y hacer un recreo en la agitada vida cotidiana de la facultad. Gracias por ser el faro que nunca me abandona, aún en los momentos más difíciles.

A **Carlos Dezar** por transmitirme su pasión por la biología molecular de plantas y darme la confianza para que analice sus datos cuando recién me iniciaba en la vida profesional. Gracias Carlos por enseñarme que el *interior* también existe, aunque se organice desde Buenos Aires.

A **Lucia**, por las miles de charlas, adentro y sobre todo afuera del laboratorio. Porque inocentemente cuando nos encontramos en Leloir por un curso, me sugirió que hable con Marcelo, que hacía genómica de plantas, justo lo que yo buscaba. Y porque la voy a extrañar, una de mis grandes amigas que me dejan estos años.

A **Elin**, una compañera de ruta. Por sus charlas y miles de consejos justos. Los pies en la tierra y la tranquilidad de la sabiduria. Gracias por todos los ratos!

Al **103** inolvidable y a las primas del **203**! A **Juli Mateos**, por sus consejos de hermana mayor, capricorniana y nómade por un tiempo. A **Gus**, porque me enseñó a pensar la vida desde otras ópticas, además de todo lo que me enseñó de *splicing* y extracción de ARN con tanta paciencia!!! A **Steve**, por su infinito cariño y paciencia. A **Marti**, porque vayamos siempre hasta las últimas consecuencias! **Caro**, por enseñarme que hay otra vida. **Andrés**, por aguantarse mi pensamiento literal que no entiende de chistes y mis modos de pocos amigos y corregirme los abstracts y papers y figuras N veces sin chistar. A **Majo**, por quitarle dramatismo a mi vida dramática. A **Javi**, que se lleva una de las partes más difíciles.. tomar la posta de ser el bioinformático del 103! A **Sole**, por sus consejos sobre la tesis, los papeles, la gente y su generosidad de compartir su trabajo. A **Beck**, casi otro hermano mayor, lástima que compartimos poco tiempo. A **Ceci**, por ser un ejemplo de vida.

A todo el labo de bioinformática. Por aguantarme dramática, exitista e intransigente con la vida. Los quiero mucho y los voy a extrañar. Siempre contarán con una compañera para un café *expreso*, una cervecita... o para hablar de ciencia obviamente

A todos los amigos que me acompañaron en Buenos Aires y a la distancia.

A **Las rompeportones**, Flor, Popi, Ale, Cuchi y Gri, eternas caricias al alma. Gracias por su paciencia y por hacerse un rato para hacerme feliz en San Genaro y en tantos lugares. Y por venir a visiarme las veces que han podido y bancarme la amiga ausente que fui todos estos años en pos de cumplir un objetivo personal.

A mis amigas de la facultad, **Romi y Marina**, que me dieron la confianza que una tesis se podía terminar. Y que estaba haciendo lo correcto cuando dejé Rosario y ahora dejo Argentina.

A la **Guille, Eli, Rolo, Juan y Eze**, la Peña Porteña. Infinitas gracias. Ninguno de nosotros será el mismo después del paso por esta Peña.

A **Belen, Meli y Sole**. Cada una a su modo me acompañó en la ciudad de la furia y en el intento de mitigar la angustia de nuestra existencia, de donde surgieron las mejores conversaciones en el depto de Laprida.

A **Nico**, porque se llevó la peor parte y aún así, su cariño sigue intacto y me consuela diciendo que siempre tuve buen *timing* (para la profesión). A **Diego**, porque se llevó la otra peor parte y en el mientras tanto, intentó mostrarme el lado B* de Buenos Aires y explicarme la vida con los modelos Bilardistas y Menotistas de las cosas. (B* de Boedo)

A mi tía **Marisa** y su familia, **Robi, Mariana, Lucca y Fede**, porque me acobijaron los primeros meses y luego me acompañaron para darme un ambiente de familia, de domingos, de cine, cafés y mucho afecto. Los quiero al infinito! Gracias Tia por ese consejo tan acertado: **Si es lo que querés, hay que ir a buscarlo hasta las últimas consecuencias.** Y acá estamos.

A mi hermana **Lucía**, ídola total, siempre me acompañó de manera incondicional y sin reclamos por las ausencias. Por lo que hemos disfrutado las visitas y por todo lo que viene por disfrutar! Gracias hermana!

A mis papás **Tuli y Hugo**, porque me dieron todo el apoyo incondicional y la libertad y día a día hacen el esfuerzo de comprender lo que significa ser científico en Argentina, sin reclamos.

Gracias a la vida, que me ha dado tanto

Contribuciones

Los resultados de esta tesis han sido publicados parcialmente en los siguientes artículos:

- **Acute effects of light on alternative splicing in light-grown plants** Estefanía Mancini, Sabrina Sánchez, Andrés Romanowski, Gustavo Schlaen, Maximiliano Sánchez-Lamas, Pablo Cerdán y Marcelo Yanovsky. *Photochemistry and Photobiology* (2016)

- **Role for LSM genes in the regulation of circadian rhythms** Soledad Pérez-Santángelo, Estefania Mancini, Lauren Francey, Gustavo Schlaen, Ariel Chernomoretz, John Hogenesch y Marcelo Yanovsky. *Proceedings National Academy of Sciences* (2014)

- **The spliceosome assembly factor GEMIN2 attenuates the effects of temperature on alternative splicing and circadian rhythms.** Gustavo Schlaen, Estefania Mancini, Sabrina Sanchez, Soledad Perez-Santángelo, Matias Rugnone, Craig Simpson, John Brown, Xin Zhang, Ariel Chernomoretz y Marcelo Yanovsky. *Proceedings National Academy of Sciences* (2015)

- **Genome wide comparative analysis of the effects of PRMT5 and PRMT4/CARM1 arginine methyltransferases on the** *Arabidopsis thaliana* **transcriptome**. Esteban Hernando, Sabrina Sanchez, Estefania Mancini y Marcelo Yanovsky. *BMC Genomics* (2015)

- **The Dengue virus NS5 protein intrudes in the cellular spliceosome and modulates splicing** Federico De Maio, Guillermo Risso, Nestor G. Iglesias, Priya Shah, Berta Pozzi, Leopoldo Gebhard, Pablo Mammi, Estefania Mancini, Marcelo Yanovsky, Raul Andino, Nevan Krogan, Anabella Srebrow y Andrea V. Gamarni. *Plos Pathogens* (2016)

Dedicatoria

¿Por qué el splicing?

Desde el momento en que tomé contacto con el laboratorio del Dr. Marcelo Yanovsky me sentí atraída por el mundo del *splicing*, ese mecanismo de regulación génica postranscripcional que estaba ahí, subyacente y que pocos se detenían a mirarlo. Venía de años intensos en el mundo de la secuenciación masiva donde ya había tomado contacto con datos de experimentos de ARN, pero nunca se me había ocurrido pensar que en esos datos había oculta tanta información sobre la naturaleza y que luego, una de nuestras misiones fue intentar revelarla.

Con el correr de los años fuimos aprendiendo muchísimo sobre este mecanismo, conservado, regulado y que sigue escondiendo una magia que aún hoy me sigue conmoviendo. La maquinaria del *splicing* es fascinante, es precisa, es deliberada en sus elecciones y se ha sostenido a través de los años.

Lo que más entusiasma, es saber que detrás de todo experimento de RNA-Seq hay algo más que podemos ver y que fuimos de los primeros en verlo. Y ahí está la magia de la ciencia, la que encandila y enciende la llama de la pasión de saber por saber nomás. Esa sensación que pocos se permiten experimentar, que implica tiempo, constancia, obstinación, agudeza, cansancio, equivocaciones y convencimiento que la natureleza tiene mucho para ofrecernos si nos detenemos a mirar bien.

Por todo eso y por lo que viene, le dedico esta tesis a todos los que se maravillan con la naturaleza, con *la vida, ese milagro*.

Capítulo 1

Introducción

1.1. Maduración del pre-ARNm y *splicing* alternativo

La producción de la cantidad correcta de proteína en la célula adecuada en el momento preciso es crucial para el crecimiento y el desarrollo de los eucariotas multicelulares y su respuesta al entorno. La regulación de la expresión génica es un componente central de este proceso. Esta regulación ocurre a múltiples niveles, incluyendo la exportación de transcriptos, el control de la estabilidad del ARNm, la traducción, las modificaciones post-traducionales de las proteínas y la degradación, que es lo que finalmente regula la cantidad de proteína en la célula [1]. Sin embargo en los últimos tiempos se ha revelado que la regulación de procesos co/post-transcripcionales, tales como el *splicing* y la poliadenilación también contribuyen a incrementar la complejidad de los mecanismos de control de la expresión génica.

La mayoría de los genes eucariotas que codifican para proteínas se transcriben en precursores del ARN mensajero (pre-ARNm) donde los exones, fragmentos de ADN que forman parte del transcripto maduro, son interrumpidos por segmentos que no estarán presentes en el transcripto maduro denominados intrones. Por lo tanto, estos pre-ARNm deben sufrir una etapa de maduración denominada *splicing* donde se remueven los intrones y se produce un transcripto traducible en la mayoría de los casos. El *splicing* alternativo ocurre cuando se seleccionan sitios de *splicing* alternativos, dando lugar a la aparición de más de un ARNm a partir de un mismo precursor (pre-ARNm). El fenómeno de *splicing* alternativo inicialmente se pensó como una forma poco común de regulación de la expresión génica [2, 3], pero con la aparición de las tecnologías de

biología molecular más modernas se fue demostrando que la mayorías de los genes eucariotas eran procesados de manera alternativa y de una manera controlada a lo largo de diferentes tejidos o estados de desarrollo. Esto indica que el *splicing* alternativo tienen un rol muy importante en el incremento de la complejidad funcional de un organismo y que es más la regla que la excepción [4].

El procesamiento del pre-ARNm se lleva a cabo en el *espliceosoma*, un complejo ribonucleoproteico dinámico con un ciclo funcional altamente regulado, que cataliza dos reacciones de transesterificación entre los sitios dador (sitio de *splicing* 5') y aceptor (sitio de *splicing* 3'). La maquinaria está compuesta por pequeñas ribonucleoproteínas nucleares (del inglés *small nuclear ribonucleoproteins, snRNPs*) y proteínas auxiliares. Los *snRNPs* son complejos ribonucleoproteicos espliceosomales compuestos por moléculas de ARN pequeñas nucleares (*snRNA*) ricas en uridinas, proteínas Sm y proteínas específicas de cada *snRNP*. De acuerdo a la composición proteínas y *snRNPs*, se presentan 2 tipos de espliceosomas:

1. **Espliceosoma mayor**: consiste en 5 *snRNPs* (U1, U2, U4, U5 y U6) y aproximadamente 250 proteínas. Procesa la mayoría de los pre-ARNm con sitios de *splicing* GT-AG.

2. **Espliceosoma menor**: se forma a partir de diferentes *snRNPs*: U11, U12, U4atac/U6atac con funciones análogas a los *snRNPs* U1, U2, U4 y U6 del complejo U2. Procesa transcriptos minoritarios que contienen los denominados intrones U12.

El *snRNP* U5 es común a ambos complejos.

El ensamblado del complejo espliceosomal se produce de manera secuencial, por unión de sus componentes sobre el intrón a remover. El primer paso en el reconocimiento de los sitios de *splicing* comprende la unión del *snRNP* U1 al sitio de *splicing* 5' y el factor auxiliar U2 (U2AF) al sitio de *splicing* 3'. La subunidad pequeña del U2AF se une a la región frontera exón-intrón y la subunidad grande se une a la región rica en pirimidinas (tracto de polipirimidina). A continuación, el *snRNP* U2 se une el punto de ramificación y el complejo preformado de los tri-*snRNPs*, U4/U6-U5 se recluta al intrón para formar el complejo espliceosomal. Durante el proceso de activación, el espliceosoma atraviesa severos rearreglos que conducen a la pérdida de los *snRNPs* U1 y U4 y la formación del complejo catalítico central a partir de los *snRNAs* U2, U5 y U6 y la proteína Prp8 [5,6] (Ver Figura 1.1).

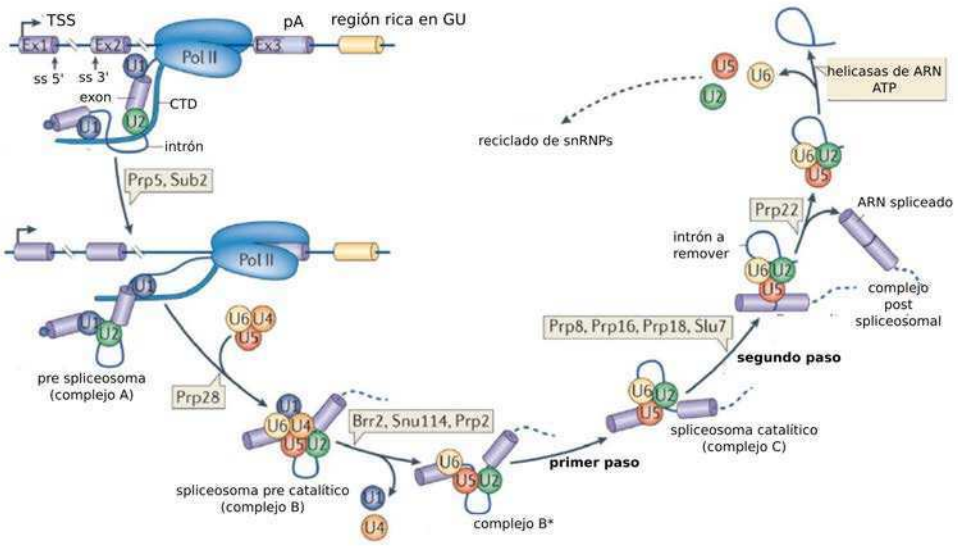


Figura 1.1: El ensamblado del espliceosoma. El ensamblado del espliceosoma ocurre en los sitios de transcripción. (a) Los *snRNPs* U1 y U2 se ensamblan en el pre-ARNm de una manera cotranscripcional, usando los sitios de *splicing* 5' (ss 5') y 3' (ss 3'), mediados por el dominio C-terminal (CTD) de la ARN-polimerasa II. Los *snRNPs* U1 y U2 interactúan entre sí para formar el pre-spliceosma (**complejo A**). Este proceso es dependiente de las helicasas DExD/H, Prp5 y Sub2. En una reacción subsecuence, en la reacción catalizada por Prp28, el complejo tri-*snRNP* preensamblado U4/U6.U5 es reclutado para formar el **complejo B**. El complejo B resultante sufre una serie de rearreglos para dar lugar al complejo catalítico activo complejo B*, que requiere de múltiples helicasas (Brr2, Snu114 y Prp2) y resultan en la liberación de los *snRNPs* U4 y U1. El Complejo B* lleva a cabo el primer paso catalítico del *splicing*, dando lugar a un **complejo C**, que contiene los exones libres 1 y el exón-intron 2 intermedio. El complejo C sufre rearreglos adicionales y lleva a cabo el segundo paso catalítico, que resultará en el complejo post-espliceosomal que contiene el intrón a remover y los exones unidos. Finalmente, se liberan los *snRNPs* U2, U5 y U6 de la molécula mRNP y se reciclan para los ciclos subsecuentes de *splicing*. Adaptado de [5]

Tipos de eventos de *splicing* alternativo

A pesar de que no hay un consenso en la clasificación de todos los tipos de eventos de *splicing* alternativo, la mayoría de los eventos resultan del uso de uno o más de cuatro módulos básicos (Figura 1.2). La forma de *splicing* alternativo más antigua reportada es la de salteo o exclusión del exón (**SE**), que consiste en la elección de uno o más exones de manera alternativa. Además, los exones pueden ser acortados o elongados a través del uso de sitios dadores (5') alternativos (**5' Alt**) y aceptores (3') alternativos (**3' Alt**). Cuando se suprime la excisión del intrón se denomina retención de intrón (**RI**). Este es uno de los eventos más difíciles de estudiar debido a la dificultad de distinguir de artefactos experimentales, que se pueden derivar por ejemplo, de analizar precursores de ARNs en vías de procesamiento. Los animales y las plantas difieren en el tipo de evento más frecuente. En humanos se ha reportado que el salteo de exón es el evento más frecuente (40 %), mientras que en plantas se reportó que ocurre en un 5 % [7] (Cuadro 1.1). Por otro lado, la retención de intrón se ha reportado en plantas como un evento muy frecuente (40 %) y mucho

menos en humanos [8]. Esta diferencia podría sugerir que las plantas y animales podrían estar reconociendo los exones e intrones en una manera diferente.

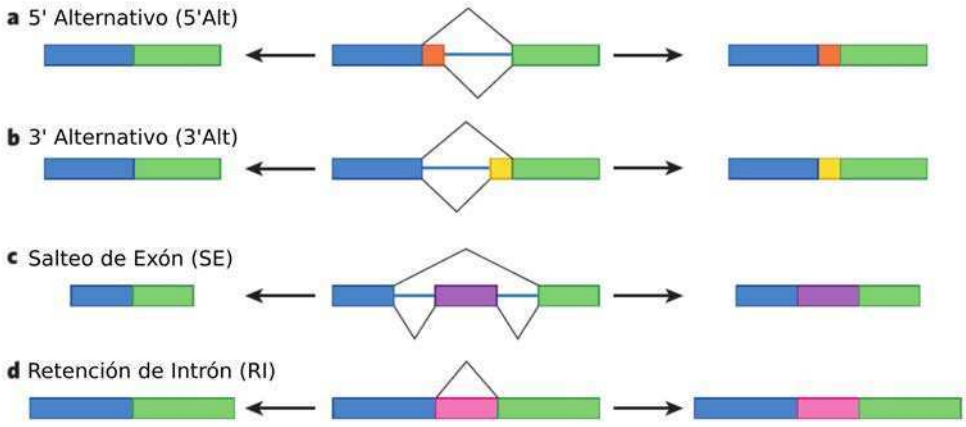


Figura 1.2: Clasificación de los eventos de *splicing* alternativo. Hay cuatro tipos de eventos de *splicing* alternativo: dador (a)(5'Alt)/aceptor (3'Alt) (b) alternativo, salteo de exón (SE)(c) y retención de intrón (RI)(d). Los rectángulos centrales representan los pre-ARNm. Para cada pre-ARNm, las líneas negras señalan las combinaciones entre los exones. Hacia la izquierda y derecha, se muestran los posibles productos finales. Adaptado de [4].

Clase	Humanos	Plantas
Salteo de exón	>40 %	8 %s
Sitio de *splicing* 5' alternativo	19 %	15.5 %
Sitio de *splicing* 3' alternativo	7.9 %	7.5 %
Retención de intrón	<5 %	40 %

Cuadro 1.1: Frecuencia de los eventos de *splicing* alternativo en humanos y plantas. Adaptado de [9]

1.2. Elementos regulatorios

El complejo del espliceosoma presenta una plasticidad destacada en el reconocimiento de sustrato y puede incorporar diferentes partes de pre-ARNm en maduros ARNs bajo la influencia de un gran número de proteínas regulatorias, muchas de las cuales tienen capacidad de unirse al ARN. La mecánica está finamente regulada por interacciones entre elementos *cis* (secuencias), presentes en los exones e intrones y factores *trans* (proteínas) que reconocen estos elementos y que se constituyen como los elementos regulatorios principales.

Elementos *cis*

Son secuencias específicas en las regiones frontera exón-intrón y dentro de los mismos, presentes en el pre-ARNm, esenciales para el *splicing* constitutivo y alternativo. Se divididen en 2 clases:

- **Señales de** *splicing*: sitios reactivos del ARN que comprenden el sitio dador 5', el sitio aceptor 3' precedido por la zona de ramificación (*branch point*) y el tracto de polipirimidinas, cada uno de los cuales es estrictamente requerido por el espliceosoma para el reconocimiento del sustrato y la catálisis. Los pre-ARNm de metazoos poseen dos tipos de intrones que se distinguen por su sitio de *splicing* característico y su excisión por complejos de espliceosoma diferentes. La mayoría de los intrones de metazoos son del tipo U2, que contienen las secuencias canónicas GT-AG en las posiciones de la juntura y son procesados por el espliceosoma mayor. El segundo tipo de intrones se denomina U12, y es catalizado por la maquinaria menor y contienen la secuencia no canónica AT-AC. Estos últimos, no son muy frecuentes. Se han descripto unos 800 en mamíferos (humano y ratón) y unos 300 en plantas [10]. Normalmente, no encontramos más de un intrón de este tipo en los transcriptos y son removidos en su mayoría de modo co-transcripcional con una dinámica más lenta [9].

- **Señales regulatorias** (del inglés *Splicing Regulatory Elements* ***(SREs)***). Son motivos[1] presentes en el pre-ARNm que tienen funciones regulatorias en la selección del sitio de *splicing*. Normalmente son sitios blancos de las proteínas que se unen al ARN.

  Se clasifican en:

  - **ESEs** (*exónic splicing enhancers*): localizados en exones. Favorecen el *splicing* y la remoción del intrón.
  - **ESSs** (*exónic splicing silencers*): localizados en exones. Inhiben el *splicing*.
  - **ISEs** (*intronic splicing enhancers*): localizados en intrones. Favorecen el *splicing*, o sea la remoción del intrón.
  - **ISSs** (*intronic splicing silencers*): localizados en intrones. Inhiben su remoción.

[1] Un motivo es un elemento conservado en la secuencia de aminoácidos o nucleótidos, que habitualmente se asocia con una función concreta. Los motivos se generan a partir de alineamientos múltiples de secuencias con elementos funcionales o estructurales conocidos [11]

En un principio, las secuencias regulatorias podrían estar en cualquier ubicación en el pre-ARNm pero la mayoría de los estudios se han concentrado en los 200-300 nucleótidos adyacentes a los sitios de *splicing* observados, que parecerían contener mayor información en las secuencias [12]. Durante las últimas décadas, la bioinformática se ha convertido en una herramienta fundamental para el estudio del *splicing* alternativo y las señales regulatorias. Los abordajes desde el lado de la computación se han vuelto poderosos y sofisticados con cada genoma que se ha secuenciado completamente y con cada avance asociado a las tecnologías para los análisis genómicos. Los métodos de análisis de enriquecimiento de motivos utilizados para la identificación de sitios de pegados de factores de transcripción pueden ser usados en teoría para analizar posibles sitios de regulación de *splicing*. A diferencia de los sitios de pegado de los factores de transcripción, los ***SREs*** suelen ser más cortos y degenerados y poseen menor contenido de información. Presentan desafíos adicionales en cuanto a que múltiples copias de un mismo sitio, incrementan las capacidades regulatorias [13–16]. Uno de los objetivos subyacentes a cada experimento que se realiza es detectar este tipo de elementos regulatorios y la manera en que actúan en conjunto para lograr responder de manera coordinada [4].

Elementos *trans*

Son proteínas que interactúan con los elementos regulatorios en *cis*. Los reguladores de *splicing* más estudiados son los miembros de la familia de proteinas **SR** y **hnRNP** (Ver Cuadro 1.2).

La familia de proteínas **SR**, llamadas así por la presencia de repeticiones de Serinas (S) y Argininas (R) en sus secuencias, son proteínas de unión al ARN que suelen ser **activadoras** del *splicing* o también llamadas **reguladoras positivas**, uniéndose a secuencias correspondientes a exones y reclutando componentes centrales de la maquinaria de *splicing*, como el *snRNP* U1 al sitio 5' y el factor auxiliar U2 (U2AF) al sitio 3' a través de interacciones proteína-proteína. Se han descripto como proteínas de alta afinidad de unión a secuencias exónicas ricas en purinas [17–19]. La familia de proteínas **hnRNP**, llamadas así por ser ribonucleoproteínas heterogéneas nucleares (del inglés *heterogeneous nuclear ribonucleoproteins)*, también son proteínas de unión al ARN y se describen como **represoras** o **reguladoras negativas** del *splicing*, dado que se unen a regiones correspondientes a exones o intones interfiriendo con la capacidad de la maquinaria de pegarse a

los sitios de *splicing*. Los mecanismos por los cuales las proteínas **hnRNP** inhiben el *splicing* están poco estudiados.

Clase	Función	Ejemplos
Familia de Proteínas SR	Típicamente activadoras, por reclutamiento de los componentes del espliceosoma	nSR100 (SRRM4), SC35, SF2 (ASF), SRM160 (SRRM1), SRp30c, SRp38, SRp40, SRp55, SRp75, TRA2alpha, TRA2beta
hnRNPs	Típicamente represoras, por mecanismos pobremente estudiados	hnRNP A1, hnRNP A2/B1, hnRNP C, hnRNP F, hnRNP G (RBMX), hnRNP H, hnRNP L, nPTB (PTBP2), PTB (PTB1)
Otras	Activadoras y represoras	CELF4 (BRUNOL4), CUGBP, ESRP1, ESRP2, FOX1 (A2BP1), FOX2 (A2BP2), HuD, MBNL1, NOVA1, NOVA2, PSF (SPFQ), quaking, SAM68 (KHDRBS1), SLM2 (KHDRBS3), SPF45 (RBM17), TIA1, TIAR(TIAL1)

Cuadro 1.2: Tipos y ejemplos de proteínas regulatorias. Entre paréntesis los nombres sinónimos. Adaptado de [4]

Es interesante mencionar que muchos de los estudios recientes sobre los efectos del ambiente sobre en el *splicing* alternativo han reparado que este tipo de proteínas suelen ser blancos de regulación. En el capítulo 4 se detalla un trabajo donde se analiza el efecto de la luz sobre el *splicing*, donde varias de los ARNm de los genes que codifican para proteínas regulatorias del tipo SR y hnRNPs han sufrido cambios en el *splicing* alternativo.

Los mecanismos bioquímicos que controlan el uso de sitios de *splicing* y en consecuencia el *splicing* alternativo son complejos y en gran parte desconocidos. Es claro que no pueden existir factores específicos de *splicing* para los más de 100000 eventos que se han reportado en células humanas. No es sorpendente que sea un número bajo de proteínas las que se hayan encontrado responsables al menos en parte de la regulación de un gran número de eventos de *splicing* alternativo [4].

Con respecto a las regiones dadoras y aceptoras de *splicing*, se han reportado grandes diferencias

en cuanto a la conservación de las bases de las regiones flanqueantes a los exones alternativos y constitutivos. Los exones alternativos tienen mucha más conservación de bases en los intrones flanqueantes comparado a los constitutivos, que proveen buenos candidatos para la búsqueda de secuencias regulatorias. Las secuencias de los exones también están involucradas en la regulación. Los estudios de genómica comparativa de los exones son más complicados, dado que hay mayor presión de selección sobre las regiones codificantes a fin de conservar la secuencia de la proteína. Se ha demostrado que la tasa de evolución es menor en las regiones cercanas a la juntura exón-intrón que en el medio del exón, por estimación de las sustituciones sinónimas y no sinómimas realizadas a partir de alineamientos de secuencias de humano y ratón [20]. La densidad de polimorfismos de tipo "nucleótido simple "(del inglés *single nucleotide polimorphism, SNP*) también es menor cerca del sitio de *splicing*, lo cual también indica que son regiones con alta presión de selección [21].

Los mecanismos de *splicing* han sido elucidados principalmente por ensayos *in vitro* y estudios genéticos en mamíferos y levaduras, y en mucha menor medida en plantas. El advenimiento de la era genómica permitió la identificación de ortólogos de proteínas y pequeños ARNs del núcleo del espliceosoma, sugiriendo que los principios de procesamiento de intrones en mamíferos son también aplicables a las plantas. Sin embargo, el hecho que los intrones de animales no puedan ser procesados en plantas, conduce a pensar que hay alguna especificidad en la maquinaria de *splicing* y en las secuencias intrónicas de las plantas [9, 12]. Por ejemplo, se ha reportado que hay una clara diferencia en la longitud de los intrones de animales y plantas. Mientras que los primeros son pequeños, con una longitud promedio de unos 160 pares de bases, en animales encontramos intrones con una longitud media de 5 kilopares de bases, sugiriendo que el reconocimiento de intrones y exones puede diferir entre ambos. En cambio, se ha encontrado que las secuencias consenso dadora (5'), aceptora (3') y de la rama de bifurcación son comparables, sugiriendo que los componentes centrales de la maquinaria actuarían de manera similar. Tanto en plantas como en animales, la presencia de secuencias ricas en uridinas hacia el 3', parecería tener un rol importante sobre la eficiencia del *splicing*, y existe una diferencia clara entre el contenido de adenina-uridina (AU) y guanina-citosina (GC) entre exones e intrones. Estas diferencias podrían estar relacionadas al posicionamiento de los nuclesomas sobre exones e intrones, lo cual podría influenciar sobre la velocidad de elongación de la ARN polimerasa II, afectando tambien el *splicing* [22].

Además de los elementos regulatorios descriptos, existen cada vez más evidencias de elementos

regulatorios comunes a eventos de *splicing* alternativos tales como por ejemplo: la longitud de los intrones y exones implicados, el estado de compactación de la cromatina, la comunicación con la maquinaria de procesamiento de microARNs, el contenido GC de las secuencia y otros. [9, 23].

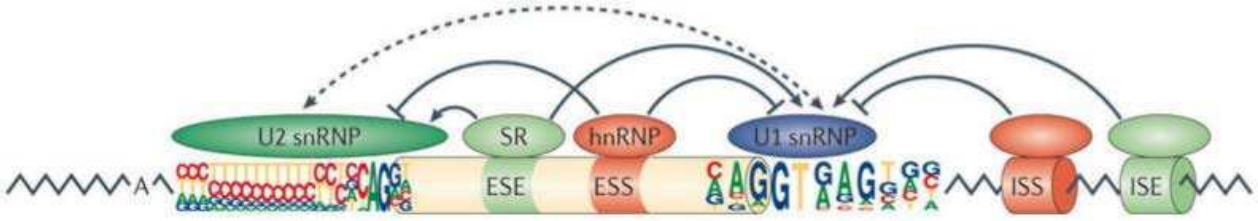


Figura 1.3: Regulación del *splicing* alternativo. La elección del sitio de *splicing* está regulada a través de elementos que actúan en *cis*, denominados SREs (ver texto principal) y de factores que actúan en *trans*. Basándonos en su ubicación relativa y sus actividades, los elementos reguladores pueden ser clasificados como activadores e inhibidores *(ESEs, ISEs, ESSs or ISSs)*. Estos SREs reclutan específicamente los factores de *splicing* para promover o inhibir el reconocimiento de los sitios de reacción. Los factores de *splicing* comunes incluyen a las proteínas SR que reconocen los *ESEs* y promueven el *splicing*, así como varios hnRNPs que típicamente reconocen *ESSs* para inhibir el splicing. Ambos elementos, suelen afectar la función de los snRNPs U1 y U2 durante el proceso de ensamblado de espliceosoma. Adapatado de [5].

La regulación en este contexto, significa que encontraremos diferentes patrones de *splicing* en diferentes ambientes celulares, condiciones, tratamientos, etcétera, abonando la idea que el *splicing* alternativo se constituye como una capa regulatoria cuya importancia es cada vez más reconocida. Es interesante notar que muchos genes que se reportan regulados por *splicing* alternativo suelen ser diferentes de aquellos que responden a nivel de expresión génica, implicando diferentes procesos biológicos y vías regulatorias, complementando de algún modo con la regulación por expresión diferencial [24–26]. En el trabajo que se describe en el capítulo 4 daremos cuenta también de las diferencias encontradas en los procesos biológicos asociados a los genes que respondieron a la luz por expresión y por *splicing*.

Cuando se aborda el tema de la regulación es importante entender cómo se activan y regulan los factores de *splicing*. Tanto en plantas como en animales, muchos de los factores de *splicing* junto con proteínas de unión a ARN, sufren cambios en sus patrones de *splicing* como respuesta a señales, que provoca que se autorregulen [27–29]. Además estas proteínas pueden ser reguladas post-traduccionalmente como respuesta a cambios en su ambiente [30]. Cambios sutiles en la concentración de los factores de *splicing* pueden modificar ligera o pronunciadamente los patrones de *splicing* de sus genes blancos. Hay algunas evidencias que sugieren que en animales y plantas la regulación ocurriría de diferente manera.

1.3. Consecuencias del *splicing* alternativo

El *splicing* alternativo tiene consecuencais importantes en la célula, principalmente a nivel de ARN o proteína. Una de las implicancias que más se ha descripto tiene que ver con la regulación de los niveles de transcripto a partir de la inclusión de codones de terminación prematuros (sin sentido) que derivarán a la degradación de los mismos por mecanismos de degradación mediado por codones sin sentido (del inglés *non sense mediated decay, NMD*). Sin embargo, esto no significa que todos los transcriptos que sufren retención de intrón y poseen codones de terminación prematuros, sean blancos de este mecanismo de degradación [31].

Otra de las consecuencias más estudiadas es la diferencia en la proteína final resultante luego de la traducción, que puede presentar regiones alternativas que impactarán, por ejemplo, en su localización subcelular, estabilidad o función. Por ejemplo, se ha reportado que muchos exones que son alternativamente incluidos codifican para regiones intrínsecamente desordenadas que generalmente residen en la superficie de la proteína y participan en las interacciones proteína-proteína. Esta implicancia funcional podría reconfigurar las redes de interacción y las vías de señalización. Además, las proteínas que se originan por maduraciones alternativas, suelen ganar o perder sitios de fosforilación, impactando de esta manera en las vías de señalización de quinasas. Existen proteínas o polipéptidos que son truncadas como consecuencia del *splicing* alternativo y que pueden actuar como inhibidores dominantes negativos de su misma proteína (debido a una interacción improductiva o impedimento de la dimerización) y se han denominado micropéptidos o pequeños péptidos de interferencia. En menor medida se han reportado que los mecanismos de *splicing* tendrían impacto sobre la poliadenilación alternativa y la maquinaria de procesamiento de ARNs pequeños (miARNs) [9].

En humanos la importancia del *splicing* se manifiesta en enfermedades genéticas hereditarias causadas por defectos en el *splicing*, por ejemplo, por mutaciones en genes que codifican en factores de *splicing* o en secuencias consenso de sitios de *splicing*. En plantas, si bien el *splicing* no tiene el impacto de causar enfermedades como en humanos, las variaciones naturales son la base de la evolución y el mejoramiento de cultivos. El *splicing* alternativo comenzó a llamar la atencion de los investigadores en plantas dado que muchos procesos del desarrollo y la respuesta al ambiente parecerían estar regulados vía *splicing* alternativo. La primera evidencia de la importancia de la

regulación por *splicing* alternativo en el desarrollo en plantas fue la descripción de un patrón de expresión diferencial de los factores de *splicing* de proteínas ricas en serina (SR) expresadas en diferentes órganos y durante el desarrollo indicando regulación órgano específica del *splicing* alternativo. Se ha reportado que el *splicing* alternativo tiene un rol crítico durante la respuesta de las plantas al estrés abiótico, biótico, desarrollo, floración, reloj circadiano, luz y temperatura [9]. El estudio de estos mecanismos puede aportar conocimiento para el entendimiento de la relación genotipo-fenotipo en las plantas.

El *splicing* alternativo se presenta entonces como un proceso a partir del cual los pre-ARNm pueden ser procesados de forma diferente para producir isoformas de ARNm con diferente estabilidad o capacidad codificante, permitiendo un control cuantitativo de la producción de proteínas y de síntesis de proteínas relacionadas con funciones cualitativamente diferentes.

1.4. Tecnologías para estudiar *splicing* alternativo

Durante los años 90, momento en el cual se produce el auge de la biología molecular en el mundo, el principal método para anotar un transcriptoma requería del lento y costoso proceso de clonar bibliotecas de fragmentos de ADNc de usualmente 300-400 pares de bases (del inglés *expressed sequenced tags, ESTs*) seguido de secuenciación capilar. Usando los alineamientos de los *ESTs* y las secuencias genómicas que estaban disponibles, se podían localizar los exones e intrones. Debido a los altos costos y a la limitada cantidad de datos disponibles sólo se podía tener una ligera idea del transcriptoma bajo estudio. El análisis de estos datos requería herramientas computacionales sofisticadas, muchas de las cuales proveyeron las bases de los *softwares* que actualmente se utilizan. Aunque fue una estrategia muy utilizada para la anotación de eventos de *splicing*, estaba limitada para resolver la detección de junturas no canónicas, exones pequeños, errores de secuenciación, etcétera. Estas técnicas proveyeron una mirada cualitativa de la ocurrencia de *splicing* ya que daban cuenta de la ocurrencia del evento pero no brindaban información cuantitativa sobre la regulación.

Con el auge de los microarreglos, originalmente pensados para cuantificar la expresión génica, se propusieron estrategias alternativas al clonado de *ESTs*. Estos métodos basados en hibridaciones, típicamente implican incubar sondas de ADN con marcadores fluorescentes en placas con oligo-

nucleótidos[2] previamente fijados en una placa. Se diseñaron algunos microarreglos especializados tales como microarreglos con oligonucleoótidos que sólo hibridan a las junturas exón-exón (microarreglos de junturas) que sirven para detectar y cuantificar isoformas conocidas. Estas tecnologías aunque representaron una mejora por sobre el clonado y secuenciación de los *ESTs* son costosas y experimentalmente muy laboriosas, además que no permiten avanzar sobre el descubrimiento de isoformas.

Los grandes avances durante la última década en las tecnologías de secuenciación del ADN han hecho posible secuenciar ADNc a partir del ARN que se extrae de las células, denominado secuenciación masiva del ARN, abreviadamente RNA-Seq [32]. La misma puede utilizarse para definir un mapa preciso de todas las isoformas alternativas que se presentan en los diferentes tipos celulares así como analizar perturbaciones y diferentes estados transcripcionales. Dentro de sus características sobresalientes, podemos mencionar la resolución a nivel de bases y el rango dinámico en los niveles de expresión que puede trabajar, además del bajo costo de secuenciación por base y la baja complejidad en la preparación de las muestras para secuenciar. Esta tecnología permite secuenciar desde ambos extremos los fragmentos y también de una manera hebra específica. Un protocolo típico de secuenciacón por síntesis de ARN implica (ver Figura 1.4):

1. Extracción del ARN: poliA o total según el experimento que se quiera realizar
2. Síntesis del ADNc, fragmentación y ligación de adaptadores
3. Amplificación en la placa de secuenciación, formación de *clusters*
4. Secuenciación de los fragmentos amplificados
5. Análisis de los resultados

La información proveniente de un experimenrto de secuenciación es enorme y compleja. En los ultimos años se han desarrollado múltiples herramientas bioinformáticas para intentar responder a las diferentes preguntas biológicas que surgen a partir de un experimeto de RNA-Seq. Para poner en contexto el trabajo que hemos realizado durante la tesis doctoral, vamos a reseñar brevemente las

[2]Pequeños fragmentos de ADN simple hebra, complementarios a las regiones bajo estudio

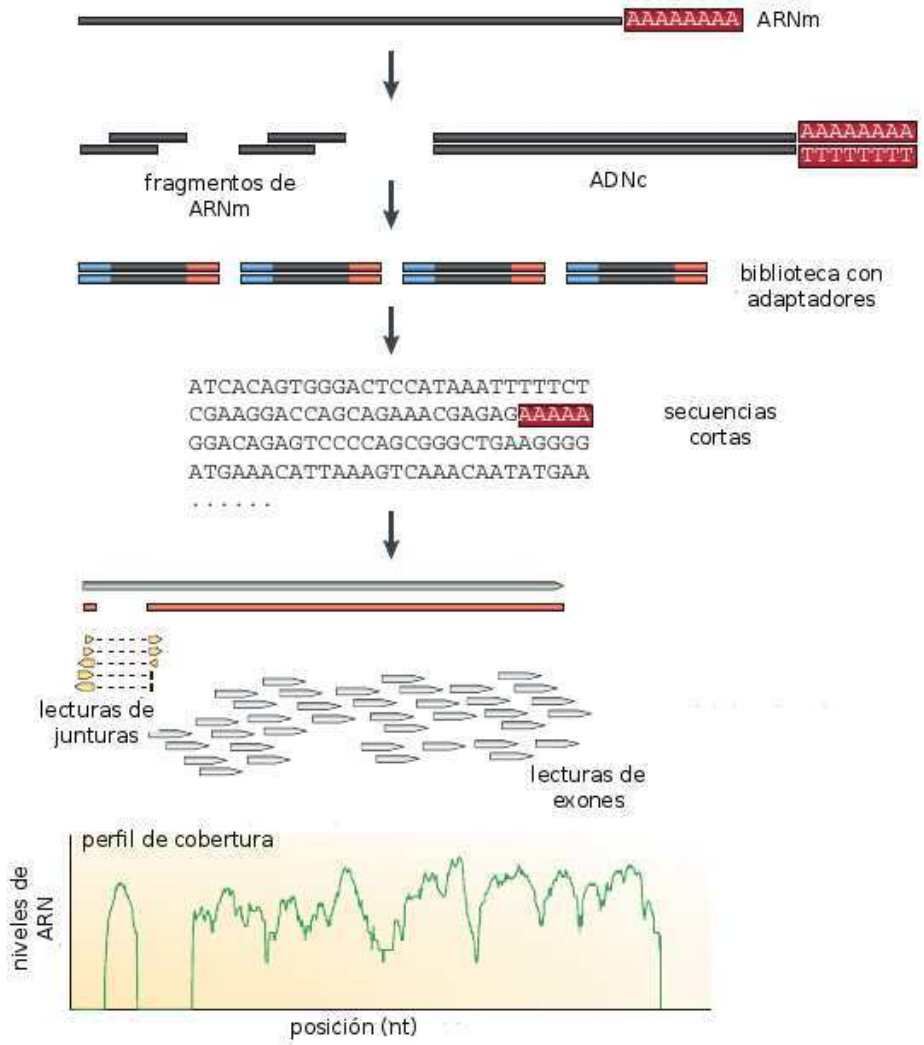


Figura 1.4: Protocolo de secuenciación masiva de ARNm. Las moléculas de ARNm son convertidas en una biblioteca de ADNc. En el momento de su fragmentación, se agregan los adaptadores (en azul y rojo). A partir de esta biblioteca, se obtienen millones de lecturas cortas usando la técnica de secuenciación, que son alineadas contra el genoma y/o transcriptoma de referencia. Las lecturas son clasificadas en exónicas o junturas (contienen *gaps*). A partir del alineamiento se contruye un perfil de cobertura para cada gen. Adaptado de [33]

diferentes preguntas y los procedimientos que se proponen una vez que se han obtenido las lecturas del secuenciador. De acuerdo a [32], los desafíos computacionales se presentan en 3 categorías que describiremos a continuación:

1. **Alineamiento de las lecturas al genoma o transcriptoma de referencia**. Una de las tareas esenciales para cualquier análisis de RNA-Seq consiste en asignar las lecturas obtenidas a su probable gen de origen. Los algoritmos para hacer los alineamientos se pueden clasificar según permiten o no alinear lecturas con la introducción de espacios vacíos o *gaps*. Es importante tener en cuenta esta distinción porque la elección del algoritmo dependerá en primer término de qué tipo de referencia tenemos disponible. Los algoritmos que permiten

la inserción de *gaps* en su alineamiento se podrán usar para alinear lecturas tanto contra un genoma como con un transcriptoma de referencia. En el caso de usar el genoma, aquellas lecturas que presentan *gaps*, serán precisamente las que provengan de regiones de junturas exón-exón, cuya importancia será remarcada a lo largo de esta tesis. Los algoritmos que no permiten *gaps* en cambio se podrán utilizar cuando tenemos como referencia secuencias de ADNc por ejemplo, pero estarán limitados en cuanto al descubrimiento de nuevos exones o junturas. En general, por cuestiones de eficiencia y tiempo, los métodos que permiten la inserción de *gaps* suelen realizar una primera fase de alineamiento utilizando un algoritmo que no permite *gaps* y luego, alinean las lecturas remanentes. Para detectar los nuevos sitios de *splicing*, la mayoría de los algoritmos que trabajan sin transcriptoma de referencia buscan la ocurrencia de los patrones de dinucleótidos canónicos para los sitios dador/aceptor: **GT-AG, GC-AG, AT-AC**. Además de la información de la búsqueda de los dinucleótidos, otras informaciones tales como las *islas de cobertura* que surgen a partir del alineamiento de las lecturas completas y la longitud mínima y máxima estimada de los intrones son de utilidad para estimar los sitios de las junturas. Los algoritmos que utilizan transcriptoma de referencia, directamente alinean contra las junturas conocidas pero son incapaces de descubrir la ocurrencia de nuevos eventos (Figura 1.5).

2. **Reconstrucción del transcriptoma**: Cuando no existe transcriptoma anotado, se intenta reconstruir las isoformas secuenciadas. Es una tarea computacional muy compleja, básicamente por las siguientes razones:

    a) La expresión de los genes ocurre en rangos de magnitudes muy variables, con algunos genes muy expresados y otros muy poco,

    b) Es posible que las lecturas se originen de transcriptos maduros e inmaduros a la vez, dependiendo del protocolo de preparación de las bibliotecas,

    c) Las lecturas son cortas y los genes pueden generar múltiples isoformas, haciendo difícil saber de que isoforma proviene cada lectura.

En este punto los métodos pueden clasificarse según utilicen o no genoma de referencia. La elección del algoritmo más apropiado dependerá de la pregunta biológica a responder y

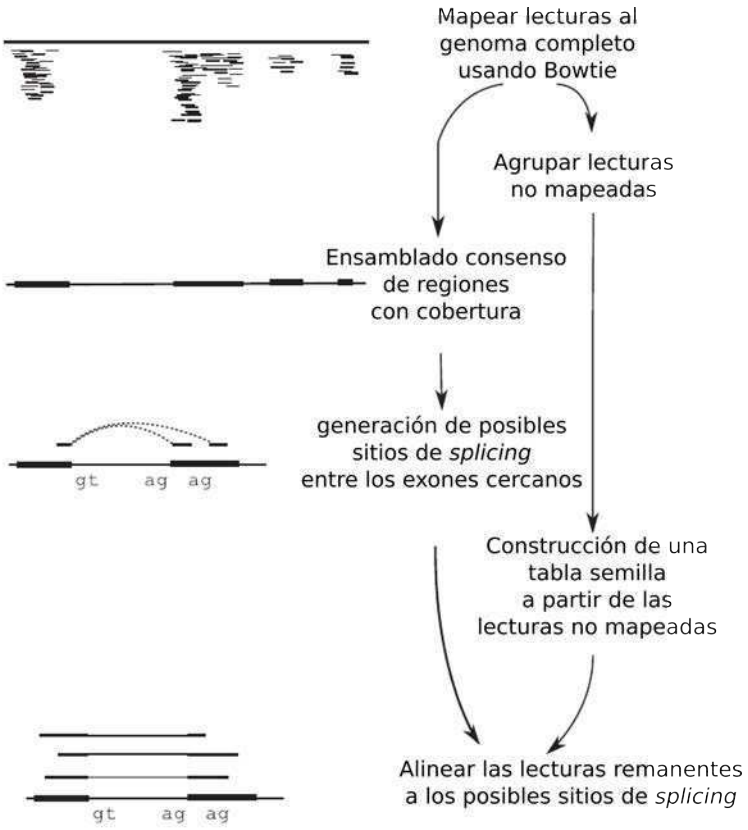


Figura 1.5: Esquema de cómo funciona un algoritmo de alineamiento de lecturas contra el genoma de referencia que permite inserción de *gaps* (TopHat). Las lecturas son alineadas sin *gaps*, usando Bowtie contra el genoma de referencia completo, separando aquellas que no se pueden alinear. A partir de las regiones que tienen lecturas se genera un consenso, sobre las cuales se intentará alinear las lecturas remanentes, ahora con inclusión de *gaps* que serán las que se identificarán como las junturas exón-exón. Adaptado de [34].

de las disponibilidades de genoma en la que nos encontremos. En el caso de no contar con genoma de referencia, no tendremos opción. En el caso de contar con el genoma, muchas veces se sugieren ensamblados mixtos, con y sin el genoma de referencia para capturar mayor información.

3. **Cuantificación de la expresión y el *splicing* alternativo**.

   - **Expresión génica**: Algunos métodos, estiman la expresión por gen, calculando la suma de la expresión de todas sus isoformas. Dado que calcular la expresión de las isoformas es computacionalmente muy costoso, podemos simplificar resumiendo la expresión a nivel de exones, usando un metódo de **unión** (todos los exones que están anotados para ese gen) o **intersección** (sólo los comunes a todas las isoformas, similar a la manera de cuantificar expresión en micro arreglos). En nuestro trabajo, la cuantificación de la

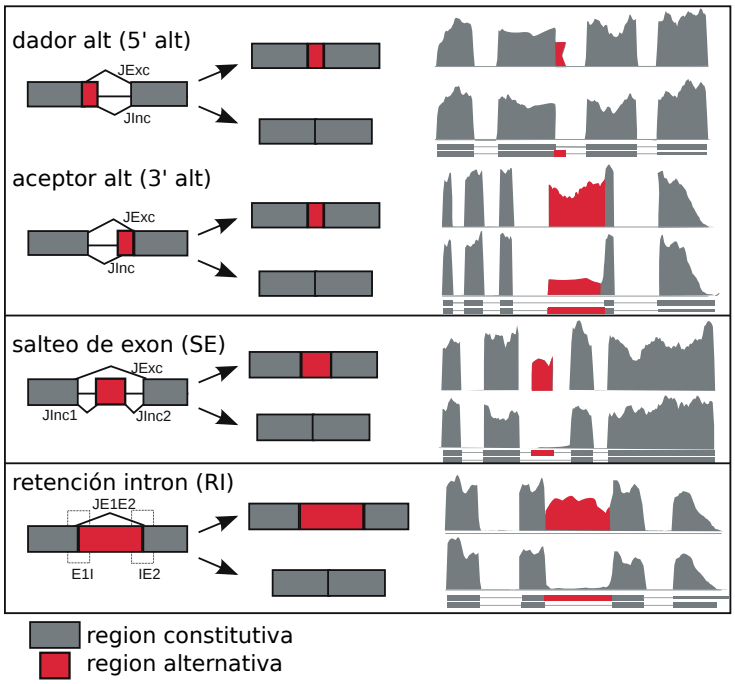


Figura 1.6: Ejemplos de cómo se visualiza el resultado del alineamiento con un navegador genómico, resaltando cómo se ven las regiones alternativas

expresión génica se realiza por el **método de unión**. Para estimar la expresión de un gen en una muestra hay que normalizar las lecturas obtenidas por la longitud del gen, dado que la fragmentación del ARN durante la construcción de la biblioteca provoca que los transcriptos más largos generen más lecturas comparado a los fragmentos más cortos. Si vamos a comparar la expresión de un gen entre dos o más condiciones, hay que normalizar debido a la variabilidad en el número de lecturas producida en cada biblioteca, que produce fluctuaciones en el número de fragmentos secuenciados en cada muestra. Los *softwares* que hacen estimación de las cantidades y de expresión diferencial, toman en cuenta estas fuentes de variablidad a la hora de cuantificar.

- *Splicing* **alternativo:** De acuerdo a [7], los métodos para detectar *splicing* alternativo pueden ser divididos en 2 esquemas de cuantificación (ver Figura 1.7):

  a) **Modelos basado en conteos**: están basados en métodos usados para cuantificar transcriptos con una sola isoforma. Transforman la pregunta sobre *splicing* alternativo en una pregunta sobre uso diferencial de unidades de conteo subgénicas. En esta estrategia, la estructura del gen suele quedar configurada por una representación sencilla de **unidades de conteo**. Estos modelos claramente no in-

tentan reconstruir las isoformas, aunque se ha postulado que podrían usarse como una buena aproximación ya que no existiría isoforma que resultara de la combinación de todas las unidades de conteo [35]. Estos modelos pueden verse como una evaluación entre dos posibles resultados de *splicing*: inclusión-exclusión de cada unidad de conteo y son altamente dependientes de la anotación existente. Algunos trabajos se han referido a estos modelos como **modelos basados en eventos (evento-céntricos)** [36].

b) **Modelos basados en resolución de isoformas**: Intentan estimar la ocurrencia de *splicing* alternativo a partir de la estimación de la abundancia de las diferentes isoformas en las diferentes muestras o condiciones. En general, estiman la proporción de cada isoforma q a partir de la maximización de una función de versosimilitud. Maximizar esta función es equivalente a maximizar la probabilidad de seleccionar una lectura o un fragmento a partir de cada transcripto. Estos modelos de resolución de isoformas intentan asignar las lecturas a los transcriptos que han dado origen introduciendo un alto grado de incertidumbre en la asignación debido a que las isoformas se superponen.

Como conclusión, podemos decir que los métodos que se utilicen son altamente dependientes del protocolo de secuenciacón que se haya utilizado y además son dependientes entre sí. Hasta la fecha se han publicado numerosas y diversas herramientas, la mayoría evaluada en humanos, ratones y otros mamíferos. En plantas todavía no existen herramientas establecidas y de referencia unánime [7].

Métricas para cuantificar eventos de *splicing* alternativo

Los primeros reportes que usaron RNA-Seq para cuantificar *splicing* alternativo, seguían un enfoque análogo al que se usaba en los microarreglos de junturas a partir de las anotaciones de los genes. En este caso, las lecturas que alinean al exón candidato se denominan lecturas de inclusión, mientras que las lecturas que mapean a los exones flanqueantes y a las junturas que lo saltean, se denominan lecturas de exclusión. Esta métrica mostró buena concordancia con la manera de evaluar los microarreglos de junturas. La métrica más utilizada actualmente se denomina **PSI**:

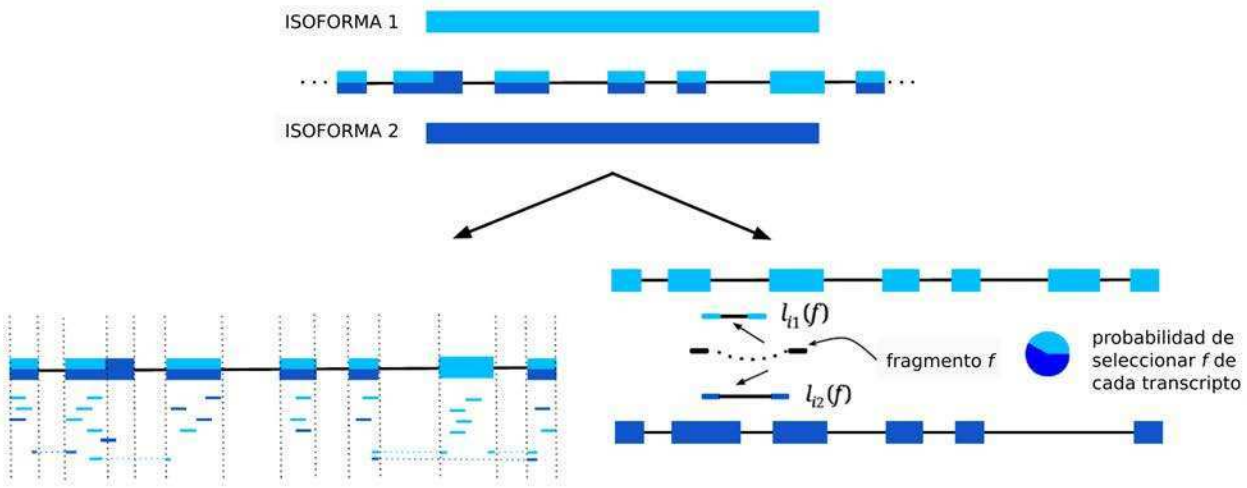


Figura 1.7: Esquemas de cuantificación. Un modelo simplificado de un gen con 2 isoformas. Los exones se colorean de acuerdo a la isoforma de origen. Se proponen 2 modelos de cuantificación. El modelo basado en conteos (a la izquierda), las lecturas se asignan a unidades de conteos (indicadas con líneas intermitentes) sin ambigüedad. Para cada unidad de conteo, el modelo puede ser entendido como 2 posibles situaciones, la de incluir-excluir una unidad de conteo. En el modelo de resolución de isoformas (a la derecha), las 2 lecturas apareadas (cajas negras conectadas por linea intermitente negra), alinean corriente arriba y corriente abajo de un sitio dador de *splicing*. Si la distribución de tamaños de fragmentos es conocida, es posible inferir de que isoforma hay más probabilidad de que provengan las lecturas apareadas. En este caso, la longitud efectiva del transcripto asi como otros parámetros (dependiendo del modelo que se use) podrían afectar la probabilidad de asignar las lecturas a las diferentes isoformas. Adaptado de [7].

(del inglés *percent spliced in*) y clásicamente refiere a la proporción de la isoforma que incluye el exón por sobre todas las isoformas presentes en esa muestra [37]. En el caso de la retención de intrón, se utiliza una métrica que se denomina **PIR** (del ingles *percent intron retention*) [37, 38]. Sin embargo, tal como se detalla en el capítulo 3 puede tener otras acepciones.

Los métodos para cuantificar los eventos en una sola condición reportan la ocurrencia del evento utilizando el PSI, o los conteos de la región normalizando 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))*. La cuantificación de isoformas puede ser expresada de un modo global, que en general se expresa como RPKM o FPKM y luego las proporciones entre ellas se expresan a nivel de PSI.

La comparación de eventos e isoformas a lo largo de dos o más condiciones provee información para el entendimiento de la regulación del *splicing* alternativo. Es importante distinguir entre la abundancia relativa entre isoformas de la expresión diferencial de isoformas. Cambios en la abundancia relativa de isoformas en lugar de cambios a nivel de expresión de la isoforma, nos hablarán de un mecanismo de probablemente relacionado al *splicing* y/o la **estabilidad** diferencial

de las mismas. Por otro lado puede haber cambios detectables en la expresión de las isoformas a largo de las condiciones que no representen alteraciones en las proporciones entre ellas y que nos hablarán de un mecanismo relacionado a la **regulación de la transcripción**.

La mayoría de los métodos que se enfocan en la cuantificación de eventos, en general utilizan un conjunto de eventos precalculados a partir de la anotación, particularmente útiles para estudiar organismos bien anotados. Para cuantificar los eventos se usan las lecturas que alinean en esa región o las métricas calculadas como PSI o PIR. Luego se estiman las abundancias relativas de los eventos en cada condición, y según el modelo estadístico de método, se le asigna un valor estadístico (p-valor) a la diferencia entre las condiciones. El p-valor puede ser reportado a nivel evento o a nivel gen, dependiendo la aproximación que se haya elegido.

La descripción de *splicing* alternativo en términos de eventos facilita la validación experimental, su caracterización en términos mecanísticos y también está motivada por las limitaciones actuales de la reconstrucción de isoformas a partir de lecturas cortas, normalmente entre 100-300 pares de bases. En general, estos métodos que reportan cambios en *splicing* alternativo a nivel evento suelen presentar una tasa de validación más alta [39]. Los métodos que trabajan con cuantificación a nivel isoformas, suelen reportar los cambios en los niveles de isoformas entre muestras pero no en sus proporciones. En el cuadro 1.3, se resumen las herramientas bioinformáticas disponibles más populares para cuantificar *splicing* alternativo, de acuerdo a criterios presentados en trabajos recientes del área [36, 40, 41].

Los últimos avances van en el sentido de la secuenciación de moléculas completas sin el proceso de fragmentación y de amplificación clonal y de la secuenciación de celulas individuales [42]. A medida que avanza la tecnología también crece la demanda de nuevos desarrollos bioinformáticos que acompañen la interpretación de los datos que las mismas producen.

Nombre	Modelo	Finalidad
Cuffdiff 2 [43]	Isoformas	Abundancia relativa de las isoformas. Comparaciones complejas, descubrimiento de isoformas, en RPKM
MISO [44]	Isoformas	Abundancia relativa de las isoformas. Sólo para comparaciones de a pares, sólo eventos anotaods
DEXSeq [45]	exón	Inclusión/Exlusión exones anotados. Usa el número de lectuas por unidad de conteo. Comparación de a pares. Puede predecir nuevos SE a partir de los exones constitutivos
MATS [46]	Junturas + exón	Reporta PSI y p-valor por exón, diseños complejos. Extrae los eventos alternativos de la anotación. No detecta nuevos eventos ni complejos
DiffSplice [47]	Junturas + exón	Reporta Módulos de Splicing Alternativo (ASM) donde hay cambios entre los transcriptos

Cuadro 1.3: Ejemplos de herramientas bioinformáticas disponibles para cuantificar *splicing* alternativo

Capítulo 2

Objetivos

El objetivo general del doctorado está orientado al desarrollo de una herramienta bioinformática que permita el descubrimiento, caracterización, cuantificación y evaluación de eventos de *splicing* alternativo de una manera global, a partir de los resultados de la secuenciación de alto rendimiento del ARNm (RNA-Seq).

Objetivos específicos

- Diseñar un protocolo de trabajo bioinformático para identificar qué eventos de *splicing* alternativo se producen durante la maduración del ARNm.
- Adaptar el protocolo a diseños experimentales complejos, tales como análisis de experimentos en el tiempo y la interacción de genotipos y tratamientos.
- Poner a disposición de la comunidad científica el protocolo de una manera ordenada y fácil de utilizar.
- Aplicar el protocolo para estudiar mecanismos de regulación del proceso de *splicing* alternativo, caracterizando el transcriptoma de organismos con niveles alterados de factores de *splicing*.
- Aplicar el protocolo en experimentos de RNA-Seq hechos en la planta modelo *Arabidopsis thaliana* para identificar qué relación existe entre los eventos de *splicing* alternativo y la regulación de procesos fisiológicos puntuales, como las respuestas a la luz y a la temperatura.

Capítulo 3

Diseño de un protocolo de trabajo bioinformático para identificar eventos de *splicing* alternativo

3.1. Introducción

El advenimineto de las tecnologías de secuenciación masiva, dió lugar al diseño de estrategias para el estudio del transcriptoma a un nivel inimaginable. La secuenciación masiva del ARN (RNA-Seq), ya sea total o poliadenilado, permite recrear un mapa de todas las isoformas que están en una muestra en un dado momento y su cuantificación. Sin embargo, las herramientas bioinformáticas que acompañan estos increíbles desarrollos de nanotecnologías no siempre son capaces de resolver todas las preguntas que surgen a partir de estos estudios. En el caso del *splicing* alternativo, los interrogantes más importantes, refieren a la descripción de las isoformas que están presentes en una muestra así como su cambio cuantitativo a lo largo de diferentes condiciones, tal como se describe en el capítulo 1.

El origen del trabajo que describiremos a lo largo del capítulo y que es la columna vertebral de esta tesis, se remonta al año 2012 cuando comencé el trabajo de doctorado. Para ese entonces, el laboratorio del Dr. Yanovsky disponía de múltiples experimentos de RNA-Seq de la planta modelo *Arabidopsis thaliana*, uno de los cuales, había sido hecha sobre una mutante de una proteína arginin-metil transferasa (Ver Capítulo 5). Utilizando tecnologías anteriores a la secuenciación masiva,

se había descubierto que el principal efecto a nivel de *splicing* que se observaba en las plantas mutantes, era la incapacidad de remover eficientemente intrones cuya secuencia nucleotídica en el sitio dador (5') se alejaba del consenso de los sitios considerados como fuertes [48]. De esta manera, se inició la búsqueda de herramientas bioinformáticas dispobibles hasta ese momento con el objetivo puesto en encontrar alguna que rápidamente permita analizar la retención de intrones globalmente en datos de RNA-Seq. Este tipo de evento, había sido reportado de ocurrir de forma frecuente en plantas y no así en mamíferos (ver Tabla 1.1), para los cuales suelen desarrollarse en primer lugar la mayoría de las herramientas bioinformáticas. Dado que no encontramos herramientas adecuadas ya desarrolladas para describir este tipo de resultados, comenzamos a desarrollar nuestro propio protocolo de análisis. En un principio, el foco estuvo puesto en cuantificar las lecturas que alineaban a las regiones de los intrones y de alguna manera, descubrir aquellos que cambiaban su nivel de retención a lo largo de las condiciones. Luego, a medida que fuimos resolviendo desafíos, incorporamos cada vez más detalles a nuestros resultados y el código se volvió sofisticado y extenso, razón por la cual surgió la necesidad de convertirlo en un paquete con sus respectivos manuales de ayuda que lo acompaña, para poder compartirlo con colegas y que sea simple de usar.

Tal como se describirá a lo largo del capítulo y en los capítulos posteriores, el desarrollo del protocolo nos habilitó a extraer enorme cantidad de información de cada experimento de RNA-Seq que llegó a nuestras manos y que nos ha permitido incrementar nuestro conocimiento, no sólo de los mecanismos de regulación del *splicing* alternativo sino también de genómica y bioinformática en general.

3.2. Materiales y métodos

Alineamientos de lecturas contra el genoma de referencia: archivos BAM

Las lecturas resultantes de cada experimento de secuenciación de ARN deben ser alineadas contra el genoma de referencia con un algoritmo que permite la inserción de *gaps* (ver Capítulo 1, Figura 1.5). El resultado del alineamiento de cada muestra contra el genoma de referencia se guarda en archivos de formato **BAM** (del inglés *Binary of sequence Alignment Map*). Estos archivos contienen información de las posiciones en el genoma de referencia donde han sido alineadas

las lecturas y el modo en que han sido alineadas, por ejemplo, si tienen algún *gap* o diferencia (*missmatch*) con respecto a la referencia.

Anotación de genomas

La anotación del genoma se provee en formato **GFF3 o GTF** (del inglés *General Feature Format*). Este formato, muy popular en bioinformática, fue establecido por el consorcio GMOD [49] y contiene la informacion genómica tabulada en 9 campos obligatorios para cada una de las regiones genómicas que queremos describir. Entre ellos:

- **nombre de la secuencia sobre la que se establecen las posiciones**. Ejemplo: cromosoma, cloroplasto, mitocondria, *contig, scaffold*.
- **Fuente de la información**. Ejemplo: TAIR10, Ensembl, NCBI
- **Tipo de región a describir**. Ejemplo: gen, exon, intrón.
- **Posición inicio**. (con respecto a la primer columna).
- **Posición de finalización**. (con respecto a la primer columna).
- *Score*. Si hubiera sigo asignada la anotación por algún algoritmo con puntajes.
- **hebra** Definida como + (directa) o - (reversa).
- **marco de lectura** Puede ser '0', '1' o '2'.
- **Atributos**: Es un campo opcional. Se puede dejar vacío o agregar descripciones separadas por **;**.

Lenguaje de programación R

El protocolo desarrollado fue escrito en el lenguaje de programación **R**. El mismo, es un lenguaje que originalmente se utilizó en el área de estadística y matemática, por lo que ha heredado mucha popularidad en el ámbito científico. Es un lenguaje gratuito, multiplataforma (puede usarse en los 3 sistemas operativos más populares Windows, Linux, MacOS) y es de código abierto (el código de las funciones está disponible para usarse y editarse). La instalación del lenguaje se hace en un modo *base* y luego se va aumentando su capacidad con la instalación de paquetes, que son **conjuntos de funciones (pequeños algoritmos) y documentación que están compilados y son fáciles de compartir**. En general, los paquetes en R se escriben utilizando un paradigma de programación que

se denomina orientada a objetos. Los objetos tienen una clase definida y métodos específicos que se aplican sobre ellos. En este paradigma, los objetos son variables donde se alojan los resultados de un método (funciones). Los paquetes suelen hacerse públicos en **repositorios especializados** donde se favorece la visibilidad y la calidad de los mismos. En el caso de paquetes genéricos de R, los paquetes de extensiones y las diferentes distribuciones del lenguaje se encuentran en el sitio del proyecto R [50] y de la red de repositorios CRAN [51].

En el caso de los paquetes orientados a problemas biológicos esencialmente de genómica, en el año 2001 se creó un consorcio denominado **Bioconductor** [52,53]. El proyecto comenzó albergando paquetes de estadística y análisis de datos relacionados a la tecnología de microarreglos, pero a medida que avanzaron las tecnologías de secuenciación y los genomas secuenciados, comenzó a crecer el número de paquetes dedicados a este tipo de experimentos. Actualmente en el repositorio hay 1288 paquetes de *software*, 934 paquetes de anotación y 304 de experimentos [54].

3.3. Resultados

De acuerdo a los objetivos planteados, nuestro mayor desafío fue diseñar una estrategia bioinformática para caracterizar el descubrimiento, caracterización y cuantificación de eventos de *splicing* alternativo de una manera global. Para alcanzar este objetivo, uno de los resultados más destacados del trabajo de tesis y donde más tiempo hemos invertido es en el desarrollo de un paquete de **R/Bioconductor** al que denominamos **ASpli**. En el mismo, hemos materializado el protocolo de trabajo desarrollado junto con toda la experiencia adquirida en el área durante el transcurso del doctoroado.

ASpli es un paquete modular, escrito dentro del paradigma de programación orientada a objetos y está disponible a la comunidad de forma gratuita desde julio de 2016 en el sitio de Bioconductor [54].

A lo largo de esta sección se describirán sus principales características, subrayando fortalezas, dificultades y ejemplos de su utilización para estudiar el *splicing* alternativo a partir de experimentos de RNA-Seq.

Módulos de ASpli

ASpli está dividido en 4 módulos que pueden ser utilizados independientemente y cuya funcionalidad será descripta en detalle a lo largo de este capítulo (Figura 3.1). Brevemente, cada módulo consiste en:

- Extracción de coordenadas genómicas (método `binGenome()`)
- Superposición de coordenadas genómicas y lecturas que fueron alineadas (método `readCounts()`)
- Análisis de *splicing* alternativo usando junturas (método `AsDiscover()`)
- Estimación de expresión diferencial de genes y de uso diferencial de *bins* (método `DUreport()`)

En cada módulo es posible exportar los resultados a archivos en formato tabular para los subsecuentes análisis.

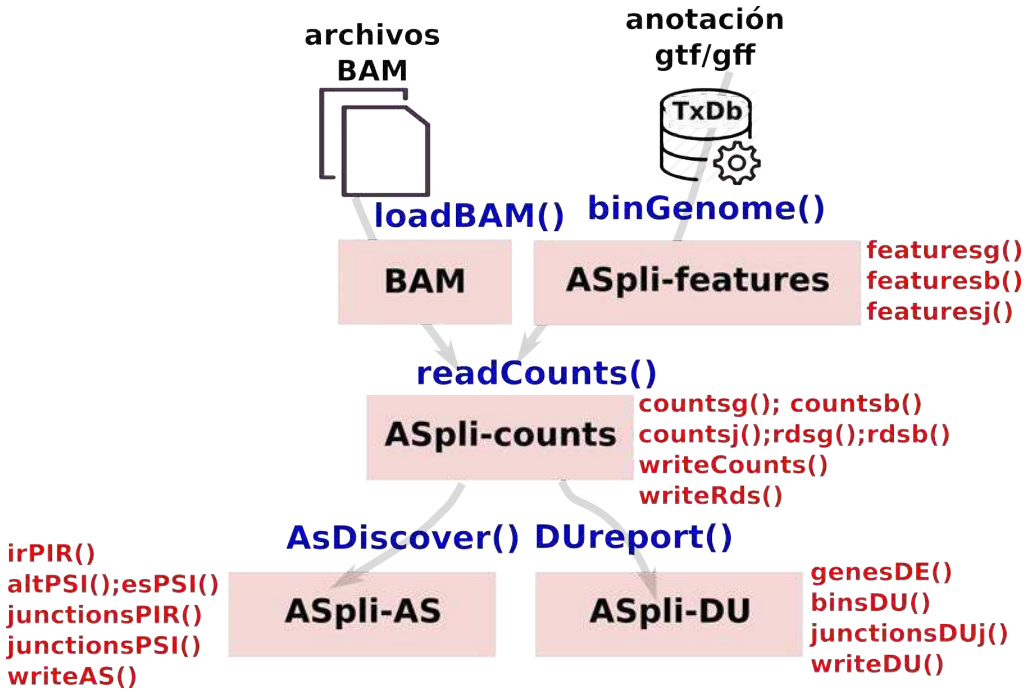


Figura 3.1: Estructura de **ASpli**. Usando la anotación (en formato GTF/GFF) y la información de las lecturas alineadas contenida en los archivos BAMs podemos llevar a cabo el análisis del experimento de RNA-Seq usando **ASpli**. Los objetos que resultan del uso de cada método (en azul), se representan en cuadrados rosa. Los métodos para exportar resultados se detallan en rojo.

3.3.1. Disección de la anotación del transcriptoma en *bins*. Método: `binGenome()`.

Uno de los primeros desafiós que se nos presentó al momento de trabajar en *splicing* alternativo en plantas, tal como se menciona en la introducción, es que la mayoría de las herramientas que

están disponibles están orientadas al evento de salteo de exón y en el caso de descubrimiento de nuevos eventos, también lo reportaban sobre exones anotados (Ver Tabla 1.3). El primer módulo de **ASpli** refleja exactamente ese resultado: el desarrollo de un método de extracción de coordenadas genómicas que permite extraer a partir de la anotación del transcriptoma, las coordenadas de los genes asi como también de todos los elementos subgénicos: exones, intrones y junturas.

Para utiilizar este módulo necesitamos contar con la anotación del organismo en estudio en formato **GTF/GFF**, que debe ser convertida en un objeto de clase **TxDb**. De este modo, la anotación queda convertida en una base de datos de tipo relacional que permite extraer información de la anotación de una manera muy eficiente. Este paso es obligaotrio y debe hacerse usando un método propio de otro paquete de R/Bioconductor denominado `GenomicFeatures` [55].

La regiones subgénicas tales como exones e intrones se analizan usando las anotaciones de los genes, utilizando la idea que introdujo el trabajo de [45], donde se proponía desarmar la anotación del transcriptoma en nuevas unidades subgénicas no superponibles llamadas *bins*. Esta estrategia, colapsa todas las isoformas de un dado gen en una estructura final, que resulta de la proyección de los extremos de todos los exones de un gen sobre una línea horizontal imaginaria. De esta manera, en algunos casos se conservan las coordenadas de los exones e intrones originales, pero en otros, cuando en la región se superponen diferentes exones, se originan nuevos rangos.

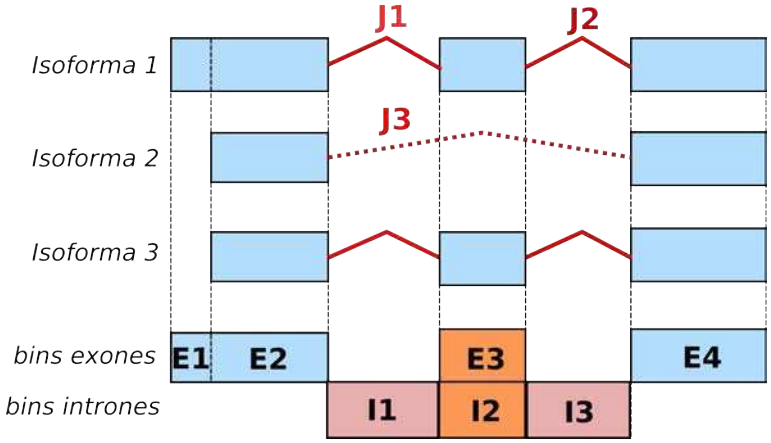


Figura 3.2: Disección de la anotación del transcritpoma. Esquema de los *bins* resultantes para un dado gen con 3 isoformas hipotéticas. En cuadros celestes, los *bins* exónicos, en cuadros rosa los intrónicos. Aquellos bins que son exónicos e intrónicos a la vez se denominarán *bins* alternativos (en naranja). En líneas rojas comunicando los diferentes exones se representa el repertorio de junturas J1,J2,J3.

En **ASpli**, siguiendo esta idea, extraemos las coordenadas de los exones e intrones de todos los **genes multiexónicos**. Si existe más de una isoforma en un dado gen, habrá exones que se superpongan y por el procedimiento explicado antes, al realizar la proyección sobre una línea imaginaria,

algunos exones e intrones se superpondrán. Cuando realizamos esta disección, obtendremos como resultado un conjunto de *bins* para cada gen. Los mismos serán clasificados en exclusivamente exónicos, exclusivamente intrónicos o alternativos, que serán los de nuestro principal interés (en la Figura 3.2, son aquellos que están coloreados en naranja). Estos *bins* alternativos se clasificarán de acuerdo al evento alternativo que asumimos le ha dado origen:

- **SE**: Salteo de exón
- **RI**: Retención de intŕon
- **5'Alt**: Uso de dador 5' alternativo
- **3'Alt**: Uso de aceptor 3' alternativo
- "*"(SE*, RI*, 5'/3'Alt*) El uso de ***** significa que esta región está involucrada en más de un evento posible. La clasificación putativa resulta de la comparación con los intrones originales.
- **external**: Externos. Son los primeros y últimos *bins* de los transcriptos. Se puede obviar analizarlos si sólo quisiéramos analizar *splicing* alternativo. Su uso diferencial estaría asociado a inicios y/o terminación alternativa de la transcripción.

Reglas de decisión para asignar el evento a un *bin* alternativo, en la posición **i**:

bin alternativo[i]	*bin* anterior[i-1]	*bin* posterior[i+1]	comentarios
SE	intrón	intrón	
5'Alt	exón	intrón	tiene en cuenta la hebra
3'Alt	intrón	exón	tiene en cuenta la hebra
RI	exón	exón	
external			coinciden con el comienzo o fin de alguno de los transcriptos
múltiple (*)	alt	alt	antes y después del *bin* en estudio hay otro *bin* alternativo

Cuadro 3.1: Reglas de clasificación de los *bins* resultantes luego de la partición de la anotación del transcriptoma. En primer lugar se identifican los *bins* alternativos. Luego se analiza la clasificación de los *bins* en la posición inmediata anterior y posterior. De esta manera, deducimos el evento putativo de *splicing* alternativo que le dio origen. Los *bins* externos son aquellos que coinciden con el inicio o fin de cada transcripto anotado para ese gen.

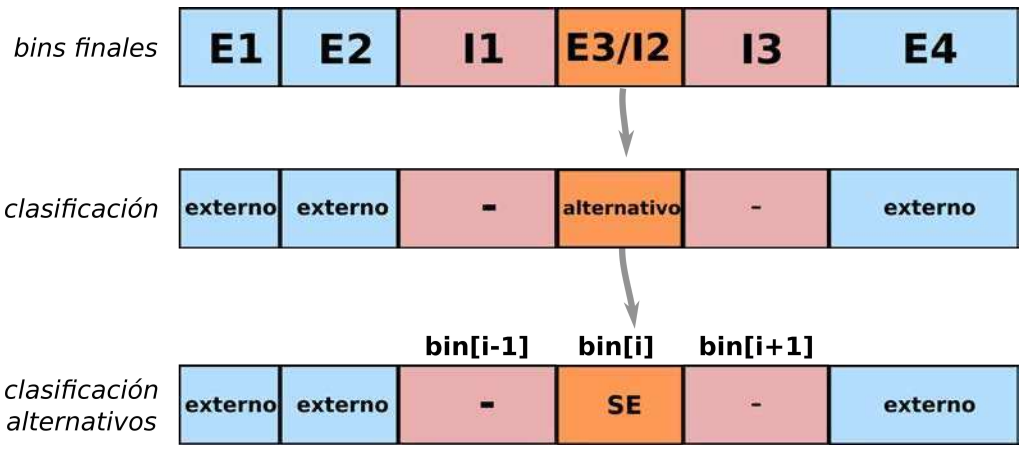


Figura 3.3: Esquema simplificado de la clasificación de los *bins*. Siguiendo con el esquema del gen de la Figura 3.2 y las reglas de clasificación de la Tabla 3.1, se ilustra la estrategia para clasificar los *bins* luego de la disección del transcriptoma

Para extraer todas las coordenadas genómicas a los niveles gen, exón, intrón y juntura se debe utilizar el método `binGenome()`. Los resultados se guardan en un objeto de clase `ASpli-features`, que contiene toda la información que es posible de extraer a partir de la anotación.

En el momento que se ejecuta este módulo, se imprimirá un archivo **ASpli_binFeatures.log** con los detalles del transcriptoma en estudio. Esta resultado en sí mismo es informativo como punto de partida del experimento bajo estudio (Figura 3.4).

```
* Number of extracted Genes =33602
* Number of extracted Exon Bins =160604
* Number of extracted intron bins =134251
* Number of extracted trascripts =41671
* Number of extracted junctions =127896
* Number of AS bins (not include external) = 7973
* Number of AS bins (include external) = 7983
* Classified as:
        ES bins =  913          ( 11 %)
        IR bins =  2230         ( 28 %)
        Alt5'ss bins =  1456    ( 18 %)
        Alt3'ss bins =  2298    ( 29 %)
        multiple AS bins =  1076        ( 13 %)
 classified as:
                        ES bins =  133          ( 12 %)
                        IR bins =  284          ( 26 %)
                        Alt5'ss bins =  329     ( 31 %)
                        Alt3'ss bins =  295     ( 27 %)
```

Figura 3.4: Se muestra el resultado del uso del módulo `binGenome()` sobre la anotación del transcriptoma de *Arabidopsis thaliana* versión TAIR10

Las coordenadas sugbenómicas se nombran de la siguiente manera (Ejemplo gen de *Arabidospsis thaliana* CASEIN KINASA II BETA CHAIN 3 (CKB3, AT3G60150):

- **CKB3:E001**: define el primer *bin* exónico

- **CKB3:I001**: define el primer *bin* intrónico
- **CKB3:J001**: define la primer juntura. Las junturas se definen como la última posición del exón dador y la primera posición del exón aceptor. Con nuestro módulo de extracción de coordenadas genómicas es posible extraer las coordenadas de todas las junturas anotadas, que se corresponderían con todos los intrones anotados.

Los *bins* y las junturas se nombran siempre en el sentido 5' ->3'. Esta nomenclatura es independiente de la hebra en que esté codificado el gen e implica que el *bin* o la juntura de más bajo orden estará siempre hacia el 5'.

Podemos resumir el método `binGenome()` en el siguiente pseudocódigo[1]:

```
INPUT: anotación en formato TxDb
1.- Obtener coordenadas genómicas  de genes a partir de todos los
exones constituyentes
2.- Obtener coordenadas genómicas  de transcriptos
3.- Obtener coordenadas genómicas  de todos los exones
4.- Obtener coordenadas genómicas  de todos los intrones
5.- Obtener coordenadas genómicas de las junturas anotadas a partir de los intrones
6.- Obtener bins: superponer coordenadas genómicas de exones e intrones
para cada gen
7.- Clasificar bins en exónicos, intrónicos, alternativos
8.- Clasificar bins alternativos
OUTPUT: Objeto clase "Aspli-features"
```

3.3.2. Superposición de las lecturas alineadas con las coordenadas genómicas. Método `readCounts()`.

En este paso contamos cuántas lecturas se superponen a cada región genómica usando el método `readCounts()`. Los resultados se guardan en un objeto del tipo `ASpli-counts`, que

[1]Se denomina así a la descripción en términos coloquiales de las acciones que deben ejecutarse en un programa o algoritmo para obtener un resultado

contiene los resultados de los conteos a los siguientes niveles:

1. **Genes**: el número de lecturas para un gen es la suma de las lecturas de los *bins* exónicos que lo constituyen. Por esta razón utilizamos el término de longitud efectiva (*effective length*) para definir la longitud del gen como la suma de las longitudes de los *bins* exónicos que lo constituyen.

2. ***Bins***: se informa el número de lecturas que se superpone a cada *bin*, tanto exónicos como intrónicos. En este punto **ASpli** se destaca de los métodos similares ya que reporta el número de lecturas para **todos los intrones**, no solo aquellos anotados como alternativos.

3. **Regiones E1I y IE2**: Para calcular la métrica de retención de intrón (PIR), se considera cada intrón como posible a ser retenido. Se definen 2 regiones de retención **E1I** (la que implica al exón 1 y el intrón en cuestión) e **IE2** (la que conecta el intrón al exón 2) y una juntura de exclusión **E1E2** [38, 56]. En este punto también incorporamos la información de la longitud de la lectura de la biblioteca secuenciada (`l`). El propósito de este parámetro es asegurar que al menos el 10 % de la lectura atraviese la regiones que separan exones de intrones **(E1I, IE2)** (ver Figura 3.7).

4. **Junturas**: Se extraen de los alineamientos. Las junturas se definen como aquellas lecturas que fueron alineadas contra la referencia de forma interrumpida (con *gaps*). Las junturas son esenciales para el descubrimiento y la cuantificación de eventos de *splicing* alternativos anotados y nuevos. Para cada juntura experimental identificada reportamos si es nueva o conocida y que *bins* atraviesa. Además, informamos si está completamente incluída en un *bin* exónico, lo que nos podría dar información de posibles exintrones recientemente descriptos [57].

En el caso de genes y *bins*, se calcula el cociente entre el número de lecturas en una dada región y la longitud de la región, lo que se denomina comúnmente *densidad de lecturas*.

Resultados en este paso. Método `writeCounts()`

Las tablas con los conteos de las lecturas a todos los niveles (gen, exón, intrón y junturas) pueden ser exportadas usando el método `writeCounts()`. A continuación detallamos el contenido

de las tablas.

Genes

	symbol	locus_overlap	gene_coordinates	start	end	length	effective_length	counts			
DX11L1	DX11L1	-	1:11874-14409	11874	14409	2536	1652				

Exones

	feature	event	locus	locus_overlap	symbol	gene_coordinates	start	end	length	counts			
DX11L1:E001	E	-	DDX11L1	-	DDX11L1	1:11874-14409	11874	12227	354				

Introne

	feature	event	locus	locus_overlap	symbol	gene_coordinates	start	end	length	counts			
DDX11L1:I001	I	-	DDX11L1	-	DDX11L1	1:11874-14409	12228	12612	385				

Región E1I

	event	locus	locus_overlap	symbol	gene_coordinates	start	end	length	counts			
DDX11L1:I001	-	DDX11L1	-	DDX11L1	1:11874-14409	12138	12318	181				

Región IE2

	event	locus	locus_overlap	symbol	gene_coordinates	start	end	length	counts			
DDX11L1:I001	-	DDX11L1	-	DDX11L1	1:11874-14409	12522	12702	181				

Junturas

	junction	gene	strand	multipleHit	symbol	gene_coordinates	bin_spanned	j_within_bin	counts			
1.11844.12010	noHit	noHit	*	-	-	-	DDX11L1:E001	NA				

Figura 3.5: Tablas de conteos. Se ilustran los encabezados de las columnas y la primera fila de todas las tablas que se exportan en este paso usando el método `writeCounts()`.

Información genómica (usando como referencia las tablas de la Figura 3.5)

Columnas comunes a todas las tablas:

- **locus_overlap**: Si existe superposición entre cualquier parte del rango genómico con otros genes, esta columna indica el nombre del gen con el que se superpone. Si el gen tiene exactamente las mismas coordendas que otro gen, se conservará el primero en orden alfabético.
- **symbol**: Indica el nombre del gen (DX11L1)
- **gene_coordinates**: Indica las coordenadas genómicas del gen en formato abreviado. En este caso: 1:11874-14409, indica el locus del gen que está en el cromosoma 1, en la región 1:11874-14409.
- **start, end** y **length**: Indica las coordenadas genómicas y longitud de la región en cuestión
- **counts**: Informa la cantidad de lecturas **por muestra** para cada rango genómico en estudio.

Información exclusiva en la tabla de genes:

- **effective_length**: Indica las la suma de las longitudes de los *bins* exónicos que constituyen ese gen.

Información exclusiva en la tabla de *bins* (exones, intrones, regiones E1I e IE2):

- **feature**: Indica la clasificación del *bin* en estudio. Los niveles son **E** (exón) ó **I** (intrón).
- **event**: Si el *bin* es clasificado como alternativo, en esta columna se informa la etiqueta asignada de acuerdo a nuestra clasificación. De esta manera, los *bins* se clasifican en salteo de exón **(SE)**, retención de intrón **(RI)**, dadores/aceptores alternativos **(5'Alt, 3'Alt)** y terminales **(external)**. Además, agregamos la etiqueta ***** para marcar aquellos involucrados en más de 1 evento. Si el *bin* no ha sido clasificado como alternativo tendrá un "- "
- **locus**: Indica las coordenadas genómica del gen

Información exclusiva en la tabla de junturas

- **junction**: En esta columna se indica si existe una juntura anotada que coincida con la experimental bajo estudio. En la mayoría de los casos existen y se informa el nombre de la misma.
- **gene**: Indica el nombre del gen donde está completamente contenida la juntura bajo estudio.
- **strand**: Indica el sentido (hebra) del gen donde está contenida la juntura bajo estudio.
- **multipleHit**: Indica si la juntura está contenida en más de un gen. En el caso que así sea, se utilizará esta información para descartar posibles junturas espúreas, probablemente resultantes de errores en el alineamiento.
- **bin_spanned**: Indica los *bins* con los que tiene contacto la juntura. Esta información es relevante para analizar los posibles casos de *splicing* alternativo. Por defecto, una juntura debe tener contacto con 2 *bins* exónicos.
- **j_within_bin**: Indica si la juntura está incluida completamente en un *bin*. Es una situación inusual, pero en el caso que suceda, podría contribuir al descubrimiento de nuevos exones, intrones o los denominados exintrones [57].

Podemos resumir este método en el siguiente pseudocódigo:

```
INPUT 1: Objeto clase "Aspli-features"
INPUT 2: Alineamientos en formato BAM
1.- Superponer coordenadas genómicas de bins y
lecturas alineadas
OUTPUT: conteos a nivel bin
2.- Sumar lecturas de bins de exones para cada gen
OUTPUT: conteos a nivel gen
3.- Extraer coordenadas de intrones y
crear los rangos artificiales E1I, IE2, usando el parámetro l
OUTPUT: conteos a nivel E1I, IE2
4.- Obtener coordenadas genómicas  de
todas las junturas presentes en los archivos BAM
5.- Contar la ocurrencia de cada una de las junturas obtenidas
OUTPUT: conteos a nivel junturas
OUTPUT FINAL: Objeto clase "ASpli-counts"
```

3.3.3. Estimación de la expresión diferencial de genes y el uso diferencial de *bins* y junturas. Método `DUreport()` .

Utilizando las tablas de conteos para genes, *bins* y junturas se estima su uso diferencial. Para ello se utiliza la función `DUreport()`. Los resultados se guardan en un objeto de tipo `ASpli-DU`.

Para estimar los cambios en expresión entre condiciones, usaremos el paquete de R **edgeR** [58]. En el mismo se propone en primer lugar, aplicar un factor de normalización para corregir posibles desbalances entre la profundidad de secuenciación de cada muestra y también la composición de ARN de cada muestra. En **edgeR** se propone realizar este ajuste calculando un factor de normalización, usando el método TMM (*trimmed mean of Mvalues*) [59] para cada par de muestras. En este tipo de experimentos, los datos se ajustan a un modelo binomial negativo.

Antes de realizar un análisis sobre la expresión diferencial, se debe estimar la dispersión. En este

modelo binomial negativo, se estima el coeficiente de variación biológica (BCV), que comprende la variabilidad debido a la técnica en sí misma y a la variabilidad biológica. Una vez estimada la dispersión y normalizadas las bibliotecas, se aplica un test estadístico (similar al test de Fisher) para evaluar si la diferencia entre las medias de cada condición es significativa. Como resultado, se informa la tasa de cambio (cociente entre las medias de una condición y otra), informado como logaritmo en base 2 y el p-valor correspondiente a la estimación de diferencias entre medias (**p-value**), corregido para mútliple testeo con el método FDR (false discovery rate) [60]. El método de **edgeR** para evaluar la expresión diferencial de genes y el uso diferencial de *bins* y junturas fue elegido por ser uno de los más sólidos y recomendados para este tipo de análisis [61, 62].

Expresión diferencial de genes

Usando las tablas de conteos de genes, la expresión diferencial se estima usando el paquete de R **edgeR** [58]. Las tablas de conteos de los genes deben ser filtradas antes de ser analizadas con el objetivo de descartar genes que no hayan recibido suficiente cantidad de lecturas. En nuestro protocolo, se propone considerar aquellos que hayan recibido en promedio un mínimo de 10 lecturas y una densidad de lecturas mayor a 0.05 en **cualquiera** de las condiciones bajo estudio. Todos estos umbrales pueden ser ajustados por el usuario para ser más o menos restrictivos.

Uso diferencial de junturas y *bins*

A partir de las tablas de junturas, se estima su uso diferencial usando el modelo propuesto por **edgeR** [58], del mismo modo que para genes. En el caso de las **junturas**, se analizan aquellas que:

- Pertenecen a los genes que se consideraron para el análisis de genes y que hayan recibido un número mínimo de lecturas en una de las condiciones. Se puede modificar este umbral usando el parámetro `threshold`, por defecto usamos 5 lecturas.
- El rango de sus coordenadas genómicas esté comprendido dentro de un solo gen.

Bins: Del mismo modo, que genes y junturas, en el caso de los *bins*, analizaremos aquellos que hayan recibido un mínimo de lecturas. En nuestro protocolo, por defecto, se aplican filtros a 2 niveles:

- Se utilizan los *bins* de los genes multiexónicos que hayan recibido como mínimo 10 lecturas en promedio en **todas** las condiciones y cuya densidad de lectura en promedio por condición sea como mínimo 0.05 en **todas** las condiciones. Se remarca la necesidad de evaluar genes expresados en todas las condiciones porque si en alguna condición el gen no se expresara, encontraríamos uso diferencial de *bins* debido a la diferencias en la **expresión** y no a diferencias en el *splicing*.

- Seleccionados los *bins* que pertenecen a esos genes, se analizan si han recibido en promedio, como mínimo, 5 lecturas en **cualquier** condición y si su relación densidad de lectura *bi*n/densidad de lectura gen, sea como mínimo 0.05 en **cualquier** condición.

- El análisis de los *bins* que se hayan definido como **externos** será opcional (se excluyen por defecto).

El uso diferencial de junturas y *bins* se analiza en primer lugar con el método estadístico propuesto en **edgeR** [58]. Dado que estos elementos subgénicos están asociados directamente a la expresión del gen que los contiene, hemos implementado un criterio de normalización de los conteos, que permite cuantificar el uso diferencial específicamente, distinguiendo de los cambios en los conteos producidos simplemente por la expresión del gen.

Dado un Bin_i, que pertenece a un Gen_A en la condición 1, los conteos del Bin_i $(G_A{:}B_i)$ se dividen por los conteos del Gen_A en la condición 1 y se multiplican por los conteos promedio de ese gen a lo largo de todas las condiciones $\tilde{G}_A$:

$$\text{Conteo por } \textit{Bin} \text{ ajustado} = (\frac{G_A : B_i}{G_A})_1 * \tilde{G}_A$$

Esta manera de estimar el uso diferencial de *bins* y junturas sería del modo **evento céntrico** tal como se describió en la introducción. A continuación describiremos un modo complementario, analizando el *bin* desde el lado de las junturas con las que se relaciona.

Resultados que se pueden exportar en esta etapa. Método `writeDU()`.

Tablas de expresión y uso diferencial:

La información genómica detallada para las tablas de conteos se repetirá en las tablas de los

	symbol	locus_overlap	gene_coordinates	start	end	length	effective_length	logFC	pvalue	gen.fdr
A1BG	A1BG	-	19:58858172-58864865	58858172	58864865	6694	1766	-0.164	0.424	0.50

(a) Tablas de expresión diferencial de genes.

Exones

	feature	event	locus	locus_overlap	symbol	gene_coordinates	start	end	length	logFC	pvalue	bin.fdr
LINC01137:E003	E	ES	LINC01137	-	LINC01137	1:37920480-37940044	783034	783186	153	-0.65	0.19	0.82

Intrones

	feature	event	locus	locus_overlap	symbol	gene_coordinates	start	end	length	logFC	pvalue	bin.fdr
LINC01128:I003	I	-	LINC01128	-	LINC01128	1:762971-794826	764485	776579	12095	0.31	0.23	0.86

(b) Tablas de uso diferencial de *bins* exónicos e intrónicos.

Junturas

	junction	gene	strand	multipleHit	symbol	gene_coordinates	bin_spanned	j_within_bin	counts	logFC	pvalue	fdr
1.14829.14970	WASH7P:J001	WASH7P	-	-	WASH7P	1:14362-29370	WASH7P:E001; WASH7P:E002	NA		-0.10	0.50	1

junction_start_hit	Jsum start	PSI c1	PSI c2	junction_end_hit	jsum	PSI c1	PSI c2
-	0	1	1	-	0	1	1

(c) Tablas de uso diferencial de junturas

Figura 3.6: Resultados que se exportan con el método `writeDU()`

pasos posteriores. En el caso de las tablas de expresión diferencial (genes) y uso diferencial (*bins* y junturas analizados por el enfoque clásico) se agregan:

- **logFC** (logaritmo *fold change*): Indica el logaritmo en base 2 de la tasa de cambio estimada entre las condiciones por el método propuesto en el paquete estadístico **edgeR** [58].
- **pvalue**: Indica el p-valor asociado al test efectuado sobre las diferencias de las condiciones
- **fdr**: Indica el p-valor corregido por múltiple testeo.

La tabla de junturas contiene columnas adicionales que brindan información sobre la relación de la juntura en estudio con otras junturas. En el caso que hubiera junturas que comparten el comienzo o el final con la juntura en estudio, se informa la métrica PSI para cada condición, tanto para las que comparten el comienzo como las que comparten final. Entonces tenemos:

- **junction_start_hit, junction_end_hit**: Si existiera, indica el nombre de las junturas que comparten el inicio o el final con la juntura bajo estudio.
- **Jsum_start,Jsum_end**: Indica la suma de lecturas asignadas a las junturas que comparten el inicio o el final con la juntura bajo estudio.

- **PSI C1, C2**: Indican la proporción de la juntura en estudio con respecto a todas las junturas que comparten el inicio o el final con la juntura en estudio en dos condiciones (C1,C2). Si la juntura en estudio no comparte inicio ni final con ninguna juntura, tendrá un PSI $= 1$. Esta es una manera rápida de buscar aquellas junturas que podrían estar cambiando su proporción a lo largo de las condiciones.

Podemos resumir este método en el siguiente pseudocódigo:

```
INPUT 1: Objeto clase "Aspli-counts"
Parámetros: uso de bins externos,
número de lecturas mínimo para gen, bin y juntura.
1.- Filtrar genes por número de lecturas y densidad de lecturas
1.- Estimar uso diferencial de genes
OUTPUT: Todos los genes con su correspondiente p-valor
corregido por múltiple testeo
2.- Filtrar genes para splicing
3.- Extraer junturas y bins de esos genes
4.- Filtrar junturas y bins
5.- Corregir conteos de junturas y bins
6.- Estimar uso diferencial de bins
OUTPUT: Todos los bins y sus junturas con su correspondiente p-valor
corregido por múltiple testeo
OUTPUT FINAL: genes, bins y junturas con sus correspondientes
p-valores ajustados en un objeto ASpli-DU
```

3.3.4. Cuantificación y descubrimiento de eventos de *splicing* usando junturas. Método `AsDiscover()`.

A partir de las tablas de conteos, podemos obtener una mirada **integradora** de los eventos de *splicing* bajo estudio. Es importante y necesario mencionar que junto con la función de extracción

de coordenadas genómicas de los *bins*, esta es la parte más importante del paquete dado que brinda información acerca de la relación entre *bins* y junturas (Figura 3.7).

En nuestro protocolo, las junturas se consideran para el análisis si:

- Están completamente incluídas en un solo gen. No se analizan aquellas que están contenidas parcialmente en más de un locus. Tal como detallamos anteriormente, es posible que se originen por errores en el alineamiento.
- Tienen un mínino de lecturas que las soportan (por defecto 5).

Análisis de *splicing* alternativo a partir de la información de la relación entre junturas y *bins.*

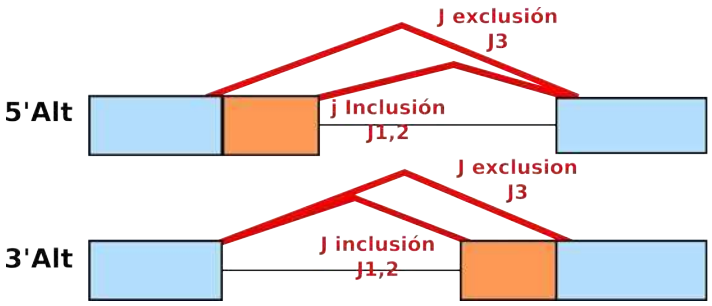
La información de las junturas se utiliza para obtener las métricas de **PSI** (porcentaje de inclusión) y **PIR** (porcentaje de retención de intrón). Estas métricas han sido usadas intensamente durante los últimos años [37] para cuantificar los eventos de *splicing*.

Para realizar estos cómputos los *bins* son separados de acuerdo a la anotación que les fue asignada en el primer módulo. Luego, se computan las métricas de acuerdo al tipo de evento, tal como se detalla en la Figura 3.7. En el caso de los *bins* **exónicos** (excluyendo los anotados como RI), los valores de **PSI** se calculan usando las junturas que comparten la posición de comienzo (J1) o fin (J2) con el *bin* así como aquella que lo excluye completamente (J3). En el caso de los *bins* exónicos anotados como RI y los *bins* intrónicos no anotados como alternativos, la métrica **PIR** se calcula utilizando las lecturas contenidas en las regiones **E1I** e **IE2** y la juntura de exclusión **E1E2** (J3). Estas métricas otorgan confianza a la caracterización del evento en cuestión y complementan el análisis **conteo céntrico** dándole soporte a la estimación del uso diferencial.

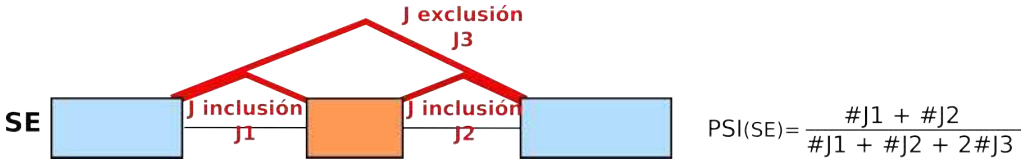
Esta información se reporta en sus correspondientes tablas (ver Figura 3.8) y las columnas respectivas contienen las siguiente información original:

En el caso de los *bins* anotados como **SE** y todos los *bins* exónicos que no fueron identificados como alternativos):

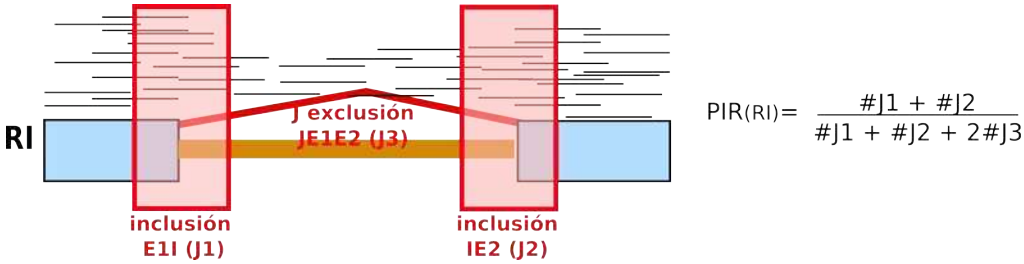
- **Counts J1 (inicio), J2 (final), J3 (exclusion)**: En el caso que sea posible, indican las sumas de las junturas por condición.



(a) Estimación del PSI para sitios de *splicing* alternativos (AltSS) y su relación con las junturas



(b) Estimación del PSI para los eventos de tipo salteo de exón (SE) y de todos los *bins* de exones no anotados como alternativos y su relación con las junturas.



(c) Estimación de PIR para los eventos de tipo retención de intrón (RI) y para todos los *bins* de intrones y su relación con las junturas

Figura 3.7: Estimación de PSI y PIR y su relación con las junturas (en rojo)

SE

	event	Counts J1	Counts J2	Counts J3	PSI C1	PSI C2
LINC01137:E003	ES					

AltSS

	event	Counts J1	Counts J2	Counts J3	PSI C1	PSI C2
PLEKHN1:E007	Alt3ss					

RI

	event	J1	Counts J1	J2	Counts J2	J3	Counts J3	PIR C1	PIR C2
DDX11L1:I001	-	DDX11L1:I001_E1I		DDX11L1:I001_IE2		DDX11L1:J001			

Figura 3.8: Tablas de PSI/PIR

- **PSI C1, PSI C2**: Indican el cálculo de PSI para 2 condiciones (C1, C2) como se indica en la Figura 3.7.

$$PSI(SE) = \frac{\#J1 + \#J2}{(\#J1 + \#J2) + 2 * \#J3} = \frac{\sum \text{Junturas Inclusión}}{\sum \text{Junturas Inclusión} + 2 * \sum \text{Junturas Exclusión}}$$

Para los *bins* exónicos anotados como **5'Alt** y **3'Alt**:

- **Counts J1 (inicio), J2 (final), J3 (exclusion)**: En el caso que sea posible, indican las sumas de las junturas por condición. En este caso, si el evento está bien anotado, en las columnas de J1 y J2, sólo encontraremos una de las 2 junturas presentes, porque la otra no debería existir.
- **PSI C1, PSI C2**: Indican el cálculo de PSI para 2 condiciones (C1, C2) como se indica en la Figura 3.7.

$$PSI(AltSS) = \frac{\#J1,2}{\#J1,J2 + \#J3} = \frac{\sum \text{Junturas Inclusión}}{\sum \text{Junturas Inclusión} + \sum \text{Junturas Exclusión}}$$

PIR: En esta tabla tenemos las siguientes columnas:

- **Counts J1, J2, J3**: Informan los conteos por condición en las regiones E1I (J1), IE2 (J2) y de la juntura de exclusión (J3).
- **PIR C1, PIR C2** Indica el cálculo de PIR para cada condición

$$PIR = \frac{\#J1 + \#J2}{(\#J1 + \#J2) + 2 * \#J3} = \frac{\sum \text{Lecturas Inclusión}}{\sum \text{Lecturas Inclusión} + \sum \text{Junturas Exclusión}}$$

Análisis de *splicing* alternativo a partir de la información de la relación entre junturas.

Utilizando solamente las junturas derivadas del alineamiento, podemos extraer información sobre la ocurrencia de *splicing* alternativo (ver Figuras 3.7, 3.9).

Por un lado, cada juntura se analiza de la misma manera que a los intrones. Es decir, se computan las lecturas que alinean a las regiones flanqueantes E1I, IE2 y la juntura en sí misma se constituye como la juntura de exclusión J3. El propósito de esta tabla es ayudar a corroborar los casos de retención de intrón que se hayan analizado desde el lado de los *bins*, pero además aportar evidencia para el descubrimiento tanto de nuevos intrones como de nuevos eventos de retención de intrón.

PIR

	hitIntron	hitIntronEvent	counts juntura	counts E1I	counts IE2	PIR C1	PIR C2
1.14829.14970	WASH7P:I001	-					

PSI

	junction	gene	strand	multipleHit	symbol	gene_coordinates	bin_spanned	j_within_bin	counts juntura
1.12697.13221	noHit	DDX11L1	+	-	DDX11L1	1:11874-14409	DDX11L1:E002;DDX11L1:E003	NA	

StartHit	counts Start	jsum Start	PSI C1 Start	PSI C2 Start	EndHit	sum End	PSI C1 END	PSI C1 END	pAS
1.12697.13225					1.12721.13221				ES

Figura 3.9: Tablas de junturas

En la tabla que contiene esta información (ver Figura 3.9), tenemos las columnas nuevas:

- **counts Juntura**: son los conteos de la juntura por cada condición
- **counts E1I**: son los conteos de la región E1I por cada condición
- **counts IE2**: son los conteos de la región IE2 por cada condición
- **hitIntron**: informa si la juntura bajo análisis coincide con algún intrón anotado
- **hitIntronEvent**: en el caso de que la juntura coincida con algún intrón anotado, indica si el intrón anotado ha sido clasificado como alternativo

Por otro lado, dada una juntura, podemos analizar si comparte el inicio, fin o ambos con otra juntura. En el caso que esto ocurra, se pueden comparar las proporciones a lo largo de las condiciones (PSI). Utilizando la información del sentido (hebra) es posible de deducir si las junturas estarían definiendo eventos alternativos putatitvos. En el ejemplo (ver Figura 3.9), la juntura bajo análisis (1.12697.13221) no coincide con ninguna conocida (noHit) y comparte inicio y final con otras junturas. Este podría ser un caso de un putativo **SE**.

En la tabla que contiene esta información, tenemos las columnas originales (ver Figura 3.9):

- **junction_start_hit, junction_end_hit**: Si existiera, indica el nombre de las junturas que comparten el inicio o el final con la juntura bajo estudio.
- **Jsum_start,Jsum_end**: Indica la suma de lecturas asignadas a las junturas que comparten el inicio o el final con la juntura bajo estudio.

- **PSI C1, C2**: Indican la proporción de la juntura en estudio con respecto a todas las junturas que comparten el inicio o el final con la juntura en estudio. Si la juntura en estudio no comparte inicio ni final con ninguna juntura, lo cual implica que tendrá un PSI $= 1$ indicando que no está asociado a un evento de *splicing* alternativo. Esta es una manera rápida de buscar aquellas junturas que podrían estar cambiando su proporción debido a la ocurrencia de *splicing* alternativo a lo largo de las condiciones.

Podemos resumir este método en el siguiente pseudocódigo:

```
INPUT: Objeto clase "Aspli-counts"
1.- Separar bins en SE, AltSS, RI, -
2.- Filtrar junturas
3.- Buscar para bins SE, -: J1, J2, J3. Calcular PSI -> Tabla EsPSI
4.- Buscar para bins AltSS, -: J1 o J2 y J3. Calcular PSI > Tabla AltPSI
5.- Buscar para bins RI, I, -: E1I, IE2 y J3. Calcular PIR -> Tabla IrPIR
6.- Unir todas las tablas -> psi_pir
OUTPUT:
7.- para cada juntura buscar si comparte comienzo/fin con otra juntura.
Calcular PSI. -> Tabla JunctionPSI
8.- considerar cada juntura como intrón putativo. Calcular PIR -> Tabla JunctionPIR
OUTPUT final: ASpli-AS
```

3.3.5. Descubrimiento de nuevos eventos de *splicing* alternativo

Utilizando la información recopilada desde el lado de las junturas combinada con la estimación de uso diferencial del modelo **conteo céntrico**, podemos identificar cambios en el uso de sitios de *splicing* alternativo que no hayan sido reportados previamente como tales.

Para ello, en primer lugar debemos analizar cada *bin* y su correspondiente clasificación a nivel de *evento*. Aquellos que no hayan sido identificados como *eventos alternativo*, no tendrán anotación (tendrán un **"-"** para identificarlos). Luego podemos analizar cómo fue su FDR y cómo fueron sus métricas PSI/PIR en cuanto se trate de *bins* exónicos o intrónicos.

Desde el punto de vista de las junturas, la identificación de nuevos eventos de *splicing* se puede analizar de 2 maneras. Para nuevos eventos de SE y sitios de *splicing* alternativos, identificamos junturas que comparten su posición de comienzo o final, o ambas con otras junturas. En ese caso se estiman las proporciones de la juntura bajo estudio, y se reporta un PSI que es simplemente el cociente de la juntura dividido todas las junturas que comparten con ella algún extremo.

Para el caso de descubrimiento de nuevos intrones, cada juntura se piensa como un intrón putativo, y se le calcula el PIR de la misma manera que para los intrones anotados. De esta forma, podemos inferir nuevos eventos de retención de intrón cuando se presente un PIR $>5\,\%$ en alguna de las condiciones.

La información que se obtiene desde el lado de las junturas es complementaria a la que se obtiene desde el lado de los conteos a nivel *bin*. Esta posibilidad de analizar eventos de *splicing* nuevos y anotados de una manera integrada es una de las fortalezas más destacable de nuestro método.

3.3.6. Criterios de selección de eventos de *splicing* alternativo

El objetivo del desarrollo fue encontrar una herramienta que permita identificar aquellos eventos de *splicing* alternativo cuyo uso se haya modificado debido a un cambio en las condiciones, por ejemplo por un cambio ambiental, mutación, etcétera. Para ello, una vez obtenidos todos los análisis debemos decidir cuáles son los eventos que han cambiado siginificativamente. Nuestro protocolo deja a libertad del usuario el criterio preciso de selección de los eventos diferencialmete usados (DU). A partir de nuestra experiencia en el análisis de datos, hemos construido el sigiente criterio para elegir los candidatos, combinando los enfoques **evento céntrico** con el de las **métricas PSI/PIR**.

Para seleccionar *bins* diferencialmente usados:

- *bins* anotados como alternativos: que tengan un valor de FDR <0.1 y una ΔPSI ó ΔPIR $>5\,\%$ y $<95\,\%$.

- *bins* NO anotados como alternativos: que tengan un valor de FDR <0.1 y un ΔPSI ó ΔPIR $>10\,\%$ y $<90\,\%$.

De todas maneras, todo el protocolo está orientado a brindar la libertad de analizar los datos y establecer relaciones entre las diferentes tablas de acuerdo a los criterios escogidos.

3.3.7. Impresión de resultados en formato tabular. Métodos `writeCounts()`, `writeDU()`, `writeAS()`, `writeAll()`

Como ya se detalló para cada módulo, los resultados a cada paso se van guardando en objetos de tipo `ASpli-package` con información accesoria importante. Es posible imprimir las tablas en un formato delimitado por tabulaciones en cada instancia del análisis. Los métodos (`writeCounts()`, `writeDu()`, `writeAS()`, `writeAll()` para imprimir las tablas, generan un directorio con diferentes niveles (Figura 3.10).

Podemos identificar en el nombre de los archivos y la ubicación, la información que contienen. En primer lugar tenemos las diferentes subcarpetas que se corresponden al nivel que estamos analizando: **genes, exones, intrones y junturas**. Luego, cada tabla se identificará según el método que les dio origen:

- **counts**: son las tablas de conteos a todos los niveles,
- **de**: identifica solo a la tabla de expresión diferencial (de) de genes
- **du**: identifica a las tablas de uso diferencial, por lo tanto estarán presentes en los niveles subgénicos bajo estudio: *bins* (exones e intrones) y junturas
- **PIR**: está presente en el subdirectorio de intrones y de junturas,
- **PSI**: está presente en el subdirectorio de exones (discriminando exones anotados como ES y exones anotados como 5' o 3' alternativos) y junturas
- **rd**: identifica a las tablas que contienen el cálculo de densidad de lecturas (*read density*).

Además, se imprimen tablas resumen que contienen información para *bins*, simplemente concatenando tablas de los niveles exón e intrón.

3.3.8. Resultados gráficos. Método `plotTopTags()`

Uno de las formas más comunes de ver gráficamente los resultados de un experimento de RNA-Seq, es utilizando un tipo de gráfico que se denomina *de cobertura* (en inglés *coverage plots*). En este tipo de gráfico en el eje X se reprenta cada posición del genoma de referencia, y

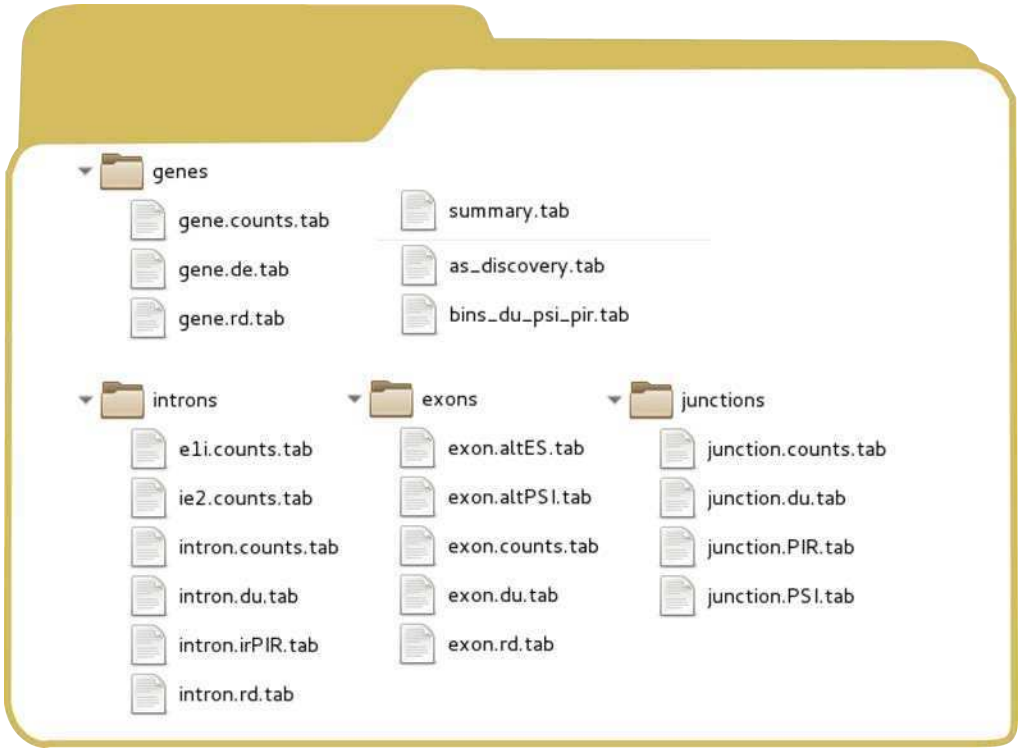


Figura 3.10: Estructura del directorio donde se imprimen las salidas de **ASpli**. Se muestran todas las tablas posibles a exportar en sus respectivos sub-directorios.

en el eje Y, la cantidad de lecturas que han alineado. El análisis gráfico de los resultados es un paso esencial para interpretar los resultados analíticos y para diseñar estrategias de validación. En **ASpli** incluímos una función `plotTopTags()` para graficar de un modo masivo los eventos que consideramos alternativos (ver Figura 3.11).

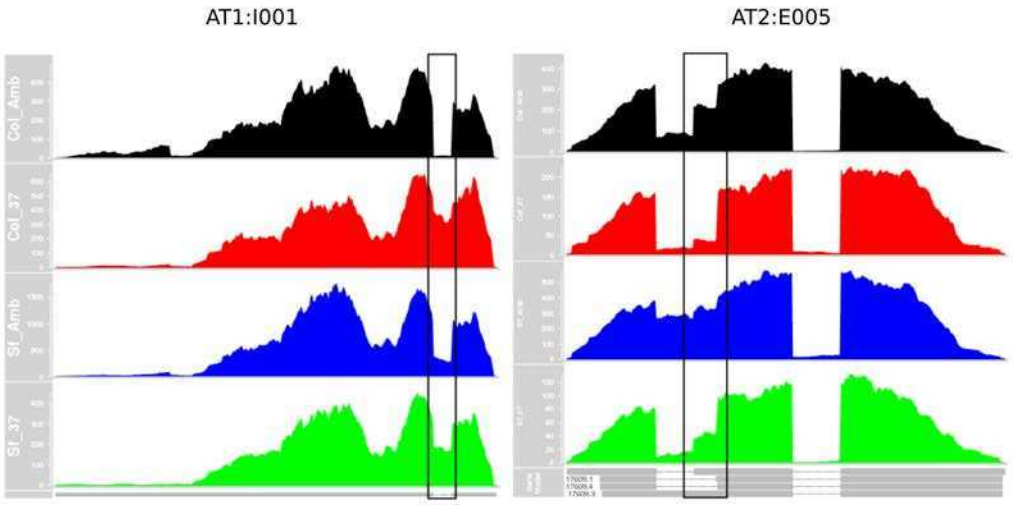


Figura 3.11: Ejemplos de la representación gráfica lograda con **ASpli**

3.4. Disponibilidad

ASpli es un paquete gratuito y de código abierto [63] y su versión 1.0.0 está disponible para su utilización en la plataforma de Bioconductor [54] desde junio 2016.

Para su aceptación, el paquete fue revisado intensamente y debe pasar múltiples controles en lo que respecta al código. La pertenencia al consorcio otorga visibilidad en el mundo bioinformático a la vez que estandariza numerosos procedimientos relacionados a formatos, instalaciones y manuales de ayuda. Previamente fue compartido a pedido y también fue utilizado en múltiples trabajos del laboratorio y de otros laboratorios en carácter de colaboración [64–68]. Para utilizarlo, se puede descargar manualmente del sitio de Bioconductor (`http://bioconductor.org/packages/release/bioc/html/ASpli.html`), o puede instalarse desde la consola de R con las siguientes instrucciones:

```
source("https://bioconductor.org/biocLite.R")
biocLite("ASpli")
```

3.5. Conclusiones

A lo largo de este capítulo hemos intentado describir el proceso de construcción de un protocolo bioinformático para la caracterización de la remodelación del transcriptoma por medio del proceso de *splicing* alternativo de un modo global, a partir de datos de experimentos de RNA-Seq.

Nuestro protocolo fue evolucionando hasta tomar forma de un paquete de R/Bioconductor. El mismo permite compartir con otros usuarios el código de manera prolija y controlada. Además la plataforma de Bioconductor brinda visibildad y un control de calidad sobre lo que se allí se publica.

Una de las fortalezas del método es la integración de los datos para el análisis de los eventos de *splicing* alternativo. Esto siginifica que el comportamiento de cada *bin* se caracteriza desde la visión **conteo céntrica** estimando su uso diferencial a partir de un estadístico y también de una manera complementaria, mediante el análisis de las junturas que lo atraviesan. Este análisis integrado nos brinda información más fidedigna a la hora de elaborar conclusiones sobre el efecto global de un tratamiento sobre la regulación del *splicing* alternativo.

Otra de las fortalezas, es la estructura modular, que habilita al usuario a tomar decisiones a cada

paso del protocolo. Dado que **ASpli** está escrito en un lenguaje popular para el mundo científico, todos los objetos que se van generando son muy fáciles de analizar e interconvertir para usar con múltiples herramientas disponibles. Así, por ejemplo, las tablas de conteos pueden ser analizadas con otros paquetes de análisis de expresión y *splicing* diferencial, realizar cómputos de distancias, de agrupamientos, correlaciones, etcétera. Cada módulo de **ASpli** tiene su propio resultado, que puede ser exportado fácilmente a archivos delimitados por tabulaciones. Los mismos, son autocontenidos y pueden manipularese fácilmente en *softwares* como planillas de cálculo convencionales.

Tal como se detalló en la introducción, el análisis de la regulación del *splicing* alternativo usando el enfoque **evento céntrico**, es más informativo a la hora de elaborar hipótesis sobre las cuestiones mecanísticas que están gobernando los cambios en el transcriptoma.

Toda la información genómica que **ASpli** provee luego de la disección de transcriptoma en *bins* es de mucha utilidad para los análisis *a posteriori* de la elección de los eventos diferencialmente usados. Por ejemplo, el análisis de secuencias regulatorias en las regiones flanqueantes a las junturas, el enriquecimiento de motivos, frecuencia de kmeros, la conservación de los sitios de *splicing*, etcétera, se hace de manera muy sencilla simplemente extrayendo la información de las posiciones, como se ejemplifica en los Capítulos 4 y 5.

Otra de las características que le han dado versatilidad al paquete, es que no está limitado a un diseño experimental. Utilizando las tablas de conteos o de cálculo de PSI/PIR, se pueden analizar diseños más complejos, como por ejemplo de medidas repetidas en el tiempo o interacción entre distintas condiciones ambientales y genotipos, muy comunes en el laboratorio donde se desarrolló esta tesis.

Capítulo 4

El efecto de la luz sobre el *splicing* alternativo en las plantas.

4.1. Introducción

Las plantas han evolucionado su capacidad de percibir cambios en la cantidad, calidad, duración y dirección de la luz para ajustar su crecimiento y desarrollo en las condiciones ambientales más propicias [69]. Las plántulas que crecen en oscuridad, dado que no han emergido a la superficie, tienen hipocotilos elongados, cotiledones cerrados y etioplastos no fotosintéticos. Una vez que las plántulas se exponen a la luz, sufren una sucesión de modificaciones fisiológicas que contribuyen a la transición entre el crecimiento heterotrófico a autotrófico, tales como la inhibición del crecimiento de los hipocotilos, la apertura de los cotiledones y el desarrollo de los cloroplastos [69]. Las plantas que crecen a la luz, detectan luego cambios en la relación de luz roja y roja lejana que les llega debido a la reflexión de los rayos de luz sobre las hojas de las plantas que las rodean y responden a estas señales con respuestas fisiológicas para evitar que las alcance el sombreado [70]. Finalmente las plantas maduras perciben cambios en la longitud del día y usan esa información para ajustar el momento de la floración hacia el momento más favorable del año [71]. Las señales de la luz se perciben a través de varias familias de fotoreceptores. Particularmente, en las plantas de *Arabidopsis thaliana*, los fotoreceptores sensoriales conocidos incluyen: 5 fitocromos (phys) que mayormente absorben en la longitud de onda del rojo y rojo lejano, 2 criptocromos (crys), 2 fototropinas y 3 miembros de la familia de proteínas zeitlupe que absorben longitudes de onda azul y ultra violeta

A (UV) y los recientemente caracterizados receptores de luz ultra violeta B (UV-B) [72].

Luego de la activación, la mayoría de estos fotoreceptores controlan múltiple procesos fisiológicos vía regulación génica [72, 73]. Los fitocromos que han sido activados por la luz, interactúan directamente con factores de transcripción de la familia de proteínas bHLH controlando su vida media (turnover) [74]. Los fitocromos también controlan la expresión de los genes activando la proteína ubiquitín-ligasa E3, *CONSTITUTIVE PHOTOMORFOGENIC 1 (COP1)*, que marca a muchas proteínas para degradación, tales como el factor de transcripción de tipo bZIP *ELONGATED HYPOCOTYL 5 (HY5)* [74]. Todos estos factores de transcripción regulados por los fitocromos luego contolan la transcripción de cientos de genes [75].

Existen evidencias crecientes que la luz regula otros aspectos de la red regulatoria de la expresión génica, tales como la estabilidad del ARN, la traducción y el *splicing* alternativo [73, 76–85]. Por ejemplo, una breve exposición a la luz roja en plántulas etioladas, es decir que nunca se expusieron a la luz previamente, regula los patrones de *splicing* alternativo en cientos de genes, incluyendo los de factores de *splicing* en sí mismos. Estos efectos están mediados por los fitocromos más abundantes phyA y phyB, al menos en las plántulas etioladas [81]. Resultados similares relacionados a la regulación del *splicing* alternativo por el fitocromo fueron obtenidos en un trabajo reciente sobre *Physcomitrella patens* [84]. La luz también modula el *splicing* alternativo en las plantas desetioladas (es decir ya expuestas a la luz). En este estado de desarrollo, la mayoría de los efectos de la luz en el *splicing* aparentan estar mediados a través/vía una señál retrógrada involucrada en una molécula intermediaria generada en el cloroplasto durante el proceso fotosintético y no por los fotoreceptores fotosensibles [80]. A pesar de que se han evaluado los efectos de la luz sobre el *splicing* alternativo de un modo global en plántulas etioladas, su efecto sobre el *splicing* alternativo en plantas desetioladas sólo ha sido investigado en un grupo reducido de genes.

En este capítulo, describiremos los resultados obtenidos utilizando la tecnología de RNA-Seq para evaluar el efecto de un breve pulso de luz dado en el medio de la noche, sobre plántulas de *Arabidopsis thaliana* crecidas en día largo (12 horas de luz, 12 horas de oscuridad). El diseño experimental intentó simular las señales lumínicas del amanecer temprano o del atardecer tardío asociadas con alargamiento del fotoperíodo y podría permitirnos identificar mecanismos posttranscripcionales involucrados en la regulación del reloj circadiano asi como también de la transición floral regulada por luz. Este trabajo fue el más reciente publicado en el laboratorio utilizando **ASpli**.

En el mismo, implementamos la versión más actual del paquete, donde la elección de eventos de *splicing* alternativo fueron elegidos según el criterio descripto en el capítulo anterior, combinando las métricas de las junturas PIR/PSI con el p-valor estimado a partir del análisis tradicional.

4.2. Materiales y métodos

Material vegetal y condiciones de crecimiento

Se prepararon tres réplicas biológicas a partir de plántulas de *Arabidopsis thaliana* del ecotipo Columbia-0. Las semillas se estratificaron por 4 días en la oscuridad a 4°C y luego se transfirieron a condiciones lumínicas de doce horas luz / doce horas de oscuridad a 22°C en luz blanca. Se utilizó el medio MS *(Murashige and Skoog)* conteniendo 0.8 % de agar. Luego de 9 días, las plantas se irradiaron en el medio de la noche (**ZT18**, es decir 6 horas luego del inicio de la noche) con un pulso de luz blanca o roja durante 2 horas, o fueron dejadas en la oscuridad como control. El pulso de luz blanca (50 $\mu molm^2s^{-1}$) fue realizado con tubos fluorescentes (Sylvania Standard F18WT8154). El pulso de luz roja (80 $\mu molm^2s^{-1}$ fue realizado con diodos emisores de luz (Kingbrigth Super Bright Leds, L934-SRC, $\lambda = 660 \pm 20nm$). Las muestras tratadas y control fueron cosechadas en el mismo momento y conservadas en nitrógeno líquido. Se esquematiza el protocolo en la Figura 4.1

Figura 4.1: Diseño experimental

Preparación de las bibliotecas para secuenciar

Se aisló ARN total siguiendo el protocolo de *RNaeasy Plant Kit (QIAGEN)*. Se determinó la concentración y la calidad de las muestras con *NanoDrop 2000c (Thermo Scientific)* y *Agilent RNA 6000 NanoKit*, respectivamente. Las bibliotecas para secuenciación se prepararon siguiendo el protocolo *TruSeq RNA Sample preparation Guidelines (Illumina)*, salvo para el paso de amplificación

Nombre de la muestra	Lecturas
control A	12011875
control B	11205481
control C	12741544
pulso A	11494706
pulso B	12281861
pulso C	8983890
promedio	11453226

Cuadro 4.1: Rendimiento de las bibliotecas

Entre muestras (min, max):		0.78	0.99
Entre réplicas (min, max):	control	0.87	0.99
Entre réplicas (min, max)	pulso	0.89	0.99

Cuadro 4.2: Correlación entre réplicas y muestras

que se usaron sólo 12 ciclos de amplificación. El control de la calidad de la biblioteca se realizó con los kits *Agilent 2010 Bioanalyzer* y *Agilent DNA 1000 Kit (Agilent technologies)*. Luego se construyeron 6 bibliotecas, que fueron secuenciadas desde un solo extremo *(single end)*, con la tecnología *Illumina GAIIx*. Se obtuvieron en promedio 11.4 millones de lecturas de 100 pares de bases para cada muestra. Ver Cuadros 4.1 y 4.2.

Alineamiento de lecturas contra genoma de referencia

Las lecturas obtenidas fueron alineadas contra el genoma de referencia de *Arabidopsis thaliana* usando el software TopHat version 2.0.9 [86, 87] con parámetros por defecto, considerando la máxima longitud de intrón esperada en 5000 pares de bases, de acuerdo a [88]. Los archivos resultantes de cada alineamiento en formato BAM fueron luego utilizados para el análisis con **ASpli**. Todos los archivos resultantes de la secuenciacón fueron depositados en la base de datos de *Gene Expression Omnibus, GEO* [89], con el número GSE68560 [90].

Selección de genes diferencialmente expresados

Utilizando los alineamientos contra las referencias, se procedió al análisis de genes y eventos tal como se detalló en el capítulo anterior. Los archivos BAM junto con la anotación fueron utilizados para generar las tablas de conteos a los niveles de genes y *bins*. Para seleccionar los

genes diferencialmente expresados, seleccionamos aquellos genes que:

- tuvieran en promedio como mínimo 10 lecturas en alguna de las condiciones
- tuvieran en promedio como mínimo una densidad de lecturas mayor a 0.05 en alguna de las condiciones

Luego de aplicar estos filtros sobre los 33602 genes anotados en TAIR10 [91], obtuvimos un total de 15925 genes que se consideraron *expresados* y que fueron analizados para estimar expresión diferencial. Los mismos se analizaron utilizando el método estadístico propuesto en el paquete **edgeR** [92], para estimar la tasa de cambio (*fold change*) con su estadístico asociado (p-valor) que fue corregido por un método de mútiple testeo (FDR) [60]. Luego, para obtener los genes que consideramos diferencialmente expresados se eligieron aquellos cuyo:

- valor absoluto del logaritmo en base 2 de la tasa cambio sea > 0.59, lo cual significa un cambio de al menos 1.5 veces entre las condiciones,
- que el valor corregido del p-valor sea < 0.05

Luego de aplicar estos criterios obtuvimos **4341** genes que satisfacían estas condiciones (etiquetados como Genes DE).

Selección de eventos de *splicing* alternativo

A partir de los alineamientos guardados en los archivos BAM, se generaron las tablas de conteos a los niveles de genes, *bins* de exónes y de intrones y de las junturas. Los *bins* que se conservaron para el análisis de uso diferencial son aquellos que satisfacían los siguientes criterios:

- Pertenecer a genes multiexónicos cuya cantidad de lecturas en promedio fuera mayor a 10 y cuya densidad de lectura fuera como mínimo 0.05 en **todas las condiciones**.
- Tuvieran en promedio más de 5 lecturas y una densidad de lecturas fuera como mínimo 0.05 en alguna de las condiciones

Las junturas que se utilizaron para analizar los *bins* son aquellas que:

- Pertenecían a genes cuya cantidad de lecturas en promedio fuera mayor a 10 y cuya densidad de lectura fuera como mínimo 0.05 en **todas las condiciones**.

- Tuvieran en promedio más de 5 lecturas en alguna de las condiciones

Utilizando estos datos, y siguiendo los criterios propuestos en **ASpli** se seleccionaron aquellos *bins* que presentaban un valor absoluto de cambio (*log 2 Fold Change*) mayor a 0.59 y un FDR inferior a 0.15. A su vez, se seleccionaron los *bins* utilizando las métrica PSI y PIR, tal como se detalla en el capítulo anterior y se seleccionaron aquellos cuyo ΔPSI o ΔPIR fuera superior a 5 % o 10 % para eventos conocidos o nuevos respectivamente, entre las condiciones en estudio.

Nivel	Total	Elevados	Disminuídos
Genes	4341	2149	2192
bins			
3'Alt	99		
Anotados	14	9	5
Nuevos	85	47	38
5'Alt	46		
Anotados	0		
Nuevos	46	25	21
SE	27		
Anotados	5	4	1
Nuevos	22	9	13
RI	232		
Anotados	16	7	9
Nuevos	216	109	107

Cuadro 4.3: Tabla resumen de genes y eventos

RT-PCR semicuantitativa

Para realizar las validaciones utilizando la técnica de RT-PCR, se obtuvieron 3 réplicas biológicas a partir de aproximadamente 50 semillas de *Arabidopsis thaliana* de ecotipo Columbia-0, así como

de mutantes *phyA-211;phyB-9 (phyAB), phyA-211;phyB-9;phyC-2;phyD-201; (phyABCDE)* [93], o *cry1-304;cry2-1 (cry1:cry2)* [56]. Las plantas fueron crecidas en placas MS-agar, estratificadas en frío 5 días y crecidas 8 días en condiciones de 12 horas luz 12 horas oscuridad. En el día 9, se aplicó un pulso de luz de 2 horas, en el punto ZT18 sobre las plántulas, tal como se detalló para el experimento de RNA-Seq. El ARN total se extrajo con Trizol y se realizó la transcripción reversa usando la enzima Superscript II *(Life Technologies)* de acuerdo al manual de usuario de la compañía. El ADNc resultante se utilizó para la amplificación con PCR, con 30 ciclos. Luego los productos de RT-PCR se analizaron en un gel de agarosa 3 % wt/vol y fueron detectados con Sybr Green como se propone en [94].

Análisis de enriquecimiento de Ontologías Génicas

Para analizar los grupos de genes y eventos seleccionados como diferencialmente expresados (DE) y diferencialmente usados (DS) se utilizó la estrategia de Ontologías Génicas (*Gene Ontology, GO*) [95]. Para asignar la ontología que le corresponde a cada gen seleccionado, se utilizó la herramienta de *Virtual Plant* [96,97]. La prueba de enriquecimiento usando la prueba de Fisher con su correspondiente corrección por múltiple testeo [60] se realizó en las 3 categorías que describe GO: Procesos biológicos (PB), Función Molecular (FM) y componente celular (CC). El factor de enriquecimiento se calculó tomando las cantidades de genes en el grupo en estudio asignadas a cada categoría sobre el total de genes analizados y los genes totales anotados en esa categoría con respecto al total de genes anotados (comúnmente denominado universo):

$$\text{Factor enriquecimiento (FE)} = \frac{\dfrac{\text{Genes anotados a una categoría GO}}{\text{Total de genes analizados}}}{\dfrac{\text{Genes anotados en esa categoría GO}}{\text{Total de genes anotados con GO}}}$$

Cuando nos referimos a grupos de genes bajo estudio, referimos a: genes diferencialmente expresados (DE) o genes con eventos de *splicing* diferencialmente usados (DS). El universo entonces serían todos los genes expresados para el primer caso, y todos los genes de los eventos analizados en el segundo caso.

4.3. Resultados

Para investigar el efecto de la luz sobre el *splicing* alternativo de una manera global en las plantas desetioladas, las plántulas de *Arabiopsis thaliana* fueron crecidas bajo condiciones de 12 horas de luz/12 horas de oscuridad durante 8 días. En el noveno dia, en la mitad de la noche (ZT18) la mitad de las plantas se expusieron a un pulso de luz durante 2 horas y la otra mitad se mantuvo en condiciones de oscuridad. Luego de las 2 horas, se cosecharon las muestras y se prepararon las bibliotecas para RNA-Seq. Se secuenciaron bibliotecas de ADNc de 3 replicados biológicos para cada condición. La lecturas resultantes de la secuenciación fueron mapeadas al genoma TAIR10 y se evaluaron los cambios en la expresión génica y en los eventos de *splicing* alternativos conocidos como nuevos. Se identificaron un total de **4341** genes diferencialmente expresados (más de 1.5 veces aumentada/disminuída su expresión) con un de p-valor corregido inferior a 0.05, en respuesta al pulso de luz. Este grupo estaba enriquecido en genes asociados con los componentes del cloroplasto, metabolismo primario y secundario, estrés oxidativo y respuestas abióticas (ver Figura 4.2). También se encontró enriquecimiento en genes asociados con los ritmos circadianos, sosteniendo la idea que los efectos de la luz sobre la red circadiana está mediada principalmente por los efectos en los niveles de ARNm de varios genes centrales del reloj.

Luego evaluamos los efectos del la luz sobre el *splicing* alternativo caracterizando sus efectos tanto en los eventos anotados como en eventos nuevos. Se identificaron un total de **382** genes que presentaban eventos de *splicing* alternativo que fueron regulados por el tratamiento del pulso de luz. Es destacable que sólo la mitad de los genes que tenían afectados los patrones de *splicing* fueron afectados también a nivel ARNm sugiriendo que al menos en parte, la luz estaría afectando el *splicing* por mecanismos independientes de la regulación de la transcripción.

Tal como se había reportado previamente sobre los efectos de la luz sobre el *splicing* alternativo en plantas etioladas [81], o en respuesta a transiciones de luz oscuridad prolongadas [80], también observamos un fuerte enriquecimiento en categorías GO asociadas con el procesamiento del ARN y el *splicing* alternativo entre esos genes cuyos patrones de *splicing* fueron afectados por el pulso de luz. Es sorprendente y coherente a la vez que estas categorías no estaban enriquecidas entre los genes que tenían alterados los niveles de expresión. Esta observación soporta la idea que la luz regula los patrones de *splicing* alternativo en su mayoría afectando el *splicing* de los factores de

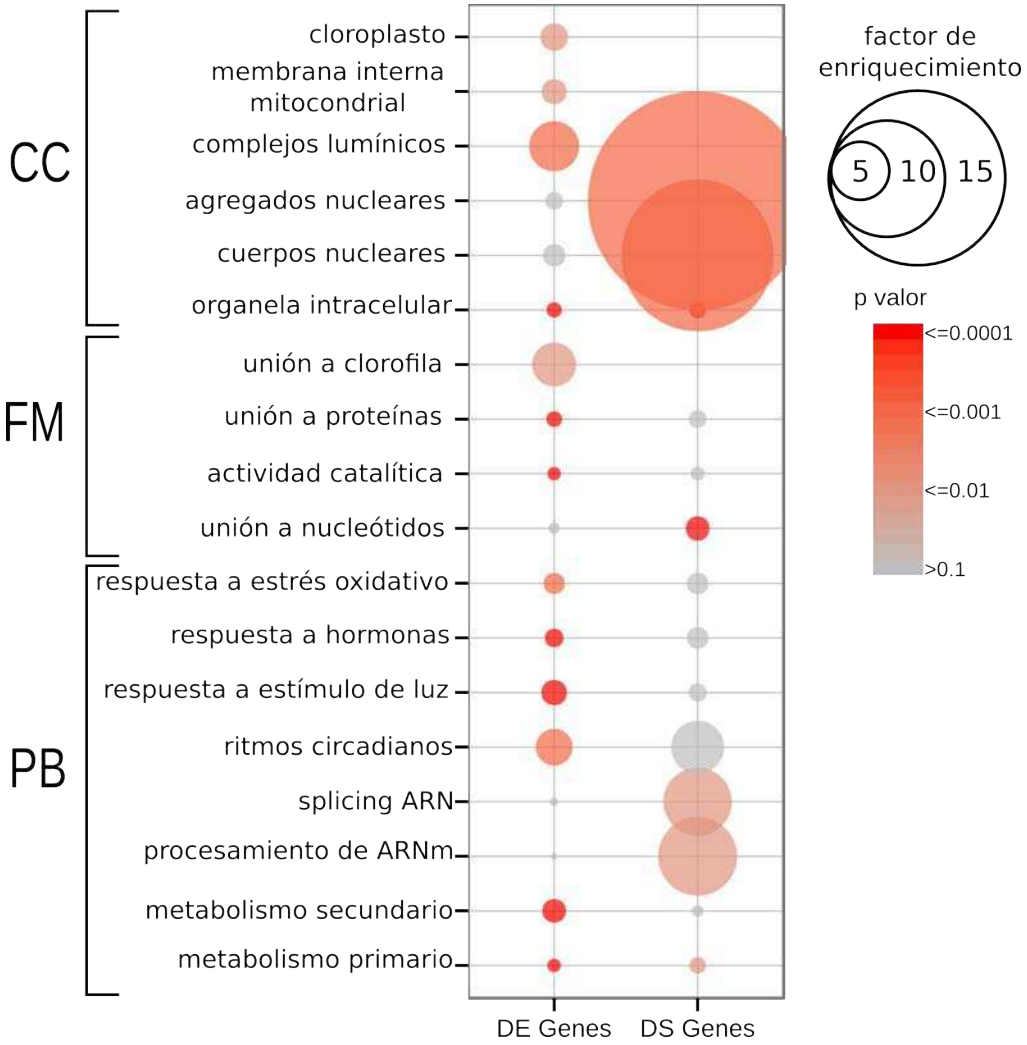


Figura 4.2: Enriquecimiento de categorías ontológicas

splicing en sí mismos.

En un estudio global sobre el efecto de la luz en la remodelación de *splicing* alternativo en el musgo *Physcomitrella patens* [84] los autores encontraron que la luz promueve la retención del intrón en la mayoría de los genes. En nuestro análisis también encontramos que la retención de intrón fue el evento que más se presentó, pero en algunos casos promoviendo la retención y en otros, la exclusión. Una observación similar se obtuvo a nivel de exónes donde algunos genes mostraron que se favorecía la inclusión y en otros la exclusión (ver Figura 4.3). Estas observaciones indican que la luz actúa como un modulador de los patrones de *splicing* y no con un efecto inhibitorio sobre el proceso de *splicing* en sí mismo.

Luego focalizamos el análisis sobre genes asociados al reloj circadiano. A pesar que la categoría de ritmos circadianos no estaba enriquecida en nuestros análisis, encontramos genes del reloj con patrones de *splicing* alterados, como *REVEILLE 8 (RVE8/LCL5, AT3G09699), JUMONJI*

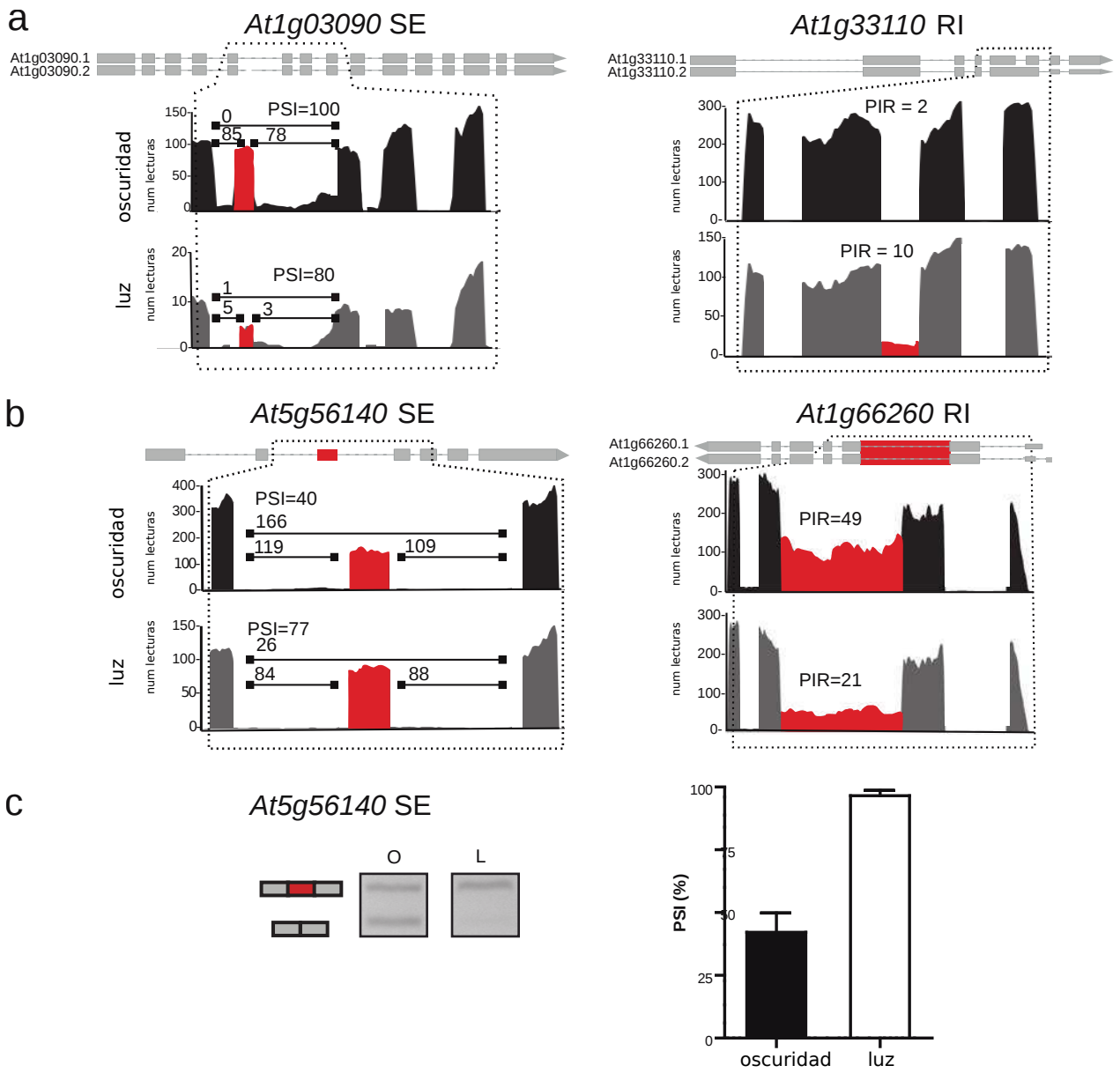


Figura 4.3: Ejemplos de retención de intron (RI) y salteo de exón (SE). Resultados de RNA-Seq de genes representativos mostrando cambios en *splicing* alternativo a nivel SE y RI. Se muestran los gráficos de cobertura con su respectivo modelo génico. (a) Ejemplo de eventos anotados (SE y RI) (b) Ejemplo de 2 eventos nuevos (SE y RI). (c) Validación de los resultados de RNA-Seq para eventos nuevos de SE usando RT-PCR. El gráfico representa el promedio de 3 réplicas independientes y su correspondiente barra de error. Las regiones alternativas se muestran en rojo.

DOMAIN CONTAINING 5 (JMJD5, AT1G01060), TIME FOR COFFEE (TIC, AT3G22380) y *CASEIN KINASA II BETA CHAIN 3 (CKB3, AT3G60150)*. Algunos de estos eventos de *splicing* habían sido previamente reportados y otros fueron descriptos por primera vez en nuestro análisis, como por ejemplo el uso de un aceptor alternativo (3'Alt) en el gen *JMD5* . Todo lo que respecta a la implicancia funcional de estos patrones resta por ser dilucidado.

Entre los resultados muy interersantes de nuestro estudio, encontramos que muchos de los

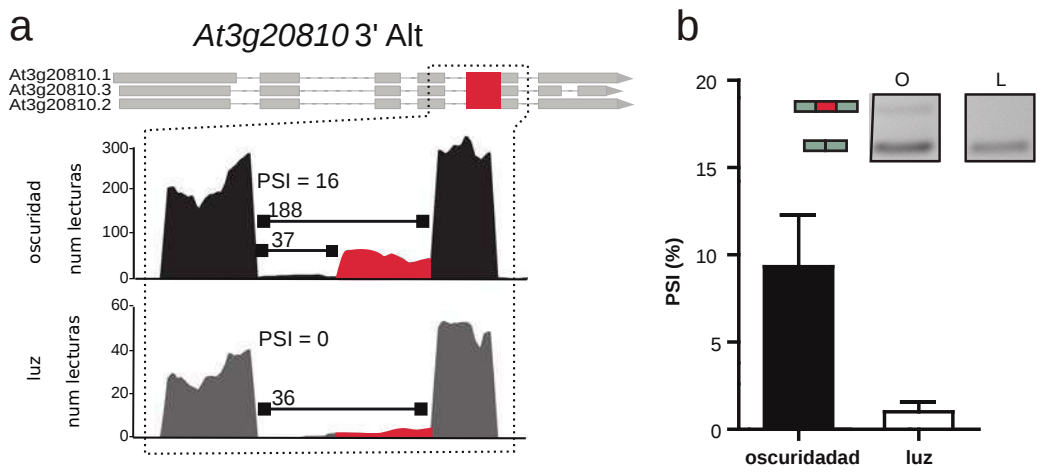


Figura 4.4: La luz regula el *splicing* alternativo del gen del reloj JMJD5. (a) Gráficos de cobertura y su modelo de gen. (b) Validación de los resultados de RNA-Seq usando RT-PCR. El gráfico representa el promedio de 3 muestras biológicas independientes y su error standard correspondiente. Las regiones alternativas se destacan en rojo.

genes donde encontramos eventos de *splicing* alternativo regulados por luz, codificaban proteínas ricas en serinas (proteínas SR) y ribonucleoproteínas heterogéneas nucleares (hnRNP) que son conocidas moduladoras del reclutamiento de los *snRNPs* a los sitios donores y aceptores [23]. Se había demostrado que muchos genes son regulados a nivel de *splicing* alternativo luego de una exposición prolongada a la luz u oscuridad en plantas desetioladas [80] o ante un tratamiento de agudo de luz roja en plantas etioladas [81]. En nuestro trabajo también mostramos que la luz induce cambios muy rápidos en el *splicing* alternativo sobre factores de *splicing* en sí mismo en las plantas crecidas en luz (ver Figura 4.5 y Tabla 1.2).

Es interesante destacar que los efectos de la exposición prolongada a la luz sobre el *splicing* alternativo no mostró ser dependiente de fotoreceptores sensoriales clásicos tales como fitocromos o criptocromos y estar mediados principalmente por los efectos de luz sobre los procesos fotosintéticos [80]. Por lo contrario, la percepción de 3 horas de luz percibida a través de los fitocromos A (phyA) y B (phyB) ha sido recientemente reportada de regular los patrones de *splicing* alternativo de varios genes, incluyendo aquellos que codifican factores de *splicing* en plantas etioladas. [81]. En nuestro trabajo, evaluamos los efectos de la luz blanca sobre el *splicing* en el factor de *splicing SERINE/ARGININE RICH PROTEIN SPLICING FACTOR30 (SR30, AT1G09140)*, en plantas *phyAB* y las mutantes dobles *cryptochrome 1,2 (cry1;cry2* (Figura 4.6). Encontramos que, de la misma manera que lo que se había reportado en el trabajo de los efectos prolongados de

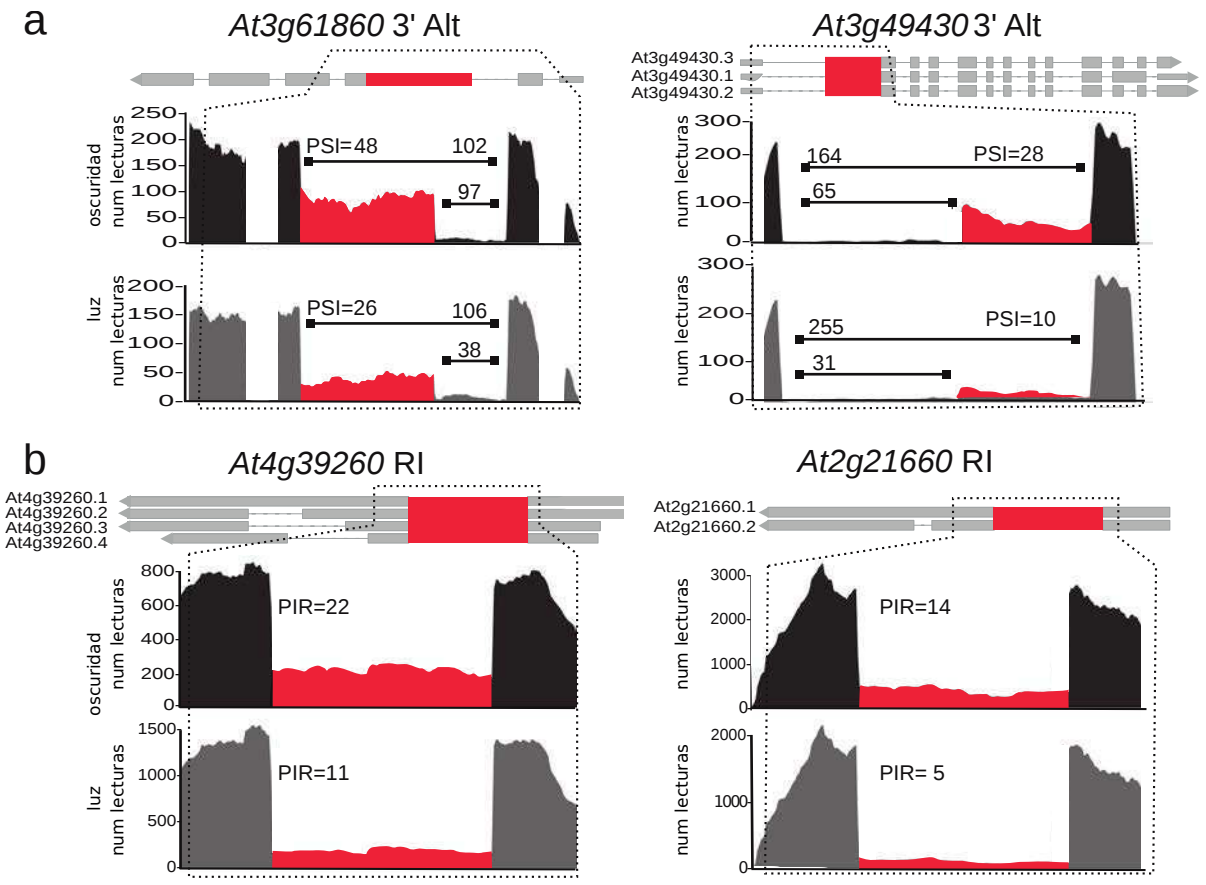


Figura 4.5: Gráficos de cobertura de eventos de *splicing* alternativo regulados por luz en genes que codifican factores de *splicing*.(a) Eventos de *splicing* asociados con genes que codifican proteínas ricas en serinas-argininas (SR) (b) Eventos de *splicing* asociados a genes que hnRNPs. Las regiones alternativas se destacan en rojo.

la luz [80], los efectos de un pulso de luz corto también presentaban efecto sobre SR30 en las mutantes, sugiriendo que la mayor parte de los efectos de la luz estarían siendo modulados por una molécula generada durante el proceso fotosintético y no por las vías tradicionales de acción de los fotoreceptores fotosensibles. De todas maneras la persistencia de los efectos de la luz blanca sobre SR30 en los mutantes *phyAB* y *cry1;cry2* podrían deberse a fotoreceptores adicionales de la misma o diferente familia. Entonces evaluamos el efecto de la luz roja sobre el *splicing* alternativo de SR30 en el quíntuple mutante *phyABCDE* que es completamente ciego a las señales de luz roja que controlan el desarrollo de las plantas. Llamativamente, encontramos un fuerte efecto sobre el patrón de *splicing* alternativo en SR30, de magnitud similar a las plantas salvaje (Figura 4.6).

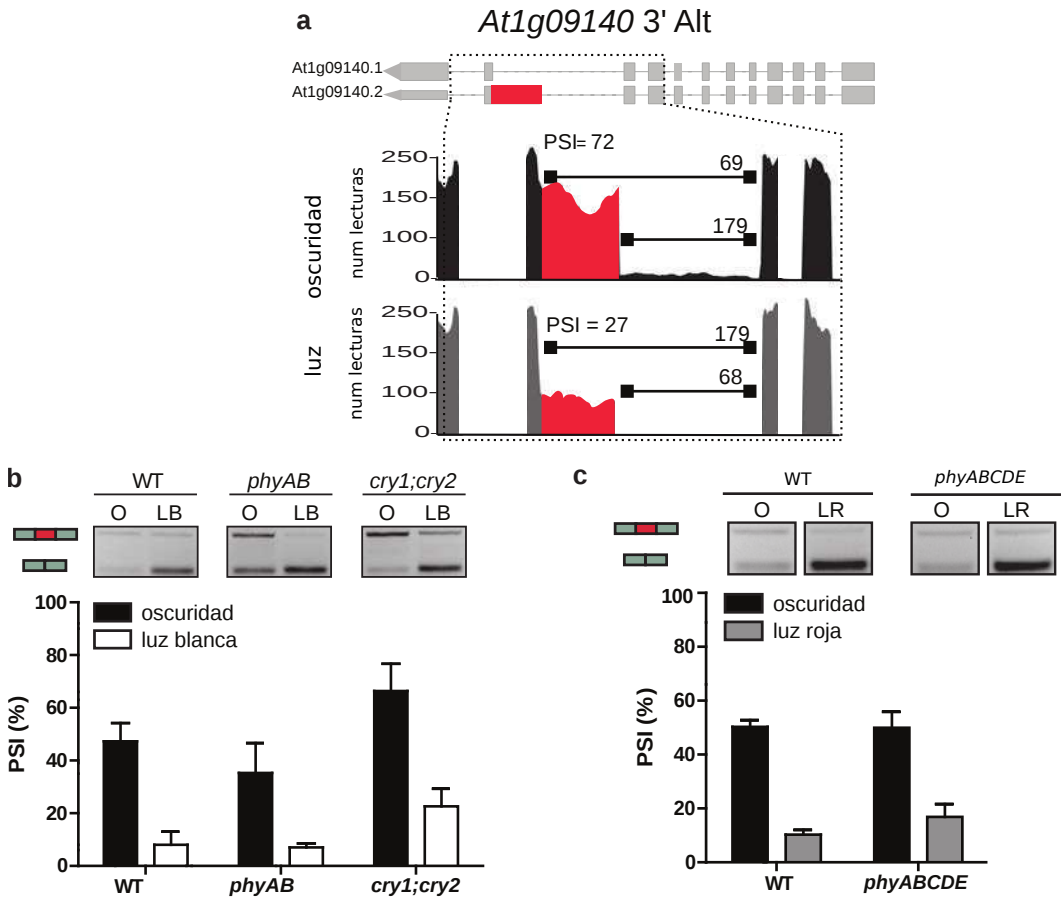


Figura 4.6: Validación experimental de la regulación de SR30 por la luz. (a) Gráficos de cobertura y sus modelos de genes asociados (b) Electroforesis en gel de los productos de *splicing* amplificados por RT-PCR de plantas Col-0, *phyAphyB* y *cry1cry2* expuestas a 2 horas de pulso de luz blanca (LB) o mantenidas en la oscuridad como control (O). (c) Electroforesis en gel de los productos de splicing amplificadds de plantas Col-0 y mutantes *phyABCDE* expuestas a 2 horas de pulso de luz roja (LR) o dejadas en la oscuridad como control (O). Los gráficos representan el promedio de 3 replicados biológicos independientes y las barras de error representan el error estandard. Las regiones alternativas se muestran en rojo.

4.4. Discusión

Las evidencias de que la luz modula el desarrollo y crecimiento de las plantas vía expresión génica son contundentes [73]. Hay muchos estudios globales que demuestran cómo la luz regula los niveles de ARNm de cientos de genes, pero muy pocos que evalúan su efecto a nivel *splicing*. El primer estudio sobre los efectos de la luz sobre el *splicing* alternativo fue realizado en arroz utilizando un microarreglo de expresión y detectó unos 10 genes con patrones alterados de *splicing* alternativo regulados por luz [98]. La mayoría de los efectos implicaban cambios en la región no traducida *(UTR)* y no en las regiones codificantes, sugiriendo que la luz podría estar controlando los niveles de ARNm a través de la estabilidad. Sin embargo, muy pocos eventos pudieron ser identificados, debido tal vez a que la tecnología disponible para ese entonces no era la más adecuada. Dos

estudios recientes evaluaron los efectos globales de la luz sobre el *splicing* en *Physcomitrella patens* y plántulas etioladas de *Arabidospsis thaliana* usando RNA-Seq [81,84]. Estos estudios reportaron cientos de genes que presentaban eventos de *splicing* alternativo regulados por luz, muchos de los cuales fueron asociados con genes que codifican factores de splicing y componentes de las vías de señalizacion. Se encontraron resultados similares en un estudio reciente que caracteriza los patrones de *splicing* alternativo en *Arabidopsis thaliana* utilizando un panel de RT-PCR de alta resolución, en transiciones de exposiciones prolongadas de luz oscuridad [80]. Se mostró que aproximadamente el 50 % de los eventos evaluados se localizaban en genes de factores de *splicing*.

En nuestro trabajo, aportamos conocimiento sobre los efectos de la luz sobre el *splicing* alternativo en las plantas de *Arabidopsis thaliana*, llevando a cabo una caracterización global de los efectos de un breve pulso de luz dado en el medio de la noche. Este tratamiento simula el alargamiento del fotoperíodo. De acuerdo con estudios previos, encontramos que aunque el pulso de luz es breve, tiene un efecto significativo sobre los genes, afectando el *splicing* en muchos que codifican factores de *splicing* en sí mismos. Todos estos descubrimientos, revelan que la luz controla el desarrollo y crecimiento de las plantas de un modo significativo impactando sobre el *splicing* alternativo en genes de factores de *splicing* y componentes de las vías de señalización, además de los ya bien estudiados efectos sobre los niveles de ARNm y proteínas. Estos efectos de la luz sobre el *splicing* alternativo tienen lugar en especies distantes evolutivamente así como también en diferentes estados del desarrollo y en respuesta a exposiciones breves o prolongadas de luz, sugiriendo que existen mecanismos ancestrales por los que la luz regula los patrones/programas de expresión génica. De un modo interesante, tanto en plantas etioladas como en *Physcomitrella patens*, los efectos del tratamiento con un pulso de luz breve fueron modulados en gran medida aunque no exclusivamente por miembros de la familia de fotoreceptores [81, 84]. Por otro lado, los efectos de la exposición prolongada a las condiciones de luz u oscuridad de las plantas de *Arabidopsis thaliana* crecidas en luz parecen operar por mecanismos independientes de la familia de fitocromos y criptocromos más abundantes [80]. En cambio, parecerían estar mediados por circuitos de señales retrógradas que conectan la actividad fotosintética del cloroplasto con la regulación del *splicing* alternativo en el núcleo. En nuestro trabajo encontramos que los efectos de un breve pulso de luz sobre el *splicing* alternativo en SR30 fue tan fuerte en las dobles mutantes *phyAB* y *cry1;cry2* como en las plantas salvajes. Además observamos que el fuerte efecto sobre el *splicing* alternativo del pulso de

2 horas de luz roja no fue diferente sobre la quíntuple mutante de fitocromos y la salvaje. Aunque no podemos decir que los fotoreceptores sensoriales no estén implicados en este efecto, es claro que al menos en los mutantes evaluados en el trabajo, el efecto no operaría por esta vía tradicional de fotorecepción. Las diferencias en las principales vías que median la fototrasducción de luz y la regulación de los patrones de *splicing* en plantas etioladas y desetioladas pueden estar asociadas a la presencia de cloroplastos completamente desarrollados en las últimas, los cuales podrían ser requeridos para la generacion de una señal retrógrada que controle el *splicing* alternativo en el núcleo [80]. Más allá de los fotoreceptores que median el efecto de la luz sobre el *splicing* alternativo, una de las preguntas más importantes que resta ser contestada es cómo la luz regula, modula y controla el *splicing* alternativo. Tal como se menciona en la introducción, el *splicing* es catalizado por una maquinaria ribonucleoproteica dinámica denominada espliceosoma [23]. Los componentes centrales interaccionan con factores de *splicing* auxiliares, tales como las proteínas SR y hnRNPs para reconocer y seleccionar diferentes elementos *cis* promoviendo o inhibiendo el reclutamiento del espliceosoma a diferentes partículas donoras y aceptoras. Entonces, la regulación del *splicing* alternativo por la luz debería disparar cambios en los niveles de estas proteínas. Un trabajo reciente reportó que el factor de *splicing* de *Arabidopsis thaliana REGULATOR OF CHROMOSOME CONDENSATION 1 (RCC1)*, que es una proteína de tipo SR, es requerida para el desenlace normal de la fotomorfogénesis bajo condiciones de luz roja. [99]. Sin embargo, el mecanismo por el cual la luz dispara los cambios iniciales que resultarán en la alteración de los patrones de *splicing* no se han podido descifrar. Existen evidencias que el *splicing* alternativo también estaría modulado por modificaciones epigenéticas que modifican las velocidades de la elongación transcripcional dando tiempo a que puedan ser recluctadas más partículas del espliceosoma en los sitios débiles. Del mismo modo se ha reportado que las modificaciones de histonas pueden regular el *splicing* alternativo potenciando el reclutamiento de factores específicos de *splicing* [23]. Finalmente, cambios en la actividad y/o concentración de los factores de *splicing* centrales pueden influenciar la regulación del *splicing* alternativo, alterando probablemente la cinética de los pasos del ensamblado del espliceosoma [100].

Además de evaluar el rol de los factores *trans*, buscamos la presencia de secuencias *cis* que pudieran estar cumpliendo algún rol en la regulación de los eventos de *splicing* alternativo regulados por luz.

Con el fin de encontrar algún elemento regulatorio común analizamos las regiones flanqueantes a las junturas de los eventos detectados como diferencialmente usados. En particular, nos concentramos en el análisis de los 232 intrones que encontramos diferencialmente retenidos en respuesta a la luz. Llevamos a cabo la búsqueda de motivos enriquecidos en las secuencias flanqueantes y usando como control un número similar de intrones no retenidos, utilizando 2 estrategias: el algoritmo MEME [101] y la comparación en frecuencia de kmeros [1] desde 3 a 7 bases. Utilizamos diferentes longitudes de ventana para seleccionar las secuencias río arriba y río abajo del sitio 5' y 3' del intrón, así como también, clasificando los intrones en aquellos que subían o bajaban su porcentaje de retención con el pulso de luz. Con ninguna estrategia pudimos encontrar algún elemento *cis* enriquecido en las secuencias bajo análisis.

Finalmente, mientras que es claro que la luz ejerce un fuerte efecto sobre la regulación del *splicing* alternativo en las plantas desetioladas a lo largo de su desarrollo, nuestro conocimiento sobre estos efectos y sus consecuencias ha sido poco estudiada. En plantas etioladas, se ha demostrado que la luz promueve la acumulación de una variante de *splicing* del gen *SUPRESSOR OF PHY A (SPA)* que codifica para una proteína truncada, que actúa de un modo dominante negativo promoviendo la desetiolación y el desarrollo fotomorfogénico [81]. En plantas desetioladas, en condiciones de baja luz, se promueve la reducción de los niveles de la isoforma larga de *SR31* y esta regulación es importante para el crecimiento y desarrollo adecuado de las plantas expuestas a estas condiciones ambientales [80]. En nuestro estudio analizamos el efecto de un breve pulso de luz dado en el medio de la noche, con el fin de identificar los potenciales mecanismos de regulación post-transcripcional que controlan al reloj y a la transición floral. Un estudio similar del transcriptoma realizado con microarreglos permitió identificar una familia de genes nuevos cuya expresion es fuertemente inducida por la luz y regulan la función del reloj y la floración [102]. En este estudio encontramos que un pulso de luz dado en el medio de la noche afectó los niveles de *splicing* alternativo en varios genes del reloj tales como *JMJD5, RVE8/LCL5, LHY, CKB3* y *TIC*. De manera sorprendente, muchos de estos genes sólo son regulados por luz a niveles post-transcripcionales y no muestran cambios en los niveles de ARNm. Queda por elucidar como modulan la transición floral y el reloj circadiano los eventos regulados por luz.

[1] En genómica computacional, el término kmero refiere a todas las posibles susecuencias de longitud k

Capítulo 5

Otros ejemplos del uso de ASpli

El objetivo principal de este capítulo es brindar ejemplos de cómo se establecieron estrategias de análisis bioinformáticos para elaborar hipótesis y conclusiones a partir de los datos de RNA-Seq. En cada uno de los trabajos planteamos estrategias diferentes según la pregunta que subyacía al experimento. El enfoque evento céntrico de **ASpli** y la posibilidad de cuantificar los cambios del porcentaje de retención de intrones fueron claves para nuestra estrategia de caracterización global del *splicing* alternativo.

5.1. Análisis del efecto de mutantes de genes centrales del espliceosoma sobre el reloj circadiano y el *splicing* alternativo.

Los relojes circadianos son mecanismos auto-regulatorios endógenos que controlan la periodicidad en múltiples procesos biológicos. Estos proporcionan a los organismos una ventaja adaptativa que permite sincronizar distintos eventos fisiológicos y del desarrollo al momento más adecuado del día. La regulación del reloj circadiano ocurre principalmente mediante circuitos de retroalimentación transcripcional [103]. Sin embargo, estudios recientes muestra que se requiere de mecanismos post-transcripcionales para el correcto funcionamiento del reloj. En plantas, varios de los genes centrales del reloj sufren *splicing* alternativo, algunos dependientes de la temperatura [103].

Cada vez más evidencias indican que el *splicing* alternativo está regulado no solo por factores

auxiliares sino también por cambios en los niveles o actividades de componentes centrales del espliceosoma mismo o de las proteínas que modulan su ensamblado.

5.1.1. El rol de los genes LSM en la regulación de los ritmos circadianos

El primer trabajo publicado del laboratorio donde utilizamos **ASpli** como protocolo fue en el análisis del rol de los genes LSM en la regulación de los ritmos circadianos y el *splicing* alternativo [64].

En este trabajo, junto a la Dra. Pérez-Santangelo caracterizamos el comportamiento circadiano de mutantes de factores de *splicing* cuyos transcriptos están regulados por el reloj. Se encontró que mutaciones en genes de los complejos LSM5 y LSM4, que codifican para el complejo espliceosomal U6 snRNP, alteran el período de los ritmos circadianos [64].

Con el objetivo de caracterizar la respuesta transcripcional en las mutantes, se secuenciaron muestras de *Arabidopsis thaliana* salvajes y de las mutantes LSM4 (*lsm4-1*) y LSM5 (*sad1/lsm5*). Se utilizó la plataforma de Illumina HiSeq 1500. Se secuenciaron bibliotecas de ADNc de tres réplicas biológicas independientes cada una de las mutantes y sus accesiones salvajes correspondientes. Se obtuvieron alrededor de 16 millones de lecturas *paired-end* de 100 pares de bases de longitud por muestra. Las lecturas fueron mapeadas al genoma de *Arabidopsis thaliana* TAIR10 [91] usando TopHat v2.0.9 [34] con los parámetros prestablecidos, a excepción del máximo del intrón que fue establecido en 5000 pb. A partir de los archivos BAM, **ASpli** fue utilizado para evaluar los cambios en la expresión de genes, en eventos de *splicing* alternativos anotados y un análisis profundo del *splicing* en todos los intrones presentes en genes expresados por sobre un umbral.

Resultados del análisis de la expresión diferencial de genes

Se identificaron cambios en la expresión de 1845 genes en la mutante *sad1/lsm5* y 6578 en la mutante *lsm4-1* y, 759 estaban compartidos en ambas mutantes de un total de 21806 genes que se consideraron expresados, una vez que aplicamos los filtros correspondientes. Entre estos se buscó identificar genes del oscilador central o genes controlados por vías de salidas del reloj, tales como genes vinculados al control del tiempo de floración y respuestas al desarrollo regulados por la luz. Con respecto al análisis de enriquecimiento ontológico de los genes alterados en las mutantes, se

encontró un enriquecimiento en genes implicados principalmente en regulación de la transcripción y respuestas de estrés.

Resultados del análisis de *splicing* en los mutantes *sad1/lsm5* y *lsm4*

Luego se evaluaron las alteraciones en eventos de *splicing* alternativo anotados en ambos mutantes. Se evaluaron 6567 eventos alternativos anotados correspondientes a genes expresados sobre un umbral mínimo en los mutantes y sus correspondientes plantas salvajes. De estos, 363 eventos estaban alterados en la mutante *lsm4-1* y 57 en la mutante *sad1/lsm5*, con 24 eventos alternativos igualmente cambiados en ambas mutantes, lo que representa una superposición 7,6 veces mayor de lo esperado por azar ($p<0.0001$, test de Fisher). En la figura se muestran los gráficos de cobertura de dos ejemplos representativos de alteraciones en el *splicing* alternativo observados en las mutantes. En plantas salvajes, el evento de *splicing* alternativo más abundante fue asociado con la selección alternativa del sitio de reconocimiento de *splicing* 3' (3'Alt: 33 %), seguido por la retención de intrón (RI: 32 %), selección alternativa del sitio de reconocimiento de *splicing* 5' (5'Alt: 22 %) y salteo de exón (SE: 13 %) (ver Figura 5.1). Entre los eventos de *splicing* alternativo afectados por las mutantes *lsm4-1/lsm5*, se encontró un aumento en la proporción de eventos de retención de intrón, lo cual es consistente con lo reportado previamente para la mutante *sad1/lsm5* [104, 105].

Por otro lado también se evaluó el *splicing* de todos los intrones presentes en los genes expresados sobre un umbral mínimo para determinar el impacto sobre el proceso de *splicing* en sí mismo, y no sólo sobre el *splicing* alternativo. En los 87241 intrones analizados, se detectaron alteraciones en el *splicing* de 553 y 4354 intrones en *sad1/lsm5* y *lsm4-1* respectivamente y 212 en común, lo que representa una superposición 7.7 veces mayor que lo esperado por azar ($p<0.001$ test de Fisher, Figura 5.1 D). En este caso se muestran ejemplos representativos de alteraciones en el *splicing* en intrones observados en las mutantes. En un análisis más profundo se observó que la proporción de eventos de retención de intrón incrementada entre las mutantes *lsm* era el doble para eventos de retención anotados como alternativos que para intrones considerados constitutivos (o al menos no anotados como alternativos) Ver Figura 5.1 F. Esto contribuye a pensar que cambios en los componentes centrales del espliceosoma podrían jugar un rol regulatorio en el control del *splicing*

alternativo asociado a vías de señalización específicas.

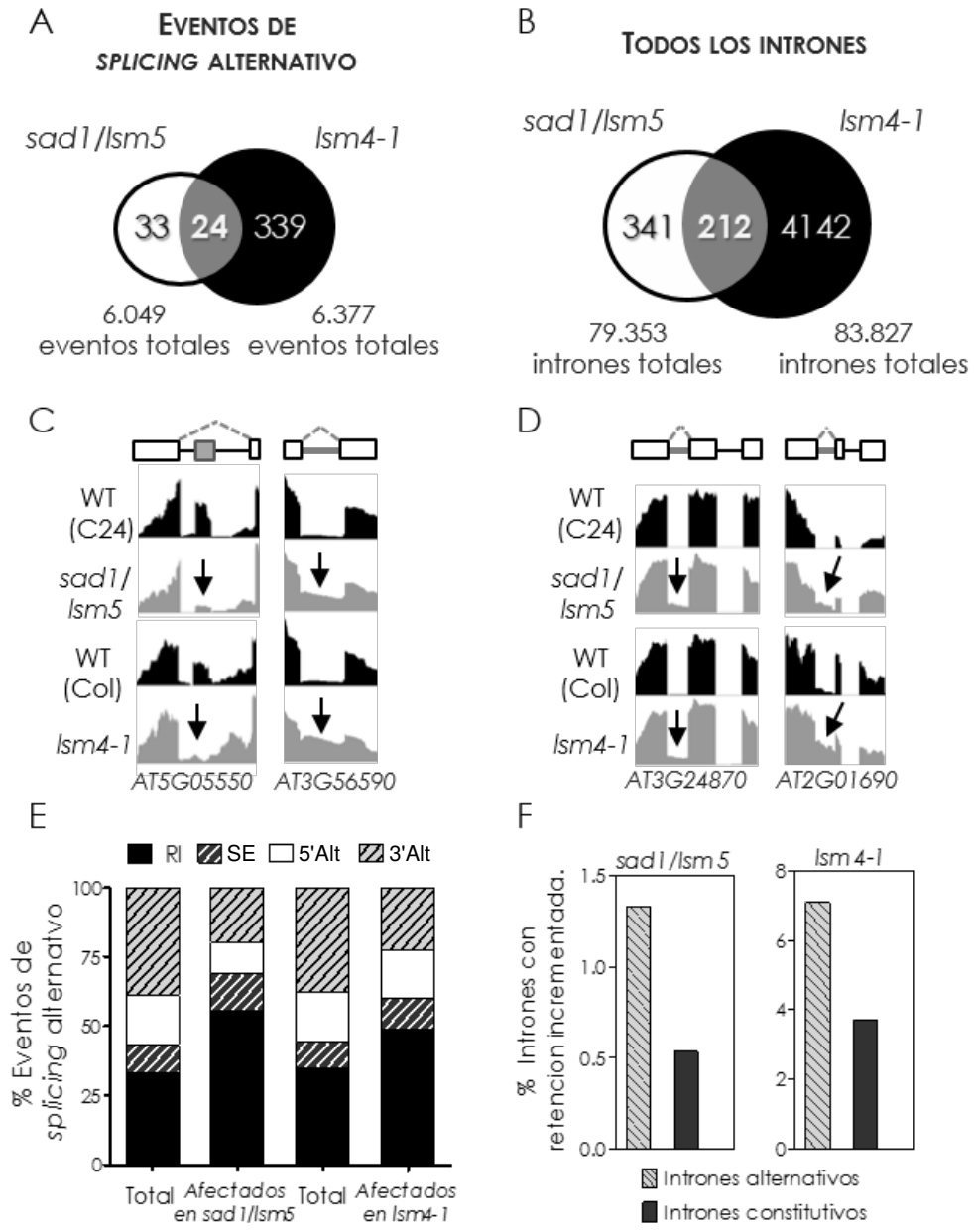


Figura 5.1: Análisis a nivel global mediante RNA-Seq del *splicing* alternativo y constutivo en las mutantes *sad1/lsm5* y *lsm4-1*

5.1.2. El rol de *GEMIN2*, un factor modulador del ensamblado del espliceosoma, los ritmos circadianos y la tolerancia al frío.

En mamíferos, el ensamblado del espliceosoma es regulado por el complejo SMN (del inglés *Survival of Motor Neuron*). En este trabajo en colaboración con el Dr. Gustavo Schlaen, demostramos que las mutantes de *Arabidopsis thaliana* para el gen *GEMIN2*, único componente del complejo

SMN conservado desde las levaduras hasta humanos, presenta floración temprana y período más corto en múltiples ritmos circadianos. Además, las mutantes tienen afectado un subgrupo específico de eventos de *splicing* alternativo, incluyendo algunos que involucran genes del reloj. Llamativamente varios de estos eventos de *splicing* alternativo se observan en plantas salvajes en respuesta a bajas temperaturas. Cabe destacar, que en dichas condiciones de baja temperatura, la supervivencia de los mutantes de *GEMIN2* se ve seriamente comprometida. Todo lo anterior, sumado a que la expresión de *GEMIN2* se induce a bajas temperaturas, sugiere que este componente formaría parte de un mecanismo compensatorio que amortigua el efecto de oscilaciones en la temperatura sobre el *splicing* alternativo, asegurando una apropiada adaptación de los procesos fisiológicos y de desarrollo.

En este trabajo, con el objetivo de caracterizar la respuesta transcripcional en las mutantes, se secuenciaron muestras de *Arabidopsis thaliana* ecotipo Columbia-0 y de las mutantes *gemin2-1*. Los experimentos de secuenciación se hicieron en 2 etapas. En una primera etapa, las muestras se secuenciaron con el protocolo *single-end* en una plataforma Illumina GAIIx y en la segunda etapa se utilizó la plataforma de Illumina HiSeq 1500, con el protocolo *paired-end*.

En todos los casos, las lecturas fueron mapeadas al genoma de *Arabidopsis thaliana* TAIR10 [91] usando TopHat v2.6 [34] con los parámetros prestablecidos, a excepción del máximo del intrón que fue establecido en 5000 pares de bases. A partir de los archivos BAM, **ASpli** fue utilizado para evaluar los cambios en la expresión de genes, en eventos de *splicing* alternativos anotados y un análisis profundo del *splicing* en todos los intrones presentes en genes expresados por sobre un umbral.

GEMIN2 y el *splicing* alternativo

El ensamblado del anillo heptamérico de proteínas Sm sobre los *snARNs* es estrictamente dependiente del complejo SMN. Dado que *Arabidopsis thaliana* cuenta con muy pocos homólogos de los componentes del complejo SMN de humanos, se esperaría que la mutante *gemin2-1* tuviera una deficiencia en los niveles de *snARNs*, ya que los *snARNs* que no se ensamblan en *snRNPs* son inestables y se degradan. En este trabajo, se demostró qe la mutación de *GEMIN2* causa una severa alteración de los *snARNs*. En particular U1 y U6 se encuentran sustancialmente disminuídos mientras que los demás *snARNs* están aumentados. Entonces, debido al impacto que tiene *GEMIN2*

sobre los niveles de *snRNPs*, los componentes principales del espliceosoma, es previsible su potencial impacto sobre el *splicing* (Figura 5.2)

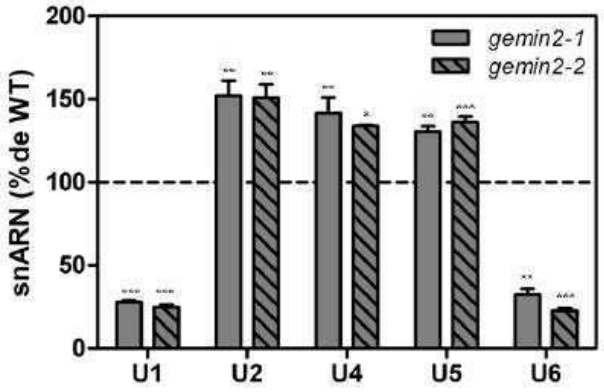


Figura 5.2: La mutación de *GEMIN2* causa una severa alteración en los *snRNPs*. Los niveles de *snRNAs* de los mutantes *gemin2-1* y *gemin2-2* se muestran como % de los valores de plantas salvajes

Tras el procesamiento de los datos se encontraron una gran cantidad de eventos de *splicing* alternativo afectados en la mutante *gemin2-1*, algunos de los cuales fueron confirmados por RT-PCR. El análisis estadístico de los resultados obtenidos por RNA-Seq muestra que los eventos alterados en la mutante se encuentran enriquecidos principalmente en retención de intrón y salteo de exón. Por otra parte, llamativamente, el *splicing* constitutivo permanece mayoritariamente intacto, evidenciando que aunque están reducidos los niveles de snRNAs U1 y U6 observados en las plantas mutantes *gemin2-1*, los mismos son suficientes para sostener el *splicing* constitutivo.

GEMIN2 y la sensibilidad al frío

En *Arabidopsis thaliana* , múltiples genes centrales del reloj sufren *splicing* alternativo en forma temperatura dependiente. Por otro lado, se observó que varios de los eventos afectados en *gemin2-1* a temperatura ambiente, son afectados en plantas salvajes por exposición a bajas temperaturas no congelantes (10º). Para evaluar cuan extensiva es la semejanza en el patrón de splicing de la mutante *gemin2-1* y plántulas salvajes expuestas al frío, y para estudiar a nivel global el efecto del frío sobre el splicing en plantas salvajes y la mutante *gemin2-1*, se realizó un experimento tratando a plantas salvajes y mutantes *gemin2-1* a 10ºC por 1 o 24 horas y plantas control que permanecieron a 22ºC.

Luego del procesamiento de los datos se obtuvieron los eventos de *splicing* alternativo anotados y los intrones que cambian su señal en la mutante respecto del genotipo salvaje o en plantas

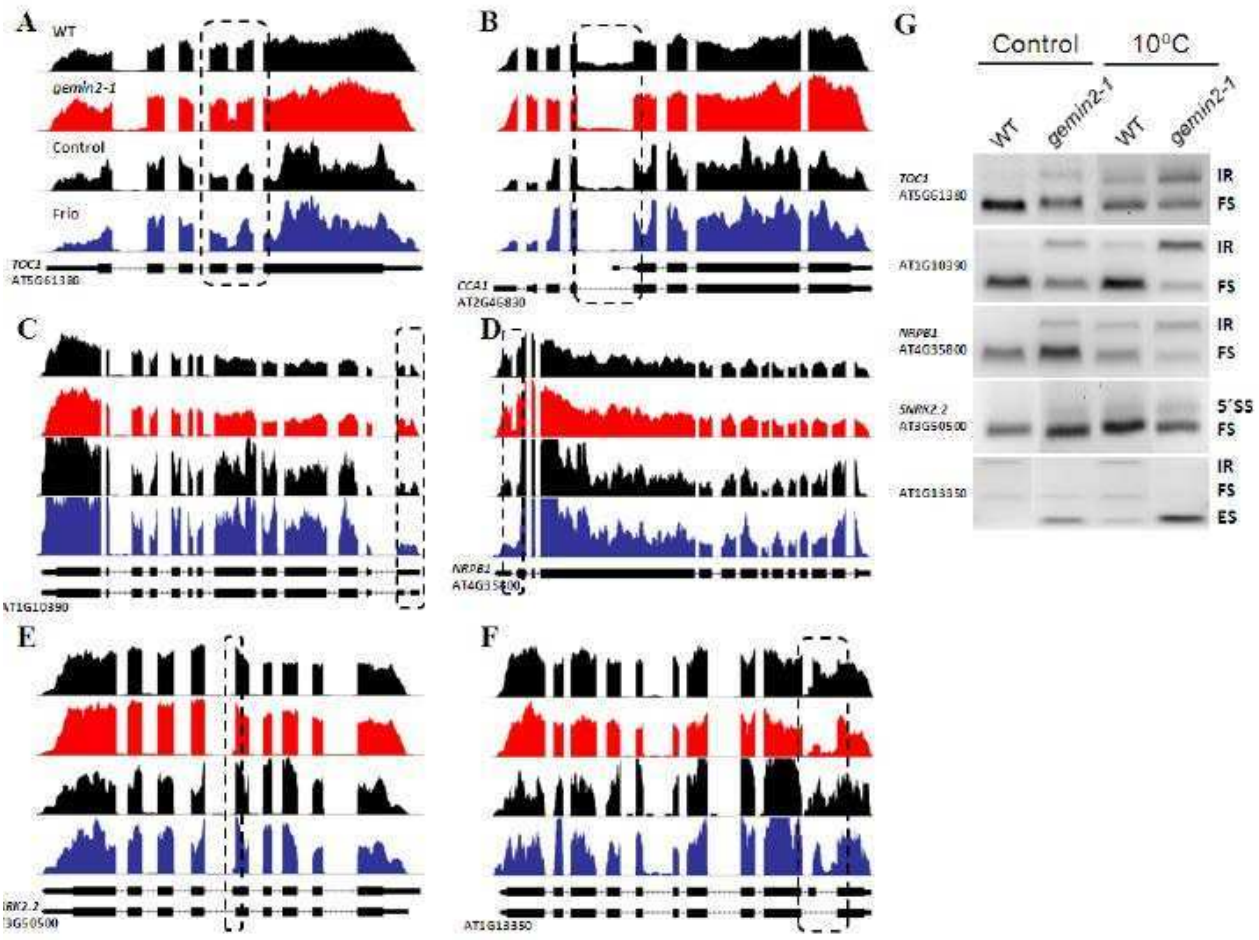


Figura 5.3: La mutante *gemin2-1* y las plántulas salvajes expuestas al frío presentan un patrón de *splicing* similar en varios genes. A-F) Gráficos de cobertura de algunos de los genes con *splicing* alternativo alterado en forma similar en la mutante *gemin2-1* y en plantas salvajes expuestas a bajas temperaturas. Las líneas punteadas encuadran la región donde ocurre el evento. Debajo de cada figura se muestran las isoformas anotadas para el gen correspondiente.G) Confirmación de algunos eventos por RT-PCR convencional. **10°C** indica plántulas tratadas con 12 horas a 10°C, mientras que **control** corresponde a plántulas que permanecieron a 22°C. RI, retención de intrón; 5'SS, sitio de splicing 5' alternativo; SE, salteo de exón; FS, mensajero canónico o constitutivo.

expuestas a 1 o 24 horas al frío respecto al control. En primer lugar se evaluó la superposición entre los eventos alterados en la mutante a 22°C y los que cambian con el frío en plantas salvajes. Se observó una superposición significativamente mayor que la esperada por azar y además, que los cambios ocurrían en el mismo sentido. Es decir, que la mayoría de los eventos en común, la inclusión de una región particular del gen en el ARNm maduro aumenta en ambas condiciones o disminuye en ambas condiciones, mientras que los eventos que aumentan en una condición y disminuyen en la otra son menos que los esperados por azar.

El análisis global del *splicing* en plantas salvajes y mutantes expuestas o no al frío demostró qe esta semejanza es bastante más amplia y que las plantas mutantes a temperatura control repro-

ducen en gran medida el fenotipo molecular de plantas salvajes expuestas a bajas temperaturas, en lo que respecta al *splicing* alternativo.

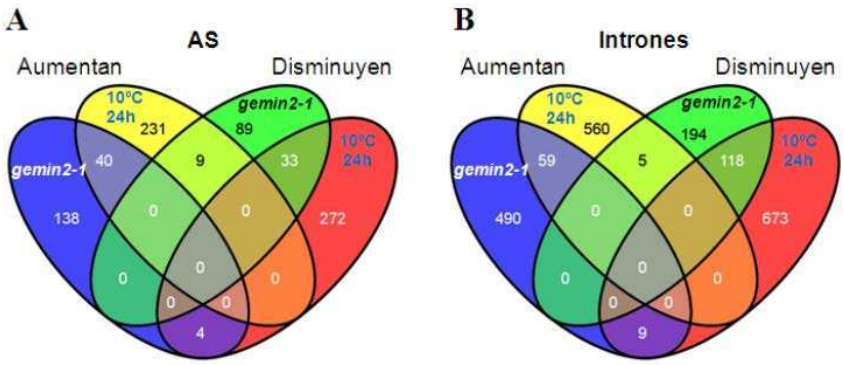


Figura 5.4: La deficiencia en GEMIN2 imita parcialmente el patrón de *splicing* alternativo de plantas expuestas a bajas temperaturas. Diagramas de Venn indicando aquellos eventos de *splicing* alternativo y todos los intrones compartidos entre las plantas mutantes de *gemin2-1* y las plantas salvajes a 1h y 2h de frío.

Estudios recientes sugieren que en condiciones ambientales adversas el *splicing* alternativo es uno de los mecanismos que regula la expresión de genes involucrados en la respuesta y tolerancia de la planta al estrés. Resulta llamativo que los mutantes *gemin2-1* reproducen al menos en parte, el *splicing* alternativo observado en plantas salvajes en respuesta al frío y aún así presentan una sensibilidad pronunciada en dicha condición. Así mismo, resultó interesante que la expresión de *GEMIN2* se indujera por el frío en plantas salvajes.

A partir de allí se planteó la hipótesis de que el conjunto de eventos compartidos por ambas condiciones (frío en plantas salvajes y mutación en *gemin2-1* a temperatura ambiente), se deba al efecto negativo de bajas temperaturas y/o a la mutación de *gemin2-1* sobre el *splicing* y no como parte de una respuesta adaptativa. En ese caso, el rol de la inducción de *GEMIN2* por bajas temperaturas sería el de aminorar o amortiguar el efecto negativo de bajas temperaturas sobre el *splicing*, restableciendo la homeostasis de un componente sensible al estrés o que se vuelva limitante al ser requerido en mayor cantidad durante el mismo. Los defectos observados en el *splicing* serían entonces la consecuencia de ese desajuste.

Para evaluar esta hipótesis se analizó el efecto del frío sobre el *splicing*, en plántulas salvajes y mutantes *gemin2-1*. Los datos correspondientes a 1 hora en frío reflejarían la respuesta inmediata de la planta al frío y/o los defectos en el *splicing* ocasionados por dicho estrés, mientras que a las 24 horas se representaría el estado luego de un día de adaptación.

Los eventos luego fueron agrupados en diferentes grupos según su perfil temporal en las plantas

salvajes y mutantes expuestas al frío. Llamativamente, los 144 eventos que aumentan en ambas condiciones se agruparon mayoritariamente en un grupo que incluye 77 eventos. En plantas salvajes, los mismos aumentan a la hora y se mantienen a las 24 horas, pero en la mutante aumentan a la hora y notablemente más a las 24 horas. (Figura 5.5).

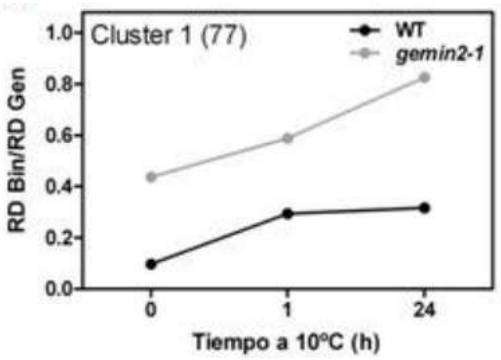


Figura 5.5: Efecto del frío sobre los eventos de *splicing* afectados simultáneamente en las plantas mutantes a temperatura control y en las plantas salvajes expuestas al frío. A partir de las tablas de densidad de lecturas se aplicó un algoritmo de agrupamiento que reveló la existencia de interesantes patrones de comportamiento entre los eventos alterados en ambas condiciones a 1h y 24h de aplicar el tratamiento de frío. Se ejemplifican los eventos que aumentan en ambas condiciones. El número de genes del grupo se indica entre paréntesis.

Estos resultados apoyan la idea de que existe un mecanismo de compensación, dependiente de *GEMIN2*, que impide un aumento exagerado en los niveles de un grupo de alteraciones en el *splicing* especialmente sensibles al frío. En las mutantes, que carecen de este sistema de amortiguación, el grupo de alteraciones aumenta abruptamente (Figuras 5.5).

5.2. Análisis comparativo de los efectos de las arginin-metiltransferasas PRMT4 y PRMT5 sobre el transcriptoma de *Arabidopsis thaliana*

La metilación de argininas (R) es una modificación post-traduccional muy importante que regula una gran variedad de procesos celulares en todos los eucariotas, tales como la transcripción, el procesamiento del ARN, la transducción de señales y la reparación de ADN, entre otros. Esta modificación es catalizada por una familia de enzimas llamadas PROTEIN ARGININE METHYLTRANSFERASES (PRMTs). Estas se clasifican como PRMTs de tipo I o de tipo II dependiendo de la posición que adopta el metilo en el grupo guanidino de la arginina metilada. En colaboración

con el Dr. Esteban Hernando, se realizó una comparación fisiológica y molecular de las PRMTs mejor caracterizadas en plantas, la PRMT de tipo I (PRMT4) conocida como CARM1 en mamíferos, y la PRMT de tipo II (PRMT5). Empleando como modelo *Arabidopsis thaliana* se observó que mutantes de expresión nula de dichos genes exhibían alteraciones similares en procesos del desarrollo tales como el tiempo de la floración, la fotomorfogénesis y la respuesta al estrés salino, mientras que sólo las mutantes *prmt5* presentaron alteraciones en los ritmos circadianos.

Con el objetivo de caracterizar la respuesta transcripcional en las mutantes, se secuenciaron muestras de *Arabidopsis thaliana* ecotipo Columbia-0 y de las mutantes *prmt5* y *prmt4a;4b* usando la tecnología de *pair-end* de la plataforma Illumina HiSeq 1500. Las lecturas obtenidas fueron mapeadas en el genoma de *Arabidopsis thaliana* TAIR10 usando TopHat v2.0.9 con los parámetros preestablecidos, a excepción de largo máximo de intrón que fue establecido en 5000 pares de bases. Los archivos BAM fueron utilizados para el análisis con **ASpli**.

A partir de los análisis sobre el RNA-Seq realizados con **ASpli** se encontró que la expresión y el procesamiento del pre-ARNm de muchos genes estaban igualmente afectados por PRMT4 y PRMT5. Además, se demostró que PRMT4 y PRMT5 corregulan la expresión y el *splicing* de genes centrales en el control de la transcripción, el procesamiento del ARN, las respuestas a la luz, la floración y la tolerancia a estrés tanto abiótico como biótico, siendo estos genes los candidatos a mediar las alteraciones fisiológicas observadas en las mutantes prmt4a;4b y prmt5. También, encontramos que las PRMTs de tipo I y tipo II más importantes, PRMT4 y PRMT5 respectivamente, controlan mayoritariamente los mismos procesos fisiológicos mediante la regulación de la expresión y el *splicing* de los mismos genes.

Análisis global del impacto de PRMT5 y PRMT4 en la expresión génica

Las mutantes *prmt5* exhibieron 2604 genes sobreexpresados y 3075 genes subexpresados, mientras que las mutantes *prmt4a;4b* presentaron 2959 genes sobreexpresados y 2545 genes subexpresados, en ambos casos relativo a plantas de genotipo salvaje. Se observó un comportamiento similar entre los genes diferencialmente expresados en ambas mutantes, observándose 1076 genes sobreexpresados y 1312 genes subexpresados en común. Por otro lado, de un total de 5679 y 5504 genes diferencialmente expresados en prmt5 y prmt4a;4b, respectivamente, solo 64 genes se comportaron

en forma antagónica. Se categorizó en grupos funcionales basados en ontologías génicas a aquellos genes que estaban diferencialmente expresados y se comportaban igual en ambas mutantes. Se seleccionaron quince categorías funcionales de interés y para cada una se determinó un factor de representación (FR) (Ver Figura 5.6).

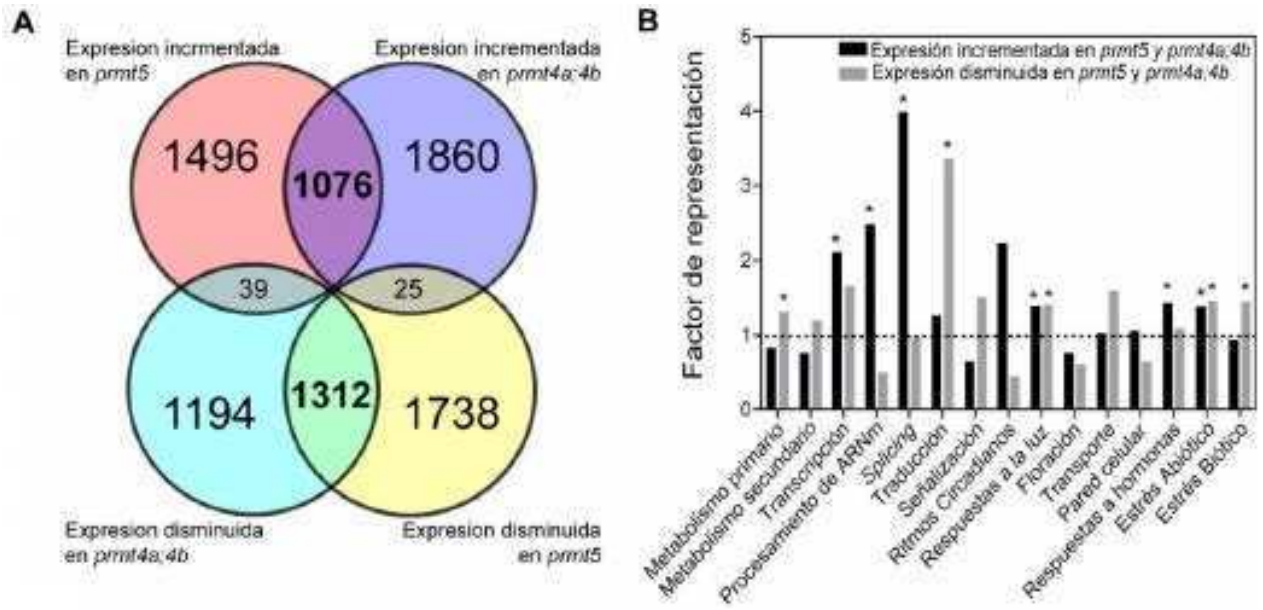


Figura 5.6: Análisis global del impacto de PRMT5 y PRMT4 en la expresión génica. (A) Superposición entre los genes diferencialmente expresados en las mutantes *prmt5* y *prmt4a;4b* (B) Factor de Representación (FR) en las categorías funcionales analizadas de genes corregulados por PRMT5 y PRMT4

Análisis global del impacto de PRMT5 y PRMT4 en el *splicing* alternativo

Se encontraron 1137 eventos de *splicing* alternativo alterados significativamente en las mutantes *prmt5* y 1290 en las mutantes *prmt4a;4b*, los cuales representan el 19 % y 21 %, respectivamente, de todos los eventos de *splicing* evaluados. Entre los eventos de *splicing* alterados identificados, 261 exhibieron un incremento en su inclusión y 352 exhibieron una disminución en su inclusión, simultáneamente en ambas mutantes y sólo 6 eventos exhibieron un comportamiento antagónico. Los eventos de *splicing* alternativo afectados en común fueron clasificados en categorías ontológicas funcionales y se evaluaron enriquecimientos. Se encontró una sobrerrepresentación significativa para las categorías de metabolismo primario, respuestas a la luz, respuestas a hormonas y estrés abiótico. A su vez, encontramos alteraciones en el procesamiento del pre-ARNm de genes específicos asociados a la tolerancia al estrés salino, al procesamiento de ARN, al *splicing* y a la regulación del tiempo de la floración. Es interesante destacar, que en las mutantes no se observaron diferencias

significativas en la distribución relativa de las distintas categorías de eventos de *splicing* alternativo. Ver Figura 5.7.

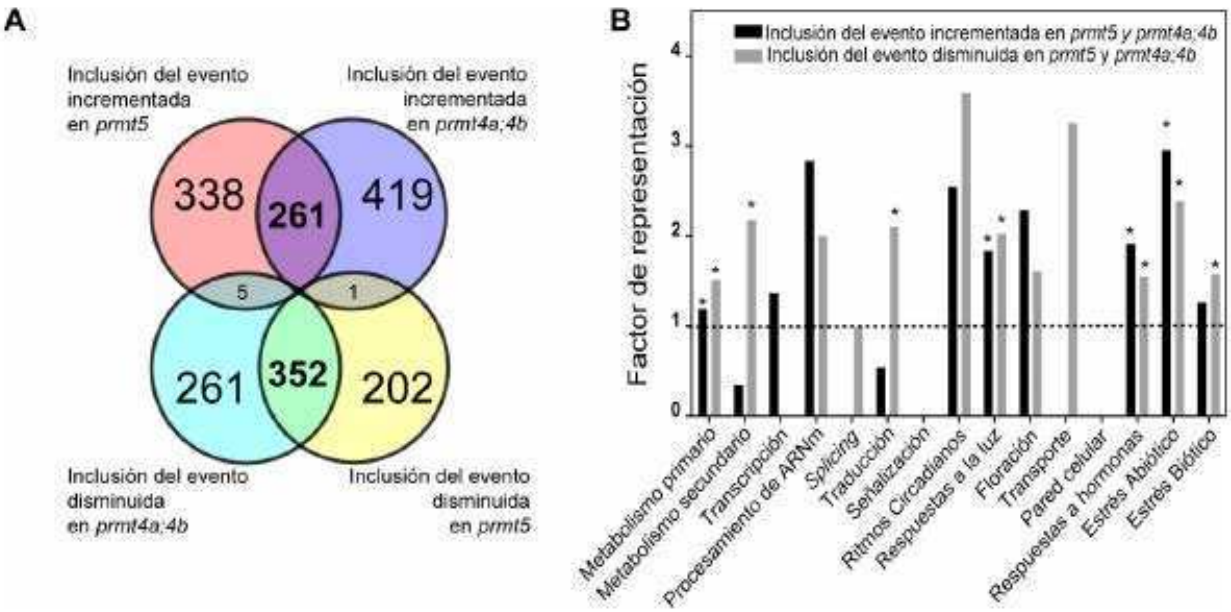


Figura 5.7: Análisis global del impacto de PRMT5 y PRMT4 en el *splicing* alternativo (A) Superposición entre los eventos de *splicing* alternativo con inclusión diferencial en las mutantes *prmt5* y *prmt4a;4b* (B) Factor de Representación (FR) en las categorías funcionales analizadas de eventos de *splicing* alternativo corregulados por PRMT5 y PRMT4.

Análisis global de los efectos de PRMT5 y PRMT4 en el *splicing* constitutivo

Con el fin de estudiar el impacto de PRMT5 y PRMT4 en el *splicing* constitutivo, se analizó el impacto de las mutaciones en estos genes sobre el procesamiento de todos los intrones no anotados como alternativos. Se encontraron 2506 intrones cuya retención se incrementó en *prmt5* y 1143 cuya retención aumentó en mutantes *prmt4a;4b*, los cuales representan el 3,1 % y el 1,4 %, respectivamente, de todos los intrones estudiados. Es interesante destacar que las alteraciones observadas en intrones "constitutivos" fueron menores que las identificadas en los intrones anotados como alternativos, para los cuales se vieron alterados el 17,7 % en las mutantes *prmt5* y el 17,8 % en mutantes *prmt4a;4b*, encontrando numerosos intrones afectados en ambas mutantes, manteniendo la tendencia observada en el análisis global de la expresión de genes y del *splicing*. Aquellos eventos de retención de intrón en común en ambas mutantes fueron analizados empleando ontologías génicas. (Figura 5.8).

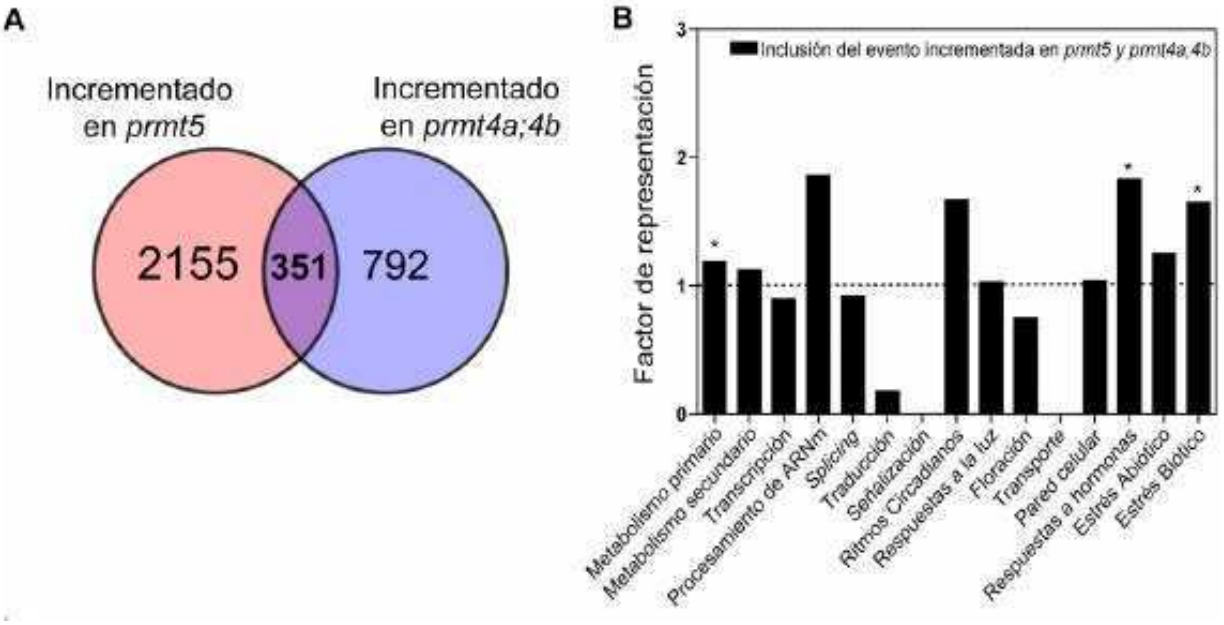


Figura 5.8: Análisis global de los efectos de PRMT5 y PRMT4 en el *splicing* constitutivo. (A) Superposición entre intrones no alternativos con retención incrementada en ambas mutantes. (B) Factor de Representación (FR) en las categorías funcionales analizadas de intrones cuya retención se vio incrementada simultáneamente en prmt5 y prmt4a;4b.

Análisis de las secuencias donoras 5' de splicing

Tal como se detalla en la introducción, en todo análisis de *splicing* con datos de RNA-Seq, se intenta encontrar patrones de regulación comunes a los eventos que se encuentran alterados. En este trabajo, podemos mostrar un ejemplo de lo que eso significa. **ASpli** es un paquete que permite fácilmente integrar la información de los eventos seleccionados con otros paquetes de R, por ejemplo, para extraer las secuencias del genoma de referencia y analizar secuencias consensos o descubrimiento de motivos sobre ellas. En este caso, para profundizar el análisis del rol de las PRMTs en el procesamiento de pre-ARNm, se estudiaron las secuencias de los sitios de *splicing* donores 5' de aquellos eventos de retención de intrón alterados en ambas mutantes y la secuencia donora 5' consenso de todos los intrones presentes en el genoma de *Arabidopsis thaliana*. Se encontró que las secuencias dadoras 5' de los eventos afectados en simultáneo en ambas mutantes, así como de eventos afectados solo en *prmt5*, exhibían una subrepresentación de los nucleótidos consenso A y G presentes en las posiciones -2 y -1 del sitio dador 5' consenso. Al contrario, las secuencias del sitio donor de *splicing* de los eventos alterados sólo en mutantes *prmt4a;4b* no presentaron diferencias significativas respecto de la secuencia consenso. Finalmente, vale mencionar que las secuencias de los sitios 3' aceptores de *splicing* no presentaron diferencias significativas respecto de la secuencia

consenso (Figura 5.9).

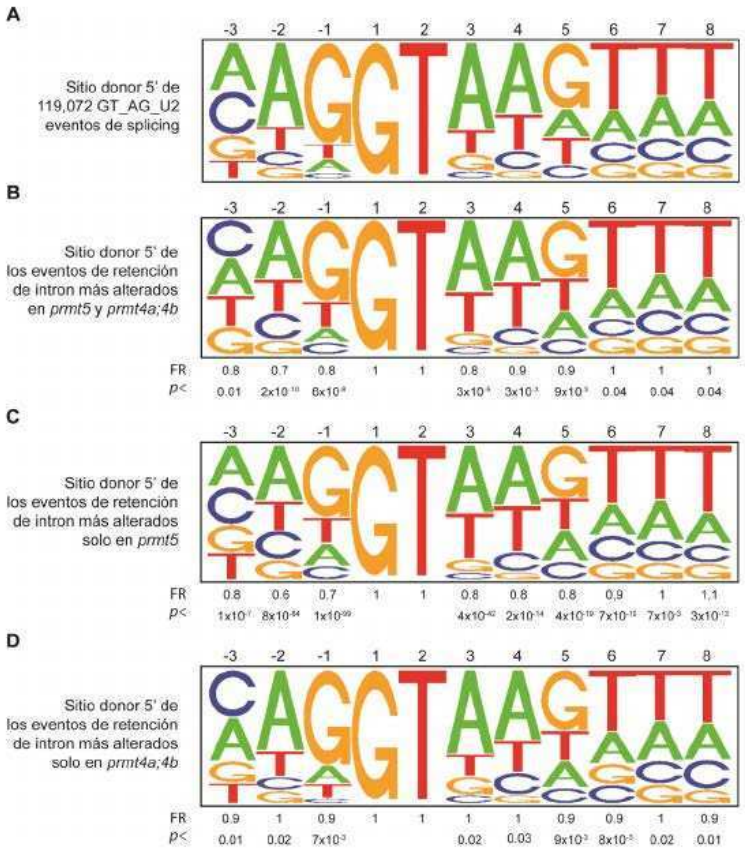


Figura 5.9: Análisis de las secuencias donoras 5' de *splicing*. Contenido de información por posición en el sitio de *splicing* donor 5' de (A) 119.072 intrones GT_AG_U2 de *Arabidopsis thaliana*, (B) los eventos de retención de intrón más alterados simultáneamente en ambas mutantes, (C) los eventos de retención de intrón más alterados sólo en *prmt5* y (D) los eventos de retención de intrón más alterados sólo en *prmt4a;4b*.

5.3. Ejemplo del uso de ASpli para caracterizar cambios en *splicing* en células humanas.

La proteína del virus de Dengue N55 se asocia al splicesoma y modula el *splicing*

Para finalizar, describiremos un breve análisis que realizamos sobre la remodelación del transcriptoma, en este caso de *Homo sapiens*, ante la infección del virus del Dengue.

La proteína N55 del virus del Dengue lleva a cabo múltiples funciones en el citoplasma de las células infectadas, permitiendo la replicación viral y contrarrestando las respuestas antivirales del

huésped. En colaboración con la Dra. Andrea Gamarnik, se ha demostrado una nueva función de la proteína N55 en el núcleo, interfieriendo en el mecanismo de *splicing*. En el trabajo se realizó un análisis global de proteómica de las células infectadas junto con estudios funcionales. Se encontró que la proteína se une al complejo del espliceosoma y modula el *splicing* endógeno así como también el *splicing* de un constructo exógeno. En particular, se mostró que la proteína N55 en el contexto de la infección viral, interactúa con los componentes centrales del U5 *snRNP*, CD2BP2 y DDX23, alterando las proporciones de inclusión/exclusión de los eventos de *splicing* alternativo así como alterando la abundancia de isoformas de factores antivirales conocidos. Utilizando nuestro método, se pudo caracterizar alteraciones globales en el *splicing*, observando un incremento en la retención de intrones luego de la infección del virus de Dengue y también un incremento en la replicación viral debido al silenciamiento de componentes específicos de U5. Diferentes estudios mecanísticos indicaron que la unión de N55 al espliceosoma reduce la eficiencia del procesado de pre-ARNm, independienteme de la actividad enzimática de N55.

Reducción de la eficiencia de *splicing* debido a la infeccion de Dengue

Se secuenciaron bibliotecas de ARN de 3 replicados biológicas de células control e infectadas luego de 24 y 36 horas del inicio de la infección. Las mismas fueron secuenciadas con la tecnología Illumina HiSeq 4000, protocolo *paired-end*. Las lecturas resultantes fueron alineadas contra el genoma de referencia de *Homo Sapiens* (UCSC, HG19 [106]). Se analizaron cerca de 175000 *bins* para cada muestra, de los cuales 26 % correspondieron a *bins* intrónicos y 74 % a *bins* exónicos. Los datos fueron filtrados para definir aquellos *bins* que habían sido utilizados diferencialmente con los criterios descriptos anteriormente. De esta manera, se encontraron cambios en 451 y 858 eventos a 24 y 36 horas luego de la infección. De un modo interesante, se observó un enriquecimiento de retención de intrón en los ARNs maduros, tanto a 24 como a 36 horas luego de la infección (Figura 5.10). Este resultado fue muy interesante para nosotros, porque a pesar que se lo considera poco frecuente en mamíferos, se ha reportado su ocurrencia en gran cantidad de células humanas [38]. En las células infectadas, la retención de intrón incrementó en un 70 % de los eventos que cambiaron a las 24 y 36 horas, indicando una reducción de la excisión del intrón en cientos de transcriptos, que podría estar asociado a una perturbación significativa en la catálisis del *splicing*.

Bins analizados

	24h	36h
bins exónicos	129763	128520
bins intrónicos	45677	45871

Eventos de *splicing* alterados (DENV)

	24hpi	36hpi
tipo de evento alterado	451	858
retención de intron RI	399	761
5'/3' Alt	18	43
salteo de exon SE	34	54

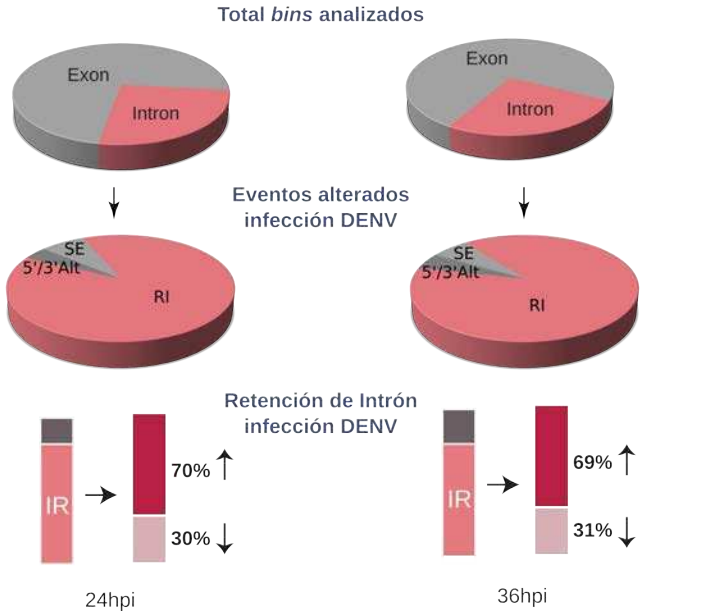


Figura 5.10: Disminución de la eficiencia de *splicing* en las células infectadas con Dengue. Datos del análisis de *splicing* en las células control e infectadas (DENV). En la tabla superior, se indican los totales de *bins* analizados y los eventos de *splicing* alternativos alterados para cada punto post infección (24/36 horas). El % de eventos alterados se muestra en los gráficos de tortas. En el panel inferior se detallan los cambios sólo para retención de intrón.

5.4. Conclusiones

A lo largo de este capítulo y el anterior, hemos ejemplificado el uso de **ASpli** para evaluar diferentres hipótesis relacionadas a la regulación del *splicing* alternativo y constitutivo. Los fragmentos de los trabajos elegidos como ejemplos representan diferentes estrategias de análisis que surgieron a partir de los resultados obtenidos con **ASpli**. La ejecución del protocolo es similar en todos los trabajos, sin embargo, cada uno mereció un análisis diferente según la pregunta biológica que se intentaba responder.

En el primer trabajo, se pretende ilustrar cómo a partir de un experimento de RNA-Seq se pudo caracterizar el efecto sobre la regulación del *splicing* alternativo y constitutivo de organismos mutantes de genes que son factores de *splicing* o reguladores del ensamblado del espliceosoma. La clave para este análisis fue el análisis de la variación de la proporción de eventos de *splicing* alternativo en las mutante y su comparación con los efectos sobre el *splicing* constitutivo.

En el segundo trabajo, se pone de manifiesto la versatilidad de los resultados de nuestro análisis

ya que permitieron indagar sobre el efecto de la mutante de *gemin2-1* en el tiempo y a diferentes temperaturas, por una estrategia de agrupamiento, donde los datos que se usaron para este análisis eran los conteos de lecturas a nivel de *bins*. Esto permitió realizar análisis de co-splicing e identificar redes de eventos quizás coregulados.

En el trabajo de caracterización del efecto de las mutantes de las proteínas PRMTs, se quiere destactar cómo a partir de un análisis global de *splicing* alternativo, se pudo caracterizar la secuencia dadora 5' de eventos alterados en los mutantes comparado a la secuencia dadora 5' consenso de todos los intrones de *Arabidopsis thaliana*.

Y finalmente, el trabajo sobre el efecto de la infección de Dengue en el *splicing* alternativo tiene varios puntos a destacar. En primer lugar, fue uno de los primeros trabajos que hicimos sobre un organismo diferente de *Arabidopsis thaliana*, validando nuestro método sin dificultades. En segundo lugar, gracias a nuestra estrategia de incluir los intrones, hemos contribuído a evidenciar que la retención de intrón es un fenómeno que ocurre también en humanos y que pudimos cuantificar una deficiencia en el procesamiento del ARN de manera global debido a la interferencia del virus del Dengue con la maquinaria de *splicing*.

Capítulo 6

Conclusiones generales

Esta tesis se presenta como el resultado de un intenso trabajo multidisciplinario y colaborativo. Se ha desarrollado en un marco científico invaluable, con la dedicación y el acompañamiento de investigadores de primer nivel tanto profesional como personal.

En primer lugar, se ha presentado la problemática de cuantificar cambios en *splicing* alternativo en general, y en particular en plantas. El desafío consistió en revisar las herramientas disponibles y luego, diseñar nuestra propia estrategia de análisis. La principal limitación que encontramos es que las herramientas disponibles, vigente en gran parte hasta el día de hoy, están orientadas a la cuantificación de eventos anotados principalmente relacionados a exones y que la estimación de los cambios se realizan por una sola estrategia. En nuestro caso, queríamos identificar, cambios a nivel global en el *splicing* incorporando conocimiento de nuevos eventos y sobre todo en aquellos relacionados a la retención de intrón.

Entonces, para cumplir parte de los objetivos propuestos para la tesis, en la primera etapa del doctorado establecimos un protocolo de trabajo bioinformático orientado al estudio del *splicing* alternativo en plantas para la tecnología de RNA-Seq. Una de las primeras tareas a resolver fue identificar las regiones alternativas dentro de los genes. Para ello, se propone la disección de los genes en unidades artificiales llamadas *bins*, cuyas coordenadas genómicas resultan de la proyección de todas las coordenadas de los exones de un dado gen sobre una línea horizontal imaginaria. En el caso de que un gen no presente transcriptos alternativos, las coordenadas de sus *bins* coinciden con las de sus exones. En cambio, en presencia de isoformas, las entidades tienen coordenadas nuevas, diferentes a las de los exones. Estos *bins* luego pueden ser clasificados según cómo se

hayan originado, siendo solo exónicos o intrónicos o alternativos cuando coinciden con regiones alternativas. Estos *bins* alternativos luego se clasifican según el evento de *splicing* que les dio origen como: dadores 5' o 3' alternativos (5'/3' Alt), salteo de exón (SE) o RI (retención de intrón). La siguiente tarea fue lograr incorporar los intrones a las tablas de conteos, porque ninguno de los métodos disponibles hasta el momento lo realizaban de una manera sencilla. En nuestro protocolo, incorporamos información de los intrones, de las regiones adyacentes a los mismos (denominadas E1I, IE2) y de las junturas. Finalmente, para cuantificar la ocurrencia de eventos de *splicing* alternativo, proponemos un método integrador, donde la evaluación de cambios en cada región alternativa se evalúa desde un enfoque tradicional **conteo céntrico** combinado con el enfoque de cuantificación usando las junturas.

El protocolo escrito en R, se materializó en un paquete de al que denominamos **ASpli** y que se hizo público en el repositorio de Bioconductor en junio de 2016. Esta estrategia permite que podamos compartir nuestro método de manera ordenada y prolija con el resto de la comunidad. Podemos decir, que con este desarrollo hemos aportado al mundo del *splicing* una herramienta de fácil uso y muy flexible para usuarios provenientes de las ciencias biológicas.

Es importante destacar, que el protocolo no refleja sólamente una nueva estrategia en el análisis de *splicing* alternativo, sino también, un conocimiento exhaustivo de las anotaciones de los organismos. Los resultados que se generan con **ASpli** son muy fáciles de analizar en cuanto a información genómica. Como ya mencionamos a lo largo de la tesis, uno de los objetivos más difíciles de lograr cuando estamos ante un experimento de RNA-Seq es descubrir qué patrones de regulación subyacen a los eventos de *splicing* alternativo que se modifican. En este punto, es importante destacar que la utilización de un enfoque **evento céntrico** permiten inferir con más claridad mecanismos de regulación compartidos entre eventos alternativos. Utilizando **ASpli** es muy senciilo por ejemplo, extraer información sobre el largo de los intrones, exones, contenido GC, enriquecimiento de motivos, análisis ontológicos, etcétera.

La estructura modular de **ASpli** permite la extensión del paquete de manera muy sencilla para, por ejemplo, incorporar nuevas estrategias de análisis. Uno de los objetivos a corto plazo es facilitar el análisis de experimentos de diseños más complejos, tales como medidas repetidas en el tiempo o interacción genotipo-tratamiento. Si bien hemos podido realizar este tipo de análisis a partir de la tabla de conteos o de PSI/PIR, tal como se muestra en el Capítulo 5, por el momento requieren

de un manejo del lenguaje R más elevado.

Toda el trabajo de tesis está atravesado por preguntas acerca de los mecanismos de regulación del *splicing* alternativo, inherentes a cada experimento que se realiza sobre un organismo, tejido, estadío, etcétera. Es por ello que **ASpli** fue utilizado en numerosos análisis de experimentos de secuenciación del laboratorio, de otros grupos en calidad de colaboración y de acceso público, obteniendo resultados provechosos para la elaboración de nuevas hipótesis y preguntas sobre la regulación del *splicing* alternativo. Damos cuenta de ello a lo largo de los capítulos 4 y 5. En el Capítulo 4, mostramos de una manera minuciosa el análisis global sobre el impacto del ambiente (en este caso la luz) sobre la regulación del *splicing* alternativo, en plantas. En el Capítulo 5, detallamos algunos ejemplos de cómo se vio afectado el *splicing* alternativo y constitutivo ante la mutación de genes que codifican para proteínas directamente relacionadas a la maquinaria de *splicing*, así como también mostramos un ejemplo de un análisis realizado sobre datos de RNA-Seq, pero en este caso de lineas celulares humanas infectadas con virus de Dengue. Este para destacar que fue nuestro primer trabajo publicado con datos en un organismo diferente a *Arabidospsis thaliana*. Además, porque aportamos evidencia de una disminución en la eficiencia de *splicing*, observado principalmente en la retención de intrones, fenómeno muy poco descripto en humanos hasta el día de hoy. Este tipo de análisis no se podría haber realizado con los paquetes clásicos que no permiten descrubrimiento de nuevos eventos o que sólo se limitan a analizar exones, como la mayoría de los *softwares* actuales, originalmente pensados para mamíferos.

Como conclusión, podemos decir que a lo largo del trabajo de esta tesis, hemos ilustrado cómo el desarrollo del paquete permitió echar luz sobre múltiples mecanismos de regulación del *splicing* alternativo, que escapan claramente al contenido de la tesis, pero que han colocado al laboratorio como uno de los referentes en el mundo en cuanto al análisis de datos de RNA-Seq y *splicing* alternativo.

Bibliografía

[1] E. J. Beckwith and M. J. Yanovsky, "Circadian regulation of gene expression: at the crossroads of transcriptional and post-transcriptional regulatory networks," Curr. Opin. Genet. Dev., vol. 27, pp. 35–42, Aug 2014.

[2] W. Gilbert, "Why genes in pieces?," Nature, vol. 271, p. 501, Feb 1978.

[3] R. E. Breitbart, A. Andreadis, and B. Nadal-Ginard, "Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes," Annu. Rev. Biochem., vol. 56, pp. 467–495, 1987.

[4] T. W. Nilsen and B. R. Graveley, "Expansion of the eukaryotic proteome by alternative splicing.," Nature, vol. 463, pp. 457–63, Jan. 2010.

[5] A. G. Matera and Z. Wang, "A day in the life of the spliceosome," Nat. Rev. Mol. Cell Biol., vol. 15, pp. 108–121, Feb 2014.

[6] D. Staiger and J. W. Brown, "Alternative splicing at the intersection of biological timing, development, and stress responses," Plant Cell, vol. 25, pp. 3640–3656, Oct 2013.

[7] R. Liu, A. E. Loraine, and J. A. Dickerson, "Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems," BMC Bioinformatics, vol. 15, p. 364, Dec 2014.

[8] Y. Marquez, J. W. Brown, C. Simpson, A. Barta, and M. Kalyna, "Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis," Genome Research, vol. 22, no. 6, 2012.

[9] A. S. Reddy, Y. Marquez, M. Kalyna, and A. Barta, "Complexity of the alternative splicing landscape in plants," Plant Cell, vol. 25, pp. 3657–3683, Oct 2013.

[10] T. S. Alioto, "U12DB: a database of orthologous U12-type spliceosomal introns," Nucleic Acids Res., vol. 35, pp. D110–115, Jan 2007.

[11] SIB Swiss Institute of Bioinformatics, "Prosite user manual." `http://prosite.expasy.org/prosuser.html`, Accedido noviembre de 2016.

[12] A. S. Reddy, M. F. Rogers, D. N. Richardson, M. Hamilton, and A. Ben-Hur, "Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements," Front Plant Sci, vol. 3, p. 18, 2012.

[13] G. S. Huh and R. O. Hynes, "Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element," Genes Dev., vol. 8, pp. 1561–1574, Jul 1994.

[14] A. J. McCullough and S. M. Berget, "G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection," Mol. Cell. Biol., vol. 17, pp. 4562–4571, Aug 1997.

[15] M. Y. Chou, J. G. Underwood, J. Nikolic, M. H. Luu, and D. L. Black, "Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing," Mol. Cell, vol. 5, pp. 949–957, Jun 2000.

[16] Z. Wang, M. E. Rolish, G. Yeo, V. Tung, M. Mawson, and C. B. Burge, "Systematic identification and analysis of exonic splicing silencers," Cell, vol. 119, pp. 831–845, Dec 2004.

[17] A. Busch and K. J. Hertel, "Evolution of SR protein and hnRNP splicing regulatory factors," Wiley Interdiscip Rev RNA, vol. 3, no. 1, pp. 1–12, 2012.

[18] X. D. Fu, "The superfamily of arginine/serine-rich splicing factors," RNA, vol. 1, pp. 663–680, Sep 1995.

[19] J. C. Long and J. F. Caceres, "The SR protein family of splicing factors: master regulators of gene expression," Biochem. J., vol. 417, pp. 15–27, Jan 2009.

[20] J. L. Parmley, A. O. Urrutia, L. Potrzebowski, H. Kaessmann, and L. D. Hurst, "Splicing and the evolution of proteins in mammals," PLoS Biol, vol. 5, pp. 1–11, 02 2007.

[21] W. G. Fairbrother, D. Holste, C. B. Burge, and P. A. Sharp, "Single nucleotide polymorphism-based validation of exonic splicing enhancers," PLoS Biol., vol. 2, p. E268, Sep 2004.

[22] G. Dujardin, C. Lafaille, M. de la Mata, L. E. Marasco, M. J. Munoz, C. Le Jossic-Corcos, L. Corcos, and A. R. Kornblihtt, "How slow RNA polymerase II elongation favors alternative exon skipping," Mol. Cell, vol. 54, pp. 683–690, May 2014.

[23] A. R. Kornblihtt, I. E. Schor, M. Allo, G. Dujardin, E. Petrillo, and M. J. Munoz, "Alternative splicing: a pivotal step between eukaryotic transcription and translation," Nat. Rev. Mol. Cell Biol., vol. 14, pp. 153–165, Mar 2013.

[24] K. A. Dittmar, P. Jiang, J. W. Park, K. Amirikian, J. Wan, S. Shen, Y. Xing, and R. P. Carstens, "Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing," Mol. Cell. Biol., vol. 32, pp. 1468–1482, Apr 2012.

[25] C. C. Warzecha, P. Jiang, K. Amirikian, K. A. Dittmar, H. Lu, S. Shen, W. Guo, Y. Xing, and R. P. Carstens, "An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition," EMBO J., vol. 29, pp. 3286–3300, Oct 2010.

[26] Q. Pan, O. Shai, C. Misquitta, W. Zhang, A. L. Saltzman, N. Mohammad, T. Babak, H. Siu, T. R. Hughes, Q. D. Morris, B. J. Frey, and B. J. Blencowe, "Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform," Mol. Cell, vol. 16, pp. 929–941, Dec 2004.

[27] M. Kalyna, S. Lopato, V. Voronin, and A. Barta, "Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins," Nucleic Acids Res., vol. 34, no. 16, pp. 4395–4405, 2006.

[28] C. Ruhl, E. Stauffer, A. Kahles, G. Wagner, G. Drechsel, G. Ratsch, and A. Wachter, "Polypyrimidine tract binding protein homologs from Arabidopsis are key regulators of alternative splicing with implications in fundamental developmental processes," Plant Cell, vol. 24, pp. 4360–4375, Nov 2012.

[29] A. L. Saltzman, Q. Pan, and B. J. Blencowe, "Regulation of alternative splicing by the core spliceosomal machinery," Genes Dev., vol. 25, pp. 373–384, Feb 2011.

[30] S. Stamm, S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T. A. Thanaraj, and H. Soreq, "Function of alternative splicing," Gene, vol. 344, pp. 1–20, Jan 2005.

[31] M. Kalyna, C. G. Simpson, N. H. Syed, D. Lewandowska, Y. Marquez, B. Kusenda, J. Marshall, J. Fuller, L. Cardle, J. McNicol, H. Q. Dinh, A. Barta, and J. W. Brown, "Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis," Nucleic Acids Res., vol. 40, pp. 2454–2469, Mar 2012.

[32] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," Nat. Methods, vol. 8, pp. 469–477, Jun 2011.

[33] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," Nat. Rev. Genet., vol. 10, pp. 57–63, Jan 2009.

[34] C. Trapnell, L. Pachter, and S. L. Salzberg, "Tophat: discovering splice junctions with rna-seq.," Bioinformatics, vol. 25, no. 9, 2009.

[35] W. Wang, Z. Qin, Z. Feng, X. Wang, and X. Zhang, "Identifying differentially spliced genes from two groups of RNA-seq samples," vol. 518, pp. 164–170, 2013.

[36] G. P. Alamancos, E. Agirre, and E. Eyras, "Methods to study splicing from high-throughput RNA sequencing data," Methods in Molecular Biology, vol. 1126, 2014.

[37] L. Chen, "Handbook of Statistical Bioinformatics," pp. 31–54, 2011.

[38] U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe, "Widespread intron retention in mammals functionally tunes transcriptomes," Genome Research, vol. 24, no. 11, 2014.

[39] G. P. Alamancos, A. Pages, J. L. Trincado, N. Bellora, and E. Eyras, "Leveraging transcript quantification for fast computation of alternative splicing profiles," RNA, vol. 21, pp. 1521–1531, Sep 2015.

[40] J. E. Hooper, "A survey of software for genome-wide discovery of differential splicing in RNA-Seq data," Hum. Genomics, vol. 8, p. 3, Jan 2014.

[41] R. Liu, A. E. Loraine, and J. A. Dickerson, "Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems," pp. 1–16, 2014.

[42] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," Nat. Rev. Genet., vol. 17, pp. 333–351, May 2016.

[43] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with RNA-seq," Nat. Biotechnol., vol. 31, pp. 46–53, Jan 2013.

[44] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, "Analysis and design of RNA sequencing experiments for identifying isoform regulation," Nature Methods, vol. 7, no. 12, 2010.

[45] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data," Genome Research, vol. 22, no. 10, 2012.

[46] S. Shen, J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing, "rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data," Proc. Natl. Acad. Sci. U.S.A., vol. 111, pp. E5593–5601, Dec 2014.

[47] Y. Hu, Y. Huang, Y. Du, C. F. Orellana, D. Singh, A. R. Johnson, A. Monroy, P. F. Kuan, S. M. Hammond, L. Makowski, S. H. Randell, D. Y. Chiang, D. N. Hayes, C. Jones, Y. Liu, J. F. Prins, and J. Liu, "DiffSplice: the genome-wide detection of differential splicing events with RNA-seq," Nucleic Acids Res., vol. 41, p. e39, Jan 2013.

[48] S. E. Sanchez, E. Petrillo, E. J. Beckwith, X. Zhang, M. L. Rugnone, C. E. Hernando, J. C. Cuevas, M. A. Godoy Herz, A. Depetris-Chauvin, C. G. Simpson, J. W. Brown, P. D. Cerdan, J. O. Borevitz, P. Mas, M. F. Ceriani, A. R. Kornblihtt, and M. J. Yanovsky, "A methyl transferase links the circadian clock to the regulation of alternative splicing," Nature, vol. 468, pp. 112–116, Nov 2010.

[49] GMOD, "Generic model organism database project." `http://gmod.org`, 2007 (accedido noviembre de 2016).

[50] R Core Team, "The r project for statistical computing." `https://www.r-project.org/`, 2016, accedido noviembre 2016.

[51] R Core Team, "The comprehensive r archive network." `https://cran.r-project.org/`, 2016, accedido noviembre 2016.

[52] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," Genome Biol., vol. 5, no. 10, p. R80, 2004.

[53] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole?, H. Pages, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan, "Orchestrating high-throughput genomic analysis with Bioconductor," Nat. Methods, vol. 12, pp. 115–121, Feb 2015.

[54] Bioconductor Core team, "Bioconductor. open source for bioinformatics." `http://bioconductor.org`, 2001, accedido noviembre de 2016.

[55] M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. Morgan, and V. Carey, "Software for computing and annotating genomic ranges," PLoS Computational Biology, vol. 9, 2013.

[56] T. C. Mockler, H. Guo, H. Yang, H. Duong, and C. Lin, "Antagonistic actions of Arabidopsis cryptochromes and phytochrome B in the regulation of floral induction," Development, vol. 126, pp. 2073–2082, May 1999.

[57] Y. Marquez, M. Hopfler, Z. Ayatollahi, A. Barta, and M. Kalyna, "Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity," Genome Research, vol. 25, no. 7, 2015.

[58] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," Bioinformatics, vol. 26, pp. 139–140, Jan 2010.

[59] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," Genome Biol., vol. 11, no. 3, p. R25, 2010.

[60] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. Roy. Statist. Soc. Ser., no. 1, 1995.

[61] S. Anders, D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber, and M. D. Robinson, "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor," Nat Protoc, vol. 8, pp. 1765–1786, Sep 2013.

[62] J. Zyprych-Walczak, A. Szabelska, L. Handschuh, K. Gorczak, K. Klamecka, M. Figlerowicz, and I. Siatkowski, "The Impact of Normalization Methods on RNA-Seq Data Analysis," Biomed Res Int, vol. 2015, p. 621690, 2015.

[63] Open Source Initiative, "Open source initiative." `https://opensource.org/`, Accedido noviembre de 2016.

[64] S. Perez-Santangelo, E. Mancini, L. J. Francey, R. G. Schlaen, A. Chernomoretz, J. B. Hogenesch, and M. J. Yanovsky, "Role for LSM genes in the regulation of circadian rhythms," Proc. Natl. Acad. Sci. U.S.A., vol. 111, pp. 15166–15171, Oct 2014.

[65] R. G. Schlaen, E. Mancini, S. E. Sanchez, S. Perez-Santangelo, M. L. Rugnone, C. G. Simpson, J. W. Brown, X. Zhang, A. Chernomoretz, and M. J. Yanovsky, "The spliceosome assembly factor GEMIN2 attenuates the effects of temperature on alternative splicing and circadian rhythms," Proc. Natl. Acad. Sci. U.S.A., vol. 112, pp. 9382–9387, Jul 2015.

[66] C. E. Hernando, S. E. Sanchez, E. Mancini, and M. J. Yanovsky, "Genome wide comparative analysis of the effects of PRMT5 and PRMT4/CARM1 arginine methyltransferases on the Arabidopsis thaliana transcriptome," BMC Genomics, vol. 16, p. 192, Mar 2015.

[67] F. A. De Maio, G. Risso, N. G. Iglesias, P. Shah, B. Pozzi, L. G. Gebhard, P. Mammi, E. Mancini, M. J. Yanovsky, R. Andino, N. Krogan, A. Srebrow, and A. V. Gamarnik, "The Dengue Virus NS5 Protein Intrudes in the Cellular Spliceosome and Modulates Splicing," PLoS Pathog., vol. 12, p. e1005841, Aug 2016.

[68] E. Mancini, S. E. Sanchez, A. Romanowski, R. G. Schlaen, M. Sanchez-Lamas, P. D. Cerdan, and M. J. Yanovsky, "Acute Effects of Light on Alternative Splicing in Light-Grown Plants," Photochem. Photobiol., vol. 92, no. 1, pp. 126–133, 2016.

[69] C. Kami, S. Lorrain, P. Hornitschek, and C. Fankhauser, "Light-regulated plant growth and development," Curr. Top. Dev. Biol., vol. 91, pp. 29–66, 2010.

[70] J. J. Casal, "Photoreceptor signaling networks in plant responses to shade," Annu Rev Plant Biol, vol. 64, pp. 403–427, 2013.

[71] Y. H. Song, J. S. Shim, H. A. Kinmonth-Schultz, and T. Imaizumi, "Photoperiodic flowering: time measurement mechanisms in leaves," Annu Rev Plant Biol, vol. 66, pp. 441–464, 2015.

[72] V. C. Galvao and C. Fankhauser, "Sensing the light environment in plants: photoreceptors and early signaling steps," Curr. Opin. Neurobiol., vol. 34, pp. 46–53, Oct 2015.

[73] J. J. Casal and M. J. Yanovsky, "Regulation of gene expression by light," Int. J. Dev. Biol., vol. 49, no. 5-6, pp. 501–511, 2005.

[74] H. Wang and H. Wang, "Phytochrome signaling: time to tighten up the loose ends," Mol Plant, vol. 8, pp. 540–551, Apr 2015.

[75] S. H. Wu, "Gene expression regulation in photomorphogenesis from the perspective of the central dogma," Annu Rev Plant Biol, vol. 65, pp. 311–333, 2014.

[76] P. Juntawong and J. Bailey-Serres, "Dynamic Light Regulation of Translation Status in Arabidopsis thaliana," Front Plant Sci, vol. 3, p. 66, 2012.

[77] M. J. Liu, S. H. Wu, J. F. Wu, W. D. Lin, Y. C. Wu, T. Y. Tsai, H. L. Tsai, and S. H. Wu, "Translational landscape of photomorphogenic Arabidopsis," Plant Cell, vol. 25, pp. 3699–3710, Oct 2013.

[78] M. J. Liu, S. H. Wu, H. M. Chen, and S. H. Wu, "Widespread translational control contributes to the regulation of Arabidopsis photomorphogenesis," Mol. Syst. Biol., vol. 8, p. 566, 2012.

[79] I. Paik, S. Yang, and G. Choi, "Phytochrome regulates translation of mRNA in the cytosol," Proc. Natl. Acad. Sci. U.S.A., vol. 109, pp. 1335–1340, Jan 2012.

[80] E. Petrillo, M. A. Godoy Herz, A. Fuchs, D. Reifer, J. Fuller, M. J. Yanovsky, C. Simpson, J. W. Brown, A. Barta, M. Kalyna, and A. R. Kornblihtt, "A chloroplast retrograde signal regulates nuclear alternative splicing," Science, vol. 344, pp. 427–430, Apr 2014.

[81] H. Shikata, K. Hanada, T. Ushijima, M. Nakashima, Y. Suzuki, and T. Matsushita, "Phytochrome controls alternative splicing to mediate light responses in Arabidopsis," Proc. Natl. Acad. Sci. U.S.A., vol. 111, pp. 18781–18786, Dec 2014.

[82] H. L. Tsai, Y. H. Li, W. P. Hsieh, M. C. Lin, J. H. Ahn, and S. H. Wu, "HUA ENHANCER1 is involved in posttranscriptional regulation of positive and negative regulators in Arabidopsis photomorphogenesis," Plant Cell, vol. 26, pp. 2858–2872, Jul 2014.

[83] Y. Wang, X. Fan, F. Lin, G. He, W. Terzaghi, D. Zhu, and X. W. Deng, "Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light," Proc. Natl. Acad. Sci. U.S.A., vol. 111, pp. 10359–10364, Jul 2014.

[84] H. P. Wu, Y. S. Su, H. C. Chen, Y. R. Chen, C. C. Wu, W. D. Lin, and S. L. Tu, "Genome-wide analysis of light-regulated alternative splicing mediated by photoreceptors in Physcomitrella patens," Genome Biol., vol. 15, no. 1, p. R10, 2014.

[85] E. Yakir, D. Hilman, M. Hassidim, and R. M. Green, "CIRCADIAN CLOCK ASSOCIATED1 transcript stability and the entrainment of the circadian clock in Arabidopsis," Plant Physiol., vol. 145, pp. 925–932, Nov 2007.

[86] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," Genome Biol., vol. 10, no. 3, p. R25, 2009.

[87] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," Bioinformatics, vol. 25, pp. 1105–1111, May 2009.

[88] X. Hong, D. G. Scofield, and M. Lynch, "Intron size, abundance, and distribution within untranslated regions of genes," Mol. Biol. Evol., vol. 23, pp. 2392–2404, Dec 2006.

[89] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," Nucleic Acids Res., vol. 30, pp. 207–210, Jan 2002.

[90] N. C. for Biotechnology Information Gene Expression Omnibus, "Ncbi geo." `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68560`, 2000, (accedido 20 de julio de 2016).

[91] The Arabidopsis Information Resource , "Tair." `http:http://www.arabidopsis.org`, Accedido noviembre de 2016.

[92] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," Bioinformatics, vol. 26, no. 1, 2010.

[93] B. Strasser, M. Sanchez-Lamas, M. J. Yanovsky, J. J. Casal, and P. D. Cerdan, "Arabidopsis thaliana life without phytochromes," Proc. Natl. Acad. Sci. U.S.A., vol. 107, pp. 4776–4781, Mar 2010.

[94] M. Llorian and C. W. J. Smith, Theory and Protocols. Wiley-VCH Verlag GmbH Co. KGaA, 2012.

[95] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," Nat. Genet., vol. 25, pp. 25–29, May 2000.

[96] V. a software platform to support systems biology research, "Virtualplant." `http://virtualplant.bio.nyu.edu/cgi-bin/vpweb/`, 2010, (accedido 1 de Abril de 2015).

[97] M. S. Katari, S. D. Nowicki, F. F. Aceituno, D. Nero, J. Kelfer, L. P. Thompson, J. M. Cabello, R. S. Davidson, A. P. Goldberg, D. E. Shasha, G. M. Coruzzi, and R. A. Gutierrez, "VirtualPlant: a software platform to support systems biology research," Plant Physiol., vol. 152, pp. 500–515, Feb 2010.

[98] K. Jung, L. Bartley, P. Cao, P. Canlas, and P. Ronald, "Analysis of alternatively spliced rice transcripts using microarray data," Rice, vol. 2, no. 1, pp. 44–55, 2009.

[99] H. Shikata, M. Shibata, T. Ushijima, M. Nakashima, S. G. Kong, K. Matsuoka, C. Lin, and T. Matsushita, "The RS domain of Arabidopsis splicing factor RRC1 is required for phytochrome B signal transduction," Plant J., vol. 70, pp. 727–738, Jun 2012.

[100] P. Papasaikas, J. R. Tejedor, L. Vigevani, and J. Valcarcel, "Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery," Mol. Cell, vol. 57, pp. 7–22, Jan 2015.

[101] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "MEME: discovering and analyzing DNA and protein sequence motifs," Nucleic Acids Res., vol. 34, pp. W369–373, Jul 2006.

[102] M. L. Rugnone, A. Faigon Soverna, S. E. Sanchez, R. G. Schlaen, C. E. Hernando, D. K. Seymour, E. Mancini, A. Chernomoretz, D. Weigel, P. Mas, and M. J. Yanovsky, "LNK genes integrate light and clock signaling networks at the core of the Arabidopsis oscillator," Proc. Natl. Acad. Sci. U.S.A., vol. 110, pp. 12120–12125, Jul 2013.

[103] A. Romanowski and M. J. Yanovsky, "Circadian rhythms and post-transcriptional regulation in higher plants," Front Plant Sci, vol. 6, p. 437, 2015.

[104] A. Golisz, P. J. Sikorski, K. Kruszka, and J. Kufel, "Arabidopsis thaliana LSM proteins function in mRNA splicing and degradation," Nucleic Acids Res., vol. 41, pp. 6232–6249, Jul 2013.

[105] P. Cui, S. Zhang, F. Ding, S. Ali, and L. Xiong, "Dynamic regulation of genome-wide pre-mRNA splicing and stress tolerance by the Sm-like protein LSm5 in Arabidopsis," Genome Biol., vol. 15, p. R1, Jan 2014.

[106] UCSC Genome Bioinformatics , "Ucsc." `http://hgdownload.cse.ucsc.edu/downloads.html`, Accedido noviembre de 2016.