Identificación, análisis y perspectivas evolutivas del SL *trans*-splicing en gusanos platelmintos

Autor: Mag. Javier Calvelo Comesaña

Tutores: Dr. Andrés Iriarte - Dr. Uriel Koziol - Dr. Hector Musto

Montevideo, Uruguay
Agosto, 2024

Índice contenidos:

Capítulo 1: Introducción y Objetivos

1.A) Introducción

Comúnmente se recomienda que el primer paso para resolver un problema sea delimitarlo. Definir cuál es su naturaleza, cuál es su origen y su extensión. Cuando se trabaja con genes en un genoma, se comienza por determinar su localización y secuencia. Identificar el primer y el último codón codificante para poder predecir su secuencia proteica, idealmente junto con las regiones no codificantes 5' y 3' UTR, promotores, e isoformas alternativas que puedan existir. Luego en base a esta información se pueden realizar otros tipos de estudios: funcionalidad, expresión diferencial bajo distintas condiciones, conservación y/o divergencia entre grupos, entre otras líneas de estudio. Con un buen modelo experimental y con suerte, estas investigaciones pueden dilucidar preguntas clave de la especie estudiada. Desde preguntas básicas como adaptaciones fisiológicas o su historia evolutiva, hasta cuestiones prácticas como identificar y caracterizar blancos vacunales y/o drogas.

En organismos con Spliced Leader (SL) *trans-splicing* es necesario agregar que los transcriptos de este gen pueden ser combinados en algún punto con una secuencia diferente. Este proceso elimina toda secuencia río arriba del sitio de inserción; la cual puede ser descartada o procesada como un gen independiente; potencialmente alterando el marco de lectura. La inserción puede ocurrir en uno o más sitios (o ninguno) y la secuencia insertada es frecuentemente desconocida cuando se estudia por primera vez este proceso en una nueva especie. Cada inserción individual puede ser producto de un error de la maduración del transcrito; o puede ser crucial para la viabilidad del organismo. Y como capa adicional de incertidumbre, puede que los datos recabados no sean suficientes por sí mismos para caracterizar apropiadamente el proceso y diversidad de transcritos producidos.

Esta tesis plantea afrontar la cuestión de SL *trans-splicing* en el grupo de gusanos planos platelmintos, particularmente en los linajes parásitos Cestoda y Trematoda. Se exploran las dificultades intrínsecas del mecanismo para su estudio; así como las asociadas a los datos disponibles; y se implementan soluciones bioinformáticas para extraer el mayor volumen y calidad de información posible. Luego se analiza la identidad, conservación e importancia biológica de los genes y sitios sometidos a SL *trans-splicing*, así como los *loci* codificantes de los SL-ARN en sí mismos.

En el presente capítulo presento los objetivos generales y específicos de la tesis doctoral, así como una introducción breve a los temas transversales de la tesis en su conjunto. En el **Capítulo 2** se presenta SLFinder, un pipeline de análisis informáticos diseñada para tomar

transcriptomas ensamblados "*de-novo*" e identificar posibles secuencias SL y sin necesidad de recurrir a conocimiento previo. El **Capítulo 3** presenta la aplicación de este pipeline en la especie modelo *Hymenolepis microstoma*. Se identifican los SLs presentes en su transcriptoma y sus sitios aceptores, para luego analizar el posible rol de SL *trans-splicing* en la expresión constitutiva de genes mono y policistrónicos. Analizando y delimitando su posible rol como mecanismo regulador del ciclo de vida de la especie. El **Capítulo 4** extiende la aplicación de los métodos desarrollados a una escala filogenética, con la investigación del complemento de *loci* SL-ARN y los genes sometidos a SL *trans-splicing* en 24 especies de Cestodos y Trematodos. Finalmente, las conclusiones generales se exponen en el **Capítulo 5**.

1.B) SL *trans-splicing* en eucariotas

El proceso de "*splicing*" de ARN refiere al corte y empalme de moléculas de ARN. En eucariotas el proceso es mediado por el spliceosoma, un complejo de 5 ARNs no codificantes catalíticos, y proteínas asociadas (Kastner et al., 2019). Representa una característica crucial y distintiva de los genes eucariotas con sus abundantes regiones intrónicas no codificantes (Roy & Irimia, 2009) y forma las bases mecanísticas para la regulación de varias características fenotípicas (Wright et al., 2022). Una clasificación básica de los procesos de splicing responde al ARN utilizado (Lei et al., 2016): *cis-splicing* en donde el empalme ocurre entre porciones de la misma molécula ARN, causando la remoción de porciones internas (**Fig. 1A**), y *trans-splicing* en el cual dos moléculas ARN diferentes son combinadas en un producto final (**Fig. 1B**). El SL *trans-splicing* es una especialización de este último donde un ARN no codificante (Spliced Leader o "SL") es incorporado en un pre-ARNm como parte de su maduración (Bitar et al., 2013; Blumenthal, 2005; Hastings, 2005; Lasda & Blumenthal, 2011). Resultando en la producción de un transcrito ARN que incorpora parte del SL-ARN en su extremo 5' (región *Leader* del SL-ARN), típicamente incluyendo una caperuza 5' de tipo trimetilguanosina (TMG) (**Fig. 1C**). La región río arriba del transcrito original puede ser eliminada o procesada en un transcrito independiente, el resto del SL-ARN (región denominada *outron* o *intron-like* en la literatura) forma un ARN ramificado "Y" (Lasda & Blumenthal, 2011).

La **Figura 2** representa la estructura de un SL-ARN clásico. La región *Leader* e *Intron-like* son separadas por un sitio donador de splicing canónico [UGGU], esta última usualmente contiene un sitio llamado SM para la interacción con el complejo de proteínas pequeñas Sm (Bitar et al., 2013; Hastings, 2005; Lasda & Blumenthal, 2011). Pero su identificación no es trivial debido a variaciones en el motivo. Más aún, en grupos como dinoflagelados podría estar incluso localizado en la región SL (Zhang et al., 2007) mientras que en otros como

Perkinsozoa parece estar ausente (Alacid et al., 2022). La porción *Leader* por su parte puede variar considerablemente entre distintos organismos; entre 16 bases en el urocordado *Ciona intestinalis* (Matsumoto et al., 2010) a 51 bases en el platelminto de vida libre *Stylochus zebra* (Davis, 1997); y no tiene motivos conservados universales, aunque si hay conservación dentro de linajes específicos (Bitar et al., 2013).

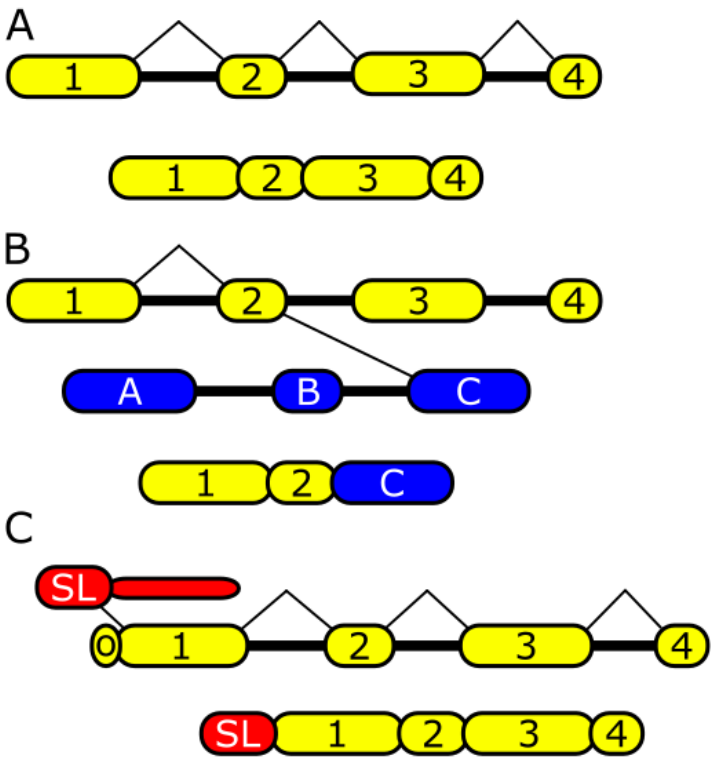

Figura 1: Clasificación básica de tipos de splicing observados en organismos eucariotas. Las cajas de colores indican exones, intrones como las líneas gruesas que los conectan, y las uniones de *splicing* con líneas delgadas. A) *Cis*-splicing donde distintas porciones del mismo transcrito son cortadas y empalmadas, terminando en la remoción de regiones internas en el ARNm final (denominadas intrones). B) *Trans-splicing*, donde dos transcritos ARN diferentes; representados con cajas amarillas numeradas y azules con letras; son cortados y empalmados juntos. Estos pueden ser isoformas provenientes del mismo gen o diferentes. C) SL *trans-splicing* es un caso especial de *trans-splicing* en el que uno de los transcritos involucrados es un gen no codificante especializado en el proceso; referido en la tesis como SL-ARN y representado en rojo. Típicamente es incorporado en el extremo 5' del pre-ARNm y conlleva el reemplazo de la región río arriba del sitio de inserción.

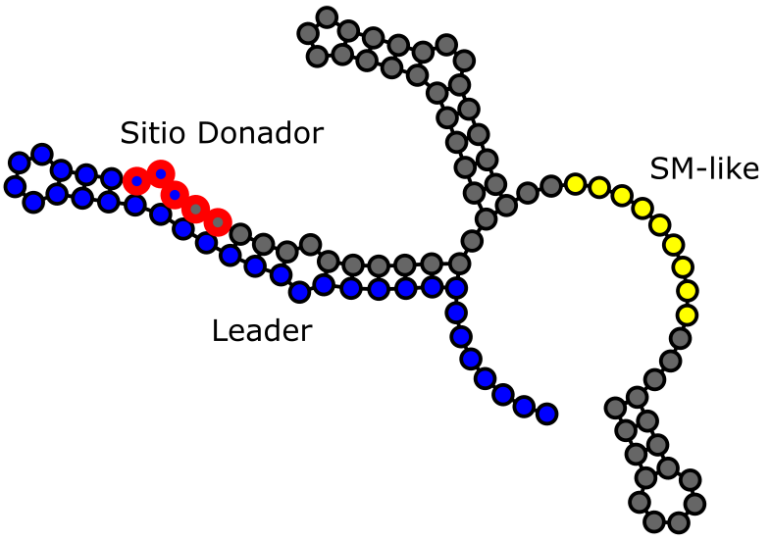

Figura 2: Representación esquemática de un SL-ARN en platelmintos. El transcrito se divide en 2 porciones: la región *Spliced Leader* (SL) que es integrada en el ARNm final y el resto de la molécula referida comúnmente como "*intron-like*", separadas por un sitio donador canónico "UGGU" en rojo. La porción intrónica suele contener un sitio SM-like.

El plegamiento de los SL-ARN a nivel de estructura secundaria muestra características similares a los ARN pequeños nucleares (snRNA) del spliceosoma sugiriendo un posible origen común, aunque no necesariamente único dentro de eucariotas (Hastings, 2005; Lasda & Blumenthal, 2011). Sin embargo, al igual que otros patrones y procesos asociados con el *splicing* de ARNs en eucariotas (ver Roy & Irimia (2009) por una revisión), no existe consenso sobre su origen singular o múltiple (Bitar et al., 2013; Douris et al., 2010; Hastings, 2005; Krchňáková et al., 2017; Lasda & Blumenthal, 2011). Esto se debe al bajo grado de conservación de los SL-ARN entre distintos grupos (Bitar et al., 2013) y a su distribución parcheada en la filogenia en eucariotas (Bitar et al., 2013; Douris et al., 2010; Hastings, 2005; Lasda & Blumenthal, 2011). A la fecha se lo ha identificado en grupos como Euglenozoa (Sather & Agabian, 1985; Tessier et al., 1991), Platelmintos (Rajkovic et al., 1990), Nematoda (Krause & Hirsh, 1987; Ross et al., 1995), Urochordata (Vandenberghe et al., 2001), Rotifera (Pouchkina-Stantcheva & Tunnacliffe, 2005), Cnidaria (Stover & Steele, 2001), Dinoflagellata (Lidie & Van Dolah, 2007), Crustaceos (anfípodos y copépodos; Douris et al. (2010)), Cercozoa (Matsuo et al., 2018), Perkinsozoa (Alacid et al., 2022); pero está ausente en grupos como Vertebrados, Insectos, Plantas y Hongos (Lasda & Blumenthal, 2011; Lei et al., 2016). La discusión sobre su origen se reactiva con su identificación en cada nuevo linaje y tentativamente se decanta sobre un origen múltiple independiente (ej: Douris et al., 2010).

Independientemente de su origen, su principal función es el procesamiento de transcriptos policistrónicos proveniente de operones (loci codificantes para múltiples genes que son transcritos en un mismo ARN primario); segmentándolo en ARNm monocistrónicos y proveyendo una caperuza 5' a los genes localizados río abajo (Blumenthal, 2005; Hastings, 2005; Lasda & Blumenthal, 2011; Pettitt et al., 2014). En grupos como nemátodos, el grupo de animales mejor estudiado poseedor de SL *trans-splicing*, existen SLs especializados para el *trans*-splicing de transcriptos monocistrónicos y del primer cistrón en transcriptos policistrónicos (SL-1) y otros especializados para los siguientes cistrones en los transcritos policistrónicos, resultando en la resolución de los operones (SL-2) (Allen et al., 2011; Harrison, 1989; Pettitt et al., 2010; Wenzel et al., 2020). El grado de especialización es tal que el ratio entre SL-1 y SL-2 en sitios específicos puede usarse para identificar operones independientemente del largo de los espacios intergénicos, y predecir operones híbridos en los que existe un inicio de transcripción interno (Allen et al., 2011). Dicho esto, y teniendo en cuenta que la gran mayoría de los genes sometidos a SL *trans-splicing* tienden a ser monocistrónicos, otras funciones propuestas incluyen la purga de secuencias deletéreas o regulatorias en los 5' UTR (Bitar et al., 2013; Hastings, 2005; Lasda & Blumenthal, 2011); así como la formación de isoformas alternativas (Agorio et al., 2003; Boroni et al., 2018; Hastings, 2005; Nilsson et al., 2010; Siegel et al., 2011). Debido en parte a una menor eficiencia del proceso del SL *trans-splicing* respecto es menos eficiente (Blumenthal, 2005), sus inserciones se concentran en sitios aceptores donde el *cis-splicing* se ve restringido (ej: por la ausencia de un sitio de *splicing* canónico eficiente). Esto ofrece un mecanismo simple para restringir las inserciones de SLs a sitios específicos, pero a su vez implica que el mecanismo es potencialmente ruidoso (Tourasse et al., 2017).

En lo referente a la maquinaria molecular participante, SL *trans-splicing* emplea la mayoría de los componentes principales del spliceosoma que *cis-splicing*, con excepción del U1 snRNP que es reemplazado por el SL-ARN (Hannon et al., 1991). A nivel de las proteínas asociadas, los estudios se han limitados a los linajes tripanosomátidos, donde el *cis-splicing* cumple un rol marginal en su expresión génica (Michaeli, 2011) y por ende no es representativo de otros eucariotas, y en nemátodos. En nemátodos, *cis-* y *trans-splicing* coexisten en la mayoría de los genes (Allen et al., 2011; Tourasse et al., 2017) y se ha caracterizado en suficiente detalle para identificar factores específicos al SL *trans-splicing* (Lasda & Blumenthal, 2011), incluyendo varias proteínas y ARNs no codificantes que participan el SL *trans-splicing* mediado por SL-1 (inicio de la transcripción) o SL-2 (resolución de operones) (Denker et al., 2002; Eijlers et al., 2024; Fasimoye et al., 2022; Macmorris et al., 2007).

Aunque el estudio del SL *trans-splicing* en especies modelo como *Caenorhabditis elegans* o *Schistosoma mansoni* puede considerarse rutinario a la fecha (Allen et al., 2011; Blumenthal, 2005; Boroni et al., 2018), su identificación en nuevas especies suele deberse a observaciones fortuitas (ej: Barnes et al., 2019). La alta variabilidad de los loci de SL-ARN entre grupos distantes limita la eficacia de comparaciones por homología a linajes cercanos, haciendo que SL-ARNs funcionales no sea encontrados por su divergencia nucleotídica. Una vez identificados los SL, el siguiente paso es identificar sus sitios aceptores. Al día de hoy la mejor aproximación a escala genómica son las tecnologías de secuenciación masiva, tanto en términos de volumen de datos como facilidades prácticas de aplicación (Conesa et al., 2016). Sin embargo, debido a su localización en el extremo 5' es difícil recuperar el SL en su totalidad, particularmente importante cuando se utilizan muestras enriquecidas por pesca poli-A (Conesa et al., 2016); el estándar en eucariotas. Solventar el sesgo en contra de las secuencias 5' solo puede ser abordado con aproximaciones metodológicas en la secuenciación, como lo es la aproximación "SL-seq" o "SL-trapping" en la que se utilizan *primers* específicos a la secuencia del SL para enriquecer la secuenciación en transcritos sometidos a SL *trans-splicing* (Cuypers et al., 2017; Nilsson et al., 2010), secuenciaciones completas de las isoformas con *reads* largos (ej: Bernard et al., 2023; Rhoads & Au, 2015) u otros métodos que recuperen principalmente la región 5' del transcritos en masa. Dicho esto, la amplia disponibilidad de datos generados mediante pesca poli-A, el desarrollo de métodos para la identificación de sitios aceptores a partir de estos datos promete es prometedor a pesar de sus limitaciones. Por otro lado, la identificación de nuevas secuencias requiere un método para identificar secuencias en el 5' de los transcritos que no concuerdan con lo esperado dada la secuencia del genoma, y luego filtrar posibles errores en la secuenciación.

En resumen, el estudio de SL *trans-splicing* en una nueva especie puede resumirse en *"secuencia desconocida localizada en una porción de baja cobertura de los transcritos secuenciados"*. Solventar estas dificultades requiere una aproximación confiable y escalable a escala genómica para su identificación, seguido de estudios exploratorios para determinar las particularidades del SL *trans-splicing* en el linaje de interés. Como el largo de los SLs, diversidad nucleotídica y su posible especialización con objetivos particulares.

1.C) Gusanos platelmintos

Los gusanos platelmintos son un grupo de acelomados, famoso por sus linajes parásitos: las clases Cestoda, Monogenea, Trematoda (Hickman et al., 2008), mientras que sus linajes basales están compuestos principalmente por organismos de vida libre (Egger et al., 2015; Riutort et al., 2012). Las clases Cestoda y Trematoda son de interés económico y social,

particularmente en países en vías de desarrollo (Giri & Parija, 2012), debido su capacidad para infectar vertebrados mamíferos, incluyendo humanos (Hickman et al., 2008; Mahmud et al., 2017).

Ambos grupos presentan un ciclo de vida complejo que involucra la transmisión entre varias especies hospedadoras, denominados intermediarios o definitivos dependiendo del lugar que ocupan en su ciclo de vida (Hickman et al., 2008). A modo de ejemplo, *Hymenolepis microstoma,* especie modelo analizada en el **Capítulo 3** de esta tesis, presenta tres estadios (Cunningham & Olson, 2010): 1) Adulto, alojado en el ducto biliar de ratones como *Mus musculus*, desde donde emite huevos que son liberados juntos las heces conteniendo 2) larvas oncósferas (primer estadío larvario) ya infectivas. Estos permanecen en el suelo, hasta ser ingeridos por escarabajos del género *Tribolium* o Tenebrio. Una vez en el escarabajo, las larvas migran al hemocele donde continúan su desarrollo en 3) cisticercoide (segundo estadío larvario), en el que permanecen hasta que el ciclo se completa cuando el escarabajo es consumido por un ratón. Otras especies presentan variaciones, por ejemplo, *Echinococcus granulosus* utiliza cánidos como hospedador definitivo y ovejas o cabras como intermediario en los que se reproduce en forma asexual (Mahmud et al., 2017). Trematodos de la sublcase Digenea incorporan estadios móviles (ej: larvas miracidio y cercarias) en los que el parásito activamente busca el siguiente hospedador en el ciclo, o un lugar donde enquistar dependiendo de la especie (Hickman et al., 2008). Independientemente de la especie, las infecciones por platelmintos tienden a pasar desapercibidas, pero presentan complicaciones dependientes del órgano donde se alojen los parásitos (intestino, músculo esquelético, hígado, corazón, cerebro, etc), estadio y abundancia (Mahmud et al., 2017).

La presencia de SL *trans-splicing* fue detectado en 1990 en la especie *S. mansoni* (Rajkovic et al., 1990), posteriormente expandido a *E. multilocularis* (Brehm et al., 2000), *Fasciola hepatica* (Davis et al., 1994); y más recientemente especies como *H. microstoma* (Olson et al., 2020) y *Opisthorchis felineus* (Ershov et al., 2019). Aunque a la fecha no se han realizado estudios comprensivos de las dinámicas evolutivas de los loci SL-ARN ni los genes sometidos a SL *trans-splicing* en el grupo. Típicamente la región SL en el grupo se aproxima a 36 pb y presenta un codón AUG terminal que en ocasiones representa un nuevo codón de inicio de la traducción (Bitar et al., 2013; Cheng et al., 2006). A pesar de su potencial rol como fuente de isoformas alternativas, muy pocos casos potenciales han sido reportados (ej: Agorio et al., 2003).

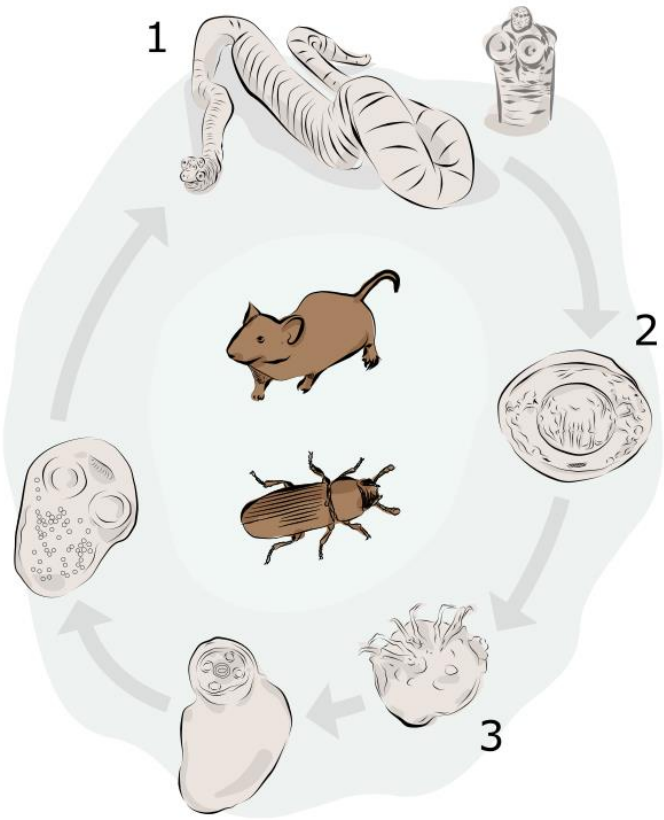

Figura 3: Ciclo de vida del cestodo *Hymenolepis microstoma*. 1) Los adultos se alojan en el ducto biliar de su hospedador del ratón (*Mus musculus*). Estos liberan huevos al ambiente conteniendo el primer estadío larvario 2) Oncosfera, y permanecen en el suelo hasta ser 3) ingeridas por un escarabajo que actúa como hospedador intermediario (típicamente un representante del género *Tribolium*).

En términos generales, la prevalencia de SL *trans-splicing* en el transcriptoma de platelmintos es relativamente reducida en comparación con nemátodos, donde la mayoría de los genes son sometidos al mecanismo (Bernard et al., 2023; Tourasse et al., 2017), en platelmintos se ha sugerido hasta un 40-55% (Boroni et al., 2018). Sin mostrar un claro enriquecimiento de vías metabólicas o clases proteicas particulares (Boroni et al., 2018; Mourão et al., 2013). Otra diferencia importante con nemátodos es la aparente ausencia de especialización de SL-ARNs en la resolución de operones (Boroni et al., 2018; Protasio et al., 2012), aunque sí se ha observado en *Schmidtea mediterranea* la existencia de un SL-ARN preferencialmente expresada en células madre (Rossia et al., 2014). Pero se desconoce si presenta propiedades funcionales diferentes a los otros SL-ARN en la especie.

1.D) Objetivos de la tesis

El objetivo general de esta tesis es identificar el complemento de loci codificantes de SL-ARN y sus transcriptos blanco en Cestodos y Trematodos. investigando su importancia biológica y sus patrones evolutivos.

Objetivos Específicos

1) Generar una herramienta capaz de identificar *de-novo* secuencias SLs, que no dependa de la homología con secuencias de referencia.
2) Generar una herramienta que permita identificar sitios aceptores de SLs en secuenciaciones de ARN masivas.
3) Identificar los SL-ARNs presentes en Céstodos y Tremátodos.
4) Identificar los sitios aceptores de SL *trans-splicing* en Céstodos y Tremátodos.
5) Evaluar la importancia biológica de los sitios aceptores de SLs en *Hymenolepis microstoma,* con énfasis en sus patrones a lo largo de los diferentes estadios del ciclo de vida.
6) Identificar patrones evolutivos en el complemento de SL-ARNs y en la organización de genes en operones en platelmintos parásitos
7) Evaluar la importancia y limitaciones de la identificación de los sitios aceptores de SL *trans-splicing* para mejorar la anotación genómica en platelmintos parásitos.

1.E) Referencias bibliográficas

Agorio, A., Chalar, C., Cardozo, S., & Salinas, G. (2003). Alternative mRNAs arising from trans-splicing code for mitochondrial and cytosolic variants of Echinococcus granulosus thioredoxin glutathione reductase. *Journal of Biological Chemistry*, *278*(15), 12920–12928. https://doi.org/10.1074/jbc.M209266200

Alacid, E., Irwin, N. A. T., Smilansky, V., Milner, D. S., Kilias, E. S., Leonard, G., & Richards, T. A. (2022). A diversified and segregated mRNA spliced-leader system in the parasitic Perkinsozoa. *Open Biology*, *12*(8), 220126. https://doi.org/10.1098/rsob.22.0126

Allen, M. A., Hillier, L. D. W., Waterston, R. H., & Blumenthal, T. (2011). A global analysis of C. elegans trans-splicing. *Genome Research*, *21*(2), 255–264. https://doi.org/10.1101/gr.113811.110

Barnes, S. N., Masonbrink, R. E., Maier, T. R., Seetharam, A., Sindhu, A. S., Severin, A. J., & Baum, T. J. (2019). Heterodera glycines utilizes promiscuous spliced leaders and demonstrates a unique preference for a species-specific spliced leader over C . elegans SL1. *Scientific Reports*, *67*(4), 1356. https://doi.org/10.1038/s41598-018-37857-0

Bernard, F., Dargère, D., Rechavi, O., & Dupuy, D. (2023). Quantitative analysis of C. elegans transcripts by Nanopore direct-cDNA sequencing reveals terminal hairpins in non trans-spliced mRNAs. *Nature Communications*, *14*(1), 1229. https://doi.org/10.1038/s41467-023-36915-0

Bitar, M., Boroni, M., Macedo, A. M., Machado, C. R., & Franco, G. R. (2013). The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. *Frontiers in Genetics*, *4*(October), 199. https://doi.org/10.3389/fgene.2013.00199

Blumenthal, T. (2005). Trans-splicing and operons. In *WormBook: the online review of C. elegans biology*. https://doi.org/10.1895/wormbook.1.5.1

Boroni, M., Sammeth, M., Gava, S. G., Jorge, N. A. N., MacEdo, A. M., MacHado, C. R., Mourão, M. M., & Franco, G. R. (2018). Landscape of the spliced leader trans-splicing mechanism in Schistosoma mansoni. *Scientific Reports*, *8*(1), 3877. https://doi.org/10.1038/s41598-018-22093-3

Brehm, K., Jensen, K., & Frosch, M. (2000). mRNA trans-splicing in the human parasitic cestode Echinococcus multilocularis. *Journal of Biological Chemistry*, *275*(49), 38311–38318. https://doi.org/10.1074/jbc.M006091200

Cheng, G., Cohen, L., Ndegwa, D., & Davis, R. E. (2006). The flatworm spliced leader 3′-terminal AUG as a translation initiator methionine. *Journal of Biological Chemistry*, *281*(2), 733–743. https://doi.org/10.1074/jbc.M506963200

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13. https://doi.org/10.1186/s13059-016-0881-8

Cunningham, L. J., & Olson, P. D. (2010). Description of Hymenolepis microstoma (Nottingham strain): A classical tapeworm model for research in the genomic era. *Parasites and Vectors*, *3*(1), 123. https://doi.org/10.1186/1756-3305-3-123

Cuypers, B., Domagalska, M. A., Meysman, P., Muylder, G. De, Vanaerschot, M., Imamura, H., Dumetz, F., Verdonckt, T. W., Myler, P. J., Ramasamy, G., Laukens, K., & Dujardin, J. C. (2017). Multiplexed Spliced-Leader Sequencing: A high-throughput, selective method for RNA-seq in Trypanosomatids. *Scientific Reports*, *7*(1), 3725. https://doi.org/10.1038/s41598-017-03987-0

Davis, R. E. (1997). Surprising diversity and distribution of spliced leader RNAs in flatworms. *Molecular and Biochemical Parasitology*, *87*, 29–48.

Davis, R. E., Singh, H., Botka, C., Hardwick, C., El Meanawy, M. A., & Villanueva, J. (1994). RNA trans-splicing in Fasciola hepatica. Identification of a spliced leader (SL) RNA and SL sequences on mRNAs. *Journal of Biological Chemistry*, *269*(31), 20026–20030.

Denker, J. A., Zuckerman, D. M., Maroney, P. A., & Nilsen, T. W. (2002). New components of the spliced leader RNP required for nematode trans-splicing. *Nature*, *417*(6889), 667–670. https://doi.org/10.1038/nature756

Douris, V., Telford, M. J., & Averof, M. (2010). Evidence for Multiple Independent Origins of trans-Splicing in Metazoa. *Molecular Biology and Evolution*, *27*(3), 684–693. https://doi.org/10.1093/molbev/msp286

Egger, B., Lapraz, F., Tomiczek, B., Müller, S., Dessimoz, C., Girstmair, J., Škunca, N., Rawlinson, K. A., Cameron, C. B., Beli, E., Todaro, M. A., Gammoudi, M., Noreña, C., & Telford, M. J. (2015). A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. *Current Biology*, *25*(10), 1347–1353. https://doi.org/10.1016/j.cub.2015.03.034

Eijlers, P., Al-Khafaji, M., Soto-Martin, E., Fasimoye, R., Stead, D., Wenzel, M., Müller, B., & Pettitt, J. (2024). A nematode-specific ribonucleoprotein complex mediates interactions between the major nematode spliced leader snRNP and its target pre-mRNAs. *Nucleic Acids Research*, *Apr 27*, gkae321. https://doi.org/10.1093/nar/gkae321

Ershov, N. I., Mordvinov, V. A., Prokhortchouk, E. B., Pakharukova, M. Y., Gunbin, K. V., Ustyantsev, K., Genaev, M. A., Blinov, A. G., Mazur, A., Boulygina, E., Tsygankova, S., Khrameeva, E., Chekanov, N., Fan, G., Xiao, A., Zhang, H., Xu, X., Yang, H., Solovyev, V., … Skryabin, K. G. (2019). New insights from Opisthorchis felineus genome: Update on genomics of the epidemiologically important liver flukes. *BMC Genomics*, *20*(1), 399. https://doi.org/10.1186/s12864-019-5752-8

Fasimoye, R. Y., Spencer, R. E. B., Soto-Martin, E., Eijlers, P., Elmassoudi, H., Brivio, S., Mangana, C., Sabele, V., Rechtorikova, R., Wenzel, M., Connolly, B., Pettitt, J., & Müller, B. (2022). A novel, essential trans-splicing protein connects the nematode SL1 snRNP to the CBC-ARS2 complex. *Nucleic Acids Research*, *50*(13), 7591–7607. https://doi.org/10.1093/nar/gkac534

Giri, S., & Parija, S. C. (2012). A review on diagnostic and preventive aspects of cystic echinococcosis and human cysticercosis. *Tropical Parasitology*, *2*(2), 99–108. https://doi.org/10.4103/2229-5070.105174

Hannon, G. J., Maroney, P. A., & Nilsen, T. W. (1991). U small nuclear ribonucleoprotein requirements for nematode cis- and trans-splicing in vitro. *Journal of Biological Chemistry*, *266*(34), 22792–22695.

Harrison, R. G. (1989). Animal Mitochondrial DNA as a Genetic Marker in Population and Evolutionary Biology. *Trends in Ecology & Evolution*, *4*(1), 6–11.

Hastings, K. E. M. (2005). SL trans-splicing: Easy come or easy go? *Trends in Genetics*, *21*(4), 240–247. https://doi.org/10.1016/j.tig.2005.02.005

Hickman, C., Roberts, L., Keen, S., Larson, A., Anson, H., & Eisenhour, D. (2008). *Integrated Principles of Zoology* (14th ed.). McGraw-Hill Higher Education.

Kastner, B., Will, C. L., Stark, H., & Lührmann, R. (2019). Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. *Cold Spring Harbor Perspectives in Biology*, *11*(11), a032417. https://doi.org/10.1101/cshperspect.a032417

Krause, M., & Hirsh, D. (1987). A trans-spliced leader sequence on actin mRNA in C. elegans. *Cell*, *49*(6), 753–761. https://doi.org/10.1016/0092-8674(87)90613-1

Krchňáková, Z., Krajčovič, J., & Vesteg, M. (2017). On the Possibility of an Early Evolutionary Origin for the Spliced Leader Trans-Splicing. *Journal of Molecular Evolution*, *85*(1–2), 37–45. https://doi.org/10.1007/s00239-017-9803-y

Lasda, E. L., & Blumenthal, T. (2011). Trans-splicing. *Wiley Interdisciplinary Reviews: RNA*, *2*(3), 417–434. https://doi.org/10.1002/wrna.71

Lei, Q., Li, C., Zuo, Z., Huang, C., Cheng, H., & Zhou, R. (2016). Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome Biology and Evolution*, *8*(3), 562–577. https://doi.org/10.1093/gbe/evw025

Lidie, K. B., & Van Dolah, F. M. (2007). Spliced leader RNA-mediated trans-splicing in a dinoflagellate, Karenia brevis. *Journal of Eukaryotic Microbiology*, *54*(5), 427–435. https://doi.org/10.1111/j.1550-7408.2007.00282.x

Macmorris, M., Kumar, M., Lasda, E., Larsen, A., Kraemer, B., & Blumenthal, T. (2007). A novel family of C. elegans snRNPs contains proteins associated with trans-splicing. *RNA*, *13*(4), 511–520. https://doi.org/10.1261/rna.426707

Mahmud, R., Ai, Y., & Lim, L. (2017). *Medical Parasitology* (1st ed.). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-319-68795-7

Matsumoto, J., Dewar, K., Wasserscheid, J., Matsumoto, J., Dewar, K., Wasserscheid, J., Wiley, G. B., Macmil, S. L., Roe, B. A., Zeller, R. W., Satou, Y., & Hastings, K. E. M. (2010). High-throughput sequence analysis of Ciona intestinalis SL trans-spliced mRNAs: alternative

expression modes and gene function correlates. *Genome Research*, *20*(5), 636–645. https://doi.org/10.1101/gr.100271.109

Matsuo, M., Katahata, A., Satoh, S., & Matsuzaki, M. (2018). Characterization of spliced leader trans-splicing in a photosynthetic rhizarian amoeba, Paulinella micropora, and its possible role in functional gene transfer. *PLoS ONE*, *13*(7), e0200961. https://doi.org/10.1371/journal.pone.0200961

Michaeli, S. (2011). Trans-splicing in trypanosomes: Machinery and its impact on the parasite transcriptome. *Future Microbiology*, *6*(4), 459–474. https://doi.org/10.2217/fmb.11.20

Mourão, M. de M., Bitar, M., Pereira Lobo, F., Paula Peconick, A., Grynberg, P., Prosdocimi, F., Waisberg, M., Coutinho Cerqueira, G., Mara Macedo, A., Renato Machado, C., Yoshino, T., & Franco, G. R. (2013). A directed approach for the identification of transcripts harbouring the spliced leader sequence and the effect of trans-splicing knockdown in Schistosoma mansoni. *Memorias Do Instituto Oswaldo Cruz*, *108*(6), 707–717. https://doi.org/10.1590/0074-0276108062013006

Nilsson, D., Gunasekera, K., Mani, J., Osteras, M., Farinelli, L., Baerlocher, L., Roditi, I., & Ochsenreiter, T. (2010). Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. *PLoS Pathogens*, *6*(8), e1001037. https://doi.org/10.1371/journal.ppat.1001037

Olson, P., Tracey, A., Baillie, A., James, K., Doyle, S., Buddenborg, S., Rodgers, F., Holroyd, N., & Berriman, M. (2020). Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection. *BMC Biology*, *18*, 165. https://doi.org/10.1101/2020.04.08.031872

Pettitt, J., Harrison, N., Stansfield, I., Connolly, B., & Müller, B. (2010). The evolution of spliced leader trans-splicing in nematodes. *Biochemical Society Transactions*, *38*(4), 1125–1130. https://doi.org/10.1042/BST0381125

Pettitt, J., Philippe, L., Sarkar, D., Johnston, C., Gothe, H. J., Massie, D., Connolly, B., & Müller, B. (2014). Operons are a conserved feature of nematode genomes. *Genetics*, *197*(4), 1201–1211. https://doi.org/10.1534/genetics.114.162875

Pouchkina-Stantcheva, N. N., & Tunnacliffe, A. (2005). Spliced leader RNA-mediated trans-splicing in phylum rotifera. *Molecular Biology and Evolution*, *22*(6), 1482–1489. https://doi.org/10.1093/molbev/msi139

Protasio, A. V., Tsai, I. J., Babbage, A., Nichol, S., Hunt, M., Aslett, M. A., de Silva, N., Velarde, G. S., Anderson, T. J. C., Clark, R. C., Davidson, C., Dillon, G. P., Holroyd, N. E., LoVerde, P. T., Lloyd, C., McQuillan, J., Oliveira, G., Otto, T. D., Parker-Manuel, S. J., … Berriman, M. (2012). A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. *PLoS Neglected Tropical Diseases*, *6*(1), e1455. https://doi.org/10.1371/journal.pntd.0001455

Rajkovic, A., Davis, R. E., Simonsen, J. N., & Rottman, F. M. (1990). A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(22), 8879–8883. https://doi.org/10.1073/pnas.87.22.8879

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Riutort, M., Álvarez-Presas, M., Lázaro, E., Solà, E., & Paps, J. (2012). Evolutionary history of the Tricladida and the platyhelminthes: An up-to-date phylogenetic and systematic account. *International Journal of Developmental Biology*, *56*(1–3), 5–17. https://doi.org/10.1387/ijdb.113441mr

Ross, L. H., Freedman, J. H., & Rubin, C. S. (1995). Structure and expression of novel spliced leader RNA genes in Caenorhabditis elegans. *The Journal of Biological Chemistry*, *270*(37), 22066–22075. http://www.ncbi.nlm.nih.gov/pubmed/7665629

Rossia, A., Jackb, E. J. R. A., & Alvarado, A. S. (2014). Molecular cloning and characterization of SL3: A stem cell- specific SL RNA from the planarian Schmidtea mediterranea. *Gene*, *533*(1), 156–167. https://doi.org/doi:10.1016/j.gene.2013.09.101. Molecular

Roy, S. W., & Irimia, M. (2009). Splicing in the eukaryotic ancestor: form, function and dysfunction. *Trends in Ecology and Evolution*, *24*(8), 447–455. https://doi.org/10.1016/j.tree.2009.04.005

Sather, S., & Agabian, N. (1985). A 5' spliced leader is added in trans to both alpha- and beta-tubulin transcripts in Trypanosoma brucei. *Proceedings of the National Academy of Sciences of the United States of America*, *82*(17), 5695–5699. https://doi.org/10.1073/pnas.82.17.5695

Siegel, T. N., Gunasekera, K., Cross, G. A. M., & Ochsenreiter, T. (2011). Gene expression in Trypanosoma brucei: Lessons from high-throughput RNA sequencing. *Trends in Parasitology*, *27*(10), 434–441. https://doi.org/10.1016/j.pt.2011.05.006

Stover, N. A., & Steele, R. E. (2001). Trans-spliced leader addition to mRNAs in a cnidarian. *Proceedings of the National Academy of Sciences*, *98*(10), 5693–5698. https://doi.org/10.1073/pnas.101049998

Tessier, L., Keller, M., Chan, R. L., Fournier, R., & Weil, J. (1991). Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena. *The EMBO Journal*, *10*(9), 2621–2625.

Tourasse, N. J., Millet, J. R. M., & Dupuy, D. (2017). Quantitative RNA-seq meta-analysis of alternative exon usage in C. elegans. *Genome Research*, *27*(12), 2120–2128. https://doi.org/10.1101/gr.224626.117

Vandenberghe, A. E., Meedel, T. H., & Hastings, K. E. M. (2001). mRNA 5'-leader trans-splicing in the chordates. *Genes & Development*, *15*(3), 294–303. https://doi.org/10.1101/gad.865401.Evidence

Wenzel, M., Johnston, C., Müller, B., Pettitt, J., & Connolly, B. (2020). Resolution of polycistronic RNA by SL2 trans-splicing is a widely conserved nematode trait. *Rna*, *26*(12), 1891–1904. https://doi.org/10.1261/RNA.076414.120

Wright, C. J., Smith, C. W. J., & Jiggins, C. D. (2022). Alternative splicing as a source of phenotypic diversity. *Nature Reviews Genetics*, *23*(11), 697–710. https://doi.org/10.1038/s41576-022-00514-4

Zhang, H., Hou, Y., Miranda, L., Campbell, D. A., Sturm, N. R., Gaasterland, T., & Lin, S. (2007). Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences*, *104*(11), 4618–4623. https://doi.org/10.1073/pnas.0700258104

Capítulo 2: Detección de nuevos Spliced Leaders con SLFinder

Resumen

Spliced Leader trans-splicing es un importante mecanismo para la maduración de ARNm en varios linajes de eucariotas, incluyendo varios grupos de parásitos de gran importancia médica y económica. Sin embargo, su estudio a lo largo del árbol de la vida se ve severamente limitado por dificultades en la identificación de las secuencias SLs utilizadas en el *trans*-splicing. En este trabajo presentamos a SLFinder, una pipeline de 4 pasos diseñada para identificar de novo secuencias SL candidatas, haciendo mínimas asunciones de sus propiedades. La pipeline toma transcriptomas ensamblados *de-novo* y un genoma de referencia como input y permite al usuario intervenir en varios puntos para corregir características inesperadas del set de datos. La estrategia y su implementación fueron probadas con datos reales de RNAseq provenientes de especies con y sin SL *trans*-splicing.

SLFinder es capaz de identificar SL candidatos con precisión y en un tiempo razonable. Es de particular utilidad en especies sin secuencias SLs conocidas, generando secuencias candidatas para futuro refinamiento y validación experimental. En el marco de esta tesis, este paper realizó dos grandes contribuciones. Primero generar una herramienta pensada desde su concepción en el manejo de datos subóptimos con la que garantizar los medios para realizar el resto de la investigación. SLFinder no necesita conocer el largo de un SL antes de encontrarlo, su secuencia o número total. En caso de ser necesario es posible utilizarla sin un genoma de referencia y filtrar las secuencias candidatas por otros medios. Y en segundo lugar, ofreció un primer pantallazo a las dificultades a afrontar en la identificación de loci SL-ARN. No es suficiente con identificar la región *Leader*, hay que filtrar loci asociados a elementos transponibles.

Mis contribuciones en este *paper* consisten en el diseño, implementación y validación de la pipeline SLFinder. Esto incluye la escritura del software y la estrategia empleada para evaluar su eficacia en la identificación de secuencias SLs y filtrar posibles contaminantes.

Calvelo et al. BMC Bioinformatics (2020) 21:293
https://doi.org/10.1186/s12859-020-03610-6

BMC Bioinformatics

SOFTWARE Open Access

SLFinder, a pipeline for the novel identification of splice-leader sequences: a good enough solution for a complex problem

Javier Calvelo[1,2,3], Hernán Juan[1], Héctor Musto[2], Uriel Koziol[3] and Andrés Iriarte[1*]

* Correspondence: airiarteo@gmail.com
[1]Laboratorio de Biología Computacional, Departamento de Desarrollo Biotecnológico, Instituto de Higiene, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay
Full list of author information is available at the end of the article

Abstract

Background: Spliced Leader trans-splicing is an important mechanism for the maturation of mRNAs in several lineages of eukaryotes, including several groups of parasites of great medical and economic importance. Nevertheless, its study across the tree of life is severely hindered by the problem of identifying the SL sequences that are being trans-spliced.

Results: In this paper we present SLFinder, a four-step pipeline meant to identify de novo candidate SL sequences making very few assumptions regarding the SL sequence properties. The pipeline takes transcriptomic de novo assemblies and a reference genome as input and allows the user intervention on several points to account for unexpected features of the dataset. The strategy and its implementation were tested on real RNAseq data from species with and without SL Trans-Splicing.

Conclusions: SLFinder is capable to identify SL candidates with good precision in a reasonable amount of time. It is especially suitable for species with unknown SL sequences, generating candidate sequences for further refining and experimental validation.

Keywords: SL trans-splicing, De novo assembly, RNAseq data

Background

Spliced Leader (SL) trans-splicing, that is, the incorporation of a short RNA (the spliced leader) on the 5′ end of a different transcript, is an important but poorly understood part of the mRNA maturation process of many eukaryotic lineages. SL genes are often encoded in tandem repeats measuring a few kilobases, close to 5S rRNA genes [1] but there are exceptions (e.g. [2]). SL transcript sequences can be divided into two regions: an exon like sequence that remains in the final trans-spliced transcript and an intron that usually contains a canonical Sm-protein-binding site (see for exceptions: [3, 4]), separated by a splice donor site [1].

While there is no clear pattern with regards to specific metabolic pathways or functions for the transcripts that are subject to these mechanisms [5–7], SL trans-splicing participates in important regulatory functions such as operon resolution, 5'UTR edition and the incorporation of modified 5′ cap [1, 5, 6, 8, 9]. At least in some cases, it has been shown that it can also play an important role generating different isoforms by facultative SL trans-splicing (e.g. [10]) or alternative SL trans-splicing acceptor sites [11]. The number of classes of SLs (i.e., Spliced Leaders with a distinct sequence) and the number of copies in each genome varies among different organisms, and at least in some cases, there is evidence of specialization. For example, *Caenorhabditis elegans* has two distinct types of SLs, one (SL-1) is incorporated at the start of the operon and the other (SL-2) is used to resolve the downstream coding sequences into different transcripts [8]. In the planarian *Schmidtea mediterranea* it has been described one particular SL that is expressed preferentially on stem cells [12].

The molecular mechanisms involved are poorly understood and are subject of continuous research (e.g. [13]) but evidence indicates that it's closely related to cis-splicing, with several shared regulatory signals [1, 14]. All identified SL transcripts share a similar secondary structure to the snRNAs (i.e., U1, U2, U4, and U5) that form the spliceosome, suggesting a common evolutionary history [1, 5, 14]. However, its evolution is a topic of debate among researchers, mainly due to the uneven distribution of SL Trans-Splicing across the phylogeny of eukaryotes [1, 5, 14].

So far SL Trans-Splicing has been reported in groups such as Euglenozoa [15, 16], Platyhelminthes [17, 18], Nematoda [19, 20], Urochordata [21], Rotifera [22], Cnidaria [23], Dinoflagellata [24], Crustaceans [25] and Amoebozoa [4]. However, it is absent in others such as vertebrates, insects, plants, Fungi and several protists [14, 26]. This brings the question if the mechanism has independently evolved several times (i.e., by modification of cis-splicing) or was present on the eukaryotic last common ancestor and lost many times [1, 5, 14, 25], with the discussion going back and forth as the mechanism is identified in new taxonomic groups (e.g. [25]).

When analyzing a new organism, the first obvious step is the identification of potential SL sequences on the mRNAs. This does not only allow to identify the presence of the mechanism in the group but having these sequences opens the possibility to use methodologies tailored toward SL Trans-Spliced transcripts. For example, "SL Trapping" [27] or "SL-seq" [28], both modified Next Generation Sequencing (NGS) protocols, allow an enriched sequencing of SL trans-spliced transcripts (e.g. [11]). Other approaches exist, but they either focus on identifying trans-splicing acceptor sites on the coding genes (e.g. [29, 30]), then requiring to be experimentally validated and providing no information about the specific SLs involved; or they require known SL sequences [31–34].

Unfortunately, the identification of SL sequences can be a significant roadblock due to technical limitations, specifically the reduced coverage of reads toward the transcript 5′ end that is typical of poly-A capture [35]. Combined with low or null sequence conservation across different phyla [5], within phylum variability, and several species with multiple SL classes with high nucleotide diversity [31, 36–38], these difficulties make the identification of SL sequences in new species a non-trivial problem. Several authors have tested different approaches to this problem with different degrees of automatization and reliance on previously known information (e.g. [5, 25, 31, 39–44]).

Nevertheless, currently, there is no standardized protocol or analysis pipeline that allows the identification of putative SL (pSL) sequences, that is why often novel SL sequences are discovered almost by chance (e.g. [31]).

Here we present SLFinder, a four-step pipeline implemented in bash designed to facilitate the identification of novel SL exonic sequences from standard NGS RNAseq data (mRNA enriched by poly-A capture following a non-strand specific protocol). The pipeline first limits the potential candidates and provides a unifying command-line environment where parameters can be quickly adjusted to fit each species and dataset characteristics; while making limited assumptions on the SL sequence and mechanism, namely: 1) the SL sequence is located in the 5′ end of the transcript, 2) the SL sequence is present on the transcripts of many genes, 3) The sequence is not a palindrome, 4) There is at most one copy of it on each transcript, and 5) When mapped to the genome there is a canonical splicing donor site after the 3′ end (GT). In addition, and despite its limitations, the analyses are designed with transcriptome sequences generated using the widely used poly-A capture protocol so it can be applied to a larger group of organisms.

In order to evaluate SLFinder, we analyzed RNAseq data from several species with and without known SL Trans-Splicing and compared our predictions with the reported sequences in the bibliography. To better represent the intended use of the software on the identification of novel SL sequences, no manual intervention was carried out to curate the results (contrary to our recommendations when using this software).

Implementation

Mandatory input data

For these analyses three inputs are necessary: 1) one or more assembled transcriptomes from the species of interest, following a de novo approach with Trinity [45, 46]; 2) a reference genome from the species and 3) an external database with Protein or cDNA (ideally from the same species or a reliable database such as SwissProt from Uniprot) for loci annotation.

Trinity can be replaced as the assembler following the instructions included on the manual, however, it is important to conduct an entirely de novo strategy to ensure that reads containing the SL sequence are not excluded of the final transcript. In addition, while we didn't thoroughly test its effect, read normalization based on kmer frequency (e.g. [46]) is discouraged since reads from multiple transcripts will have the SL sequence and could potentially be partially discarded in some datasets. The longer the species SL sequences, the greater this issue is expected to be.

Strategy

Basically, the pipeline recovers potential SL exonic 3′ regions by looking for frequent kmers on the transcripts ends, extends them as much as possible by attempting to assemble them in contigs, and then filters out likely false positives based on sequence orientation, abundance, genomic data and overlap with annotation to known proteins. In practice, however, there are two issues to solve to implement such a straightforward approach. First, false positives due to biased kmer counts, which can be a result of the reconstruction of more than one isoform for a gene and other biological factors such as

very similar transcripts from different genes of the same multigenic family. Second, the loss of strand information during sequencing in standard RNAseq sequencing, so that each transcript can be assembled either as the sense strand or as its reverse complement. Both main issues are addressed by our pipeline.

The pipeline overview is presented on Fig. 1. First, the redundancy in the de novo assembly transcriptome is reduced in SLFinder-Step0, hereafter referred to as Step0, by retrieving the 5′ and 3′ ends of the longest isoforms of each gene. Isoforms of the same gene are identified based on Trinity's contigs name convention. Alternative strategies can be implemented (e.g. clustering based on sequence identity) following the manual instructions. Regardless of the chosen method, once redundant sequences are filtered the next step is to identify SLs among the more commonly observed sequences in the transcripts ends. In an ideal situation, the SL sequence should be located at the exact

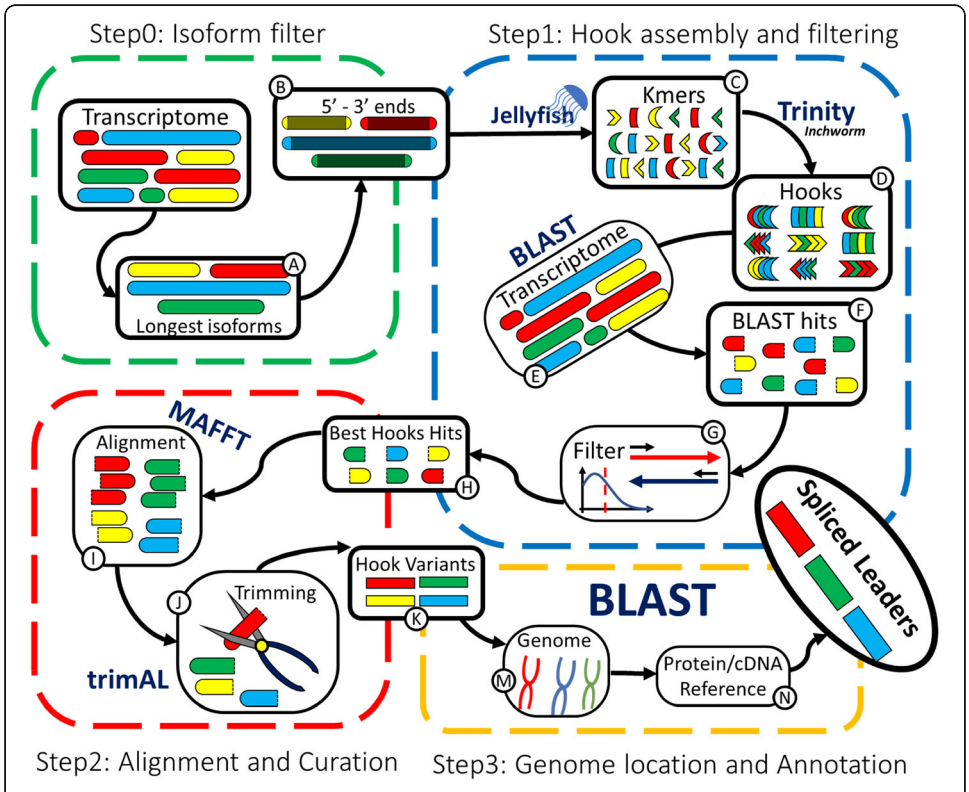

Fig. 1 SLFinder workflow. The pipeline is divided into four scripts (Step0 to 3) to facilitate user intervention. Step0 reduces a Trinity assembly redundancy based on the contigs IDs and extract transcripts ends. Step1 generates Hook sequences for the putative Splice Leaders based on the Kmers present on the filtered sequences generated in Step0, identifies matching sequences in the original assembly and selects the Best Hooks based on their overall sequence orientation and frequency. The Best Hook Hits sequences are then aligned and trimmed in Step2 to generate Hook Variants, and Step 3 localizes them on a Reference genome and classifies hits according to the presence of a potential Donor Site. In detail, longer isoforms are retrieved from the transcriptome (A), both terminals regions are keep (B) and kmers are counted with Jellyfish and selected according to their abundance (C). Selected kmers are assembled into Hooks using Inchworm (D) and then Hooks are blasted against the original unfiltered transcriptome (E) and then filtered according to coi and abundance. The Best Hook matching sequences are retrieved (H). These sequences are then aligned with MAFFT (I) and trimmed automatically with trimAL (J), results are Hook Variants that represent possible versions of the repeated sequence that generated the Hooks (including both real polymorphisms and sequencing errors) (K). Step3 locates locus codding for the Hook Variants in the species genome using BLAST and identifies potential donor sites (M). Finally, putative coding sequences are discarded using a third BLAST search against a protein-coding or protein sequences reference databases (N)

beginning of the assembled transcripts and cleaner results should be obtainable by retrieving from the transcript end a fragment similar in size to the expected length of the SL. However, we noticed that the assemblies used to validate this pipeline often presented non-conserved, mainly low-quality, sequences that preceded the known SL sequence. Instead, Step1 achieves this by counting the kmers present on the filtered sequences (and their reverse complement) with Jellyfish software [47]. Those kmers observed below a given threshold are discarded (by default 0.0005% of the total contigs after filtering, in practice ≈10 contigs, depending on the dataset), and then assembled in longer sequences with Inchworm, Trinity's first module. The resulting sequences, hereafter referred to as "Hooks", are a collection of true SL sequences (if present) and every other common sequence found on the filtered transcriptome.

To further narrow down candidates, Step1 analyzes the location and orientation of matching sequences in the original transcriptome. A Blast search [48] is conducted to "fish out" similar sequences among the transcriptomic contigs (with the "-task blastn-short" option and "-evalue 1e-2"). Hooks are selected according to their number of hits and sequence orientation in the assembled transcriptomes. Since we are working without information on the strand, the transcripts can be assembled either sense or anti-sense; and so, can be the Hooks that are generated from these transcripts. However, if our assumption 3 holds (i.e., the sequence is not a palindrome) and the Hook represents a true SL (meaning that it is located at the 5′ end on the transcript) its hits on the transcriptome should be found in two mutually exclusive configurations, depending on the orientation of the Hook: 1) forward-oriented at the Start of a sense assembled transcript or reverse oriented at the End in an anti-sense one if the Hook was generated in a sense configuration (named Sf and Er orientations, respectively); or 2) reverse oriented at the start of a sense assembled transcript and Forward oriented at the end of an anti-sense one (named Sr and Ef orientations, respectively) (Fig. 2a).

With this in mind, we created a simple consistency orientation index (coi) to evaluate each potential Hook (Fig. 2). A coi equal to 1 means that the Hook's hits are all oriented in one and only one of the valid configurations. Testing, however, shows Hooks for known SLs can have some hits that do not follow these rules but their coi is always high (i.e., above 0.95; see Results). In addition, tests show that Hooks with few hits on the transcriptome often have a high coi by chance, even when they are not SL sequences. To compensate we also introduce an Observation Count Cutoff (occ) filter that is simply the median of all identified Hooks. Finally, transcripts with multiple hits for the same Hook are excluded and reported separately for user inspection. These transcripts may represent chimeric sequences generated during the assembly process [49, 50] or short repeated sequences. Hooks that pass these filters are selected for further analysis.

The sequences of these selected Hooks, hereafter referred to as Best Hooks, are retrieved for further analysis. The transcript's ends with a BLAST match (with the "-task blastn-short" option and "-evalue 1e-2") to each Best Hook are retrieved (from the transcript end until two bases after the match end in order to recover as much sequence from the pSL sequence).

The next filtering step in the pipeline consists of locating and analyzing genomic loci matching the Best Hooks from which they are potentially transcribed (putative SL genes). However, first, it is necessary to address three issues: high redundancy, noisy

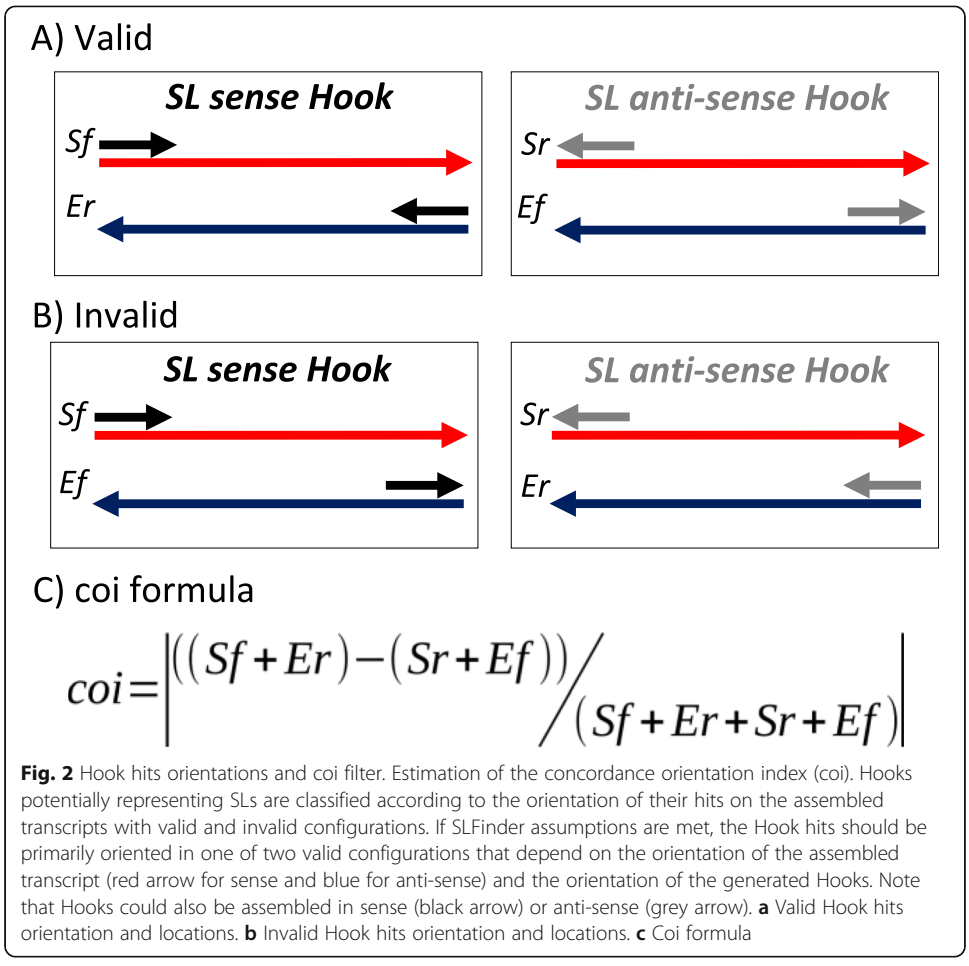

Fig. 2 Hook hits orientations and coi filter. Estimation of the concordance orientation index (coi). Hooks potentially representing SLs are classified according to the orientation of their hits on the assembled transcripts with valid and invalid configurations. If SLFinder assumptions are met, the Hook hits should be primarily oriented in one of two valid configurations that depend on the orientation of the assembled transcript (red arrow for sense and blue for anti-sense) and the orientation of the generated Hooks. Note that Hooks could also be assembled in sense (black arrow) or anti-sense (grey arrow). **a** Valid Hook hits orientation and locations. **b** Invalid Hook hits orientation and locations. **c** Coi formula

sequences coming from sequencing and assembly errors, and imprecise pSL delimitation. Without knowing the SL sequence there is no reliable way to address these problems with a one-base precision, nevertheless, Step2 minimizes them by first clustering all sequences according to sequence identity with CD-HIT-EST [51] with a 100% identity threshold by default, followed by alignment with MAFFT [52] using the accuracy oriented method G-INS-I "--globalpair --maxiterate". Finally, sequences are automatically trimmed with trimAL [53]. The resulting sequences, referred to as "Hook Variants", represent possible versions of the repeated sequence that generated the Hook (including both real polymorphisms and sequencing errors). Depending on the data, it might be necessary to re-run this step several times with different parameters or even manually curate the sequences before continuing with Step3 (see the software manual for detailed instructions). To facilitate this process, Step2 also generates sequence logos before and after trimming with Weblogos3 [54].

Step3 carries out a BLAST (–task blastn-short) search of the Hook Variants against the provided reference genome to identified pSL coding loci. Since some level of noise is expected in the Hook Variants sequences, even when they represent true SL (see below), the BLAST search is configured with a 100% identity threshold, ungapped, and a high query coverage (90% by default). In practice, these thresholds allow mismatches in the terminal region of the Hook Variant. Once identified, Step3 searches for the

existence of a potential donor site and attempts to annotate the region with an external CDS or protein reference with either blastn or tblastx. As a final fail-safe to check the inaccuracy in the pSL delimitation, Step3 takes the following considerations when reporting a potential donor site: 1) It analyses 4 bp surrounding each Hook Variant 3′ end hit in the genome (excluding mismatches in the extremes) looking for a possible splice donor site ("GT"). If one "GT" is found, step3 reports either "5prima" or "3prima" depending on the hit orientation, simplified to "Clear donor site" in this paper. 2) If the longest matching Hook Variant with a donor site overlaps with possible splice donor site (i.e., the sequence ends with a "G" or "GT" that matches with the splice donor site) an "*" is included in the report to indicate that manual inspection is advised. 3) If a potential donor site is found in fewer than 80% of Hook Variants matching a locus, the site is reported as "Unclear". Finally, a BLAST search between the region surrounding each locus and the provided Protein/cDNA reference dataset is conducted, and loci with matches are discarded. In every step the user can check the Hook Variants and blast results to reconsider or inspect some discarded Hooks.

A putative SL coding Locus was considered valid if a potential donor site was identified and there were no known protein-coding sequences located close to the locus (by default 100pb, this parameter can be changed by the user). Sequences for loci with and without a clear donor site are clustered with CD-HIT-EST in pSL sequences (100% identity Threshold). The final output also includes multiple sequence alignment of each locus done with MAFFT (G-INS-I "--globalpair --maxiterate") and its original Hook Variants to facilitate manual inspection.

Test data

Test data was selected from species according to known presence or absence of SL Trans-splicing, the existence of a reference genome and availability of RNAseq following a poly-A capture protocol. The final species list comprised *Aplysina aerophoba*, *C. elegans* [19], *Ciona intestinalis* [21], *Drosophila melanogaster*, *Hydra vulgaris* [23], *Mus musculus*, *Saccostrea glomerata*, and *Schistosoma mansoni* [18]. *Schistosoma mansoni,* has been reported to have a single SL class with a long sequence (36 bp), represents an ideal scenario to test SLFinder. Meanwhile, *C. intestinalis* with single short SL (16 pb) allows investigating how the pipeline behaves with shorter sequences. Finally, *C. elegans* and *H. vulgaris* have multiple SL sequences (some of them with known sequence diversity among their coding SL-RNAs, e.g. SL2 in *C. elegans*) which will test SLFinder ability to identify and retrieve different SLs when present.

Transcriptomic and genomic data used on these analyses are detailed in Table 1. Non-control samples from experimental studies (i.e. response to pathogens or other stimulus) were discarded. Genomic locus annotation was carried out with the Swiss-Prot database from Uniprot (Downloaded 01/05/2019). In addition, since several genome assemblies are available for *S. mansoni*; mostly based on Protasio et al. 2012 work [55] but improved and annotated following different methodologies for their curation and annotation; SLFinder was tested using 2 reference genomes: one from Wormbase Parasite (WBPS) improved with PacBio data and one from GeneDB that is more fragmented but with several SL-RNA genes annotated.

Table 1 Datasets utilized to validate and evaluate SLFinder

Species	Taxon	RNAseq BioProject	Ref. Genome Assembly	Reported SLs
A. aerophoba	Porifera	PRJEB26562	GCA_900275595.1[a]	No
C. elegans	Nematoda	PRJNA270896	PRJNA13758[b]	Yes
C. intestinalis	Urochordata	PRJNA396771	GCF_000224145.3[a]	Yes
D. melanogaster	Insecta	PRJNA318586	GCF_000001215.4[a]	No
H. vulgaris	Cnidaria	PRJNA497966	Hm105[c]	Yes
M. musculus	Vertebrata	PRJNA319673	GCF_000001635.26[a]	No
S. glomerata	Molusca	PRJNA487836	GCA_003671525.1	No
S. masoni	Plathelminthes	PRJNA225599	PRJEA36577[b] and GeneDB	Yes

[a] Available on NCBI database
[b] Available on WBPS database
[c] Hydra 2.0 Genome Project

Read quality for RNAseq data was assessed with FastQC [56] and low quality bases along with adapter sequences were removed with Trimmomatic v0.36 [57] (options: ILLUMINACLIP: TruSeq3-PE.fa:2:30:10, SLIDINGWINDOW: 5:20 and MINLEN: 50). Transcriptomes were de novo assembled with Trinity v2.8.3, without read normalization.

Bioinformatic analysis and pipeline evaluation

Analyses were carried out in a desktop computer with 96Gb of RAM and 32 threads/16 cores (only 4 threads were used on each run). Program versions used are listed on Table 2 with default parameters (with exception of *C. intestinalis*). Pipeline accuracy was tested by sequence comparisons with known SL sequences (Additional file 1), verifying the match of the predicted SL locus with the annotated SL within 100 bp range. This comparison was done using gffread [58]. In addition, each potential locus was manually inspected, and "seqkit locate" was utilized to verify the transcripts carrying specific pSL sequences in order to detect and categorize artefacts. Figures of sequence alignments were generated with BioEdit v7.0.5.3 [59].

Results

A total of 32 transcriptomes (9 from *A. aerophoba*, 6 from *C. elegans*, 3 from *C. intestinalis*, 4 from *D. melanogaster*, 5 from *H. vulgaris*, 2 from *M. musculus*, 1 from *S.*

Table 2 List of programs and software packages utilized by SLFinder, including the version utilized in this paper and the basic tasks they carry out

Program	Version	Tasks
Blast	v2.6.0	Sequence searches against Transcriptome assemblies, Genome and Protein reference database.
cd-hit-est	v4.7	Sequence clustering to simplify results and reduce runtimes
Jellifish	v2.2.6	Kmer counts
MAFFT	v7.307	Sequence Alignment
Seqkit	v0.10.0	Basic sequence manipulation
trimAl	v1.2rev59	Hook Variant generation by automatic trimming
Trinity	v2.8.3	Hook assembly from Kmers
Weblogos	v3.6.0	Sequence Logos generation to facilitate manual curation

glomerate, and 2 *S. mansoni*) were assembled and analyzed (Basic descriptor metrics are shown in Additional file 2). Running times per step where highly dependent on the dataset (Table 3) mainly depending on the number of reads to process. No Hook sequence passed the coi filter in Step1 for the species without known SLs *A. aerophoba*, *D. melanogaster*, *M. musculus* and *S. glomerata* (Additional file 3).

Positive results were identified for the species *C. elegans*, *Hydra vulgaris* and *S. mansoni*, all with previously described SL. SLFinder also identified the SL reported for *C. intestinalis* after changing the parameters to account for short SLs (15-base kmer length, 14 Inchworm assembly kmer, and no filtering according to the median count value).

In the following sections we will describe the results obtained by each Step of SLFinder (Step 1 Hook generation and filtering, Step 2 Hook Variant trimming, and Step 3 putative SL (pSL) loci identification) on each positive dataset. Since the intent of this software is novel SL identification, we will focus on features of SLFinder reports that depart from the true SL sequence (e.g. longer/shorter sequences than expected and potential false positives results).

Caenorhabditis elegans dataset

A total of 13 hooks were generated in Step1, three of which passed both the coi and the occ filters and resulted in 246 different Hook Variants after Step2. Comparison with known SL sequences showed that the Best Hooks "a1–9915" and "a11–1448" corresponded to the known SL-1, while "a10–178" to SL-2 (Fig. 3a).

Step3 identified 26 putative pSL loci in the reference genome, 18 of which were previously reported as SL-RNA genes in the genome annotation (Additional file 4). Thirteen pSL were reported as having a clear potential donor site and were later clustered into 8 sequences; hereafter referred to as Celegans_pSL-(1 to 8). Another 10 loci were reported as Unclear due to the presence of several bases in the 3′ region in several variants for the Hook "a1–9915" that overlapped with the splice donor site (Additional file 5). Most of these loci were located on Chromosome V in a cluster of ≈13 kb and were grouped into a single sequence identical to Celegans_pSL-7. In addition, Locus-5 and -26 were reported without a donor site and Locus-25 was not analyzed because SLFinder failed to determine its orientation due to low count numbers in the transcriptome. Manual inspection showed that both Locus-25 and Locus-26 have a potential donor site masked by a three base extension in the 3′ end (GTA) of the only

Table 3 SLFinder steps performance for all datasets

Data Set	Step0	Step1	Step2	Step3	Total
A. aerophoba	32 m 12 s	0 m 34 s	X	X	32 m 46 s
C. elegans	4 m 23 s	0 m 24 s	13 m 25 s	1 m 08 s	19 m 46 s
C. intestinalis	4 m 29 s	1 m 54 s	0 m 03 s	0 m 43 s	7 m 09 s
D. melanogaster	5 m 36 s	0 m 20s	X	X	5 m 56 s
H. vulgaris	15 m 17 s	57 m 26 s	19 m 58 s	14 m 35 s	1 h 47 m 16 s
M. musculus	7 m 14 s	2 m 10s	X	X	7 m 24 s
S. glomerata	24 m 24 s	0 m 25 s	X	X	24 m 59 s
S. mansoni	6 m 53 s	0 m 38 s	0 m 06 s	3 m 03 s	10 m 40s

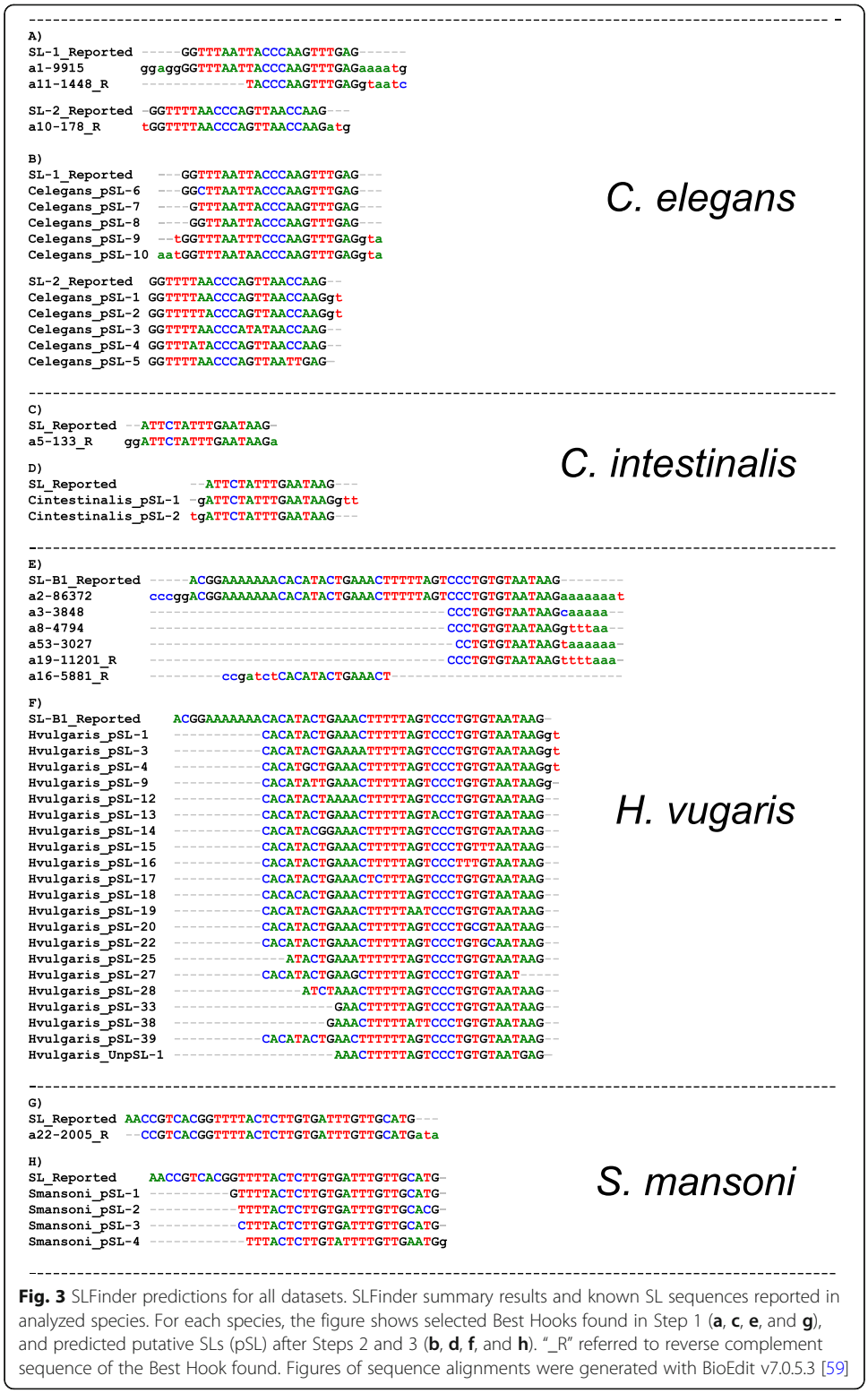

Fig. 3 SLFinder predictions for all datasets. SLFinder summary results and known SL sequences reported in analyzed species. For each species, the figure shows selected Best Hooks found in Step 1 (**a**, **c**, **e**, and **g**), and predicted putative SLs (pSL) after Steps 2 and 3 (**b**, **d**, **f**, and **h**). "_R" referred to reverse complement sequence of the Best Hook found. Figures of sequence alignments were generated with BioEdit v7.0.5.3 [59]

matching Hook Variant for each, hereafter referred to as Celegans_pSL-9 and -10, respectively (see similar cases in Additional file 5).

Potential Spliced Leaders Celegans_pSL-6, – 7, – 8, – 9, and – 10 match the previously described SL-1 and their nucleotide differences were limited to the 5′ region, whereas

Celegans_pSL-1, − 2, − 3, − 4 and − 5 represent different variants of SL-2 and are slightly more diverse in their nucleotide sequences (Fig. 3b). These observations are in concordance with the genome annotation and previous results for *C. elegans* [2].

Site-specific observations of these loci are included in Additional file 4. Of particular relevancy are Locus-12 and -20, both display partial repetitions of SL-1 following the reported hit (Additional file 6). Verification of the functionality of these SLs loci is beyond the scope of this paper and the capabilities of SLFinder, a pattern search against the reads only identified six read pairs bearing Locus-12 repeat across all samples.

In Summary, SLFinder identified both SL classes, SL-1 and SL-2, previously reported for *C. elegans* and located several of their described SL-RNAs loci (10 for SL-1 and 8 for SL-2), in addition to five not previously annotated copies of SL. While verifying the functionality of these new SL-RNAs is beyond the scope of this paper, our results suggest that at least two of them (Locus-12 and -20) are presumed to be pseudogenes due to the presence of fragments from SL-1 following the 3′ end.

Ciona intestinalis dataset

Using the modified parameters (15-base kmer length, 14-base Inchworm assembly kmer and removing occ filtering), Step1 generated 81 Hooks but only "a5–133" passed the coi filter. Twenty-three Hook Variants were identified for "a5–133" in Step2. Results show that the Hook matches the sequence of the previously reported SL [21] (Fig. 3c).

Step3 identified 38 putative loci in the genome, 23 of which have a predicted protein-coding gene in the matched region according to the available annotation (Additional file 7). Long non-coding RNA (lncRNA) are annotated surrounding Locus-8, − 12, and − 15 in this dataset (XR_717275.3, XR_003396022.1 and XR_003396339.1 respectively), but their functions are unknown and only Locus-12 is encompassed by the lncRNA included by its hit. Fourteen pSL were reported with a clear potential donor site and were grouped into two clusters; hereafter named as Cintestinalis_pSL-1 and -2; that differ on their extension toward both sequence ends (Fig. 3d). The former shows a 3′ extension "GTT" which overextends the expected donor site for the SL and ends next to another "GT" in the genome. Detailed observations are included in Additional file 7.

Despite its shorter size, once the software parameters were properly fine-tuned, SLFinder was able to recover the reported SL sequence for this species. However, the results are not as clear as other datasets analyzed, indicating that these conditions are near the limits of what is possible to obtain with this strategy.

Hydra vulgaris dataset

For this species Step1 generated 31 Hooks, 6 of which passed both coi and frequency filters and their matches in the transcriptome were processed on 385 Hook Variants. Comparison with known SL sequences for the species shows that the longest Hook "a2–86,372" matches SL-B1 (reported in [36]); while hooks "a3–3848", "a8–4794", "a16–5881", "a19–11,201" and "a53–3028" match only the terminal region (Fig. 3e).

Unfortunately, the 5′ region of the observed Hook Variants, from 16 to 32 bases, was lost in Step2 during trimming (Additional file 5d).

Step3 identified 239 loci in the genome many of which were found in close proximity to annotated protein-coding regions (Additional file 8). 93 loci were reported with a clear donor site ("Clear") and 59 with an unclear donor site ("Unclear"). The former was clustered in 37 pSL sequences, Hvulgaris_pSL-(1 to 37), and the latter in 10, Hvulgaris_UnpSL-(1 to 10). Hvulgaris_UnpSL-4 has an identical sequence to Hvulgaris_pSL-1, Hvulgaris_UnpSL-3 to Hvulgaris_pSL-6, and Hvulgaris_UnpSL-7 to Hvulgaris_pSL-9 (Additional file 9). In addition, 5 loci were not analyzed because SLFinder failed to determinate their orientation due to low counts in the transcriptome. Among these, Locus-46 and -112 display a potential donor site and are included in the further discussion as Hvulgaris_pSL-38 and -39 respectively.

The manual inspection revealed several issues that suggest they are most likely non-functional versions of SLs that guarantee further analysis. For the purposes of presenting the tool, however, they were considered non-functional. Removing them reduces the pSL unique sequences to 21 (Fig. 3f) (see the full set of pSLs generated by SLFinder in Additional file 9 and detailed observations in Additional file 8). Note that many pSL loci displayed a donor site that overlaps with the known last base of the SL (as previously described for *C. elegans*), while others presented extensions that led to an alternative "GT" (as previously described for *C. intestinalis*) without including the expected donor site. While is possible that the latter pSLs represent longer than already reported SL sequences, testing this will require additional studies that are beyond the scope of this paper. Furthermore, an inspection of the transcripts bearing these sequences indicates that these pSL loci match some transcripts for several bp after the pSL sequence (Data not shown) raising further doubts on their functionality. Lastly, 28 loci showed partial repeats of SL sequences, including some of the previously reported SLs (SL-D, –F, and -G) that were not recovered by SLFinder (Additional file 6).

In summary, *H. vulgaris* was the most complex dataset analyzed, with several potential pseudogenes for the SL-RNAs identified. This is, likely in no small extent, related to their complex evolutionary history [36]. Unfortunately, SLFinder failed to identify the other 6 SL reported for the species [36]. A pattern search with seqkit locate of the terminal region of these SLs on the original fastq files indicates a marginal presence of SL-B2, SL-B3, SL-B4, SL-D and SL-G in the dataset, so the most probable cause of this false negatives is their low prevalence in the analyzed RNAseq data (Data not shown).

Schistosoma mansoni dataset

One Hook ("a22–2005") out of 30 generated in Step1 passed both coi and frequency filters and was then processed into 11 Hook Variants by Step2. Comparison with the known SL sequence for *S. mansoni* shows that this Hook represents the reverse complement of the described SL in almost its entirety (Fig. 3g). As with the *H. vulgaris* dataset, part of the 5′ region of the Hook that was recovered was lost during trimming in Step2 due to the poor alignment quality of this region. This could be at least partially explained by missing information and high variability among the retrieved sequences in the transcriptome assemblies (Data not shown).

When using the WBPS reference genome, Step3 identified 132 pSL loci, only 13 in the proximity to protein-coding genes (Additional file 10). Most of them showed a clear donor site and were clustered in 3 groups; hereafter referred to as Smansoni_pSL-(1 to 3). The remaining 9 loci were reported as lacking a potential donor site. This was confirmed by manual inspection in all cases except for Locus-128, in which the donor site was masked by the retention of 3 bp on the 3′ end of the generated Hook Variant; hereafter referred to as Smansoni_pSL-4. All four pSL are shown in Fig. 3h while loci coordinates and observations are reported in Additional file 10. Note that only Smansoni_pSL-1 was encoded by several loci. On the other hand, Smansoni_pSL-2 had a substitution in the terminal ATG of the SL. This ATG was reported as completely conserved in all studied Platyhelminthes (see [5]). A pattern search of the terminal region of this pSL reveals a marginal presence on the reads from both sequenced samples, indicating very low expression of this SL variant in the dataset (Data not shown).

Surprisingly, only 22 pSL loci were identified when using the GeneDB reference genome (Additional file 11). Fifteen of these presented a clear potential donor site and were clustered in the same three pSL classes found using the WBPS reference genome (see above), including Smansoni_pSL-4 (Data not shown). Five pSL coding loci were already reported as SL-RNA coding genes, including one locus that was reported without a donor site because of missing information in the reference genome.

In the case of *S. mansoni*, SLFinder identified the known SLs, including one possible pseudogene, with the only drawback of a partial recovery of its 5′ region. Results also show the importance of the Reference Genome, as illustrated by the number of pSL loci found in the assemblies of WBPS and GeneDB.

Discussion

Considerations when using SLFinder

The strategy presented here, although effective, has shortcomings that originate from the input data and the minimal assumptions regarding the SL sequences. SLFinder requires enough SL exon sequences to be present in the de novo transcriptome assembly. This may be an important issue when considering the widely distributed poly-A enrichment strategy for RNAseq in eukaryotes, nevertheless, our results clearly show that identifying SL sequences and loci is possible in real datasets. Short SL sequences, poor data quality, and the inappropriate reference genome, or a combination of the three may also be issues to consider. See for instance the results of *C. intestinalis* dataset, which could be handled however with specific parameters settings. When dealing with such cases we recommend changing kmer size, ideally using similar organisms a guideline, and annotate every hit for a Hook with a high coi value. Bear in mind that because of these limitations, negative results should not be considered evidence of absence of SL Trans-Splicing.

In addition, the lack of reliance on known SL sequences combined with the approach taken to generate Hook Variants are the source of the issues in identifying the donor site described in Results. Basically, the problem is how to answer the question: “Where does the sequence end when the sequence is unknown?”. SLFinder solves this issue by trimming according to alignment quality and then localizing them in the reference

genome for further pinpointing the SL extension. While most of the issues with automatic trimming (see [60]) don't apply in this context, a side effect of this strategy is the addition of non-SL bp if they are present in enough transcripts, along with a common loss of the SL 5′ region during the trimming of Hook hits (both observed in the Results section). Nevertheless, both drawbacks can be properly addressed with an informed user intervention that is facilitated by SLFinder modularity, either by adjusting trimAL parameters or manually processing the alignments (Note that these modifications might affect Step3 results as some divergent pSL Loci will be lost).

The quality of the reference genome and other biological features of the species play an important role in SLFinder accuracy and performance. As stated before, the reference genome is a key piece of information when pinpointing the pSL sequence and filtering out Hook Variants generated due to sequencing and/or assembly errors. This is clearly shown in the analyses of *H. vulgaris* and *S. mansoni* datasets. On the one hand, SL prediction in *H. vulgaris* was far from straightforward given the high abundance of pSL coding loci found, many of which are likely false positives. This result may be explained, at least in part, by the high prevalence of transposable elements in their genome [36]. In the case of *S. mansoni* the differences observed between WBPS and GeneDB genome assemblies may explain the different results obtained with SLFinder for this species. A better assembly may help identify more SL loci, as is the case of WBPS assembly. Note that PacBio technology was used to improve assembly quality in this assembly [61].

In the absence of a reference genome, the Hook sequences generated during Step1 and the Hook Variants in Step2 offer a good alternative, but it would require validation based on homology (SL sequences from other closely related taxa) or wet lab experimental approaches.

Advantages of SLFinder

Taxon sampling bias has been a constant issue in the study of SL trans-splicing across the tree of life. For example, Bitar et al. study conducted a study based on BLAST searches against public databases and identified mostly SL-1 like sequences in the phylum Nematoda. Results included species like *Globodera rostochiensis* that possess known divergent SL sequences [38] and *Heterodera glycines* for which more SL classes were latter described [31]. SLFinder represents a solution to this problem by providing a straightforward method to identify pSL sequences that is not based on sequence homology.

The use of over-represented kmers to identify regulatory regions is not a new approach for exploratory analysis of DNA sequences [62] and was applied to identify SL sequences before [36]. However, the novel but simple filters implemented in SLFinder allowed the easy recovery of known SL exonic sequences of the four species with this splicing mechanism in just a few hours; and in the case of *C. elegans* and *S. mansoni* even identifying the known SL-RNA coding loci. Only the *C. intestinalis* dataset required a fine-tuning of SLFinder parameters to account for a shorter than expected SL sequences.

Potential SLs sequences identified with this pipeline can be validated through experimental procedures like RT-PCR or 5′ RACE (e.g. [31, 37]) or can be used as input data

for other informatics analyses like the ones implemented by SLQuant [34] and, UTRme [33]. Even a simple pattern search (e.g. [31]) could be used to identify the acceptor genes in order to further analyze mRNA maturation in the species of interest. The identified putative SLs coding loci can be used to further validate the SL-RNA by looking for the sm site or the RNA secondary structure [1, 5, 37].

Conclusion

SLFinder offers a practical alternative for the discovery of novel SL sequences aside from homology searches or fortuitous identification. This modular pipeline was proved with freely available RNAseq data for organisms with and without reported Splice Leader sequences with very good results. Putative SLs found by SLFinder can be later refined regarding their exact length and confirmed through additional bioinformatics analyses and wet lab experiments. This software represents a step forward toward a more comprehensive understanding of the distribution of SL Trans-Splicing in the tree of life, its evolutionary history and importance.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03610-6.

Additional file 1: Supplementary Table 1. Reported SL sequences for the species *C. elegans*, *C. intestinalis*, *H. vulgaris* and *S. mansoni*.

Additional file 2: Supplementary Table 2. Basic descriptors of the analyzed datasets. Data Set, GenBank SRA ID, N° Reads, N° Trimmed Reads, Assembled Bases, Total Transcripts, N50, Median Contig Length, Average Contig Length, Genes predicted by Trinity (Trinity Genes) and GC content.

Additional file 3: Supplementary Table 3. Hook hits obtained for each dataset (Step 1 - F in Fig. 1). Hit counts according to position and orientation in the contig, coi score and observation count cutoff (occ) are indicated. Selected Best Hooks are highlighted in bold. Results for *C. intestinalis* include those obtained with default and custom values (see main text).

Additional file 4: Supplementary Table 4. Putative SLs loci identified by SLFinder in the *C. elegans* dataset. Previously reported SL-RNAs are also indicated.

Additional file 5: Supplementary Figure 1. Common issues to consider when utilizing SLFinder.

Additional file 6: Supplementary Figure 2. Non-functional SL loci found during SLFinder analyses.

Additional file 7: Supplementary Table 5. Putative SL loci identified by SLFinder in the *C. intestinalis* dataset. Annotated lncRNA and protein-coding genes closer than 100 pb are also indicated.

Additional file 8: Supplementary Table 6. Putative SL loci identified by SLFinder in the *H. vulgaris* dataset. Annotated protein-coding genes closer than 100 pb are also indicated.

Additional file 9: Supplementary Figure 3. Sequences of putative SL identified in *H. vulgaris* dataset.

Additional file 10: Supplementary Table 7. Putative SL loci identified by SLFinder in the *S. mansoni* dataset using Wormbase's genome assembly as reference (WBPS). Annotated protein-coding genes closer than 100 bp are indicated.

Additional file 11: Supplementary Table 8. Putative SL loci identified by SLFinder in the *S. mansoni* dataset using GeneDB's genome assembly as reference. Annotated protein-coding genes closer than 100 bp are indicated.

Abbreviations

coi: Consistency orientation index; Ef: Hook match at the End of the transcript in Forward orientation; Er: Hook match at the End of the transcript in Reverse orientation; NGS: Next Generation Sequencing; occ: Observation Count Cutoff; pSL: Putative Splice Leader; Sf: Hook match at the Start of the transcript in Forward orientation; SL: Spliced Leader; Sr: Hook match at the Start of the transcript in Reverse orientation; WBPS: Wormbase Parasite

Acknowledgements

J.C. is a recipient of a doctoral scholarship from Agencia Nacional de Investigación e Innovación (ANII), Uruguay. H.M., U.K. and A.I. are members of the Uruguayan National Researchers System (SNI), and PEDECIBA, Uruguay.

Availability and requirements

Project name: SLFinder
Project home page: https://github.com/LBC-Iriarte/SLFinder.git
Operating system(s): Linux
Programming language: BASH

Other requirements: BLAST 2.6.0 or higher, cd-hit-est 4.7 or higher, Jellifish 2.2.6 or higher, MAFFT 7.307 or higher, Seqkit 0.10.0 or higher, trimAl 1.2rev59, Trinity 2.8.3 or higher and Weblogos 3.6.0 or higher,
License: Creative Commons Attribution License (CC BY 4.0).
Any restrictions to use by non-academics: None

Authors' contributions

J.C. and A.I. conceived of the presented work with support from U.K. and H.M.. J.C., U.K. and A.I. designed the experiments. J.C. and H.J. performed the bioinformatics studies, software's Manual editing and tests. J.C., H.J., U.K., H.M. and A.I. analyzed and interpreted the results. J.C., U.K. and A.I. wrote the manuscript with support from H.M. All the authors discussed the results and commented on the manuscript. The author(s) read and approved the final manuscript.

Funding

This work was supported by grant FCE_3_2016_1_125297 to A.I. from Agencia Nacional de Investigación e Innovación (ANII), Uruguay. The funding agency played no role in the design of the study, analysis, interpretation of data, or in writing the manuscript.

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Laboratorio de Biología Computacional, Departamento de Desarrollo Biotecnológico, Instituto de Higiene, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay. [2]Unidad de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay. [3]Sección Biología Celular, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay.

Received: 16 January 2020 Accepted: 17 June 2020
Published online: 08 July 2020

References

1. Hastings KEM. SL trans-splicing: easy come or easy go? Trends Genet. 2005;21:240–7.
2. Stricklin SL. *C. elegans* noncoding RNA genes. In: WormBook; 2005. https://doi.org/10.1895/wormbook.1.1.1.
3. Ganot P, Kallesøe T, Reinhardt R, Chourrout D, Thompson EM. Spliced-leader RNA trans splicing in a chordate, Oikopleura dioica, with a compact genome. Mol Cell Biol. 2004;24:7795–805.
4. Matsuo M, Katahata A, Satoh S, Matsuzaki M. Characterization of spliced leader trans-splicing in a photosynthetic rhizarian amoeba, Paulinella micropora, and its possible role in functional gene transfer. PLoS One. 2018;13:e0200961.
5. Bitar M, Boroni M, Macedo AM, Machado CR, Franco GR. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. Front Genet. 2013, 199;4(October). https://doi.org/10.3389/fgene.2013.00199.
6. Matsumoto J, Dewar K, Wasserscheid J, Matsumoto J, Dewar K, Wasserscheid J, et al. High-throughput sequence analysis of Ciona intestinalis SL trans-spliced mRNAs: alternative expression modes and gene function correlates. Genome Res. 2010;20:636–45.
7. de Moraes Mourão M, Bitar M, Pereira Lobo F, Paula Peconick A, Grynberg P, Prosdocimi F, et al. A directed approach for the identification of transcripts harbouring the spliced leader sequence and the effect of trans-splicing knockdown in *Schistosoma mansoni*. Mem Inst Oswaldo Cruz. 2013;108:707–17.
8. Pettitt J, Harrison N, Stansfield I, Connolly B, Müller B. The evolution of spliced leader trans-splicing in nematodes. Biochem Soc Trans. 2010;38:1125–30. https://doi.org/10.1042/BST0381125.
9. Pettitt J, Philippe L, Sarkar D, Johnston C, Gothe HJ, Massie D, et al. Operons are a conserved feature of nematode genomes. Genetics. 2014;197:1201–11.
10. Agorio A, Chalar C, Cardozo S, Salinas G. Alternative mRNAs arising from trans-splicing code for mitochondrial and cytosolic variants of Echinococcus granulosus thioredoxin glutathione reductase. J Biol Chem. 2003;278:12920–8.
11. Boroni M, Sammeth M, Gava SG, Jorge NAN, MacEdo AM, MacHado CR, et al. Landscape of the spliced leader trans-splicing mechanism in Schistosoma mansoni. Sci Rep. 2018;8:3877.
12. Rossia A, Jackb EJRA, Alvarado AS. Molecular cloning and characterization of SL3: a stem cell- specific SL RNA from the planarian Schmidtea mediterranea. Gene. 2014;533:156–67.
13. Philippe L, Pandarakalam GC, Fasimoye R, Harrison N, Connolly B, Pettitt J, et al. An in vivo genetic screen for genes involved in spliced leader trans-splicing indicates a crucial role for continuous de novo spliced leader RNP assembly. Nucleic Acids Res. 2017;45(14):8474–83.
14. Lasda EL, Blumenthal T. Trans-splicing. Wiley Interdiscip Rev RNA. 2011;2:417–34.
15. Sather S, Agabian N. A 5′ spliced leader is added in trans to both alpha- and beta-tubulin transcripts in Trypanosoma brucei. Proc Natl Acad Sci U S A. 1985;82:5695–9. https://doi.org/10.1073/pnas.82.17.5695.
16. Tessier L, Keller M, Chan RL, Fournier R, Weil J. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena. EMBO J. 1991;10:2621–5.
17. Brehm K, Jensen K, Frosch M. mRNA trans-splicing in the human parasitic cestode Echinococcus multilocularis. J Biol Chem. 2000;275:38311–8.

18. Rajkovic A, Davis RE, Simonsen JN, Rottman FM. A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni. Proc Natl Acad Sci U S A. 1990;87:8879–83. https://doi.org/10.1073/pnas.87.22.8879.
19. Krause M, Hirsh D. A trans-spliced leader sequence on actin mRNA in *C. elegans*. Cell. 1987;49:753–61.
20. Ross LH, Freedman JH, Rubin CS. Structure and expression of novel spliced leader RNA genes in Caenorhabditis elegans. J Biol Chem. 1995;270:22066–75 http://www.ncbi.nlm.nih.gov/pubmed/7665629.
21. Vandenberghe AE, Meedel TH, Hastings KEM. mRNA 5′-leader trans-splicing in the chordates. Genes Dev. 2001;15:294–303.
22. Pouchkina-Stantcheva NN, Tunnacliffe A. Spliced leader RNA-mediated trans-splicing in phylum rotifera. Mol Biol Evol. 2005;22:1482–9.
23. Stover NA, Steele RE. Trans-spliced leader addition to mRNAs in a cnidarian. Proc Natl Acad Sci. 2001;98:5693–8. https://doi.org/10.1073/pnas.101049998.
24. Lidie KB, Van Dolah FM. Spliced leader RNA-mediated trans-splicing in a dinoflagellate, Karenia brevis. J Eukaryot Microbiol. 2007;54:427–35.
25. Douris V, Telford MJ, Averof M. Evidence for multiple independent origins of trans-splicing in Metazoa. Mol Biol Evol. 2010;27:684–93.
26. Lei Q, Li C, Zuo Z, Huang C, Cheng H, Zhou R. Evolutionary insights into RNA trans-splicing in vertebrates. Genome Biol Evol. 2016;8:562–77.
27. Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, et al. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. PLoS Pathog. 2010;6:e1001037.
28. Cuypers B, Domagalska MA, Meysman P, De Muylder G, Vanaerschot M, Imamura H, et al. Multiplexed spliced-leader sequencing: a high-throughput, selective method for RNA-seq in Trypanosomatids. Sci Rep. 2017;7:0–11..
29. Gopal S, Awadalla S, Gaasterland T, Cross GAM. A computational investigation of kinetoplastid trans-splicing. Genome Biol. 2005;6:R95.
30. Kelly S, Wickstead B, Maini PK, Gull K. Ab initio identification of novel regulatory elements in the genome of Trypanosoma brucei by Bayesian inference on sequence segmentation. PLoS One. 2011;6:e25666.
31. Barnes SN, Masonbrink RE, Maier TR, Seetharam A, Sindhu AS, Severin AJ, et al. *Heterodera glycines* utilizes promiscuous spliced leaders and demonstrates a unique preference for a species-specific spliced leader over C . elegans SL1. Sci Rep. 2019;67:1356. https://doi.org/10.1038/s41598-018-37857-0.
32. Fiebig M, Gluenz E, Carrington M, Kelly S. Molecular & biochemical parasitology SLaP mapper: a webserver for identifying and quantifying spliced-leader addition and polyadenylation site usage in kinetoplastid genomes. Mol Biochem Parasitol. 2014;196:71–4. https://doi.org/10.1016/j.molbiopara.2014.07.012.
33. Radío S, Fort RS, Garat B, Sotelo-silveira J. UTRme: a scoring-based tool to annotate untranslated regions in trypanosomatid genomes. Front Genet. 2018;9:671.
34. Yague-sanz C, Hermand D. SL-quant: a fast and flexible pipeline to quantify spliced leader trans-splicing events from RNA-seq data. Gigascience. 2018;7:1–7..
35. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13. https://doi.org/10.1186/s13059-016-0881-8.
36. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, et al. The dynamic genome of Hydra. Nature. 2010;464:592–6.
37. Pettitt J, Mu B, Stansfield I, Connolly B. Spliced leader trans-splicing in the nematode Trichinella spiralis uses highly polymorphic, noncanonical spliced leaders. RNA. 2008;14:760–70.
38. van Bers NEM. Characterization of genes coding for small hypervariable peptides in Globodera rostochiensis: Wageningen University; 2008. http://edepot.wur.nl/16343.
39. Guo Y, Bird DM, Nielsen DM. Improved structural annotation of protein-coding genes in the Meloidogyne hapla genome using RNA-Seq. Worm. 2014;16:e29158.
40. Roy SW. Genomic and Transcriptomic analysis reveals spliced leader trans-splicing in Cryptomonads. Genome Biol Evol. 2017;9:468–73.
41. Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, et al. The genomes of four tapeworm species reveal adaptations to parasitism. Nature. 2013;496:57–63.
42. Wasik K, Gurtowski J, Zhou X, Ramos OM, Delás MJ, Battistoni G, et al. Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano*. Proc Natl Acad Sci. 2015;112:12462–7. https://doi.org/10.1073/pnas.1516718112.
43. Yang F, Xu D, Zhuang Y, Yi X, Huang Y, Chen H, et al. Spliced leader RNA trans-splicing discovered in copepods. Sci Rep. 2015;5:17411. https://doi.org/10.1038/srep17411.
44. Zhang H, Dungan CF, Lin S. Introns, alternative splicing, spliced leader trans-splicing and differential expression of pcna and cyclin in Perkinsus marinus. Protist. 2011;162:154–67.
45. Grabher MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2013;29:644–52.
46. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512. https://doi.org/10.1038/nprot.2013.084.
47. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.
48. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
49. Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics. 2013;14:328.
50. Wang S, Gribskov M. Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics. 2017;33:327–33.
51. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
52. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.

53. Capella-gutiérrez S, Silla-martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
54. Crooks GE, Hon G, Chandonia J, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188–90.
55. Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, et al. A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. PLoS Negl Trop Dis. 2012;6:e1455.
56. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 6 June 2018.
57. Bolger AM, Lohse M, Usadel B, Planck M, Plant M, Mühlenberg A. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
58. Johns Hopkins University, Center for Computational Biology. GFF utilities. http://ccb.jhu.edu/software/stringtie/gff.shtml. Accessed 6 June 2018.
59. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. 1990;41:95–8.
60. Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, et al. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst Biol. 2015;64:778–91.
61. WBPS14. *Schistosoma mansoni*. 2019. https://parasite.wormbase.org/Schistosoma_mansoni_prjea36577/Info/Index/. Accessed 11 Dec 2019.
62. Hampson S, Kibler D, Baldi P. Distribution patterns of over-represented k-mers in non-coding yeast DNA. Bioinformatics. 2002;18:513–28.

Publisher's Note

Capítulo 3: SL *trans-splicing* en *Hymenolepis microstoma*.

Resumen

Hymenolepis microstoma fue escogida para la primera aplicación de SLFinder por ser una interesante especie modelo para el estudio de Cestodos. Gracias a su fácil mantenimiento y segura de manejar, seguridad de manejo y por la disponibilidad de datos genómicos y transcriptómicos con los que trabajar. Más aún, se dispone de análisis independientes del SL *trans-splicing* con los que validar y comparar los resultados obtenidos, y su cercanía filogenética a *Echinococcus granulosus* y *E. multiloculares* permite comparar sus SL-ARN para confirmar su identidad y extensión, a pesar de la variabilidad intrínseca de estos loci. Por estos motivos, *H. microstoma* demostró ser el mejor caso posible para el diseño y análisis de SL trans-splicing en el grupo platelmintos

En este estudio se identificaron 4 SL-ARN en la especie, uno de ellos desconocido antes de este trabajo, cientos de genes sometidos a SL trans-splicing y sus sitios aceptores (SL-ACE), y se validaron experimentalmente varios loci policistrónicos. Los principales resultados son la confirmación del rol constitutivo del SL *trans-splicing* en la especie, su rol en la resolución de loci policistrónicos y su conservación filogenética con *Echinococcus multilocularis*. Así como la identificación de múltiples artefactos en la anotación del genoma de la especie, causados por la mala identificación de loci policistrónicos con splicing complejo, y como la identificación de SL-ACE puede usarse para remediarlos.

En el marco de esta tesis, este estudio representa el modelo ideal a seguir para el análisis de SL *trans-splicing* en cualquier especie de interés. De la identificación de los loci SL-ARN al análisis de los SL-ACE y caracterización de sus genes. Como tal forma las bases metodológicas aplicada en el análisis de platelmintos Cestodos y Trematodos. Mis contribuciones consisten en el procesamiento y análisis de los datos transcriptómicos. Incluyendo la estrategia empleada para identificar SL-ACEs, y la identificación y procesamiento de artefactos en la anotación, así como la confirmación experimental por RT PCR de trans-splicing y transcritos policistrónicos.

International Journal for Parasitology 53 (2023) 103–117

Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

Trans-splicing in the cestode *Hymenolepis microstoma* is constitutive across the life cycle and depends on gene structure and composition

Javier Calvelo [a,b], Klaus Brehm [c], Andrés Iriarte [a,*], Uriel Koziol [b,*]

[a] Laboratorio Biología Computacional, Departamento de Desarrollo Biotecnológico, Instituto de Higiene, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay
[b] Sección Biología Celular, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
[c] University of Würzburg, Institute of Hygiene and Microbiology, Würzburg, Germany

ARTICLE INFO

Article history:
Received 9 August 2022
Received in revised form 31 October 2022
Accepted 10 November 2022
Available online 5 January 2023

Keywords:
Spliced leader
Post-transcriptional regulation
Tapeworm
Flatworm
Platyhelminthes

ABSTRACT

Spliced leader (SL) *trans*-splicing is a key process during mRNA maturation of many eukaryotes, in which a short sequence (SL) is transferred from a precursor SL-RNA into the 5′ region of an immature mRNA. This mechanism is present in flatworms, in which it is known to participate in the resolution of polycistronic transcripts. However, most *trans*-spliced transcripts are not part of operons, and it is not clear if this process may participate in additional regulatory mechanisms in this group. In this work, we present a comprehensive analysis of SL *trans*-splicing in the model cestode *Hymenolepis microstoma*. We identified four different SL-RNAs which are indiscriminately *trans*-spliced to 622 gene models. SL *trans*-splicing is enriched in constitutively expressed genes and does not appear to be regulated throughout the life cycle. Operons represented at least 20% of all detected *trans*-spliced gene models, showed conservation to those of the cestode *Echinococcus multilocularis*, and included complex loci such as an alternative operon (processed as either a single gene through *cis*-splicing or as two genes of a polycistron). Most insertion sites were identified in the 5′ untranslated region (UTR) of monocistronic genes. These genes frequently contained introns in the 5′ UTR, in which *trans*-splicing used the same acceptor sites as *cis*-splicing. These results suggest that, unlike other eukaryotes, *trans*-splicing is associated with internal intronic promoters in the 5′ UTR, resulting in transcripts with strong splicing acceptor sites without competing cis-donor sites, pointing towards a simple mechanism driving the evolution of novel SL insertion sites.

1. Introduction

Trans-splicing is a mechanism by which two different RNAs are spliced together during mRNA maturation (Lasda and Blumenthal, 2011; Lei et al., 2016). Several eukaryotes display a specialized version of the mechanism known as spliced leader (SL) *trans*-splicing where a specific short sequence (SL) is transferred from a precursor SL-RNA into an immature mRNA, commonly on the 5′ untranslated region (UTR), resulting in the removal of all sequence upstream of the splicing site (Hastings, 2005; Lasda and Blumenthal, 2011). Since its discovery (Sather and Agabian, 1985), the evolutionary origin of *trans*-splicing, its functions, and overall relevance in eukaryotic evolution have been a hard-to-crack mystery. While there is a clear connection to *cis*-splicing in regards to the molecular machinery and regulatory signals involved, basic questions such as if SL *trans*-splicing has a single evolutionary origin remain a matter of debate (Hastings, 2005; Lasda and Blumenthal, 2011; Bitar et al., 2013; Krchňáková et al., 2017). This uncertainty is caused mainly by the lack of sequence conservation of the SL-RNA (Hastings, 2005; Bitar et al., 2013) and its scattered distribution across the eukaryotic phylogenetic tree (Sather and Agabian, 1985; Krause and Hirsh, 1987; Rajkovic et al., 1990; Brehm et al., 2000; Vandenberghe et al., 2001; Pouchkina-Stantcheva and Tunnacliffe, 2005; Lidie and Van Dolah, 2007; Douris et al., 2010; Matsuo et al., 2018).

Abbreviations: GO, gene ontology; mTP, mitochondrial transit peptide; pSL, putative SL-RNA; RT-PCR, reverse transcription-PCR; SJ, splice junction; SL, spliced leader; SM-site, site for interacting with the spliceosome Sm proteins; SP, signal peptide; TPM, transcripts per million; UTR, untranslated region; WRS test, Wilcoxon Rank Signed test.

* Corresponding authors at: Sección Biología Celular, Facultad de Ciencias, Iguá 4225, CP11400 Montevideo, Uruguay (U. Koziol); Laboratorio Biología Computacional, Departamente de Desarrollo Biotecnológico, Instituto de Higiene, Dr. Alfredo Navarro 3051, CP11600 Montevideo, Uruguay (A. Iriarte).

E-mail addresses: airiarte@higiene.edu.uy (A. Iriarte), ukoziol@fcien.edu.uy (U. Koziol).

https://doi.org/10.1016/j.ijpara.2022.11.006
0020-7519/

SL-RNAs are divided into two regions (Hastings, 2005): a leader region that is integrated into the final mRNA, and an intron-like sequence with a site for interacting with the spliceosome Sm proteins (SM-site); both are separated by a canonical donor site (GT). Their secondary structure shows similar features to other small nuclear RNAs that form the spliceosome (i.e., U1, U2, U4, and U5), suggesting a common origin (Hastings, 2005; Lasda and Blumenthal, 2011). Furthermore, SL *trans*-splicing makes use of the same molecular machinery involved in *cis*-splicing, minus U1 (Hannon et al., 1991). SLs are mostly incorporated into the pre-mRNA molecules on splice acceptor sites not associated with a strong upstream donor site (Hastings, 2005; Lasda and Blumenthal, 2011), with the SL-RNA being expended in the process. Since both mechanisms share the same components, both compete in the cell during mRNA transcription for almost the same precursors (Hastings, 2005). However, since SL *trans*-splicing is less efficient (Blumenthal, 2005), SLs are inserted in places where *cis*-splicing is impaired by a non-optimal donor site (Hastings, 2005) and/or is coupled with the cleavage of the previous cistron in an operon (Blumenthal, 2004).

The main confirmed function of SL *trans*-splicing is operon resolution in eukaryotes by dividing polycistronic pre-mRNA into individual cistrons, and providing a 5′ cap to the downstream cistrons of the operon (Hastings, 2005; Lasda and Blumenthal, 2011; Pettitt et al., 2014). Several nematodes take this further by having different SLs specialized for the first gene of operons and monocistronic genes (named SL-1) and for the resolution of downstream genes of operons (SL-2) (Harrison et al., 2010; Pettitt et al., 2010; Wenze et al., 2019). Although the great majority of SL *trans*-splicing in many species occurs in monocistronic genes, its physiological importance is less clear (Lasda and Blumenthal, 2011; Boroni et al., 2018) although one commonly suggested function is 5′ UTR 'sanitation' (Hastings, 2005; Lasda and Blumenthal, 2011; Bitar et al., 2013). Since the insertion of the SL removes the pre-mRNA 5′ region before the insertion site, SL *trans*-splicing can remove deleterious sequences located in these regions (Lasda and Blumenthal, 2011). Thus, SL insertions relax the selective constraints of the 5′ UTR region. In this sense, understanding SL *trans*-splicing is crucial to comprehend the dynamics shaping UTR evolution in animals. In addition, this splicing mechanism has been proposed as a potential source of alternative isoforms by generating novel translation start sites (Agorio et al., 2003; Hastings, 2005; Nilsson et al., 2010; Boroni et al., 2018). Although there is conflicting evidence toward its overall importance in the diversification of transcriptomes (Soulette et al., 2019), this might be particularly important in Platyhelminthes since their SLs bear a 3′ terminal "AUG" motif which can function as an initiation codon (Cheng et al., 2006).

Platyhelminthes are an interesting group to study this mechanism. Similar to *Caenorhabditis elegans* (Pettitt et al., 2014), they possess SL *trans*-splicing but without a clear SL-RNA specialization in regard to their target genes (Protasio et al., 2012; Tsai et al., 2013, but see Rossi et al., 2014). In particular, cestodes possess several SL variants that seem to be used interchangeably (Tsai et al., 2013; Olson et al., 2020). Furthermore, flatworm parasites (classes Monogenea, Cestoda and Trematoda (Hickman et al., 2008)) represent a significant burden on economic activities and human health (Mahmud et al., 2017; Webb and Cabada, 2017). Given the need for new therapeutic options (Cabada et al., 2016; Trainor-moss and Mutapi, 2016), SL *trans*-splicing is a promising target for novel drugs to treat eukaryotic parasites bearing this mechanism (Pandarakalam et al., 2019) since it is absent in vertebrates (Lei et al., 2016). Understanding the biological importance of the mechanism, its components, and its target genes will support these efforts.

In this study, we conducted an in-depth analysis of the usage of *trans*-splicing in the cestode *Hymenolepis microstoma*. This cestode is a classic cestode model species due to the easy maintenance of its life cycle in the laboratory (Macnish et al., 2003; Cunningham and Olson, 2010). Recently, Olson et al. (2020) described the presence of SL *trans*-splicing in this species from three SL-RNA precursors. In this work, we analyzed in depth this process in this species with the objective of identifying SL-RNAs, their targets, and their potential roles in gene regulation throughout its life cycle.

2. Materials and methods

2.1. Transcriptomic and genomic data

RNAseq data covering the three life stages of the species were previously published by Preza et al. (2021). These included four biological replicates for: eggs containing infective oncospheres, cysticercoids, and adults (pools of animals in all cases), with between 12,937,021 and 14,680,727 raw read pairs per replicate. In the case of adults, the posterior-most segments had been removed before processing in order to avoid contamination with the RNA of developing oncospheres. The reference genome and associated gene models of *H. microstoma* were retrieved from the WormBase Parasite database v14 (Howe et al., 2017) (ID: HMN_v3, BioProject: PRJEB124) that was more recently described in Olson et al. (2020). Genes were annotated according their orthological relationships with *Echinococcus multilocularis* (also retrieved from the WormBase Parasite ID: EMULTI002, BioProject: PRJEB122) using SynChro (Drillon et al., 2014). Estimated syntenic blocks are summarized in Supplementary Table S1. SL acceptor genes and potential operons were compared with Tsai et al. (2013), accounting for ID changes when necessary.

2.2. SL identification and characterization

SLFinder was utilized to identify SL sequences with default settings (Calvelo et al., 2020). Read quality was first evaluated with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 6.6.18)), and low quality bases together with adapter sequences were removed with Trimmomatic v0.36 (Bolger et al., 2014) with the parameters (SLIDINGWINDOW:5:20, MINLEN:25). Transcriptomes were then de novo assembled with Trinity v2.8.3 (Haas et al., 2013) with read normalization disabled. Once novel putative SLs were identified, we conducted a manual inspection of the loci to determine the limits of the putative SL-RNA (pSL). Specifically, the 5′ end of the assembled contigs bearing each pSL were compared with their respective genomic loci in MEGA7 (Kumar et al., 2016) to recover more of the SL 5′ region; the extension of the intron-like region was determined based on the reported SL genes for *H. microstoma* (Olson et al., 2020) and *E. multilocularis* (Tsai et al., 2013). Specifically, we searched for a 3′ terminal "GRGGGCC" like motif that seems to be conserved in the class Cestoda (see Section 3.2).

Conservation of *H. microstoma* SL-RNAs was evaluated relative to known platyhelminth SL genes: the trematode *Schistosoma mansoni* retrieved from the GeneDB database (Logan-Klumpler et al., 2012), the cestodes *Echinococcus granulosus*, *Echinococcus multilocularis* from Tsai et al. (2013) and *Taenia solium* from the NCBI. Sequences were aligned using Muscle (Edgar, 2004) implemented in MEGA7, and the NCBI genes were trimmed accordingly. Potential SM binding sites (5′-ARU4-6GR-3′) were identified with Seqkit v0.12.1 (Shen et al., 2016), allowing for a single mismatch. Conserved sections of the SL-RNA were manually identified from the aligned sequences. From this alignment, a basic Neighbor-Joining

phylogeny was estimated. Pairwise distance matrix and average distances within groups (*Hymenolepis*, taeniids, and *S. mansoni*) were also calculated. The secondary structure was predicted by the Minimum Free Energy method (Zuker and Stiegler, 1981) using RNAfold from the ViennaRNA webserver (Lorenz et al., 2011) with default values but constraining the SM site to be single-stranded, and then visualized in Forna (Kerpedjiev et al., 2015).

2.3. Identification and classification of SL acceptor genes

Acceptor genes were identified by mapping reads bearing the 3′ terminal bases of each of the leader regions to the genome (hits of 15 bases without mismatches), as implemented on the script SLFinder-Genes of the SLFinder pipeline (available at https://github.com/LBC-Iriarte/SLFinder). Reads with the SLs were identified with Seqkit and the leader sequences were trimmed with Cutadapt v2.9 (Marcel, 2011). Then, reads were mapped to the genome with Bowtie2 v2.3.5.1 (Langmead and Salzberg, 2012) with the options "--no-mixed", "--no-discordant" and "--end-to-end". SL reads associated with canonical acceptor sites ("AG") and mapped to known genes were then selected from SLFinder-Genes output and counted. A gene was considered subject to *trans*-splicing if there was an acceptor site supported by at least three reads distributed on three samples or more. Acceptor sites were then classified according to their location respective to the model of the closest gene: "5' UTR" (before the coding sequence), "predicted internal acceptor site" (insertion on a reported *cis*-splicing site), "novel acceptor site in a reported exon" (insertion inside reported exon) or "intron" (insertion within a reported intron sequence). When multiple isoforms were reported for the gene and the nature of the acceptor site was not clear the site was classified as "ambiguous". For comparison purposes, the same procedure was used for Olson et al. (2018) RNAseq data and SL reads were counted when meeting our criteria.

2.4. SL usage analysis

Variation in the usage of each detected SL type between life stages was statistically tested by ANOVA. Evidence of SL specialization was then analyzed at a single site level by conducting Fisher tests (Fisher, 1934) between the mapped SL reads (for sites supported by more than 10 reads between the four biological replicates considered) and their expected counts given the SL frequencies for each life stage. Sites selected for adult, cysticercoid, and egg samples were further analyzed in the same manner but relative to their expected values based on the average transcripts per million (TPM) in each life stage reported for each gene by Preza et al. (2021).

2.5. Read coverage of gene model features and splice junctions

SL sequences were removed from the RNAseq reads and were then mapped to the genome with STAR v2.7.1a (Dobin et al., 2013), and coverage measurements conducted with Samtools depth v1.10 (Li et al., 2009). To identify splice junctions (SJs) over zones of interest, a second pass alignment was conducted over the initial results, following STAR's User Manual recommendations, and visualized with sashimi plots (Katz et al., 2015) generated with ggsashimi v1.0.0 (Garrido-Martín et al., 2018). This data was used as the basis of several analyses detailed below.

2.6. Analysis of SL trans-splicing in the 5′ UTR of monocistronic genes

The 5′ UTR structure of SL *trans*-spliced genes was analyzed by manually evaluating the 50 SL acceptor sites with highest read support that were not predicted to be downstream genes within operons (see Section 2.9). The 5′ regions were visualized on IGV (Thorvaldsdóttir et al., 2013) to delimitate them and then generate sashimi plots for evaluation. Only SJs supported by at least five reads were considered. 5′ UTR regions were then classified into "Putative Operons" if manual annotation suggested that the gene was part of a polycistron with an upstream gene model in the same strand (regardless of whether it met our strict criteria detailed in Section 2.9), "Intronic 5' UTR" if there was evidence of a *cis*-splicing SJ near or involving the SL acceptor sites, or "Simple 5' UTR" if neither. The 5' UTRs of the top 500 most highly expressed genes (by TPM values) were classified in a similar manner. Another read alignment with STAR was used to identify strongly supported junctions, from reads with an overhang of at least 10 bases across the splice junction (second pass with options --alignSJoverhangMin and --alignSJDBoverhangMin set to 10) and indicated in a case-by-case manner.

2.7. Identification of chimeric gene models

The sequences of genes *trans*-spliced internally, according to their annotated CDS, were divided at the SL insertion. Homologous protein sequences in *E. multilocularis* were identified by tblastn searches, using each part as a query (Altschul et al., 1990). Genes from *H. microstoma,* for which their divided sequence matched different genes in *E. multilocularis* for at least one of their internal SL insertion sites, were classified as possible chimeric gene models and inspected on sashimi plots for reads connecting both putative genes across the predicted SJ.

2.8. Gene expression and GO term enrichment

Gene expression data and genes differentially expressed between life stages were retrieved from Preza et al. (2021). Gene Ontology (GO) annotation of the longest annotated transcript of each gene of *H. microstoma* was predicted by EggNOG mapper tool v2 with default parameters (Cantalapiedra et al., 2021) using the reported cDNA sequenced as a query (accessed in September 2020). Detected chimeric gene models (Sections 2.7 and 3.5) were divided before and after the SL insertion. Cases with multiple SL acceptor sites were divided at the insertion sequence that maximized the sequence length of both halves. Enrichment analysis was then carried out with Fisher's exact test against the expressed genes reported in Preza et al. (2021), using the package topGO v3.11 (Alexa and Rahnenfuhrer, 2019, topGO: Enrichment Analysis for Gene Ontology. R package version 2.36.0). GO terms with a *P*-value < 0.005 were considered significant.

2.9. Identification of potential operons

Gene clusters were selected as potential operon candidates if they: i) were separated by less than 300 bp, ii) were encoded on the same strand, iii) were not located inside the intron of a larger gene, and iv) the read coverage of the intergenic region didn't fall below six reads at any position.

2.10. Experimental validation by reverse transcription-PCR (RT-PCR) of SL trans-splicing and polycistronic transcription

Total cellular RNA from adult samples was isolated and purified with TRI Reagent (Sigma Aldrich, Germany) in combination with the Direct-zol RNA Miniprep (Zymo Research, United States), then reverse transcribed to cDNA with SuperScript II Reverse Transcriptase (Thermo, United States). Primers for the PCRs were designed with Primer3 (Untergasser et al., 2012) and selected based on their compatibility with the SL3-specific primer (the SL sequence most commonly found in adults in our samples, see Supplementary

Table S2). PCR was performed with HighTaq DNA Polymerase (Bioron, Germany). All reactions were conducted following the program: initial strand separation at 94 °C (2 min); 35 cycles of strand separation at 94 °C (10 s), annealing at 54 °C (20 s), and extension at 72 °C (1 min); and a final extension at 72 °C (1 min). PCR products were analyzed by electrophoresis in 2% agarose gels, and selected amplification products were purified with a Monarch® DNA Gel Extraction Kit. In some cases, the purified amplicons were sequenced directly by the Sanger method at Macrogen.Inc, South Korea. In other cases, the amplicons were cloned using the TA Cloning™ Kit, Dual Promoter, with pCR™II Vector (Thermo, United States) before sequencing. Negative control experiments without reverse transcription were performed for selected operon candidates.

2.11. Identification of removal or exposure of localization signals associated with SL insertions

The protein sequences of genes *trans*-spliced internally were analyzed for the presence of signal peptides (SPs) and mitochondrial transit peptides (mTPs), using SignalP v5.0 (Armenteros et al., 2019b) and TargetP v2.0 (Armenteros et al., 2019a), respectively, before and after the SL insertion, accounting for potential changes in the reading frame of the mRNA and the addition of the ATG triplet in the SL. This was done by calculating novel open reading frames with getorf from the EMBOSS package (Rice et al., 2000). Open reading frames were selected for analysis if they met the following criteria: i) were at least 10 codons in length, ii) there was a possible ATG start codon not further than 30 bases from the SL insertion, and iii) no stop codons were detected until the end of the reported exon. Cases where the SP or mTP showed dependence on the SL insertion (removal, exposure, or change in the reliability prediction) were manually inspected on IGV for classification.

2.12. Properties of internal introns upstream of internal SL trans-splicing sites

Sequence length and GC content of introns associated with internal SL insertions not flagged as annotation artifacts (Section 2.7) were analyzed relative to those exclusively subject to *cis*-splicing in the genome. Intron sequences were first retrieved based on the genome annotation and filtered by their coordinates to remove redundancy across isoforms and/or overlapping genes. Since neither of both measurements showed a normal distribution (data not shown), the non-parametric Wilcoxon Rank Signed (WRS) test was applied to identify significant differences (Wilcoxon, 1945).

3. Results

3.1. Identification of SL sequences and trans-splicing acceptor sites

We applied our recently published pipeline, SLFinder (Calvelo et al., 2020), to search for novel SL sequences in the model cestode *H. microstoma*, using the transcriptomic data that we previously generated (four biological samples of adults, cysticercoids, and infective eggs; Preza et al., 2021). We assembled de novo transcriptomes (reported in Supplementary Data S1), from which SLFinder identified four SL sequences shown in Fig. 1A; SL-1 is encoded in two loci on the chromosome HMN_01_pilon (positions 9215203 – 9215292 and 9215948 – 9216037), SL-2 on HMN_03_pilon (positions 20837612 – 20837715), SL-3 on HMN_01_pilon (positions 5649186 – 5649090), and a new SL-RNA referred in this work as SL-4 on HMN_01_pilon (positions 7329289 – 7329201). SL-1, -2 and -3 match the *H. microstoma* SL sequences recently reported by Olson et al. (2020). To evaluate SL usage and insertion sites we defined the last 15 bases of the leader region as unique tags for each SL (Fig. 1A). A total of 38,078 reads were found bearing these tags and successfully mapped to reported gene models at a canonical splicing acceptor site. They corresponded to 1650 acceptor sites distributed among 1305 genes, of which 714 sites (distributed among 622 genes) were supported by reads from more than three biological samples. SL insertions without a canonical acceptor site or that did not match the assigned gene orientation were excluded. We classified these highly supported acceptor sites according to their position within their respective genes. As expected, most of the SL acceptor sites were found in the 5′ UTR (393 sites), but we also identified internal acceptor sites. Most of these internal sites were previously predicted *cis*-splicing acceptor sites (236), novel acceptor sites in exons (41), and novel acceptor sites in introns (31). A small portion were classified as ambiguous (13 sites) as they are associated with different annotated isoforms for the gene (Supplementary Table S3). However, as is discussed below (Section 3.5), a large fraction of the genes bearing internal sites corresponded to incorrectly predicted gene models in which two different genes were fused together. In addition, 62 genes presented multiple acceptor sites near each other (less than 100 bp from each other), but in all cases one of them was predominant over the others by taking on average 84% of all SL bearing reads mapped to these regions.

Comparing our results with previous studies, we found that 84% of the *trans*-spliced genes identified with high confidence by Olson et al. (2020) were also identified in this work (417 of the 496 reported genes). Furthermore, 402 *trans*-spliced genes had orthologues with *trans*-splicing in *E. multilocularis* according to Tsai et al. (2013), indicating robustness in the methodology employed in the three studies. There is, however, a key difference between our results and those of Olson et al. (2020), who described a higher number of SL reads and *trans*-spliced genes in mid-metamorphosis larvae in comparison with adults. Here, we found an approximately equal number of SL reads across life stages, with 13,068 raw reads assigned to a known gene and displaying a canonical donor site in adults, 13,900 in cysticercoids, and 11,110 in eggs, as well as similar numbers of *trans*-spliced genes across life stages (374, 393, and 327 genes supported by at least three biological replicates for each life stage, respectively). In order to ascertain the source of these discrepancies, we re-analyzed the RNAseq samples first published by Olson et al. (2018) using our pipeline and found very few SL-bearing reads in the adult samples regardless of the specific sample (Supplementary Table S4). Thus, the differences are not due to the different methods of analysis and appear to be present in the original RNA samples.

To measure relative usage between SLs among genes and life stages we first selected sites supported by at least 10 reads among the four samples of each stage (252 for adults, 283 cysticercoids, and 214 for eggs). All SLs were present in the transcriptomes for all life stages; SL-2 and SL-3 together represented between 70 and 75% of reads in all stages; however, the relative abundance of the two changed between adults and the studied larval stages (Fig. 1B**)**. Differences between life stages were statistically verified by ANOVA tests for all SL-types, and the pairwise T-student test showed significant differences between adults and both larval stages (Supplementary Table S5). In our re-analysis of the data of Olson et al. (2020), we also observed a relative increase in the use of SL-3 in adults in comparison to larvae, but SL-1 is more prevalent in all samples in comparison to our data (Supplementary Table S4**)**.

Focusing on our dataset, no individual acceptor site showed SL type frequencies that deviated from the abundance found in each life stage (as determined by Fisher Tests on sites supported by

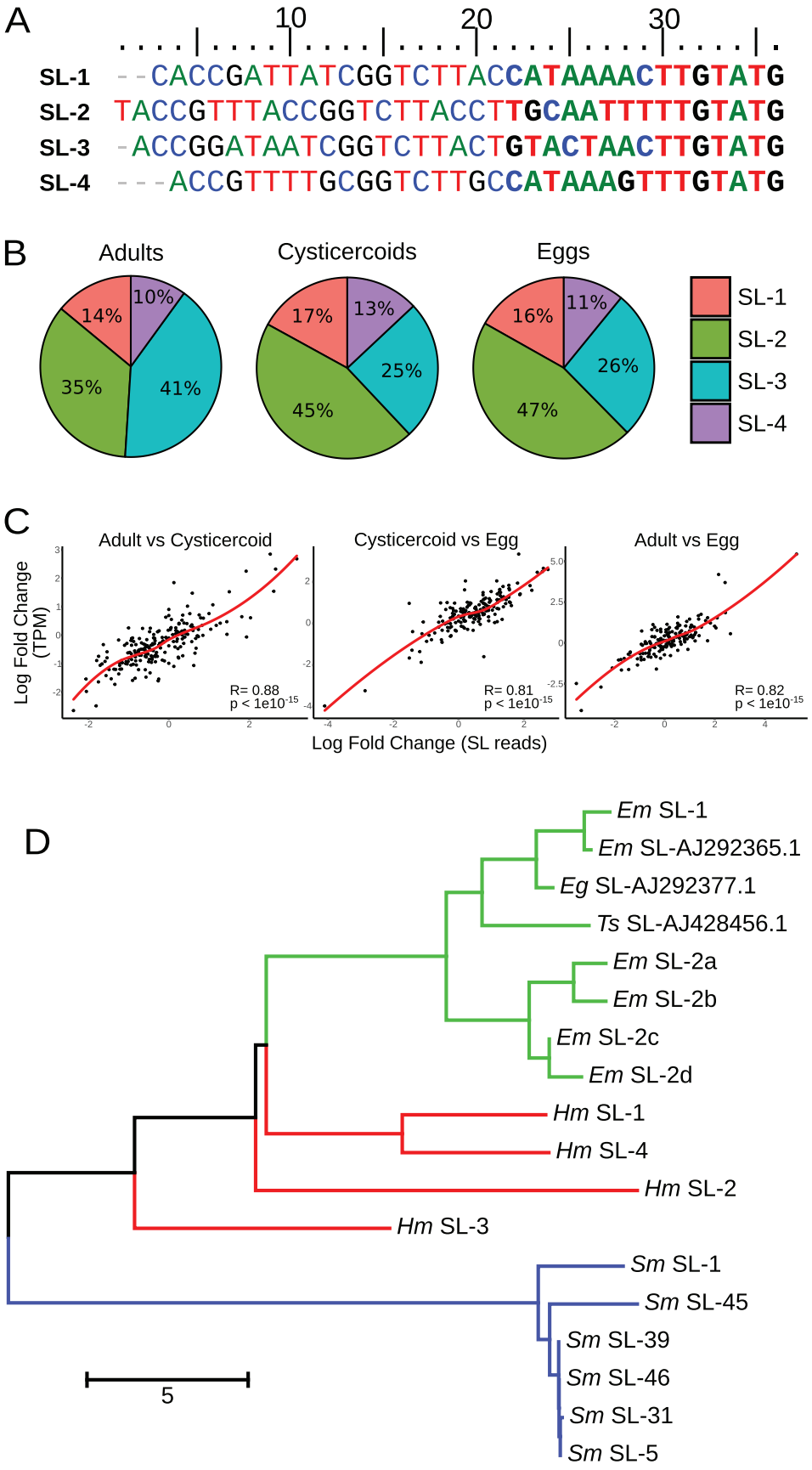

Fig. 1. *Hymenolepis microstoma* spliced leader (SL) sequences. (A) SL sequences after manual extension of the SLFinder output. The 3′ tag used for identifying and trimming of SL bearing reads is highlighted in bold. (B) Relative frequencies of each SL found in each life stage. (C) Scatterplots comparing the Log2 fold change of the mean number of SL reads and the Log2 fold change of their mean transcripts per million (TPM), for all pairwise life-stage comparisons. Only genes with at least 10 SL reads in both life stages, among all of their acceptor sites, were considered. (D) Neighbor-joining phylogenetic relationships between SL-RNA sequences. SL sequences of cestodes and trematodes are included: *Echinococcus granulosus* (*Eg*), *Echinococcus multilocularis* (*Em*), *H. microstoma* (*Hm*), *Schistosoma mansoni* (*Sm*), and *Taenia solium* (*Ts*).

more than 10 SL reads among the four samples for each life stage, Supplementary Tables S6-S8). In order to identify if any gene showed a higher prevalence of *trans*-splicing in a particular life stage, we analyzed 161 sites with at least 10 SL reads for the three life stages. We found that in all three stages, there is a strong correlation between the average TPM reported by Preza et al. (2021) and the total number of SL reads identified for the same genes as determined by Pearson's Correlation (*P*-value < 0.001, correlation factors 0.42, 0.35 and 0.54 for adults, cysticercoids and eggs, respectively). Similarly, fold-changes of gene expression and of abundance of SL reads are highly correlated in all pairwise comparisons between life stages (Fig. 1C). Individual Fisher Tests against the expected read counts also failed to identify outliers (Supplementary Table S9). Altogether, these results show that SL bearing reads follow the expected biases, indicating that different SLs are inserted on their targets indiscriminately and that *trans*-splicing is largely constitutive across the life cycle.

3.2. Sequence and structure conservation of H. microstoma SLs

To evaluate relevant conserved regions, *H. microstoma* SL-RNA sequences were aligned with known SL-RNA sequences of other cestodes (*E. granulosus*, *E. multilocularis*, and *T. solium*) and of the trematode *S. mansoni* (Supplementary Fig. S1). The donor site with the "AUG" codon reported for platyhelminths is conserved in all sequences, although the immediately surrounding area shows clear differences between both lineages, except for a weakly conserved uridine-rich tract immediately downstream of the donor site. SM sites were also easily identified (within positions 82–93 in the alignment) with a clear difference in size between the cestode species and *S. mansoni,* all with a single mismatch in the U-rich central region. Unexpectedly all *H. microstoma* SL-RNAs display different putative SM sites, with SL-3 having a minor deletion on its 5' portion: SL-1 AAUUAUUUGG; SL-2 AAUUCUUUGG; SL-3 ACUUUUUGG; and SL-4 AAUUGUUUGG. Differences in SL-RNA sequences concentrate on the space between donor sites and the SM sites (position 44 in the alignment up to the SM site). Comparisons between the classes Cestoda and Trematoda in this region are misleading due to a lack of sequence similarity, but relative to the other cestodes the more abundant SL-2 and SL-3 are the most similar in terms of sequence length and composition. SL-1 and SL-4 share the same indels in this region which translate to minor structural differences (see below). To have a better overview of sequence similarity between SL-RNAs, we estimated a basic Neighbor-Joining phylogeny from this alignment. There is a high divergence between *S. mansoni* and the cestode species (Fig. 1D). *Hymenolepis microstoma* SL-RNA sequences arrange themselves basal to taeniids as a paraphyletic group, although the actual support for this is arguable due to sequence divergence. It is remarkable, however, that the distances between *H. microstoma* SL-RNAs are considerably greater (20 differences on average) than those observed within the family Taeniidae (six differences), suggesting they share a more ancient common ancestor than the SLs from this latter group. When the potential SM sites are constrained to be single stranded, all four *H. microstoma* SL-RNAs display a similar secondary structure consisting of two hairpins before the SM site and a smaller third one after it; the donor site is located mid-way on the first hairpin six bases upstream of a conserved bulge (Fig. 2). The major distinctive feature between frequently and less frequently spliced SL-RNAs is the second hairpin: SL-1 and -4 have a simple hairpin and possess a comparatively more unstable predicted structure (minimum free energy −28.60 and −26.60 kcal/-mol, respectively), while SL-2 and −3 bear a longer hairpin with an internal loop and are more stable (−34.70 and −34.50 kcal/mol, respectively).

3.3. SL acceptor genes are largely constitutively expressed and enriched for ubiquitous biological processes

All SL acceptor genes had expression values reported in Preza et al. (2021) except for one (HmN_003004400). Interestingly *trans*-spliced genes are enriched in genes that are not differentially expressed (χ^2 statistic 37.30, *P*-value < 0.001), see details in Supplementary Table S10. Thus, SL *trans*-splicing appears to be largely dedicated to processing constitutively expressed genes. To further explore the potential functions of SL *trans*-spliced genes we identified enriched GO terms among the genes subjected to this mechanism (Table 1**)**. SL *trans*-spliced genes showed several categories of functional enrichment. Enriched biological processes are dominated by terms related to protein synthesis and modification, such as "Protein Maturation" (GO:0051604), "Ubiquitin-dependent Protein Catabolic Process" (GO:0006511), "GPI Anchor Biosynthetic Process" (GO:0006506), and "Mitochondrial Translational Termination" (GO:0070126); these are ubiquitous processes that are likely to be largely constitutive.

3.4. 5′ UTR introns are a common feature of many monocistronic SL trans-spliced genes

In two well-studied examples of SL *trans*-splicing in cestodes (genes *elp* and *tgr* from *Echinococcus* spp.), *trans*-splicing of monocistronic genes has been correlated to the presence of introns in the 5′ UTR containing internal transcription start sites. This arrangement provides a strong splice acceptor site in the absence of a corresponding cis donor site for transcripts initiated from the internal (intronic) promoter. In the *elp* gene of *E. multilocularis* (Brehm et al., 2000), two mRNA isoforms are generated with identical coding sequences, one that is not *trans*-spliced and begins from a non-coding 5′ exon, and a second, SL *trans*-spliced isoform in which the SL is inserted directly into the second exon of the gene (Fig. 3A). In the *tgr* gene of *E. granulosus*, two *trans*-spliced transcripts have been described that result in two protein isoforms with different subcellular localizations: one that begins in the first coding exon and codes for a mitochondrial protein, and one that begins in the second coding exon, removing the mitochondrial localization signal and resulting in a cytoplasmic protein (Agorio et al., 2003; Otero et al., 2010; Fig. 3B). By analyzing available RNA-Seq data of *E. granulosus*, we have identified an additional, non-coding exon in the 5′ region of the *eg-tgr* gene, which gives rise to new transcripts for the cytoplasmic and mitochondrial isoforms, as it is spliced to either of the previously known initial exons (Fig. 3B). These isoforms are presumably not *trans*-spliced as they were not amplified by RT-PCR using an SL forward primer by Agorio et al. (2003). Therefore, the incorporation of the SL in the previously described transcripts for mitochondrial and cytoplasmic isoforms of *eg-tgr* occurs in acceptor sites that are also used for *cis*-splicing.

We compared the gene structure and transcriptional isoforms of the ortholog genes in *H. microstoma*. We manually corrected the predicted gene models since they did not agree with the available RNASeq data (the incorrect automatic predictions are likely resulting from the complex organization of these genes in the 5′ region, and similar problems were detected in the Wormbase Parasite predictions of the *Echinococcus* spp. genes when compared with their experimentally validated versions, Supplementary Table S11). In the case of *hm-elp* (HmN_000608600), a similar exon/intron structure was found in the 5′ region as in *em-elp*, and both RNASeq and RT-PCR experiments (using specific reverse primers in combination with a forward primer corresponding to the SL-3 sequence) confirmed the existence of SL *trans*-splicing at a position homologous to that found in *E. multilocularis* **(**Fig. 3A**)**. In contrast, we found no evidence from RNASeq or RT-PCR of *trans*-splicing for

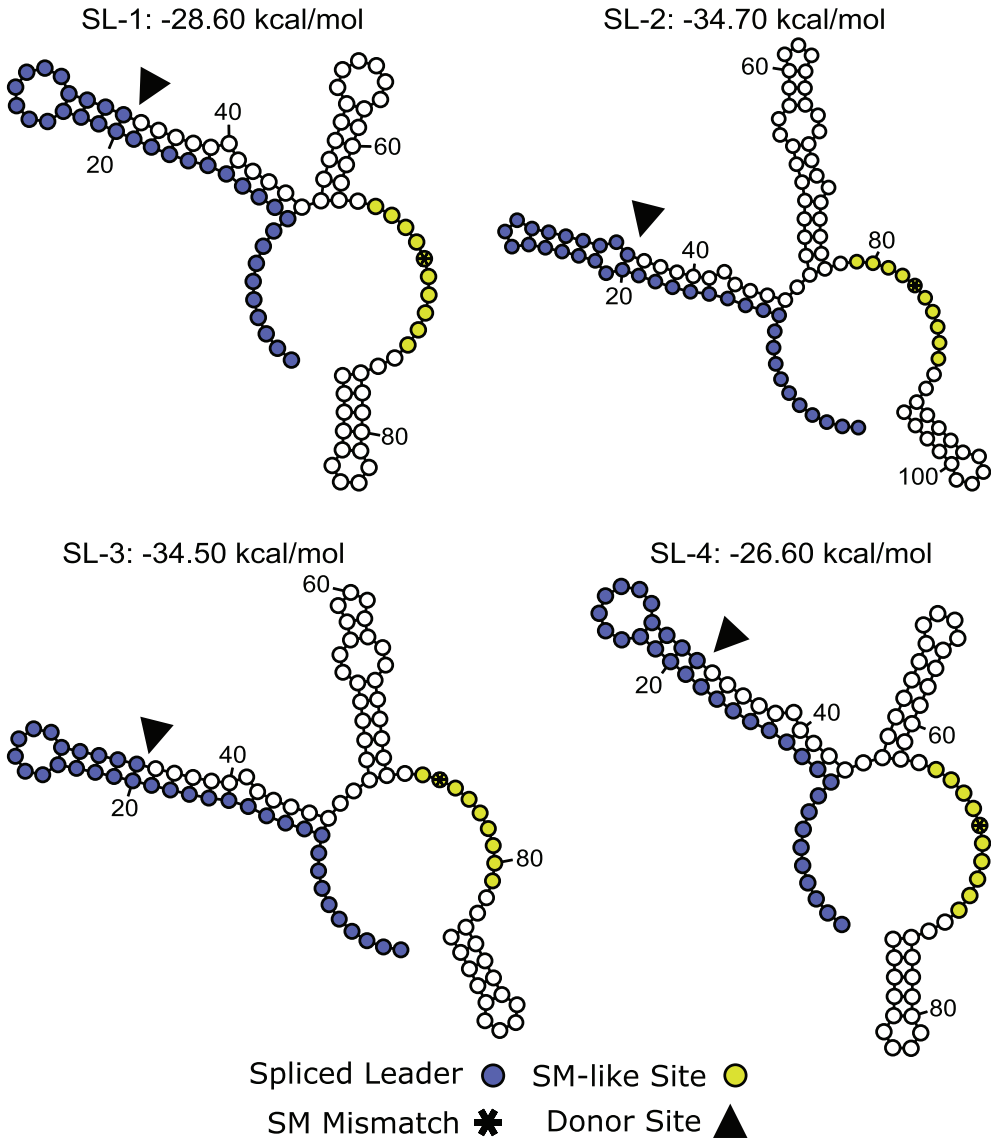

Fig. 2. Secondary structure of *Hymenolepis microstoma* spliced leader (SL)-RNA sequences generated by RNAFold. The spliced leader and the site for interacting with the spliceosome Sm proteins (SM site) are colored in blue (dark grey) and yellow (light grey), respectively; mismatches with the SM site consensus are indicated with an "*".

Table 1
Enriched Gene Ontology (GO) terms in the spliced leader (SL) *trans*-spliced genes with respect to the expressed transcriptome.

Class	GO Term	Term	Ref.	List	Expected	*P*-value
BP	GO:0006695	Cholesterol Biosynthetic Process	29	5	2.38	0.001
	GO:0070126	Mitochondrial Translational Termination	55	12	4.52	0.001
	GO:0018205	Peptidyl-lysine Modification	156	20	12.81	0.002
	GO:0006400	tRNA Modification	48	12	3.94	0.002
	GO:0006506	GPI Anchor Biosynthetic process	18	6	1.48	0.002
	GO:0014066	Regulation of Phosphatidylinositol 3-kinase Signaling	53	8	4.35	0.002
	GO:0006672	Ceramide Metabolic Process	24	5	1.97	0.002
	GO:0036099	Female Germ-line Stem Cell Population Maintenance	13	5	1.07	0.003
	GO:0097502	Mannosylation	13	5	1.07	0.003
	GO:0006511	Ubiquitin-Dependent Protein Catabolic Process	283	35	23.23	0.003
	GO:1903008	Organelle Disassembly	54	9	4.43	0.004
MF	GO:0000030	Mannosyltransferase Activity	12	5	1.01	0.001
CC	GO:0005762	Mitochondrial Large Ribosomal Subunit	38	9	2.99	0.002
	GO:0019898	Extrinsic Component of Membrane	126	12	9.71	0.003

BP, biological process; MF, molecular function; CC, cellular component.

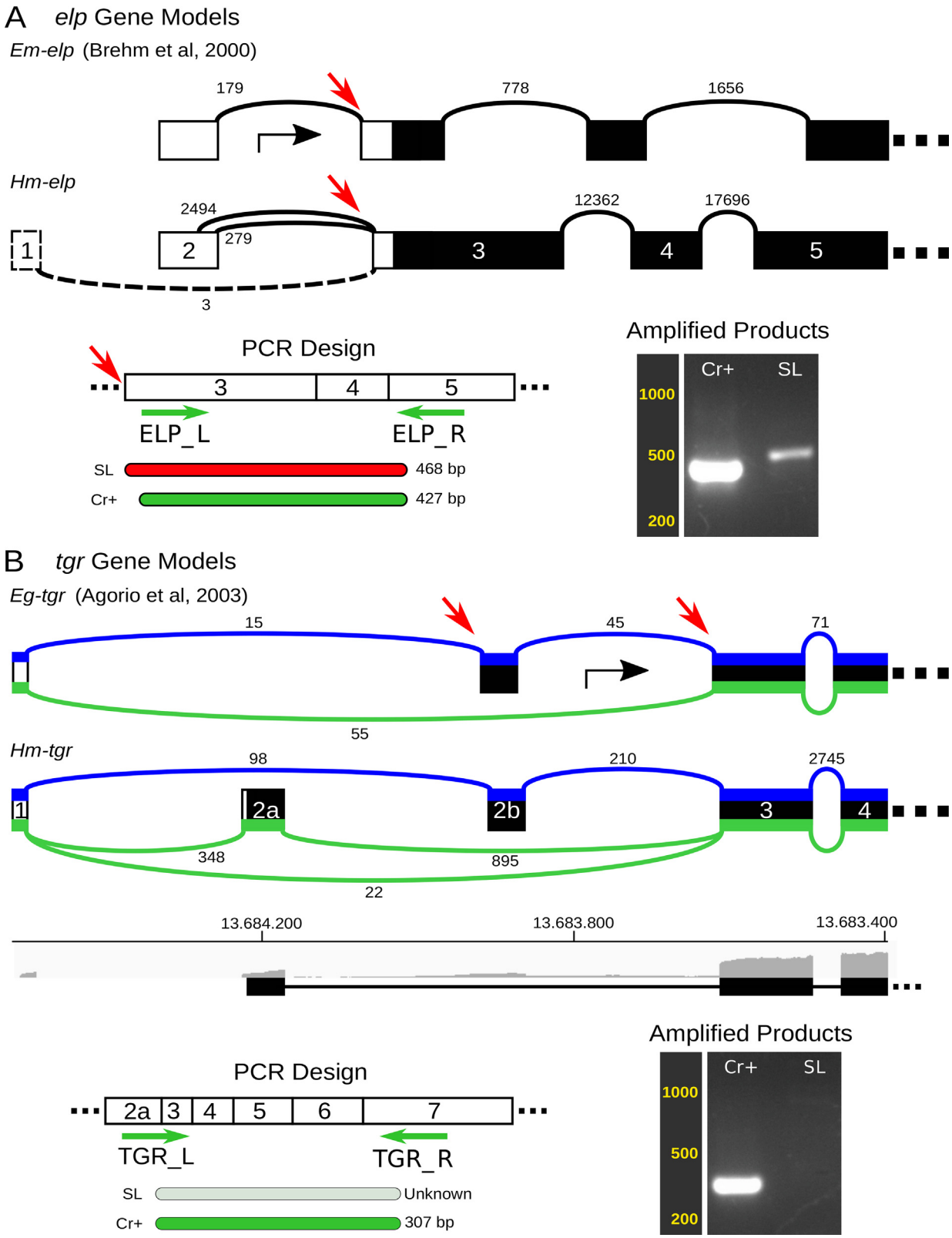

Fig. 3. Comparison of monocistronic genes known to be subjected to spliced leader (SL) *trans*-splicing in *Echinococcus* spp. and their orthologs in *Hymenolepis microstoma*. (A) *em-elp as* reported by Brehm et al. (2000) in *Echinococcus multilocularis* and (B) *eg-tgr* as reported by Agorio et al. (2003) in *Echinococcus granulosus* (including an additional initial exon described in this work). Untranslated region (UTR) and protein-coding regions are represented in white and black boxes, respectively. SL insertion sites are indicated with arrows. Splice Junctions are represented with arches with the number of supporting reads next to them. Internal promoters identified in the original references are indicated with rightward arrows. Below each comparison, the reverse transcriptase-PCR design utilized for experimental validation is represented, indicating the location of primers, the expected amplicons and their sizes in bp, together with the experimental results (amplified products, including the amplicons obtained using a forward SL primer (SL), and the amplicons obtained using a forward gene-specific primer (Cr +)). Exons for *H. microstoma* gene models are numbered accounting for mutual exclusion. The first exon of *hmi-elp* and the associated splice junction are dotted because, while being present in the WormBase Parasite annotation, they are poorly supported by read coverage. For *tgr* orthologs, isoforms encoding mitochondrial and cytoplasmic proteins are represented with arches above and below, respectively. In addition, the observed read coverage in the 5′ region of this gene is included below.

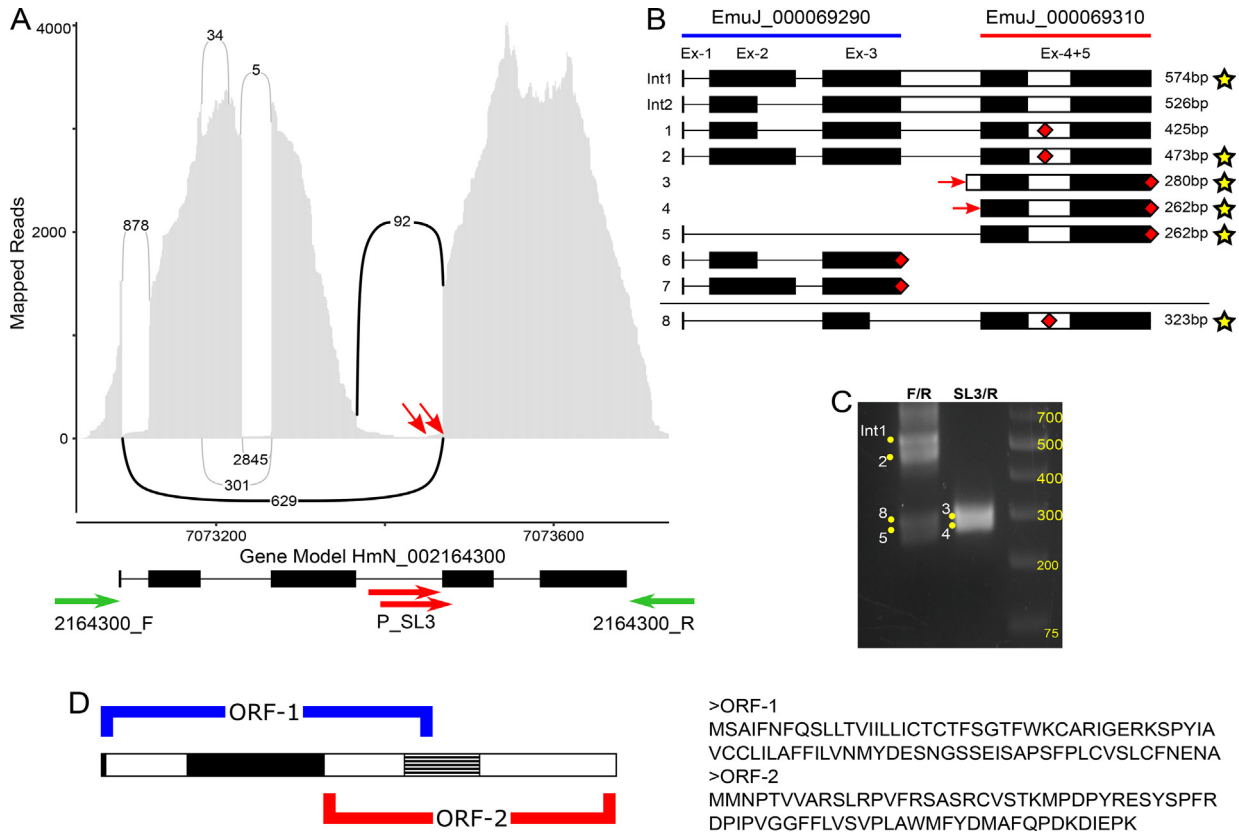

Fig. 4. Revised gene model for gene HmN_002164300 based on our RNASeq and reverse transcriptase (RT)-PCR data. (A) Sashimi plot showing the read coverage and splice junctions in the region. Spliced leader (SL) insertion sites are indicated with arrows. The current annotated gene model for the gene in WormBase Parasite is displayed below, together with the primer locations for RT-PCR validation. (B) Possible intermediaries and isoforms predicted for the locus and their regions of similarity to homologous gene models in *Echinococcus multilocularis*. Boxes represent coding regions (black for predicted exons in the annotation and white for retained/incorrectly annotated intron) and diamonds represent stop codons. The expected product lengths for the isoforms targeted by the PCR amplification are included, and successfully amplified products are highlighted with stars. Isoform 8, detected by cDNA cloning, is also included. (C) Obtained RT-PCR products for amplifications with 2164300_F and 2164300_R primers (lane F/R) and with P_SL3 and 2164300_R primers (lane SL3/R). (D) Schematic representation of the open reading frames (ORFs) encoded in the locus and the exons they encompass (alternating black and white boxes), lengths based on isoform 1; the incorrectly annotated intron between exons 4 and 5 in the annotation is represented with stripes. ORF 1 is coded by isoforms 1, 2, 6, and 7 with relatively minor changes due to alternative splicing over exon 2. ORF-2 is coded by isoforms 3, 4, and 5 with some variations in its 5′ UTR.

the *H. microstoma hm-tgr* gene (HmN_000673800) (Fig. 3B). Furthermore, the *H. microstoma* gene has an additional exon that codes for the N-terminal region of another cytoplasmic isoform. This exon is mutually exclusive with the exon coding for the mitochondrial localization signal (exons 2a and 2b in Fig. 3B). Thus, the genes from both species generate transcripts coding for mitochondrial and cytoplasmic isoforms, but some cytoplasmic isoforms are achieved differently in each species (bypassing the mitochondrial signal sequence by initiating transcription from a downstream intronic promoter, coupled to *trans*-splicing, in *E. granulosus*, versus mutually exclusive alternative *cis*-splicing of coding exons in *H. microstoma*). The lack of *trans*-splicing in *hm-tgr* could indicate the absence of internal promoters in the 5′ introns in this species.

These selected examples show that although the presence of 5′ UTR introns can be linked to SL *trans*-splicing, it is not always necessarily the case. We therefore sought to analyze this phenomenon on a larger scale in *H. microstoma* to determine how common it is. First, we manually explored the 5′ UTR of SL *trans*-spliced genes that were not predicted to be part of operons (see Section 3.6), focusing on the top 50 genes by total SL bearing read counts. Of these, 23 genes had evidence of 5′ UTR introns with *cis*-splicing junctions involving the SL acceptor sites (Supplementary Table S12; Supplementary Data S2), and 20 other genes appeared to be polycistronic loci that were missed on our operon search (see Section 3.6). Only two genes were monocistronic and did not have any evidence of introns in their 5′ UTR. Another interesting observation is that several genes possess multiple putative *cis*-splicing donor sites, which can be indicative that cis-splicing in the 5′ UTR regions is less specific than other splice junctions in the same gene.

Second, we analyzed how common 5′ UTR introns are in *H. microstoma*. To this end, we manually analyzed the top 500 expressed genes (from Preza et al., 2021) to search for evidence from RNASeq data of the presence of 5′ UTR introns in their gene structure (highly expressed genes were selected in order to have good coverage of the 5′ region). After discarding genes encoded too close to their upstream neighbors for reliable inspection (including operons) and other difficulties (e.g., bad read alignment, incorrect gene models, etc.) 403 genes remained, of which 82 had evidence of *cis*-splicing in the 5′ UTR region (see notes in Supplementary Table S13). From these 403 genes, 12 were SL-*trans*-spliced, and all of these had evidence of *cis*-splicing in the 5′ UTR region. This association between 5′ UTR *cis*-splicing and SL *trans*-splicing is highly significant, *P*-value < 0.001, Fisher Exact Test. Therefore, although 5′ UTR introns are not inevitably linked to SL *trans*-splicing, they appear to be a major factor in determining which transcripts are *trans*-spliced. The presence of internal promoters in some of these 5′ UTR introns, which would result in the production of primary transcripts with strong splicing acceptor

sites in the absence of competing splicing donor sites, could be one of the main drivers of SL *trans*-splicing in a subset of these genes.

3.5. Chimeric gene models in the genomic annotation of H. microstoma

Given the high number of internal SL insertions observed (Section 3.1), we explored the possibility of error in the genome annotation regarding gene delimitation. Specifically, we divided the coding sequences that are *trans*-spliced internally before and after the insertion site, and conducted BLAST searches against *E. multilocularis* gene predictions available in the Wormbase Parasite database. A total of 102 genes in *H. microstoma* matched different genes in *E. multilocularis*, according to this analysis. Most of them had weak or absent evidence of *cis*-splicing connecting both halves, as determined from the presence of few or no reads across splice junctions associated with the *trans*-splicing acceptor site (Supplementary Table S14). These results indicate that for most of these 102 genes, the SL insertion sites that we identified do not correspond to internal insertions, but rather to 5′ insertions of downstream genes in operons, which are incorrectly fused to the upstream gene in the current gene predictions. The source of these discrepancies between species likely is a combination of different choices in strategy between Tsai et al. (2013) and Olson et al. (2020), and the inherent biological complexity of some of these operon arrangements.

To illustrate the latter, we selected the deceptively simple gene model HmN_002164300 that encodes for a small protein of unknown function. According to our results, it has an intronic SL insertion on site 7073450 (supported by four reads) and another on 7073468 in the predicted acceptor site for *cis*-splicing (supported by 389 reads); both located in the third intron (Fig. 4). According to our BLAST searches, this gene matches two different genes (EmuJ_000069290 and EmuJ_000069310) in *E. multilocularis*, which comprise operon N° 11 in Tsai et al. (2013). There is no evidence supporting the existence of the intron between exons 4 and 5 as predicted in WormBase Parasite, and we refer to this last exon, including the incorrectly predicted intron, as "exon 4 + 5", which fully contains an open reading frame homologous to EmuJ_000069310.

As shown in Fig. 4A and B, our RNASeq data indicate that this locus encodes for at least seven potential isoforms: two isoforms are generated by transcription from exon 1 and include all downstream exons, with an alternative donor site in exon 2. Two other isoforms are generated by SL *trans*-splicing of the acceptor sites upstream of exon 4 + 5, and another isoform (similar to the *trans*-spliced isoforms) is generated by the direct *cis*-splicing of exon 1 (encoding a single "ATG" codon) with exon 4 + 5. Finally, two other potential isoforms comprised only the first three exons, which would be generated indirectly by downstream *trans*-splicing (an alternative stop codon is observed almost immediately after the end of exon 3).

RT-PCR experiments are consistent with the existence of isoforms 2–5, including the two *trans*-spliced transcripts (Fig. 4C; our amplification strategies cannot, by design, amplify isoforms 6 and 7). Cloning and sequencing of amplicons confirmed the existence of isoforms 2 and 5, as well as an intermediary product (int1) in which *cis*-splicing was completed but *trans*-splicing had not yet occurred (Fig. 4B). Additionally, an eighth isoform was identified during cloning and sequencing, consisting of Exon 1, a shorter form of Exon 3 and Exon 4 + 5 (Fig. 4B). Isoforms 1, 2, 6, and 7 code for an open reading frame homologous to EmuJ_000069290, whereas isoforms 3, 4, and 5 code for an open reading frame homologous to EmuJ_000069310. In summary, locus HmN_002164300 is very complex and appears to function sometimes as a single gene with alternative *cis*-splicing, or as two genes of a polycistron that is resolved through *trans*-splicing. In this sense, it is reminiscent of the alternative operons that have been described in *C. elegans* (Morton and Blumenthal, 2011; Blumenthal et al., 2015).

3.6. Trans-splicing in operons of H. microstoma

To estimate the role of the SL insertions in the processing of polycistronic transcripts, we searched for SL *trans*-splicing of genes in operons in the genome of *H. microstoma*. A total of 192 gene clusters met our stringent criteria to be considered putative operons (genes encoded in the same strand, with an intergenic distance shorter than 300 bp with continuous RNASeq coverage above five reads), 82 of which are associated with SL insertions (Supplementary Table S15). Only 12 SL acceptor genes were found at the start of putative operons, of which seven bear internal SL insertion sites (genes HmN_000117840, HmN_000465900, HmN_000484800, HmN_002068700, HmN_003008490, HmN_002052000, HmN_000107500, HmN_000179600, and HmN_002164300) and are among the fusion artifacts described in Section 3.5 (Supplementary Table S13), thus actually corresponding to insertions in the second gene of the operon. The vast majority of *trans*-splicing events were detected in downstream genes, as expected for a major role in polycistron resolution. Of the original 192 operon candidates for *H. microstoma*, 111 matched homologous operons reported in *E. multilocularis* by Tsai et al. (2013), of which 69 showed SL insertions in both species (Supplementary Table S15). Adding the chimeric gene models described in Section 3.5, the number of homologous operons in both species rises to 146 (Supplementary Table S13) involving 125 gene models subjected to *trans*-splicing, although the distance between them is often above the 300 bp mark. We conclude that operon organization appears to be highly conserved between these two cestode species.

Eight putative operons were selected and successfully validated by RT-PCR (Op_18, Op_36, Op_44, Op_59 and Op_67, Op_137, Op_147 and Op_151) (Fig. 5). In summary, the SL acceptor site identified by RNAseq data was confirmed in all cases using an SL-3 forward primer and a gene-specific reverse primer, and an additional SL acceptor site which was not detected by our in-silico analysis was also identified for Op_151. Furthermore, using a strategy similar to that used by Tsai et al. (2013), RT-PCR products could be identified spanning the final exon of the first gene, the intergenic region, and the first exons of the second gene in the operon, confirming the existence of polycistronic transcripts. Many of these RT-PCR products included *cis*-splicing events, as expected from incompletely processed transcripts in which *cis*-splicing has already begun before the less efficient *trans*-splicing is completed (Brehm et al., 2000; Tsai et al., 2013). Additional RT-PCR products, consistent with retention of intron sequences, were observed for the intermediary amplicons of Op_137, Op_147, and Op_151 and the *trans*-spliced amplicons of Op_18, Op_44, and Op_67. The lack of amplification in control reactions, without reverse transcription, discarded genomic contamination. A detailed exploration of each operon is shown in Supplementary Fig. S2.

3.7. Intron properties associated with internal SL acceptor sites

After excluding the internal acceptor sites associated with confirmed annotation artifacts (Section 3.5), 206 internal sites remained. Internal *trans*-splicing is thought to be rare, as it would usually be out-competed by *cis*-splicing. Therefore, the intronic regions upstream of the internal acceptor sites were analyzed in search of distinctive characteristics relative to exclusively *cis*-spliced introns. Globally, we detected a slight increase in size and a lower GC content relative to exclusively *cis*-spliced introns (Fig. 6A).

In addition, given the existence of internal acceptor sites across the entire lengths of genes, we verified if SL insertions toward both

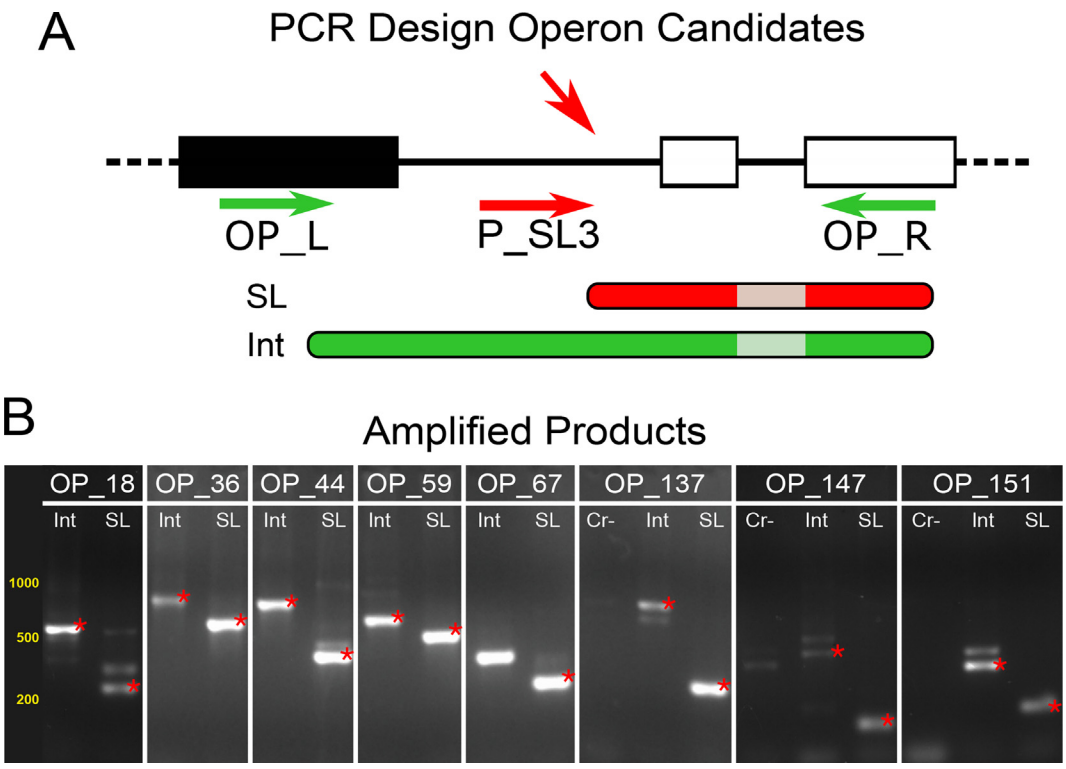

Fig. 5. Experimental validation of operons. (A) Diagram of the target sequences in the genome and primer locations. The last exon of the upstream gene is represented in a black box and the first two exons of the downstream one in white boxes. The location of the spliced leader (SL) insertion is represented by the arrow. For all operons, the left primer (OP_L) was designed in the last exon of the first gene and the right primer (OP_R) on the second exon of the following gene. (B) Amplified products for each putative operon (SL indicates amplicons obtained using a SL forward primer, and Int indicates amplicons of intermediate splicing products obtained using a forward primer located in the last exon of the first gene of the operon). Those indicated with red asterisks were confirmed by Sanger sequencing. Operons OP_137, OP_147, and OP_151 were selected to control the absence of genomic DNA contamination, with additional negative controls lacking reverse transcriptase (Cr-).

ends were equal in terms of their surrounding introns. Insertions located at the first and last three exons were compared with the immediately upstream and downstream introns in terms of length and GC content (Fig. 6B and C). According to our results, introns upstream of the SL insertion are significantly longer in the 3′ end of the genes, assuming that current gene models are accurate (but see Section 4.3), and have significantly lower GC% in the 5′ end of genes.

3.8. Some internal SL insertions are associated with the generation of isoforms with different cellular localizations

Internal SL *trans*-splicing addition could result in truncated proteins, with different possible functional consequences. Although inactivation is the most obvious one, other possibilities include the inclusion or removal of signal sequences that could modify subcellular localization. We searched for secretion signal peptides (SP) and mitochondrial transit peptides (mTP) exposed or removed by the SL insertions with SignalP and TargetP. These resulted in two strong examples of internal SL insertions affecting the subcellular localization of the resulting isoforms (Supplementary Fig. S3). In one case (HmN_00248500, N-acyl-phosphatidylethanolamine-hydrolyzing phospholipase d), a secreted isoform would be produced exclusively through *cis*-splicing, and a cytosolic form would be generated by internal SL *trans*-splicing. In the other (HmN_000530500, "Iron dependent peroxidase"), a mTP is exposed by the internal SL insertion.

Therefore, although a few potential examples of switching of subcellular localization by SL *trans*-splicing could be identified, it seems clear that this is not a major function of internal SL *trans*-splicing in this species.

4. Discussion

4.1. Four different SL sequences are trans-spliced to a large set of common transcripts in H. microstoma

In broad strokes, our results align with and expand those of Olson et al. (2020): *H. microstoma* harbors a limited complement of SL coding loci that are used interchangeably across all acceptor sites. The most notable difference between our results is the relative relevance of SL *trans*-splicing in adult worms and larvae, as we did not detect significant differences between life stages. It must be noted that both studies used different larval samples (larvae mid-metamorphosis in the case of Olson et al. (2020) and fully developed and infective cysticercoids in this work), which could explain in part the observed differences. However, reads containing SL were very rare in adult samples from Olson et al. (2020), suggesting that other factors, such as RNA integrity of these specific samples, could be at play. Another observed difference between both studies is the relative abundance of the different SL RNAs in the transcriptome, with SL-1 possessing an abundance more similar to SL-2 and SL-3 in Olson et al. (2020) samples. This difference is harder to explain since lower data quality in RNA samples should translate to overall lower SL reads, regardless of their sequence. The samples were obtained under slightly different conditions in the two different laboratories where the specimens were bred, which could indicate an effect of the environment on the relative abundance of SL types. It is important to note that SL-1 is encoded in two loci, which could have different expression levels. At this time, there is not enough data to support any of these hypotheses, and the differences observed between both studies are subtle.

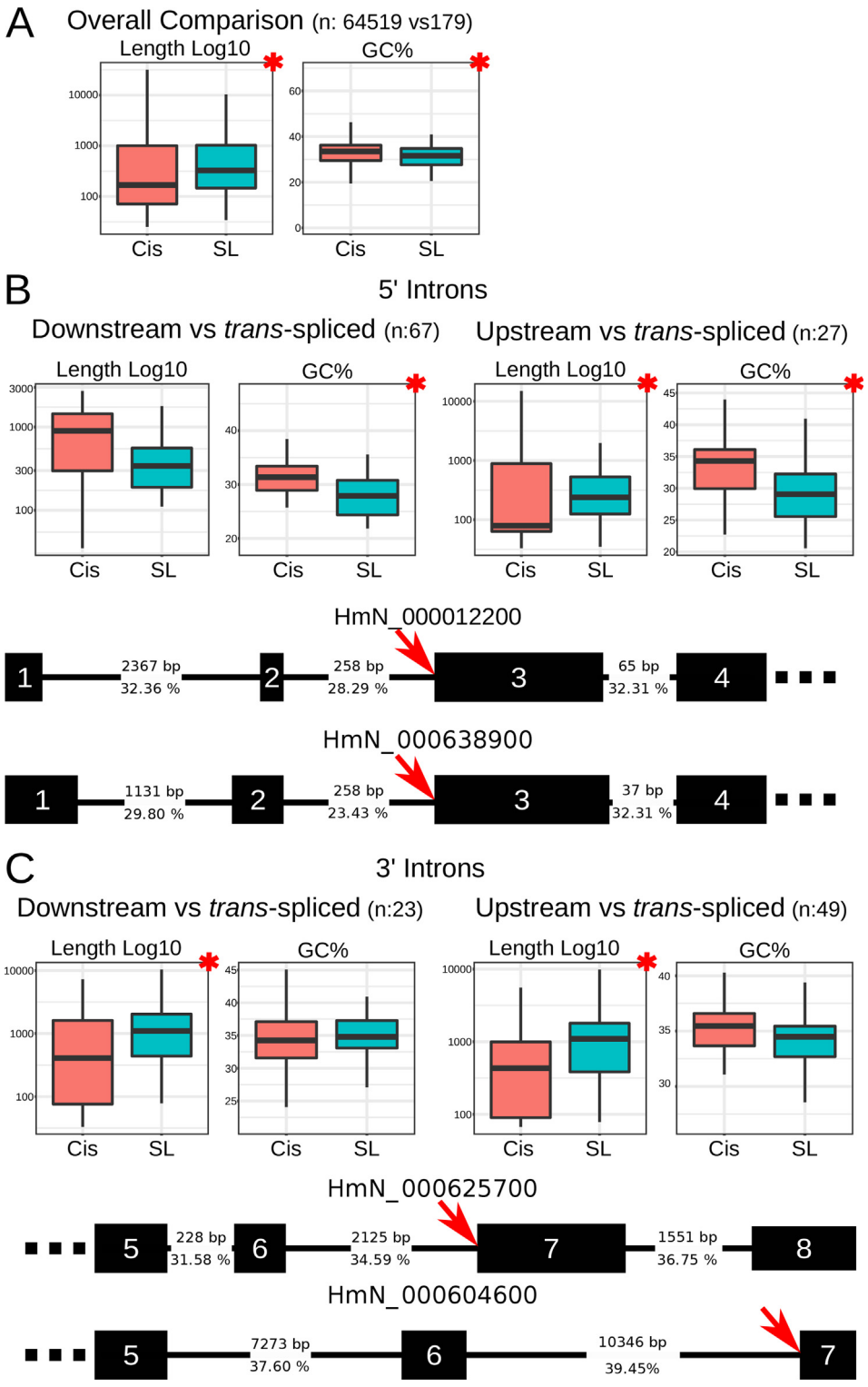

Fig. 6. Differences in length and GC content between introns exclusively subjected to *cis*-splicing or subjected to spliced leader (SL) *trans-splicing*, significant differences by Wilcoxon signed-rank test (*P*-value < 0.05) are indicated with an asterisk. (A) Overall differences between all reported introns. (B) SL *trans*-spliced introns located on the 5′ region of the gene compared with their immediate downstream and upstream introns. Two illustrative examples are included. (C) SL *trans*-spliced introns located on the 3′ region of the gene compared with their immediate downstream and upstream introns. Two illustrative examples are included.

In terms of the predicted structure, all SLs in *H. microstoma* contain three hairpins, with the donor site being located in the first hairpin and the first portion of the SL being left unpaired, similar to most SLs described by Krchňáková et al. (2017) and previously reported for other cestodes (Brehm et al., 2000). It is important to note, however, that direct SL RNA structure comparison between different studies can be misleading because they can change greatly depending on the specific methodology applied (see Krchňáková et al. (2017) compared with Bitar et al. (2013)). A full examination will require characterizing the SL-RNA complete sequence and must factor in the hypermethylated 5′ cap. Nevertheless, it is interesting to note that SL-2 and SL-3, the most abun-

dant SLs, have a similar structure in terms of hairpin size and a lower minimum free energy value than SL-1 and SL-4. At the sequence level, the well-known high variability of SL-RNAs is clear in our analysis, although there are distinguishable conserved short regions at smaller phylogenetic scales that offer interesting insights. The SM-like sites, crucial for the interaction with the spliceosome, concentrate their mismatches on their 5′ portions, suggesting that this area is less constrained by purifying selection. This should be factored into future studies since the identification of an SM site has been taken as an important factor in the identification of new SL-RNAs (Wenzel et al., 2021; Islas-Flores et al., 2021). In any case, it is important to note that the observed SM variants do not correlate with the SL abundance.

Intriguingly, *Hymenolepis* SL RNAs are very divergent relative to the taeniid species included in this study. One possibility is that SL-RNA loci within Taeniidae experienced a bottleneck after the divergence and later expanded in copy number in *E. multilocularis*, followed by lineage-specific duplications.

4.2. SL trans-splicing is not related to specific gene functions

We observed that SLs are incorporated mainly on constitutively expressed genes across life stages, and only a fraction was associated with operons. Thus, in line with previous works, we concluded that there is no clear functional specialization among the genes subjected to SL *trans*-splicing. Furthermore, GO term enrichment of *trans*-spliced genes was largely restricted to ubiquitous biosynthetic pathways. We also observed that genes where the SL insertion clearly affected protein cellular localization were limited. As Soulette et al. (2019) observed for trypanosomes, the bulk of splicing events are constitutive rather than leading to alternative isoforms. Exceptions that cannot be attributed to annotation artifacts are limited, and the actual role of SL *trans*-splicing in cases such as *eg-tgr* is probably secondary to the presence of alternative promoters. This may be due, in part, to practical limitations for the identification of alternative *trans*-splicing isoforms. As reported in this study and other organisms such as *C. elegans* (Allen et al., 2011), SLs can be incorporated at the terminal exons of some genes when separated from the rest by a long intron. It is difficult to determine if these *trans*-splicing events result from SL *trans*-splicing competing with inefficient *cis*-splicing donors or from the presence of cryptic promoters in these introns (both possibilities would be compatible with the longer length of these introns). Examples such as *tgr* in *Echinococcus,* and other candidates with SP or mTP signals affected by SL insertions, indicate that *trans*-splicing can play an important role in shaping the final protein function and/or location. However, the sparsity of clear examples suggests that the overall picture is more consistent with a neutral or even semi-deleterious effect, akin to the mis-splicing events described for *cis*-splicing (Saudemont et al., 2017). Even if internal insertions are background noise, however, an intriguing potential side effect is that terminal exons subjected to a minimal level of SL *trans*-splicing could act as "seeds" for novel genes. These would at first be encoded on operons with long intergenic spaces similar to those confirmed by Morton and Blumenthal (2011) in *C. elegans*, eventually shortening to more typical distances by deletion or becoming fully independent loci. The viability and potential predominance of this mechanism need to be addressed in a more comprehensive study.

4.3. Conservation of operons and polycistronic processing by SL trans-splicing in cestodes

The high level of conservation between operons of *E. multilocularis* and *H. microstoma*, especially after accounting for gene fusion artifacts, and the shared locations of the SL acceptor sites within them, highlights the importance of SL *trans*-splicing in operon resolution, in concordance with previous studies (Hastings, 2005; Lasda and Blumenthal, 2011; Bitar et al., 2013). Although many of them implied longer intergenic distances than those used by Tsai et al. (2013) as the limit in *E. multilocularis*, long polycistronic transcripts have also been observed, for example, in *C. elegans* (Morton and Blumenthal, 2011).

Our detailed analysis showed that the presence of operons, and their inherent complexity, has resulted in some artifacts in the automatic gene predictions of *H. microstoma*. The problem stems from the low efficiency of automatic annotation pipelines to properly process operons from RNAseq data, particularly in cases where the differentiation between genes and isoforms becomes almost arbitrary. The existence of chimeric gene models is troublesome given the frequency at which novel genomes are annotated by transferring known genes from similar species (see Kamenetzky et al., 2022).

Due to several biases, the extent of the issue is hard to assess from our work alone. First, we probably underestimate the number of SL insertion sites due to the methodology employed to sequence the transcriptome (no method was used to enrich the RNASeq data for SL trans-spliced transcripts, in comparison to, for example, Cuypers et al. (2017) and Boroni et al. (2018)). Additionally, the strategy relies on the correctness of the *E. multilocularis* annotation since only cases where both genes were identified and properly annotated in this species will be detected in our BLAST-based strategy. On the other hand, the divide-and-BLAST approach can lead to false positives in cases where one of the halves codes for a conserved domain that is common in multiple genes of the same protein family, with alternative splicing not annotated in a comparable manner in both species. Despite these caveats, our work shows that precise identification of SL acceptor sites is a valuable input for gene annotation, as it can be used as a key element to delineate genes within operons and help guide the manual annotation of more complex loci.

4.4. Association of trans-splicing of monocistronic transcripts with 5′UTR introns

Our survey of monocistronic genes shows many SL insertions utilizing the same acceptor sites at the 5′ UTR (or at the beginning of the gene coding region) as those used by conventional *cis*-splicing, rather than insertions at outrons. The latter is typical for other eukaryotes. The presence of identifiable *cis*-splicing junctions and in such high frequency among these SL acceptor sites indicates that *trans*-splicing and *cis*-splicing on these acceptor sites are, in many cases, functionally equivalent. Similar examples have also been described in the trematode *S. mansoni* (e.g., HMG CoA Reductase gene, Rajkovic et al., 1990; Davis et al., 1995). In the case of the *elp* and *tgr* genes of *Echinococcus* spp., these *trans*-splicing events have been linked to the presence of internal promoters in introns (Brehm et al., 2000; Agorio et al., 2003). This results in transcripts lacking a donor splice site but with a strong acceptor site that is ideally suited for the acceptor-first syntax required for efficient *trans*-splicing (Hastings, 2005). The structure of the 5′ UTR region of *elp* in *H. microstoma* is surprisingly conserved and results in a *trans*-spliced isoform in which the SL is inserted at the beginning of the second exon. In contrast, the *tgr* ortholog of *H. microstoma* lacks *trans*-splicing in all its isoforms, although it also has 5′ UTR introns. These differences could be explained if the corresponding introns of *H. microstoma* lack internal promoters or have only weak transcriptional activity, which would be consistent with their shorter length compared with the corresponding introns in *Echinococcus* spp. In the future, expanding the analysis of these well-studied genes to other flatworms could provide a primer for the evolutionary analysis of the gain and loss of SL *trans*-splicing.

To the best of our knowledge, this strong association of *trans*-splicing with *cis*-splicing sites in the 5′ UTR region has not been described in other eukaryotes. Future work will be needed to

determine if this phenomenon is true for other flatworms. It is possible that it is also relevant, but has not been noticed, in other eukaryotes for which 5′ UTR regions are not well defined experimentally. However, this would be unlikely for well-studied models such as *C. elegans*. Which reasons could explain this difference between *trans*-splicing in cestodes and other organisms? One possible explanation could be that *trans*-splicing in these species may require strong splicing acceptor sites (beyond the minimal AG consensus), which would be unlikely to appear by chance in the 5′ UTR region of genes. Thus, a large proportion of *trans*-splicing could be related to pre-existing *cis*-splicing acceptor sites that become available for *trans*-splicing if the donor site is not part of the same transcript. Alternatively, transcription initiation may be relatively lax, leading to the evolutionary appearance of many internal promoters from intronic regions. Finally, the presence of an AUG codon at the 3′ end of the spliced leader could facilitate the evolutionary incorporation of acceptor sites for internal exons in the beginning of the coding region, as it would allow translational initiation in some cases.

Another interesting observation is the abundance of *trans*-spliced 5′ UTRs with multiple *cis*-splicing donor sites (see, for instance, gene models HmN_000162500, HmN_000205900, and HmN_000234700), especially when compared with the splice junctions within the coding regions of the gene models. In addition to the acceptor and donor sites, *cis*-splicing is also regulated by the presence of enhancer and repressor motifs located both in both exons and introns of the gene (Lee and Rio, 2015; Ule and Blencowe, 2019), mutation over which can lead to alterations of the normal splicing of the locus (Park et al., 2018). Given these multiple donor sites and the lack of a clear association between SL *trans*-splicing and gene function (see Section 4.2), it is unlikely they represent isoforms with distinctive gene functions. Instead, they can be evidence of relaxation of the selective pressures over the *cis*-splicing regulatory elements upstream of the SL *trans*-splicing site or a side effect of the weakened *cis*-splicing at the site that allows *trans*-splicing to occur.

Acknowledgements

J.C. is a recipient of a doctoral scholarship from Agencia Nacional de Investigación e Innovación (ANII), Uruguay. U.K. and A.I. are members of the Uruguayan National Researchers System (SNI), and PEDECIBA, Uruguay. This work was supported by grant FCE_3_2016_1_125297 to A.I. from ANII, Uruguay. The funding agency played no role in the design of the study, analysis, interpretation of data, writing of the manuscript or in the decision to submit the manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpara.2022.11.006.

References

Agorio, A., Chalar, C., Cardozo, S., Salinas, G., 2003. Alternative mRNAs arising from trans-splicing code for mitochondrial and cytosolic variants of *Echinococcus granulosus* thioredoxin glutathione reductase. J. Biol. Chem. 278, 12920–12928. https://doi.org/10.1074/jbc.M209266200.

Allen, M.A., Hillier, L.D.W., Waterston, R.H., Blumenthal, T., 2011. A global analysis of *C. elegans* trans-splicing. Genome Res. 21, 255–264. https://doi.org/10.1101/gr.113811.110.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Armenteros, J.J.A., Salvatore, M., Emanuelsson, O., Winther, O., Von Heijne, G., Elofsson, A., Nielsen, H., 2019a. Detecting sequence signals in targeting peptides using deep learning. Life Sci. Alliance 2, e201900429 https://doi.org/10.26508/lsa.201900429.

Armenteros, J.J.A., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., Nielsen, H., 2019b. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. 37, 420–423. https://doi.org/10.1038/s41587-019-0036-z.

Bitar, M., Boroni, M., Macedo, A.M., Machado, C.R., Franco, G.R., 2013. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. Front. Genet. 4, 199. https://doi.org/10.3389/fgene.2013.00199.

Blumenthal, T., 2004. Operons in eukaryotes. Brief. Funct. Genomic. Proteomic. 3, 199–211. https://doi.org/10.1093/bfgp/3.3.199.

Blumenthal, T., Davis, P., Garrido-Lecca, A., 2015. Operon and non-operon gene clusters in the *C. elegans* genome. https://doi.org/10.1895/wormbook.1.17.

Blumenthal, T., Trans-splicing and operons. In: WormBook: The Online Review of C. elegans Biology. https://doi.org/10.1895/wormbook.1.5.1.

Bolger, A.M., Lohse, M., Usadel, B., Planck, M., Plant, M., Mühlenberg, A., 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Boroni, M., Sammeth, M., Gava, S.G., Jorge, N.A.N., MacEdo, A.M., MacHado, C.R., Mourão, M.M., Franco, G.R., 2018. Landscape of the spliced leader trans-splicing mechanism in *Schistosoma mansoni*. Sci. Rep. 8, 3877. https://doi.org/10.1038/s41598-018-22093-3.

Brehm, K., Jensen, K., Frosch, M., 2000. mRNA trans-splicing in the human parasitic cestode *Echinococcus multilocularis*. J. Biol. Chem. 275, 38311–38318. https://doi.org/10.1074/jbc.M006091200.

Cabada, M.M., Lopez, M., Cruz, M., Delgado, J.R., Hill, V., 2016. Treatment Failure after Multiple Courses of Triclabendazole among Patients with Fascioliasis in Cusco, Peru: A Case Series. PLoS Negl. Trop. Dis. 10, e0004361.

Calvelo, J., Juan, H., Musto, H., Koziol, U., Iriarte, A., 2020. SLFinder, a pipeline for the novel identification of splice-leader sequences: A good enough solution for a complex problem. BMC Bioinform. 21, 293. https://doi.org/10.1186/s12859-020-03610-6.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., Huerta-Cepas, J., 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. 1–5. https://doi.org/10.1093/molbev/msab293.

Cheng, G., Cohen, L., Ndegwa, D., Davis, R.E., 2006. The flatworm spliced leader 3′-terminal AUG as a translation initiator methionine. J. Biol. Chem. 281, 733–743. https://doi.org/10.1074/jbc.M506963200.

Cunningham, L.J., Olson, P.D., 2010. Description of *Hymenolepis microstoma* (Nottingham strain): A classical tapeworm model for research in the genomic era. Parasites Vectors 3, 123. https://doi.org/10.1186/1756-3305-3-123.

Cuypers, B., Domagalska, M.A., Meysman, P., Muylder, G.D., Vanaerschot, M., Imamura, H., Dumetz, F., Verdonckt, T.W., Myler, P.J., Ramasamy, G., Laukens, K., Dujardin, J.C., 2017. Multiplexed Spliced-Leader Sequencing: A high-throughput, selective method for RNA-seq in Trypanosomatids. Sci. Rep. 7, 3725. https://doi.org/10.1038/s41598-017-03987-0.

Davis, R.E., Hardwick, C., Tavernier, P., Hodgson, S., Singh, H., 1995. RNA trans-splicing in flatworms. Analysis of trans-spliced mRNAs and genes in the human parasite *Schistosoma mansoni*. . J Biol Chem 270, 21813–21819.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Douris, V., Telford, M.J., Averof, M., 2010. Evidence for Multiple Independent Origins of trans-Splicing in Metazoa. Mol. Biol. Evol. 27, 684–693. https://doi.org/10.1093/molbev/msp286.

Drillon, G., Carbone, A., Fischer, G., 2014. SynChro: A fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. PLoS One 9, e9262.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340.

Fisher, R.A., 1934. Statistical methods for research workers. Oliver and Boyd, Edinburgh.

Garrido-Martín, D., Palumbo, E., Guigó, R., Breschi, A., 2018. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. PLoS Comput. Biol. 14, e1006360.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512. https://doi.org/10.1038/nprot.2013.084.

Hannon, G.J., Maroney, P.A., Nilsen, T.W., 1991. U small nuclear ribonucleoprotein requirements for nematode cis- and trans-splicing in vitro. J. Biol. Chem. 266, 22792–122695.

Harrison, N., Kalbfleisch, A., Connolly, B., Pettitt, J., Muller, B., 2010. SL2-like spliced leader RNAs in the basal nematode *Prionchulus punctatus*: New insight into the evolution of nematode SL2 RNAs. RNA 16, 1500–1507. https://doi.org/10.1261/rna.2155010.

Hastings, K.E.M., 2005. SL trans-splicing: Easy come or easy go? Trends Genet. 21, 240–247. https://doi.org/10.1016/j.tig.2005.02.005.

Hickman, C., Roberts, L., Keen, S., Larson, A., Anson, H., Eisenhour, D., 2008. Integrated Principles of Zoology. McGraw-Hill Higher Education, Boston, Massachusetts.

Howe, K.L., Bolt, B.J., Shafie, M., Kersey, P., Berriman, M., 2017. WormBase ParaSite – a comprehensive resource for helminth genomics. Mol. Biochem. Parasitol. 215, 2–10. https://doi.org/10.1016/j.molbiopara.2016.11.005.

Islas-Flores, T., Galán-Vásquez, E., Villanueva, M.A., 2021. Screening a spliced leader-based *Symbiodinium microadriaticum* cDNA library using the yeast-two hybrid system reveals a hemerythrin-like protein as a putative smicRACK1 ligand. Microorganisms 9, 791. https://doi.org/10.3390/microorganisms9040791.

Kamenetzky, L., Maldonado, L.L., Cucher, M.A., 2022. Cestodes in the genomic era. Parasitol. Res. 121, 1077–1089. https://doi.org/10.1007/s00436-021-07346-x.

Katz, Y., Wang, E.T., Silterra, J., Schwartz, S., Wong, B., Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., Airoldi, E.M., Burge, C.B., 2015. Quantitative visualization of alternative exon expression from RNA-seq data. Bioinformatics 31, 2400–2402. https://doi.org/10.1093/bioinformatics/btv034.

Kerpedjiev, P., Hammer, S., Hofacker, I.L., 2015. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics 31, 3377–3379. https://doi.org/10.1093/bioinformatics/btv372.

Krause, M., Hirsh, D., 1987. A trans-spliced leader sequence on actin mRNA in *C. elegans*. Cell 49, 753–761. https://doi.org/10.1016/0092-8674(87)90613-1.

Krchňáková, Z., Krajčovič, J., Vesteg, M., 2017. On the Possibility of an Early Evolutionary Origin for the Spliced Leader Trans-Splicing. J. Mol. Evol. 85, 37–45. https://doi.org/10.1007/s00239-017-9803-y.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874. https://doi.org/10.1093/molbev/msw054.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923.

Lasda, E.L., Blumenthal, T., 2011. Trans-splicing. Wiley Interdiscip. Rev. RNA 2, 417–434. https://doi.org/10.1002/wrna.71.

Lee, Y., Rio, D.C., 2015. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu. Rev. Biochem. 84, 291–323. https://doi.org/10.1146/annurev-biochem-060614-034316.

Lei, Q., Li, C., Zuo, Z., Huang, C., Cheng, H., Zhou, R., 2016. Evolutionary Insights into RNA trans-Splicing in Vertebrates. Genome Biol. Evol. 8, 562–577. https://doi.org/10.1093/gbe/evw025.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Data, G.P., Sam, T., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Lidie, K.B., Van Dolah, F.M., 2007. Spliced leader RNA-mediated trans-splicing in a dinoflagellate, *Karenia brevis*. J. Eukaryot. Microbiol. 54, 427–435. https://doi.org/10.1111/j.1550-7408.2007.00282.x.

Logan-Klumpler, F.J., De Silva, N., Boehme, U., Rogers, M.B., Velarde, G., McQuillan, J. A., Carver, T., Aslett, M., Olsen, C., Subramanian, S., Phan, I., Farris, C., Mitra, S., Ramasamy, G., Wang, H., Tivey, A., Jackson, A., Houston, R., Parkhill, J., Holden, M., Harb, O.S., Brunk, B.P., Myler, P.J., Roos, D., Carrington, M., Smith, D.F., Hertz-Fowler, C., Berriman, M., 2012. GeneDB-an annotation database for pathogens. Nucleic Acids Res. 40, D98–D108. https://doi.org/10.1093/nar/gkr1032.

Lorenz, R., Bernhart, S.H., Siederdissen, C.H.Z., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L., 2011. ViennaRNA Package 2.0. Algorithms Mol. Biol. 6, 26. https://doi.org/10.1093/nar/gkz164.

Macnish, M.G., Ryan, U.M., Behnke, J.M., Thompson, R.C.A., 2003. Detection of the rodent tapeworm *Rodentolepis* (=*Hymenolepis*) *microstoma* in humans. A new zoonosis? Int. J. Parasitol. 33, 1079–1085. https://doi.org/10.1016/S0020-7519(03)00137-1.

Mahmud, R., Ai, Y., Lim, L., 2017. Medical Parasitology. Springer International Publishing, Cham, Switzerland. https://doi.org/https://doi.org/10.1007/978-3-319-68795-7.

Marcel, M., 2011. CUTADAPT removes adapter sequences from high-throughput sequencing reads. EMBnet 17, 10–12 https://doi.org/10.14806/ej.17.1.200.

Matsuo, M., Katahata, A., Satoh, S., Matsuzaki, M., 2018. Characterization of spliced leader trans-splicing in a photosynthetic rhizarian amoeba, *Paulinella micropora*, and its possible role in functional gene transfer. PLoS One 13, e0200961.

Morton, J.J., Blumenthal, T., 2011. Identification of transcription start sites of trans-spliced genes: Uncovering unusual operon arrangements. RNA 17, 327–337. https://doi.org/10.1261/rna.2447111.

Nilsson, D., Gunasekera, K., Mani, J., Osteras, M., Farinelli, L., Baerlocher, L., Roditi, I., Ochsenreiter, T., 2010. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. PLoS Pathog. 6, e1001037.

Olson, P., Tracey, A., Baillie, A., James, K., Doyle, S., Buddenborg, S., Rodgers, F., Holroyd, N., Berriman, M., 2020. Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection. BMC Biol. 18, 165. https://doi.org/10.1101/2020.04.08.031872.

Olson, P.D., Zarowiecki, M., James, K., Baillie, A., Bartl, G., Burchell, P., Chellappoo, A., Jarero, F., Tan, L.Y., Holroyd, N., Berriman, M., 2018. Genome - wide transcriptome profiling and spatial expression analyses identify signals and switches of development in tapeworms. Evodevo 9, 1–29. https://doi.org/10.1186/s13227-018-0110-5.

Otero, L., Bonilla, M., Protasio, A.V., Fernández, C., Gladyshev, V.N., Salinas, G., 2010. Thioredoxin and glutathione systems differ in parasitic and free-living platyhelminths. BMC Genomics 11, 237. https://doi.org/10.1186/1471-2164-11-237.

Pandarakalam, G.C., Speake, M., McElroy, S., Alturkistani, A., Philippe, L., Pettitt, J., Müller, B., Connolly, B., 2019. A high-throughput screen for the identification of compounds that inhibit nematode gene expression by targeting spliced leader trans-splicing. Int. J. Parasitol. Drugs Drug Resist. 10, 28–37. https://doi.org/10.1016/j.ijpddr.2019.04.001.

Park, E., Pan, Z., Zhang, Z., Lin, L., Xing, Y., 2018. The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am. J. Hum. Genet. 102, P11–P26. https://doi.org/10.1016/j.ajhg.2017.11.002.

Pettitt, J., Harrison, N., Stansfield, I., Connolly, B., Müller, B., 2010. The evolution of spliced leader trans-splicing in nematodes. Biochem. Soc. Trans. 38, 1125–1130. https://doi.org/10.1042/BST0381125.

Pettitt, J., Philippe, L., Sarkar, D., Johnston, C., Gothe, H.J., Massie, D., Connolly, B., Müller, B., 2014. Operons are a conserved feature of nematode genomes. Genetics 197, 1201–1211. https://doi.org/10.1534/genetics.114.162875.

Pouchkina-Stantcheva, N.N., Tunnacliffe, A., 2005. Spliced leader RNA-mediated trans-splicing in phylum rotifera. Mol. Biol. Evol. 22, 1482–1489. https://doi.org/10.1093/molbev/msi139.

Preza, M., Calvelo, J., Langleib, M., Hoffmann, F., Castillo, E., Koziol, U., Iriarte, A., 2021. Stage-specific transcriptomic analysis of the model cestode *Hymenolepis microstoma*. Genomics 113, 620–632. https://doi.org/10.1016/j.ygeno.2021.01.005.

Protasio, A.V., Tsai, I.J., Babbage, A., Nichol, S., Hunt, M., Aslett, M.A., de Silva, N., Velarde, G.S., Anderson, T.J.C., Clark, R.C., Davidson, C., Dillon, G.P., Holroyd, N.E., LoVerde, P.T., Lloyd, C., McQuillan, J., Oliveira, G., Otto, T.D., Parker-Manuel, S.J., Quail, M.A., Wilson, R.A., Zerlotini, A., Dunne, D.W., Berriman, M., 2012. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. PLoS Negl. Trop. Dis. 6, e1455.

Rajkovic, A., Davis, R.E., Simonsen, J.N., Rottman, F.M., 1990. A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. Proc. Natl. Acad. Sci. U. S. A. 87, 8879–8883. https://doi.org/10.1073/pnas.87.22.8879.

Rice, P., Longden, L., Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.

Rossi, A, Ross E.J., Jack, A, Alvarado, A.S., 2014. Molecular cloning and characterization of SL3: A stem cell- specific SL RNA from the planarian *Schmidtea mediterranea*. Gene 533, 156–167. https://doi.org/10.1016/j.gene.2013.09.101.

Sather, S., Agabian, N., 1985. A 5' spliced leader is added in trans to both alpha- and beta-tubulin transcripts in *Trypanosoma brucei*. Proc. Natl. Acad. Sci. U. S. A. 82, 5695–5699. https://doi.org/10.1073/pnas.82.17.5695.

Saudemont, B., Popa, A., Parmley, J.L., Rocher, V., Blugeon, C., Necsulea, A., Meyer, E., Duret, L., 2017. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biol. 18, 208. https://doi.org/10.1186/s13059-017-1344-6.

Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One 11, e0163962.

Soulette, C.M., Oliverio, O., Roy, S.W., 2019. On the Function of Trans-Splicing: No Evidence for Widespread Proteome Diversification in Trypanosomes. Genome Biol. Evol. 11, 3014–3021. https://doi.org/10.1093/gbe/evz217.

Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. https://doi.org/10.1093/bib/bbs017.

Trainor-moss, S., Mutapi, F., 2016. Schistosomiasis therapeutics: whats in the pipeline? Expert Rev. Clin. Pharmacol. 9, 157–160. https://doi.org/10.1586/17512433.2015.1102051.

Tsai, I.J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K. L., Tracey, A., Bobes, R.J., Fragoso, G., Sciutto, E., Aslett, M., Beasley, H., Bennett, H.M., Cai, J., Camicia, F., Clark, R., Cucher, M., De Silva, N., Day, T.A., Deplazes, P., Estrada, K., Fernández, C., Holland, P.W.H., Hou, J., Hu, S., Huckvale, T., Hung, S.S., Kamenetzky, L., Keane, J.A., Kiss, F., Koziol, U., Lambert, O., Liu, K., Luo, X., Luo, Y., MacChiaroli, N., Nichol, S., Paps, J., Parkinson, J., Pouchkina-Stantcheva, N., Riddiford, N., Rosenzvit, M., Salinas, G., Wasmuth, J.D., Zamanian, M., Zheng, Y., Cai, X., Soberon, X., Olson, P.D., Laclette, J.P., Brehm, K., Berriman, M., Morett, E., Portillo, T., Jose, M.V., Carrero, J.C., Larralde, C., Morales-Montor, J., Limon-Lason, J., Cevallos, M.A., Gonzalez, V., Ochoa-Leyva, A., Landa, A., Jimenez, L., Valdes, V., 2013. The genomes of four tapeworm species reveal adaptations to parasitism. Nature 496, 57–63. https://doi.org/10.1038/nature12031.

Ule, J., Blencowe, B.J., 2019. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. Mol. Cell 76, 329–345. https://doi.org/10.1016/j.molcel.2019.09.017.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G., 2012. Primer3-new capabilities and interfaces. Nucleic Acids Res. 40, e115.

Vandenberghe, A.E., Meedel, T.H., Hastings, K.E.M., 2001. mRNA 5'-leader trans-splicing in the chordates. Genes Dev. 15, 294–303. https://doi.org/10.1101/gad.865401.

Webb, C., Cabada, M.M., 2017. Intestinal cestodes. Curr. Opin. Infect. Dis. 30, 504–510. https://doi.org/10.1097/QCO.0000000000000400.

Wenze, M., Johnston, C., Müller, B., Pettitt, J., Connolly, B., 2019. Deep evolutionary origin of nematode SL2 trans-splicing revealed by genome-wide analysis of the *Trichinella spiralis* transcriptome. bioRxiv, 1–36.

Wenzel, M.A., Müller, B., Pettitt, J., 2021. SLIDR and SLOPPR: flexible identification of spliced leader trans-splicing and prediction of eukaryotic operons from RNA-Seq data. BMC Bioinform. 22, 140. https://doi.org/10.1186/s12859-021-04009-7.

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. Biometrics Bull. 1, 80–83.

Zuker, M., Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 9, 133–148. https://doi.org/10.1093/nar/9.1.133.

Capítulo 4: Evolución de genes SL-ARN y sus objetivos de splicing en gusanos planos parásitos

Solucionadas las cuestiones de detección y análisis, se dispone de todo lo necesario para abordar cuestiones a escala filogenética: un análisis detallado de la diversidad, conservación y tendencias evolutivas del SL *trans-splicing* en platelmintos. Examinar SL-ARNs, sitios y genes objetivo y como han acompañado la diversificación del grupo. Con este fin se aplicaron las lecciones aprendidas en capítulos previos a 24 especies de los grupos Cestoda y Trematoda. Ambos linajes de parásitos obligatorios con ciclos de vida complejos, y con miembros capaces de transmitirse a humanos. Y crucial para el desarrollo de esta tesis: de las que se dispone públicamente de datos genómicos y transcriptómicos en los que basar la investigación.

Metodológicamente el estudio presentado en este capítulo es una extensión del Capítulo 3, abordando un mayor número de especies y afrontando preguntas a una escala filogenética. Los resultados alcanzados confirman muchos de los reportes previos realizados por otros autores en estudios enfocados en una única especie y los ponen en un contexto filogenético. Revelando no sólo un trasfondo de conservación de genes sometidos a SL *trans-splicing* en platelmintos parásitos, pero también un quiebre en virtualmente cada aspecto del proceso en el grupo Cyclophyllidea. Cambios en la diversidad de SL-ARN funcionales en cada especie, asociación de SLs con elementos transponibles, sus genes blanco y loci policistrónicos conservados. En definitiva mostrando que el SL *trans-splicing* es un proceso dinámico a escalas evolutivas.

Mis contribuciones consisten en el diseño de las estrategias de análisis, los scripts informáticos realizados y el análisis de los resultados obtenidos.

Evolution of SL-RNA genes and their splicing targets in parasitic flatworms

Javier Calvelo[1,2,3] Héctor Musto[2], Uriel Koziol[3] and Andrés Iriarte[1]

1. Laboratorio de Biología Computacional, Departamento de Desarrollo Biotecnológico, Instituto de Higiene, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay
2. Unidad de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay.
3. Sección Biología Celular, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay.

Abstract

Spliced Leader (SL) *trans*-splicing is a key step in the processing of many mRNAs in parasitic flatworms (Neodermata), including the processing of polycistronic transcripts into mature monocistronic mRNAs. Despite its importance, efforts for its identification and characterization in this phylum have remained a collection of single species studies with little exploration at a wider phylogenetic context. In this work we present a comprehensive analysis of this process, based on the available genomic and transcriptomic data of 24 cestode and trematode species, including the identification of the SL-RNA sequences of each species and their splicing acceptor transcripts and sites. Our analysis explores the scope and limits of available datasets and hints at large scale evolutionary changes of SL-RNA loci and splice acceptor sites. We identified a main pattern of concerted evolution of SL-RNA loci in most species, as well as divergence of SL-RNA loci in selected species. SL *trans*-splicing can be detected for a limited number of mRNAs in all species (<31%), and there is extensive use of the same splice acceptor sites for *cis*-splicing, especially for monocistronic transcripts. Ancestral SL *trans*-splicing sites can be found in many highly conserved genes throughout the Neodermata, including in putative ancestral operons, but novel acceptor sites for SL *trans*-splicing can also be identified at shorter evolutionary scales. Particular trends were observed in the cestode order Cyclophyllidea, whose genome size reduction is correlated with a partial remodeling of the SL *trans*-splicing landscape of the group: This included an increased SL-RNA diversity within each species, a loss of many ancestral SL *trans*-splicing acceptor sites, as well as novel SL gene targets exclusive for this group.

Introduction

Spliced leader *trans*-splicing is a form of splicing between two RNA molecules, a specialized short RNA named Spliced Leader RNA (SL-RNA) and a pre-mRNA, to form a mature mRNA. In this process the sequence of the mRNA precursor upstream of the splicing site (outron) is replaced by a spliced leader (SL) sequence typically containing a 5´ trimethylguanosine cap. In many eukaryotes that contain operons, SL *trans*-splicing

represents a key molecular mechanism behind the resolution of polycistronic transcripts into individually capped monocistronic mRNAs (Bitar et al., 2013; Douris et al., 2010; Hastings, 2005; Lasda & Blumenthal, 2011). In addition to participating in polycistron resolution, this mechanism can also result in the removal of deleterious sequences located in the 5' UTR (Bitar et al., 2013; Hastings, 2005; Lasda & Blumenthal, 2011) and in particular cases as the source of alternative mRNA isoforms (e.g. Agorio et al., 2003; Brehm et al., 2000; Siegel et al., 2011).

The distribution of SL *trans*-splicing in the eukaryotic phylogenetic tree is dispersed, as this mechanism is absent in many large clades, including vertebrates. Furthermore, the extent of SL *trans*-splicing is variable in different groups in which the mechanism is present. For example, in trypanosomatids almost all coding genes are found in operons and require SL *trans*-splicing for expression (Michaeli, 2011), and in the nematode *Caenorhabditis elegans* SL *trans*-splicing occurs in almost all mRNAs, even though only a fraction of its genes are organized in operons (Allen et al., 2011) (Allen et al., 2011; Bernard et al., 2023). In contrast, in flatworms only a fraction of mRNAs are targeted by *trans*-splicing (estimated to be between 11-47%, in different species, using different methodologies; Boroni et al., 2018; Brehm et al., 2000; Ershov et al., 2019; Protasio et al., 2012; Tsai et al., 2013), including not only genes organized in operons but also many monocistronic genes.

In some phylogenetic groups that have several different SL-RNA genes there is evidence for their specialization. In *C. elegans*, different SL types are specialized for particular targets, commonly referred to as SL-1 (targeted to the first genes in operons and to monocistronic genes) and SL-2 (involved in operon resolution by targeting the downstream genes) (Blumenthal, 2005; Lasda & Blumenthal, 2011). Other examples include the flatworm *Schmidtea mediterranea* in which a particular SL-RNA is enriched in a subpopulation of stem cells (Rossia et al., 2014), and *Parvilucifera sinerae*, a parasitoid dinoflagellate displaying clear segregation in the splicing of different SL sequences depending on the function of the acceptor mRNAs (Alacid et al., 2022).

Despite the crucial roles of SL *trans*-splicing, its evolutionary origin remains uncertain (Brehm et al., 2000; Douris et al., 2010; Krause & Hirsh, 1987; Lidie & Van Dolah, 2007; Marlétaz et al., 2008; Matsuo et al., 2018; Pouchkina-Stantcheva & Tunnacliffe, 2005; Rajkovic et al., 1990; Ross et al., 1995; Sather & Agabian, 1985; Steele et al., 2004; Tessier et al., 1991; Vandenberghe et al., 2001). Its scattered distribution suggests a complex history of independent origin, loss, and/or repeated horizontal transfer (Bitar et

al., 2013; Douris et al., 2010; Hastings, 2005; Krchňáková et al., 2017; Lasda & Blumenthal, 2011). Furthermore, few comparative studies have been performed within phyla, and the evolutionary dynamics of *trans*-splicing gain and loss in eukaryotes are not well understood (Bitar et al., 2013; Douris et al., 2010; Hastings, 2005; M. Wenzel et al., 2020). A significant roadblock in this field is the low conservation of the SL sequence at large phylogenetic scales, which makes SL-RNA identification non-trivial (Bitar et al., 2013; Douris et al., 2010; Hastings, 2005; Lasda et al., 2010). While some efforts have been made to facilitate the identification of novel SL sequences, (Calvelo et al., 2020; Radío et al., 2018; M. A. Wenzel et al., 2021; Yague-sanz & Hermand, 2018) comprehensive studies involving multiple species remain scarce. The primary sequences of SL-RNAs from distant clades do not show any similarity except for small motifs related to their interaction with spliceosomal components. However, their sequence can be partially conserved within the same clade (Bitar et al., 2013), and they share the same overall structure (Bitar et al., 2013; Douris et al., 2010; Hastings, 2005; Lasda et al., 2010): a leader region that is incorporated into the pre-mRNA and an intronic region with an Sm-like motif, both separated by a canonical splicing donor site. In addition, their secondary structure typically displays two hairpins surrounding the Sm-like motif in the intron and has been proposed as a criterion for SL-RNA prediction (M. A. Wenzel et al., 2021), although exceptions have been reported (Alacid et al., 2022; Zhang et al., 2007). Identifying the SL-RNA sequences and the repertoire of mRNAs subjected to this form of *trans*-splicing is key to understanding the importance of this mechanism in different clades (Bitar et al., 2013; Islas-Flores et al., 2021; Nilsson et al., 2010) and provides crucial information to improve their genome annotation (Calvelo et al., 2023; M. Wenzel et al., 2020).

Because of the central role of *trans*-splicing in the expression of many different genes, it has been proposed as a potential drug target against parasites possessing this mechanism, particularly helminths (Pandarakalam et al., 2019). Here, we present for the first time a comprehensive approach to identify the SL-RNA complement and a catalog of *trans*-splicing acceptor genes throughout the parasitic flatworm groups Cestoda and Trematoda. Both groups are endoparasitic platyhelminthes (subphylum Neodermata), and pose significant threats to human and animal health (Giri & Parija, 2012; Mahmud et al., 2017; Webb & Cabada, 2017). Several parasitic flatworm species have been shown to possess SL *trans*-splicing (**Supplementary Table 1**), in some cases with several SL-RNAs present in each species, (Calvelo et al., 2023; Olson et al., 2020; Rossia et al., 2014; Tsai et al., 2013) without clear evidence of SL-RNA specialization (Calvelo et al., 2023; Olson et al., 2020; Tsai et al., 2013). A key feature, however, is that all reported SLs

for the group possess a 3' terminal "ATG" motif (Cheng et al., 2006) that can potentially provide an alternative start codon for translation.

Our study expands on previous works by offering a global view of SL trans-splicing in parasitic flatworms. Highlighting patterns and evolution of SL-RNA gene diversity, SL trans-splicing acceptor sites and their widespread involvement with cis-splicing, as well as the organization of trans-spliced genes into operons. We observed the existence of conserved SL trans-splicing acceptor sites (SL-ACE) and operons across this parasite group, as well as more dynamic lineage specific examples.

Results and Discussion

Candidate SL identification and selection

A total of 24 species of parasitic flatworms were selected based on the availability of a reference genome assembly and RNAseq data (**Table 1**) and processed following the strategy described in **Figure 1**. In short, our first step was to compile known SL-RNAs from the literature, and to identify new SL-RNA candidates with SLFinder (Calvelo et al., 2020).

Species	Genome Bioproject ID	RNAseq Bioproject ID
Echinococcus granulosus	PRJEB121	PRJNA254535
Echinococcus multilocularis	PRJEB122	PRJNA254535
Hymenolepis diminuta	PRJEB30942	PRJNA546293
Hymenolepis microstoma	PRJEB124	PRJNA637398
Mesocestoides corti	PRJEB510	PRJNA433559
Schistocephalus solidus	PRJEB527	PRJNA304161
Sparganum proliferum	PRJEB35374	PRJDB8966
Spirometra erinaceieuropaei	PRJEB1202	PRJDB8967
Taenia asiatica	PRJNA299871	PRJNA299871
Taenia multiceps	PRJNA307624	PRJNA307624
Taenia saginata	PRJNA71493	PRJNA71493
Taenia solium	PRJNA170813	PRJNA328007
Clonorchis sinensis	PRJNA386618	PRJNA386618
Fasciola gigantica	PRJNA230515	PRJNA230515
Fasciola hepatica	PRJEB25283	PRJEB6948
Fasciolopsis buski	PRJNA284521	PRJNA212796
Opisthorchis felineus	PRJNA413383	PRJNA257351
Paragonimus heterotremus	PRJNA284523	PRJNA284523
Paragonimus westermani	PRJNA454344	PRJNA219632
Schistosoma bovis	PRJNA451066	PRJNA491632
Schistosoma haematobium	PRJNA78265	PRJNA491632
Schistosoma japonicum	PRJNA520774	PRJNA579703
Schistosoma mansoni	PRJEA36577	PRJNA361136
Trichobilharzia regenti	PRJEB4662	PRJNA292737

Table 1: Bioproject IDs for the genome and RNAseq data utilized in this work.

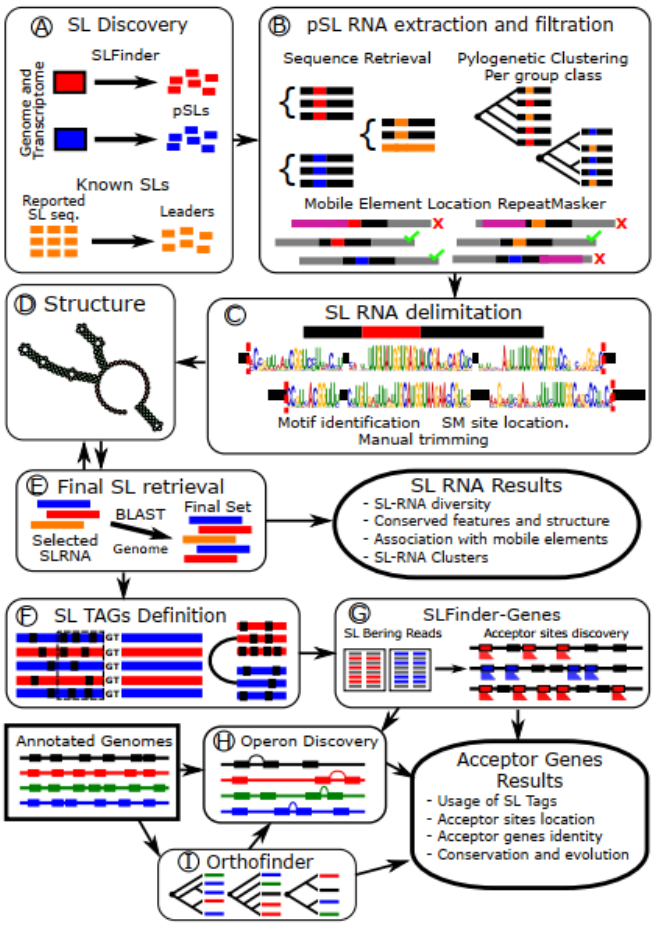

Figure 1: Summary of the analysis pipeline used in this work. A) First, candidate SLs were identified, either *de novo* with the SLFinder pipeline or retrieved from the literature, and hits for their leader region identified in the genome with BLAST searches. B) The surrounding sequence of the candidate SL hits in the genome was retrieved and potential SL-RNAs were identified based on their sequence similarity with each other and the reference sequences. Absence of known repeated elements in proximity is needed to consider a potential SL-RNA. C) Conserved motifs were identified and used to isolate the SL-RNA from the surrounding genome sequence. Candidates whose conserved motifs appeared out of order relative to the majority were excluded and the analysis repeated. D) The SL-RNA secondary structure was estimated and the sequence SL-RNAs identified based on conservation, that is, features present in the reference sequences and in others. E) To recover stragglers potentially missed in the initial steps of analysis, an additional BLAST search was conducted against genomes, using the identified SL-RNAs as queries. Hits were added to the pool if their secondary structure was conserved. F and G) SL tags were defined from the selected SL-RNAs and the reference sequences, and used to identify SL-ACEs. H) In parallel, potential operons were identified based on the gene proximity, coding strand and presence of SL-ACEs. I) The orthology relationships between

the annotated genes in the different genomes was determined and potential conserved operons were identified.

The initial combined dataset of candidates obtained from SL-Finder and blast searches consisted of 1179 loci bearing a putative SL sequence. These were then filtered down to 312 SL-RNA loci across all species, guided by four sequential criteria: 1) functional SL-RNAs share enough sequence similarity to be grouped together in a neighbor-joining phylogeny, clustering as a monophyletic groups with known SL-RNA sequences, as defined using TreeCluster, 2) true SL-RNA loci are not closely associated with mobile genetic elements (≤ 500 bp), 3) the mature portion of the SL-RNAs has specific conserved motifs (see below) in a specific order, and 4) novel SL-RNAs share secondary structure elements commonly observed in known SL-RNAs. These filtering procedures were necessary to differentiate true SL-RNAs from loci, other than the SL-RNA genes, that share the SL sequence, resulting from reverse transcription of *trans*-spliced transcripts followed by genetic integration into the genome (Barnes et al., 2019; Jaeckisch et al., 2011; Slamovits & Keeling, 2008). Lastly, a BLAST search was conducted between the selected SL-RNAs and the genomes to search for additional SL-RNA loci that may have been missed in the initial search, followed by further filtering based on secondary structure analysis.

In particular, a large proportion of the potential SL-RNA loci that were discarded in our dataset were associated with repeated elements (597 loci), primarily with LTR and LINE elements, which are the most common mobile genetic elements found in Platyhelminthes (Coghlan et al., 2019). The only identifiable SL-RNA feature in these loci were the partial leader sequences. These loci may be the result of genomic integration of retrotransposons whose transcripts had received the spliced leader sequence by *trans*-splicing. We hypothesize that the SL *trans*-splicing of retrotransposon transcripts may play a role in containing the propagation of these mobile genetic elements, as SL *trans*-splicing of their RNA intermediates could prevent their integration to the genome or hinder future transposition events by the removal of key regions. It is noteworthy that cestodes of the order cyclophyllidea (including *Echinococcus*, *Taenia* and *Hymenolepis*) largely lacked this type of loci, and at the same time are characterized by a reduction in genome size and mobile genetic element content (Coghlan et al., 2019).

SL-RNA sequence and structural conservation

A total of 71 non-redundant SL-RNA sequences were identified among the 312 SL-RNA loci in different flatworm species (**Supplementary File 1** and **Supplementary Table 2;** see **Supplementary Table 3** for the full registry of potential loci identified in the analyzed genomes), and an additional 4 unique sequences were found among previously reported full SL-RNAs. Two conserved regions are present among these sequences, separated by a highly variable region. The first region comprises the characteristic "AUG" motif found at the 3' end of the leader sequences of flatworms, followed by a canonical splice donor site "GU". This motif was widely conserved, with a few exceptions: in two cases, Unique_SL-1 (*C. sinensis*) and -48 (*S. bovis*), the AUG motif is not conserved, and in three cases, Unique_SL-18 (*S. bovis*), -20 (*S. haematobium*), and -37 (*F. hepatica*), the splice donor site is not conserved. These latter examples are unlikely to constitute functional SL-RNAs. In these cases the transcriptomic evidence of *trans*-splicing was negligible (see below), and we interpret these loci as possible SL-RNA pseudogenes. The second conserved region corresponds to the Sm-like motif near the terminal end of the SL-RNAs. The most conserved part of this motif is the 3' portion, a TVTTTGG 3' sequence, while the 5' end is considerably more variable, as we previously described in *H. microstoma* (Calvelo et al., 2023).

The predicted secondary structure of SL-RNAs from cestodes and trematodes can be distinguished into two large groups: a predominant arrangement with two hairpins upstream the Sm-like site, that was previously observed in *E. multilocularis* (Brehm et al., 2000), *F. hepatica* (Davis et al., 1994) and *Opisthorchis felineus* (Ershov et al., 2019); or a single large hairpin **(Figure 2),** initially reported in *S. mansoni* (Rajkovic et al., 1990). The latter was only predicted in schistosomatid species and is considerably more unstable, with an average Maximum Expected Accuracy (MEA) of -15.54 kcal/mol vs. -30.67 kcal/mol **(Supplementary Table 2)**. Full structure predictions are provided on **Supplementary File 2**. Regardless, in both types of structures there are two conserved features: a) the splice donor site is located at a comparable position on the first hairpin, typically associated with an internal loop or mismatch, and b) several SL RNAs displayed another hairpin immediately downstream the Sm-like motif. Both consistent with previous observations on several flatworms (Brehm et al., 2000; Davis et al., 1994; Ershov et al., 2019; Rajkovic et al., 1990). The latter was not recovered consistently in our analysis, but this is likely due to limitations in the delimitation of the 3' end of the SL-RNA transcript.

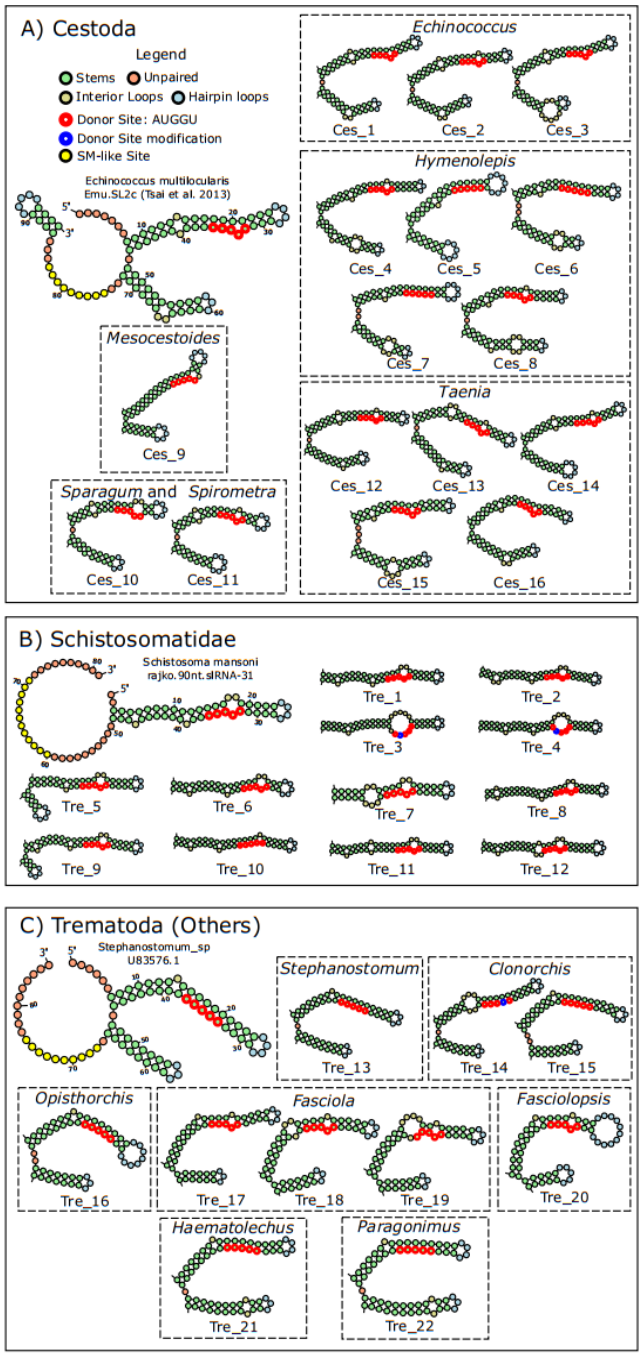

Figure 2: Secondary structural elements in the SL-RNAs of parasitic flatworms A) SL-RNAs from cestodes, B) SL-RNAs from trematodes of the family Schistosomatidae and c) SL-RNAs from other trematodes.

Twelve SL-RNAs had atypical predicted secondary structures but were included in the analysis because they were highly similar to known SL-RNAs when both sequences were trimmed at the same positions (see details in **Supplementary File 3**). The most common elements in these atypical structures include: 1) self-annealing within the leader region,

forming an extra hairpin, and 2) annealing of the 3' portion (that forms the terminal hairpin) with the second hairpin or with the leader region. Trimming tests showed that these outlier structures are very sensitive to the predictions of the 5' and 3' ends of the SL-RNA boundaries, down to the inclusion or removal of a single base (see details in **Supplementary File 4**), suggesting that the typical structure found in other SL-RNAs may also be present in these outliers.

The Neighbor Joining phylogeny of the SL-RNAs mirrors the split in SL-RNA structure within trematode species, with a clear division between Schistosomatidae and the other trematodes (**Figure 3**). Furthermore, the tree topology roughly follows the species tree (Coghlan et al., 2019), as paralog sequences from each species and genus typically show greater similarity to each other than to the SL-RNA of other species. The observed branching pattern of paralog sequences suggests that SL-RNA loci undergo a constant process of expansion and replacement, or that sequence similarity between different paralogs in each species is maintained by concerted evolution. The latter process is more likely to occur as a result of unequal crossing-over or gene conversion in tandem copies of genes, an arrangement found for SL-RNA genes in many eukaryotes (Hastings, 2005), including some flatworms (Rajkovic et al., 1990; Vandenberghe et al., 2001).

SL-RNA tandem repeats

As mentioned, SL-RNA are often clustered as tandem repeats. Here, SL-RNA tandem repeats (defined as SL-RNA loci encoded on the same strand and separated by fewer than 5000 bases), were identified for species *E. multilocularis*, *F. gigantica*, *F. hepatica*, *O. felineus*, *S. bovis*, *S. haematobium*, *S. japonicum*, *S. mansoni*, and *T. multiceps* (**Supplementary Table 4**). Searching for shared homologous genes in a 100 kb radius it was possible to identify synteny for tandem repeats between *E. multilocularis* and *T. multiceps* (SLClus-7 with SLClus-8), tentatively between two clusters within *T. multiceps* (SLClus-8 with SLClus-9), and between the schistosomatid species *S. bovis* (SLClus-14), *S. haematobium* (SLClus-15) and *S. mansoni* (SLClus-17 and -18) (**Figure 4)**.

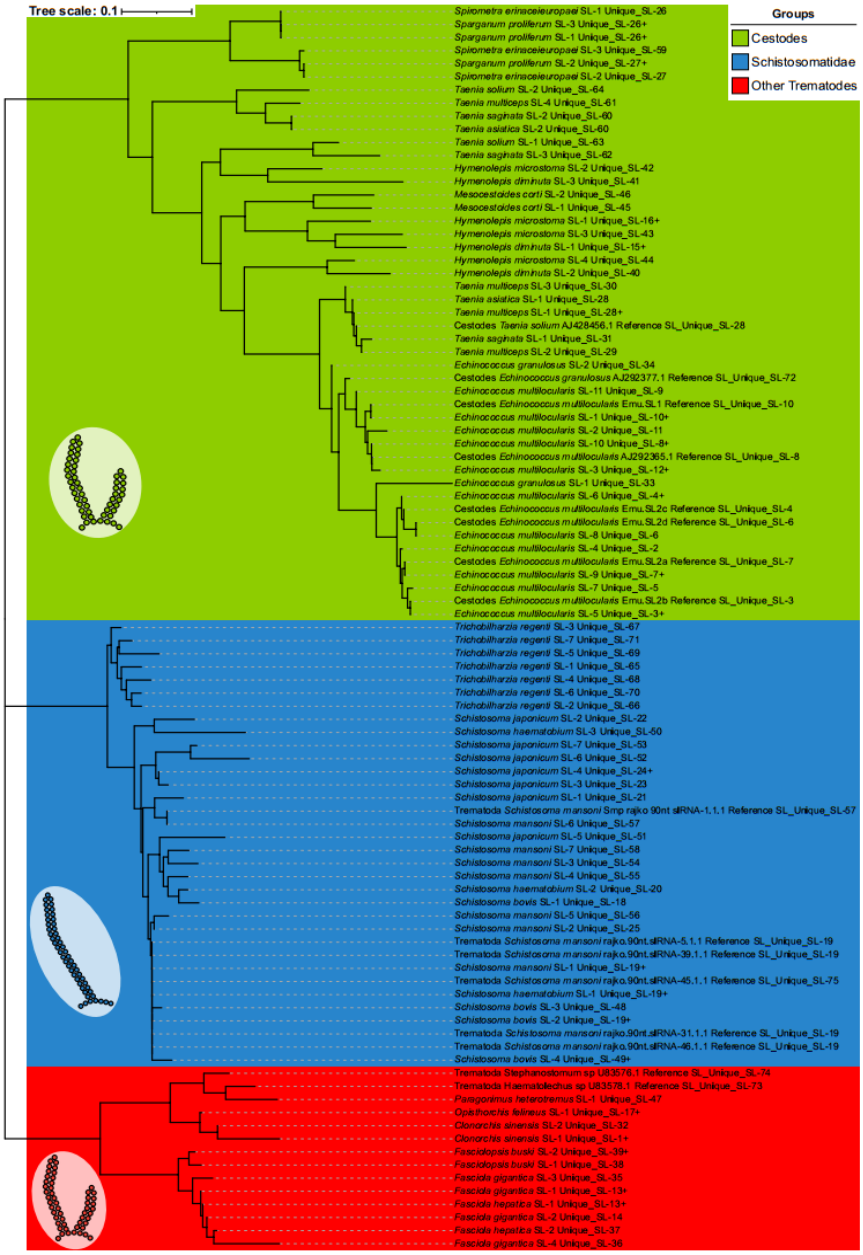

Figure 3: Neighbor Joining phylogeny of flatworm SL-RNAs. Previously described sequences are indicated as “Reference SL”. Three groups are highlighted with a schematic representation of their structure: Cestode SL-RNAs (green), schistosomatid SL-RNAs (blue) and SL-RNAs from other trematodes (red). Unique SL-RNA sequences with more than one locus in the species are indicated with a “+”.

The ability to identify these types of clusters is significantly impacted by the quality of the genome assembly, as demonstrated by *S. bovis* and *S. haematobium*, which have 2 SL-RNAs each, compared to *S. mansoni* with 111 SL-RNAs. However, species like *Hymenolepis microstoma*, with a high-quality assembled genome encoding only 5 non-clustered SL-RNAs, suggest that the presence of these clusters is not universally found in Platyhelminthes. The shared clusters identified on *E. multilocularis* (SLClus-7) and *T. multiceps* (SLClus-8) are noteworthy. Their surrounding orthologous genes show a conserved synteny, but the SL-RNA genes within each species are more similar to each other than to those of the other species (sequences Unique_SL-8, 10 and 12 from *E. multilocularis* and Unique_SL-28, 29 and 30 from *T. multiceps*, see **Figure 2**). These results support the existence of concerted evolution of tandem SL-RNA loci in these species. An interesting observation is that the orthologous SL-RNA clusters identified are largely homogeneous internally (they encode mostly the same SL-RNA), but these sequences are different between species, thus suggesting that after the speciation event the SL-RNA loci were homogenized by gene conversion.

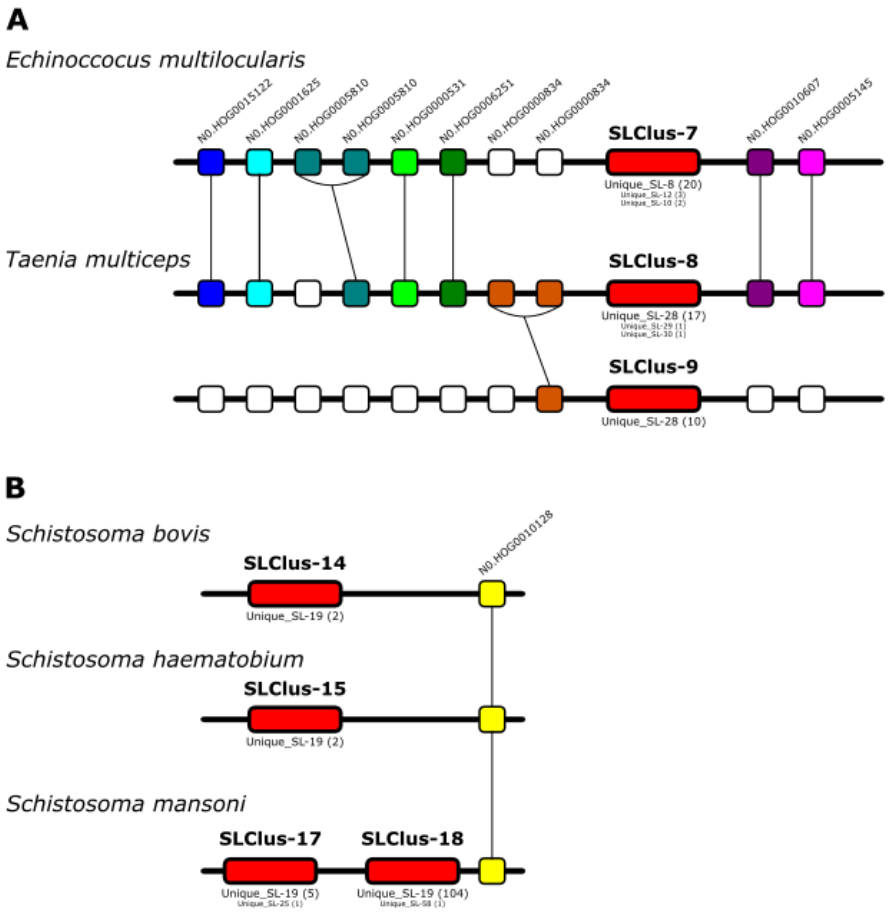

Figure 4: Shared Phylogenetic Hierarchical Orthogroups (HOGs) IDs found in a 100 kb radius around syntenic SL-RNA tandem repeats.

Because of the importance of high-quality genome assemblies for the identification of SL-RNA gene clusters, we searched the location of the SL-RNAs described in this work on a new *E. granulosus* genome assembly (ID: ASM2155672v1) generated with long read data by Korhonen et al. (2022) In contrast with assembly ASM52419v1 generated by Tsai et al. (2013), 93 SL-RNA loci were identified, concentrated in 2 clusters on chromosome 1 and contig8_egr. These clusters comprised mostly sequences similar to Unique_SL-8 from *E. multilocularis* (**Supplementary Table 5)**. The surrounding gene synteny confirms that the SL Cluster on chromosome 1 is orthologous to the SL-RNA gene clusters found in *E. multilocularis* and *T. multiceps* (**Supplementary Table 6**). While similar, the fact that they are not exact copies of Unique_SL-8 is further evidence for the frequency of gene conversion on these genes.

SL trans-splicing Acceptor Sites

With the goal of identifying and analyzing individual SL *trans*-splicing acceptor sites (SL-ACEs), we defined SL tags as the last 15 bases of the SL portion (upstream the donor site) in order to search for reads supporting *trans*-splicing in RNA-Seq data. In total, 30 unique SL tags were identified from the selected SL-RNA loci pool (19 for Cestoda and 11 for Trematoda; note that many SL-RNAs may share the same tag). These SL tags were used to identify SL-ACEs using the SLFinder-Genes pipeline (Calvelo et al., 2023). Although SL-bearing reads in different species and datasets represented only ~0.001-0.1% of the total RNAseq data, as previously found in *S. mansoni* (Boroni et al., 2018) and *H. microstoma* (Calvelo et al., 2023), this still resulted in thousands of reads for most species, revealing several interesting insights. First, SL tag use on each species is largely in line with the observed presence or absence of SL-RNA candidate loci containing the respective tags, with few exceptions. In particular, although we could not find any SL-RNA loci in the highly fragmented *S. solidus* and *P. westermani* genome assemblies, we detected the Cestoda_M and Cestoda_P tags in *S. solidus*, and the Trematoda_H tag in *P. westermani*, indicating the existence of the corresponding loci in their genomes. Cestode transcriptomes display more diverse intraspecific SL-RNA complements compared to trematodes, which are mainly dominated by a single SL tag (**Figure 5** and **Supplementary Table 7**). Second, while basal cestode species (*S. solidus*, *S. proliferum*, and *S. erinaceieuropaei*) share the same two SL tags, taeniid and *Hymenolepis* species concentrate the diversity of SL tags in cestodes, including differences between species within the same genus. It is important to note that differences between SL tags are often a single nucleotide. Therefore, rare SL tags represented by <<1% of the SL bearing reads

for a species, but which were not identified in any SL-RNA loci in that species, may be the result of sequencing errors and not necessarily missing SL-RNA loci.

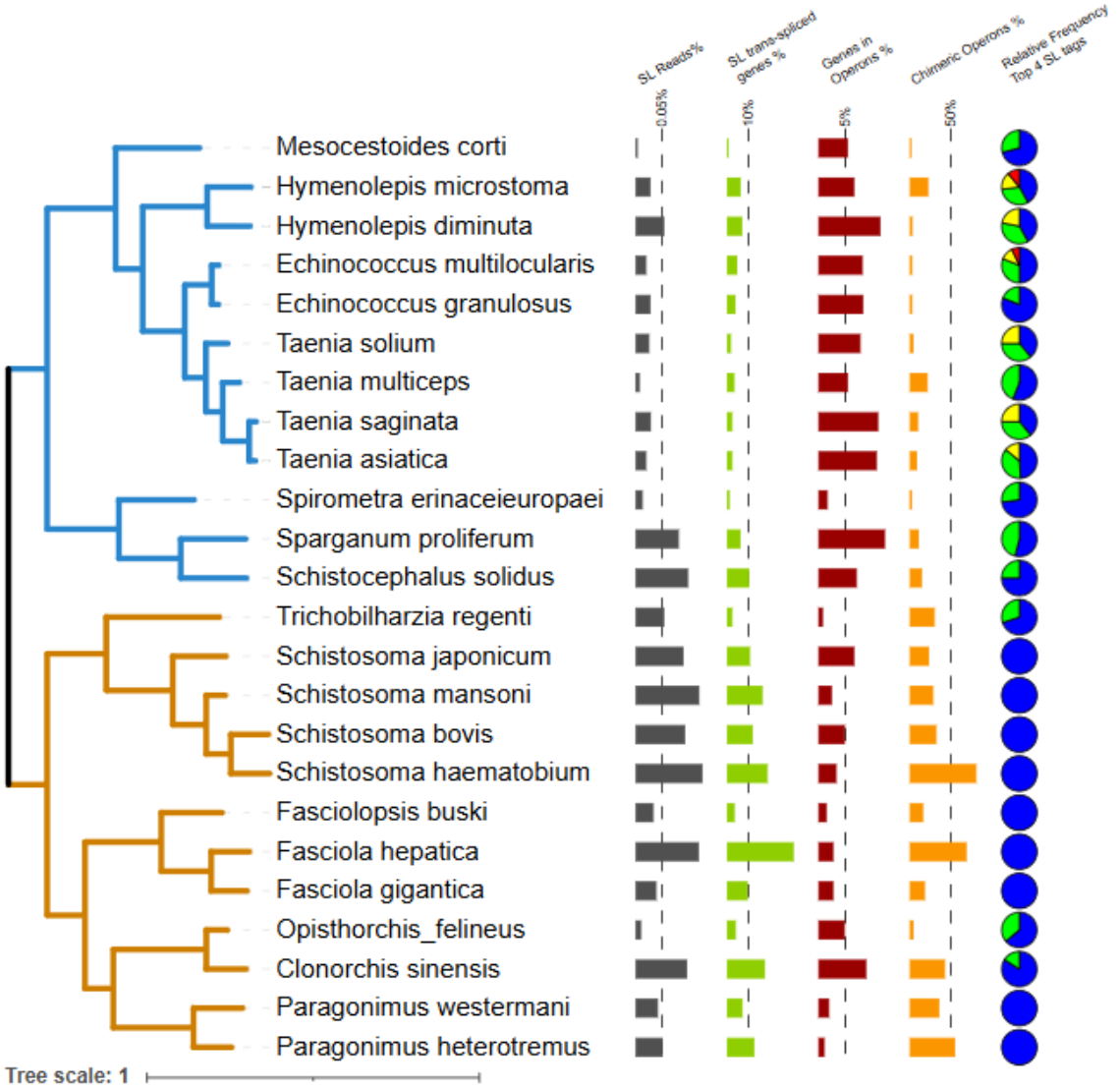

Figure 5: SL trans-splicing across phylogeny of parasitic platyhelminthes (Neodermata). The species tree was estimated by Orthofinder using the STAG method. For each species it is displayed the percentage of SL bearing reads relative to the full RNAseq dataset for the species (max ~0.12%), observed prevalence of SL trans-spliced genes in their genome (max ~30%), genes encoded in potential operons in our study (max ~13%), proportion of operon candidates made of chimeric gene models (max ~80%) and the relative abundance of the top 4 SL tags for the species (from highest to lowest: Blue, Green, Yellow and Red). SL tags with less than 5% SL bearing reads in the species were excluded.

Rare SL tags with less than 0.1% of SL-bearing reads in any species; namely Trematoda_D (corresponding to Unique_SL-25 from *S. mansoni*), Trematoda_I (corresponding to Unique_SL-48 from *S. bovis*) or Trematoda_J (corresponding to Unique_SL-51 from *S. japonicum*); are likely cases of sub-functional or non-functional SL-RNAs, either due to structural mutations and/or very limited expression. The low levels of reads found matching these tags may correspond at least in part to sequencing errors (as tags may differ in only a single base). Therefore, special cases such as Unique_SL-1 (*C. sinensis*, represented by tag Trematoda_E) and Unique_SL-48 (*S. bovis*, represented by tag Trematoda_I), which bear a modified "ATG" are unlikely to be functional. Similarly, it is impossible to confirm or reject *trans*-splicing of SL-RNAs Unique_SL-18 (*S. bovis*) and Unique_SL-20 (*S. haematobium,* both containing a modified donor site "GC"), or Unique_SL-37 (*F. hepatica*, with a "CT" donor site) as they share the same tag as other, more conserved SL-RNAs.

The total SL-ACE counts for each species were correlated with the total number of observed reads (Pearson correlation: 0.89), with cestode species displaying lower overall percentages of SL-bearing read counts than trematodes (**Figure 5**, see **Supplementary Table 8** for a detailed summary). SL-ACEs supported by more than 3 reads were initially classified according to their positions relative to the annotated gene models with which they were associated. Most of them were found upstream of their assigned gene model CDS, or as internal sites. In broad terms this basic classification separates two distinctive types of SL aces: the predominant upstream SL-ACEs in both number and SL bearing reads support, and the less abundant internal sites (**Figure 6**). In addition, a small proportion of SL-ACES were located downstream of the gene model to which they were assigned and/or on the opposite strand, especially in *S. mansoni* (10% of all acceptor sites), *P. heterotremus* (17%), and *F. buski* (19%). It is likely that these result from missing and/or overlapping genes in the genome annotation.

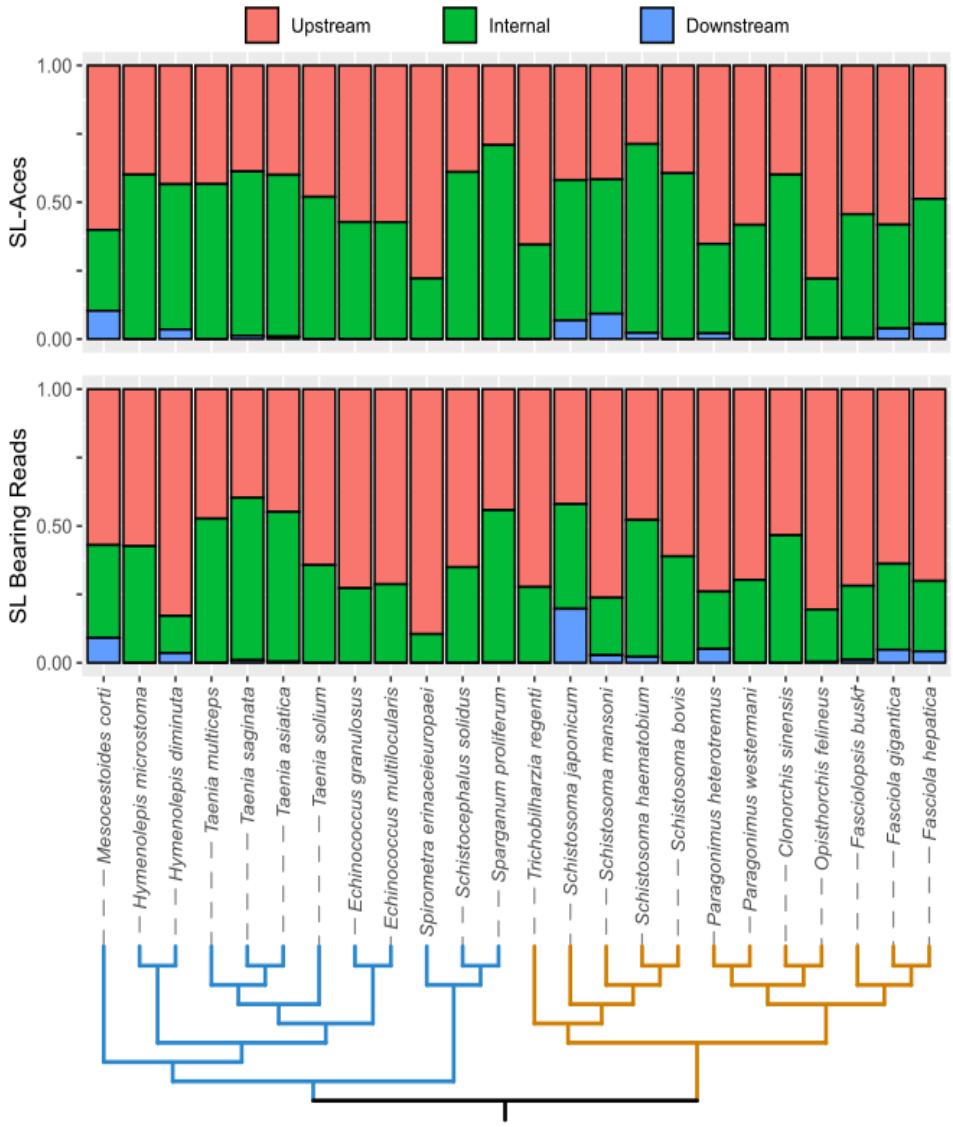

Figure 6: Summary of the SL-ACEs Upstream, Internal and Downstream relative to the CDS of their assigned gene model. For each species it is detailed the relative abundance of the three types of sites and supporting SL Bearing and *cis*-splicing reads observed in those sites. SL-ACEs assigned to more than one gene model, or in the opposite strand were excluded.

To determine the importance of *trans*-splicing in the expression of highly conserved genes (presumably involved in essential processes), we verified how many Benchmarking Universal Single-Copy Orthologs (BUSCO) marker genes contained SL-ACEs in each species. All species had BUSCO marker genes with SL-ACEs, and on average 18% of all BUSCO genes had evidence of *trans*-splicing (the values were similar in cestodes and trematodes; **Supplementary Table 9**). Furthermore, we found 35 genes with *trans*-splicing in *S. mansoni* (with at least one SL-ACE supported by more than 3 reads) among

the list of 195 genes that produced fully penetrant detachment RNAi phenotypes in the screening performed by Wang et al. (2020). These results indicate that affecting *trans*-splicing in parasitic flatworms could interfere with numerous essential processes.

Improvement of the annotation of polycistronic genes based on SL-ACE locations.
As observed in our previous work (Calvelo et al., 2023), because of their short intergenic distance with high RNA-Seq coverage, polycistronic genes can be mis-annotated and fused into single gene models in the currently available genome annotations. To identify these cases and properly discriminate between different genes, we selected four species from each group to be used as references, selecting the best assembly quality (N50 values) for their genus: *E. multilocularis*, *H. microstoma*, *S. proliferum*, and *Taenia multiceps* for Cestoda; and *C. sinensis*, *F. hepatica*, *O. felineus*, and *S. mansoni* for Trematoda. First, to identify potential chimeras, genes were divided into two parts by each internal SL-ACE, and hits on the proteins of the reference species were identified by blastx searches. Cases where each half had a different set of hits in one of the references were pre-selected, and the BLAST search was repeated with the full gene. A gene model was only considered chimeric if on all cases the SL-ACE separated both sets of hits, with a maximum overlap of 30 bases. When multiple isoforms were reported for the same gene, we selected the longest one with clear evidence of chimerism for further analysis.

The initial assessment identified that a significant portion of the internal SL-ACEs could correspond to *trans*-splicing in the 5' region of genes in operons that have been mis-annotated and fused into single gene models, with significant variance between the considered reference species. The range varies widely, from as low as 4% when comparing *E. granulosus* with *E. multilocularis*, to over 70% when comparing *T. saginata* with *E. multilocularis* (**Supplementary Tables 10 and 11**). These differences likely result from methodological decisions during the annotation pipeline, particularly the transfer of annotation between genomes, as is the case between *E. granulosus* and *E. multilocularis* (Tsai et al., 2013). Nevertheless, it is essential to note that only a fraction of all candidates found were ultimately confirmed as chimeras in the second stage (45% in average in cestodes, and 33% in trematodes), indicating that the divide and BLAST approach by itself is susceptible to false positive errors.

Extensive sharing of acceptor sites between cis and trans-splicing in parasitic flatworms
In other eukaryotes, the SL *trans*-splicing process is known to utilize much of the same machinery as *cis*-splicing (Hastings, 2005; Lasda & Blumenthal, 2011), and some competition between both processes has been described for splicing acceptor sites in

other eukaryotes (Allen et al., 2011), resulting in low levels of *trans*-splicing at a minority of internal *cis*-splicing sites.

Within flatworms, coexistence of the two forms of splicing over the same acceptor sites was previously reported for *S. mansoni* (Boroni et al., 2018); and in *H. microstoma* we found several cis-splicing donor sites within the 5' UTR, where the impact to the protein coding sequence is mitigated or non-existent (Calvelo et al., 2023). In particular, we observed that many monocistronic genes with *trans*-splicing at the beginning of the transcript also have introns interrupting the 5´UTR region, and the 3´splicing acceptor site was the same for *cis* and *trans*-splicing. This result suggested that most *trans*-splicing of monocistronic genes occurred as a result of transcription from internal promoters within the 5´ UTR of these genes, resulting in strong splicing acceptor sites in the 5´UTR without any associated splicing donor sites (*i.e*. outrons).

To explore how widespread this phenomenon is across parasitic flatworms, we measured *cis*-splicing reads over three categories of SL-ACEs according to their location: 1) upstream the coding region of a monocistronic gene or of the first gene of an operon ("Upstream Mono"), 2) upstream the coding region of a downstream gene in an operon (and thus related to polycistron resolution; "Upstream Poly"), and 3) within the coding region of their gene (Internal). Most species exhibit a high degree of co-existence of SL *trans*-splicing and *cis*-splicing at the same acceptor sites; the proportion of all SL *trans*-splicing acceptor sites showing any evidence of *cis*-splicing ranged from 20% in *Spirometra erinaceieuropaei* up to 95% in *S. mansoni* (the average was 66% among all species; **Supplementary Table 12**). This proportion decreased moderately when we required three or more reads supporting *cis*-splicing for inclusion (range from 9.5% to 91%, average 52% among all species). These huge differences likely correspond to a combination of data quality, and different annotation biases.

In almost all analyzed species, SL-ACEs classified as Upstream Poly had the lowest levels of *cis*-splicing, followed by Upstream Mono SL-ACEs with intermediate levels of *cis*-splicing (and with a high dispersion), whereas Internal SL-ACEs had the highest levels of *cis*-splicing (**Figure 7** and **Supplementary Table 13).** This pattern indicates that the existence of *cis*-splicing in SL-ACEs related to polycistron resolution is strongly prevented, as may be expected, given their importance to preserve the reading frame of each cistron. The low levels of *cis*-splicing in SL-ACEs related to polycistron resolution may result from the existence of strong motifs in the outrons that recruit the SL-RNA and associated trans-

splicing proteins (SL-RNP) (Denker et al., 2002), and more generally from the absence of available donor splicing sites in the vicinity of these SL-ACEs. The greater level of *cis*-splicing observed in Upstream Mono SL-ACEs could originate from the presence of internal promoters within 5´UTR introns, as we previously proposed in *H. microstoma*, which would result in the generation of a subset of pre-mRNA with strong outrons (with a strong splicing acceptor site available for *trans*-splicing in the absence of competing *cis*-splicing donor sites). Unfortunately, the absence of high-quality 5´UTR annotation data prevented us from further exploring this hypothesis. On the other hand, alternating between *cis*-splicing and *trans*-splicing within the 5´UTR of these genes would be less likely to result in non-functional transcripts, which could lead to a relaxation of the selective pressure on SL-RNP recruitment. Finally, the Internal SL-ACEs displayed the highest levels of *cis*-splicing, indicating that many of these are minor *trans*-splicing sites which may correspond to biological “noise”, equivalent to the levels of erroneous *cis*-splicing that result in non-functional transcripts in other species (Melamud & Moult, 2009; Wan & Larson, 2018). Nevertheless, the existence of strongly SL *trans*-spliced internal sites on some genes observed in most species does support a putative role on alternative isoform formation as it has been proposed for some specific cases (i.e: Agorio et al., 2003; Boroni et al., 2018). On an annotation quality note, we observed radically different behaviors of the SL-ACEs assigned to chimeric gene models between species (data not shown). We speculate that they are caused by a different proportion of Upstream Mono, Upstream Poly and Internal sites among these SL-ACEs. Overall, the extensive sharing of acceptor sites between *cis* and *trans*-splicing (especially upstream of monocistronic transcripts), in combination with a relatively small complement of *trans*-spliced genes, are very different from the patterns of *trans*-splicing observed in the best studied nematode models such as *C. elegans*, for which such widespread coexistence of *cis* and *trans* splicing processes on the same acceptor sites has not been described.

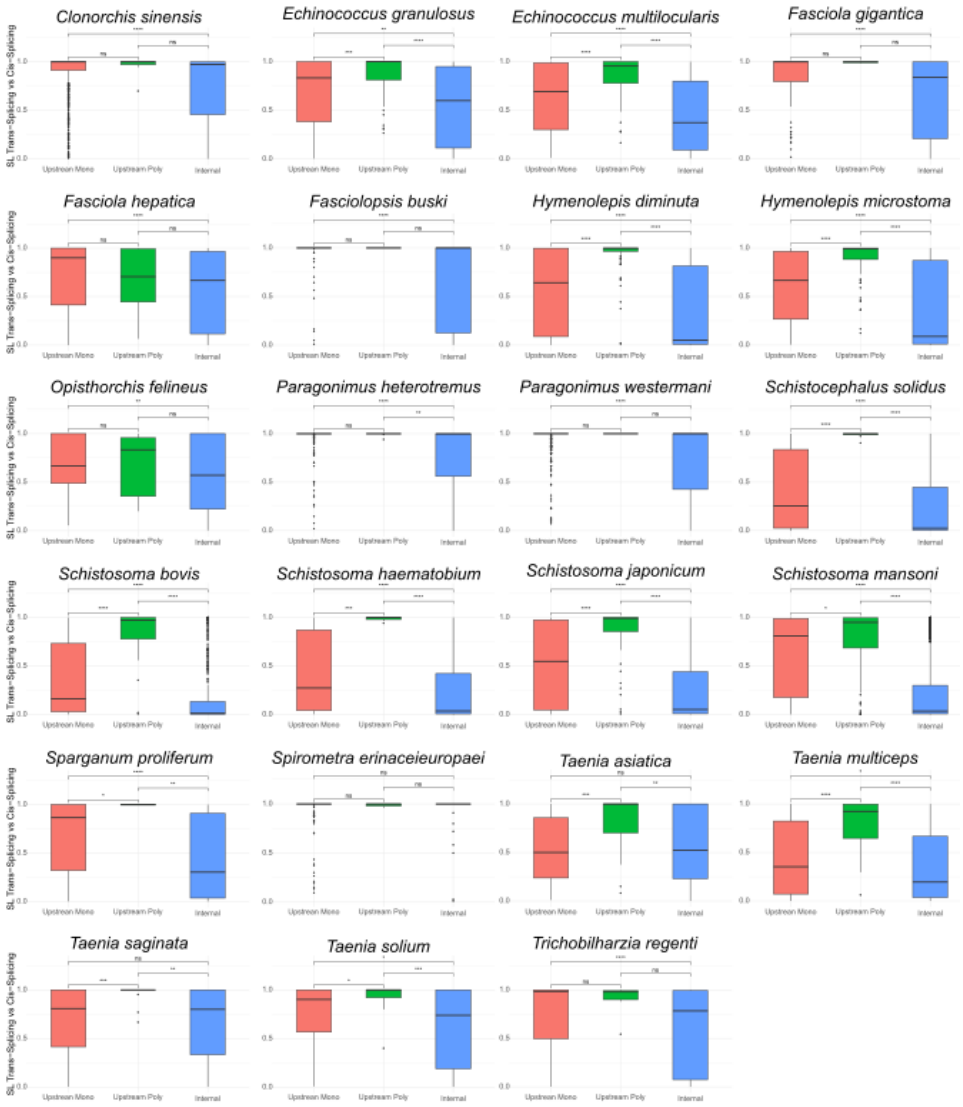

Figure 7: Proportion of SL bearing reads of all splicing reads observed on different sets of SL-ACEs: 1) Upstream Mono: upstream the coding region of a monocistronic gene or of the first gene of an operon. 2) Upstream Poly: upstream the coding region of a downstream gene in an operon, and thus related to polycistron resolution. And 3) Internal: the SL-ACEs are located within the coding sequence. SL-ACEs assigned to a chimeric gene model were excluded, along with *Mesocestoides corti* due to limited data. Pairwise comparisons were carried out with the Wilcoxon Rank Sum and Signed Rank Tests, significant levels of the pairwise comparisons are displayed with the following code: ns (p > 0.05), * (p <= 0.05), ** (p <= 0.01), *** (p <= 0.001), and **** (: p <= 0.0001)

Lack of target specialization of SL-RNAs in parasitic flatworms

For species with two or more SL tags, we searched for evidence of SL-RNA specialization at particular SL-ACEs relative to the overall usage across all samples. Chi-square tests were conducted on highly supported SL-ACEs (with at least 5 observed and expected read counts for all SL tags). Consistent with our previous analysis in *H. microstoma* (Calvelo et al., 2023), very few sites exhibited statistically significant biases in their use of different SL tags relative to the overall average, and even in these cases the biases were very small (summary in **Supplementary Table 14**; full details in **Supplementary Table 15**). Furthermore, we explored whether any SL-RNAs may be specialized for operon resolution, as shown for SL2 in nematodes. When comparing all Upstream Mono SL-ACEs, versus Upstream Poly SL-ACEs, only some species had small biases in their SL tags with significant results under a Wilcoxon rank sum test. However, their biological relevance is questionable given the slight differences in their median proportions (**Supplementary Table 16**). In summary, we did not detect any clear evidence of SL-RNA specialization among parasitic flatworms.

Conservation of SL trans-splicing acceptor genes in parasitic flatworms

The species of parasitic flatworms included in our analysis includes many distant branches of parasitic flatworms, providing us with the opportunity to study the conservation of *trans*-splicing target genes across these species. The conservation of SL *trans*-splicing target genes between different species was determined based on their orthology. For this purpose, the longest isoform reported for each species gene, after the chimeras were identified and excluding gene models with undetermined amino acids, was processed using Orthofinder (Emms & Kelly, 2019). In total, 338,481 sequences were sorted into 28,636 Phylogenetic Hierarchical Orthogroups (HOGs) (**Supplementary Table 17**). The species tree estimated using the STAG method (Emms & Kelly, 2018) aligns with other phylogenetic studies (Coghlan et al., 2019).

A HOG was considered to have conserved SL trans-splicing between two species if both had member genes identified with SL-ACEs. Conservation values were relatively low, with two species sharing on average 11% of their HOGs subjected to SL *trans*-splicing. This probably reflects both true gain-and-loss of *trans*-splicing for different HOGs in different lineages, as well as partial discovery of *trans*-splicing acceptor sites in most species. Nevertheless, there are prominent clusters visible in a heatmap of the proportion of shared *trans*-spliced HOGs relative to the total number of *trans*-spliced HOGs from both species (“Symmetric Heatmap”, **Figure 8A**). These clusters are largely concordant with the phylogeny of the species, indicating that there is a strong phylogenetic signal in regard to the genes subjected to SL *trans*-splicing. The division between cyclophyllidean cestodes

and trematodes is clearly visible, but basal cestodes (the diphyllobothriideans *Sparganum proliferum* and *Schistocephalus solidus*) show higher similarity to the trematode species. An additional heatmap showing the proportion of shared *trans*-spliced HOGs relative to the total number of *trans*-spliced HOGs from each species ("Asymmetric heatmap", **Figure 8B**) reveals that although a large proportion of cyclophyllidean *trans*-spliced HOGs are also *trans*-spliced in trematodes and basal cestodes, the reverse is not true. This occurs even for cyclophyllidean species with high RNA-Seq coverage (e.g. *Echinococcus* and *Hymenolepis* spp.) indicating that it is not an artifact stemming from low rates of discovery in cyclophyllidean species. It is likely that this pattern results from a real differential inheritance of ancestral SL *trans*-spliced genes in both groups. An examination of the mutually *trans*-spliced HOGs between flatworm groups (Cyclophyllidea, Diphyllobothriidea, Schistosomatidae and Other Trematoda) confirms that diphyllobothriideans share more *trans*-spliced HOGs with Trematodes than with Cyclophyllidea, despite being more closely related to the latter.

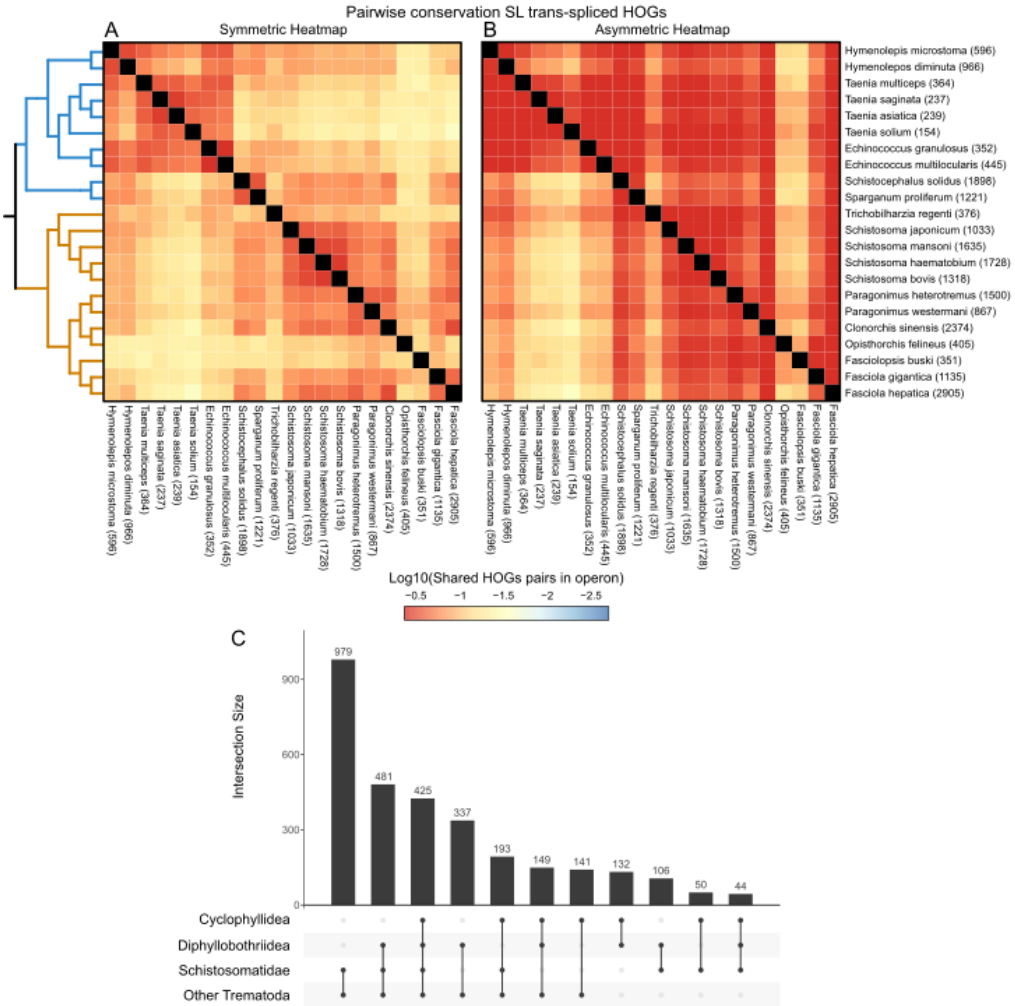

Figure 8: Conservation of SL trans-splicing acceptor genes in parasitic flatworms. Heatmaps showing the pairwise proportion of HOGs subjected to SL trans-splicing in logarithmic scale. The scale goes from -2.7 (~0.002%) to -0.37 (~42.658%). The number of HOGs considered per species is displayed on the row's IDs. A) Symmetrical heatmap displaying the proportion of shared trans-spliced HOGs between the species in the row and column, relative to the total number of *trans*-spliced HOGs in these species. B) Asymmetrical heatmap displaying the proportion of shared *trans*-spliced HOGs between the species in the row and column, relative to the number of *trans*-spliced HOGs in the species in the row alone. The species *Mesocestoides corti* and *Spirometra erinaceieuropaei* were excluded due to very low numbers of detected SL-ACEs. Both heatmaps are based on the similarity between each species dataset. C) Upset plot of the shared trans-spliced HOGs between the groups Cyclophyllidea, Diphyllobothriidea, Schistosomatidae and Other Trematoda. HOGs were included if they were identified as *trans*-spliced in at least one species of two or more groups.

Conservation of operons in flatworms

Genes putatively encoded in operons were defined as two or more genes encoded on the same strand, separated by no more than 300 bases, as well as those gene models identified as chimeric (see above). Genes with nearly complete overlaps with other genes (above 90%) were excluded. Cestode genomes exhibit more than double the number of operon candidates compared to Trematodes, with a median of 448.5 versus 182, respectively (**Supplementary Table 18**, full details in **Supplementary Table 19**). This difference is largely driven by the Cyclophyllidean species and likely reflects both assembly quality and their smaller genomes. Notably, the number of operons candidates that originated from chimeric gene models varied greatly between different species, from 1% to 21% in cestodes, and from 3 to 81% in trematodes (**Figure 5**).

Similar to the conservation of SL *trans*-spliced genes, conserved operons were defined based on the homology of their member genes: pairs of HOGs found within a putative operon in different species, with the added requirement of evidence of SL *trans*-splicing in the second HOG of the pair in both species. The similarity of the operon complement of cyclophyllidean species is visible in the data, while clustering in Trematoda is limited to Schistosomatidae (**Figure 9A**). Furthermore, unlike HOGs subjected to SL *trans*-splicing (**Figure 8B**), the asymmetric heatmap does not reveal as much similarity between Cyclophyllidea and other platyhelminthes (**Figure 9B**), and there are no clear stronger similarities between the considered flatworm lineages beyond a slight enrichment between Schistosomatids and Other Trematoda (**Figure 9C**). This suggests that while there is a

set of HOGs whose SL *trans*-splicing is conserved across parasitic flatworms, their organization into specific operons is not widely conserved. Furthermore, the high conservation of operons in cyclophyllidean species correlates with the massive size reduction of their genomes (Coghlan et al., 2019). This genome size reduction was associated with the reduction of intergenic regions, which may have resulted in the reorganization of many genes into operons.

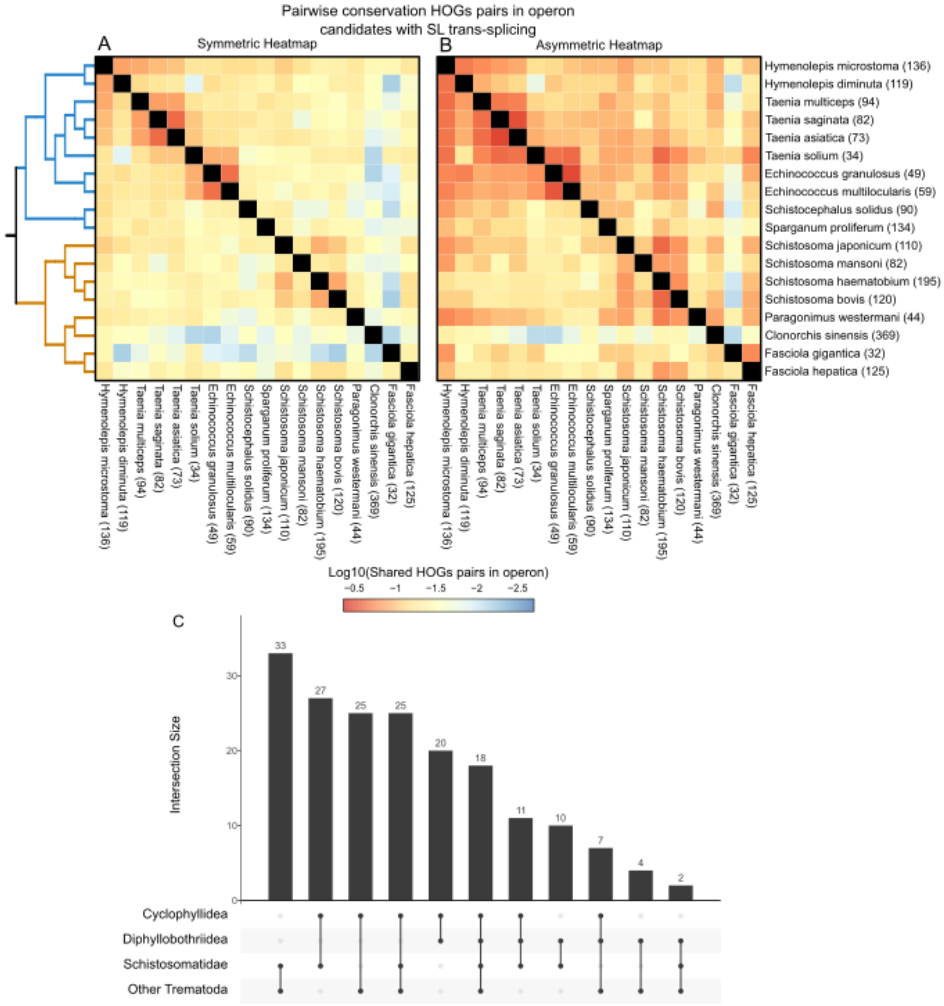

Figure 9: Conservation of operons in parasitic flatworms. Heatmaps showing the pairwise proportion (in logarithmic scale) of HOG pairs in operon candidates with evidence of SL *trans*-splicing. The scale goes from -2.7 (~0.002%) to -0.37 (~42.658%). The number of pairs of HOGs considered per species is displayed on the row's IDs. A) Symmetrical heatmap displaying the proportion of shared HOG pairs between the species in the row and column, relative to the total number of HOG pairs in these species. B) Asymmetrical heatmap displaying the proportion of HOG pairs on the species in the row that are shared with the species on the column, relative to the HOG pairs in the species in the row alone. Species with fewer than 30 HOG pairs for comparison were excluded. Both heatmaps are based on the similarity between each species dataset. C) Upset plot of the shared *trans*-

spliced HOG pairs between the groups Cyclophyllidea, Diphyllobothriidea, Schistosomatida and Other Trematoda. HOG pairs were included if they were identified as *trans*-spliced in at least one species of two or more groups.

Analysis of particular cases of gain and loss of trans-splicing and polycistronic loci in parasitic flatworms

In order to detect particular cases of evolutionary gain and loss of *trans*-splicing, we first focused on a subset of HOGs which: 1) had evidence of SL *trans*-splicing in at least 5 species. 2) included representatives of at least 2 of the reference species used to identify chimeric gene models, and 3) at least one of the genes from each represented species had evidence of SL *trans*-splicing or had relatively high expression levels (equal or higher than the median TPM values of genes with confirmed SL trans-splicing in each species). For these highly expressed genes, the lack of detection of *trans*-splicing is less likely to be a false negative.

In total, 41 HOGs were selected for manual inspection (**Supplementary Table 20** and **Supplementary File 5**). When inference was possible, most analyzed HOGs indicated that SL *trans*-splicing is the ancestral condition for parasitic flatworms (including cases in which *trans*-splicing was detected in almost all species), followed by *trans*-splicing gain within trematodes. These patterns indicate the maintenance of *trans*-splicing in several genes throughout the evolution of neodermatan flatworms (corresponding to a divergence time of at least 259 million years based on the fossil record, which is probably much older (Baets & Littlewood, 2015; Laumer et al., 2015). Interestingly, several examples including HOGs N0.HOG0002707, N0.HOG0007613, N0.HOG0009887, N0.HOG0010318, N0.HOG0010643 and N0.HOG0010915 suggest a loss or reduction of SL *trans*-splicing in the group Cyclophyllidea, consistent with the results of pairwise comparisons shown in **Figure 8**.

In order to identify gains and losses of operons during the evolution of parasitic flatworms, we paid special attention to HOG pairs that were well conserved operons that were exclusive to Cestodes or Trematodes (observed in at least 4 genera of cestodes but not in trematodes), or vice versa; or associated with one of the selected HOGs for their expression levels and evidence of SL *trans*-splicing (discussed above) and had a manageable number of representative genes per species. In total 26 putative operons were selected (**Supplementary Table 21 and Supplementary File 6**). In addition to the direct loss of one or more HOGs in the operon, differences between Cestodes and

Trematodes include changes in synteny (see for examples: Op3, Op8 and Op22) and an increase of the intergenic distances that often exceed the 300 bases set on this study (see for example Op2 and Op5). However, it should be noted that the 300 bases threshold distance is likely to be very conservative, in light of the chimeric gene models found. For example in the case of conserved operon Op18, the HOG pair N0.HOG0008418__N0.HOG0010248 has closely related species for which the pair was reported either as a chimeric gene model or as separate genes with intergenic distance ranging from ~1500 to over 8000 bases, along with strong evidence of SL trans-splicing over N0.HOG0010248 representatives in both cases. Likewise, Op6 shows the same pattern within *Schistosoma* spp. where *S. bovis* and *S. mansoni* have an intergenic distance of 2412 and 4295 bases respectively, but their orthologues on *S. haematobium* and *S. japonicum* were reported as chimeric gene models.

Another driving mechanism of the divergence from Cestodes and Trematodes is the replacement of members of the operon. In 4 of the 10 cases observed however the alternative HOGs are related to each other by sequence similarity and derivative of the same original Orthogroup estimated by Orthofinder, indicating divergence between groups. These are cases where the sequence is conserved enough to be identified as orthologues, but divergent enough to be assigned to different HOGs.

It is noteworthy that among this limited sample there are 5 potential species-specific gains of SL trans-splicing through genetic rearrangements. First there are Op5, Op12 and Op24 have potential gains of SL trans-splicing through the inclusion of a novel first gene in the operon on *C. sinensis*; always involving a highly expanded HOG on this species (N0.HOG0000286, N0.HOG0000285 and N0.HOG0000648 respectively). Second Op7 and Op11 feature reductions of the intergenic space between HOGs on Hymenolepis sp. that led to the formation of novel operon candidates and the gain of SL *trans*-splicing on the new downstream genes. This anecdotal evidence indicates that the operon composition can be altered in flatworms with relative ease, while keeping the genes functional thanks to SL *trans*-splicing.

Lastly, one interesting case is the HOG pair N0.HOG0011542__N0.HOG0002707, annotated as “Cytochrome b-c1 complex subunit 7” and “Enolase” respectively. Both HOGs have evidence of SL *trans*-splicing in some Cestode and Trematode species, with the latter being rare in cestodes but pervasive in trematodes, which is supported by previous studies, including some of the first descriptions of *trans*-splicing in neodermatans (Davis, 1997; Davis & Hodgson, 1997). Furthermore, the operon arrangement between

both genes is only present in Trematodes (see Op8 on **Supplementary File 6**). Thus, it appears that either the Enolase transcript gained SL *trans*-splicing processing when it formed an operon in Trematodes, or partially lost it when it became monocistronic in Cestodes.

Conclusions

In this work we analyzed the intricacies of the diversity and evolution of SL *trans*-splicing among parasitic flatworms, exploring how much can be gleaned from the available datasets. The genomic assemblies and transcriptomic datasets that are available constrain the detection capacity of SLs (identification of SL-RNAs and their acceptor sites) and using dedicated RNA Sequencing methods (such as SL trapping: Nilsson et al., 2010) may impact the clarity of observed evolutionary patterns. However, it is worth noting that when SL trapping was applied to the trematode *S. mansoni*, most of the novel sites detected were only weakly *trans*-spliced (Boroni et al., 2018). If observations on *cis*-splicing are applicable (Melamud & Moult, 2009; Wan & Larson, 2018) it is likely that many weak trans-splicing sites represent “biological noise” (i.e. errors) in the SL *trans*-splicing process. Thus, for those species with high RNA Seq coverage available, our analysis is likely to cover a high percentage of the functionally relevant SL-ACEs.

The observed diversity in SL-RNA between species is both expected and intriguing. While SL-RNA loci are known to be highly variable at large evolutionary distances, for them to function they need to be able to interact with the spliceosome, which may be mediated by their associated proteins (Blumenthal, 2005; Denker et al., 1996, 2002; Fasimoye et al., 2022). This selection constraint however hasn’t stopped them from developing a considerable variability in their primary and secondary structures, including the apparent fusion of hairpins in schistosomatids. It is unclear if this is related to functional differences or its biological relevance since. For example, it is possible that all flatworm SL-RNAs acquire a similar conformation when they interact with their associated proteins.

In any case, this structural variability between species is accompanied with evidence of a constant homogenization of the SL-RNA loci within each species, and a lack of specialization on their target genes. An intriguing possibility is that both are causally linked and self-reinforcing: concerted evolution by gene conversion between loci is allowed by natural selection because all SL-RNAs in the genome are functionally interchangeable, and the evolution of SL-RNA specialization is hampered in flatworms because loci conversion is prevalent. In this scenario, the first evolutionary step toward the evolution of

any SL-RNA specialization would be to stop or at least reduce the rate of loci conversion and allow SL-RNAs to evolve independently. For instance, the loss or fragmentation of SL-RNA clusters should reduce the impact of concerted evolution. Something like this might have happened within Cyclophyllidea and would explain the higher SL-RNA diversity observed in the group. Either way, better genome assemblies of additional Cestodes are required in order to confirm the origin of the SL-RNA clusters observed in taeniidae, and the other particularities associated with Cyclophyllidea, relative to the other flatworms: 1) the relative low abundance of SL bearing reads recovered and smaller number of SL-ACEs, 2) the low abundance of SL sequences incorporated on repeated elements. 3) loss of *trans*-splicing of many ancestral *trans*-spliced HOGs, and 4) an increase in the number of conserved HOG pairs observed in operons. While we cannot confirm it at this time, our manual inspection of selected cases shows that acquisition and potential loss of SL *trans*-splicing on a gene is a dynamic process that can be altered by genome rearrangements, and these are likely to have been prevalent during the genome size contraction that occurred at the base of this group (Coghlan et al., 2019).

In more applied matters, our results offer key insights for the future refinement of annotation pipelines in flatworms. Current approaches clearly struggle to discriminate between individual genes encoded in polycistronic loci, which then hampers follow up studies. The identification of where SL-ACEs are located can be used to, at a minimum, diagnose the troublesome loci and correct the issue, analogous to works realized with nematodes (see Allen et al., 2011). A full automation pipeline however will require overcoming several roadblocks. First is the acquisition of reliable data on the specific SL-ACEs, standard RNAseq dataset clearly are enough to improve the annotation quality of some loci but as of now it is a haphazard process. SL trapping is the obvious solution with its increased coverage of these sites, but additional data would exacerbate the next roadblock: discriminating between biologically important sites from splicing noise. Simply setting a cut off on read support could address the bulk of them but it will require fine-tuning, given that internal sites involved in alternative isoforms might have low SL bearing read counts. Lastly is the matter of sorting SL-ACEs used for operon resolution from those important from alternative isoforms. One strategy that isn't reliant on previous annotations can be to exploit the strong depletion of *cis*-splicing reads on SL-ACEs key to operon resolution, but it too needs proper calibration before any unsupervised application.

Overall, our results concur with previous studies performed on individual species when it comes to the key features of SL *trans*-splicing in parasitic Platyhelminthes but offer for the first time a global view in these species of the patterns and evolution of SL-RNA gene

diversity, SL *trans*-splicing acceptor sites and their widespread involvement with cis-splicing, as well as the organization of *trans*-spliced genes into operons. Our results show the existence of conserved trans-splicing acceptor sites and operons throughout the evolution of this group of parasites, as well as other examples of highly dynamic changes at shorter evolutionary distances. The high prevalence of SL trans-splicing of universally conserved genes in all neodermatans reinforces the idea that this mechanism is likely to be crucial for these parasites and warrants further research into its molecular mechanisms.

Methods

Data selection, quality assessment and transcriptome assembly

RNAseq data and reference genomes for 12 cestode and 12 trematode flatworm species were downloaded from NCBI (NCBI Resource Coordinators, 2015) and the Wormbase Parasite database (Howe et al., 2017). To minimize biases between samples, only paired-end RNAseq data generated with Illumina technology were utilized. Read quality was assessed using FastQC (Andrews, 2010). Adapter sequences and low-quality bases were removed with Trimmomatic v.0.36 (Bolger et al., 2014) using parameters: SLIDINGWINDOW:5:20, MINLEN:25. Lastly, transcriptomes were assembled *de-novo* using Trinity v2.12.0 (Grabher et al., 2013) with read normalization disabled.

SL Sequences identification and retrieval

Novel putative SLs were identified using the SLFinder v1.09 pipeline (Calvelo et al., 2020) with the option "-me False" to maximize potential discovery. Filters based on the donor site and proximity to proteins were not applied. Possible artifacts generated during the initial Hook sequences generation were manually identified and addressed following Calvelo et al. (2020) recommendations. Novel predictions were augmented with Blast searches of previously known SL sequences from other works (**Supplementary Table 1**) against the genomes (Options: -task blastn-short, -perc_identity 95, -ungapped, and -qcov_hsp_perc 75). Lastly, based on the expected total lengths of SL-RNA for parasitic flatworms (Brehm et al., 2000; Davis et al., 1994; Rajkovic et al., 1990), putative SL-RNA sequences were defined as the surrounding region around the initial hits: 20 bases upstream and 100 bases downstream (~140 bases in total) and were retrieved for further analysis. Potential SL-RNA loci located too close to the one of the contigs boundaries to retrieve these sequences, or had more than 30 unknown bases, were excluded from further analysis.

SL-RNA selection and delimitation

Potentially coding SL-RNA loci were filtered based on three assumptions: 1) The mature portion of the SL-RNAs is more conserved than the surrounding sequence in the genome. 2) True SL-RNA loci are not closely associated with mobile elements. 3) Functional SL-RNAs share enough sequence similarity to be grouped together in a Neighbor Joining clustering along with already reported sequences. To achieve this, first, nearly identical putative SL-RNA sequences on each flatworm group were clustered using cd-hit-est v4.8.1 (Fu et al., 2012) with default parameters (>90% identity). All representative sequences of each cluster were aligned with MAFFT v7.475 (Katoh & Standley, 2013) along with known full SL-RNA sequences. Next, a basic phylogenetic tree was estimated using Neighbor Joining phylogeny with Ninja v1.2.2 (Wheeler, 2009), and groups were clustered with Treecluster v1.0.3 (Balaban et al., 2019) in 'med_clade' mode and -t 0.5. Simultaneously, mobile elements were identified for each genome using RepeatMasker v4.1.1 (Smit et al., 2015) with a custom library generated for each through RepeatModeler v2.0.1 (Smit & Hubley, 2015). Candidate SL-RNAs were selected based on: 1) Their inclusion in a phylogenetic cluster with known SL-RNAs and 2) No mobile element identified within 500 bases of the candidate SL-RNA, unless the locus is part of a cd-hit-est cluster with at least one valid member.

Conserved motifs among the filtered sequences for Cestodes and Trematodes, in addition to the known full SL-RNA for each group, were identified using MEME v4.11.2 (Bailey & Elkan, 1994) and utilized to guide manual trimming around conserved regions. An overview of the process is provided on (**Supplementary Figure 1**). SM-like sites were manually identified with the assistance of the 'locate' function from the Seqkit package v0.14.0 (Shen et al., 2016) that matches the motif RAU4-7GR, allowing up to 2 mismatches. Secondary structure predictions were made with RNAfold v2.4.18 (Lorenz et al., 2011) setting the SM-like sites as unpaired, and visualized in Forna (Kerpedjiev et al., 2015). SL-RNA candidates with structures compatible with known SL sequences, having a similar number and disposition of hairpins, were selected for further analysis. Lastly, to rescue more potentially viable SL-RNA sequences, we conducted a new BLAST search with the pre-selected SL-RNA sequences against the reference genomes (-perc_identity 90, -qcov_hsp_perc 85), disregarding mobile element proximity or phylogenetic affinity, and processed the results as described before.

Selected SL-RNA sequences were clustered by sequence identity, leaving a unique representative per species and aligned using MAFFT (Katoh & Standley, 2013) with the L-INS-I method (options --localpair and --maxiterate 1000). Then they were clustered based on similarity by Neighbor Joining using the MEGA 11 Package (Tamura et al.,

2021). Further annotations to the tree were carried out on ITOL(Letunic & Bork, 2021), and additional annotations were added using Inkscape software (Inkscape Project, 2020). A complete registry is provided in **Supplementary Table 3**, and the neighbor-joining trees used for selection are provided in **Supplementary Files 7 and 8**.

SL-RNAs in Echinococcus granulosus genome assembly ID ASM2155672v1

Putative SL-RNAs were determined by blast searches between the unique SL-RNA sequences defined in this study and the genome assembly. Reported coordinates are the summary of all overlapping hits for each loci. Orthology with E. multilocularis and T. multiceps was estimated using Orthofinder v2.5.4 (Emms & Kelly, 2019) with default settings.

SL Acceptor sites identification

SL-bearing reads were identified based on the presence of the last 15 bases preceding the donor site of the Leader region of a selected SL-RNA, using SLFinder-Genes from the SLFinder package (described in Calvelo et al. 2023); located at 40 bases from the read boundaries (35 assumed known bases plus 5 of tolerance). Acceptor sites associated with a reported gene were pre-classified based on their location relative to their associated gene model in Upstream (before the first start codon annotated for the gene's isoforms), Internal (within the CDS of at least one of the isoforms) and downstreamSites assigned to a single gene, that were supported by 4 SL bearing reads or more, had the same strand as their assigned gene, and were not classified as Downstreamwere selected for further analysis.

Chimeric gene models identification

Potential chimeric gene models were identified in a two-stage process, aimed to first identify potential chimeric gene models and then filter out false positives that could arise due to fragmentation of the gene models in the annotation (the opposite bias): First, genes were divided into two by their SL acceptor sites (one at time and with a minimum length of any half: 60 bases) and compared to four reference species using blastx with "-qcov_hsp_perc 40" (*E. multilocularis*, *H. microstoma*, *S. proliferum*, and *Taenia multiceps* for Cestodes; and *C. sinensis*, *F. hepatica*, *O. felineus*, and *S. mansoni* for Trematoda). Genes whose halves matched a different set of genes in at least one of the species were pre-selected as potential chimeric gene models. Second, the BLAST search was repeated with the whole gene, and the position of the acceptor site relative to the hits was evaluated. A gene model was considered chimeric if the acceptor site separated two gene hits in one of the reference species at the acceptor site (maximum overlap tolerated: 30 bases). If

two or more acceptor sites were found, the gene model was subsequently subdivided, as long as each segment contained one gene hit within it. Open reading frames within each valid segment were predicted with getorfs from the Emboss package v6.6.0.0 (Rice et al., 2000). Only the longest isoform with evidence of chimerism was used for further analysis. Schematic representation of different chimera cases are presented in **Supplementary Figure 2.**

Orthology relationships and gene target conservation

Orthology relationships among all the species' genes were estimated using Orthofinder with default setting. Using based on the longest annotated isoform for standard gene models, or the corrected annotation for chimeric genes. Genes containing unknown amino acids were excluded. The predicted Phylogenetic Hierarchical Orthogroups (HOGs) were considered a rough approximation of protein family membership. Functional annotation was carried out with Interproscan v5.62-94.0 (Jones et al., 2014). Interpro entries assigned to one member were assumed to be valid for the entire HOG. Conservation of SL trans-splicing across different species was defined as genes from the same HOG with SL-ACEs identified in both species. Pairwise comparisons were performed using in-house scripts, and heatmaps were generated using the R package pheatmap (Raivo Kolde, 2019). Similarities between the groups Cyclophyllidea (species *E. granulosus*, *E. multilocularis*, *H. diminuta*, *H. microstoma*, *M. corti*, *T. asiatica*, *T. multiceps, T. saginata* and *T. solium*), Basal Cestodes (species *S. solidus*, *S. proliferum* and *S. erinaceieuropaei*), Schistosomatids (species *S. bovis*, *S. haematobium*, *S. japonicum*, *S. mansoni* and *T. regenti*) and Other Trematoda (species *C. sinensis*, *F. gigantica*, *F. hepatica*, *F. buski*, *O. felineus*, *P. heterotremus* and *P. westermani*) with Upset graphs using the UpSetR package v1.4.0 (Conway et al., 2017). Genes showing extensive overlap (≥90%) in their genomic coordinates with another were excluded, as this made the assignment of the SL-ACE unclear (see Methods: Operon identification). The first gene in a chimeric gene model was only included if it possessed an additional SL-ACE to that used to split the chimera into new different gene models.

Operon identification

Operon candidates were defined as genes encoded on the same strand and with an intergenic distance ≤ 300 bases, measured from their most proximal annotated isoforms to each other. Isolated chimeric genes were considered operons regardless of the intergenic distance, based on the assumption that they were miss-annotated initially due to the detection of polycistronic pre-mRNAs in the original work. Mitochondrial genes and gene models overlapping more than 80% with another gene model were discarded.

Candidate operon conservation between two species was defined as the conserved pairs of HOGs present in both genomes, and species both displayed evidence of SL trans-splicing in the gene of the second HOG in the pair. Pairwise comparisons were conducted with in-house scripts, and heatmaps were generated using the R package pheatmap. Similarities between the groups Cyclophyllidea, Basal Cestodes, Schistosomatids and Other Trematoda with Upset graphs using the UpSetR package v1.4.0 (Conway et al., 2017).

Gene expression and read coverage at cis-splicing junctions

Reads were mapped to the reference genomes using STAR v2.7.10b (Dobin et al., 2013) with the suggested parameters for the species genome size (--genomeChrBinNbits and --genomeSAindexNbases set to 12). Read counts were calculated using htseq-count v2.0.2 (Putri et al., 2022) with strand specificity disabled ("-s no"). To minimize the impact of miss-annotations, reads were considered if located anywhere within the maximum extension of each gene (the most distal coordinates of the annotated elements). Then, a second-pass protocol was conducted to identify cis-splicing reads on SL-ACEs, as described in the STAR user manual.

Competition between SL trans-splicing and cis-splicing was evaluated by classifying SL-ACEs based on their location relative to their gene model suspected function: 1) Upstream mono: the SL-ACE is located upstream the CDS of a monocistronic or the first gene in the operon); 2) Upstream Poly: is located upstream the CDS of a gene encoded in a operon loci, excluding the first. And 3) Internal Sites: the SL-ACE is within a CDS. Differences in the ratio between SL trans-splicing and cis-splicing (defined: as "SL bearing reads"/("SL bearing reads" + "cis-splicing reads") using the Kruskal-Wallis rank sum test, implemented on the R package Stats v4.2.2 (R Development Core Team, 2022). Boxplots were generated using the packages ggplot (Wickham, 2016) and ggpubr (Kassambara, 2023) SL-ACEs assigned to chimeric gene models were discarded.

SL TAG specialization on specific sites and operon resolution

Species with 2 or more SL tags in their transcriptome (comprising at least 5% of the observed read counts) were analyzed to explore SL specialization. Expected frequencies were calculated based on the observed read counts on SL-ACEs, assigned to a single gene, were in the same strand, and did not belong to the group downstream, and. Sites with at least 5 read counts observed and expected for all SL tags were subjected to a "Chi-squared test for given probabilities", and p-values were corrected by False Discovery Rate (FDR). Calculations were carried out with the R package Stats.

SL tags from SL-RNA specialized in operon resolution were identified by comparing the proportion of each one ("Number of Reads per SL TAG/Total SL Reads" observed at the site) on sites potentially involved in operon resolution versus all others, using a "Wilcoxon rank sum test with continuity correction", implemented on the R package. SL-ACEs assigned to multiple genes, located on the opposite strand, or found after their assigned gene model, were discarded.

Selection of particular cases of gain and loss of trans-splicing on individual HOGs

HOGs with particularly good evidence of SL trans-splicing were defined as: a) Had gene members from at least 2 species used to identify chimeric gene models on Cestodes (species *E. multilocularis*, *H. microstoma*, *S. proliferum* and *T. multiceps*) and Trematodes (species *C. sinensis*, *F. hepatica*, *O. felineus* and *S. mansoni*). b) At least one of the genes from each represented species in the HOG had evidence of SL trans-splicing or have a TPM equal or higher than the median TPM of genes with confirmed SL *trans*-splicing. c) At least 5 species represented in the HOG have genes subjected to SL. The protein sequences of the Selected PHOs were aligned with MAFFT and their phylogeny estimated with IQ-TREE v2.2.2.3 (Nguyen et al., 2015). Tree figures and annotation of TPM and detected SL reads were generated in iTOL v6. For these evaluations all SL-ACEs were considered, regardless of read counts. Internal SL-ACEs on chimeric gene models were manually evaluated and re-assigned to the upstream portion when their position relative to the overall read mappings and exon structure suggested so.

Selection of particular cases of gain and loss of trans-splicing on polycistronic loci

Candidate polycistronic loci were selected for manual inspection in two groups: by the presence of conserved HOG pairs that had evidence of SL *trans*-splicing and were exclusive for either Cestoda or Trematoda, or were associated with a HOG selected in the previous section. In the case of the former, pairs were considered if they were observed on at least 5 species, distributed among 4 genuses. Once completed the selection, additional pairs were added if they involved one of the HOGs from the selected set, and repeated until exhaustion. Then the location of every gene member HOG was verified and schematized on Inkscape.

Software availability

The scripts and supplementary material are available on the github repository: https://github.com/J-Calvelo/Material_Suplementar_Tesis_Javier_Calvelo.git

References

Agorio, A., Chalar, C., Cardozo, S., & Salinas, G. (2003). Alternative mRNAs arising from trans-splicing code for mitochondrial and cytosolic variants of Echinococcus granulosus thioredoxin glutathione reductase. *Journal of Biological Chemistry*, *278*(15), 12920–12928. https://doi.org/10.1074/jbc.M209266200

Alacid, E., Irwin, N. A. T., Smilansky, V., Milner, D. S., Kilias, E. S., Leonard, G., & Richards, T. A. (2022). A diversified and segregated mRNA spliced-leader system in the parasitic Perkinsozoa. *Open Biology*, *12*(8), 220126. https://doi.org/10.1098/rsob.22.0126

Allen, M. A., Hillier, L. D. W., Waterston, R. H., & Blumenthal, T. (2011). A global analysis of C. elegans trans-splicing. *Genome Research*, *21*(2), 255–264. https://doi.org/10.1101/gr.113811.110

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.* https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Baets, K. de, & Littlewood, T. (2015). *Advances in Parasitology Fossil parasites* (Issue December). https://doi.org/10.1016/s0065-308x(15)x0005-4

Bailey, T., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, *2*, 28–36. https://www.researchgate.net/publication/15615537

Balaban, M., Moshiri, N., Mai, U., Jia, X., & Mirarab, S. (2019). TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS ONE*, *14*(8), e0221068. https://doi.org/10.1371/journal.pone.0221068

Barnes, S. N., Masonbrink, R. E., Maier, T. R., Seetharam, A., Sindhu, A. S., Severin, A. J., & Baum, T. J. (2019). Heterodera glycines utilizes promiscuous spliced leaders and demonstrates a unique preference for a species-specific spliced leader over C . elegans SL1. *Scientific Reports*, *67*(4), 1356. https://doi.org/10.1038/s41598-018-37857-0

Bernard, F., Dargère, D., Rechavi, O., & Dupuy, D. (2023). Quantitative analysis of C. elegans transcripts by Nanopore direct-cDNA sequencing reveals terminal hairpins in non trans-spliced mRNAs. *Nature Communications*, *14*(1), 1229. https://doi.org/10.1038/s41467-023-36915-0

Bitar, M., Boroni, M., Macedo, A. M., Machado, C. R., & Franco, G. R. (2013). The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. *Frontiers in Genetics*, *4*(October), 199. https://doi.org/10.3389/fgene.2013.00199

Blumenthal, T. (2005). Trans-splicing and operons. In *WormBook: the online review of C. elegans biology*. https://doi.org/10.1895/wormbook.1.5.1

Bolger, A. M., Lohse, M., Usadel, B., Planck, M., Plant, M., & Mühlenberg, A. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics.*, *30*(15), 2114–2120. https://doi.org/doi: 10.1093/bioinformatics/btu170

Boroni, M., Sammeth, M., Gava, S. G., Jorge, N. A. N., MacEdo, A. M., MacHado, C. R., Mourão, M. M., & Franco, G. R. (2018). Landscape of the spliced leader trans-splicing mechanism in Schistosoma mansoni. *Scientific Reports*, *8*(1), 3877. https://doi.org/10.1038/s41598-018-22093-3

Brehm, K., Jensen, K., & Frosch, M. (2000). mRNA trans-splicing in the human parasitic cestode Echinococcus multilocularis. *Journal of Biological Chemistry*, *275*(49), 38311–38318. https://doi.org/10.1074/jbc.M006091200

Calvelo, J., Brehm, K., Iriarte, A., & Koziol, U. (2023). Trans-splicing in the cestode Hymenolepis microstoma is constitutive across the life cycle and depends on gene structure and composition. *International Journal for Parasitology*, *53*(2), 103–117. https://doi.org/10.1016/j.ijpara.2022.11.006

Calvelo, J., Juan, H., Musto, H., Koziol, U., & Iriarte, A. (2020). SLFinder, a pipeline for the novel identification of splice-leader sequences: a good enough solution for a complex problem. *BMC Bioinformatics*, *21*, 293. https://doi.org/10.1186/s12859-020-03610-6

Cheng, G., Cohen, L., Ndegwa, D., & Davis, R. E. (2006). The flatworm spliced leader 3′-terminal AUG as a translation initiator methionine. *Journal of Biological Chemistry*, *281*(2), 733–743. https://doi.org/10.1074/jbc.M506963200

Coghlan, A., Tyagi, R., Cotton, J. A., Holroyd, N., Rosa, B. A., Tsai, I. J., Laetsch, D. R., Beech, R. N., Day, T. A., Hallsworth-Pepin, K., Ke, H. M., Kuo, T. H., Lee, T. J., Martin, J., Maizels, R. M., Mutowo, P., Ozersky, P., Parkinson, J., Reid, A. J., … Berriman, M. (2019). Comparative genomics of the major parasitic worms. *Nature Genetics*, *51*(1), 163–174. https://doi.org/10.1038/s41588-018-0262-1

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

Davis, R. E. (1997). Surprising diversity and distribution of spliced leader RNAs in flatworms. *Molecular and Biochemical Parasitology*, *87*, 29–48.

Davis, R. E., & Hodgson, S. (1997). Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in Schistosoma mansoni. *Molecular and Biochemical Parasitology*, *89*, 25–39.

Davis, R. E., Singh, H., Botka, C., Hardwick, C., El Meanawy, M. A., & Villanueva, J. (1994). RNA trans-splicing in Fasciola hepatica. Identification of a spliced leader (SL) RNA and SL sequences on mRNAs. *Journal of Biological Chemistry*, *269*(31), 20026–20030.

Denker, J. A., Maroney, P. A., Yu, Y. T., Kanost, R. A., & Nilsen, T. W. (1996). Multiple requirements for nematode spliced leader RNP function in trans-splicing. *RNA*, *2*(8), 746–755.

Denker, J. A., Zuckerman, D. M., Maroney, P. A., & Nilsen, T. W. (2002). New components of the spliced leader RNP required for nematode trans-splicing. *Nature*, *417*(6889), 667–670. https://doi.org/10.1038/nature756

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Douris, V., Telford, M. J., & Averof, M. (2010). Evidence for Multiple Independent Origins of trans-Splicing in Metazoa. *Molecular Biology and Evolution*, *27*(3), 684–693. https://doi.org/10.1093/molbev/msp286

Emms, D. M., & Kelly, S. (2018). STAG: Species Tree Inference from All Genes. *BioRxiv*. https://doi.org/https://doi.org/10.1101/267914

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*, 238. https://doi.org/10.1186/s13059-019-1832-y

Ershov, N. I., Mordvinov, V. A., Prokhortchouk, E. B., Pakharukova, M. Y., Gunbin, K. V., Ustyantsev, K., Genaev, M. A., Blinov, A. G., Mazur, A., Boulygina, E., Tsygankova, S., Khrameeva, E., Chekanov, N., Fan, G., Xiao, A., Zhang, H., Xu, X., Yang, H., Solovyev, V., … Skryabin, K. G. (2019). New insights from Opisthorchis felineus genome: Update

on genomics of the epidemiologically important liver flukes. *BMC Genomics*, *20*(1), 399. https://doi.org/10.1186/s12864-019-5752-8

Fasimoye, R. Y., Spencer, R. E. B., Soto-Martin, E., Eijlers, P., Elmassoudi, H., Brivio, S., Mangana, C., Sabele, V., Rechtorikova, R., Wenzel, M., Connolly, B., Pettitt, J., & Müller, B. (2022). A novel, essential trans-splicing protein connects the nematode SL1 snRNP to the CBC-ARS2 complex. *Nucleic Acids Research*, *50*(13), 7591–7607. https://doi.org/10.1093/nar/gkac534

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Giri, S., & Parija, S. C. (2012). A review on diagnostic and preventive aspects of cystic echinococcosis and human cysticercosis. *Tropical Parasitology*, *2*(2), 99–108. https://doi.org/10.4103/2229-5070.105174

Grabher, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X. F., Raychowdhury, L. R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Palma, F. di, W., B., Friedman, N., & Regev, A. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883.Trinity

Hastings, K. E. M. (2005). SL trans-splicing: Easy come or easy go? *Trends in Genetics*, *21*(4), 240–247. https://doi.org/10.1016/j.tig.2005.02.005

Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P., & Berriman, M. (2017). WormBase ParaSite − a comprehensive resource for helminth genomics. *Molecular & Biochemical Parasitology*, *215*, 2–10. https://doi.org/10.1016/j.molbiopara.2016.11.005

Inkscape Project. (2020). *Inkscape* (1.0.2-2). https://inkscape.org

Islas-Flores, T., Galán-Vásquez, E., & Villanueva, M. A. (2021). Screening a spliced leader-based Symbiodinium microadriaticum cDNA library using the yeast-two hybrid system reveals a hemerythrin-like protein as a putative smicRACK1 ligand. *Microorganisms*, *9*(4), 791. https://doi.org/10.3390/microorganisms9040791

Jaeckisch, N., Yang, I., Wohlrab, S., Glöckner, G., Kroymann, J., Vogel, H., Cembella, A., & John, U. (2011). Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate Alexandrium ostenfeldii. *PLoS ONE*, *6*(12), e28012. https://doi.org/10.1371/journal.pone.0028012

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kassambara, A. (2023). *ggpubr: "ggplot2" Based Publication Ready Plots (0.6.0)*. https://cran.r-project.org/web/packages/ggpubr/index.html

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kerpedjiev, P., Hammer, S., & Hofacker, I. L. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, *31*(20), 3377–3379. https://doi.org/10.1093/bioinformatics/btv372

Korhonen, P. K., Kinkar, L., Young, N. D., Cai, H., Lightowlers, M. W., Gauci, C., Jabbar, A., Chang, B. C. H., Wang, T., Hofmann, A., Koehler, A. V., Li, J., Li, J., Wang, D., Yin, J., Yang, H., Jenkins, D. J., Saarma, U., Laurimäe, T., … Gasser, R. B. (2022).

Chromosome-scale Echinococcus granulosus (genotype G1) genome reveals the Eg95 gene family and conservation of the EG95-vaccine molecule. *Communications Biology*, *5*(1), 199. https://doi.org/10.1038/s42003-022-03125-1

Krause, M., & Hirsh, D. (1987). A trans-spliced leader sequence on actin mRNA in C. elegans. *Cell*, *49*(6), 753–761. https://doi.org/10.1016/0092-8674(87)90613-1

Krchňáková, Z., Krajčovič, J., & Vesteg, M. (2017). On the Possibility of an Early Evolutionary Origin for the Spliced Leader Trans-Splicing. *Journal of Molecular Evolution*, *85*(1–2), 37–45. https://doi.org/10.1007/s00239-017-9803-y

Lasda, E. L., Allen, M. A., & Blumenthal, T. (2010). Polycistronic pre-mRNA processing in vitro: snRNP and pre-mRNA role reversal in trans-splicing. *Genes and Development*, *24*(15), 1645–1658. https://doi.org/10.1101/gad.1940010

Lasda, E. L., & Blumenthal, T. (2011). Trans-splicing. *Wiley Interdisciplinary Reviews: RNA*, *2*(3), 417–434. https://doi.org/10.1002/wrna.71

Laumer, C. E., Hejnol, A., & Giribet, G. (2015). Nuclear genomic signals of the "microturbellarian" roots of platyhelminth evolutionary innovation. *ELife*, *2015*(4), 1–31. https://doi.org/10.7554/eLife.05503

Letunic, I., & Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Lidie, K. B., & Van Dolah, F. M. (2007). Spliced leader RNA-mediated trans-splicing in a dinoflagellate, Karenia brevis. *Journal of Eukaryotic Microbiology*, *54*(5), 427–435. https://doi.org/10.1111/j.1550-7408.2007.00282.x

Lorenz, R., Bernhart, S. H., Siederdissen, C. H. Z., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26. https://doi.org/10.1093/nar/gkz164

Mahmud, R., Ai, Y., & Lim, L. (2017). *Medical Parasitology* (1st ed.). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-319-68795-7

Marlétaz, F., Gilles, A., Caubit, X., Perez, Y., Dossat, C., Samain, S., Gyapay, G., Wincker, P., & Le Parco, Y. (2008). Chætognath transcriptome reveals ancestral and unique features among bilaterians. *Genome Biology*, *9*(6), R94. https://doi.org/10.1186/gb-2008-9-6-r94

Matsuo, M., Katahata, A., Satoh, S., & Matsuzaki, M. (2018). Characterization of spliced leader trans-splicing in a photosynthetic rhizarian amoeba, Paulinella micropora, and its possible role in functional gene transfer. *PLoS ONE*, *13*(7), e0200961. https://doi.org/10.1371/journal.pone.0200961

Melamud, E., & Moult, J. (2009). Stochastic noise in splicing machinery. *Nucleic Acids Research*, *37*(14), 4873–4886. https://doi.org/10.1093/nar/gkp471

Michaeli, S. (2011). Trans-splicing in trypanosomes: Machinery and its impact on the parasite transcriptome. *Future Microbiology*, *6*(4), 459–474. https://doi.org/10.2217/fmb.11.20

NCBI Resource Coordinators. (2015). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, *43*(Database issue), D6–D17. https://doi.org/10.1093/nar/gks1189

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nilsson, D., Gunasekera, K., Mani, J., Osteras, M., Farinelli, L., Baerlocher, L., Roditi, I., & Ochsenreiter, T. (2010). Spliced leader trapping reveals widespread alternative splicing

patterns in the highly dynamic transcriptome of Trypanosoma brucei. *PLoS Pathogens*, *6*(8), e1001037. https://doi.org/10.1371/journal.ppat.1001037

Olson, P., Tracey, A., Baillie, A., James, K., Doyle, S., Buddenborg, S., Rodgers, F., Holroyd, N., & Berriman, M. (2020). Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection. *BMC Biology*, *18*, 165. https://doi.org/10.1101/2020.04.08.031872

Pandarakalam, G. C., Speake, M., McElroy, S., Alturkistani, A., Philippe, L., Pettitt, J., Müller, B., & Connolly, B. (2019). A high-throughput screen for the identification of compounds that inhibit nematode gene expression by targeting spliced leader trans-splicing. *International Journal for Parasitology: Drugs and Drug Resistance*, *10*, 28–37. https://doi.org/10.1016/j.ijpddr.2019.04.001

Pouchkina-Stantcheva, N. N., & Tunnacliffe, A. (2005). Spliced leader RNA-mediated trans-splicing in phylum rotifera. *Molecular Biology and Evolution*, *22*(6), 1482–1489. https://doi.org/10.1093/molbev/msi139

Protasio, A. V., Tsai, I. J., Babbage, A., Nichol, S., Hunt, M., Aslett, M. A., de Silva, N., Velarde, G. S., Anderson, T. J. C., Clark, R. C., Davidson, C., Dillon, G. P., Holroyd, N. E., LoVerde, P. T., Lloyd, C., McQuillan, J., Oliveira, G., Otto, T. D., Parker-Manuel, S. J., … Berriman, M. (2012). A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. *PLoS Neglected Tropical Diseases*, *6*(1), e1455. https://doi.org/10.1371/journal.pntd.0001455

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., & Zanini, F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics*, *38*(10), 2943–2945. https://doi.org/10.1093/bioinformatics/btac166

R Development Core Team, R. (2022). R: A Language and Environment for Statistical Computing. In R. D. C. Team (Ed.), *R Foundation for Statistical Computing*. R Foundation for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7

Radío, S., Fort, R. S., Garat, B., & Sotelo-silveira, J. (2018). UTRme: A Scoring-Based Tool to Annotate Untranslated Regions in Trypanosomatid Genomes. *Frontiers in Genetics*, *9*(December), 671. https://doi.org/10.3389/fgene.2018.00671

Raivo Kolde. (2019). *pheatmap: Pretty Heatmaps* (1.0.12). https://CRAN.R-project.org/package=pheatmap

Rajkovic, A., Davis, R. E., Simonsen, J. N., & Rottman, F. M. (1990). A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(22), 8879–8883. https://doi.org/10.1073/pnas.87.22.8879

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Ross, L. H., Freedman, J. H., & Rubin, C. S. (1995). Structure and expression of novel spliced leader RNA genes in Caenorhabditis elegans. *The Journal of Biological Chemistry*, *270*(37), 22066–22075. http://www.ncbi.nlm.nih.gov/pubmed/7665629

Rossia, A., Jackb, E. J. R. A., & Alvarado, A. S. (2014). Molecular cloning and characterization of SL3: A stem cell- specific SL RNA from the planarian Schmidtea mediterranea. *Gene*, *533*(1), 156–167. https://doi.org/doi:10.1016/j.gene.2013.09.101. Molecular

Sather, S., & Agabian, N. (1985). A 5’ spliced leader is added in trans to both alpha- and beta-tubulin transcripts in Trypanosoma brucei. *Proceedings of the National Academy of*

Sciences of the United States of America, *82*(17), 5695–5699. https://doi.org/10.1073/pnas.82.17.5695

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Siegel, T. N., Gunasekera, K., Cross, G. A. M., & Ochsenreiter, T. (2011). Gene expression in Trypanosoma brucei: Lessons from high-throughput RNA sequencing. *Trends in Parasitology*, *27*(10), 434–441. https://doi.org/10.1016/j.pt.2011.05.006

Slamovits, C. H., & Keeling, P. J. (2008). Widespread recycling of processed cDNAs in dinoflagellates. *Current Biology*, *18*(13), R550–R552. https://doi.org/https://doi.org/10.1016/j.cub.2008.04.054

Smit, A., & Hubley, R. (2015). *RepeatModeler Open-1.0*. http://www.repeatmasker.org. http://www.repeatmasker.org

Smit, A., Hubley, R., & Green, P. (2015). *RepeatMasker Open-4.0*. http://www.repeatmasker.org

Steele, R. E., Hampson, S. E., Stover, N. A., Kibler, D. F., & Bode, H. R. (2004). Probable horizontal transfer of a gene between a protist and a cnidarian. *Current Biology*, *14*(8), R298–R299. https://doi.org/10.1016/j.cub.2004.03.047

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027. https://doi.org/10.1093/molbev/msab120

Tessier, L., Keller, M., Chan, R. L., Fournier, R., & Weil, J. (1991). Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena. *The EMBO Journal*, *10*(9), 2621–2625.

Tsai, I. J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K. L., Tracey, A., Bobes, R. J., Fragoso, G., Sciutto, E., Aslett, M., Beasley, H., Bennett, H. M., Cai, J., Camicia, F., Clark, R., Cucher, M., De Silva, N., Day, T. A., … Valdes, V. (2013). The genomes of four tapeworm species reveal adaptations to parasitism. *Nature*, *496*(7443), 57–63. https://doi.org/10.1038/nature12031

Vandenberghe, A. E., Meedel, T. H., & Hastings, K. E. M. (2001). mRNA 5'-leader trans-splicing in the chordates. *Genes & Development*, *15*(3), 294–303. https://doi.org/10.1101/gad.865401.Evidence

Wan, Y., & Larson, D. R. (2018). Splicing heterogeneity: Separating signal from noise. *Genome Biology*, *19*(1), 86. https://doi.org/10.1186/s13059-018-1467-4

Wang, J., Paz, C., Padalino, G., Coghlan, A., Lu, Z., Gradinaru, I., Collins, J. N. R., Berriman, M., Hoffmann, K. F., & Collins Iii, J. J. (2020). Large-scale RNAi screening uncovers therapeutic targets in the parasite Schistosoma mansoni. *Science*, *369*(6511), 1649–1653. https://doi.org/10.5061/dryad.zs7h44j4v

Webb, C., & Cabada, M. M. (2017). Intestinal cestodes. *Current Opinion in Infectious Diseases*, *30*(5), 504–510. https://doi.org/10.1097/QCO.0000000000000400

Wenzel, M. A., Müller, B., & Pettitt, J. (2021). SLIDR and SLOPPR: flexible identification of spliced leader trans-splicing and prediction of eukaryotic operons from RNA-Seq data. *BMC Bioinformatics*, *22*, 140. https://doi.org/10.1186/s12859-021-04009-7

Wenzel, M., Johnston, C., Müller, B., Pettitt, J., & Connolly, B. (2020). Resolution of polycistronic RNA by SL2 trans-splicing is a widely conserved nematode trait. *Rna*, *26*(12), 1891–1904. https://doi.org/10.1261/RNA.076414.120

Wheeler, T. J. (2009). Large-scale neighbor-joining with NINJA. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5724*, 375–389. https://doi.org/10.1007/978-3-642-04241-6_31

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. In *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2nd ed.). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-0-387-98141-3

Yague-sanz, C., & Hermand, D. (2018). SL-quant: a fast and flexible pipeline to quantify spliced leader trans-splicing events from RNA-seq data. *GigaScience*, *7*(7), 1–7. https://doi.org/10.1093/gigascience/giy084

Zhang, H., Hou, Y., Miranda, L., Campbell, D. A., Sturm, N. R., Gaasterland, T., & Lin, S. (2007). Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences*, *104*(11), 4618–4623. https://doi.org/10.1073/pnas.0700258104

Suplementary Figures

- **Supplementary Figure 1:** Overview of the trimming process based on MEME motifs. A) In Cestodes the process was straightforward as 4 highly conserved motifs were identified. The trimming was decided to be conducted at the start of Motif-3 and inside Motif-2. This was done because the second less conserved half was less recognizable outside Taeniidae. Trematoda on the other hand required one round of filtration B) as only two motifs were predicted with unclear boundaries. After the exclusion of sequences with poor matches for Motif-2, like *Paragonimus westermani* Def_Loci-44, the analysis was repeated in a second round C). Here 3 Motifs were identified consistently in both the selected sequences and the reference SL-RNAs for Trematoda. The boundaries of Motifs -2 and -3 were again selected as the boundary of SL-RNAs. The loci Def_Loci-85 and -86 of *Clonorchis sinensis* were included in the final analysis despite lacking Motif-2 because it's SL TAG "Trematoda_E" was found in the species *C. sinensis*, *Fasciola gigantica*, *Fasciola hepatica* and *Fasciolopsis buski*. Despite this SL TAG low numbers, its phylogenetic distribution suggests there are SL-RNAs with "Trematoda_E" exists within this lineage, even if they possess limited functionality, and no better representative could be found.
- **Supplementary Figure 2**: Schematic representation of the resolution of several gene models identified in this work. The annotated transcript is represented with a black, with the SL Acceptor sites represented with arrows above, BLAST hits with other genes bellow with colored boxes and the selected chimeric portions with arches. A) In the classical case there is a single SL-ACE that separated two halves with different sets of hits. B) If there are additional SL-ACEs but it is not reflected on changes on the hits it was assumed the SL-ACE was ignored. C) Cases where there are multiple SL-ACEs that separate both sets of hits were used to narrow down the intergenic space by excluding the sequence between them. In an analogous case, if SL-ACEs were D)

upstream or E) downstream areas, separating genes with no BLAST hits, those sequences were trimmed down. F) A third gene in the chimeric gene model required that the SL-ACEs separated three discrete areas with their own sets of BLAST Hits. G) Cases where subdivisions where the boundary between genes was unclear were excluded.

Supplementary Files

- **Supplementary File 1:** Sequences of the unique SL-RNA loci identified for each species.
- **Supplementary File 2:** Full secondary structures predicted by RNAfold for the unique SL-RNAs sequences described in their work and the original reference sequences. Each registry is composed of 6 lines: 1) Sequence ID, 2) Sequence, 3) unpaired sequences set as input (the SM-like site), 4) Maximum Expected Accuracy (MEA) prediction, 5) Minimum Free Energy (MFE) prediction and 6) Ensemble Free Energy prediction.
- **Supplementary File 3:** Description of the outlier secondary structures of SL-RNA sequences.
- **Supplementary File 4**: Observations on the trimming sensitivity on the predicted SL-RNA structures.
- **Supplementary File 5:** Maximum Likelihood phylogenetic trees of the selected Phylogenetically Hierarchical Orthogroups (HOGs) based on their expression level and evidence of SL trans-splicing. For each gene it is indicated if it had evidence of SL trans-splicing above or below the 4 read threshold on the same acceptor site (Black Star or White star respectively). Followed by a Normalized TPM per species relative to the median TPM of SL trans-spliced genes and the total amount of SL bearing reads in the gene in logarithmic scale, displayed with green and red bars respectively. Both are capped (10 for the normalized TPM and 100 SL bearing reads), indicated with circles. Genes identified as members of operon candidates are indicated with a blue check mark
- **Supplementary File 6**: Sketches for the manually analyzed operon candidates that were selected based on their HOG pairs exclusiveness for Cestodes or Trematodes; and/or their association with one with the HOGs selected by their expression level.
- **Supplementary File 7:** Neighbor joining three of the potential SL-RNA candidates found in Cestodes. The ID format for SL-RNA candidates is 1) CD-HIT Cluster ID, 2) Species 3) Loci ID 4) Repeat elements found within a 500 bases. Points 1 and 4 are only included if relevant.

- **Supplementary File 8:** Neighbor joining three of the potential SL-RNA candidates found in Trematodes. The ID format is 1) CD-HIT Cluster ID, 2) Species 3) Loci ID 4) Repeat elements found within a 500 bases. Points 1 and 4 are only included if relevant.

Supplementary Tables

- **Supplementary Table 1:** Sequence summary of the known SL Sequences from plathelminth species and the source paper where they were initially reported. Species with an "*" were not used for SL TAG definition due to their ambiguous nucleotides on their last 15 bases.
- **Supplementary Table 2:** Summary of the structural features and metrics for the SL-RNAs identified in this word. For each SL-RNA it is detailed the species where it was found and the total number of coding loci identified, length, GC% content, SL TAG, Donor site (with a "." Indicating the canonical "ATGGT"), sequence of the SM-like site, structure type (see Figure 2 in the main text), and the secondary structure metrics estimated by RNAfold about their stability. These are the Minimum Free Energy (MFE); Centroid Energy and Distance; Maximum Expected Accuracy (MEA) energy, value and Structural Frequency; and Ensemble diversity. The standard reported structures are shown on Figure 2 in the main text and the outliers on Support File 2. The Unique_SL_1 to 72 were found in the analyzed genomes (see Supplementary Table 2 for their loci locations), the remaining sequences were found among reference sequences only.
- **Supplementary Table 3:** Location of all putative Spliced Leader (pSL) loci considered in this work. For each one it is detailed its location, SL-TAG, how it was detected: SLFinder (SLF), BLAST with known Sequences (BL), and the Recovery BLAST (RBL); Loci marked with a "*" where located too close to the Contig boundary and were discarded. Then is information important for pre-filters: types of Repeated elements within a 500 radius, number of mismatches and the cluster IDs generated by CD-HIT and Treecluster. And lastly for those identified as likely functional SL-RNAs it is detailed their non-ambiguous ID, refined coordinates, and hairpin structure. Reference sequences were assigned to Treecluster groups: Ces_Tree_clus-1, Tre_Tree_clus-11, Tre_Tree_clus-15 and Tre_Tree_clus-20. Trees provided on Supplementary files 1 and 2.
- **Supplementary Table 4:** SL-RNA tandem repeats identified in the analyzed species.

- **Supplementary Table 5:** SL-RNA loci identified on the *Echinococcus granulosus* genome assembly ASM2155672v1, their SL tag and their best match in our study determined by BLAST searches.
- **Supplementary Table 6**: Genes surrounding the SL-RNA cluster on chromosome 1 on the *Echinococcus granulosus* genome assembly ASM2155672v1 orthologous with the genes that identified the homology of SLClus-7 and SLClus-8 in *E. multilocularis* and *Taenia miltiliceps*.
- **Supplementary Table 7**: Summary of the SL bearing counts for all studied species (total SL counts and counts for each SL tag). SL tags for which a SL-RNA locus was identified containing that sequence in the genome of each species are highlighted in green.
- **Supplementary Table 8**: Full registry of the acceptor sites for SL trans- splicing (SL-ACEs) supported by at least 1 SL bearing read. For each site it is detailed its location (coordinate, chromosome/scaffold/contig, strand) total SL bearing reads, cis-splicing reads, cis-splicing donor sites and to what group was assigned for the splicing analysis. Then information associated with the assigned genes (one or more due to overlaps): if it matches the SL-ACE strand, position relative to the gene model, and how many SL bearing reads were assigned to it and from what SL TAG (following the format Tag1; Tag2; Tag3…). And lastly the subdivision of the chimeric gene model, operon position if applicable or if it was discarded from the Operon Search due to extensive overlaps; if they are applicable.
- **Supplementary Table 9**: Summary of SL trans-spliced genes matching BUSCO gene markers.
- **Supplementary Table 10**: Result summary for chimeric gene model search on genes with internal SL Acceptor Sites (SL-ACE) that could be potential mis-annotated operons. The analysis was divided in two stages. First a BLAST search against four reference genomes: *E. multilocularis*, *H. microstoma*, *S. proliferum*, and *T. multiceps* for Cestodes; and *C. sinensis*, *F. hepatica*, *O. felineus*, and *S. mansoni* for Trematoda. In the Second stage the BLAST search was repeated against the entire gene and the position of the SL-ACEs evaluated relative to the BLAST hits downstream and upstream.
- **Supplementary Table 11**: Complete registry of the chimeric gene models subdivisions identified in this work. Including the cut points in the transcript and the genome, and the gene IDs matching each portion. “Start” and “End” refer to the coordinates of the transcript in the species genome annotation, numbers to SL Acceptor sites.

- **Supplementary Table 12:** Summary of cis-splicing detected on SL Acceptor Sites (SL-ACEs) involved on Transcript Initiation (Start: SL-ACE located upstream the predicted monocistronic gene model, or is the first in the operon), Operon Resolution (SL-ACE located putatively important for the operon loci resolution) and Others. First it is detailed SL-ACEs with any evidence of cis-splicing and then the better supported with more than 3 reads.
- **Supplementary Table 13**: Results for the "Kruskal-Wallis rank sum test" evaluating the enrichment in proportion of SL trans-spliced reads relative to all splicing events. SL-ACEs are classified as 1) Upstream Mono: upstream the coding region of a monocistronic gene or of the first gene of an operon. 2) Upstream Poly: upstream the coding region of a downstream gene in an operon, and thus related to polycistron resolution. And 3) Internal: the SL-ACEs are located within the coding sequence; "Others"). The species *Mesocestoides corti* due to limited data.
- **Supplementary Table 14**: Summary of the chi-square results evaluating sites that did not followed the average SL Tag Frequencies.
- **Supplementary Table 15**: Results of the chi-square results evaluating sites that did not followed the average SL Tag Frequencies.
- **Supplementary Table 16**: Results for the "Kruskal-Wallis rank sum test" evaluating the differential SL tag use on SL-ACEs classified as 1) Upstream Mono: upstream the coding region of a monocistronic gene or of the first gene of an operon. 2) Upstream Poly: upstream the coding region of a downstream gene in an operon, and thus related to polycistron resolution. 3) Internal: the SL-ACEs are located within the coding sequence; "Others"). And 4) Chimeras: any SL-ACE assigned to a chimeric gene model. The species Mesocestoides corti due to limited data.
- **Supplementary Table 17**: Phylogenetically Hierarchical Orthogroups (HOGs) identified by Orthofinder. Including the original orthogroup ID, the node ID for which it was defined and the member genes per species and the assigned Interpro IDs for each HOG (at least one of its members. For each species it is also summarized how many member genes of the HOGs were found and how many of them have assigned SL-ACEs.
- **Supplementary Table 18:** Candidate operon summary across the studied species and it's relation with SL trans-splicing.
- **Supplementary Table 19:** All identified operon candidates in the studied species, their location, member genes and assigned Phylogenetically Hierarchical Orthogroups (HOGs). Genes and HOGs are sorted from upstream to downstream, with

the intergenic space indicated between parentheses. Genes with assigned SL Acceptor Sites (SL-ACEs) are indicated with "*", chimeric gene models with "[C]".

- **Supplementary Table 20:** Selected Phylogenetically Hierarchical Orthogroups (HOGs) based on their expression level and evidence of SL trans-splicing. The total number of species represented in both Cestodes and Trematodes, how many of them have evidence of SL trans-splicing (above 3 SL bearing reads), a basic interpretation of where the SL trans-splicing was gained in the phylogeny given the observed species, Interpro annotation and notes.
- **Supplementary Table 21**: Description of the manually analyzed operon candidates that were selected based on their of HOG pairs exclusiveness for Cestodes or Trematodes; and/or their association with one with the HOGs selected by their expression level. For each candidate it is detailed the HOGs displayed on their respective sketch in **Supplementary File 6**, along with their interpro annotation, total number of cestodes and trematode species where they are present and how many of them have evidence of SL. For the whole operon candidate it is also included flags for apparently novel SL trans-splicing in a gene due to operon re-arrangements, if different lineages display alternative HOGs and if those derive from the same raw orthogroup, and basic notes.

Discusión y conclusiones generales

5.A) Dificultades y oportunidades en el estudio de *SL trans-splicing*

La investigación presentada en esta tesis explora extensivamente las dificultades asociadas con el análisis de SL *trans-splicing* en cualquier organismo. El primer paso es identificar los SLs presentes en el transcriptoma de la especie, comenzando por el problema engañosamente simple: ¿Cuáles son las secuencias de los SLs, cuando estas son desconocidas? La solución alcanzada fue SLFinder, presentada en el **Capítulo 2**. Una herramienta útil, pero con limitaciones a tener presente: (1) SLs menores a 20 bases son difíciles de detectar debido a la secuenciación y/o ensamblado incompleto de estas secuencias, muy relevante en grupos como tunicados (Matsumoto et al., 2010), (2) la presencia de múltiples SLs con secuencias similares puede afectar el ensamblado de secuencias candidatas (los *"*Hook*"*) y enmascarar la presencia de SLs reales de baja frecuencia. Ejemplificado por la detección incompleta de las variantes de SL2 en *Caenorhabditis elegans* (Calvelo et al., 2020). Más importante, la publicación de SLIDR (Wenzel et al., 2021) vuelve nuestra aproximación parcialmente obsoleta al ofrecer una alternativa más eficaz computacionalmente y con mayor sensibilidad, gracias a su enfoque directamente en el mapeo de *reads* al genoma. Dicho esto, SLFinder posee la ventaja de no depender de asunciones sobre la estructura de los SL-ARN o la conservación de sitios particulares para filtrar contaminantes. Como por ejemplo la presencia de un sitio SM-like con una secuencia específica para la especie, rodeada por dos horquillas. Más aún, ofrece un punto de partida para identificar SLs en especies de las que se carece de un ensamblado genómico confiable. El usuario solo requiere implementar otra estrategia para descartar falsos positivos, una tarea más manejable que la inspección manual de secuencias.Tarea de utilidad en sí misma ya que permite el diseño de *primers* específicos y su uso en tecnologías basadas en amplificaciones por reacciones PCR. Facilitando, a modo de ejemplo, el análisis de transcriptos específicos a partir de muestras complejas. Como por ejemplo aquellas contaminadas con otros organismos debido a su naturaleza (ej: muestras ambientales o

parásitos embebidos en los tejidos del hospedador). En este contexto, es importante considerar las observaciones realizadas en el **Capítulo 4**: especies de grupos como Cyclophyllidea (principal grupo de Cestodos en diversidad de especies) pueden mostrar variabilidad en estas secuencias dentro de un mismo género. Dependiendo de la cantidad y localización de estas divergencias potencialmente podrían afectar la calidad y viabilidad de estas amplificaciones.

El siguiente paso consiste en determinar la localización y extensión de los SL-ARN, discriminando los loci funcionales de pseudogenes y otros loci poseedores de la secuencia SL como elementos transponibles o genes retrotranscriptos al genoma. En el **Capítulo 3** se recurrió a la homología a secuencias conocidas (Olson et al., 2020; Tsai et al., 2013); mientras que en el **Capítulo 4** se implementó un proceso semi-manual guiado por la conservación de secuencia de SL-ARNs entre especies. La primera solución es definitivamente la más confiable y sencilla de implementar, pero frecuentemente no es viable y puede llevar a errores de interpretación. Más aún, como se discute en el **Capítulo 4**, la variabilidad intrínseca de los loci SL-ARN vuelve el uso de referencias cuestionable para su delimitación.

Una vez identificados los SLs, el siguiente paso es identificar los blancos de SL *trans-spilcing* a nivel de genes y sitios aceptores SLs (SL-ACEs). Este paso es conceptualmente sencillo, identificar las secuencias con evidencia de SL *trans-splicing* y localizarlas en el genoma. Una vez localizados, es una cuestión determinar a qué gen está asociado cada SL-ACE y evaluar su posible rol biológico. Sin embargo, como es observado en el **Capítulo 3** y expandido en el **Capítulo 4**, los “pipelines” automáticos utilizados para anotar los genomas de platelmintos tienden a reportar loci policistrónicos como un único modelo génico quimérico. En consecuencia, la anotación del genoma puede: 1) carecer de genes importantes para la fisiología de la especie, 2) perjudicar estudios de expresión génica al mezclar los patrones de expresión de dos o más genes, y 3) sugerir la existencia de familias génicas artefactuales mediante la combinación de todos los dominios proteicos de los genes quiméricos en una

única predicción. La solución implementada en el **Capítulo 3** se basa en la fragmentación en los sitios de inserción y posterior blasteo a una referencia que es tomada como el estándar a alcanzar (*Echinococcus multilocularis* en este caso); expandida en el **Capítulo 4** con el uso de 4 referencias por linaje y un mejor control de casos con “hits” solapantes en búsquedas BLAST.

La re-anotación de estos loci en los **Capítulos 3 y 4** demuestra el potencial de incorporar evidencia de SL *trans-splicing* en los *pipelines* de anotación genómica. Con las inserciones del SL actuando como indicadores claros de la existencia de un gen adicional en el loci, posiblemente expresados en conjunto como un operón policistrónico. Sin embargo, es necesario reiterar que la aproximación metodológica presentada en esta tesis no es suficiente para su implementación a gran escala. Los modelos génicos quiméricos descritos fueron detectados únicamente por la identificación de SL-ACEs internos, que luego fueron asociados a “hits” de BLAST claramente definidos con referencias afectadas por el mismo problema de quimerismo. Una solución escalable y confiable requiere la recuperación eficaz de los sitios sometidos a SL *trans-splicing* y su correcta clasificación entre inicio de transcripción, isoforma alternativa o ruido biológico. La descripción esporádica realizada en esta tesis, basada en el análisis del ≲0.12% de los *reads* con evidencia de SL *trans-splicing*, permitiría identificar estos artefactos únicamente en genes con niveles altos de expresión que por lo tanto se encuentran suficientemente representados en estas librerías.

Finalmente, la identificación de múltiples secuencias SL asociadas a elementos transponibles en el **Capítulo 4** sugiere un rol del SL *trans-splicing* en el control de la transposición de elementos transponibles en un genoma. Aunque la presente tesis se centró en procedimientos para filtrar estos loci, su existencia sugiere que los SLs pueden ser integrados en esta clase de elementos durante el proceso de transposición. Aunque los mecanismos aún deben ser estudiados, una explicación es que los SLs son incorporados en los ARN de estos elementos móviles tras su transcripción, potencialmente degradando la capacidad de

replicación de las copias resultantes (los loci observados) o incluso previniendo su inserción en el genoma. De confirmarse, SL *trans-splicing* jugaría un rol en la evolución de los elementos transponibles en los genomas de estas especies.

5.B) SL *trans-splicing* en platelmintos

Los resultados de la tesis confirman en gran medida el conocimiento previo en SLs en el grupo platelmintos (Bitar et al., 2013; Lasda & Blumenthal, 2011): su importancia en la resolución de operones, la presencia de un codón "AUG" en el SL conservado, seguido de un sitio de *splicing* canónico "UGGU", y la presencia de un sitio SM-like. Este último presenta una considerable variación nucleotídica entre SL-ARNs, inclusive dentro de la misma especie y sin un claro impacto en su predominancia en el transcriptoma (Ver SL-2 y SL-3 en **Capítulo 3**). Aunque es importante señalar que las variaciones nucleotídicas se concentran en la porción 5' del motivo, la porción 3' es más conservada y presumiblemente más importante. En términos de estructura secundaria, los SL-ARN encontrados siguen las estructuras reportadas por otros trabajos en *E. multilocularis* (Brehm et al., 2000), *F. hepatica* (Davis et al., 1994) y *S. mansoni* (Rajkovic et al., 1990) pero con una considerable diversidad en términos de la posición de los bucles y largo de *hairpins*. La relevancia biológica de estas predicciones es cuestionable. No solo por no considerarse las interacciones con otros elementos del spliceosoma, sino por la alta sensibilidad de las predicciones a la extensión del SL-ARN considerada (discutido en **Capítulo 4**). Sin embargo, la existencia de una gran variabilidad a nivel nucleotídico, y la evidente funcionalidad del SL-ARN observado en Schistosomatidae a pesar de la fusión de las horquillas previas al sitio SM, sugiere que hay una considerable tolerancia a variaciones en el plegamiento del SL-ARN. En este contexto, es posible que las dificultades encontradas en la delimitación de los SL-ARN respondan a la falta de conservación de estas regiones, y que representen regiones espaciadoras.

Otro aspecto importante para considerar es el número y disposición de loci SL-ARN en los genomas. Especies con ensamblados de buena calidad como *H. microstoma* y *S. mansoni*

muestran que su número puede ser variable, por reducciones y potencialmente expansiones en su número de copias, y el *cluster* descrito en Taeniidae muestra importantes consecuencias evolutivas. Los resultados indican que la evolución concertada de las copias en tándem causa que todo el *cluster* evolucione como una unidad, y su mayor número de copias podría promover al SL-ARN codificado en él a convertirse en el SL predominante en el transcriptoma de la especie rápidamente.

Funcionalmente, la evidencia recabada en los **Capítulos 3 y 4** indica que los diferentes SL-ARN expresados en platelmintos son incorporados en transcritos indistintamente de la identidad del gen aceptor. Los distintos SL-ARN estarían siendo incorporados en una proporción presumiblemente determinada por sus niveles de expresión en la célula y/o sus interacciones con el spliceosoma. La falta de especialización en transcritos objetivos probablemente representa la condición basal de cualquier grupo que adquiere SL *trans-splicing*. En el **Capítulo 4** se especula un mecanismo por el cuál la especialización de SL-ARN no ha ocurrido en platelmintos. En resumen, la alta tasa de conversión génica en loci SL-ARN previene la evolución independiente de los varios SL-ARN, limitando su divergencia nucleotídica y por extensión manteniéndolos funcionalmente equivalentes. Es necesario señalar sin embargo que un prolongado tiempo de evolución independiente no garantiza la emergencia de especialización funcional. Este parece ser el caso en el género Hymenolepis donde la divergencia nucleotídica entre loci SL-ARN indica una capacidad para evolucionar de forma independiente, pero donde no se encontró evidencia de especialización de los SL-ARNs. Aunque no se puede descartar cierta tasa de evolución concertada dado su agrupamiento en la filogenia.

A pesar de esta generalización en sus objetivos, las especies con múltiples SL-ARNs indican que sus proporciones relativas pueden ser variables. Este es el caso en *Schmidtea mediterranea* entre distintos tipos celulares (Rossia et al., 2014) y en *H. microstoma*. Aunque las discrepancias entre Calvelo et al. (2023) y Olson et al. (2020) llaman a la precaución. Las

diferencias entre estadios larvales y adultos en ambos estudios sugieren un origen biológico en el uso de SL-ARNs, pero la discrepancia en la proporción de SL-1 en sugiere artefactos metodológicos entre estudios. Ya sean diferencias en las condiciones de mantenimiento de los parásitos y sus hospedadores entre Preza et al. (2021) y Olson et al. (2020); o durante el procesamiento y secuenciación de las muestras ARN. Análogo a reportes *C. elegans* usando la plataforma Nanopore (Bernard et al., 2023), pero con un efecto diferencial entre distintos SL-ARNs. Dilucidar entre estas posibilidades es clave, aunque laborioso metodológicamente. Se requiere es la secuenciación transcriptómica de múltiples muestras del mismo estadio sometidas a distintas condiciones (ej: adultos colectados de ratones con distintas dietas) y procesados con diferentes plataformas de secuenciación y/o kits de extracción. Así como una secuenciación de los loci SL-ARN de los especímenes analizados.

Por último, a nivel de SL-ACEs y genes aceptores, la información recabada en platelmintos apunta a un proceso dinámico de ganancia y posible pérdida de sitios aceptores, posiblemente facilitado por la prevalencia de inserciones SLs facultativas. Evidenciados con la alta prevalencia de SL-ACEs con coexistencia de SL *trans-splicing* y *cis-splicing*. Una posible explicación, en línea con observaciones de errores en el *cis* y *trans-splicing* reportado en *C. elegans* (Tourasse et al., 2017) es que los SL-ARNs de platelmintos sean menos eficientes que en grupos como nemátodos. Facilitando de este modo la formación significativa de transcritos no sometidos a SL *trans-splicing* en estos sitios. Pero confirmar esta hipótesis requiere una mejor comprensión del spliceosoma en el grupo. La pérdida de SL-ACEs es posible por los mismos motivos y el patrón general sugiere han ocurrido en el ancestro común de cyclophyllidea, pero analizar casos particulares es difícil debido a la alta tasa de falsos negativos.

5.C) Perspectivas

SL *trans-splicing* en platelmintos es clave para entender su biología tanto a nivel genético, como evolutivo. En el primer caso por la caracterización de las isoformas alternativas posibles

y los mecanismos detrás de su transcripción y procesamiento en ARNm maduros. En el segundo por las dinámicas de formación, fragmentación y reordenamiento de loci policistrónicos. En nuestra perspectiva, los desafíos a afrontar para dar el siguiente paso en el estudio de este proceso en platelmintos parásitos son significativos. Requiriendo la re-secuenciación del transcriptoma de múltiples especies con metodologías especializadas, seguido del desarrollo de criterios adecuados para la identificación de sitios de relevancia biológica. Así como mejorar los estándares de anotación genómica en el grupo.

A nivel de especie, la caracterización del SL *trans-splicing* en profundidad ofrecerá una mejor imagen de su complemento genético y patrones de expresión. A medida que se expanda la cobertura de Cestodos, Trematodos y otros linajes de platelmintos, será posible estudiar las dinámicas evolutivas en mayor detalle. Por ejemplo, permitirá evaluar patrones de ganancia y pérdida de SL-ACEs sin el velo de falsos negativos que limitan la presente investigación. Así como analizar las presiones selectivas y procesos que gobiernan la diversificación y/o homogeneización de SL-ARN en platelmintos. Determinar en qué linajes predomina la conversión génica entre loci (ej: Schistosomatidae), en cuales su divergencia independiente (ej: *Hymenolepis sp.*) y en cuales ambos arreglos coexisten (ej: Taeniidae). Así como qué consecuencias tiene en la evolución de SL-ACEs. Preguntas como ¿La diversidad de SL-ARNs favorece la ganancia de nuevos SL-ACEs o la suprime? ¿Cómo afecta el número de loci SL-ARN funcionales la tasa de SL *trans-splicing*? ¿Cómo afecta el SL *trans-splicing* la expansión de elementos transponibles a grandes escalas evolutivas?

Ampliar los estudios en torno al SL *trans-splicing* en platelmintos ofrece oportunidades para ampliar el conocimiento del mecanismo en sí mismo. A diferencia de grupos como nemátodos, la ausencia de especialización en el uso de SL-ARNs sugiere que su regulación es comparativamente simplificada. Identificar los factores que sean exclusivos de SL *trans-splicing* en el grupo, o confirmar su inexistencia, puede ofrecer importantes pistas para la emergencia y regulación del mecanismo. Particularmente en lo referente a la restricción de

SL *trans-splicing* a sitios específicos. En este contexto, linajes como el género *Hymenolepis* con múltiples SLs distintivos codificados en pocos loci (4 SLs en 5 loci) ofrecen la oportunidad de analizar la importancia funcional de distintas regiones del SL-ARN analizando mutantes. Pero primero es necesario solventar las dificultades prácticas de identificar, aislar, criar y potencialmente generar mutantes de cestodos.

5.D) Referencias bibliográficas

Bernard, F., Dargère, D., Rechavi, O., & Dupuy, D. (2023). Quantitative analysis of C. elegans transcripts by Nanopore direct-cDNA sequencing reveals terminal hairpins in non trans-spliced mRNAs. *Nature Communications*, *14*(1), 1229. https://doi.org/10.1038/s41467-023-36915-0

Bitar, M., Boroni, M., Macedo, A. M., Machado, C. R., & Franco, G. R. (2013). The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. *Frontiers in Genetics*, *4*(October), 199. https://doi.org/10.3389/fgene.2013.00199

Brehm, K., Jensen, K., & Frosch, M. (2000). mRNA trans-splicing in the human parasitic cestode Echinococcus multilocularis. *Journal of Biological Chemistry*, *275*(49), 38311–38318. https://doi.org/10.1074/jbc.M006091200

Calvelo, J., Brehm, K., Iriarte, A., & Koziol, U. (2023). Trans-splicing in the cestode Hymenolepis microstoma is constitutive across the life cycle and depends on gene structure and composition. *International Journal for Parasitology*, *53*(2), 103–117. https://doi.org/10.1016/j.ijpara.2022.11.006

Calvelo, J., Juan, H., Musto, H., Koziol, U., & Iriarte, A. (2020). SLFinder, a pipeline for the novel identification of splice-leader sequences: a good enough solution for a complex problem. *BMC Bioinformatics*, *21*, 293. https://doi.org/10.1186/s12859-020-03610-6

Davis, R. E., Singh, H., Botka, C., Hardwick, C., El Meanawy, M. A., & Villanueva, J. (1994). RNA trans-splicing in Fasciola hepatica. Identification of a spliced leader (SL) RNA and SL sequences on mRNAs. *Journal of Biological Chemistry*, *269*(31), 20026–20030.

Lasda, E. L., & Blumenthal, T. (2011). Trans-splicing. *Wiley Interdisciplinary Reviews: RNA*, *2*(3), 417–434. https://doi.org/10.1002/wrna.71

Matsumoto, J., Dewar, K., Wasserscheid, J., Matsumoto, J., Dewar, K., Wasserscheid, J., Wiley, G. B., Macmil, S. L., Roe, B. A., Zeller, R. W., Satou, Y., & Hastings, K. E. M. (2010). High-throughput sequence analysis of Ciona intestinalis SL trans-spliced mRNAs: alternative expression modes and gene function correlates. *Genome Research*, *20*(5), 636–645. https://doi.org/10.1101/gr.100271.109

Olson, P., Tracey, A., Baillie, A., James, K., Doyle, S., Buddenborg, S., Rodgers, F., Holroyd, N., & Berriman, M. (2020). Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection. *BMC Biology*, *18*, 165. https://doi.org/10.1101/2020.04.08.031872

Preza, M., Calvelo, J., Langleib, M., Hoffmann, F., Castillo, E., Koziol, U., & Iriarte, A. (2021). Stage-specific transcriptomic analysis of the model cestode Hymenolepis microstoma. *Genomics*, *113*(2), 620–632. https://doi.org/10.1016/j.ygeno.2021.01.005

Rajkovic, A., Davis, R. E., Simonsen, J. N., & Rottman, F. M. (1990). A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni. *Proceedings of the*

National Academy of Sciences of the United States of America, *87*(22), 8879–8883. https://doi.org/10.1073/pnas.87.22.8879

Rossia, A., Jackb, E. J. R. A., & Alvarado, A. S. (2014). Molecular cloning and characterization of SL3: A stem cell- specific SL RNA from the planarian Schmidtea mediterranea. *Gene*, *533*(1), 156–167. https://doi.org/doi:10.1016/j.gene.2013.09.101. Molecular

Tourasse, N. J., Millet, J. R. M., & Dupuy, D. (2017). Quantitative RNA-seq meta-analysis of alternative exon usage in C. elegans. *Genome Research*, *27*(12), 2120–2128. https://doi.org/10.1101/gr.224626.117

Tsai, I. J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K. L., Tracey, A., Bobes, R. J., Fragoso, G., Sciutto, E., Aslett, M., Beasley, H., Bennett, H. M., Cai, J., Camicia, F., Clark, R., Cucher, M., De Silva, N., Day, T. A., … Valdes, V. (2013). The genomes of four tapeworm species reveal adaptations to parasitism. *Nature*, *496*(7443), 57–63. https://doi.org/10.1038/nature12031

Wenzel, M. A., Müller, B., & Pettitt, J. (2021). SLIDR and SLOPPR: flexible identification of spliced leader trans-splicing and prediction of eukaryotic operons from RNA-Seq data. *BMC Bioinformatics*, *22*, 140. https://doi.org/10.1186/s12859-021-04009-7

Agradecimientos

Quiero agradecer a mis tutores y compañeros de trabajo por su apoyo y guía realizando esta tesis. A la "Agencia Nacional de Investigación e Innovación" y la "Comisión Asesora de Posgrado" por su apoyo financiero. Pero más que nada a mi familia. Sin sus sacrificios y paciencia jamás habría llegado tan lejos.