Tesis para la obtención del título de Magíster en Bioinformática

---

Caracterización de elementos transponibles en Fasciolidae

Estudiante:
Lic. Agustín Felipe Bilat Damasco

Tutor y co-tutor:
Dr. José F. Tort y Dr. Anna V. Protasio

PEDECIBA – Universidad de la República

Montevideo, Diciembre de 2024

Agradecimientos

A todas las personas que compartieron parte de esta estapa conmigo:

A mis tutores Anna y Pepe por las enseñanzas, la guía, la paciencia, el buen humor y la oportunidad de compartir esta etapa de mi carrera con ellos.

A Gabriel y Santiago por entusiasmarse con el proyecto de "Schisto", lo cual me va a permitir continuar especializándome en el área de la parasitología.

Al Laura, Fernando ("el Pelo") y Sebastián por aceptar conformar el tribunal de tesis.

A todos los compañeros y compañeras del Departamento de Genética, por hacer de este un espacio laboral humano, ameno y motivante.

A mis estudiantes, por enseñarme e inspirarme a ser un mejor docente.

A los amigos de la vida. Especialmente a Juan, Leandro y Stephy.

A mis hermanos Pedro, José y Victoria. A mis sobrinos Emi y Pieri. A mi tía, Felipe y al resto de los primos. Al abuelo, a la abuela y a mi padre, a quienes guardamos en el corazón. Muy especialmente, a mi madre. Gracias a todos por el apoyo y el amor de siempre.

A todas las instituciones que me apoyaron y me formaron profesionalmente:

A la UdelaR y la Comisión Académica de Posgrado "CAP" por financiar mi Maestría. También al resto de agencias financiadoras de investigación del país. A la University of Cambridge, por permitirme usar su servidor. Al Programa de Desarrollo de las Ciencias Básicas (PEDECIBA), y particularmente a PEDECIBA Bioinformática. A todas las personas e instituciones que componen el sistema científico nacional.

A todos de corazón, gracias!!

Lista de abreviaturas

ADN	Ácido Desoxirribonucleico
AP	Aspartic Proteinase
APE	Apurinic Endonuclease
ARN	Ácido Ribonucleico
EN (giy-yig)	Endonuclease with giy-yig motif
F. hepatica	*Fasciola hepatica*
F. gigantica	*Fasciola gigantica*
F. buski	*Fasciolopsis buski*
INT	Integrase
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
NEJ	Newly Excysted Juvenile
PLE	Penelope Like Element
RT	Reverse Trasncriptase
RH	RNase H
SINE	Short Interspersed Nuclear Element
Tase (DDE)	Transposase with DDE motif
TE	Transposable Element
TIR	Terminal Inverted Repeat
TSD	Target Site Duplication

Índice de contenidos

Capítulo 1.

Introducción

1.1 Fascioliasis

En el año 1379 el pastor francés Jean de Brie identificó a *Fasciola hepatica* como el agente causal de la "putrefacción del hígado" en ovejas, siendo así el primer trematodo en ser descrito. La fascioliasis es la enfermedad causada por la infección con este parásito o con *Fasciola gigantica*, la cual presenta una alta prevalencia en mamíferos rumiantes. Por este motivo tiene un fuerte impacto sobre el sector productivo, causando pérdidas anuales que oscilan en el entorno de los 3.200 millones de dólares a nivel global (Carmona & Tort, 2017; Mehmood et al., 2017). En Uruguay, se ha estimado que la seroprevalencia de *Fasciola* en el ganado es mayor al 50% (Sanchís et al., 2015), constituyendo asimismo un importante problema económico para otros países de sudamérica (Carmona & Tort, 2017).

Esta enfermedad tiene también efectos perjudiciales para la salud humana, presentando una distribución cosmopolita (Figura 1.1), y siendo reconocida por la Organización Mundial de la Salud como una zoonosis emergente en 51 países. Esta gran dispersión de la enfermedad es mayormente provocada por *F. hepatica* —lo que posiblemente se vincule a su capacidad para infectar un mayor número de hospederos— en tanto que *F. gigantica* parece presentar una distribución geográfica más restringida (Carmona & Tort, 2017). Entre las poblaciones más afectadas por la fascioliasis se encuentran los habitantes del altiplano boliviano y de otras regiones andinas (Mas-Coma et al., 2020). A nivel global, las estimaciones más moderadas dan cuenta de unas 2.4 millones de personas infectadas (Carmona & Tort, 2017). La morbilidad causada por esta infección es también notoria, presentando un impacto anual estimado entre 35 a 90 mil años de vida ajustados por discapacidad (AVAD) (Lalor et al., 2021; Torgerson et al., 2015).

Las manifestaciones clínicas y la gravedad de la fascioliasis varían no solo debido a la carga parasitaria, sino también en función de factores como el hospedero infectado —ganado bovino u ovino, o humanos— y de la fase de la infección. La fase aguda coincide con la migración del parásito inmaduro a través del parénquima hepático, mientras que la crónica se asocia al establecimiento del adulto en los conductos biliares, en donde deposita miles de huevos cada día (ver descripción del ciclo de vida más abajo) (Caravedo & Cabada, 2020; Carmona & Tort, 2017; Cwiklinski et al., 2021; González-Miguel et al., 2021; Lalor et al., 2021).

Existen, principalmente, dos estrategias generales para el control de esta zoonosis. La primera incluye todas aquellas vinculadas a la prevención de la infección, principalmente basadas en eliminar o limitar la dispersión de los moluscos que actúan como vectores de los gusanos, así como las campañas de higiene y concientización sobre el riesgo asociado al consumo de plantas acuáticas crudas en las zonas endémicas. La segunda (posiblemente la más extendida), se basa en la administración de drogas anti-helmínticas a los hospederos mamíferos. Triclabendazole es el principal agente quimioterapéutico utilizado para combatir la fascioliasis, si bien existen otras drogas como el albendazole que han mostrado ser efectivas (Carmona & Tort, 2017). Sin embargo, es importante destacar que se han reportado múltiples focos de emergencia de parásitos resistentes a estas quimioterapias, lo cual resulta particularmente preocupante (Carmona & Tort, 2017). Por otra parte, resulta llamativo que la droga Praziquantel sea inutil para tratar la fascioliasis teniendo encuenta su eficacia para combatir otras trematodiasis como la schistosomiasis e incluso la fasciolopsiasis (provocada por un parásito de la misma familia que *Fasciola*) (Siles-Lucas et al., 2021).

La caracterización de distintas biomoléculas importantes para la infección y establecimiento de estos parásitos (por ejemplo, las catepsinas) constituyen blancos potenciales para su tratamiento, lo que supone una alternativa (o una herramienta de control adicional) a las quimioterapias clásicas. En este sentido, se han realizado ensayos prometedores basados en el uso de estas biomoléculas para el desarrollo de vacunas, si bien hasta el momento presentan grados variables de efectividad en función del hospedero, el uso de proteína nativa o recombinante, entre otros factores (Carmona & Tort, 2017). Dada la relevancia de estos parásitos, así como los desafíos asociados a la aparición de resistencia, es fundamental continuar aportando al conocimiento sobre la biología de estos organismos.

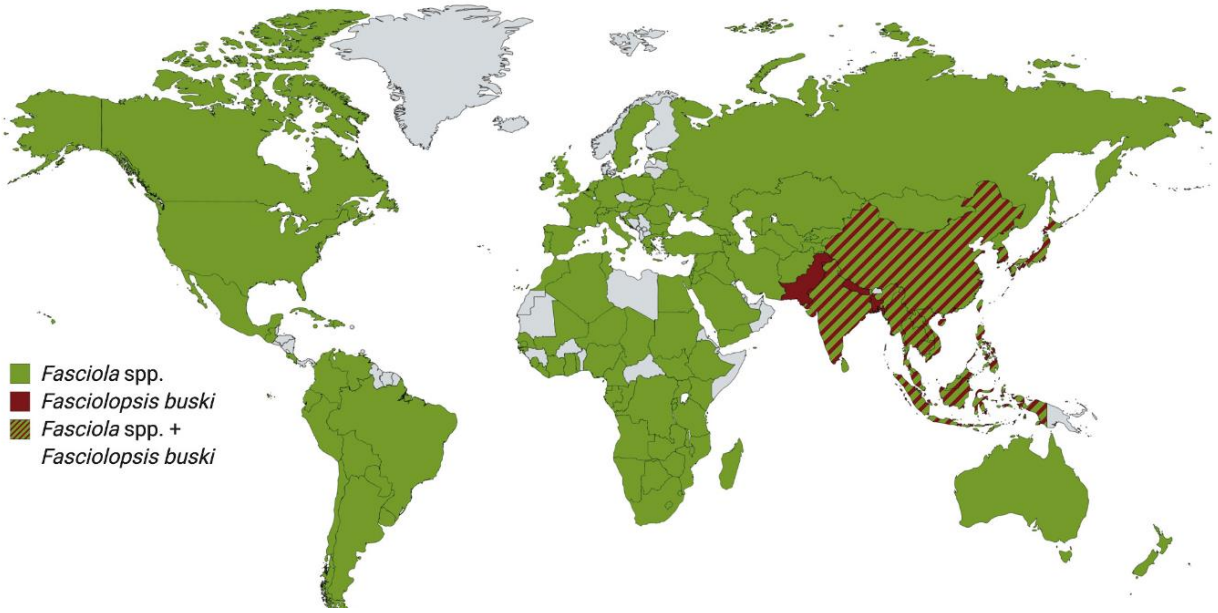


Figura 1.1 | **Distribución geográfica de Fascioliasis y Fasciolopsiasis en humanos.** En color verde se indican las regiones donde existen reportes de infección en humanos por *F. hepatica* o *F. gigantica*, en color bordó aquellas donde únicamente hay reportados casos de infecciones con la especie emparentada *F. buski*, y con ambos colores aquellas donde ambas enfermedades han sido reportadas.

1.2 Ciclo de vida de *Fasciola* spp.

Las especies del género *Fasciola* presentan un ciclo de vida heteroxeno, en el cual es necesario que infecten a dos hospederos para completar su desarrollo (Figura 1.2). Los rumiantes, o en menor frecuencia otros mamíferos (incluyendo al humano y a roedores como el carpincho) actúan de hospederos definitivos al alojar a las formas sexualmente maduras del gusano. Los hospederos mamíferos liberan los huevos del parásito al ambiente junto con las heces (Fig. 1.2A), eclosionando embriones ciliados conocidos como miracidios tras su contacto con un cuerpo de agua dulce. Utilizando reservas de glucógeno, el miracidio debe alcanzar rápidamente al hospedero intermediario (un caracol de la familia *Lymnaeidae*) para poder sobrevivir (Fig. 1.2B) (Alba et al., 2019; Lalor et al., 2021). Para ello, la larva nada hacia el caracol atraída por estímulos quimio-sensoriales, adhiriéndose y penetrando en el mismo con la ayuda de enzimas digestivas (Cruz-Mendoza et al., 2006; González-Miguel et al., 2021; Lalor et al., 2021). En el interior del

caracol pierde la cubierta ciliada, transformándose en un saco de células germinativas conocido como esporocisto, que dará orígen a sucesivas generaciones de redias. Al final de esta etapa de multiplicación asexual al interior del hospedero intermediario, se liberan al agua un gran número de cercarias (Fig. 1.2C), las cuales propulsadas por una cola natatoria se terminan por enquistar en la vegetación actuática (González-Miguel et al., 2021). La metasarcaria infectiva es eventualmente ingerida junto con la vegetación por el hospedero definitivo (Fig. 1.2D), desenquistandose en el duodeno, donde se libera el primer estadio larvario exclusivo de la etapa intra-mamífero, conocido como NEJ (por la sigla de *Newly Exysted Juvenile*) (Cwiklinski et al., 2021; González-Miguel et al., 2021). Esta forma larvaria no es capaz de sobrevivir mucho tiempo en la luz del digestivo, por lo que comienza una "peregrinación" dentro del hospedero hasta su localización definitiva. En las primeras horas post-desenquiste atraviesa la pared intestinal (Fig. 1.2E), y migra hacia el hígado al cuál ingresa un par de días post-infección atravesando la cápsula de Glisson. El juvenil migra luego durante varias semanas a través del parénquima hepático (Fig. 1.2F), al tiempo que incrementa su tamaño y desarrolla las estructuras reproductoras y digestivas (Cwiklinski et al., 2021; González-Miguel et al., 2021; Moazeni & Ahmadi, 2016). El gusano alcanza su madurez sexual al llegar a los canalículos biliares (Fig. 1.2G), donde comienza a producir miles de huevos diarios, que son finalmente liberados al ambiente, comenzando así con un nuevo ciclo.

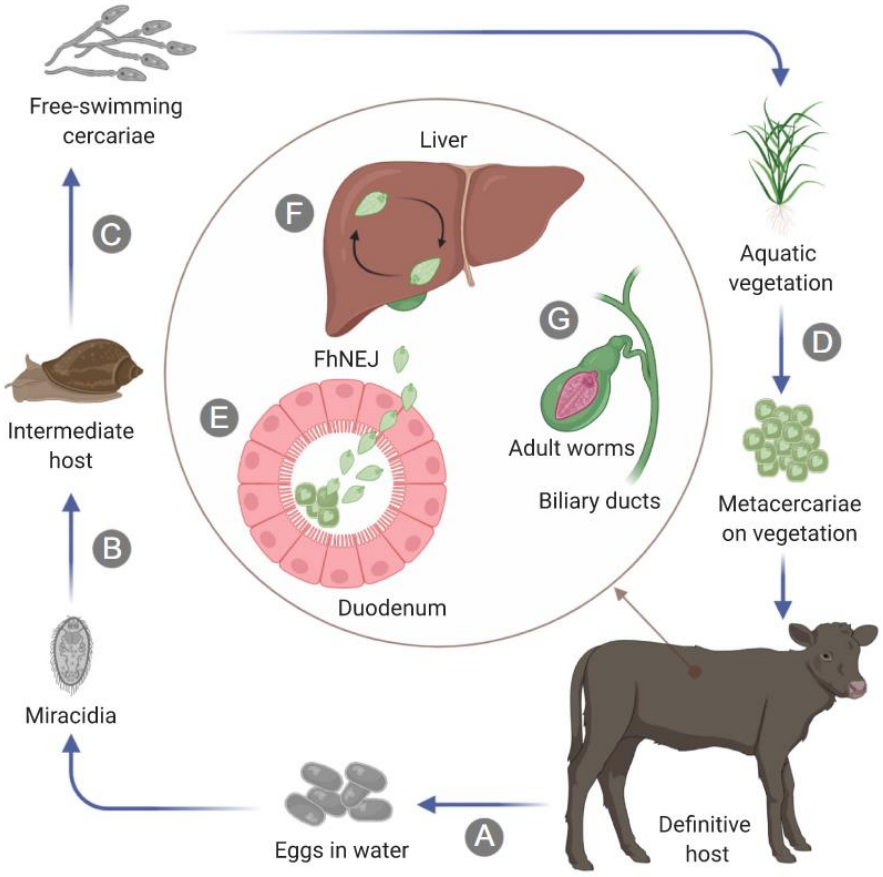


Figura 1.2 | **Ciclo de vida del parásito hepático *Fasciola* spp.** La descripción del ciclo se detalla en el cuerpo del texto. Figura tomada de (González-Miguel et al., 2021).

1.3 La familia Fasciolidae

Fasciolidae es una pequeña familia de trematodos digenéticos conformada por nueve especies descritas, tres de las cuales infectan a humanos: *F. hepatica*, *F.* gigantica y *Fasciolopsis buski* (Siles-Lucas et al., 2021). A partir de secuencias de ADN ribosomal y de un gen mitocondrial, se estudiaron las relaciones filogenéticas de siete especies de fasciólidos, incluyendo a *Echinostoma* como grupo externo (Figura 1.3) (Lotfy et al., 2008). Lo anterior permitió determinar que las innovaciones adaptativas más prominentes dentro de Fasciolidae ocurrieron en el ancestro común de *Fasciola* y *Fascioloides* —subfamilia Fasciolinae—, luego de la divergencia con el linaje que dio lugar al género *Fasciolopsis*. Uno de los cambios ocurridos tiene que ver con la localización del parásito adulto dentro del hospedero definitivo: los Fasciolinae son hepáticos pero *F. buski* y las otras especies de la familia son intestinales. Además, las especies hepáticas utilizan como hospedero intermediario a caracoles de la familia Lymnaeidae mientras que las especies intestinales infectan caracoles de la familia Planorbidae. Por lo demás, los ciclos de vida son esencialmente iguales (Figura 1.2), salvo por el tipo de hospedero definitivo que cada fasciólido es capaz de infectar, y por la distribución geografica de estos patógenos (Figura 1.3). Otros rasgos derivados compartidos por las especies hepáticas incluyen una mayor ramificación del intestino, testículo y ovarios, y un menor tamaño relativo de la ventosa ventral en comparación con la ventosa oral (Lotfy et al., 2008). Estas innovaciones morfológicas podrían tener que ver con el cambio en la localización definitiva del adulto desde el intestino hacia el hígado.

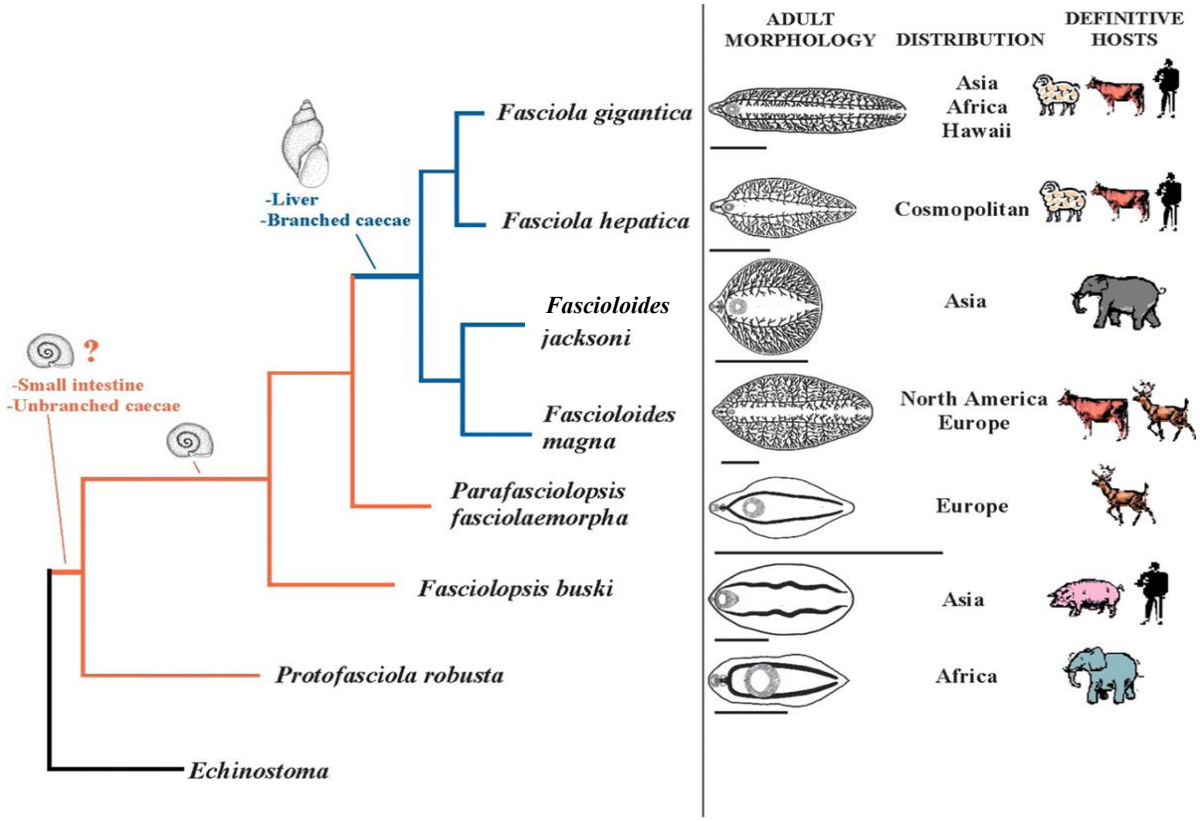


Figura 1.3 | **Historia evolutiva de la familia Fasciolidae.** Filogenia de Fasciolidae. El azul corresponde a la subfamilia Fasciolinae y el naranja indica las especies intestinales. Las flechas muestran las innovaciones evolutivas más notorias. En el panel derecho se muestra un esquema con la morfología y tamaño de los gusanos adultos, así como la distribución geográfica y hospederos definitivos reportados. Imagen tomada y modificada a partir de otro artículo (Lotfy et al., 2008)

¿Qué cambios a nivel del ADN dieron lugar a las sinapomorfías que caracterizan a las especies hepáticas? Si bien estamos lejos de responder a dicha pregunta, la disponibilidad —en años recientes— de los ensamblados genómicos de tres especies de Fasciolidae (Choi et al., 2020; Cwiklinski et al., 2015; Luo et al., 2021; McNulty et al., 2017; Pandey et al., 2020) nos permite comenzar a investigar exhaustivamente las diferencias moleculares al interior de esta familia. En este sentido, en el primer estudio comparado de estas especies se observó un gran incremento del tamaño genómico de *F. hepatica* y *F. gigantica* en relación a *F. buski*, pero no se observaron cambios sustantivos en el número de genes. Sin embargo, sí se apreció un mayor número de copias de ciertas familias multigénicas relevantes para el estilo de vida parasitario de estos helmintos (Choi et al., 2020). Más aún, se observó que el incremento del tamaño genómico en las especies hepáticas se debe a un considerable incremento del ADN repetitivo, y que dicho aumento es causado mayormente por la expansión de elementos transponibles (Choi et al., 2020; Luo et al., 2021). Sin embargo, la caracterización detallada de la fracción repetitiva de los genomas de estos organismos sigue siendo todavía un área muy poco explorada.

1.4 Elementos transponibles

1.4.1 Clasificación de TEs

Durante los años 40s y 50s del siglo pasado, la genetista estadounidense Bárbara McClintock demostró que ciertos elementos genéticos eran capaces de cambiar su posición cromosómica en el maíz (*Zea mayz*). Actualmente, sabemos que estas secuencias de ADN móviles denominadas elementos transponibles (TEs por su abreviación del inglés), están presentes en prácticamente todos los eucariotas y han tenido un rol importante en la evolución del tamaño genómico de los mismos (Chénais et al., 2012).

Los TEs se dividen en elementos de **clase I** (retrotransposones), cuya replicación requiere de un intermediario de ARN, y en elementos de **clase II** (transposones de ADN), los cuales no utilizan un intermediario de ARN para movilizarse (Bourque et al., 2018; Wells & Feschotte, 2020). En función del mecanismo específico de replicación e inserción utilizado, los retrotransposones se subdividen en las categorías LINE, LTR, PLE, entre otras; mientras que los transposones de ADN se conforman principalemnte por los tipos TIR, Crypton y Helitron (Wells & Feschotte, 2020; Wicker et al., 2007). Dichas categorías (denominadas órdenes o subclases, dependiendo del autor) son a su vez filogenéticamente clasificadas en distintas superfamilias. A un nivel más basal se define la familia, que corresponde a un conjunto de copias (inserciones) dispersas en el genoma y derivadas de la amplificación de una secuencia progenitora ancestral. Operativamente, una familia es usualmente definida por la regla “80-80-80”, la cual establece que dos inserciones de TE pertenecen a la misma familia si comparten al menos 80 pb y al menos 80% de identidad de secuencia sobre no menos del 80% del largo de la secuencia más corta (Wicker et al., 2007). La figura 1.4 resume el sistema de clasificación de TEs.

Los TEs pueden ser también clasificados en función de su capacidad de transposición (Wells & Feschotte, 2020; Wicker et al., 2007). Los elementos **autónomos** son aquellos que codifican todas las proteínas necesarias para movilizarse por sí mismos. Los elementos **no-autónomos** no producen todas las proteínas necesarias para su transposición. Estos últimos dependen de proteínas producidas por elementos autónomos para movilizarse. En la sección de resultados se especifican los criterios utilizados en el presente trabajo para clasificar una familia de TEs como autónoma o no-autónoma.

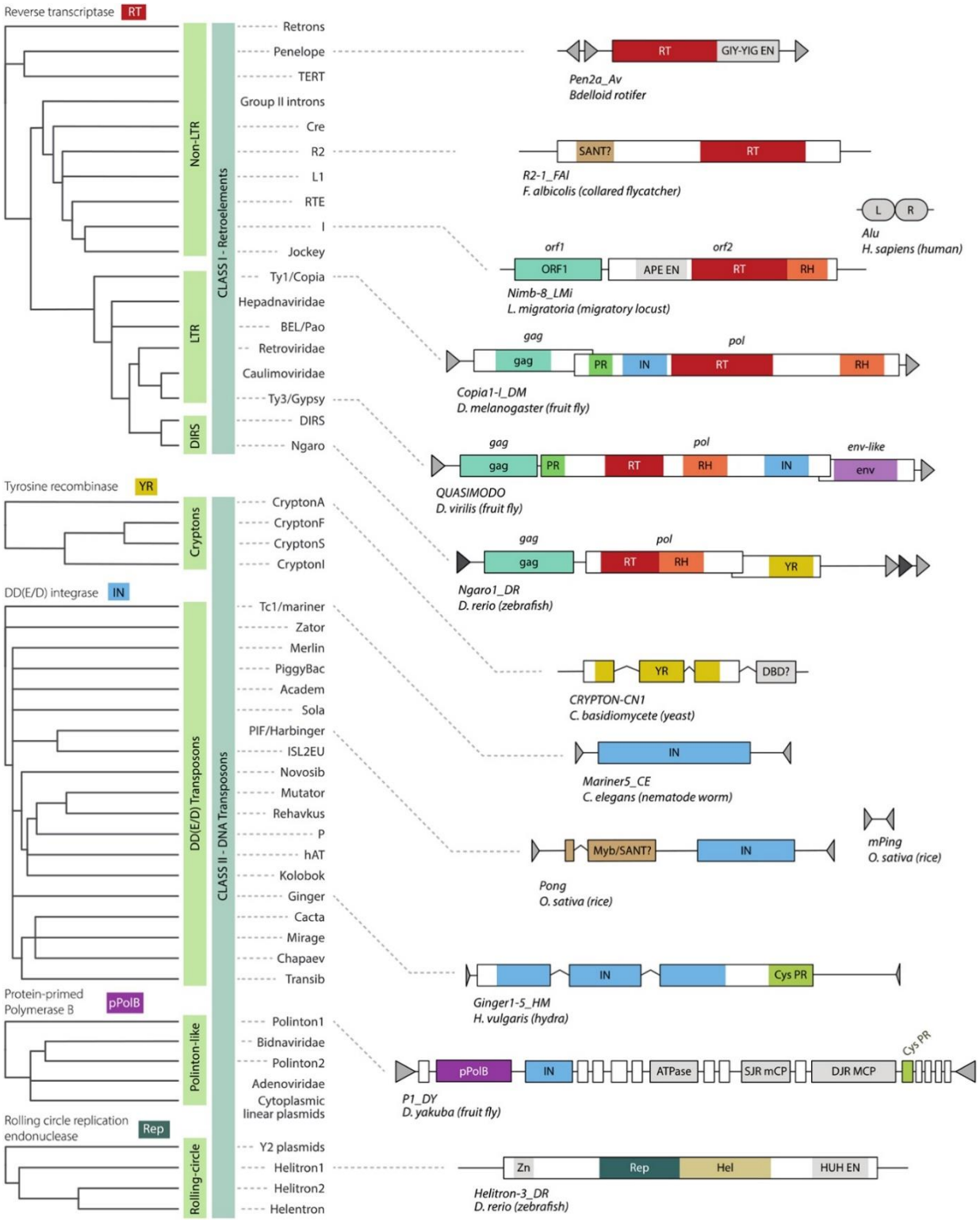


Figura 1.4 | **Estructura y clasificación de TEs.** Se muestran del lado izquierdo cladogramas de transposones con las presuntas relaciones entre distintas superfamilias, lo que se obtuvo a partir de la transcriptasa reversa de los elementos de clase I, y de distintas proteínas presentes en distintas subclases (u órdenes) de elementos de clase II. A la derecha se muestran esquemas de la estructura y dominios proteicos representativos. Los elementos indicados como "non-LTR" incluyen los elementos LINE y PLE referidos en distintas partes del texto. Las abreviaciones se detallan en el artículo original de donde extrajo la figura (Wells & Feschotte, 2020).

1.4.2 Descubrimiento y anotación de TEs

Existen tres estrategias principales para identificar elemetos transponibles a partir de un ensamblado genómico (J. Storer et al., 2022). La identificación *de novo* (*ab initio*) consiste en descubrir familias de TEs explotando la naturaleza repetitiva de estos elementos móviles, lo cual no requiere conocimiento previo de su estructura o secuencia. Un segundo método (*signature-based*) consiste en identificar dominios o características estructurales conocidas de determinados tipos de TEs. Por último, los métodos basados en el uso de bibliotecas (*library-based*) consisten en buscar secuencias en el genoma que sean similares (homólogas) a otro conjunto de secuencias de referencia de TEs. Esta estrategia se usa típicamente para anotar las coordenadas genómicas de las inserciones de TEs a partir de la biblioteca de repetidos (la "base de datos"), usualmente mediante la herramienta RepeatMasker.

La anotación basada enteramente en homología (utilizando directamente una biblioteca de TEs de un taxón cercano a la especie en cuestión) permite tener una aproximación del contenido de repetidos en el nuevo genoma, pero puede dar lugar a errores en la estimación de la diversidad de secuencias de TE presentes en dicho genoma (Platt et al., 2016). Más aún, si no se dispone de una base de datos de repetidos de una especie cercana al organismo de interés, el uso exclusivo de este método puede dar como resultado un panorama alejado de la representación real del contenido genómico de repetidos en el genoma (Platt et al., 2016).

Por lo anterior, en nuevos genomas para los cuales no hay buenas bases de datos de repetidos cercanas a la especie, suele ser preferible realizar en primer lugar la identificación de familias de repetidos mediante herramientas "generalistas" tales como RepeatModeler (Flynn et al., 2019), EDTA (Ou et al., 2019) o REPET (Flutre et al., 2011). Estas herramientas generan de forma automática una biblioteca de repetidos a partir del propio ensamblado genómico. Esto se realiza normalmente integrando varios algoritmos en un único *pipeline* que permite obtener una representación relativamente exaustiva de las familias de TEs presentes en el ensamblado analizado. También suelen realizar una clasificación automática, entre otras utilidades. Finalmente, la biblioteca generada se suele usar para la anotación genómica con RepeatMasker.

Si bien esta estrategia permite una caracterización razonable del contenido de repetidos en un nuevo genoma, las bibliotecas automáticamente generadas se componen en buena parte de modelos imperfectos e incompletos de las familias de repetidos. Por lo tanto, una anotación más precisa requiere de un proceso de curación manual. Esto permite obtener una representación más realista de las familias de TEs presentes en el genoma analizado, eliminandose también secuencias redundantes, y mejorando la caracterización y anotación genómica de los repetidos (Goubert et al., 2022; Platt et al., 2016; J. M. Storer et al., 2021).

1.4.3 Elementos transponibles en Fasciolidae

McClintock demostró que al movilizarse, los TEs activaban o apagaban genes cercanos. Esta idea de los TEs como elementos con un impacto a nivel de la regulación de la expresión génica tiene cada vez más adeptos (Chuong et al., 2017). Más aún, se acepta que los TEs pueden generar duplicaciones génicas (Cerbin & Jiang, 2018), alterar de diversas formas la arquitectura y función de los genomas (Bourque et al., 2018), e impactar en el desarrollo de los organismos (Senft & Macfarlan, 2021). Motivados por la acumulación de evidencia relativa al potencial impacto funcional y evolutivo de los TEs, así como por los trabajos que sentaron precedentes sobre el contenido de TEs en los genomas de fasciólidos (Choi et al., 2020; Luo et al., 2021), nos propusimos como objetivo general realizar una caracterización detallada de los elementos transponibles en Fasciolidae.

El presente trabajo se diferencia por seguir una estrategia basada en la curación manual (Goubert et al., 2022) de las familias de TEs identificadas a partir de los genomas de Fasciolidae. Esto nos permitió realizar una caracterización detallada y analizar en mayor profundidad las diferencias en el contenido de TEs de las distintas especies. Más aún, la inclusión de ensamblados de lecturas cortas y lecturas largas en el análisis nos permitió ponderar mejor posibles sesgos metodológicos a la hora de analizar las diferencias en el contenido genómico de TEs entre las distintas especies.

Nuestro trabajo confirmó la existencia de muchas olas de transposición involucrando distintas superfamilias de TEs durante la evolución de Fasciolidae. También pudimos identificar familias de TEs putativamente activas, así como confirmar que distintos tipos de TEs presentan una distribución heterogenea con respecto a los intrones de genes codificantes. El presente trabajo es un primer paso firme hacia la exploración del papel evolutivo y funcional de los TEs en este grupo de trematodos con relevancia para la salud humana y animal.

1.5 Hipótesis de trabajo

Las expansión de elementos transponibles ha jugado un papel importante en la evolución de la familia Fasciolidae, y podría estar asociada a las innovaciones adaptativas más notorias que distinguen a las especies hepáticas de las intestinales.

1.6 Objetivos

1.6.1 Objetivo General

- Caracterizar detalladamente el contenido de TEs en fasciólidos hepáticos e intestinales.

1.6.2 Objetivos Específicos

- Identificar, curar y caracterizar las familias de TEs a partir de los ensamblados genómicos de las especies *F. hepatica*, *F. gigantica* y *F. buski*.

- Estimar la composición genómica (cantidad de copias y cobertura) de las familias de TEs caracterizadas.

- Analizar la dinámica evolutiva de TEs en estas especies.

- Estudiar el impacto de los TEs a nivel de los genes codificantes de proteínas.

Capítulo 2.

Materiales y métodos

2.0 Convenciones usadas en el texto y disponibilidad de los datos generados

Los archivos y tablas suplementarias están disponibles en: https://onedrive.live.com/?id=C3D95EDB1FD82FAA%21s6f521ca551f1455dbc7a6b1157dc5e36&cid=C3D95EDB1FD82FAA. Estos se especifican en el texto bajo el nombre en inglés: "Supplementary File X" o "Supplementary Table X", donde X representa un número entero.

Los ***scripts*** que generamos se encuentran disponibles en: https://github.com/agustin-bilat/Bilat2024_TEs-Fasciolidae. Estos se referencian en el manuscrito usando la fuente "`courier new`" resaltada en color gris. El resto de programas ejecutados y sus parámetros se identifican en el texto usando la fuente "`courier new`" pero sin resaltar con ningún color.

2.1 Obtención de ensamblados y datos de anotación de genes

Los ensamblados genómicos y archivos de anotación de genes utilizados en el presente trabajo se obtuvieron a partir de bases de datos públicas (CNCB-NGDC Members and Partners et al., 2024; Howe et al., 2017) (Tabla 2.1).

Las secuencia cromosómica del ensamblado de *Fasciola gigantica* no se encontraba públicamente disponible de modo que se obtuvo descargando el ensamblado de los *contigs* (CNCB-NGDC Members and Partners et al., 2024), así como los datos de secuenciación de captura de conformación de los cromosomas (Hi-C) almacenados en un archivo con formato AGP (Luo et al., 2021). Este último fue editado para eliminar las filas correspondientes a los *contigs* con nombres "ctg00025", "ctg00369", "ctg00642", "ctg00791", "ctg00877", "ctg00946" y "ctg01016". También se agregó un *contig* artificial (ctg00565) compuesto por una secuencia de 2,566 nucleótidos genéricos con la letra "N" en el archivo FASTA, ya que el mismo está incluido –de acuerdo al archivo AGP– en el cromosoma denominado como "group1". A partir del archivo AGP editado y el ensamblado con los *contigs* se obtuvo el ensamblado a nivel cromosómico mediante el comando `assemble` del paquete `agptools` (https://github.com/WarrenLab/agptools). Luego, con el comando `sed` de Linux se modificó el prefijo "group" a "chr" en los *headers* del archivo FASTA. Sobre este ensamblado (Supplementary File 1 - Fgig_2.fna), al que denominamos *Fgig_2*, se realizó la identificación y anotación de los repetidos.

Para los cuatro ensamblados restantes, se tomaron las secuencias en formato FASTA tal y como se encuentran públicamente disponibles (Tabla 2.1). El valor **L90** del ensamblado *Fhep_2* equivale a 14 *scaffolds*, indicando que los mismos están también a nivel cuasi-cromosómico. Los otros ensamblados, basados principalmente en lecturas cortas (*Fhep_1*, *Fgig_1* y *Fbus*), están más fragmentados pero presentan parámetros de calidad similares entre ellos (Tabla 2.1).

Tabla 2.1 | **Fuente de datos y estadísticas asociadas a los ensamblados de fasciólidos.**

Species	*Fasciola hepatica*	*Fasciola hepatica*	*Fasciola gigantica*	*Fasciola gigantica*	*Fasciolopsis buski*
Data Source	WBPS19	WBPS19	WBPS19	GWH (CNCB-NGDC)	WBPS19
Assembly name	F_hepatica_1.0.allpaths.pg	FhHiC23	F_gigantica_1.0.allpaths	GXU_Fgig_PAC.1.0	F_buski_1.0.allpaths-lg
BioProject ID	PRJNA179522	PRJEB58756	PRJNA230515	PRJNA691688	PRJNA284521
GeneBank accession	GCA_002763495.2	GCA_948099385.1	GCA_006461475.1	GCA_018104335.1	GCA_008360955.1
GWH accession	-	-	-	GWHAZTT00000000	-
Sequencing methods	Illumina HiSeq	Pacbio Sequel II; Hi-C	Illumina HiSeq; PacBio RSII	Pacbio Sequel II; Hi-C	Illumina HiSeq; PacBio RSII
Assemblies' abbreviations used by us	*Fhep_1*	*Fhep_2*	*Fgig_1*	*Fgig_2*	*Fbus*
Used as input into RepeatModeler	NO	YES	NO	YES	YES
Total genome size (Gbp)	1.14	1.51	1.13	1.35	0.75
Number of scaffolds*	23,604	573	16,040	**10**+24	11,829
Scaffold L90	8,040	14	6,749	8	4,701
Scaffold N50 (Mbp)	0.161	131.5	0.179	133	0.180
Number of contigs	95,033	752	76,748	1,022	31,948
Contig N50 (Mbp)	0.027	9.8	0.027	4.9	0.051
Coding genes	11,217	16,865	12,669	12,503	11,837
GC	44.1%	44.1%	44.1%	44.0%	42.6%
BUSCO (complete)	66.5%	71.1%	62.1%	83.1%	61.2%
BUSCO (fragmented)	8.2%	5.7%	7.7%	6.9%	10.9%

*en **negrita** se indican los scaffolds a nivel cromosómico del ensamblado "*Fgig_2*"

2.2 Búsqueda *de novo* y curación de los repetidos

Los ensambldos que denominamos *Fbus* (*Fasciolopsis buski*), *Fgig_2* (*Fasciola gigantica*) y *Fhep_2* (*Fasciola hepatica*) (Tabla 2.1) fueron utilizados por separado en tres ejecuciones del programa `RepeatModeler 2.0.2` con el parámetro `"-LTRStruct"` (Flynn et al., 2019). Las bibliotecas de secuencias consensos de repetidos que se obtuvieron tras la corrida, fueron sometidas a un proceso de curación manual. La curación se basó en las indicaciones generales reportadas en un trabajo previo (Goubert et al., 2022) según las especificaciones indicadas en el repositorio asociado a este trabajo: https://github.com/agustin-bilat/Bilat2024_TEs-Fasciolidae. A continuación se describe brevemente este proceso, llevado a cabo en forma separada (en paralelo) a partir de cada una de las tres bibliotecas de consensos obtenidas automáticamente. Los principales pasos se esquematizan en la Figura 2.1. Primero, las secuencias redundantes se agruparon utilizando `cd-hit-est` (Li & Godzik, 2006). Para ello se usaron parámetros respetando los umbrales definidos por la regla 80-80-80 (Wicker et al., 2007), con el fin de definir si dos secuencias pertenecen o no a una misma familia de TEs. Cada secuencia consenso no redundante (Fig. 2.1A) se utilizó como consulta en búsquedas por homología mediante la herramienta `BLAST` (Altschul et al., 1990), contra el ensamblado genómico a partir del cual se obtuvo dicho consenso (Fig. 2.1B). Las secuencias de los *hits* de BLAST junto con las regiones flanqueantes adyacentes (Fig. 2.1C), se extrajeron en archivos multi-FASTA utilizando comandos de `BEDTools` (Quinlan & Hall, 2010) y se alinearon con `MAFFT` (Katoh & Standley, 2013) (Fig. 2.1D). Cada alineamiento se curó y uso para generar una secuencia consenso (Fig. 2.1E). En nuestro caso, curamos manualmente los alineamientos con `AliView` (Larsson, 2014), pero también realizamos un paso de edición automática adicional con `CIAlign` (Tumescheit et al., 2022). Finalmente, se generaron las secuencias consenso curadas con la utilidad `cons` del paquete `EMBOSS(v 6.6.0.0)` (Rice et al., 2000). Cualquier redundancia remantente entre el conjunto de secuencias consenso curadas se eliminó ejecutando `cd-hit-est` por segunda vez.

Las familias de TEs curadas se caracterizaron y clasificaron de acuerdo a lo descrito en la sección siguiente. Consensos correspondientes a alineamientos de muy baja calidad, con muy pocos *hits* de BLAST y sin ningún dominio proteico fueron eliminados. Se conservaron, sin embargo, algunos consensos con bajo número de copias en tanto presentaran características estructurales claramente definidas.

Se identificaron repetidos en tándem complejos, clasificados simplemente como ADN satélite ("Satellite") independientemente del número de copias ni del largo total del motivo de repetición. Se extrajo la secuencia consenso del motivo de repetición en estos repetidos. Por último, se incluyeron en la bibliotecas secuencias de repetidos de identidad desconocida, y se clasificaron simplemente como "Unknown".

Para determinar si dos consensos cualquiera obtenidos a partir de distintos ensamblados correspondían a la misma familia, se corrió nuevamente `cd-hit-est` con los mismos parámetros que antes. No se filtraron las secuencias agrupadas, sino que se les asignó el mismo nombre. También se les asignaron los prefijos "bus-", "gig-" y "hep-" para distinguir los consensos obtenidos a partir de los ensamblados *Fbus*, *Fgig_2* y *Fhep_2* respectivamente. Las tres bibliotecas curadas, diferenciadas por dichos prefijos, se disponen en un mismo archivo multi-FASTA (Supplementary File 2 - TE_libraries). A modo de ejemplo, las secuencias hipotéticas "hep-nombre-X#Clasificación" y "gig-nombre-X#Clasificación" pertenecen a la misma familia pero se obtuvieron a partir de los ensamblados *Fhep_2* y *Fgig_2* respectivamente.

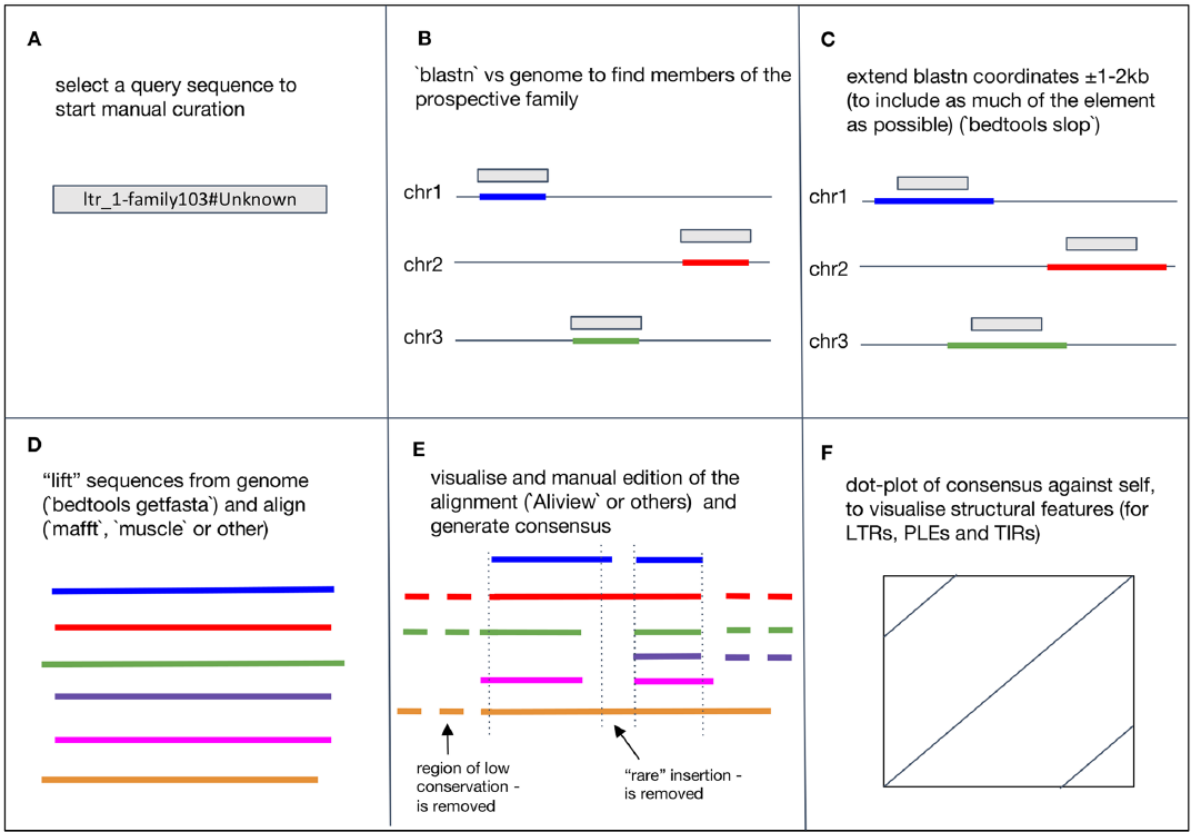


Figura 2.1 | **Esquema del proceso de curación manual.** Imagen tomada del artículo usado como guía para la curación manual (Goubert et al., 2022). En la sección 2.2 de "Materiales y métodos", se describen los pasos de la curación (**A-E**) y en el sitio https://github.com/agustin-bilat/Bilat2024_TEs-Fasciolidae se especifican los comandos utilizados. El panel **F** esquematiza un *dotplot* a partir de uno de los consensos curados en el cual se identifican repetidos terminales directos.

2.3 Caracterización y clasificación de TEs

Se realizó una **caracterización estructural** de los consensos curados. Los dominios proteicos se identificaron utilizando `pfam_scan.pl` (Mistry et al., 2021; https://github.com/aziele/pfam_scan) tomando como entrada las secuencias consenso traducidas en los seis marcos de lectura con el comando `transeq` de `EMBOSS`. A partir de alineamientos de dominios PFAM inicialmente encontrados, se realizaron nuevas búsquedas de las proteínas sobre las secuencias traducidas. Para ello se usaron los comandos `hmmbuild` y `hmmscan` del paquete `HMMER v3.4` (obtenido en 2023, de http://hmmer.org/). Esto permitió identificar un mayor número de dominios con respecto a los inicialmente detectados con `pfam_scan.pl`. Se conservaron los *hits* de `hmmscan` usando como valor umbral un *e-value* de $1x10^{-5}$. Las estructuras repetitivas internas en los extremos de cada secuencia consenso, se visualizaron mediante auto-comparaciones por *dotplots* (Fig. 2.1F), generados con las herramientas `Gepard` (Krumsiek et al., 2007) y `TE-Aid` (ver ejemplo de caracterización con TE-Aid en Figura 2.2) (Goubert et al., 2022). La segunda se corrió con los parámetros `"--e-value 1e-20 --min-orf 300"`, modificando el *script* para incluir el parámetro `"-task 'blastn' "`. La caracterización estructural permitió clasificar las secuencias consenso en transposones de tipo LTR, PLE, LINE o de ADN.

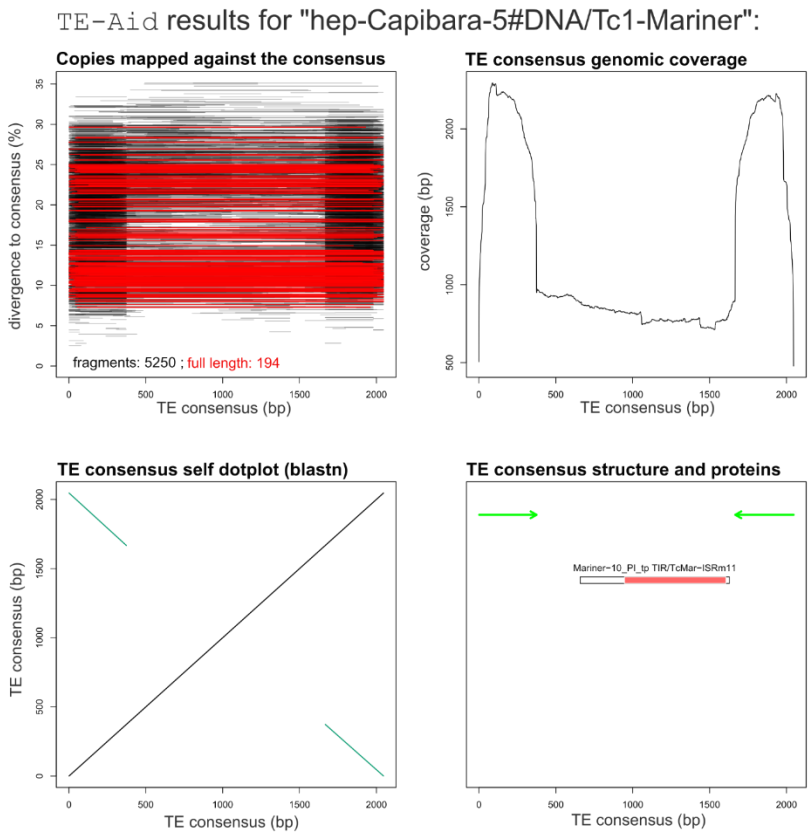


Figura 2.2 | **Salida de TE-Aid.** Se muestra a modo de ejemplo la caracterización estructural a partir de la herramienta `TE-Aid` de una de las familias curadas. Observar que los repetidos terminales internos parecen estar expandidos en comparación al los elementos completos.

Identificamos también consensos que correspondían a la subparte terminal de los LTR (solo-LTRs) para las cuales no se pudieron detectar las correspondientes contrapartes autónomas (Figura 2.3). Estos elementos se clasificaron como LTR y se nombraron como "soloLtr-X", donde "X" corresponde a un número entero positivo. También se detectaron dos familias de elementos de tipo SINE, a partir de características tales como secuencias conservadas, el largo y el número de copias (Figura 2.4). Algunas secuencias consenso adicionales se clasificaron de forma tentativa pese a no disponer de elementos suficientes para una clasificación robusta. Estas comprenden principalmente las familias de clase II clasificadas como CMC, Mutator-like y Helitron, así como algunos otros retrotransposones, principalmente no-autónomos. Las repeticiones en tándem complejas se clasificaron como ADN satélite, y los repetidos de identidad desconocida se clasificaron como "Unknown", tal como fue mencionado antes.

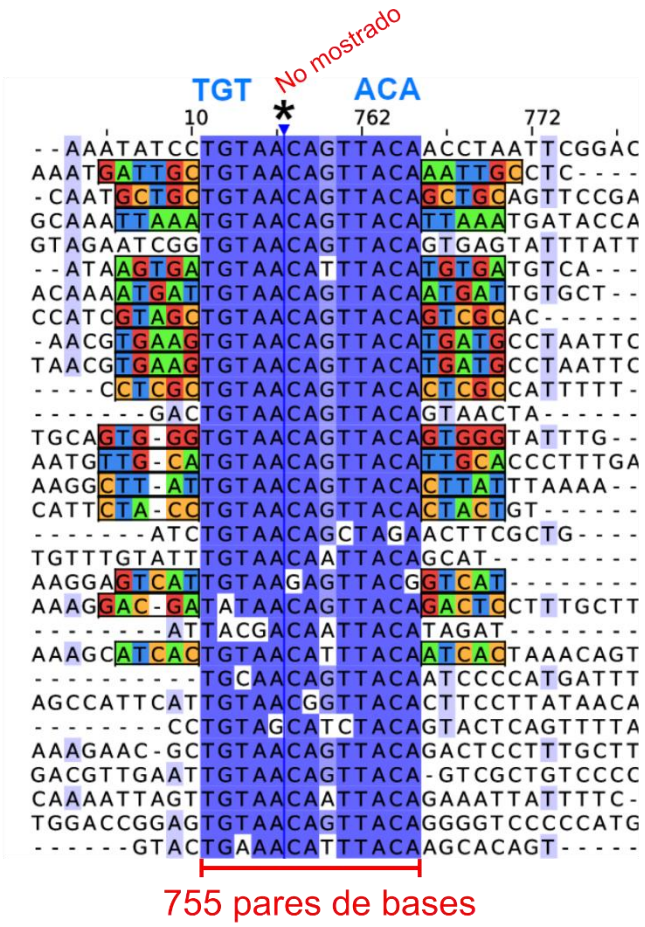


Figura 2.3 | **Características analizadas para definir si una familia corresponde elementos de tipo solo-LTR.** Alineamiento realizado con algunas de las inserciones correspondientes a la familia hep-soloLtr-7#LTR (en la región de los extremos) usada como ejemplo. Los tonos azules reflejan la identidad de secuencia, mientras que las duplicaciones en sitios blanco (más conocidas por su sigla en inglés como TSD) se colorean en base a los nuecleótidos presentes. Se muestran también en letras de color azul los residuos conservados que típicamente se encuentran hacia los extremos de los elementos LTR (sobre todo TG- y -CA).

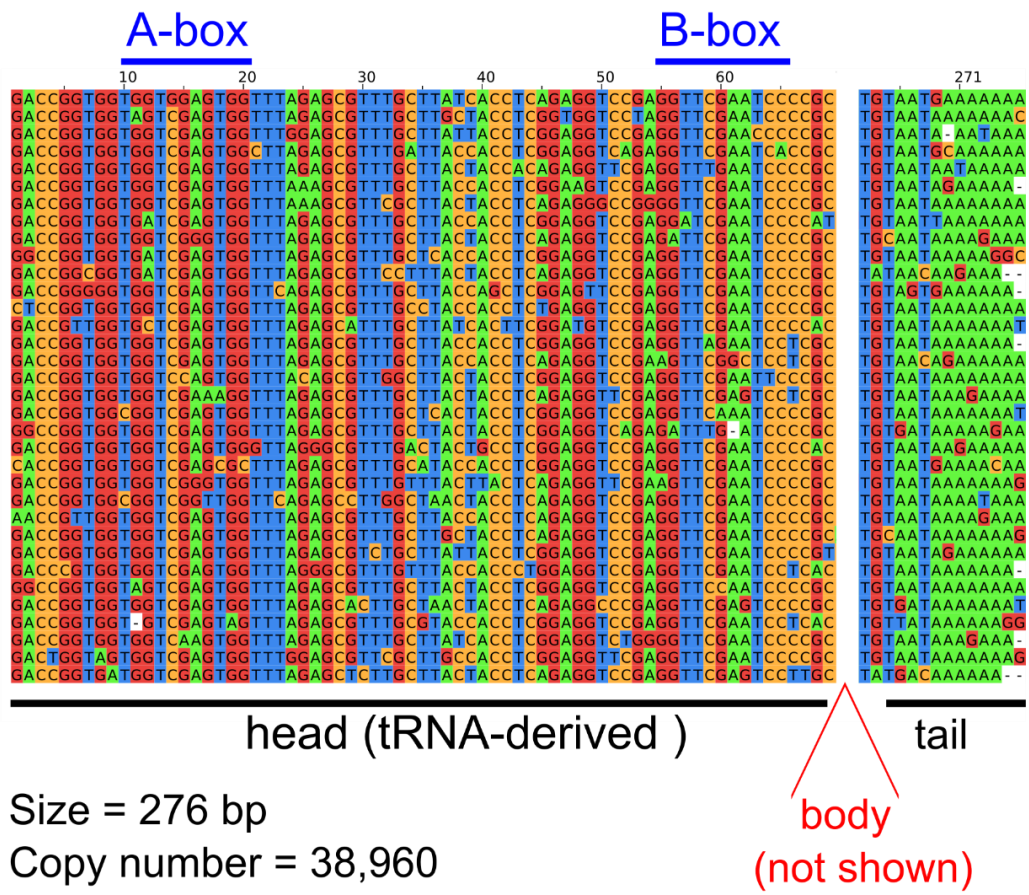


Figura 2.4 | **Características analizadas para identificar elementos de tipo SINE.** Alineamiento realizado con algunas de las inserciones de la familia "hep-Botija", utilizada como ejemplo. Se indica la región de las secuencias conservadas del promotor de la ARN polimerasa III (cajas A y B), así como el número de copias y tamaño del alineamiento. Se observa un pequeño trecho rico en adeninas (cola) al final del alineamiento.

Las filogenias de las distintas familias de TEs se estimaron mediante el método de máxima verosimilitud, partiendo de la secuencia aminoacídica de la RT (para el caso de elementos LTR, LINE y PLE) o la transposasa (para tranposones de ADN) extraída a partir de los consensos curados. Para cada caso se realizó el mismo procedimiento. Primero se tradujeron en los seis marcos de lectura las secuencias consensos, y se identificaron los dominios proteicos con `pfam_scan.pl`. Posteriormente, las utilidades `"slop"` y `"getfasta"` de `BEDTools` permitieron extraer las secuencias peptídicas de la transcriptasa reversa (o transposasa según corresponda) y 100-150 pares de base adicionales que flanquean las coordenadas de los dominios proteicos automáticamente identificados. Se incluyeron secuencias de referencia (Supplementary Table 1 - references) para orientar la clasificación de las familias curadas. La herramienta `MAFFT` con parámetros `"--localpair --maxiterate 1000"` fue usada para generar los alineamientos. Recortamos los extremos de los alineamientos con `Aliview` conservando el *core* de la RT y el dominio DDE de la transposasa. Con estos alineamientos se estimaron las filogenias mediante el método de máxima verosimilitud utilizando el programa `iqtree(v1.6.12)` (Minh et al., 2020). Mediante los parámetros `"-m MFP -bb 1000 -bnni"` se seleccionó un modelo de sustitución óptimo (con el método ModelFinder) y se estimó el soporte estadístico de las ramas por el método *ultrafast bootstrap* con 1000 réplicas. Las filogenias se visualizaron con `evolview v3` (Subramanian et al., 2019).

2.4 Clasificación funcional de los consensos

A partir de la caracterización previa se determinó qué consensos eran autónomos de acuerdo a las definiciones especificadas en la sección de resultados del manuscrito. Se observaron los alineamientos de los dominios identificados para las distintas proteínas, para asegurarse que tengan las secuencias conservadas características del dominio. Luego se contabilizaron manualmente el total de familias de TE autónomas y no-autónomas para cada superfamilia. También se contaron la cantidad de consensos correspondientes a repetidos sin clasificar (Unknown) y Satélites. La tabla así obtenida (Supplementary Table 2 – autonomous) se usó como entrada de `waffle_plots.R` para visualizar la totalidad de consensos que componen nuestras bibliotecas (Supplementary File 2 - TE_libraries) bajo dicha categorización funcional. Se agregaron anotaciones de forma manual a la imagen para indicar la cantidad de familias de LTR que fueron estrictamente clasificadas como no-autónomas (de acuerdo a los criterios mencionados) pero con indicios de que podrían ser autónomas.

2.5 Anotación y composición genómica

Para **anotar** los repetidos en los cinco ensamblados genómicos (Tabla 2.1) se realizaron ejecuciones por separado del programa `RepeatMasker` (disponible en http://repeatmasker.org) con los parámetros "`-e rmblast -lib {TE-library}.fa -s -a -gff -dir out_repMask -cutoff 250 {genome_assembly}.fna`". En cada caso, se utilizó como entrada (mediante el parámetro '`-lib`') una de las tres bibliotecas de repetidos manualmente curadas, correspondiente a la misma especie del ensamblado usado como entrada de `RepeatMasker`. Los archivos de anotación sobre dichos ensamblados obtenidos con `RepeatMasker` se encuentran disponibles (Supplementary File 3 - RepMask).

Los valores de **cobertura total** de ADN repetitivo y ADN no-repetitivo usados para generar la figura 3.2A se obtuvieron a partir de los resúmenes de anotación (archivos .tbl) de las salidas de `RepeatMasker`.

La **cobertura** de TEs para cada familia (longitud total en el ensamblado correspondiente), así como la porción superpuesta con las coordenadas de los genes codificantes y de las regiones intergénicas, se calculó a partir de los archivos de anotación (.out) de TEs obtenidos con `RepeatMasker`, y de la anotación de genes asociada al mismo ensamblado (Tabla 2.1). Los archivos de anotación génica se convirtieron primeramente en formato BED, manteniendo únicamente las líneas correspondientes a los genes codificantes. Esos archivos de anotación de repetidos y genes se usaron luego como entrada del *script* `coverage.sh` bajo los parámetros "`{repeat_masker}.out {genes_annotation}.bed {size_assembly(bp)} {chromosome_prefix}`". Las salidas generadas por cada ensamblado se editaron y unieron en un único archivo (Supplementary Table 3 – coverage), que fue usado como entrada de `coverage_plot.R` para generar así los gráficos de barras de cobertura genómica en las regiones intrónicas e intergénicas. A partir de esa misma tabla se obtuvo el porcentaje que cada familia ocupa en regiones génicas del genoma, permitiendo analizar dicha distribución de forma normalizada por el tamaño de cada familia.

El **largo de las copias** genómicas anotadas a partir de los consensos, se obtuvo de forma separada para cada ensamblado, mediante el *script* `cpyLen.sh`. Todas las salidas —una por cada ensamblado— se juntaron en un único archivo, el cual fue formateado para incluir la clasificación y el largo de las secuencias consensos correspondientes (ver Supplementary Table 4 - insertions).

El **número de copias** totales y completas de cada familia de TEs (Supplementary Table 5 - TEcopies) y los gráficos de barras asociados, se obtuvieron tomando como entrada el archivo "Supplementary Table 4" por el *script* `cpyNum.R`.

La **prueba U de Mann-Whithney** (de una cola) se realizó para evaluar en cada una de las distintas familias si hubo un aumento de la longitud de las inserciones (copias) entre los ensamblados de alta calidad (*Fgig_2* y *Fhep_2*) y los ensamblados más fragmentados (*Fgig_1* y *Fhep_1*). Se ajustaron los *p-valores* por comparaciones múltiples (prueba de Benjamini-Hochberg). Estos análisis se realizaron tomando como entrada el archivo "Supplementary Table 4" por el *script* `cpyLen_MWhitney_test.R`, el cual devuelve como salida tres archivos: uno con la media del largo de copias por familia (Supplementary Table 6 - median_cpyLen) y otro con los resultados de la prueba estadística junto con los *p-valores* ajustados (Supplementary Table 7 - MannWithney), así como un gráfico de tortas para visualizar la proporción de familias en las cuales el *test* dio significativo (*p-valor* ajustado menor a 0.01). El archivo con el largo de las inserciones (Supplementary Table 4) también se usó para generar mediante la librería `"ggplot2"` de `R`, gráficos de cajas (*box-plots*) y violines (*violin-plot*) superpuestos, con el fin de visualizar la distribución de los tamaños de las copias en distintos ensamblados para algunas familias arbitrariamente seleccionadas.

2.6 Dinámica evolutiva de los TEs

Mediante el método de máxima verosimilitud se estimó la filogenia para comparar el conjunto de secuencias consensos correspondientes a las familias curadas de Fasciolidae; más concretamente, de aquellas que tuviesen la región codificante para la transcriptasa reversa o la transposasa (ver sección 2.3: "Caracterización y clasificación de TEs"). Lo anterior permitió realizar hipótesis sobre la presencia o expansión de distintos tipos y familias de TEs a lo largo de la evolución de esta familia (es decir, en el ancestro común a las especies estudiadas o luego de cierto evento de especiación particular). También se realizaron estudios para estimar los períodos de actividad de las distintas superfamilias y la edad de cada una de familias individuales, tal como se describe a continuación.

Los **períodos de actividad** relativos entre distintos tipos de repetidos se evaluaron usando el modelo de sustitución de Kimura de 2 parámetros (excluyendo sitios CpG), el cual permite estimar la distancia genética entre las copias de TEs (es decir, las inserciones) y el correspondiente consenso usado para su anotación. Estos valores se extrajeron de los archivos de alineamiento (.aln) de la salida de `RepeatMasker` usando el *script* `getKimura_from_aln.sh`, y las tablas resultantes se combinaron con el comando `"cat"` de bash, para luego generar los gráficos de paisajes de repeticiones ("repeat landscape") con `getKimura_plots.R`. Los "picos" del gráfico reflejan los mencionados períodos de actividad de los distintos tipos de TE, en escala de tiempo relativa.

La **edad relativa** de cada familia de TE se estimó de forma directa (sin comparar con los consensos) usando de guía un trabajo previo (Chang et al., 2022) con algunas modificaciones. Primero, las secuencias nucleotídicas de las inserciones se extrajeron tomando como entrada los archivos de anotación (.out) de `RepeatMasker` y la secuencia del ensamblado correspondiente, y ejecutando de manera secuencial los *scripts* `insertions_fastas.sh` e `insertions_moveFiles.sh`. Esto generó archivos multi-FASTA para cada familia de TEs con un máximo de 500 secuencias aleatorias correspondientes a inserciones, conservando únicamente aquellas con al menos la mitad del tamaño del consenso correspondiente. Se generaron posteriormente alineamientos con la herramienta `MAFFT`, que fueron editados con `CIAlign` para eliminar las inserciones. Se infirió la filogenia de cada familia de forma rápida a partir de cada alineamiento usando un método aproximado al método de máxima verosimilitud implementado por el programa `fasttree v. 2.1.11` (Price et al., 2010), con los parámetros "`-nt -gtr -pseudo -gamma -nopr`". Los árboles obtenidos se usaron como entrada de `branch_lengths.py` (el cuál se obtuvo modificando un *script* previamente publicado (Chang et al., 2022)) para calcular el valor de la media del largo de las ramas terminales. Las salidas (una por cada ensamblado) se unieron en un único archivo (Supplementary Table 8 - TEage), y junto con el archivo "Supplementary Table 5" se usaron como entradas del *script* `branch_lengths_plots.R` para generar los gráficos de puntos de las edades (relativas) de las familias en función del número de copias.

2.7 Correlación entre variables cuantitativas

Se calculó el coeficiente de correlación de Spearman (no paramétrico) para evaluar la correlación entre las siguientes variables: diferencia de cobertura, del número de copias, del número de copias completas, del largo medio de las copias y de la edad de las familias de TEs curadas, entre los valores calculados a partir de la anotación sobre ensamblados de lecturas largas con respecto a los de lecturas cortas para las especies *F. hepatica* y *F. gigantica*. Esto se realizó con el *script* `correlograms.R`, el cual genera la figura de los 'correlogramas' tomando como entrada distintos archivos generados anteriormente como se indica a continuación: '`--cov <coverage.csv> --cpy <TEcopies.csv> --len <median_cpyLen.csv> --age <TEage.csv>`'.

2.8 Solapamiento de TEs e intrones en función de sus posiciones en los genes.

El análisis de enriquecimiento de TEs según la posición al interior de los genes se baso en un trabajo previo (Philippsen & DeMarco, 2019). Primero filtramos de las anotaciones génicas cualquier par de genes con solapamientos parciales entre sí, así como aquellos que solapen buena parte de su CDS con las coordenadas de TEs. En segundo lugar, calculamos el porcentaje de intrones con alguna ocurrencia (1 o más) de TEs de acuerdo a la posición de los intrones en los genes (a partir de la anotación de RepeatMasker y de la anotación de genes filtrada). Por ejemplo, si tenemos 100 intrones y solo 50 de estos solapan con algún TE (sin importar el tipo) el porcentaje con alguna ocurrencia de TE equivale a 50%. Este cálculo se hace tomando el primer intrón de los genes, luego tomando el segundo intrón, etcétera. En tercer lugar se repite el cálculo considerando una superfamilia de TEs a la vez (en lugar de contar ocurrencias con cualquier TEs). Usamos una verisión anterior del ensamblado de *F. hepatica*, y para *F. gigantica* solamente el ensamblado *Fgig_2*.

Capítulo 3.

Resultados

3.1 Diversidad de TEs en Fasciolidae

Las familias de TEs suelen almacenarse como secuencias consenso en formato FASTA, obtenidas de los alineamientos de las copias genómicas (también conocidas como inserciones). Este formato es compatible con una amplia variedad de programas bioinformáticos, como aquellos comúnmente usados para mapear las coordenadas genómicas de los elementos repetitivos. Con el fin de realizar un análisis detallado del contenido genómico de TEs en Fasciolidae, realizamos en primer lugar una extensa **curación manual** (Goubert et al., 2022; Platt et al., 2016; J. M. Storer et al., 2021) de las familias de TEs identificadas *de novo* en los ensamblados de *F. hepatica*, *F. gigantica* y *F. buski* (ver "Materiales y métodos"). Las bibliotecas curadas se encuentran disponibles (Supplementary File 2). Para caracterizar la composición y diversidad de las familias que componen las bibliotecas, así como para evaluar la calidad de estas últimas, realizamos distintos análisis filogenéticos y estructurales cuyos resultados son descritos a continuación.

3.1.1 Diversidad de clases de TEs

Las familias curadas se clasificaron en distintas **superfamilias** al interior de los órdenes LTRs, LINEs, PLEs y de los transposones de ADN. Para lograr esto se generaron filogenias con los consensos curados de TEs, correspondientes a los transposones de clase II (Figura Suplementaria 1) y a los distintos órdenes de retrotransposones (Figura Suplementaria 2), como se describe en la sección de "Materiales y métodos". La comparación con secuencias de clasificación conocida (Supplementary Table 1) permitió realizar la asignación de superfamilias a cada consenso. Nuestra clasificación confirma así la presencia de distintas superfamilias de TEs en Fasciolidae que habían sido reportadas en base a métodos enteramente automáticos (Choi et al., 2020). El análisis de las filogenias de los consensos da cuenta de **"clados" de TEs** bien diferenciados —con buenos soportes estadísticos usando método *ultrafast bootstrap*— al interior de todas las superfamilias. Estos resultados representan una caracterización más detallada de la diversidad filogenética de TEs en fasciólidos. Discutiremos aspectos de la evolución de los TEs en Fasciolidae más adelante.

En nuestras bibliotecas curadas también detectamos retrotransposones no-autónomos como solo-LTRs (Figura 2.3) y SINEs (Figura 2.4), así como algunos pocos consensos correspondientes a fragmentos degradados e incompletos de retrotransposones autónomos. También identificamos familias de TEs de **clase II**. Estas últimas fueron clasificadas a partir de la detección de la proteína transposasa, y en algunos casos se pudieron detectar las secuencias de repetidos terminales invertidos, los cuales son característicos de la mayoría de estos transposones, conocidos como TIRs (Wicker et al., 2007). Sin embargo, una fracción de las familias que asignamos a la categoría de transposones de ADN, fueron clasificadas de manera automática, debido a la dificultad para detectar en las mismas, proteínas o estructuras claramente definidas (ver "Materiales y métodos").

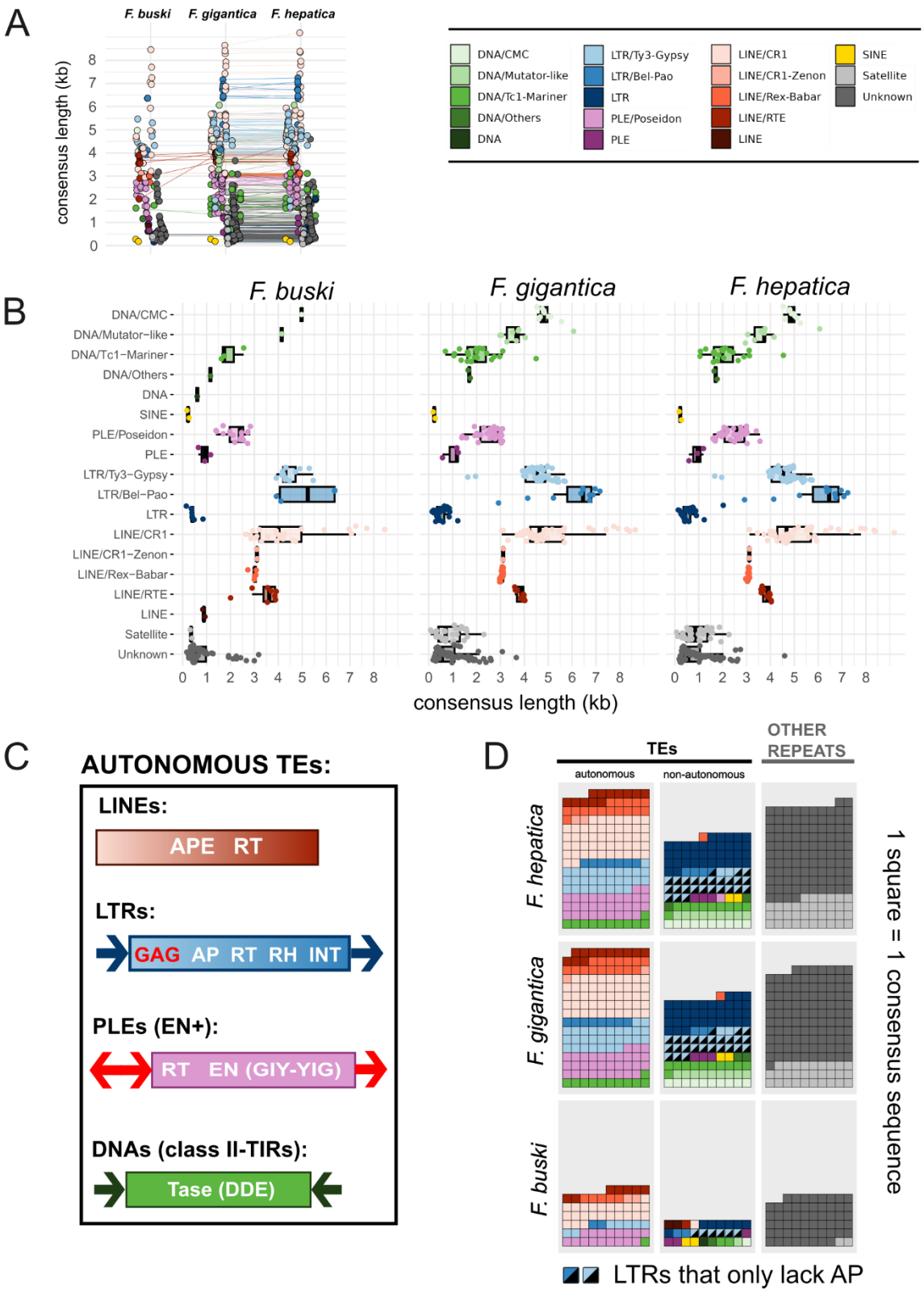
3.1.2 Diversidad de tamaños de TEs

Las bibliotecas generadas a partir de *F. hepatica* y *F. gigantica* se componen de las mismas familias de TEs, de acuerdo a la regla "80-80-80" (Wicker et al., 2007) (Figura 3.1A). Esto incluye no solo a los elementos autónomos, sino también a los TEs clasificados como no-autónomos. También comparten repetidos de identidad desconocida ("Unknown") así como repetidos en tandem complejos ("Satellite") (Figura Suplementaria 3A). Por su parte, solo algunas de las familias curadas en *Fasciola* spp. son compartidas con *F. buski* (si bien casi todos los clados de TEs antes mencionados contienen representantes de las tres especies). Para cada una de las familias (secuencias consenso) se observa un largo en pares de base que es esencialmente constante entre las distintas especies (Figura 3.1A). No obstante, existen aún pequeñas diferencias tras la curación en algunas familias. Por ejemplo, las familias de CR1 mayores a 6000 pb presentan una variación de tamañó de aproximadamente el 5% (unos 300-400 pb en 6250-6500 pb) entre los modelos obtenidos a partir de *F. hepatica* y *F. gigantica* (Figura Suplementaria 3B).

Globalmente, los diversos tipos de repetidos identificados exhiben un amplio rango de tamaños en sus secuencias consenso, con un perfil general muy similar en las tres especies (Figura 3.1B). Los disitintos órdenes en estas especies suelen diferenciarse en sus tamaños, siendo los PLEs normalmente más pequeños que los LINEs, y estos últimos más pequeños que los LTRs (a excepción de la superfamilia CR1) (Figura 3.1B).

A nivel de superfamilias también se observan diferencias. En general, vemos que dentro de los LTRs la superfamilia Bel-Pao se compone de consensos mayores a 6 kb mientras que los consensos de tipo Ty3-Gypsy suelen ser menores a los 5 kb. Algo análogo se observa en otros grupos, como los LINEs, donde la longitud media de secuencias consenso ronda los 3 kb para las superfamilias Rex-Babar y CR1/Zenon, 3.5-4 kb para los RTE, y mas de 4 kb en los CR1. Teniendo en cuenta que la superfamilia CR1 presenta más clados (Figura suplementaria 2A) no resulta sorprendente que haya un mayor rango de variación en el largo de los consensos que la componen. En particular, las familias del clado "sam6" tienen un tamaño considerable, siendo el largo medio mayor a 6 kb en cualquiera de las especies, mientras que el resto de los CR1 son menores a 5 kb. Por último, vemos varios consensos curados de TEs con tamaños pequeños (Figura 3.1B). Exceptuando algunas familias de ADN/Tc1-Mariner, todos los consensos de TEs menores a 1.5kb corresponden a elementos no-autónomos. Los repetidos clasificados como "Unknown" son también secuencias muy fragmentadas en general (largo medio de aproximadamente 0.6 kb). Estas observaciones indican —como podría esperarse suponiendo un orígen común— que las familias de TEs vinculadas filogenéticamente por pertenecer a un mismo clado o superfamilia, suelen mantener también una longitud comparable en la secuencia consenso usada para su representación.

Figura 3.1 | **Caracterización de las bibliotecas.** La parte **A** muestra la intersección entre bibliotecas. Cada círculo representa un consenso del tamaño indicado en el eje de las ordenadas, obtenida a partir del genoma de la especie indicada en el eje de las abcisas. Dos círculos cualquiera conectados por una línea corresponden a una misma familia. El color indica el tipo o clasificación del repetido, de acuerdo a lo indicado en la parte derecha de la figura. El mismo código de colores se mantiene a lo largo del manuscrito. La parte **B** muestra la distribución de tamaños de las secuencias consensos curadas en forma de gráfico de cajas (*box-plot*), superpuesta con gráfico de puntos para visualizar mejor la cantidad de familias que componen cada "caja". Se definieron como autónomos los consensos con las estructuras mostradas (no se evaluó las de color rojo) (**C**). Se representan con cuadrados de colores la cantidad de consensos autónomos y no-autónomos TEs, así como repetidos de identidad desconocida y Satélites (**D**). Un símbolo especial (un triángulo negro) fue usado para indicar la cantidad de familias de LTR para las cuales únicamente no fue posible detectar la proteína AP. AP, aspartil proteasa; APE, Endonucleasa apurínica; EN (Giy-YIG), endonucleasa con dominio de tipo "giy-yig", INT, integrasa; RT, transcriptasa reversa, RH, RNasa H, Tase (DDE), transposasa con dominio conservado "DDE".


A
F. buski
F. gigantica
F. hepatica
consensus length (kb)
0
1
2
3
4
5
6
7
8
DNA/CMC
DNA/Mutator-like
DNA/Tc1-Mariner
DNA/Others
DNA
LTR/Ty3-Gypsy
LTR/Bel-Pao
LTR
PLE/Poseidon
PLE
LINE/CR1
LINE/CR1-Zenon
LINE/Rex-Babar
LINE/RTE
LINE
SINE
Satellite
Unknown
B
F. buski
F. gigantica
F. hepatica
DNA/CMC
DNA/Mutator−like
DNA/Tc1−Mariner
DNA/Others
DNA
SINE
PLE/Poseidon
PLE
LTR/Ty3−Gypsy
LTR/Bel−Pao
LTR
LINE/CR1
LINE/CR1−Zenon
LINE/Rex−Babar
LINE/RTE
LINE
Satellite
Unknown
0 1 2 3 4 5 6 7 8
consensus length (kb)
C
AUTONOMOUS TEs:
LINEs:
APE RT
LTRs:
GAG AP RT RH INT
PLEs (EN+):
RT EN (GIY-YIG)
DNAs (class II-TIRs):
Tase (DDE)
D
TEs
OTHER REPEATS
autonomous
non-autonomous
F. hepatica
F. gigantica
F. buski
1 square = 1 consensus sequence
LTRs that only lack AP

3.1.3 Diversidad funcional de TEs

Para evaluar si el proceso de curación nos permitió obtener consensos completos de TEs potencialmente activos, categorizamos a las familias desde un punto de vista funcional en **autónomas** y **no-autónomas**. Para esto nos basamos en la definición general reportada en un trabajo previo (Wicker et al., 2007). La misma expresa que un elemento es autónomo si codifica todos los dominios necesarios para su transposición (lo cual no implica que sea activo). Utilizamos un criterio más flexible para simplificar el análisis, focalizando en algunos dominios/estructuras. La Figura 3.1C resume los criterios utilizados para cada tipo de TEs. Las familias de **LTRs** se clasificaron como autónomas si presentaban todos los dominios del gen *pol* (AP, RT, RH e INT) y la subparte LTR terminal. Para los retrotransposones **no-LTRs**, requerimos que presenten la proteína RT y las proteínas APE (en el caso de los **LINEs**) y EN (giy-yig) (en el caso de los **PLEs**). Los **transposones de ADN** se clasificaron como autónomos si codificaban para la transposasa y tenían los repetidos terminales invertidos (TIRs). La Figura 3.1D resume la cantidad de consensos de TEs clasificados como autónomos y no-autónomos bajo estos criterios. En paralelo se muestra la cantidad de consensos de satélites y de repetidos sin clasificar.

En términos generales, un 60 a 70 % —dependiendo la especie— de los consensos de TEs fueron clasificados como autónomos. Dentro de los **transposones de ADN** (clase II), la mitad de los consensos de la superfamilia Tc1-Mariner corresponden a elementos autónomos. Los restantes tipos de esta clase fueron en su mayor parte clasificados automáticamente al no haberse podido identificar claramente el dominio *pfam* de la proteína transposasa ni los repetidos terminales invertidos. La porción de las bibliotecas correspondietes a los retrotransposones **LINEs** y **PLEs** se componen casi en su totalidad de familias autónomas, contabilizando unos 55 consensos autónomos en *F. buski* y el doble —aprox. 110 consensos autónomos— en cualquiera de las dos especies del género *Fasciola*. La mitad de los **LTRs** fueron clasificados como autónomos. Sin embargo, en casi todos los Ty3-Gypsy clasificados como no-autónomos, un único dominio no fue identificado (correspondiente a la proteasa aspártica, AP). Ello sugiere que podrían ser efectivamente elementos autónomos, pero con un dominio de AP divergente o dificil de detectar.

Resumiendo, podemos afirmar que la curación manual resultó en la obtención de bibliotecas de alta calidad compuestas por tres conjuntos —uno por especie— de secuencias consensos de TEs comparables, sumado a otros tipos de repetidos de pequeño tamaño. Nuestros análisis permitieron además caracterizar la diversidad filogenética y estructural de las familias de TEs curadas.

3.2 Análisis de la composición genómica de TEs y de su evolución en Fasciolidae

Estudios previos revelaron que los genomas de las *Fasciola* spp. están entre los más grandes de los helmintos parásitos; consecuencia del alto contenido de elementos repetitivos (Choi et al., 2020; International Helminth Genomes Consortium, 2019). Por otra parte, la estimación del ADN repetitivo total en *F. gigantica* mostró estar fuertemente afectada por la calidad del ensamblado genómico (Luo et al., 2021). Por este motivo utilizamos distintos ensamblados disponibles para la anotación genómica de TEs (Tabla 2.1). Las bibliotecas curadas permitieron analizar el efecto de la calidad de los ensamblados sobre la anotación de los TEs. Comparamos la composición genómica de TEs entre fasciólidos, ponderando posibles sesgos metodológicos sobre distintas variables cuantitativas. Los resultados de estos análisis describen a continuación.

3.2.1 Sesgos en la anotación de TEs asociados a la calidad de los ensamblados

La fracción de ADN no repetitiva es aproximadamente constante (entre los 430 a 480 Mb) mientras que la fracción de ADN repetitiva cambia más del 40% (más de 250 Mb) entre el ensamblado de lecturas largas (*Fgig_2* o *Fhep_2*) y el de lecturas cortas (*Fgig_1* o *Fhep_1*) en la misma especie (Tabla 3.1 y Figura 3.2A). Esto coincide con lo reportado por los autores que generaron el ensamblado de alta calidad de *F. gigantica* (Luo et al., 2021), si bien la magnitud del cambio estimada por nosotros es considerablemente mayor. Las diferencias entre ambos estudios podrían deberse al uso de distintas bibliotecas para anotar los repetidos.

En la fracción repetitiva, prácticamente todos los tipos de TEs presentan sesgos en la cobertura asociados a la calidad de los ensamblados, tanto en las regiones génicas como intergénicas (Figura 3.2B). Algo similar ocurre con otros repetidos, como los ADN satélites. Si bien se observa un aumento en el número de copias totales en los ensamblados de mejor calidad (Figura 3.2C), la magnitud de dicho cambio parecería ser insuficiente para explicar completamente las grandes diferencias que vemos a nivel de la cobertura.

Tabla 3.1 | **Cobertura genómica de repetidos en Fasciolidae.** Cobertura genómica para distintos tipos de repetidos, obtenidos sumando el tamaño individual de las familias de: Supplementary Table 3. Solapamientos parciales entre distintos repetidos anotados pueden dar una ligera diferencia (de hasta aprox. 1%) en los valores de ADN repetitivo total así estimados con respecto al total de bases efectivamente enmsacaradas.

Classification		Coverage (bp)				
		Fbus	***Fgig_1***	***Fgig_2***	***Fhep_1***	***Fhep_2***
DNA	CMC	1,170,945	9,106,842	14,770,693	8,552,869	12,036,113
DNA	Mutator-like	1,664,638	5,848,870	9,915,278	5,037,316	7,372,015
DNA	Tc1-Mariner	2,070,300	17,700,034	19,796,406	19,212,869	22,104,267
DNA	Merlin	8,532	-	-	-	-
DNA	Helitron	-	299,802	36,529	350,291	48,043
DNA	PiggyBac	-	819,659	961,025	951,108	1,121,797
DNA	-	186,949	-	-	-	-
Total DNA		5,178,152	33,775,207	45,808,692	34,104,453	43,114,622
LINE	CR1	61,986,969	166,488,058	257,638,586	181,936,211	242,420,097
LINE	CR1-Zenon	6,959,378	7,674,530	8,352,010	8,319,909	9,699,834
LINE	Rex-Babar	18,354,003	31,041,521	35,386,056	35,237,327	42,459,973
LINE	RTE	96,087,915	118,962,349	135,128,674	118,777,620	146,814,950
LINE	-	1,497,621	-	-	-	-
Total LINE		184,885,886	324,166,458	436,505,326	344,271,067	441,394,854
LTR	Bel-Pao	6,448,478	63,628,713	115,166,978	59,126,255	105,575,148
LTR	Ty3-Gypsy	16,912,948	85,598,395	115,700,155	89,535,632	120,338,322
LTR		1,965,925	16,230,965	18,954,363	18,508,282	21,750,389
Total LTR		25,327,351	165,458,073	249,821,496	167,170,169	247,663,859
PLE		793,001	1,334,640	1,544,620	1,445,608	1,658,192
PLE	Poseidon	19,842,929	37,248,803	46,510,667	38,154,730	53,278,173
Total PLE		20,635,930	38,583,443	48,055,287	39,600,338	54,936,365
Total SINE		8,841,646	8,804,560	9,553,576	9,671,060	10,394,615
Total Unknown		39,885,683	56,565,568	71,461,330	61,570,140	74,947,847
Total Satellite		7,539,568	14,858,746	40,325,244	15,976,337	124,475,726
Total Simple repeat		6,178,895	7,631,089	14,504,433	5,051,775	29,923,842
Repetitive DNA		298,473,111	649,843,144	916,035,384	677,415,339	1,026,851,730
Non-repetitive DNA		449,684,519	477,910,745	431,830,406	460,913,662	480,107,636
Total		**748,157,630**	**1,127,753,889**	**1,347,865,790**	**1,138,329,001**	**1,506,959,366**

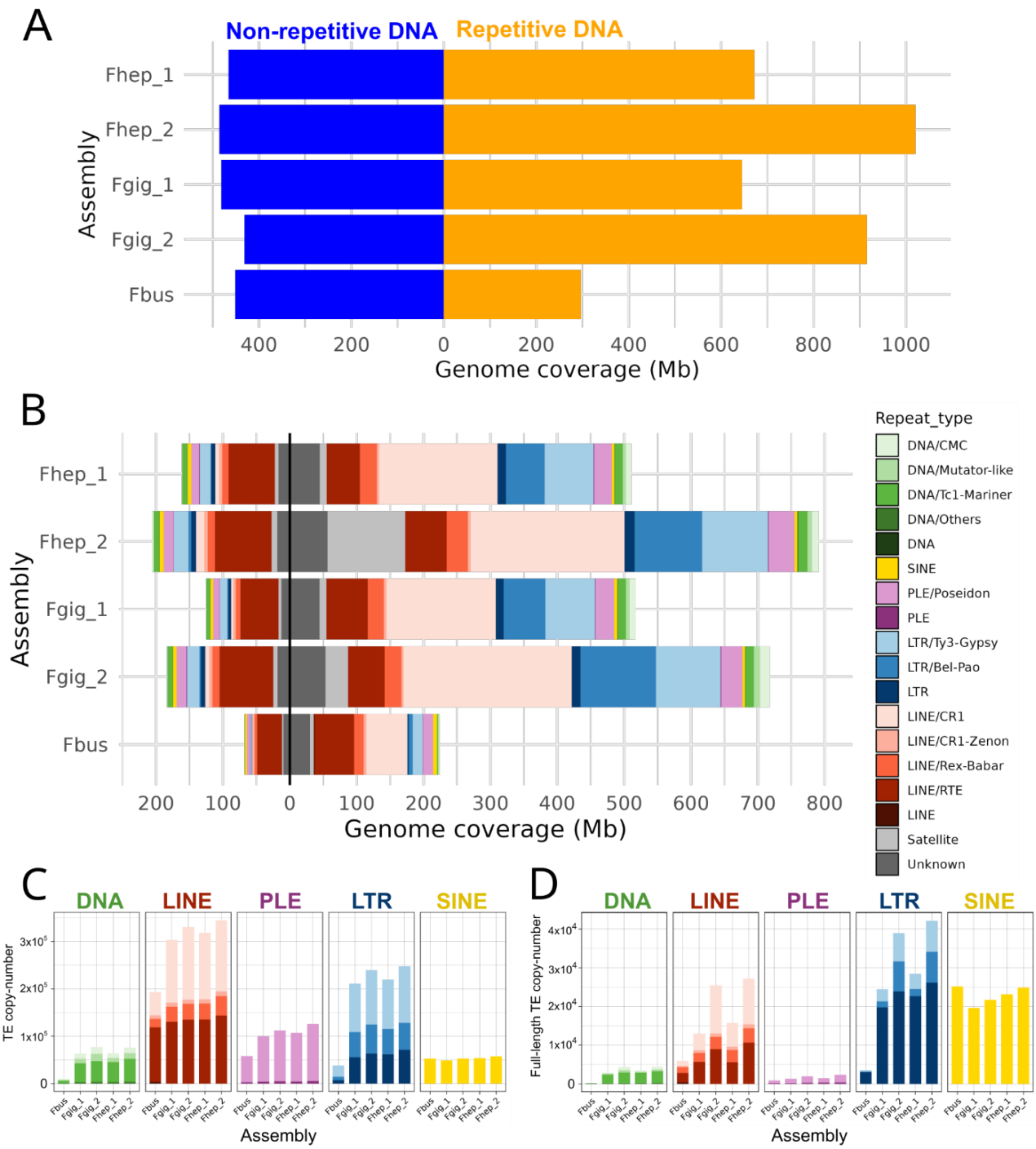


Figura 3.2 | **Composición genómica de TEs en Fasciolidae.** Cobertura del ADN repetitivo y no-repetitivo total (**A**). Cobertura de distintos tipos de repetidos según su localización en región intrónica o intergénica (a la izquierda y derecha de la linea vertical negra respectivamente) (**B**). Para los TEs con clasificación conocida, calculamos el número de copias totales (**C**) y copias completas (**D**).

Las observaciones anteriores nos llevaron a hipotetizar que no solo existe un incremento en el número de copias de las familias, sino también en el largo de las copias, por efecto del ensamblado. De hecho, los gráficos de distribución de los tamaños de las copias realizados a partir algunas familias arbitrariamente elegidas, muestran un incremento global en el tamaño de las inserciones por efecto de la mejora de los ensamblados, que es particularmente notorio en la fracción de copias completas (Figura 3.3).

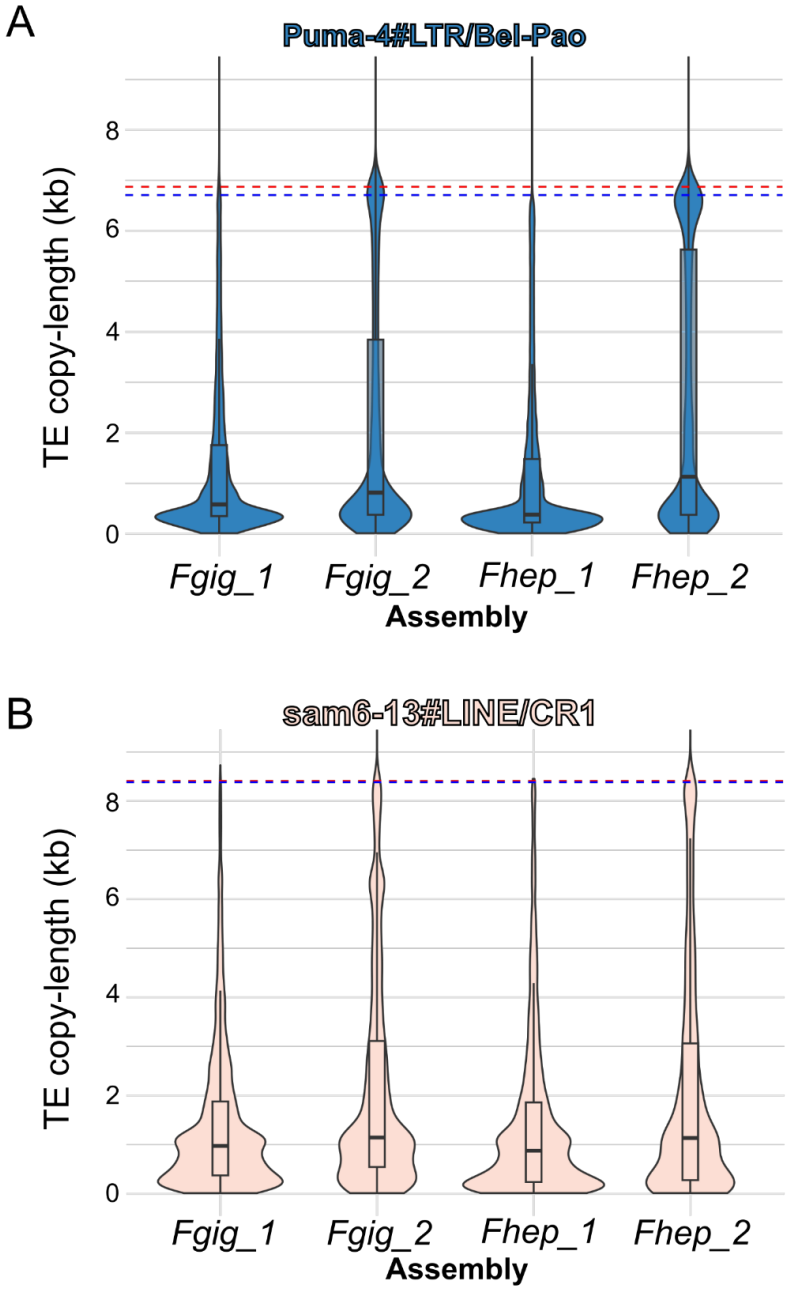


Figura 3.3 | **Distribución de largos de las copias.** Las familias Puma-4# (**A**) y sam6-13# (**B**) se seleccionaron arbitrariamente para ejemplificar los cambios en el tamaño de las inserciones entre las versiones distintas de ensamblados genómicos. La línea punteada indica el tamaño de los consenso obtenidos de *Fgig_2* (rojo) y *Fhep_2* (azul) durante el proceso de curación manual, y que fueron usados luego —por separado— como entrada de `RepeatMasker` para realizar las anotaciones sobre los ensamblados de la especie correspondiente. Los gráficos de caja permiten visualizar la distribución de tamaño de las copias genómicas por cuartiles, mientras que amplitud de los gráficos con forma de violin permiten ver los rangos de tamaño enriquecidos en inserciones.

Para tener un panorama global, realizamos un *test* no paramétrico para cada familia de TE, con el cual estimamos que alrededor del 25% de ellas presenta incrementos significativos en el largo de copias en el ensamblado de lecturas largas (Figura 3.4). Similarmente, la cuantificación global del número de copias completas muestra cambios sensiblemente mayores (en términos relativos) a los observados para el número total de copias por efecto del ensamblado, en la mayoría de superfamilias de TEs Figura 3.2D).

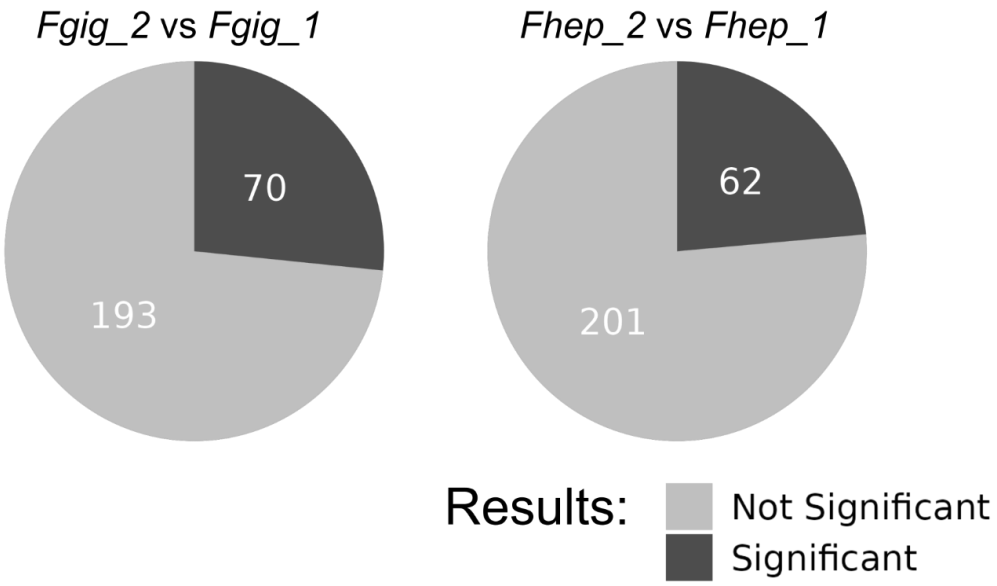


Figura 3.4 | **Cantidad de familias con cambio significativo en el largo de sus copias.** Mediante la prueba U de Mann-Whitney encontramos que cerca del 25% de las familias de TEs incrementan el largo de sus copias por efecto del incremento en la calidad del ensamblado. No se realizó el análisis para la especie *F. buski* ya que se cuenta con un único ensamblado de dicha especie. Se consideraron significativos los resultados con p-valor ajustado por FDR menor a 0.01. FDR, *False Discovery Rate*.

La calidad de los ensamblados también afecta la estimación de la edad de las familias (Figura 3.6). Uno de los motivos posibles que expliquen estas observaciones, es que las inserciones más recientes se encuentren enriquecidas en el subconjunto de copias completas en comparación al total de copias de una familia, al menos en lo que respecta a ciertas superfamilias de elementos autónomos. Así, dadas las dificultades para incluir copias más completas en los ensamblados de *contigs* generados con lecturas cortas, resulta que la calidad del ensamblado termina por afectar también la estimacion de la edad.

Finalmente, hicimos un análisis que nos permitió confirmar estadísticamente que el incremento en la cobertura de las familias de TEs causado por el efecto del ensamblado (Luo et al., 2021) se correlaciona de forma positiva con el cambio en el número de copias (totales y completas) y del largo medio de las mismas, y de forma negativa con el cambio en la edad estimada de las familias (Figura 3.5). En resumen, nuestros análisis confirman en forma detallada que la calidad del ensamblado afecta fuertemente las estimaciones realizadas a partir de los datos de anotación de TEs, con implicancias particularmente relevantes en lo referente al estudio de expansiones recientes, que estarían mayormente representadas por conjuntos de copias completas y con baja divergencia a nivel de secuencia.

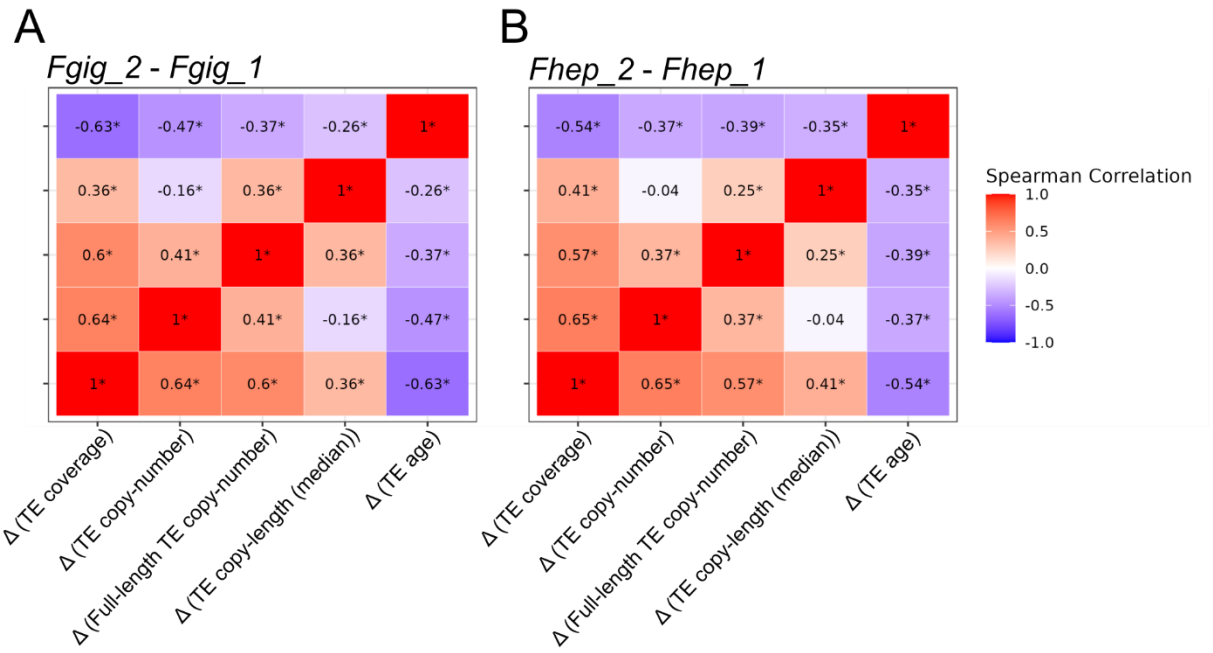


Figura 3.5 | **Correlogramas.** Para *F. gigantica* (**A**) y *F. hepatica* (**B**) calculamos la correlación entre la diferencia de cobertura (pb), número de copias, número de copias completas, largo medio de copias (pb) y edad de las familias de TE curadas, entre los ensamblados de lecturas largas (*Fgig_2* y *Fhep_2*) con respecto a los de lecturas cortas (*Fgig_1* y *Fhep_1*). La edad relativa de cada familia se estimó como el largo medio de las ramas terminales de los árboles obtenidos con *FastTree* (ver M. y M.). Se indican los valores del coeficiente de correlación de Spearman. Los asteriscos señalan diferencias significativas (p-valor de 0.01).

3.2.2 Diferencias de TEs entre *Fasciola* y *Fasciolopsis*.

La **cobertura total de TEs** en *Fasciola* es entre 130 a 140 % mayor (unos 325 a 350 Mb) que en *Fasciolopsis*, si se comparan anotaciones con ensamblados de calidades similares (*Fhep_1* o *Fgig_1* con respecto a *Fbus*) (Tabla 3.1 y Figura 3.2B). A modo comparativo, el tamañó del genoma del parásito humano *Schistosoma mansoni* (version SM_V10) (Howe et al., 2017) es de 391 Mb. Es decir que las diferencias en el contenido de TEs entre *Fasciola* y *Fasciolopsis* por sí solas, son aproximadamente iguales que el tamaño genómico de este otro trematodo. De hecho, es posible que dichas diferencias de tamaño genómico entre ambos géneros sean incluso más pronunciadas, dado que los TEs en los ensamblados *Fhep_2* y *Fgig_2* cubren al menos unos 200 Mb adicionales.

Al analizar la cobertura para distintos tipos de TEs observamos que *Fasciola* tiene al menos 6 a 7 veces más **transposones de ADN** y de **LTRs** (que sumados representan un incremento de entre 170 a 190 Mb), mientras que posee 2 veces más del conjunto restante de retrotransposones **no-LTRs** (aproximadamente 160 a 180 Mb de diferencia) (Tabla 3.1). A nivel de superfamilias, vemos que algunas presentan cambios más pronunciados que otras. Por ejemplo, en el órden de los LINEs, la superfamilia CR1 aumenta unas 3 veces mientras que los RTE aumentan apenas un 20% (Tabla 3.1).

Una mayor cantidad de familias curadas (es decir, una mayor diversificación a nivel de secuencia) acompaña el incrementeo en la cobertura de las superfamilias en *Fasciola* (Figuras suplementarias 1 y 2). Es posible que ensamblados de mejor calidad en *F. buski* puedan dar lugar a la obtención de más familias curadas a partir de regiones repetitivas que estén colapsadas en el ensamblado *Fbus*. Sin embargo, las copias más antiguas de cada familia deberían poder recuperarse igualmente en un genoma fragmentado, lo que

relativiza dicha posibilidad. Más aún, es claro que los genomas de *Fasciola* spp. son al menos 380 a 390 Mb mayores que el de *F. buski,* lo cual es coherente con que haya más inserciones genómicas y consecuentemente más diversificación de secuencia para los distintos tipos de TEs. Además, un trabajo previo utilizando los mismos ensamblados (*Fbus*, *Fgig_1*, *Fhep_1*) también observó un aumento en la cobertura de muchas superfamilias, si bien no se realizó el proceso de curación manual (Choi et al., 2020). Por último, la estimación de las edades de las familias a partir de anotaciones en los ensamblados *Fhep_1*, *Fgig_1* y *Fbus*, apunta también a la presencia de expansiones más recientes en *Fasciola* con respecto a *Fasciolopsis* (Figura 3.6). Con todo, se puede afirmar que *Fasciola* presenta un incremento prominente en la mayoría de las superfamilias de TEs con respecto a *Fasciolopsis*, posiblemente acompañado de una mayor diversificación de secuencias y aparición de nuevas familias.

3.2.3 Expansiones al interior de *Fasciola* spp.

A partir de los datos de anotación en los ensamblados *Fhep_2* y *Fgig_2* analizamos la cobertura de las familias de TEs de *F. hepatica* y *F. gigantica*. Lo primero que observamos es que el tamaño de las familias al interior de cada genoma no es homogeneo (Supplementary Table 3 - coverage). Una forma de ver esto es que solo el 10% de las familias de TEs más grandes cubren más de la mitad del tamaño ocupado por todos los TEs en su conjunto para el correspondiente ensamblado genómico. Estas son unas 25 familias, principalmente correspondientes a retrotransposones clasificados como CR1, RTE, Bel-Pao y Ty3-Gypsy, cuyos tamaños oscilan entre 7 y 55 megabases de cobertura genómica.

Quisimos explorar además si las familias de TEs más prominentes presentan expansiones diferenciales al interior del género *Fasciola*. Si bien los ensamblados *Fhep_2* y *Fgig_2* tienen calidades similares, podrían aún existir pequeñas diferencias en la calidad que afecten los valores estimados de cobertura y otras variables cuantitativas (eso sí, en mucho menor medida que al comparar estos datos con los obtenidos a partir de los ensamblados de lecturas cortas). Por ello, en base a los datos disponibles, establecimos como criterio para identificar si una familia presenta una expansión especie-específica que exista una diferencia mayor al 20 % en la cobertura y una diferencia con signo opuesto también mayor al 20 % en la edad, al comparar los ensamblados de *Fhep_2* y *Fgig_2*. Este doble criterio permite controlar que el aumento de tamaño genómico esté acompañado por la aparición de una cantidad significativa de copias muy similares entre sí. Más aún, dado que nos interesaba explorar los cambios ocurridos en familias más prominentes, hicimos dicha comparación tomando únicamente las familias de gran tamaño (al menos 10 Mb en *Fgig_2*). Bajo estos criterios, identificamos cinco familias de TEs expandidas en *F. gigantica* y una expandida en *F. hepatica* (indicadas con sombra de color gris en la Table 3.2), todas ellas de gran tamaño y pertenecientes a retrotransposones del clado “sam6” de la superfamilia CR1 o de la superfamilia Bel-Pao. Este primer análisis sugiere fuertemente que existen además de las expansiones reportadas entre los géneros *Fasciola* y *Fasciolopsis*, importante expansiones al interior del género *Fasciola*. Queda por explorar mediante métodos filogenéticos cuales son las copias que conforman las subfamilias reciéntemente expandidas en las familias identificadas o bien en otras familias, así como su distribución en el genoma y el análisis a nivel transcriptómico para determinar si están activas.

Tabla 3.2 | **Expansiones especie-específicas de TEs en *Fasciola* spp.**

Classification	TE family-name	Coverage (bp)		Percent change	Age (median terminal branch length)		Percent change
		Fgig_2 (A)	*Fhep_2* (B)	100 x (A - B) / B	*Fgig_2* (C)	*Fhep_2* (D)	100 x (C - D) / D
LTR/Bel-Pao	Puma-4#	54,131,906	55,824,028	-3%	0.068	0.048	42%
LINE/RTE	Mantis-a-5#	37,163,546	41,160,299	-10%	0.105	0.078	35%
LINE/RTE	Mantis-a-7#	31,716,900	32,923,233	-4%	0.166	0.15	11%
LINE/CR1	sam6-13#	26,701,333	17,996,552	48%	0.051	0.081	-37%
LTR/Bel-Pao	Puma-1#	26,199,432	15,285,354	71%	0.057	0.096	-41%
LINE/CR1	sam6-16#	20,360,589	17,710,859	15%	0.067	0.108	-38%
LINE/RTE	Mantis-a-6#	20,338,724	22,746,264	-11%	0.163	0.153	7%
LINE/RTE	Mantis-a-2#	18,634,525	20,460,621	-9%	0.207	0.188	10%
LINE/CR1	Pana-2a#	18,139,950	18,635,821	-3%	0.084	0.09	-7%
LINE/CR1	sam6-1#	14,906,465	13,517,962	10%	0.071	0.112	-37%
LINE/CR1	sam6-15#	14,048,028	10,807,234	30%	0.088	0.116	-24%
LINE/CR1	sam6-6#	12,765,587	16,077,467	-21%	0.088	0.069	28%
LTR/Bel-Pao	Puma-7#	11,557,782	8,431,374	37%	0.052	0.079	-34%
LTR/Ty3-Gypsy	CSRN-like-10#	11,399,467	9,353,418	22%	0.05	0.052	-4%
LINE/CR1	sam6-3#	10,814,884	8,909,581	21%	0.056	0.085	-34%
LINE/CR1	FCR1-a#	10,687,473	11,474,636	-7%	0.09	0.064	41%
LINE/CR1	sam6-20#	10,658,700	8,068,147	32%	0.038	0.047	-19%
LTR/Ty3-Gypsy	CSRN-like-3#	10,091,084	9,353,418	8%	0.094	0.101	-7%

3.2.4 Dinámica evolutiva de los TEs

Los clados de TEs identificados al interior de las distintas superfamilias están representados por las tres especies; es decir, se componen en su gran mayoría por familias de TEs de *F. hepatica*, *F. gigantica* y *F. buski* (Figuras suplementarias 1 y 2). Esto sugiere que no solo las superfamilias analizadas estarían representadas en el ancestro común de Fasciolidae como sugerían análisis previos (Choi et al., 2020; Luo et al., 2021) sino también que la mayoría de los clados identificados en este trabajo estarían también presentes en dicho ancestro.

La mayor diversificación de *Fasciola* con respecto a *Fasciolopsis* (Figuras suplementarias 1 y 2) sugiere que existieron múltiples expansiones en la evolución para distintos tipos de repetidos. Mediante los gráficos de "paisaje de repeticiones" observamos múltiples eventos de expansión en el transcurso de la evolución de estos parásitos (Figura 3.6A). Cuando se analiza por separado las distintas superfamilias, se aprecia claramente que existieron múltiples eventos de gran expansión de TEs en distintos momentos evolutivos dependiendo el tipo de TEs (Figura 3.7). Por ejemplo, los retrotransposones SINE, RTE y Rex-Babar, y los transposones Tc1-Mariner presentan eventos de expansión relativamente antiguos ("picos" con valor de Kimura mayor a 10) mientras que otros transposones y retrotransposones presentan expansiones más recientes —aún cuando se analizan los ensamblados de lecturas cortas. Estos múltiples eventos de expansión habrían ocurrido en buena parte luego de la divergencia entre los linajes *Fasciolopsis* y *Fasciola*, dando lugar al incremento en el genoma de las últimas (Figura 3.2B). El hecho de que las familias de TEs de *F. hepatica* y *F. gigantica* son las mismas y comparten un perfil de composición genómica muy similar de acuerdo a los distintos análisis que mostramos, sugiere además que la mayor parte de dichas expansiones ya estaban presentes en el ancestro común de estas dos especies. Sin embargo, detectamos familias "muy jóvenes" de retrotransposones Bel-Pao, Ty3-Gypsy, CR1 y PLEs y transposones de ADN clasificados como CMC o Mutator-like (Figura 3.6B), sugieriendo que podrían presentar expansiones especie-específicas, al igual que vimos ya para algunas familias en concreto.

En resumen, nuestros analisis sugieren que distintos tipos de TEs estaban mayormente presentes en el ancestro común de Fasciolidae, sirviendo de base para las prominentes expansiones de distintas superfamilias de retrotransposones (y en menor medida de transposones de clase II) ocurridas tras la diversificación de *Fasciola* y *Fasciolopsis*. Más aún, la identificación de familias jóvenes y expansiones especie-específicas sugieren la presencia de elementos activos, lo que aún resta por ser confirmado.

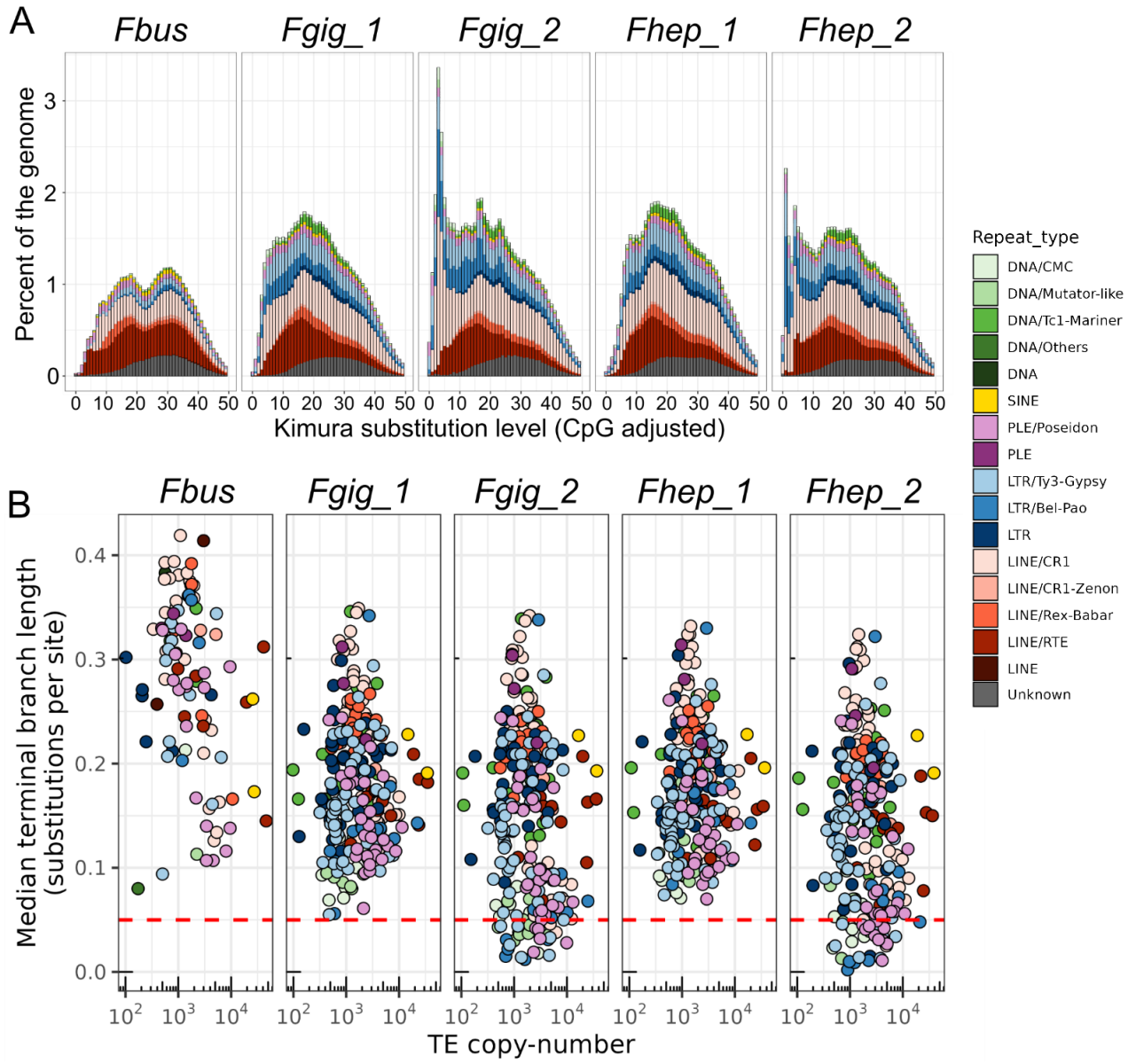


Figura 3.6 | **Dinámica evolutiva de los TEs.** Gráficos de paisaje de repeticiones ("repeat landscape plot") **(A).** Se indica en porcentaje de bases con respecto a tamaño total del genoma en función de la distancia genética de los consensos con respecto a las copias genómicas (inserciones) para distintos tipos de repetidos (bajo el modelo de sustitución de Kimura de 2 parámetros, excluyendo sitios CpG). La cobertura se obtiene sumando los tamaños de todas las inserciones en *bins* de una unidad de distancia genética, distinguiendo con color según el tipo de repetido. Edad de familias de TEs (**B**). Para cada familia de TEs se estimó la edad —en términos relativos— de acuerdo a la metodología descrita. El eje de las ordenadas indica las sutituciones nucleotídicas por sitio (valor medio de ramas terminales) y el eje de las abcisas el número de copias de la familia correspondiente.

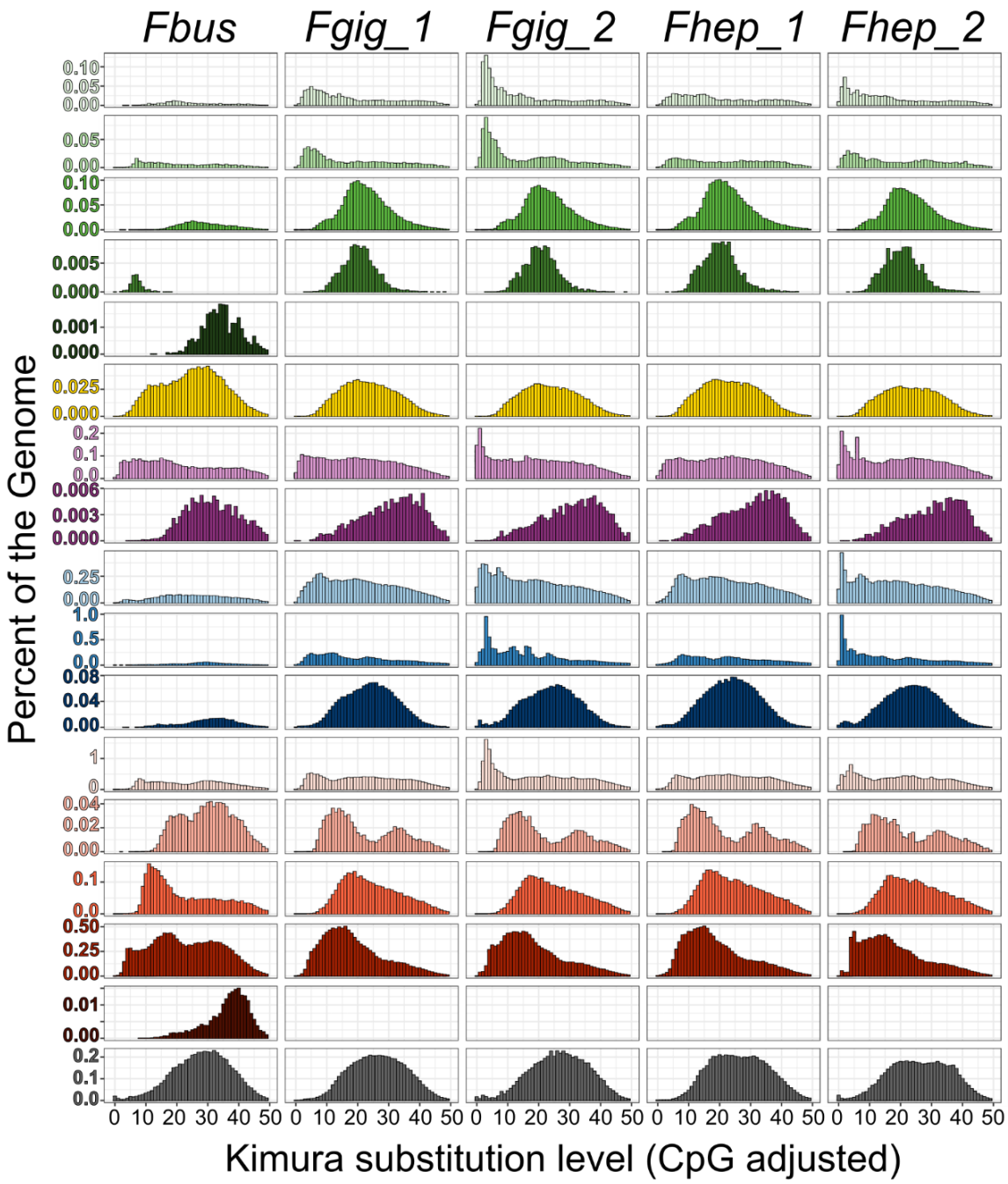


Figura 3.7 | **Gráfico de kimura.** Se indica para distintas superfamilias de TEs la cobertura total —medida como porcentaje del correspondiente ensamblado genómico— en función de la distancia genética de las copias (o inserciones) con respecto al consenso utilizado para la anotación (ver métodos por más detalles).

3.3 Impacto de TEs en genes codificantes de proteínas

3.3.1 Algunas superfamilias de TEs están enriquecidas en las regiones génicas.

Los elementos transponibles suelen presentar una distribución **no-aleatoria** en el genoma, encontrándose con mayor frecuencia en unas regiones que en otras, como resultado de un balance entre fuerzas selectivas contrapuestas que les permitan propagarse sin provocar efectos deletereos para la función celular (Bourque et al., 2018). Motivados por observaciones previas (Choi et al., 2020; Luo et al., 2021; Philippsen & DeMarco, 2019) y aprovechando nuestras bibliotecas curadas, quisimos investigar con mayor detalle la distribución de los TEs en relación a la localización genómica; más concretamente, en relación a las coordenadas de los intrones de los genes codificantes de proteínas.

Cuando analizamos el contenido genómico total, ya habíamos podido apreciar que algunas superfamilias se encuentran más enriquecidas en los genes que otras (Figura 3.2B), lo cual había sido también notado por los autores recién mencionados. Sin embargo, con el fin de realizar un análisis más preciso, calculamos para cada familia de TEs la fracción total que se solapa con las coordenadas de los intrones de genes codificantes, de acuerdo a los datos de anotación públicos disponibles para cada ensamblado genómico (Tabla 2.1). Los resultados de este análisis se muestran en la figura 3.8. Como generalidades, cabe notar que las posibles diferencias existentes en la calidad de los ensamblados y en la anotación de los genes, parecen no afectar sustancialmente los patrones generales observados. Además, que a nivel biológico (es decir, más allá de los aspectos técnicos) se observa el mismo patrón global entre las tres especies. Nuestro análisis fino, indica que todas las familias individuales dentro de una superfamilia tienden a presentar —en general— una fracción génica similar, tal como se describe a continuación.

Los retrotransposones no-LTR, la superfamilia RTE y los elementos SINEs están muy enriquecidos en los intrones (~40 a 60%), los PLE y Rex-Babar están medianamente enriquecidos (~20 a 40%), y los CR1 están prácticamente ausentes (menos del 10%, en general) (Figura 3.8). Los LTRs de la superfamilia Ty3-Gypsy no presentan un patrón de distribución claro, ni tampoco los solo-LTRs, en tanto que los de la superfamilia Bel-Pao sí, estando mayormente depletados de la región génica. En cuanto a los elementos de clase II, vemos que las familias del tipo Tc1-Mariner son las que están más enriquecidas al interior de los genes (~25 a 50%), mientras que las pertenecientes a las superfamilias CMC y Mutator-like están menos presentes en estas regiones (menos del 20%). Estos datos confirman que existe una distribución no-aleatoria de los TEs en relación a los genes.

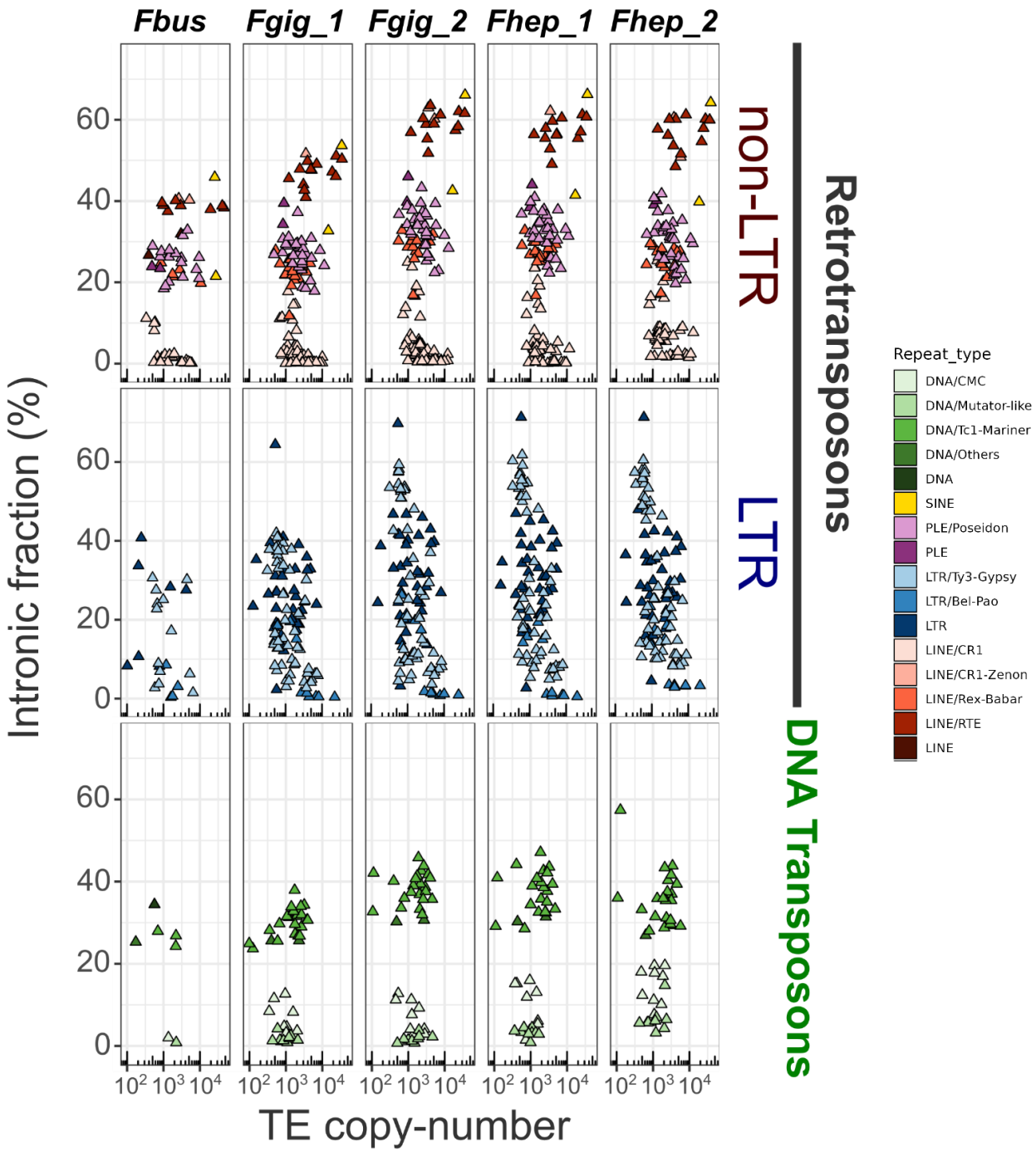


Figura 3.8 | **Fracción intrónica de TEs.** Cada triángulo representa la fracción intrónica de una familia de TEs. Esta se calculó midiendo la fracción solapada con las coordenadas de genes codificantes anotados en los correspondientes ensamblados. Se separan en tres paneles para una mejor visualización.

3.3.2 La inserción de TEs provoca un sesgo direccional en la arquitectura interna de los genes

Schistosoma mansoni, el trematodo mejor caracterizado, muestra un menor ocurrencia de TEs en intrones próximos al extremo 5´ de los genes. Según se ha sugerido, este patrón podría asociarse a una mayor selección negativa cercana a secuencias reguladoras en esa región, de forma que en el transcurso de la evolución se observe una disminución del tamaño de esos intrones en comparación a los intrones más internos o próximos al extremo 3´, contrario a lo observado en otros eucariotas (Philippsen & DeMarco, 2019). Quisimos explorar si este patrón estaba presente en fasciólidos, lo cual sugeriría que sería un patrón extendido en trematodos, y no solo en Schistosomatidae.

Nuestros resultados indican efectivamente que los intrones cercanos al extremo 5´ de los genes se encuentran menos enriquecidos en transposones, y asimismo, que presentan un tamaño medio más pequeño que la mayoría de intrones (paneles medio y superior de: Figura 3.9). Observamos que dicho patrón está más acentuado en *Fasciola* que en *Fasciolopsis*, pero nos resta por analizar aún todos los ensamblados de *Fasciola* spp. para corroborar que esto último no sea un sesgo metodológico asociado a los ensamblados o la anotación. Resulta interesante que se observa una pequeña caida en la proporción de intrones con ocurrencia de TEs (panel medio de la figura) acompañada en paralelo de una caida en el largo medio de los intrones (panel superior) a partir del intron 6 aproximadamente. Más allá de esta caida, los primeros dos intrones continúan siendo los más depletados de TEs en todas las especies. Estos datos sugieren que la acumulación preferencial de TEs en el extremo 3´ genera un sesgo direccional en los tamaños de los intrones de los genes, con intrones más pequeños hacia el extremo 5´.

Llamativamente no todos los tipos de TE presentan el mismo patrón de distribución al interior de los genes (Figura 3.9, panel inferior). Mientras que los repetidos mayoritarios tales como superfamilia RTE siguen —como cabría esperar— el mismo patrón global de los TEs, algunas superfamilias presentan patrones distintos e incluso opuestos. Por ejemplo, la superfamilias de retrotransposones Ty3-Gypsy y CR1 están mucho más enriquecidas en el extremo 5´ que hacia el 3´ de los genes. Por otra parte, los elementos SINE tienen una distribución similar a los RTE, pero no así a otros elementos LINE. Queda pendiente determinar la importancia funcional y evolutiva de estas observaciones.

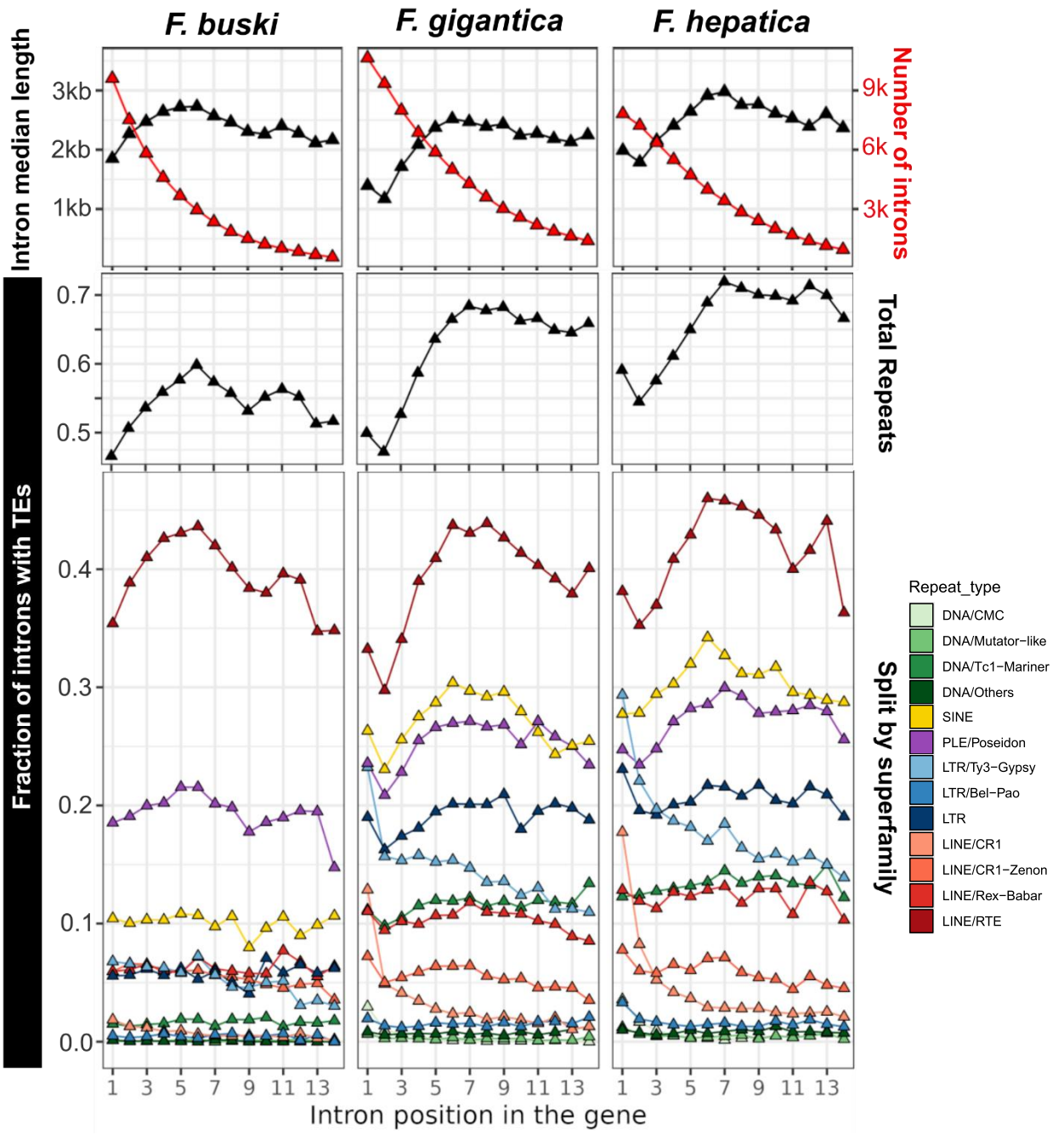


Figura 3.9 | **Inserciones de TEs de acuerdo a la posición de los intrones.** Cada triángulo en los paneles medio e inferior representan el porcentaje de intrones en la posición indicada que intersectan con al menos un TE del tipo que sea (panel medio) o la fracción que solapa con al menos un TE de la superfamilia indicada (panel inferior). En la parte superior de la imagen se indican el número de intrones y tamaño medio en pares de bases de los mismos.

Capítulo 4.

Discusión y perspectivas

F. hepatica es un parásito conocido hace más de 600 años, pero las secuencias genómicas de este y otros fasciólidos se obtuvieron recién en la última década (Choi et al., 2020; Cwiklinski et al., 2015; Luo et al., 2021; McNulty et al., 2017; Pandey et al., 2020). Los estudios morfológicos, ecológicos y moleculares permiten actualmente tener una idea más clara de la historia evolutiva la familia Fasciolidae, revelando que las innovaciones adaptativas más prominentes al interior de este grupo habrían ocurrido hace aproximadamente 56 a 65 millones de años, en el linaje que condujo a la aparición de la subfamilia Fasciolinae (Figura 1.3) (Choi et al., 2020; Lotfy et al., 2008).

Los ensamblados genómicos han permitido realizar estudios comparativos adicionales entre las especies hepáticas (*F. hepatica* y *F. gigantica*) e intestinales (*F. buski*) (Choi et al., 2020). Entre otros hallazgos, se encontró que ciertas familias multigénicas relevantes para el parasitismo (por ejemplo, las catepsinas L y B) se encuentran expandidas en *Fasciola*, si bien el número de genes totales en estas especies no cambia de manera sustantiva (Choi et al., 2020). Paralelamente, se determinó que hay un incremento considerable de ADN repetitivo en las especies hepáticas, debido principalmente a la expansión de TEs (Choi et al., 2020). **¿Qué rol han tenido las expansiones de TEs en la evolución de estos helmintos?** Algunos autores sugieren que la acumulación desproporcionada de retrotransposones LINE en regiones intrónicas durante ciertos períodos evolutivos particulares, podría asociarse con algunas de las innovaciones adaptativas de estos parásitos (Luo et al., 2021).

En el presente trabajo nos propusimos **caracterizar el contenido de TEs en Fasciolidae** de manera más detallada, como un primer avance hacia la comprensión del papel evolutivo de los elementos móviles en las adaptaciones parasitarias de esta familia. Para ello, primero hicimos una búsqueda de las familias de TEs en los genomas de *F. hepatica*, *F. gigantica* y *F. buski* (Tabla 2.1) con `RepeatModeler`. Las bibliotecas automáticamente generadas se curaron manualmente (Goubert et al., 2022) antes de realizar la anotación de las coordenadas genómicas con `RepeatMasker`. La búsqueda por homología usando estas bibliotecas curadas se realizó sobre cinco ensamblados en total (Tabla 2.1), lo que permitió ponderar mejor los sesgos metodológicos.

Las **bibliotecas obtenidas** (Supplementary File 2) se componen de secuencias consenso curadas de las familias de TEs, así como de algunos tipos de repetidos adicionales (Zhang et al., 2020). Disponer de modelos para cada familia de TEs, nos permitió comparar en gran detalle la diversidad de secuenicas móviles presentes en fasciólidos a nivel filogenético (Figuras suplementarias 1 y 2), así como a niveles estructural y funcional (Figura 3.1). Esta metodología permitió realizar una **clasificación confiable** —basadas en evidencias estructurales y filogenéticas— de la mayoría de las familias de elementos móviles, corroborando así la existencia de distintos órdenes/superfamilias de TEs en fasciólidos que habían sido previamente reportados únicamente en base a métodos automáticos (Choi et al., 2020; Luo et al., 2021).

Pudimos además identificar distintos **clados de TEs** al interior de las superfamilias, lo que supone un análisis más fino de la diversidad de elementos presentes en estos genomas. No solo eso, sino que las comparaciones filogenéticas aportaron pistas sobre la historia evolutiva de los TEs. El hecho de que casi todos los clados de TEs estén compuestos por repetidos de las tres especies analizadas —si bien presentan más familias de *Fasciola*—, sugiere que existieron familias de TEs ancestrales a cada uno de los mismos, las cuales habrían experimentado múltiples expansiones en distintos momentos de la evolución (Figura 3.7) acompañadas de una diversificación de las secuencias y aparición de nuevas familias, junto al incremento en la cobertura y número de copias en los genomas de las especies hepáticas (Figura 3.2). Estas observaciones se contraponen a la hipótesis de una posible adquisición de TEs por transferencia horizontal, por ejemplo a partir de alguno de los hospederos de estos parásitos (como sí parece haber ocurrido en otros helmintos (Suh, 2016)). Sin embargo, no descartamos completamente que haya ocurrido algún evento de estas características. En tal caso la evidencia indica que la transferencia horizontal —de existir— habría tenido un efecto menor, y que la gran mayoría de TEs se habrían expandido a partir de familias adquiridas por herencia vertical a partir del ancestro común de Fasciolidae en el linaje que condujo al género *Fasciola*. Además, el hecho de que *F. hepatica* y *F. gigantica* tengan las mismas familias (Figura 3.1A) indica que las aparición de estas familias estaban ya presentes en el último ancestro común de ambas especies.

Nuestros resultados coinciden con observaciones previas que apuntan a una gran expansión de TEs, algunos de los cuales habrían ocurrido en la ventana temporal en que aparecieron los rasgos apomórficos novdeosos que caracterizan a las especies hepáticas de la familia. Se ha propuesto que la inserción desbalanceada de LINEs en regiones génicas podría haber conferido una ventaja adaptativa a los parásitos en períodos concretos de la evolución (Luo et al., 2021). En base a análisis de enriquecimiento de funciones, los autores sugieren que las inserciones de LINE estarían sobrerrepresentadas en genes asociados a proteínas de membrana o proteínas asociadas a membrana, pudiendo haber alterado el transporte de sustancias a través de membranas y la transducción de señales.

Con esos antecedentes, quisimos calcular en más detalle la distribución de TEs en relación a los genes. Así, calculamos la distribución de cada familia individual en relación a las coordenadas de los intrones, y estudiamos **qué superfamilias están enriquecidas en los mismos**. En términos globales observamos un patrón de inserción similar entre las tres especies, lo cual se debe en buena parte a que muchas inserciones son antiguas (Figura 3.7) y por ende ancestrales a más de un taxón (es decir, ancestrales a *Fasciola* o ancestrales a Fasciolidae).

Vimos que distintas superfamilias están enriquecidas en distinta proporción en la región de los intrones (Figura 3.8). Además, las superfamilias mayormente depletadas en intrones parecerían coincidir con las que presentan expansiones más recientes (por ejemplo, CR1 y Bel-Pao), y aquellas más antiguas (RTE, Rex-Babar) parecen estar más enriquecidas en los intrones (Figura 3.6B). Aún así, sigue sin ser del todo claro que exista una correlación entre la edad de las familias y el enriquecimiento en los intrones, ya que por ejemplo, los CR1 se componen de familias de todas las edades pero están en su mayor parte depletados.

¿A qué se debe entonces que algunas familias estén sobrerepresentadas en los intrones? Como hipótesis posibles planteamos que la mayor proporción de elementos de superfamilias particulares en las regiones génicas podría ser debida a múltiples factores: los distintos mecanismos de transposición (Wells & Feschotte, 2020), las diferencias en los tamaños de los elementos completos que se movilizan (Figura 3.1B),

las edades de las inserciones, y/o el balance de fuerzas purificadoras con relación a las ventajas selectivas de las inserciones. Nos resta aún por determinar la implicancia funcional y evolutiva de la distribución heterogenea de las distintas superfamilias en relación a los genes. En resumen, estos análisis dan lugar a una apreciación más fina en relación a la distribución de los TEs en genes. Resta aún por determinar si un conjunto particular de genes o categoría funcional determinada está enriquecida entre los genes que tienen inserciones de uno u otro tipo de TEs.

La inserción de los TEs en los intrones ha producido un aumento del tamaño de los mismos en *Fasciola* (Choi et al., 2020; Luo et al., 2021). Sin embargo, se ha visto en otros trematodos que las inserciones ocurren mayormente en los intrones localizados hacia el extremo 3´ del gen. Se piensa que estas inserciones provocaron la asimetría de la estructura de los genes, caracaterizada por intrones más pequeños hacia el extremo 5´ prima (Philippsen & DeMarco, 2019). Quisimos verificar si esto también se observaba en nuestros organismos de estudio.

Efectivametne, nuestros datos verifican que hay una mayor fracción de intrones con inserciones de TEs hacia el extremo 3´ de los genes de fasciólidos (Figura 3.9: "Total Repeats"). Esto coincide con el hecho de que el largo medio de los primeros intrones sean pequeños, lo cual apoya la hipótesis de **que la inserción heterogenea de transposones habría alterado la estructura de los genes en Fasciolidae**. Adicionalmente, nuestros resultados indican que más allá del patrón general, no todas las superfamilias presentan el mismo tipo de distribución (Figura 3.9: "Split by superfamily"). Es así que algunas superfamilias como las RTE suelen estar sobrerepresentadas hacia el 3´ (al menos en *Fasciola*) mientras que los CR1 o Ty3-Gypsy están más presentes en los primeros intrones. Desconocemos las causas de dicha distribución heterogenea. Reportes previos daban ya cuenta de un incremento en el tamaño de los intrones en *Fasciola* con respecto a *Fasciolopsis*, posiblemente dado por la mayor proporción de TE resultantes de la expansión de estos repetidos luego de la divergencia entre ambos linajes (Choi et al., 2020). Nuestro trabajo aporta datos adicionales relativos a las inserciones intragénicas, y las observaciones generales expanden algunas observaciones realizadas en trematodos sanguíneos (Philippsen & DeMarco, 2019), lo cual podría sugerir un patrón general de arquitectura génica de trematodos, con intrones más pequeños hacia el extremo 5`.

La movilización de los transposones es fuente importante de mutaciones representando así una amenaza para el hospedero. En modelos experimentales como la mosca (*Drosophila melanogaster*) y el ratón (*Mus musculus*) se ha observado respectivamente que en el entorno del 10% y 50% de los fenotipos mutantes aislados del laboratorio son el resultado de inserciones *de novo* de TEs (Cosby et al., 2019). De forma análoga, existen unas 120 patologías humanas de herencia monogénica descritas, cuyo orígen es atribuído a eventos de transposición *de novo* (Cosby et al., 2019). Estas características han dado orígen a denominaciones tales como secuencias de "ADN egoísta" o "ADN parasitario" para referir a los TEs. Por tal motivo los organismos suelen tener sistemas de defensa, ya que la movilización descontrolada tendría efectos deletereos para la función de la célula. Curiosamente, el principal mecanismo conocido de silenciamiento de TEs (la vía Piwi) se encuentra incompleto en los platelmintos parásitos, sugiriéndose que mecanismos alternativos para silenciar los TEs operarían en estos organismos, los cuales estarían aparentemente basados en un clado novedoso de proteínas Argonautas catacterísticos de estos gusanos (Fontenla et al., 2017; Skinner et al., 2014).

Contrariamente a lo que ocurre a nivel individual, a nivel poblacional una mayor movilización de TEs podría implicar una fuente de variabilidad genética heredable, en el caso de ocurrir en la línea germinal. En este sentido, tal como sugiriera McClintock (McClintock, 1984), la mayor expresión y movilización de TEs podría facilitar la adaptación de las poblaciones a ambientes cambiantes y condiciones de estrés (Chénais et al., 2012). Con esto en mente, quisimos saber si existían indicios de elementos activos en la actualidad (o en tiempos recientes). Efectivamente, encontramos familias muy jóvenes pertenecientes a retrotransposones PLE, LINE/CR1 y LTR así como algunos elementos de clase II (Figura 3.6B). Más aún, encontramos que algunas de las familias más prominentes están diferencialemente expandidas entre *F. hepatica* y *F. gigantica* (Tabla 3.2). Estos resultados indican: 1) que si bien estas dos especies presentan las mismas familias de TEs, existen algunas de ellas diferencialmente expandidas entre ambas (resta por realizar un análisis filogenético más fino para determinar qué inserciones/subfamilias serían las responsables de estas diferencias); 2) que *Fasciola* tendría *a priori* una mayor plasticidad adaptativa que *Fasciolopsis* basados en los efectos de los TEs comentados; 3) que podrían existir familias activas en *F. hepatica* y *F. gigantica* (tener en cuenta además que la mayoría de las familias curadas —incluyendo las "jóvenes"— presentan todos los dominios proteicos necesarios para la movilización (Figura 3.1D).

Cabe destacar la importancia de haber incluído ensamblados realizados a partir de lecturas largas en nuestros análsis. De acuerdo a otros autores (Luo et al., 2021), la hipótesis más razonable para explicar el incremento en la cobertura entre ensamblados de distinta calidad se debe a que los *contigs* generados con lecturas largas permiten cubrir regiones más grandes de repetidos, las cuales se encontrarían colapsadas en los *contigs* construidos en base a lecturas cortas. Por esto suponemos que el uso de ensamblados de lecturas largas de buena calidad nos proporcionó una estimación más realista del tamaño genómico de las familias de TEs. Además, vimos que los eventos recientes de transposición no fueron correctamente detectados usando los ensamblados anteriores (Figuras 3.7 y 3.8). De hecho, en un trabajo anterior en cual no se disponía de estos nuevos ensamblados, se llegó a concluir que había "evindencia rasonable" que indicaba la "baja actividad de TEs en la actualidad". En nuestro análisis también hubieramos llegado a dicha conclusión a partir del análisis exclusivo de los ensamblados de lecturas cortas, dado que fue gracias a los datos de lecturas largas que pudimos revelar la existencia de eventos muy recientes de transposición (Figuras 3.7 y 3.8).

La mutagenicidad de los TEs puede ocurrir también por elementos inactivos, por ejemplo, promoviendo rearreglos genómicos entre inserciones dispersas, generando distintos tipos de variantes estructurales (Bourque et al., 2018). En este sentido cabe notar que muchos de los LTR, así como alguna de las familias Tc1-Mariner, presentaron una expansión mucho mayor de los repetidos terminales internos en comparación al número de copias del elemento autónomo completo (ver ejemplo en Figura 2.2). Además, los elementos identificados como solo-LTRs —representados con color azul oscuro en las figuras, como Figura 3.7— así como los Tc1-Mariner, presentan picos de expansión relativamente antiguos (distancia de Kimura entre 20-25). Otros elementos no-autónomos que muestran un pico de expansión en una ventana temporal similar, son los SINEs. Es importante resaltar que las familias SINEs están entre las más abundantes cuando se analiza el número de copias (Supplementary Table 5 - TEcopies) y que estos suelen localizarse al interior de los genes. Este conjunto de observaciones parece indicar que los elementos no-autónomos e inactivos han tenido también un impacto en la expansión y arquitectura del genoma, restando por caracterizar y evaluar el impacto funcional de dichas alteraciones estructurales.

Más allá de los TEs, cabe mencionar que se identificó una proporción llamativa de **ADN satélites**, sobre todo en el ensamblado de lecturas largas de *F. hepatica* (*Fhep_2*) (Figura 3.2B). Los motivos de las diferencias en la cantidad de ADN satélite entre *Fhep_2* y *Fgig_2* no son claros, pero podrían deberse a diferencias en la calidad de los *contigs* (Tabla 2.1). En cualquier caso, es evidente que la cantidad de ADN satélite estimada también aumenta en ambas especies al comparar con los ensamblados de lecturas cortas correspondientes (*Fhep_1* y *Fgig_1*). Dado que el foco de nuestro trabajo eran los TEs, no realizamos mayores análsis de estos repetidos. Sin embargo, observaciones preliminares parecen indicar una co-localización de las principales familias multigénicas expandidas en *Fasciola* (Choi et al., 2020) con las coordenadas de ADN repetidos en tandem. A modo de ejemplo, mostramos la región de uno de los *clusters* de catepsinas L en uno de los ensamblados genómicos estudiados (Figura suplementaria 4). Estas observaciones son sugerentes respecto a que la duplicación en tandem de estas familias multigénicas podría haber estado mediada por repetidos en tandem próximales, si bien para poder realizar una afirmación de este tipo faltaría realizar un análisis mucho más detallado y exhaustivo.

En resumen, nuestro trabajo aportó información valiosa y detallada relativa al contenido de repetidos en fasciólidos, permitiendo evaluar con mayor precisión las similitudes y diferencias entre especies hepáticas e intestinales. Mucho trabajo es necesario aún para poder determinar si existe alguna asociación entre las expansiones de TEs y las innovaciones evolutivas ocurridas al interior de esta familia. La irrupción de las tecnologías de secuenciación masiva está transformando todas las áreas de la biología, pero la acumulación expansiva de datos multi-ómicos requiere de recursos humanos capacitados para su análisis. En tal sentido, esperamos que los organismos financiadores y las comunidades científicas sigan volcando recursos y tiempo al estudio de las enfermedades desatendidas, tales como las causadas por helmintos parásitos. Los frutos —ya notorios— de estos esfuerzos económicos y humanos, apenas comienzan a vislumbrarse.

Capítulo 5

Anexos

5.1 Figuras suplementarias

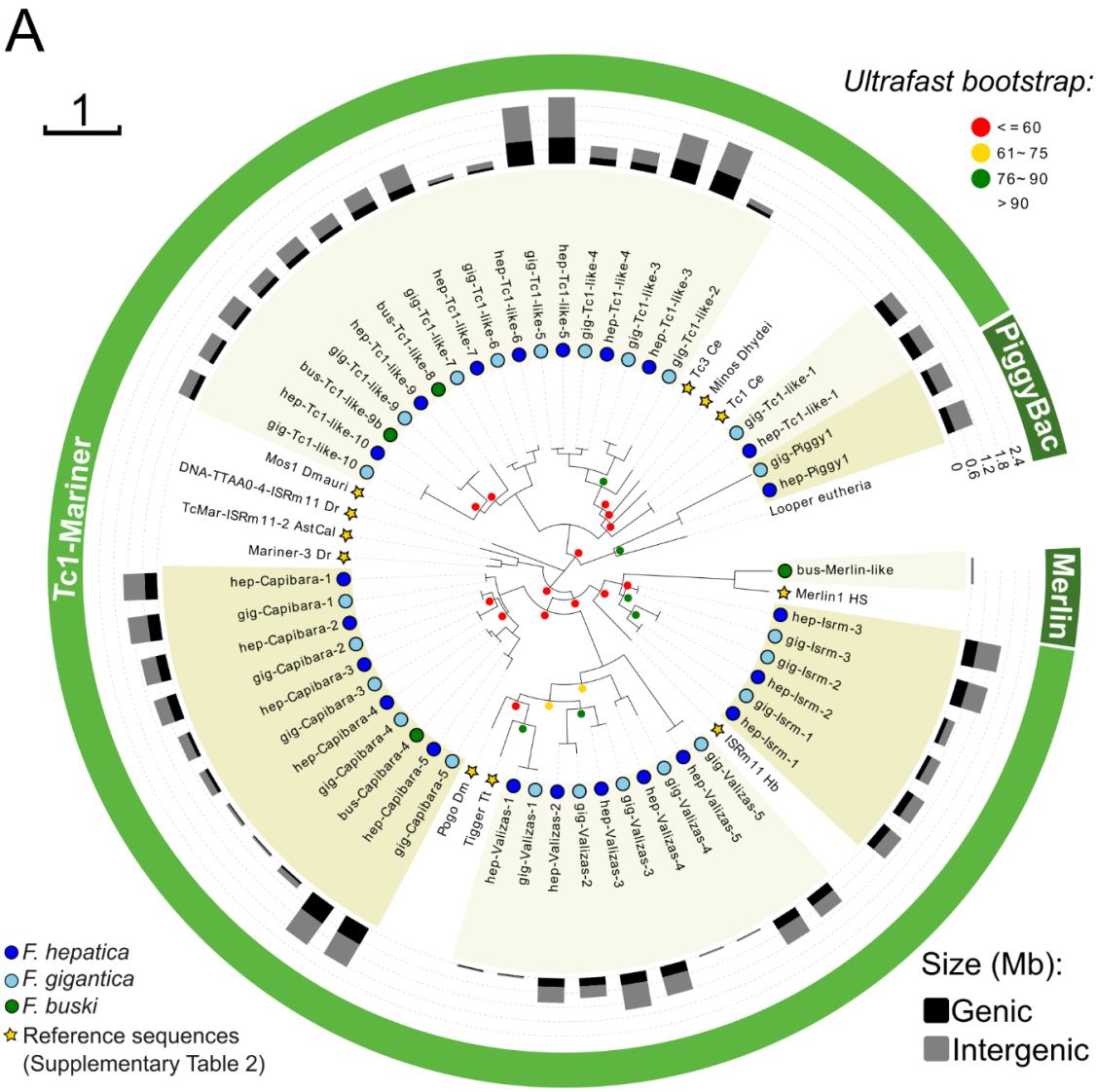
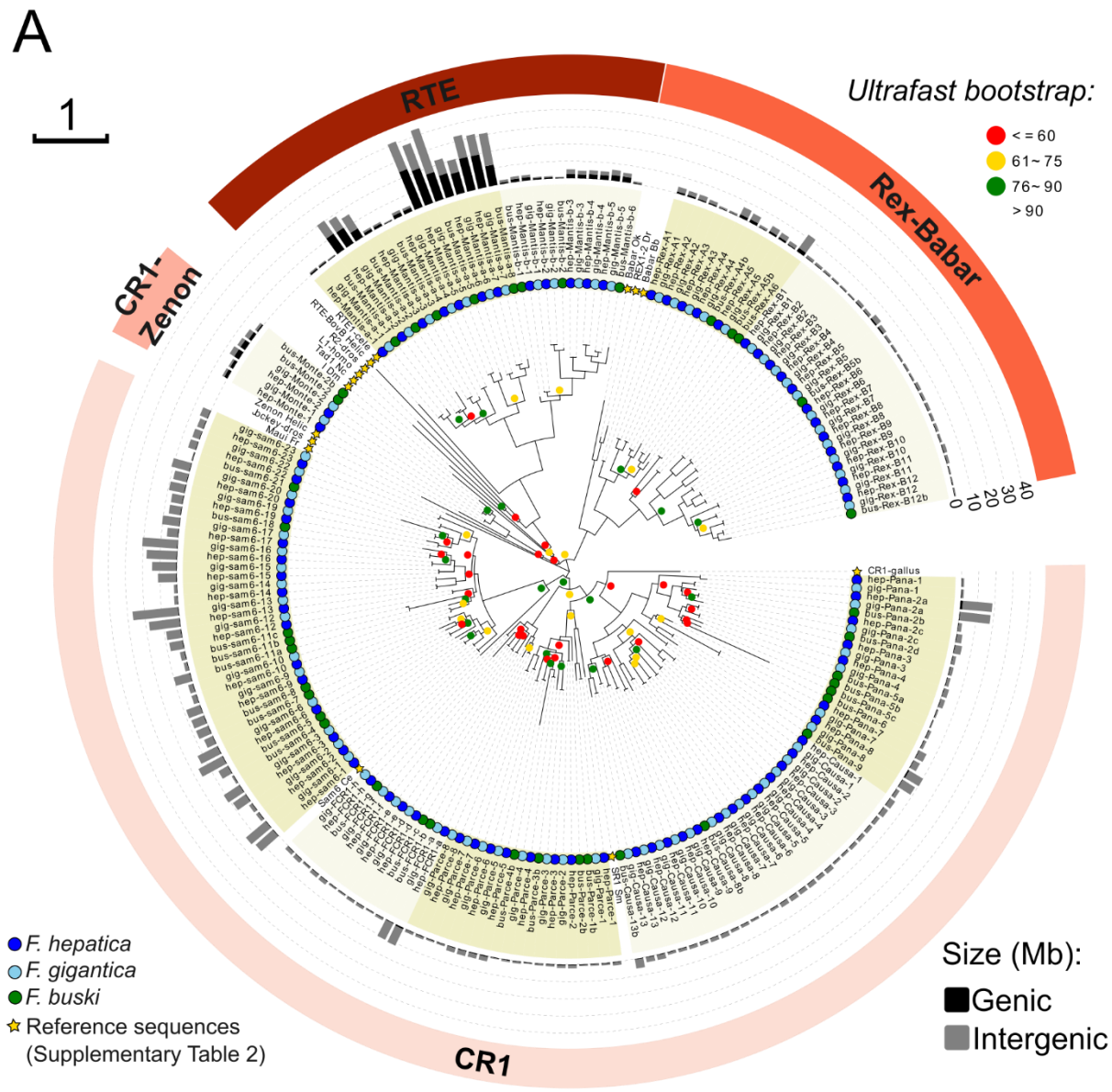
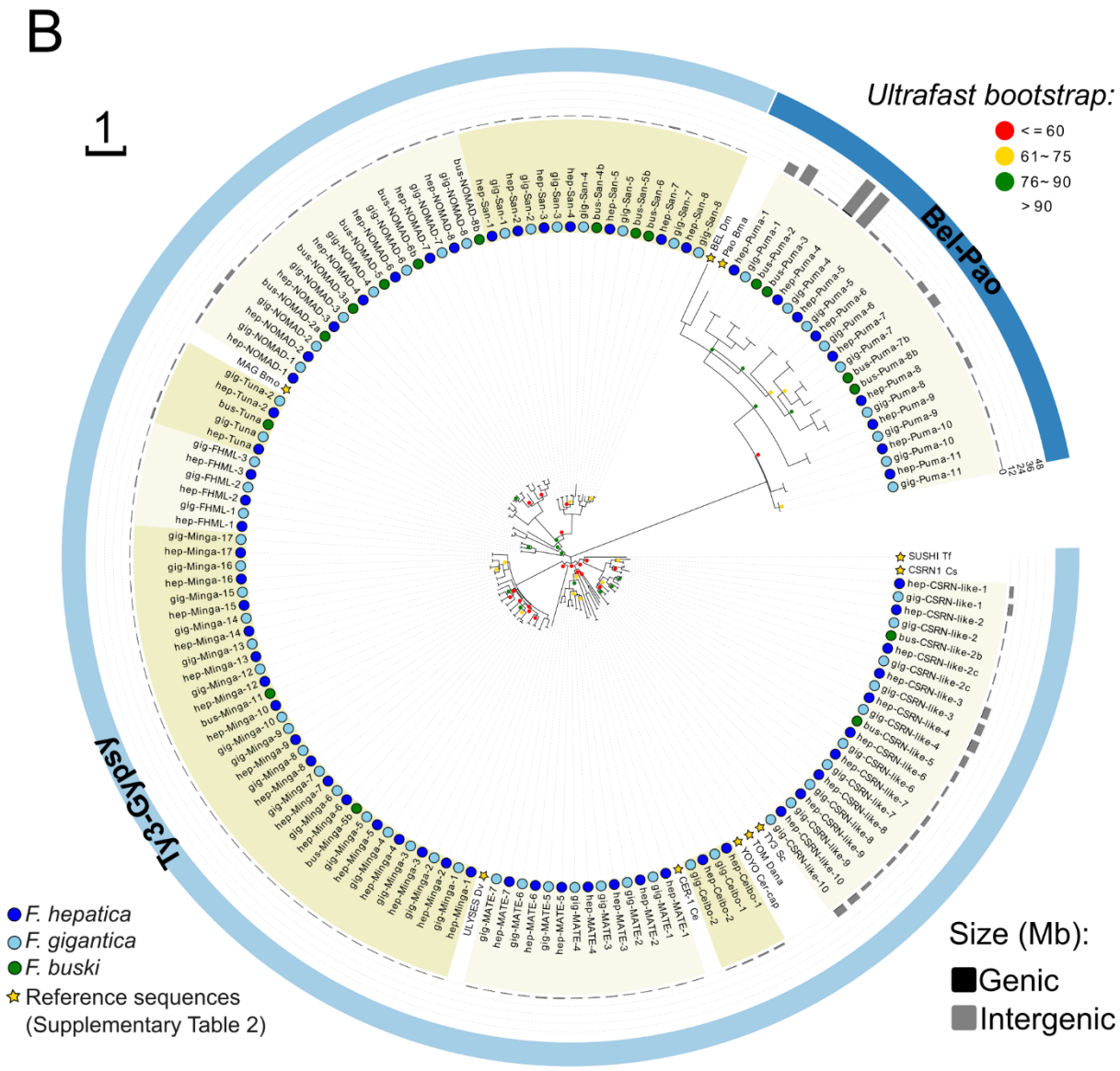


Figura Suplementaria 1 | **Filogenia de transposones de ADN.** A partir de la secuencia aminoacídica de la transposasa de los consensos curados se estimó la filogenia mediante el método de máxima verosimilitud. Las secuencias de referencia (indicadas con estrellas doradas) pertenecen a las superfamilias indicadas en los círculos de color externos. Por ejemplo, tanto "Tigger Tt" como "Tc1 Ce" son de la superfamilia Tc1-Mariner. Estas fueron usadas para guiar la clasificación de los consensos curados. Los tamaños —cobertura genómica en megabases— de cada familia se indican con barras de color gris (región intergénica) y negro (región intrónica). En tonos *beige* se agrupan familias pertenecientes a clados de TEs bien distinguibles, con un soporte estadístico (*ultrafast bootstrap*) en el entorno de 90 tomando el nodo más basal. A las familias de un mismo clado se les asignaron nombres similares, pero el sufijo numérico no es indicativo de la similitud entre las mismas. La escala en la parte superior izquierda indica las sustituciones aminoacídicas por sitio.


A
1
RTE
Rex-Babar
CR1-Zenon
CR1
Ultrafast bootstrap:
< = 60
61~ 75
76~ 90
> 90
0
10
20
30
40
F. hepatica
F. gigantica
F. buski
Reference sequences
(Supplementary Table 2)
Size (Mb):
Genic
Intergenic


B
1
Ultrafast bootstrap:
< = 60
61~ 75
76~ 90
> 90
Bel-Pao
Ty3-Gypsy
F. hepatica
F. gigantica
F. buski
Reference sequences
(Supplementary Table 2)
Size (Mb):
Genic
Intergenic

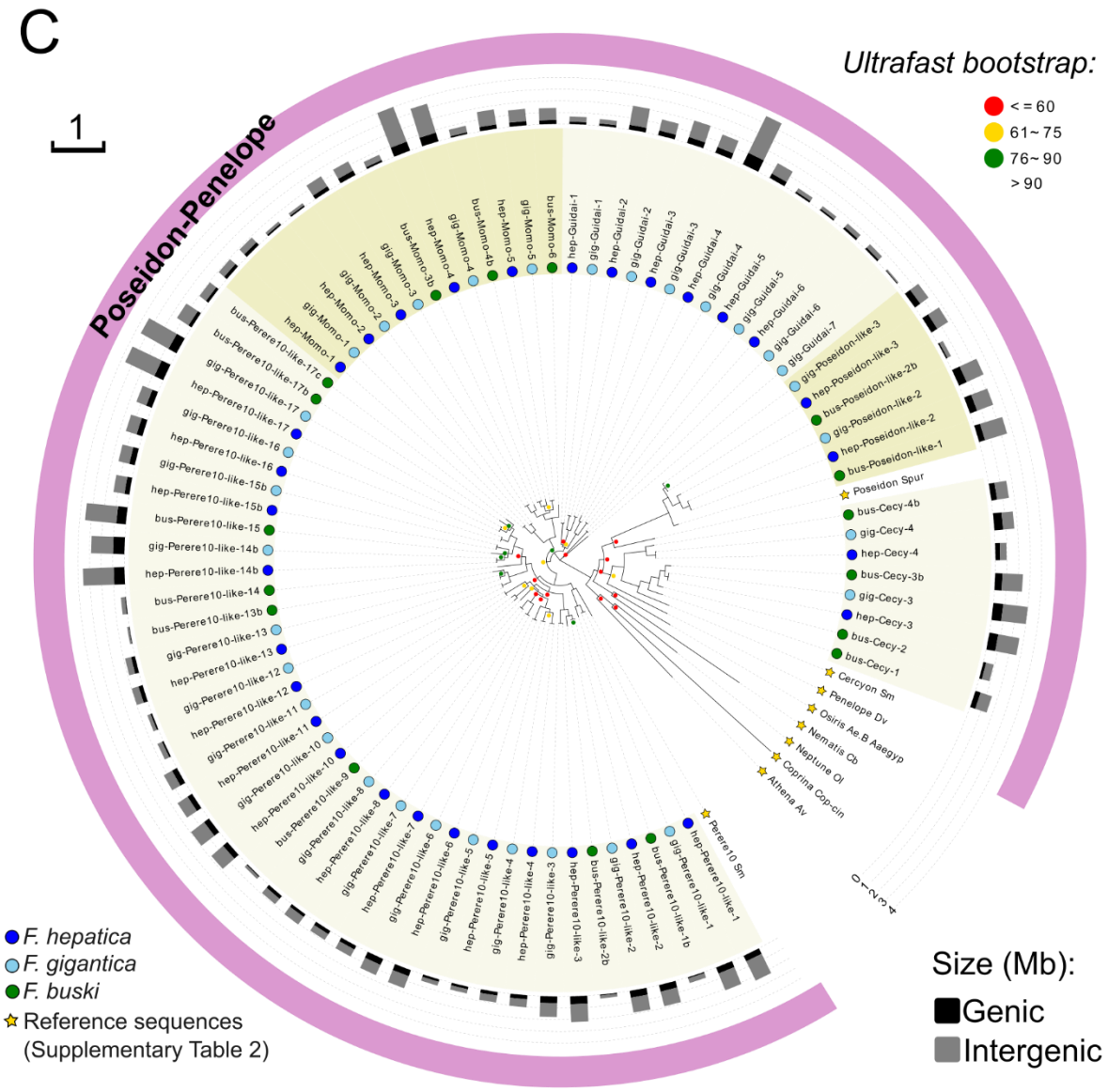


Figura Suplementaria 2 | **Filogenia de LINEs (A), LTRs (B) y PLEs (C).** A partir de la secuencia aminoacídica de la transcriptasa reversa de los consensos curados se estimaron las filogenias mediante el método de máxima verosimilitud. Las secuencias de referencia (indicadas con estrellas doradas) pertenecen a las superfamilias indicadas en los círculos de color externos y fueron usadas para clasificar los consensos curados. Los tamaños —cobertura genómica en megabases— de cada familia se indican con barras de color gris (región intergénica) y negro (región intrónica). En tonos *beige* se agrupan familias pertenecientes a clados de TEs bien distinguibles, con un soporte estadístico (*ultrafast bootstrap*) en el entorno de 90 tomando el nodo más basal. A las familias de un mismo clado se les asignaron nombres similares, pero el sufijo numérico no es indicativo de la similitud entre las mismas. Las escala en la parte superior izquierda de cada filogenia indica las sustituciones aminoacídicas por sitio.

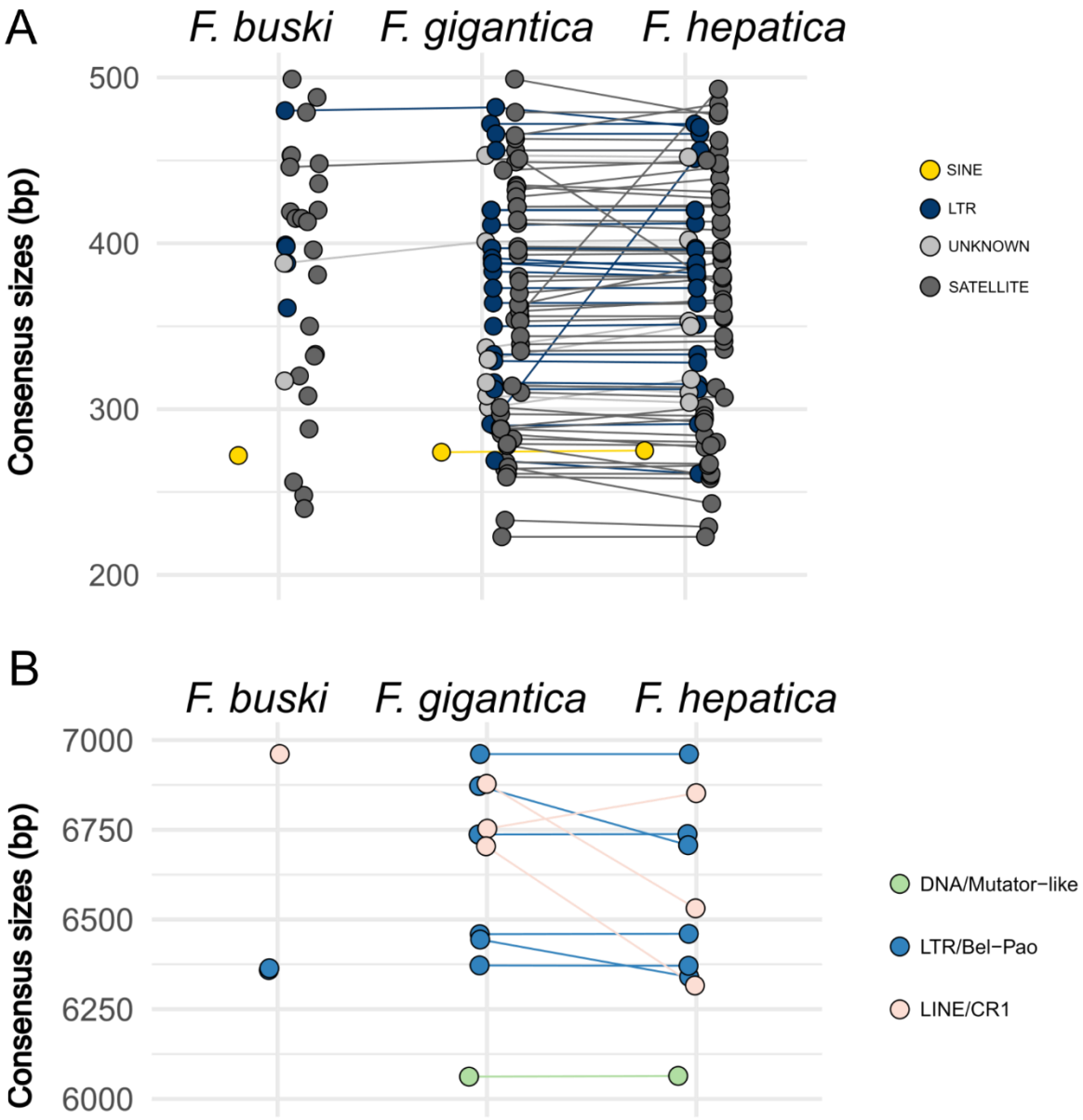


Figura Suplementaria 3 | **Intersección entre bibliotecas curadas.** Se muestran con mayor detalle los consensos o familias compartidas entre las bibliotecas curadas para tamaños que van desde 0.2 a 0.5 kb (**A**) y desde 6 a 7 kb (**B**). Cada círculo representa una secuencia consenso. La clasificación de los mismos se indica a la derecha. Dos círculos cualquiera conectados por una línea corresponden a una misma familia, que en la biblioteca (Supplementary File 2) se distinguen por el uso de distintos prefijos ("bus-" , "gig-" o "hep"-).

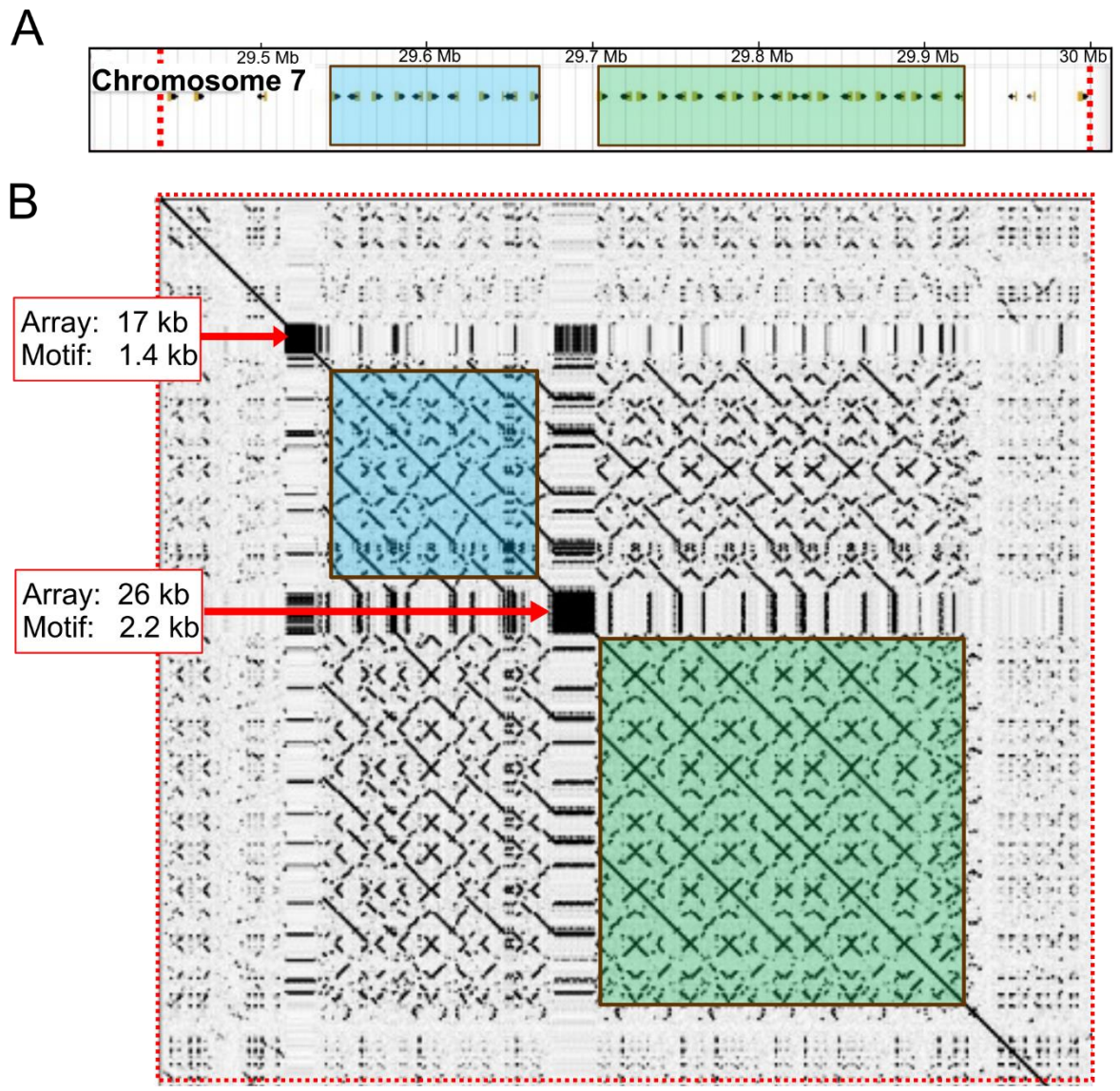


Figura Suplementaria 4 | **Co-localización de ADN satélites y *clusters* de catepsinas**. Se muestra a modo de ejemplo de la co-localización genómica de repetidos en tandem con familias multi-génicas, una región genómica de uno de los *clusters* más grandes de catepsinas L. En el panel superior **(A)** se muestran las coordenadas de los genes de catepsina L en el cromosoma 7 del ensamblado *Fgig_2*, que fueron identificadas mediante búsqueda por homología usando secuencias aminoacídicas de catepsinas conocidas (mediante `tblastn`). En el panel inferior **(B)** se muestra la estructura de repetidos de esta región; imagen obtenida por *dotplot* con la herramienta `Gepard`. Se destacan en colores celeste y verde la región en donde se localizan la mayor parte de los genes. Se observa en los bordes de dichas regiones arreglos de un repetido en tandem complejo de tamaño 17 o 26 kb con motivos de repetición relativamente grandes (mayores a 1.4 y 2.2 kb respectivamente). Además, en las coordenadas internas de dichos bloques a la altura de dichos arreglos de satélite, se observan varias "lineas" negras que corresponden a duplicaciones de estos repetidos. Estas duplicaciones se intercalan con las posiciones de los genes duplicados. Por último las "cruces" en el *dotplot* indican que hubo varias duplicaciones e inversiones en forma simultanea. Similarmente, en el panel superior se muestra que la mayoría de los genes están duplicados en tandem de forma invertida, es decir un patron de duplicaciones adyacentes orientadas "cabeza con cola".

5.2 Tablas suplementarias

Supplementary Table 1 | **Secuencias de referencia.** Se muestra la clasificación —con el esquema "DFAM"—de las familias utilizadas como referencia, incluídas en los árboles de: i) LINE, ii) PLE, iii) LTR o iv) DNA. También se indican las bases de datos de donde se obtuvieron las mismas y el correspondiente número de acceso. Las abreviaturas indicadas en la primer columna se corresponden con las mostradas en las figuras de los árboles generados con `evolview`.

i) LINE:

Abbraviation	Organism	Class	Order	Type	Subtype	Database	Accession
I_Dm	*Drosophila melanogaster*	I_Retrotransposition	LINE	R1-group	I	NCBI	M14954.2
Babar_Bb	*Battrachocottus baikalensis*	I_Retrotransposition	LINE	CR1-group	Rex-Babar	NCBI	U18939.1
Babar_Ok	*Oncorhynchus keta*	I_Retrotransposition	LINE	CR1-group	Rex-Babar	NCBI	AF063216
REX1-2_Dr	*Danio rerio*	I_Retrotransposition	LINE	CR1-group	Rex-Babar	Dfam (release 3.7)	DF0002295
SR1_Sm	*Schistosoma mansoni*	I_Retrotransposition	LINE	CR1-group	CR1	NCBI	U66331.1
Zenon_Heli	Heliconius butterfly genus	I_Retrotransposition	LINE	CR1-group	CR1-Zenon	Dfam (release 3.7)	DF0006357
Sam6_Ce	*Caenorhabditis elegans*	I_Retrotransposition	LINE	CR1-group	CR1	NCBI	Z82275.1
Maui_Fr	*Fugu rubripes*	I_Retrotransposition	LINE	CR1-group		NCBI	AF086712.1
Tad1_Nc	*Neospora crassa*	I_Retrotransposition	LINE	R1-like	Tad1	NCBI	L25662.1
RTE-BovB_Heli	Heliconius butterfly genus	I_Retrotransposition	LINE	RTE-group	BovB	Dfam (release 3.7)	DF0006923
L1-homo	*Homo sapiens*	I_Retrotransposition	LINE	L1-group	L1	NCBI	AAC51279.1
Jockey-dros	*Drosophila melanogaster*	I_Retrotransposition	LINE	R1-group	Jockey	NCBI	AAA28675.1
R2-dros	*Drosophila melanogaster*	I_Retrotransposition	LINE	R2-like	R2	NCBI	CAA36225.1
RTE1-cele	*Caenorhabditis elegans*	I_Retrotransposition	LINE	RTE-group	RTE	NCBI	AAC72298.1
CR1-gallus	*Gallus gallus*	I_Retrotransposition	LINE	CR1-group	CR1	NCBI	AAC60281.1

ii) PLE:

Abbraviation	Organism	Class	Order	Type	Subtype	Database	Accession
Neptune_Ol	*Oryzias latipes*	I_Retrotransposition	PLE	Neptune	-	NCBI	BAAF02119984.1
Poseidon_Spur	*Strongylocentrotus purpuratus*	I_Retrotransposition	PLE	Poseidon	-	NCBI	AAGJ02033054.1
Perere10_Sm	*Schistosoma mansoni*	I_Retrotransposition	PLE	Poseidon	-	NCBI	BN000801.1
Coprina_Cop-cin	Coprinopsis cinerea	I_Retrotransposition	PLE	Coprina	-	NCBI	AACS01000397.1
Athena_Av	*Adineta vaga*	I_Retrotransposition	PLE	Athena	-	NCBI	EF485018.1
Penelope_Dv	*Drosophila virilis*	I_Retrotransposition	PLE	Poseidon	-	NCBI	AAA92124.2
Cercyon_Sm	*Schistosoma mansoni*	I_Retrotransposition	PLE	Poseidon	-	NCBI	BK000685
Naiad_Hcon	*Haemonchus contortus*	I_Retrotransposition	PLE	Naiad	-	Dfam (release 3.7)	DF0289931
Osiris_Ae.B_Aaegyp	*Aedes aegypti*	I_Retrotransposition	PLE	Poseidon	-	NCBI	AAGE02017473
Nematis_Cb	*Caenorhabditis briggsae*	I_Retrotransposition	PLE	Nematis	-	NCBI	CAAC01000421

iii) LTR:

Abbraviation	Organism	Class	Order	Type	Subtype	Database	Accession
CSRN1_Cs	*Clonorchis sinensis*	I_Retrotransposition	LTR	Ty3-gypsy	-	NCBI	AAK07486.1
TOM_Dana	*Drosophila ananassae*	I_Retrotransposition	LTR	Ty3-gypsy	-	NCBI	CAA80824.1
SUSHI_Tf	*Takifugu flavidus*	I_Retrotransposition	LTR	Ty3-gypsy	-	NCBI	TWW62587.1
TY3_Sc	*Saccharomyces cerevisiae*	I_Retrotransposition	LTR	Ty3-gypsy	-	UniProtKB	Q7LHG5
YOYO_Cer-cap	*Ceratitis capitata*	I_Retrotransposition	LTR	Ty3-gypsy	-	UniProtKB	Q17318
MAG_Bmo	*Bombyx mori*	I_Retrotransposition	LTR	Ty3-gypsy	-	NCBI	X17219.1_1 [1045 - 4365]
ULYSES_Dv	*Drosophila virilis*	I_Retrotransposition	LTR	Ty3-gypsy	-	NCBI	CAA39967.1
Pao_Bma	*Bombyx mandarina*	I_Retrotransposition	LTR	Bel-Pao	-	NCBI	XP_028041443.1
BEL_Dm	*Drosophila melanogaster*	I_Retrotransposition	LTR	Bel-Pao	-	NCBI	AAB03640.1

iv) DNA:

Abbraviation	Organism	Class	Order	Type	Subtype	Database	Accession
Minos_Dh	*Drosophila hydei*	II_DNA_Transposition	Transposase	Tc1-Mariner	Tc1	NCBI	X61695.1
Tc1 _Ce	*Caenorhabditis elegans*	II_DNA_Transposition	Transposase	Tc1-Mariner	Tc1	NCBI	X01005.1
Tc3_Ce	*Caenorhabditis elegans*	II_DNA_Transposition	Transposase	Tc1-Mariner	Tc1	Dfam (release 3.7)	DF0004136
Mos1_Dmauri	*Drosophila mauritana*	II_DNA_Transposition	Transposase	Tc1-Mariner	Mariner	NCBI	M14653.1
Pogo_Dm	*Drosophila melanogaster*	II_DNA_Transposition	Transposase	Tc1-Mariner	Tc2-group	Dfam (release 3.7)	DF0001682
Tigger_Tt	*Trichuris trichiura*	II_DNA_Transposition	Transposase	Tc1-Mariner	Tc2-group	NCBI	CBXK010001312.1
Mariner-3_Dr	*Danio rerio*	II_DNA_Transposition	Transposase	Tc1-Mariner	ISRm11	Dfam (release 3.7)	DF0002344
DNA-TTAA0-4_Dr	*Danio rerio*	II_DNA_Transposition	Transposase	Tc1-Mariner	ISRm11	Dfam (release 3.7)	DF0003301
ISRm11_Hb	*Heligmosomoides bakeri*	II_DNA_Transposition	Transposase	Tc1-Mariner	ISRm11	DFAM (release 3.7)	DF003447838
TcMar-ISRm11-2_AstCal	*Astatotilapia calliptera*	II_DNA_Transposition	Transposase	Tc1-Mariner	ISRm11	DFAM (release 3.7)	DF003571753
Merlin1_HS	*Homo sapiens*	II_DNA_Transposition	Transposase	Merlin	-	DFAM (release 3.7)	DF000001000
Looper_eutheria	Eutheria	II_DNA_Transposition	Transposase	PiggyBac	-	Dfam (release 3.7)	DF000000369

Capítulo 6.

Bibliografía

Alba, A., Tetreau, G., Chaparro, C., Sánchez, J., Vázquez, A. A., & Gourbal, B. (2019). Natural resistance to Fasciola hepatica (Trematoda) in Pseudosuccinea columella snails: A review from literature and insights from comparative "omic" analyses. *Developmental & Comparative Immunology*, *101*, 103463. https://doi.org/10.1016/j.dci.2019.103463

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Caravedo, M. A., & Cabada, M. (2020). Human Fascioliasis: Current Epidemiological Status and Strategies for Diagnosis, Treatment, and Control. *Research and Reports in Tropical Medicine*, *Volume 11*, 149–158. https://doi.org/10.2147/RRTM.S237461

Carmona, C., & Tort, J. F. (2017). Fasciolosis in South America: Epidemiology and control challenges. *Journal of Helminthology*, *91*(2), 99–109. https://doi.org/10.1017/S0022149X16000560

Cerbin, S., & Jiang, N. (2018). Duplication of host genes by transposable elements. *Current Opinion in Genetics & Development*, *49*, 63–69. https://doi.org/10.1016/j.gde.2018.03.005

Chang, N.-C., Rovira, Q., Wells, J., Feschotte, C., & Vaquerizas, J. M. (2022). Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Research*, *32*(7), 1408–1423. https://doi.org/10.1101/gr.275655.121

Chénais, B., Caruso, A., Hiard, S., & Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, *509*(1), 7–15. https://doi.org/10.1016/j.gene.2012.07.042

Choi, Y.-J., Fontenla, S., Fischer, P. U., Le, T. H., Costábile, A., Blair, D., Brindley, P. J., Tort, J. F., Cabada, M. M., & Mitreva, M. (2020). Adaptive Radiation of the Flukes of the Family Fasciolidae Inferred from Genome-Wide Comparisons of Key Species. *Molecular Biology and Evolution*, *37*(1), 84–99. https://doi.org/10.1093/molbev/msz204

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, *18*(2), 71–86. https://doi.org/10.1038/nrg.2016.139

CNCB-NGDC Members and Partners, Bai, X., Bao, Y., Bei, S., Bu, C., Cao, R., Cao, Y., Cen, H., Chao, J., Chen, F., Chen, H., Chen, K., Chen, M., Chen, M., Chen, M., Chen, Q., Chen, R., Chen, S., Chen, T., … Zuo, Z. (2024). Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Research*, *52*(D1), D18–D32. https://doi.org/10.1093/nar/gkad1078

Cosby, R. L., Chang, N.-C., & Feschotte, C. (2019). Host–transposon interactions: Conflict, cooperation, and cooption. *Genes & Development*, *33*(17–18), 1098–1116. https://doi.org/10.1101/gad.327312.119

Cruz-Mendoza, I., Naranjo-García, E., Quintero-Martínez, M. T., Ibarra-Velarde, F., & Correa, D. (2006). Exposure to Fasciola hepatica Miracidia Increases the Sensitivity of Lymnaea (Fossaria) humilis to High and Low pH. *Journal of Parasitology*, *92*(3), 650–652. https://doi.org/10.1645/GE-3542RN.1

Cwiklinski, K., Dalton, J. P., Dufresne, P. J., La Course, J., Williams, D. J., Hodgkinson, J., & Paterson, S. (2015). The Fasciola hepatica genome: Gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution. *Genome Biology*, *16*(1), 71. https://doi.org/10.1186/s13059-015-0632-2

Cwiklinski, K., Robinson, M. W., Donnelly, S., & Dalton, J. P. (2021). Complementary transcriptomic and proteomic analyses reveal the cellular and molecular processes that drive growth and development of Fasciola hepatica in the host liver. *BMC Genomics*, *22*(1), 46. https://doi.org/10.1186/s12864-020-07326-y

Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE*, *6*(1), e16526. https://doi.org/10.1371/journal.pone.0016526

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2019). *RepeatModeler2: Automated genomic discovery of transposable element families* [Preprint]. Genomics. https://doi.org/10.1101/856591

Fontenla, S., Rinaldi, G., Smircich, P., & Tort, J. F. (2017). Conservation and diversification of small RNA pathways within flatworms. *BMC Evolutionary Biology*, *17*(1), 215. https://doi.org/10.1186/s12862-017-1061-5

González-Miguel, J., Becerro-Recio, D., & Siles-Lucas, M. (2021). Insights into Fasciola hepatica Juveniles: Crossing the Fasciolosis Rubicon. *Trends in Parasitology*, *37*(1), 35–47. https://doi.org/10.1016/j.pt.2020.09.007

Goubert, C., Craig, R. J., Bilat, A. F., Peona, V., Vogan, A. A., & Protasio, A. V. (2022). A beginner's guide to manual curation of transposable elements. *Mobile DNA*, *13*(1), 7. https://doi.org/10.1186/s13100-021-00259-7

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

International Helminth Genomes Consortium. (2019). Comparative genomics of the major parasitic worms. *Nature Genetics*, *51*(1), 163–174. https://doi.org/10.1038/s41588-018-0262-1

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

Krumsiek, J., Arnold, R., & Rattei, T. (2007). Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, *23*(8), 1026–1028. https://doi.org/10.1093/bioinformatics/btm039

Lalor, R., Cwiklinski, K., Calvani, N. E. D., Dorey, A., Hamon, S., Corrales, J. L., Dalton, J. P., & De Marco Verissimo, C. (2021). Pathogenicity and virulence of the liver flukes *Fasciola hepatica* and *Fasciola Gigantica* that cause the zoonosis Fasciolosis. *Virulence*, *12*(1), 2839–2867. https://doi.org/10.1080/21505594.2021.1996520

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lotfy, W. M., Brant, S. V., DeJong, R. J., Le, T. H., Demiaszkiewicz, A., Rajapakse, R. P. V. J., Perera, V. B. V. P., Laursen, J. R., & Loker, E. S. (2008). Evolutionary Origins, Diversification, and Biogeography of Liver Flukes (Digenea, Fasciolidae). *The American Journal of Tropical Medicine and Hygiene*, *79*(2), 248–255. https://doi.org/10.4269/ajtmh.2008.79.248

Luo, X., Cui, K., Wang, Z., Li, Z., Wu, Z., Huang, W., Zhu, X.-Q., Ruan, J., Zhang, W., & Liu, Q. (2021). High-quality reference genome of Fasciola gigantica: Insights into the genomic signatures of transposon-mediated evolution and specific parasitic adaption in tropical regions. *PLoS Neglected Tropical Diseases*, *15*(10), e0009750. https://doi.org/10.1371/journal.pntd.0009750

Mas-Coma, S., Buchon, P., Funatsu, I. R., Angles, R., Artigas, P., Valero, M. A., & Bargues, M. D. (2020). Sheep and Cattle Reservoirs in the Highest Human Fascioliasis Hyperendemic Area: Experimental Transmission Capacity, Field Epidemiology, and Control Within a One Health Initiative in Bolivia. *Frontiers in Veterinary Science*, *7*, 583204. https://doi.org/10.3389/fvets.2020.583204

McClintock, B. (1984). The Significance of Responses of the Genome to Challenge. *Science, New Series*, *226*(4676), 792–801.

McNulty, S. N., Tort, J. F., Rinaldi, G., Fischer, K., Rosa, B. A., Smircich, P., Fontenla, S., Choi, Y.-J., Tyagi, R., Hallsworth-Pepin, K., Mann, V. H., Kammili, L., Latham, P. S., Dell'Oca, N., Dominguez, F., Carmona, C., Fischer, P. U., Brindley, P. J., & Mitreva, M. (2017). Genomes of Fasciola hepatica from the Americas Reveal Colonization with Neorickettsia Endobacteria Related to the Agents of Potomac Horse and Human Sennetsu Fevers. *PLOS Genetics*, *13*(1), e1006537. https://doi.org/10.1371/journal.pgen.1006537

Mehmood, K., Zhang, H., Sabir, A. J., Abbas, R. Z., Ijaz, M., Durrani, A. Z., Saleem, M. H., Ur Rehman, M., Iqbal, M. K., Wang, Y., Ahmad, H. I., Abbas, T., Hussain, R., Ghori, M. T., Ali, S., Khan, A. U., & Li, J. (2017). A review on epidemiology, global prevalence and economical losses of fasciolosis in ruminants. *Microbial Pathogenesis*, *109*, 253–262. https://doi.org/10.1016/j.micpath.2017.06.006

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Moazeni, M., & Ahmadi, A. (2016). Controversial aspects of the life cycle of Fasciola hepatica. *Experimental Parasitology*, *169*, 81–89. https://doi.org/10.1016/j.exppara.2016.07.010

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, *20*(1), 275. https://doi.org/10.1186/s13059-019-1905-y

Pandey, T., Ghosh, A., Todur, V. N., Rajendran, V., Kalita, P., Kalita, J., Shukla, R., Chetri, P. B., Shukla, H., Sonkar, A., Lyngdoh, D. L., Singh, R., Khan, H., Nongkhlaw, J., Das, K. C., & Tripathi, T. (2020). Draft Genome of the Liver Fluke *Fasciola gigantica*. *ACS Omega*, *5*(19), 11084–11091. https://doi.org/10.1021/acsomega.0c00980

Philippsen, G. S., & DeMarco, R. (2019). Impact of transposable elements in the architecture of genes of the human parasite Schistosoma mansoni. *Molecular and Biochemical Parasitology*, *228*, 27–31. https://doi.org/10.1016/j.molbiopara.2018.12.007

Platt, R. N., Blanco-Berdugo, L., & Ray, D. A. (2016). Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biology and Evolution*, *8*(2), 403–410. https://doi.org/10.1093/gbe/evw009

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rice, P., Longden, I., & 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

Sanchís, J., Hillyer, G., Madeira De Carvalho, L., Macchi, M., Gomes, C., Maldini, G., Stilwell, G., Venzal, J., Paz-Silva, A., Sánchez-Andrade, R., & Arias, M. (2015). Riesgo de exposición a Fasciola hepática en ganado vacuno en extensivo de Uruguay y Portugal determinado mediante ELISA y un antígeno recombinante. *Archivos de medicina veterinaria*, *47*(2), 201–208. https://doi.org/10.4067/S0301-732X2015000200011

Senft, A. D., & Macfarlan, T. S. (2021). Transposable elements shape the evolution of mammalian development. *Nature Reviews Genetics*, *22*(11), 691–711. https://doi.org/10.1038/s41576-021-00385-1

Siles-Lucas, M., Becerro-Recio, D., Serrat, J., & González-Miguel, J. (2021). Fascioliasis and fasciolopsiasis: Current knowledge and future trends. *Research in Veterinary Science*, *134*, 27–35. https://doi.org/10.1016/j.rvsc.2020.10.011

Skinner, D. E., Rinaldi, G., Koziol, U., Brehm, K., & Brindley, P. J. (2014). How might flukes and tapeworms maintain genome integrity without a canonical piRNA pathway? *Trends in Parasitology*, *30*(3), 123–129. https://doi.org/10.1016/j.pt.2014.01.001

Storer, J., Hubley, R., Rosen, J., & Smit, A. (2022). Methodologies for the De novo Discovery of Transposable Element Families. *Genes*, *13*(4), 709. https://doi.org/10.3390/genes13040709

Storer, J. M., Hubley, R., Rosen, J., & Smit, A. F. A. (2021). Curation Guidelines for *de novo* Generated Transposable Element Families. *Current Protocols*, *1*(6), e154. https://doi.org/10.1002/cpz1.154

Subramanian, B., Gao, S., Lercher, M. J., Hu, S., & Chen, W.-H. (2019). Evolview v3: A webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Research*, *47*(W1), W270–W275. https://doi.org/10.1093/nar/gkz357

Suh, A. (2016). Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. *Nature Communications*, *48*(7), 829–834. https://doi.org/10.1038/pj.2016.37

Torgerson, P. R., Devleesschauwer, B., Praet, N., Speybroeck, N., Willingham, A. L., Kasuga, F., Rokni, M. B., Zhou, X.-N., Fèvre, E. M., Sripa, B., Gargouri, N., Fürst, T., Budke, C. M., Carabin, H., Kirk, M. D., Angulo, F. J., Havelaar, A., & De Silva, N. (2015). World Health Organization Estimates of the Global and Regional Disease Burden of 11 Foodborne Parasitic Diseases, 2010: A Data Synthesis. *PLOS Medicine*, *12*(12), e1001920. https://doi.org/10.1371/journal.pmed.1001920

Tumescheit, C., Firth, A. E., & Brown, K. (2022). CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ*, *10*, e12983. https://doi.org/10.7717/peerj.12983

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, *54*(1), 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973–982. https://doi.org/10.1038/nrg2165