Tesina de grado

Facultad de Ciencias - Universidad de la República, Montevideo - Uruguay

Licenciatura en Biología

Análisis *in silico* de los retrotransposones del linaje Athila/Tat en *Acca sellowiana*

Mathias Joaquín Mangino
Orientadora: Dra. Luisa Berná
Coorientadora: Dra. Clara Pritsch

Tribunal:
Dra. Clara Pritsch
Dra. Magdalena Vaio
Dr. Andres Parada

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FACULTAD DE CIENCIAS
UDELAR | fcien.edu.uy

FACULTAD DE AGRONOMÍA
UNIVERSIDAD DE LA REPÚBLICA

AGRADECIMIENTOS

Agradezco a mi familia por su confianza, paciencia y apoyo a lo largo de este proceso.

Quiero expresar mi gratitud a Luisa por compartir su experiencia y brindarme valiosas instancias de aprendizaje. Tu tiempo y conocimientos han sido esenciales en este camino de formación, tanto a nivel profesional como personal.

A Clara, le agradezco por dedicar su valioso tiempo y ofrecer sugerencias que han sido fundamentales para este trabajo.

Agradezco a Magdalena y Andrés por sus correcciones y enriquecedores comentarios, que han contribuido al desarrollo de este trabajo.

A Sofi, le agradezco por su constante compañía y apoyo en cada uno de mis proyectos, brindándome ánimo y motivación en todo momento.

Sus contribuciones han dejado una huella en este trabajo y en el desarrollo académico y personal.

CONTENIDO

AGRADECIMIENTOS **2**
CONTENIDO **3**
RESUMEN **4**
INTRODUCCIÓN **5**
Clasificación de los elementos transponibles 7
Importancia evolutiva según integración y organización 10
Retroelementos LTR 11
Estructura de retroelementos LTR 11
Mecanismo de retrotransposición LTR 13
Retroelementos LTR en plantas 14
Inserción de retroelementos LTR en plantas 15
Retroelementos LTR en angiospermas/Myrtaceae 16
Retroelemento Athila/Tat 17
Acca sellowiana *18*
Genoma del Guayabo 18
OBJETIVOS **20**
Objetivo general 20
Objetivos específicos 20
MATERIALES Y MÉTODOS **21**
Datos 21
Formatos de archivos 21
Software 21
Flujo de trabajo 26
RESULTADOS **30**
Selección de base de datos y estadísticos descriptivos 30
Candidatos de RT LTR Athila/Tat 33
Visualización de RT LTR 36
Caracterización de elementos candidatos de RT LTR Athila/Tat 36
Búsqueda recursiva y clasificación de RT LTR Ty3/Gypsy 40
Otro flujo de trabajo para abordar el problema 46
DISCUSIÓN **49**
Flujo de trabajo 49
Selección de base de datos 50
Recopilación de datos preliminares para el flujo de trabajo 50
Dominio RT 51
Búsqueda recursiva de RT LTR 52
Clasificación de Ty3/Gypsy basado en filogenias 52
Otro flujo de trabajo para abordar el problema 54
CONCLUSIONES **56**
REFERENCIA BIBLIOGRÁFICA **57**
MATERIAL SUPLEMENTARIO **66**

RESUMEN

Los retrotransposones con repeticiones terminales largas (RT-LTR) constituyen un grupo diverso de elementos transponibles que se replican a través de un intermediario de ARN, generando copias que se integran en el genoma del huésped. Estos elementos son predominantes en genomas eucariotas, siendo especialmente abundantes en las plantas. Asimismo, comparten similitudes estructurales y mecanismos de replicación con los retrovirus. Dentro de la familia de RT-LTR Ty3/Gypsy, el linaje Athila/Tat está ampliamente distribuido en genomas de plantas terrestres. *Acca Sellowiana*, también conocida como Guayabo del país, es una planta nativa perteneciente a la familia *Myrtaceae*. Esta especie es diploide, con un número cromosómico de 2n=22 y un genoma de aproximadamente 345 Mpb/1C. Recientemente, se secuenció su genoma en nuestro país. Se desarrolló un protocolo de extracción de RT-LTR y una caracterización bioinformática del linaje Athila/Tat en *A. sellowiana*. Se trabajó con secuencias de genes de retrotranscriptasas Athila/Tat de las bases de datos Gypsy Database 2.0 y RepeatExplorer database. Para la búsqueda en el genoma de *A. sellowiana* se utilizó BLAST, Mafft, trimAl, Iqtree, RepeatMasker, REannotate y programas en Perl y R. Esto permitió establecer un protocolo de búsqueda de datos que facilita la identificación y el análisis supervisado de los RT-LTR en un genoma de interés basado en la homología y las características estructurales. Se identificaron 124 elementos Athila y Ogre en el genoma, siendo Ogre el linaje predominante. Dada la cantidad y relación filogenética de Ogre, se infiere que está activo en el genoma de *Acca sellowiana*. Finalmente, nuestra estrategia de búsqueda por homología de dominios con inferencia filogenética aumentó la sensibilidad, permitiendo la detección de secuencias divergentes sin perder especificidad. En contraste, RepeatMasker y REannotate no pudieron lograrlo eficazmente.

INTRODUCCIÓN

A partir de que la citogenetista Barbara McClintock hizo el descubrimiento de los transposones mediante sus investigaciones sobre quiebres de cromosomas en el maíz (McClintock, 1948), los elementos transponibles (ET), secuencias de ADN con capacidad de cambiar su posición en el genoma, han sido objeto de interés e investigación para entender su funcionalidad y rol en la evolución de los genomas. Barbara McClintock nombra “elementos control” a los ET, debido a su capacidad de alterar la expresión de genes cercanos al sitio de inserción de los ET (McClintock, 1956). Sin embargo, en su tiempo hubo gran escepticismo sobre estos elementos denominados saltarines. Sus hallazgos fueron aceptados y ampliados tiempo después y gracias a esto Barbara McClintock pasó a ser la primera mujer en recibir un Premio Nobel de Medicina en 1983.

Durante las últimas décadas los avances en las tecnologías de secuenciación han dado lugar a la posibilidad de secuenciar el genoma completo de diferentes organismos (Hu et al., 2021). Gracias a la secuenciación y disponibilidad de nuevos genomas se ha obtenido mayor información sobre los ET y se ha recobrado un gran interés por investigar la diversidad y evolución de estos elementos que se mueven a través del genoma (Bennetzen & Wang, 2014). Los ET forman parte del repitoma (repertorio de secuencias repetidas) de los genomas también conocido como ADN “basura” o parásito” y en general corresponden con la mayor parte del repitoma. Además de los ET, el repitoma comprende virus endógenos, secuencias repetidas en tándem (como ADNr y ADN satélite) y ADN “no anotado” también llamado “materia oscura” (Maumus & Quesneville, 2016).

Elementos Transponibles

Los elementos transponibles son secuencias de ADN móvil capaces de replicarse y cambiar de posición en el genoma. Hay dos clases principales de ET, los retrotransposones y los transposones de ADN. Los retrotransposones, son secuencias de ADN capaces de “copiar y pegar” su secuencia en otro sitio del genoma. Esto lo hacen a través de la transcripción a un intermediario de ARN, que luego por una transcriptasa reversa pasa a ADN y se establece en otro sitio del genoma. Los transposones de ADN poseen varios mecanismos de transposición; el principal de estos es de “cortar y pegar”, el transposón se escinde de su sitio y se establece en otro sitio del genoma (Bourque et al., 2018).

Los ET son capaces de replicarse y cambiar de posición en el genoma en diferentes circunstancias: en la línea germinal, aumentando su frecuencia a través de herencia

vertical y a través de transferencia horizontal (sin involucrar gametos). La transferencia horizontal entre especies ha sido un factor importante en la distribución y expansión de los ET (Wells & Feschotte, 2020). El cambio de posición se define como transposición y puede generar diferentes variaciones en el genoma, entre las que se destacan los cambios en el tamaño genómico y el aumento de mutaciones puntuales como rearreglos cromosómicos (Bourque et al., 2018).

Los ET son una de las mayores fuentes de variabilidad de los genomas impulsores de la evolución del genoma (Lisch, 2013). Los ET no se distribuyen de manera azarosa en el genoma, sino que poseen ciertas regiones preferentes de inserción. Esta preferencia es regida por la presión selectiva en distintas regiones del genoma (Bourque et al., 2018).

Los ET autónomos son elementos que codifican las enzimas requeridas para su movilización genómica, y los ET no autónomos no codifican sus propias enzimas y su transposición depende de la maquinaria de elementos autónomos (Wicker et al., 2007).

Abundancia de ET y tamaño de genoma

Los genomas de Bacteria, Archaea y Eukarya albergan gran cantidad, y diversidad de elementos transponibles (ET) (Filée et al., 2007). Entre estos, los genomas eucariotas son los que hospedan la mayor diversidad de ET. Los ET se encuentran en todo los eucariotas, con algunas excepciones, por ejemplo los protistas del phylum *apicomplexa* como *Plasmodium falciparum*, causante de la malaria, *Toxoplasma gondii*, *Encephalitozoon intestinalis* y *Theileria parva* entre otros, al parecer eliminaron por completo los ET de sus genomas (DeBarry & Kissinger, 2011). Existe una relación entre el tamaño del repitoma, en especial dado por la abundancia de ET y el tamaño del genoma para algunas especies (Elliott & Gregory, 2015). En particular en algunas plantas y otros eucariotas pueden ocupar la mayoría del genoma, llegando a comprender el 85% del tamaño total del genoma, observado en especies con genomas grandes (Fig. 1). Cabe resaltar que los eucariotas que presentan los genomas más reducidos conocidos al momento, son justamente del phylum apicomplexa que como se mencionó no contienen ET (Corradi et al., 2010).

A pesar de la gran diversidad de los ET en los genomas, el número de proteínas involucradas en la replicación y transposición de los mismos es reducido, compartiendo una estructura altamente conservada. En particular existe una estructura denominada motivo de reconocimiento de ARN (MRR), la cual tiene un rol preponderante en la transposición de los ET. Asimismo, la distribución filogenética general de estos dominios proteicos indica que los ET son anteriores a la aparición de los eucariotas (Wells & Feschotte, 2020).

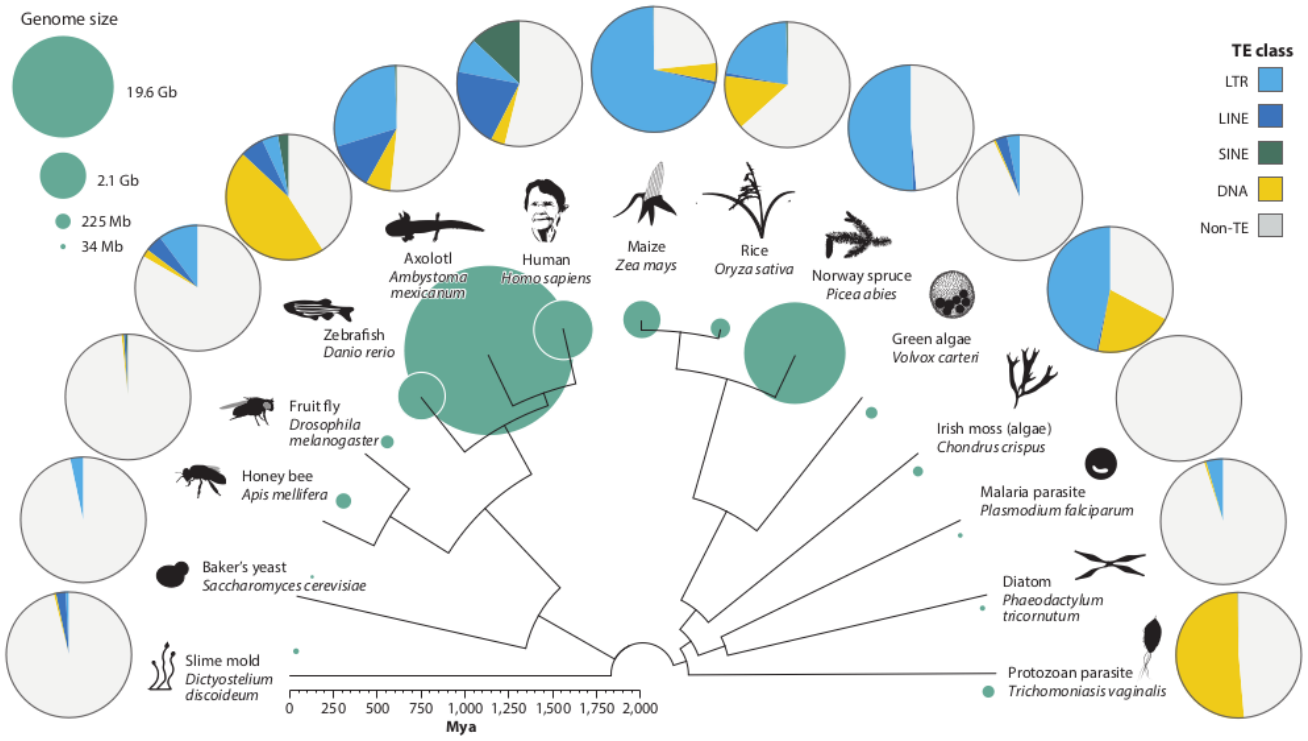

Figura 1. El tamaño de los genomas de referencia se identifica con círculos verdes, asimismo el tamaño varía drásticamente entre los eucariotas, existiendo una correlación con la cantidad de elementos transponibles. Para facilitar la visualización, los elementos DIRS se han incluido con LTR y todos los elementos de Clase II incluidos en "ADN" (Wells & Feschotte, 2020).

Clasificación de los elementos transponibles

Clasificación

La principal división o clasificación de ET la introdujo David Finnegan en 1989 (Finnegan, 1989). Este autor, separa en dos clases principales a los ET basándose en el tipo de molécula intermediaria en la transposición. La Clase I corresponde a los retrotransposones, tienen intermediario de ARN y Clase II los define como transposones de ADN, que carecen de intermediario de ARN. En 2007, Wicker y colaboradores crean una nueva clasificación jerárquica a partir de la clasificación inicial de Finnegan (Wicker et al., 2007). En esta se propone clase, subclase, orden, superfamilia, familia y subfamilia. La subclase separa según el mecanismo que utilizan los ET para transponerse, por un lado están los elementos que se copian a sí mismos para insertarse (copy-and-paste), y por otro, los elementos que salen de la región donante y se insertan en otro lugar (cut-and-paste). Por su parte, el orden en la clasificación divide a los ET según la estrategia de integración al genoma huésped. La superfamilia dentro de un orden comparte estrategia de replicación, sin embargo se diferencian según las estructuras proteicas o dominios no codificantes que utilizan. Además, las superfamilias se distinguen por la presencia y tamaño de

las duplicaciones del sitio target (TSD). Los TSD son secuencias repetidas directas que se generan en el margen de los ET una vez que se insertan. A su vez la superfamilia se divide en familias, estas se distinguen por compartir una alta similitud o identidad de la secuencia de ADN. La conservación en la secuencia se debe a las regiones codificantes, el resto de secuencia tiene una conservación mínima. En este sentido, Wicker y colaboradores (2007) nos ofrecen una forma pragmática de definir que dos elementos son de la misma familia. Los elementos deben compartir 80% o más de identidad de secuencia en al menos un 80% o más de su región codificante, o dentro de la región de repetición o en ambas regiones. Además, la secuencia debe tener una longitud mayor a 80 pb, estos requisitos definen la regla 80-80-80. El término subfamilia se aplica para casos de poblaciones de elementos no autónomos derivados de una familia (Wicker et al., 2007). En este trabajo la clasificación de familias se menciona como linajes.

Los ET de clase II se subdividen en dos subclases. La subclase I está comprendida por ET clásicos de "cut-and-paste" y por dos órdenes: "terminal inverted repeat" (TIR) y Crypton. Por su parte, la subclase II son subyugados a un proceso de transposición que implica la replicación sin escisión bicatenaria, comúnmente conocida como "copy-and-paste". Esta última subclase está comprendida por los órdenes Helitron y Maverick (Bourque et al., 2018).

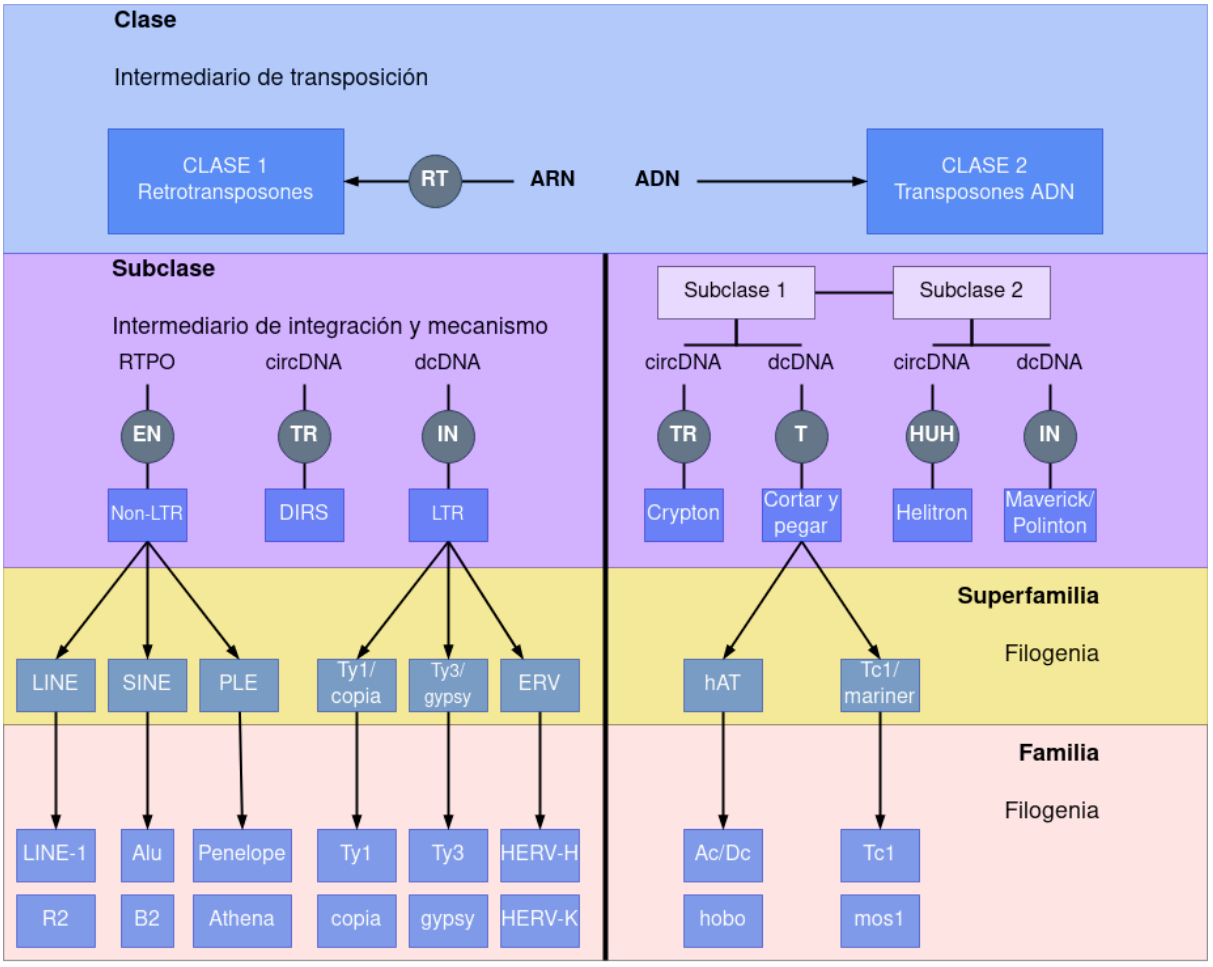

Figura 2. Clasificación de elementos transponibles (ET) en eucariotas. La clasificación es jerárquica y divide los ET en dos principales clases basándose en la presencia o ausencia de un intermediario de ARN en su proceso de transposición. Esquema que muestra las características principales y las relaciones entre clases, subclases, superfamilias y familias. Los círculos en color gris representan enzimas codificadas por los ET. Abreviaturas y codificación: EN = endonucleasa, IN = integrasa, HUH = proteína Rep/Helicasa con actividad endonucleasa HUH, T = transposasa, TR = transcriptasa inversa, RTPO = transcripción inversa primed por objetivo, circDNA = Intermediario circular de ADN, dcADN = intermedio de doble cadena lineal de ADN, LTR = repeticiones terminales largas, Non-LTR = Sin repeticiones terminales largas, DIRS = secuencia repetitiva Dictyostelium, LINE / SINE = elementos nucleares intercalados Largos y Cortos, PLE = elementos tipo Penelope y ERV = retrovirus endógenos (Bourque et al., 2018).

Los ET de clase I, o retrotransposones, no requieren de una división de subclase, ninguno de los integrantes son capaces de escindir o transferir su hebra de ADN desde un sitio donante. Sin embargo, el intermediario de ARN es transcrito a partir de una copia genómica, luego mediante una transcriptasa inversa es retrotranscripto a ADN. Debido a estas características los ET de clase I producen una nueva copia en cada ciclo de replicación. En consecuencia, los retrotransposones son los principales contribuyentes a la fracción repetitiva en los genomas. Estos ET de clase I se pueden dividir en 5 órdenes “Long terminal repeat” (LTR), “Dictyostelium intermediate repeat sequence” (DIRS), “Penelope-like-elements” (PLE), “long/short interspersed nuclear element” (LINE y SINE).Los últimos también conocidos como retrotransposones no-LTR.

Biología general de los elementos

La transposición de los ET y su posterior integración dentro del genoma tienen consecuencias de gran impacto en la arquitectura del genoma huésped, tanto a nivel individual como de especie. Los ET pueden generar una gran diversidad de cambios que van desde cambios en la expresión génica debido a la activación de genes que se encuentran cercanos al lugar de inserción, a la inactivación de genes si se insertan entre genes funcionales, al alterar los patrones de hiper/hipometilación de las región del genoma en la cual se insertan. Los ET también pueden promover reordenamientos cromosómicos como inversiones y translocaciones. También pueden provocar la duplicación génica, participar en la generación de genes de *novo*, y modificar el paisaje epigenético (Bennetzen & Wang, 2014). Los ET además tienen un rol en la evolución de las poblaciones al actuar como mecanismos de aislamiento reproductivo y adaptación (Capy, 2005). Sin embargo, la mayoría de las copias de un ET no tienen un rol tan protagónico y con el tiempo desde el evento de inserción, acumulan errores de manera independiente por deriva genética. Este decaimiento es debido al efecto acumulativo de mutaciones puntuales, inserciones anidadas, deleciones, e inserciones y deleciones cortas (indels) (Maumus & Quesneville, 2016). Consecuentemente, los ET que permanecen se vuelven con el tiempo, cada vez más fragmentados y mutados, al punto que finalmente forman parte de las secuencias conocidas como secuencias basura o en inglés ‘junk

sequences'. Estas secuencias son un mosaico de secuencias alteradas de ET que quedan en el genoma sin capacidad de transponerse de manera autónoma.

A partir de los rastros de ET se puede identificar los polimorfismos acumulados en cada elemento y estimar tiempos de inserción (proliferación o amplificación) en base a la divergencia de las secuencias. Para ello, se construye una biblioteca de secuencias consenso generada por las múltiples copias de cada repetido en el genoma, la cual se asume como ancestral, teniendo en cuenta que las mutaciones ocurren aleatoriamente entre los ET. La comparación de cada elemento con la biblioteca de secuencias consenso permite estimar el tiempo de inserción incluso para el caso de inserciones muy antiguas. De una manera más acotada, el tiempo de inserción de los elementos retrotransposones LTR completos, puede estimarse, por comparación de identidad de secuencias, como los repetidos LTR, En este caso, la estimación queda acotada a eventos recientes ya que las secuencias LTR deben estar intactas. Con esta estrategia se puede inferir la "edad" de ET particulares; esto es, estimar el tiempo transcurrido desde la inserción transposicional (Maumus & Quesneville, 2016).

Importancia evolutiva según integración y organización

Los ET no se distribuyen de manera azarosa en el genoma, sino que poseen ciertas regiones preferentes de inserción. Esta preferencia está regida por la presión selectiva en distintas regiones del genoma, reteniendo o expulsando a los ET (Bourque et al., 2018). La inserción de los ET es un evento fundamental para la replicación y transposición de los mismos. Asimismo, se evidencian tres patrones de inserción en los genomas. i) Los ET que poseen baja preferencia de sitios de inserción; ii) los ET que poseen preferencia por regiones genómicas donde minimizan los efectos nocivos; y iii) los ET dirigidos a ciertas regiones que faciliten su posterior propagación (Wells & Feschotte, 2020). El mecanismo de inserción de los ET está dirigido por las nucleasas que catalizan la integración cromosómica. Dado que las nucleasas, que son codificadas por los ET, difieren entre sí, también va diferir el grado de especificidad a sus sustratos.

En el nivel más bajo de preferencia por un sustrato están los ET que reconocen motivos de secuencias cortas o muy degeneradas; este patrón de inserción se asemeja a una distribución aleatoria (Wells & Feschotte, 2020). Por otro lado, algunos ET muestran cierto grado de preferencia por regiones cercanas a los genes, dirigiéndose a regiones 5' corriente arriba de estos genes. Esto implica un beneficio evolutivo para los ET y tal vez para el huésped. El beneficio para los ET yace en colocarse en un entorno donde la cromatina está promoviendo la expresión, asegurando su replicación y transposición. Muchos transposones de clase II han adoptado esta estrategia (Wells & Feschotte, 2020). Por último, muchos ET se insertan en regiones donde sea poco probable un efecto negativo sobre el control

de funciones celulares; así se distribuyen en “refugios seguros”, donde su inserción no tenga consecuencias perjudiciales inmediatas y sea más probable su supervivencia y reproducción (Wells & Feschotte, 2020).

Retroelementos LTR

Los retrotransposones LTR (RT LTR) están ampliamente distribuidos en plantas, hongos y animales. Estos se caracterizan por la presencia de repetidos terminales largos (LTR) directos que van desde 100 pb hasta más de 5 kb. Los LTR son fundamentales para la transcripción y replicación. Adicionalmente los RT LTR poseen otras regiones importantes para su replicación que incluyen: un sitio de unión de primers (PBS) y un intervalo de polipurinas (PPT), ambas cercanas a los LTR (Macas & Neumann, 2007). Los diferentes RT LTR pueden poseer hasta dos marcos abiertos de lectura u ORFs (por open reading frames) asociados a los genes *gag* y *pol* o, alternativamente ningún ORF. El gen *gag* codifica una poliproteína con una cápside y un dominio nucleocápside. Las proteínas GAG forman partículas similares a virus en el citoplasma, donde ocurre la transcripción inversa. El gen *pol* tiene dominios transcriptasa inversa (RT), ribonucleasa H (RNaseH) e integrasa (INT). Estos genes aseguran la replicación del retrotransposon, a través de la transcripción inversa, y la integración de la nueva copia en el genoma (Macas & Neumann, 2007). Eventualmente, algunos retroelementos del orden LTR poseen un ORF adicional con función desconocida (Neumann et al., 2003).

Los retroelementos LTR se agrupan en cinco familias: Ty1/Copia, Ty3/Gypsy, Bel/Pao, Retroviridae y Caulimoviruses. Las dos primeras, Gypsy *y Copia*, se consideran superfamilia. Estas superfamilias difieren en el orden en que se ubican la RT y la INT, en el gen *po*l La organización de ORF gag-pol-env en el genoma de Ty3/Gypsy y Bel/Pao es similar a la de Retroviridae, siendo estos retrovirus simples.

Estructura de retroelementos LTR

La región flanqueante LTR de los retrotransposones tienen las señales de terminación de la transcripción, el promotor y se dividen en tres áreas funcionales: U3, R y U5. U3 contiene una región potenciadora y promotora en el extremo 3' del transcrito; R contiene los sitios de inicio y terminación para la transcripción; U5 solo existe en la transcripción del terminal 5′. El proceso de transcripción va desde el límite izquierdo de la región LTR U3/R hasta el límite derecho de la región LTR R/U5 para producir una molécula de ARN con ambos extremos de región R (Fig. 3)(Havecker et al., 2004).

.

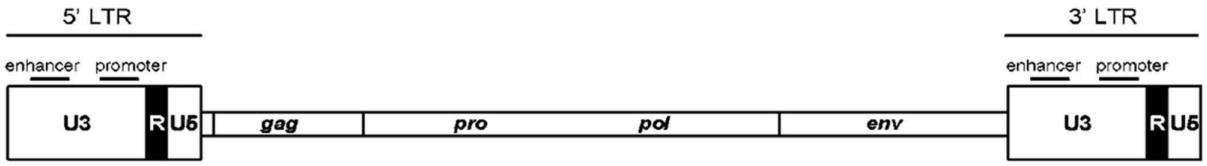

Figura 3. Organización canónica de un retroelemento LTR. Abreviatura y codificación del core: gag = proteínas virales estructurales de la cápside, pro = proteasa, pol = polimerasa, env = envoltura. Aunque el gen env está presente en algunos retrotransposones LTR (RT LTR), su detección es poco probable; pudiendo existir un ORF adicional en su posición con funciones no conocidas. La región LTR se subdivide en 3 regiones: U3, R y U5. U3 contiene las secuencias potenciadoras (enhancer) y promotoras (promoter) que impulsan la transcripción. El dominio R codifica secuencias de capping 5' (cap 5') y la señal poli A (pA) (Zhang et al., 2014).

Sin embargo, los linajes divergen considerablemente en sus secuencias de ADN y organización genómica (Fig. 4). En particular difieren en la cantidad de ORF, si presentan o no genes env-like u otros.

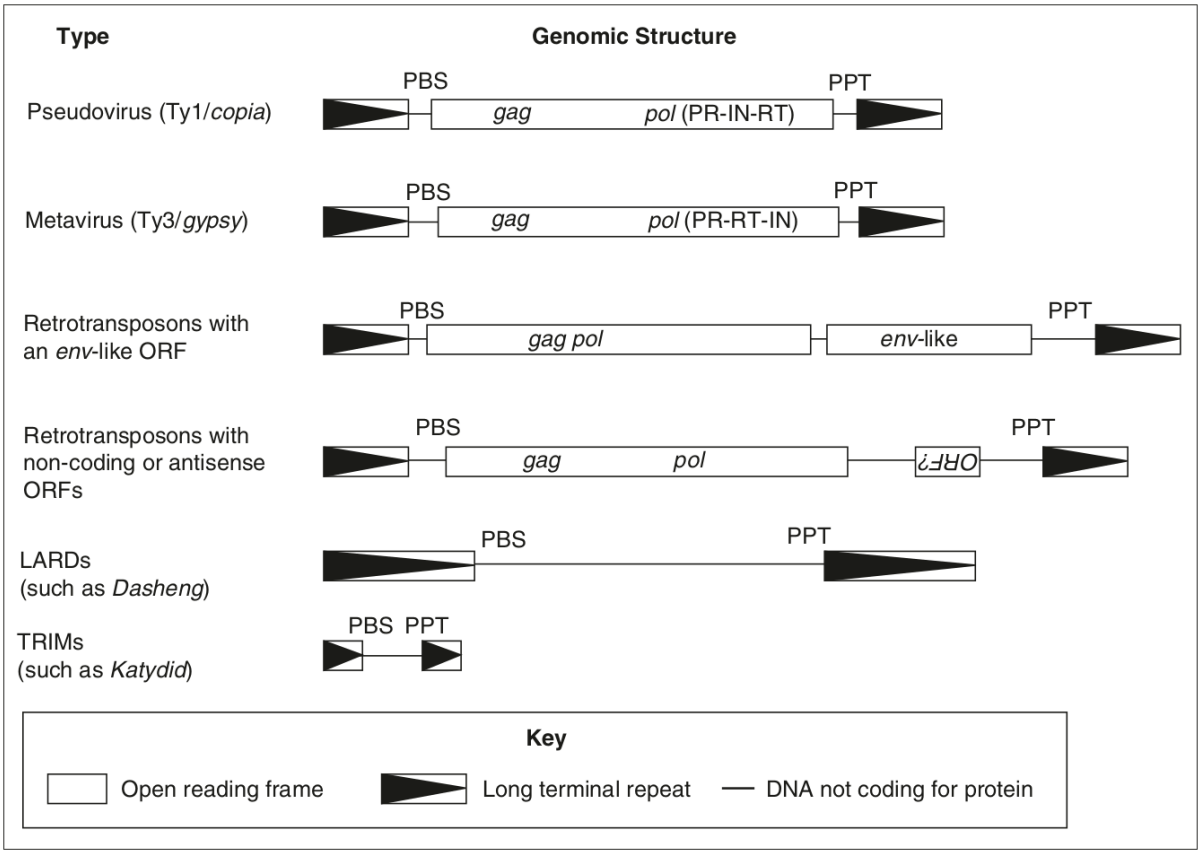

Figura 4. Organización genómica de los diferentes retrotransposones de tipo LTR (RT LTR). Abreviaturas: IN, integrasa; LARD, derivados de retrotransposones grandes; ORF, marco de lectura abierto; PBS, sitio de unión del cebador; PPT, tracto de polipurina; PR, proteasa; RT, transcriptasa inversa (proteína modular con dos dominios separados: uno para actividad polimerasa (transcriptasa reversa) y otro para ribonuclease (ribonuclease H); TRIM, retrotransposones de repetición terminal en miniatura. El texto al revés indica que el ORF se transcribe en la dirección antisentido. Consulte el texto para ver las descripciones de cada tipo de elemento (Havecker et al., 2004).

Mecanismo de retrotransposición LTR

Como se mencionó, la organización estructural y mecánica de los RT LTR presenta gran similitud a los Retroviridae, esta familia de virus de ARN envueltos es muy diversa y son conocidos por suscitar cuantiosas enfermedades en vertebrados. Para realizar la transposición, los retrotransposones en primer lugar deben sintetizar ARNm, éste será exportado fuera del núcleo y posteriormente traducido a proteínas relacionadas a la replicación y, al mismo tiempo el ARNm actúa como molde para la replicación. En el citoplasma, los policistrones son traducidos y las proteínas resultantes procesadas por medio de proteasas. El ARN molde es entonces encapsulado mediante proteínas de cápside, junto con el ARNt (que va a servir como cebador para la síntesis de ADNc) y las enzimas de replicación e integración (Servant & Deininger, 2016). El proceso es similar al de un virus citoplasmático ensamblado a partir de proteínas codificadas por el gen *gag* (Wilhelm & Wilhelm, 2001). El ARNt empareja e hibrida con las secuencias del PBS, comenzando con la síntesis del ADNc a partir de la RT. La RNasa H degrada el híbrido de ARN-ADN dejando fragmentos en la PPT para la síntesis de una segunda cadena de ADNc. Utilizando la nueva hebra de ADNc como molde, la transcriptasa inversa sintetiza ADN de doble cadena del RT LTR completo (Fig. 5)(Finnegan, 2012).

El complejo ADNc doble cadena unido a la INT (enzima que facilita la integración de material genético en el genoma de un organismo) es seguidamente importado al núcleo, donde luego se insertará en el ADN cromosómico. El mecanismo de inserción es mediado por la INT, y la copia del RT LTR puede integrarse en un nuevo locus o, alternativamente, en otro donde ya se encuentra un RT LTR integrándose en este caso por recombinación homóloga (Servant & Deininger, 2016). La INT inserta el retrotransposon LTR en el ADN cromosómico (Fig. 5)(Finnegan, 2012).

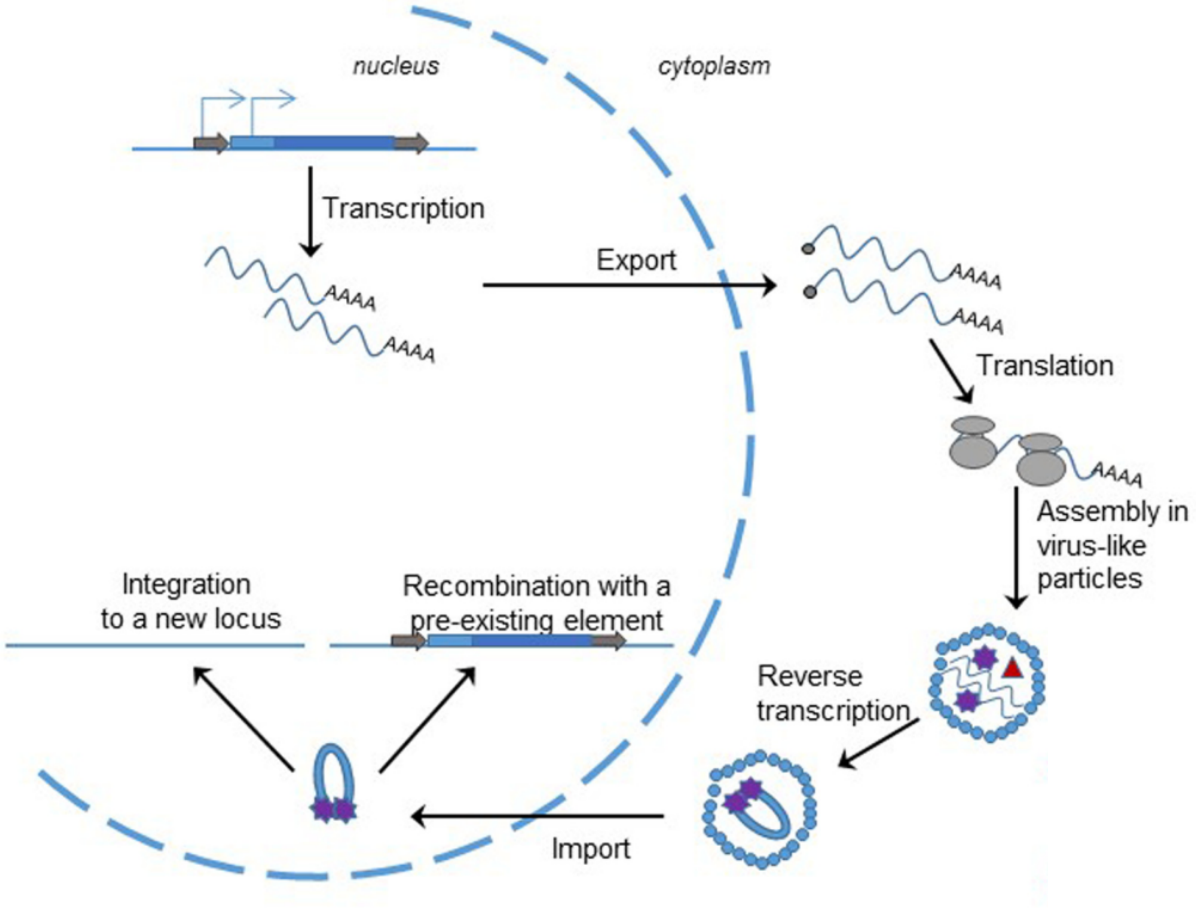

Figura 5. Mecanismo de retrotransposición de un retrotransposon LTR. Líneas azules rectas representan hebras de ADN. Los colores azul claro y azul de las cajas representan los dominios GAG, POL y ORF del retrotransposon LTR. Flechas de color gris pertencen a los lados flanqueantes LTR. Flechas con ángulo de 90 corresponden al sitio de inicio de la transcripción de la secuencia (izquierda de los genes). El ARNm está representado con su región CAP 5', la cola PoliA como una línea azul claro con ondas. Los dos círculos grises unidos pertenecen a las dos subunidades del ribosoma. Las partículas virales se encuentran representadas por círculos azul claro, la organización de estos círculos conforma la cápside similar a la de los virus. El triángulo rojo pertenece a la transcriptasa inversa y las estrellas violetas representan las integrasas (Servant & Deininger, 2016).

Retroelementos LTR en plantas

En muchos organismos se han evidenciado olas de expansión y contracción de ET. Estas olas, que ocurren en períodos acotados, suelen generar grandes diferencias en la arquitectura de los genomas, incluso provocar diferencias en el tamaño de los genomas de especies que están estrechamente relacionadas. En general, este proceso involucra sólo una o pocas familias de ET (por eso el proceso se denomina olas de transposición). Bajo este proceso, los grandes genomas serían en parte resultado de grandes y recientes ráfagas de transposición. Ya sea debido a olas de expansión u otros mecanismos de expansión, se considera a los ET en general, como los principales responsables de los aumentos del tamaño genómico de varios

organismos. En particular, los RT LTR juegan un rol principal dentro de la variación del tamaño genómico, evidenciado por varios autores (El Baidouri & Panaud, 2013).

Por otro lado, se ha demostrado que en algunos casos, los LTR son eliminados eficientemente de los genomas huésped a través de varios mecanismos que involucran deleciones y recombinaciones (Ma et al., 2019). Este modelo postula que el tamaño del genoma y el paisaje de RT LTR en un momento dado, es el resultado de dos fuerzas que se contrarrestan. Por un lado, las olas de retrotransposición que aumentan en gran número los RT LTR y por otro, los mecanismos de eliminación a través de deleciones y recombinaciones (El Baidouri & Panaud, 2013).

Los RT LTR son particularmente importantes en plantas, donde se ha reportado que gran porcentaje del genoma vegetal está compuesto por ellos (El Baidouri & Panaud, 2013; Ma et al., 2004; Piegu et al., 2006). De hecho las plantas presentan un amplio rango de tamaños de genoma, explicado básicamente por la acumulación de RT LTR, que determinan la expansión del tamaño del genoma. En consecuencia, la paradoja del valor C, en la que el tamaño del genoma de una especie no se correlaciona con la complejidad fisiológica del organismo, estaría asociado a la abundancia de RT LTR (Feschotte et al., 2002; Kidwell, 2002). Además, el tamaño del genoma no se correlaciona directamente con la diversidad de los retrotransposones LTR que contienen. En particular, observamos que algunos gimnospermas tienen genomas de hasta 20 GB, pero presentan una baja diversidad de tipos de retrotransposones LTR, aunque estos tipos específicos pueden estar altamente expandidos en el genoma. Esto contrasta con otros modelos de plantas de algunas angiospermas, donde todas las inserciones son recientes y poseen una gran diversidad de familias de ET (Wells & Feschotte, 2020).

Inserción de retroelementos LTR en plantas

En plantas, los sitios de inserción y escisión de los RT LTR corresponden mayoritariamente a regiones heterocromáticas, regiones pobres en genes, y regiones pericentroméricas (Tsukahara et al., 2012; Weber & Schmidt, 2009). Dado su rango de distribución, la mayoría de RT LTR se encuentran insertos de manera anidada en regiones de heterocromatina (SanMiguel et al., 1998) que se encuentra compactada en la mayor parte del ciclo celular. La heterocromatina es una característica fundamental de la arquitectura de los cromosomas eucariotas (Vergara et al., 2017). Estas regiones cromosómicas proporcionan propiedades funcionales clave, desempeñando un papel crucial en la regulación de elementos móviles, el aislamiento de la reparación de ADN en regiones repetitivas y la segregación cromosómica (Allshire & Madhani, 2018). La heterocromatina se divide en dos tipos: heterocromatina constitutiva y heterocromatina facultativa, en función de los componentes de la secuencia presentes en la región heterocromática. La

heterocromatina constitutiva está compuesta principalmente por elementos repetitivos y se encuentra en regiones como los centrómeros y las regiones teloméricas. En contraste, la heterocromatina facultativa se localiza en regiones genéticamente activas y su estado puede cambiar en respuesta a señales celulares y actividad génica (Grewal & Jia, 2007).

A pesar que los ET son más frecuentes en la heterocromatina (Bonchev & Parisod, 2013), cada superfamilia de RT LTR posee distintos patrones de distribución a lo largo de los cromosomas (Gao et al., 2015). En plantas, se encontró que RT LTR Ty1/Copia se distribuye principalmente a lo largo de los cromosomas con una cierta preferencia a la región de la eucromatina (Vicient & Casacuberta, 2017), pudiendo actuar como factor elemental en los eventos de reordenamientos cromosómicos, ganancia, pérdida de genes y cambios en las marcas epigenéticas (de Setta et al., 2014). Por el contrario, los RT LTR Ty3/Gypsy se distribuyen preferencialmente en regiones de heterocromatina desempeñando un factor clave para mantener la estabilidad cromosómica y silenciamiento de la heterocromatina (Gao et al., 2015). Similar a Gypsy, los elementos LINEs poseen una distribución a lo largo de las regiones centroméricas y/o en regiones pericentroméricas (Li et al., 2017).

Se cree que los RT LTR distribuidos en las regiones centroméricas pueden ser el origen de algunas de las repeticiones cortas en tándem implicadas con la función de los centrómeros (Sharma et al., 2013). Asimismo, se ha propuesto que los RT LTR pueden proporcionar los promotores para la transcripción de productos de ARN bicatenario suficiente para crear un cambio estructural epigenético del centrómero (Lippman et al., 2004).

Retroelementos LTR en angiospermas/Myrtaceae

Entre los RT LTR se destacan las superfamilias Ty3/Gypsy y Ty1/Copia (también conocidos como Metaviridae y Pseudoviridae, respectivamente). Los elementos del grupo Ty3/Gypsy, ampliamente distribuidos en angiospermas, presentan una mayor similitud de secuencia y estructural con los retrovirus (Capy, 2005). Sin embargo, difieren en que los retrotransposones, completan su ciclo de vida únicamente dentro de la célula, y los retrovirus infecciosos, se propagan fuera de la célula protegidos por la envoltura proteica codificada por el gen *env*. En diversos retroelementos, se ha observado la presencia de un marco de lectura abierta (ORF) adicional que ocupa la misma posición que el gen *env* presente en los retrovirus (Capy, 2005).

Los retroelementos Ty3/Gypsy incluyen varias familias o linajes, entre ellos el Athila/Tat. Dentro de este linaje, lo distintivo de Athila/Tat es que, en algunos de sus elementos se ha identificado un ORF similar al gen *env* (env-like) de los retrovirus no presente en Ogre (Capy, 2005). En el caso de Athila, además, de la presencia del elemento env-like, su localización preferencial en los centrómeros de *A. thaliana*,

ha despertado el interés en comprender la relación de Athila, con la función y estructura del centrómero, el control epigenético, la segregación cromosómica y mecanismos de especiación (Fukagawa & Kakutani, 2023; Naish et al., 2021; Shimada et al., 2023).

En este trabajo, nos enfocamos en profundizar en el estudio del linaje hermano Athila/Tat/Ogre (OTA), un grupo de RT LTR Ty3/Gypsy que se caracteriza por presentar patrones de distribución entre organismos, y estructura y función únicas comparado con otros retroelementos.

Retroelemento Athila/Tat

Los elementos Athila/Tat son RT LTR que pertenecen a la familia Ty3/Gypsy, la cual se define por el orden de los dominios que codifican proteínas dentro del elemento. Asimismo, los elementos Athila/Tat se encuentran ampliamente distribuidos en plantas terrestres, así como en los hongos *Mucoromycota y Zoopagomycota* (Wang et al., 2021). Recientemente Qin Wang y colaboradores plantearon que la adquisición de esta familia de retrotransposones por parte de las plantas terrestres se debe a la transferencia horizontal a partir de hongos, brindando así una ventaja evolutiva durante la colonización de la tierra a las plantas hace 400 millones de años (Wang et al., 2021). Dentro del clado Tat, se ha identificado un subclado denominado Ogre (Macas & Neumann, 2007). Los elementos del superclado OTA se caracterizan por poseer tamaños grandes, con una región interna entre 10.5 y 25 kilobases (kb) de largo, flanqueados por LTR de entre 1.8 a 6.4 kb a cada lado. Esta región interna codifica dos proteínas: la proteína GAG estructural de la cápside y la Pol, que lleva los dominios de proteasa, RT e INT esenciales para la duplicación de elementos (Macas & Neumann, 2007; Slotkin, 2010). Específicamente la localización de los elementos Athila en el genoma es la más próxima al centrómero comparado con otros TE (Pereira, 2004). Particularmente en *Arabidopsis*, estos elementos se concentran en las regiones cercanas al cinetocoro, en la heterocromatina pericentromérica y también en la región core del centrómero, intercalado con secuencias satélites que corresponden a repetidos en tándem (Fukagawa & Kakutani, 2023; Naish et al., 2021; Shimada et al., 2023; The Arabidopsis Genome Initiative, 2000). Los elementos Tat se detectaron en regiones pericentromericas en café (Cintra et al., 2021) y se ha sugerido que regiones no codificantes de estos elementos pueden ser el origen de secuencias satélites (repetidos en tándem) en estos sitios cromosómicos.

Acca sellowiana

En nuestro país existe un creciente interés por el desarrollo, cultivo e investigación en las plantas nativas. Entre ellas *Acca sellowiana* ha adquirido relevancia debido a sus propiedades nutricionales, valores nutracéuticos, aroma, aspecto y sabor.

En los últimos tiempos, se han logrado avances significativos en el estudio y cultivo de *Acca sellowiana*, también conocido como "el Guayabo del País". Las investigaciones en Uruguay se han abordado desde diferentes perspectivas, la caracterización, evaluación, genética, mejoramiento, propagación y también trabajos sobre diversidad genética, taxonomía y genómica (Quezada et al 2014; Quezada et al 2022; Pritsch et al., 2008;Puppo et al., 2009; Ross et al., 2017).

Acca sellowiana es una Myrtaceae que pertenece a la tribu *Myrteae*. Es una especie diploide 2n = 22, alógama con un genoma de aproximadamente 345 Mpb/1C, considerado un genoma muy pequeño conservado a su vez, en la familia de las Myrtaceae (Quezada et al., 2014). Esta tribu es relevante por su gran diversidad, compuesta por más de 2500 especies, por su producción de frutos, y creciente interés por su consumo. Dentro de las más conocidas se encuentran la guava, pitanga, camu-camu y arazá, y junto con *A. sellowiana* son especies atractivas por sus frutos comestibles. Dentro de las Myrtaceae se encuentra el *Eucalyptus grandis* que a su vez pertenece a la tribu *Eucalypteae*. Actualmente, sabemos que entre los mapas genéticos de *Eucalyptus grandis* y *A. sellowiana,* hay un alto nivel de conservación de sintenia (Quezada et al., 2022). De hecho, se ha reportado para dos especies de Myrtaceae tamaños genómicos similares*: Psidium guajava* “guava” con 2n=22 y 247 Mpb/1C (da Costa et al., 2008), y *E. grandis* 2n=22 y 611 Mpb/1C (Praça et al., 2009). *Eucalyptus grandis* cuenta con un genoma de referencia (Myburg et al., 2014), con una muy buena calidad de anotación. Recientemente, se determinó que en Eucalyptus y en Syzygium (tribu Syzygiae) la familia más importante dentro de la superfamilia Ty3/Gypsy fue Tat/Ogre, representando entre 25 a 33 % del total de linajes Gypsy y con evidencias de abundantes inserciones en los últimos 5 MYA aproximadamente (Ouadi et al., 2022).

Genoma del Guayabo

El equipo de la Dra. Clara Pritsch en la Facultad de Agronomía, Udelar (Fagro) secuenció recientemente el genoma de *Acca sellowiana* en el marco de un proyecto de investigación. El genoma fue realizado mediante una estrategia de secuenciación de moléculas únicas largas con PacBio, obteniéndo 80x de cobertura de secuenciado. El genoma de *A. sellowiana* secuenciado tiene un largo total de 401 Mb distribuidos en 2,075 contigs, mayores a 1 Kb. Presenta un N50 de 2,615,883 y un L50 de 46.

La caracterización de la estructura del genoma permitió la identificación y utilización de genes de interés. Además, permitió visualizar los eventos de reorganización del genoma, procesos de domesticación y mejoramiento genético (Pritsch com. pers.). En este contexto, es de gran relevancia el estudio de los ET en *Acca sellowiana* para comprender el desarrollo de esta especie y motivar el desarrollo de nuevas variantes interesantes de esta planta. Uno de los objetivos del grupo de investigación de Clara Pritsh y donde se enmarca el presente proyecto de finalización de carrera, es confirmar el contenido de ET y mapear la distribución de los linajes de ET más relevantes a lo largo del genoma de *Acca sellowiana* y saber si estos se encuentran activos. Es importante conocer la dinámica de los ET en el genoma de *Acca sellowiana*, desde la detección de los sitios de inserción, hasta la datación de estas inserciones y su nivel de actividad actual.

OBJETIVOS

Objetivo general

El presente trabajo tiene por objetivo realizar un análisis y caracterización detallada de los retroelementos LTR de linaje Athila, Tat y Ogre en el genoma del guayabo del país (*Acca sellowiana*). Para ello, se llevarán a cabo estudios bioinformáticos que permitan identificar, clasificar y caracterizar estos elementos en términos de su distribución, organización y estructura. Se espera que los resultados obtenidos contribuyan a una mejor comprensión de los mecanismos evolutivos que rigen la dinámica de estos retroelementos LTR en el genoma del guayabo del país. Pudiendo tener implicaciones importantes en la mejora genética y la conservación de esta especie que es de gran importancia económica y ecológica.

Objetivos específicos

- Identificar la presencia de elementos completos de los linajes Athila, Tat y Ogre en *A. sellowiana* a través de la identificación de las secuencias de sus dominios proteicos.
- Realizar un mapeo preciso de los linajes de los elementos Athila, Tat y Ogre dentro del genoma de *A. sellowiana* con el fin de comprender mejor su distribución y evolución en esta especie.
- en el genoma de *A. sellowiana* de fácil aplicación a otras especies, permitiendo una identificación más eficiente y precisa de estos elementos.
- Analizar la variabilidad y la relación filogenética existente entre los linajes que conforman el superclado OTA en *A. sellowiana*, lo que permitirá comprender mejor la evolución de estos elementos y su posible impacto en la diversidad genómica de esta especie.
- Desarrollar un flujo de trabajo *in silico* para la extracción de elementos LTR con todos sus genes y dominios.

MATERIALES Y MÉTODOS

Datos

El genoma de *Acca sellowiana* fue recientemente secuenciado por el equipo de Clara Pritsch de la Facultad de Agronomía, Udelar (Fagro). El genoma fue realizado mediante una estrategia de secuenciación de moléculas únicas largas con PacBio, obteniendo 80X de cobertura de secuenciado. Para el presente trabajo contamos con el genoma ensamblado y bases de datos de ET. Para realizar la búsqueda inicial de elementos pertenecientes al superclado OTA se utilizaron ET de referencia, los cuales fueron obtenidos de dos bases de datos. REXdb una base de datos de dominios proteicos de ET y Gypsy Database 2.0 (GyDB). Ambas bases de datos fueron descargadas y se utilizó el superclado OTA.

Formatos de archivos

Durante esta tesina se trabajó con tres tipos de archivos diferentes para el análisis de datos. El formato fasta, el cual es ampliamente utilizado en bioinformática para representar secuencias de nucleótidos o péptidos. Este formato de texto proporciona una estructura clara, lo que facilita su manipulación y análisis con diversas herramientas.

Además del formato fasta, se utilizó el formato Newick. Este formato es empleado específicamente para representar árboles filogenéticos. Este tipo de archivo permite visualizar y analizar las relaciones evolutivas entre diferentes secuencias, brindando información sobre la divergencia entre estas.

También se trabajó con archivos en formato txt. Aunque menos específico para la bioinformática, el formato txt es ampliamente utilizado para almacenar y manipular datos de texto en general. Durante el desarrollo de la tesina, se utilizó este formato para diversos propósitos, como la organización y el procesamiento de datos auxiliares o la generación de informes y resultados intermedios.

Software

Búsqueda por similaridad

Para identificar los ET en el genoma del guayabo se implementaron dos estrategias de búsqueda: por similaridad de secuencias con dominios de referencia en bases de datos de ET; y mediante los programas RepeatMasker y REannotate.

Para la búsqueda por similitud de secuencias en base a algunos dominios del elemento, por ejemplo el RT (retrotranscriptasa), se utilizó el programa BLAST 2.12.0 (Basic Local Alignment Search Tool), y bases de datos específicas de ET.

BLAST es un programa bioinformático de alineamiento de secuencias, ya sea de ADN o proteínas. Este es capaz de comparar una secuencia problema contra una gran cantidad de secuencias que se encuentran en una base de datos. A todos los resultados obtenidos en la comparación del balanceo se le realiza un tratamiento estadístico asignándole una puntuación. Existen cinco tipos de aplicaciones posibles: blastp, blastn, blastx, tblastn y tblastx.

La aplicación blastn crea una consulta de nuestro genoma (query) en nucleótidos a nuestras secuencias (subject) que se encuentran en nucleótidos.

El comando utilizado para la búsqueda por homología del dominio RT fue el siguiente:

```
nohup blastn -query <$genome_reference> -subject
<$database_retrotranscriptasa> -evalue 1e-5 -max_target_seqs 1
-max_hsps 1 -outfmt '6 std qlen slen' > <$output>
```

A blastn se le brindan dos archivos fasta, en este caso uno con el genoma de *Acca sellowiana* y otro con los dominios RT de los linajes Athila/Tat/Ogre extraídos de las bases de datos consultadas. El e-value describe el número de aciertos que uno puede "esperar" ver por azar al buscar en un archivo fasta de un tamaño particular. En otras palabras, es el número de veces que se espera que un query coincida con las secuencias subject por mera casualidad. En todo momento se mantuvo un valor como máximo de e-value de 1e^-5.

La ecuación que modela el e-value en el programa BLAST:

E = Kmn * e^(-lambda * S)

donde:

S es el puntaje de alineamiento obtenido entre la secuencia de query y la secuencia del subject.
K, **m** y **n** son constantes de ajuste determinadas por el tamaño del subject y la longitud del query.
Lambda es la tasa de decaimiento exponencial de la probabilidad de encontrar una coincidencia aleatoria.
e es el número de Euler, es la base del logaritmo natural, aproximadamente igual a 2.71828.

E es el e-value, que indica la expectativa de encontrar una coincidencia aleatoria en una búsqueda de un subject dado.
La opción **-outfmt** determina el formato de la salida (output) del programa. resulte en una tabla de resultados con columnas separadas por tabulador (“\t”) en formato txt. El parámetro **6** especifica el archivo de salida, una tabla de datos, los parámetros **qlen** y **slen** refieren a el largo de la secuencia query y del subject, respectivamente; el parámetro **std** es un especificador de formato estándar.
La opción **-max_target_seqs** define el número de alineamientos que queremos obtener en la búsqueda de blast.
La opción **-max_hsps** es el número máximo de alineamientos que se deben conservar para cualquier par único de consulta. El número máximo de alineamientos mostrado será los mejores resultados a juzgar por el e-value. Este número debe ser un número entero que va de uno o un número mayor.

Se utilizó el programa GScissors.pl en perl para extraer las secuencias con homología con RT de los multifasta que representan el genoma del guayabo, según sus coordenadas. El parámetro <coordinate file> requiere del identificador (ID) de la secuencia, coordenada inicio y final. Para ejecutar el programa se utilizó el siguiente comando.

```
perl GScissors.pl <input fasta file> <coordinate file> <output fasta file>
```

En el proceso de búsqueda de elementos repetidos, también se empleó el programa RepeatMasker, haciendo uso de la misma base de datos utilizada con blast. Se utilizó la versión 4.1.2-p1 de RepeatMasker, y es posible ejecutarlo localmente (https://www.repeatmasker.org/RepeatMasker/). A continuación se muestra el comando utilizado para ejecutar este programa:

```
RepeatMasker -lib <data base> <input fasta file>
```

El resultado <input fasta file>.out del RepeatMasker es procesada por REannotate. REannotate es una herramienta computacional que procesa la anotación de RepeatMasker, generando datos adicionales sobre los elementos detectados. Sus funciones principales incluyen: i) reconocimiento de elementos repetitivos, ii) la resolución del orden temporal de inserciones en elementos anidados y iii) la estimación de la edad de los RT LTR. Este programa se ejecuta de forma local (http://www.bioinformatics.org/reannotate/download.html). A continuación se muestra el comando para ejecutar el programa por defecto:

```
REannotate <input file .out> <input fasta file>
```

Alineamiento de secuencias

Todos los alineamientos de secuencias se hicieron con el software MAFFT con la versión v7.490. MAFFT puede utilizarse de forma online

(https://mafft.cbrc.jp/alignment/server/) o localmente (https://mafft.cbrc.jp/alignment/software/). El comando utilizado fue el que trae el software por defecto. En este sentido, el programa selecciona un algoritmo de alineamiento más óptimo, de acuerdo al tamaño de secuencia y la variación que posee.

```
mafft <archivo fasta> > <archivo salida alineado>
```

Para el curado automático de alineamientos se utilizó TrimAl con la versión (1.2rev59). trimAl es una herramienta para el recorte de alineamientos de forma automatizada, es especialmente adecuada para análisis a gran escala. La velocidad y la posibilidad de ajustar parámetros hace adecuado para curar secuencias a gran escala. TrimAl se ha implementado en C++ posee una versión local que se puede descargar (http://trimal.cgenomics.org/introduction) y un sitio web (http://trimal.cgenomics.org/introduction) con interfaz gráfica. El comando utiliza los parámetros -htmlout, -gt y -st.

```
trimal -in <inputfile> -out out_aligment_RT_abs -htmlout
<ouputfileHTML> -gt 0.8 -st 0.001
```

El parámetro -in indica el archivo de entrada, este acepta varios formatos (clustal, fasta, NBRF/PIR, nexus, phylip3.2, phylip). -out indica el nombre del archivo salida en el mismo formato que el de entrada . El parámetro -gt 0.8 específica que se conservan solo las columnas con gaps en menos del 20% de las posiciones. El parámetro -st indica al software la similitud media mínima permitida.

Transeq es una herramienta bioinformática perteneciente a European Molecular Biology Open Software Suite (EMBOSS). EMBOSS es una suite bioinformática, open source creada por EMBnet que es empleada fundamentalmente en análisis *in silico*. Transeq lee una o más secuencias de nucleótidos y escribe las traducciones de las secuencias a los aminoácidos correspondientes en un archivo de salida.

```
transeq -sequence <input file fasta> -outfile <output file
fasta>
```

YASS (“genomic similarity search tool”) es una herramienta de búsqueda de similitudes genómicas y visualización por dot-plot, para secuencias nucleotídicas en formato fasta. YASS utiliza kmeros para detectar posibles regiones de similitud y luego intenta extenderlas a alineamientos locales. Ademas, de tener un sitio web (https://bioinfo.lifl.fr/yass/index.php) donde hacer las consultas al servidor, se puede clonar el repositorio de github (https://github.com/laurentnoe/yass) y utilizar de

manera local. Permite la detección de los repetidos directos LTR que se esperan en los elementos completos.

Análisis filogenéticos

La construcción de las filogenias se realizó con el software IQ-TREE con una versión multicore 2.0.7 para Linux 64-bit construida en Jan 21 2022. IQ-TREE utiliza un algoritmo estocástico rápido y efectivo para reconstruir árboles filogenéticos por máxima verosimilitud.

La máxima verosimilitud (Maximum Likelihood, ML en inglés) es un método utilizado para inferir filogenias. La idea básica detrás del método es encontrar el árbol filogenético que maximiza la probabilidad de observar las secuencias que se están comparando, asumiendo un modelo de evolución específico. El proceso de inferencia comienza con la construcción de un árbol filogenético inicial, y luego se evalúa la verosimilitud de ese árbol asumiendo ciertas tasas de evolución y la probabilidad de que cada secuencia evolucione a lo largo de cada rama del árbol. Luego, se ajustan los parámetros de las tasas de evolución y se calcula la verosimilitud del árbol nuevamente. Este proceso se repite varias veces hasta que se encuentra el árbol con la máxima verosimilitud.

Para calcular la verosimilitud de un árbol dado, se utiliza la fórmula de verosimilitud, que es la probabilidad de observar las secuencias dadas las condiciones del árbol y del modelo de evolución utilizado. Este cálculo se realiza mediante el uso de algoritmos de optimización numérica, como el método de Newton-Raphson, que buscan los valores óptimos para los parámetros del modelo de evolución.

IQ-TREE, permite realizar la selección de modelo, pruebas de rama, pruebas de topología de árbol, mapeo de verosimilitud. IQ-TREE toma como input un archivo multifasta previamente alineado, y devuelve varios archivos con construcciones filogenéticas en diferentes formatos. Los archivos principales son 3: archivo.iqtree, archivo.treefile y archivo.log. Archivo.iqtree presenta un informe principal que es auto legible con resultados computacionales y contiene una representación árbol final. El archivo.treefile contiene la filogenia por máxima verosimilitud en formato NEWICK. Finalmente, el archivo.log lleva el registro de toda la ejecución del software y un reporte de errores.

```
iqtree2 -s <input file alignment> -B 1000 -m MFP
```

El parámetro -s especifique el archivo de alineación de entrada que puede estar en formato en PHYLIP, FASTA, NEXUS, formato CLUSTAL o MSF. El modelo evolutivo por defecto del IQ-TREE puede no ajustarse muy bien a los datos. Por lo tanto, IQ-TREE permite determinar automáticamente el modelo de mejor ajuste a través

del parámetro -m con distintas opciones. La opción MFP busca el mejor modelo que se ajuste a nuestros datos.

El árbol filogenético con el método máxima verosimilitud se analizó con IQ-TREE (Nguyen et al., 2015), se empleó el flag -B para especificar la generación de un conjunto de 1000 valores de arranque (bootstrap) con el fin de evaluar el árbol. IQ-TREE se ha implementado en C++ y C posee una versión local que se puede descargar (http://www.iqtree.org/) y un sitio web (http://iqtree.cibiv.univie.ac.at/) con interfaz gráfica.

Visualización y manipulación de datos

La visualización, manipulación y estadísticas de datos se realizaron mediante el lenguaje R, en el entorno de RStudio. R es un lenguaje para realizar análisis estadísticos y gráficos. Se encuentra disponible como Software Libre bajo los términos de la Licencia Pública General GNU de la Free Software Foundation en forma de código fuente. RStudio es un entorno de desarrollo integrado. Se utilizó la versión de R 4.2.1. Para la manipulación de los datos con R se utilizaron las librerías tidyverse, dplyr y gridExtra. La visualización de los datos se hizo con las siguientes librerías de R ggplot2, viridis, RColorBrewer y gggenes. La construcción de las filogenias se hizo con las librerías de R BiocManager, ape, tidytree y ggtree.

Flujo de trabajo

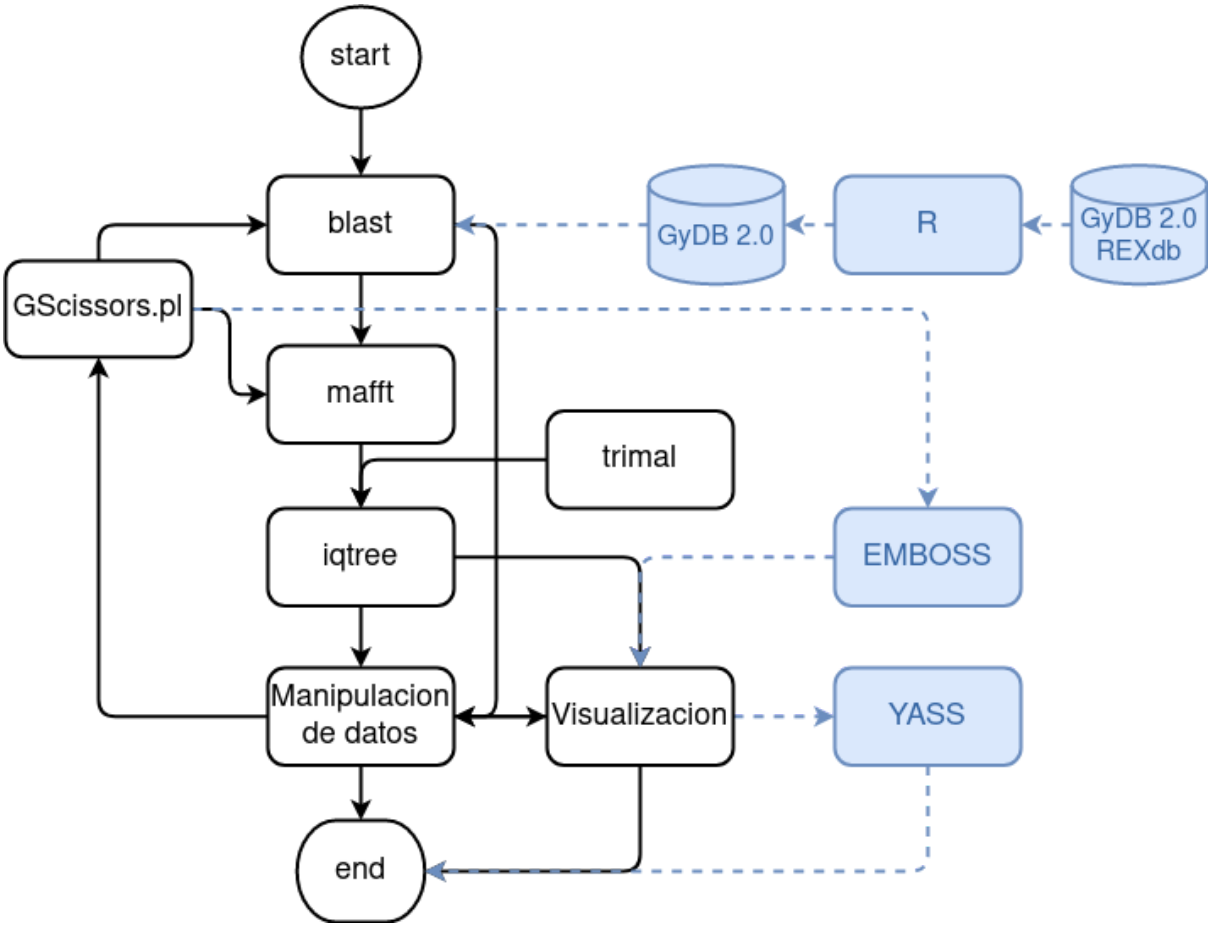

Figura 6. Grafo dirigido acíclico (GDA) para la obtención de retrotransposones long terminal repeat (RT LTR). Las flechas representan la dirección y el sentido del flujo de datos. Las estructuras cilíndricas representan las bases de datos, los círculos representan el inicio y el final del GDA. Los rectángulos representan los diferentes programas, finalmente la estructura en azul son programas y procesos por fuera del flujo de trabajo para obtener información adicional.

El flujo de trabajo utilizado se basó en un trabajo previo de Qin Wang y colaboradores (Wang et al., 2021). Se describe en la Figura 6 y puede resumirse en los siguientes pasos: 1) Búsqueda, expansión del dominio RT y extracción de candidatos 2) Búsqueda de repetidos directos y otros dominios 3) Verificación del linaje en base al dominio RT 4) Búsqueda recursiva y clasificación mediante análisis filogenéticos.

Como paso inicial se realiza la descarga de las bases de datos GyDB, REXdb al entorno de trabajo. Posteriormente, se utilizan los comandos básicos de shell script para obtener las secuencias completas de los retroelementos y secuencias de la RT correspondientes al linaje específico de retrotransposones de interés.

1) Búsqueda, expansión del dominio RT y extracción de candidatos

Inicialmente se generan estadísticos descriptivos de las bases de datos seleccionadas, para los elementos del clado OTA, en particular el largo de los elementos completos, el largo de la región LTR y el largo del dominio RT. A partir de

secuencias del dominio RT de las bases de datos se realiza una búsqueda por similitud utilizando BLAST en el genoma de interés.

Posteriormente, se obtienen las posiciones correspondientes a posibles RT en los contigs del ensamblado. Los identificadores de contigs o cromosomas, y las posiciones son utilizados para extraer las secuencias. Para cada secuencia obtenida se determinó el largo y porcentaje de guanina y citosina (%GC).

Se realiza un alineamiento múltiple con las secuencias obtenidas y un análisis filogenético. A partir de la filogenia se define un subgrupo de candidatos que agrupan mejor con los RT de la base de datos usada, en otras palabras, se descartan secuencias que no agrupan con las bases de datos.

2) Búsqueda de repetidos directos y otros dominios

A partir de cada secuencia aislada (posibles ET) que corresponde a un dominio RT, se obtuvieron las secuencias contiguas de forma de capturar el elemento completo, incluyendo otros dominios y los dos LTR flanqueantes. En cada elemento, las secuencias contiguas se extendieron corriente arriba y abajo en base a la longitud máxima de los RT LTR presentes en la base de datos. La identificación de la región repetida LTR de cada secuencia se llevó a cabo buscando dominios repetidos directos por medio del programa BLAST (Fig. 7). Para validar y visualizar los dominios LTR se realizaron dotplots de cada secuencia contra sí misma, para esto se utilizó la herramienta YASS.

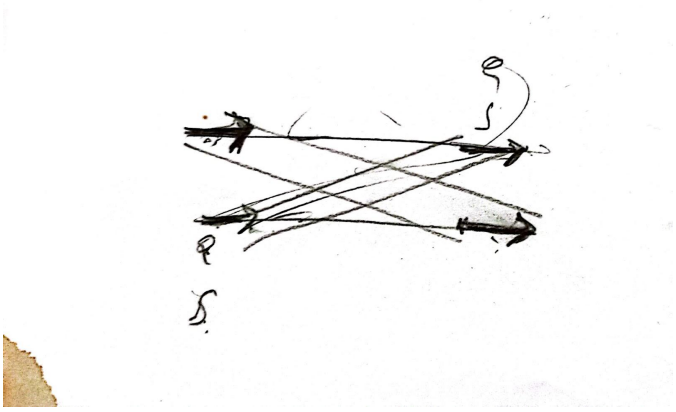

Figura 7. Representación gráfica de la estrategia utilizada para la búsqueda de los repetidos directos. Las líneas paralelas con las flechas representan dos elementos iguales alineados, las flechas oscuras representan los repetidos terminales directos. Las líneas paralelas que se cruzan indican la similitud de los repetidos directos de Q y S.

Para realizar una descripción abarcativa de RT LTR completos, se procedió a identificar los dominios restantes necesarios para que los RT LTR sean autónomos

en su replicación (INT, GAG, RT, RNaseH, AP, ENV y LTR), mediante el programa BLAST. Cada uno de los dominios fueron buscados de forma independiente usando secuencias obtenidas a partir de la base de datos. Se utilizó R (librería tidyverse) para concatenar los resultados de blast de todos los dominios y de los LTR. Finalmente, se realizaron estadísticos descriptivos, largo de secuencia de cada dominio, de los dominios LTR y el gen GAG.

3) Verificación del linaje en base al dominio RT

Para verificar el linaje se realiza un nuevo alineamiento del dominio RT y una filogenia de las secuencias filtradas (obtenidas en el paso 1) y se verifica la presencia del superclado OTA. Aquellos elementos que no agruparon con los linajes buscados fueron eliminados, dejando únicamente los RT LTR del linaje deseado.

4) Búsqueda recursiva y clasificación mediante filogenia

A partir de las secuencias correspondientes a los RT LTR Gypsy/OTA identificados en *A. sellowiana* se realizó una nueva búsqueda. Para esto se realizó un blast utilizando los elementos RT LTR de *A. sellowiana* como “query”. Las nuevas secuencias putativas identificadas se procesan como se describe en el punto 2. Se verifica la presencia de repetidos directos y de los dominios necesarios para la replicación. Finalmente, se realiza un alineamiento múltiple global con todas las secuencias identificadas y se construyen filogenias de los dominios RT, RNaseH e INT para clasificar los elementos de forma individual y concatenada.

RESULTADOS

Para realizar la identificación y análisis de la distribución de RT LTR Ty3/Gypsy/no-cromovirus de las familias Athila, Tat y Ogre en *Acca sellowiana* se utilizó una estrategia híbrida. La misma consistió en una búsqueda por similitud de secuencias para el dominio RT en el genoma de *A. sellowiana*, combinada con un enfoque filogenético de acuerdo a la metodología utilizada por Wang y colaboradores (Wang et al., 2021).

Selección de base de datos y estadísticos descriptivos

Se evaluaron dos bases de datos de ET para identificar cuál se ajustaba mejor para la búsqueda de ET de tipo LTR Athila, Tat y Ogre en el genoma del Guayabo. Por un lado GyDB 2.0 (Gypsy Database), una base de datos abierta empleada para el estudio de la relación evolutiva de virus, elementos genéticos móviles (MGEs) y las repeticiones genómicas. Por otro lado la base de datos asociada al programa RepeatExplorer (REXdb), una base de datos de referencia de dominios proteicos de elementos transponibles de plantas. Para la evaluación comparativa de las bases, se usó la cantidad de resultados de blast (hits) de identificación del dominio RT del clado OTA con las bases de datos de forma individual o usándolas a ambas juntas (Tab. 1.1). Como se puede observar, se obtuvo un número mayor de hits usando sólo la base de datos GyDB (Tab. 1.1). En particular, GyDB obtuvo 432 hits, REXdb género 426 hits y usando GyDB y REXdb conjuntamente se obtuvieron 382 hits (Tab. 1.1). Al contrastar los resultados obtenidos con las diferentes búsquedas, la base de datos GyDB fue la que presentó mayor cantidad de hits en el genoma del Guayabo en diferentes posiciones. Consecuentemente, se utilizó GyDB a lo largo de nuestro flujo de trabajo.

Tabla 1.1	Hits obtenidos (e-value< 1e^-5)
GyDB	432
REXdb	426
GyDB y REXdb	382

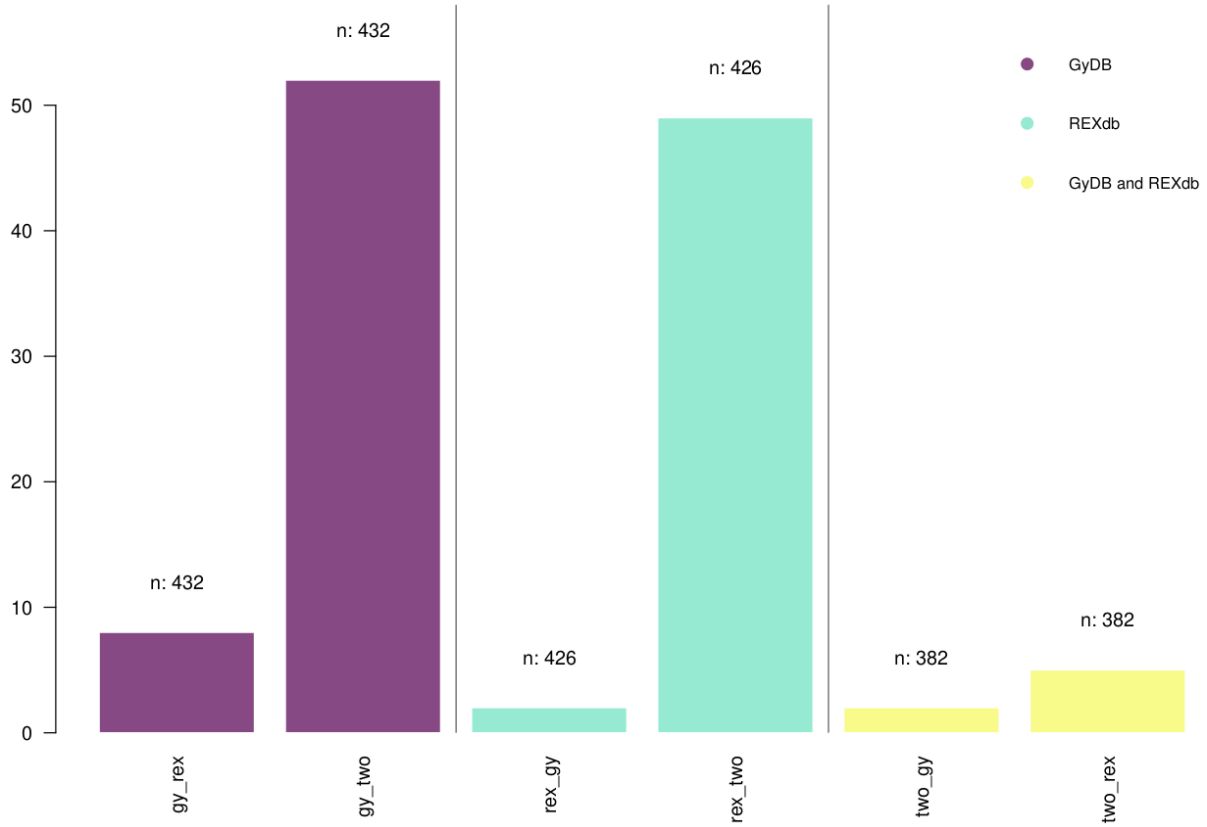

Figura 8. Diferencias de resultados obtenidos a partir de blast con las diferentes bases de datos. Codificación de colores: en violeta es la base de datos GyDB y las diferencias con las otras bases de datos. En celeste es REXdb y las diferencias con las otras bases de datos. En amarillo son GyDB junto con REXdb y las diferencias contra cada una de las base de datos por separado. Abreviaturas: gy_rex = resultados distintos entre GyDB y REXdb, gy_two = resultados distintos entre GyDB y GyDB junto con REXdb, rex_gy = resultados distintos entre REXdb y GyDB, rex_two = resultados distintos entre REXdb y GyDB junto con REXdb, two_gy = resultados distintos entre GyDB junto con REXdb y GyDB; y two_rex = resultados distintos entre GyDB junto con REXdb y REXdb.

Los elementos LTR poseen dos genes, Gag y Pol. El gen Pol a su vez está compuesto por dominios INT, RNaseH, AP y RT. La RT es un dominio codificante el cual posee una alta conservación de secuencia. Dada esta característica la búsqueda de los RT LTR se enfocó en este dominio.

Como primera aproximación para describir los elementos Athila/Tat/Ogre (OTA) consenso incluidos en las base de datos GyDB, se realizaron estadísticos descriptivos para los elementos completos que comprenden las variables: largo de la región LTR, largo de RT y largo de los elementos totales (Fig. 8.1 y 8.2). El total de elementos OTA encontrados en GyDB fue 13. El elemento Athila/Tat más corto observado es Bagy-2 con 5 Kb aproximadamente. Los elementos con el largo máximo lo registraron B1147A04.5 y Ogre con aproximadamente 23 Kb de largo (Fig. 8.1). La media del largo del superclado OTA es de 13352 pb, con un desvío estándar de 4576 pb (Fig. 8.1).

El largo promedio de la región LTR del superclado OTA de GyDB es de 1330 pb, con un desvío estándar de 1298 pb, siendo esta región la que posee mayor variabilidad en cuanto a su longitud. Los elementos Ogre son los que presentan regiones LTR de mayor envergadura con aproximadamente 5 Kb. El resto de las regiones LTR presentan un largo que va entre 500 pb y 2Kb (Fig. 8.2).

El largo promedio de la RT es de 691 pb, con un desvío estándar de 31 pb, siendo ésta la región más conservada en cuanto a la longitud (Fig. 8.2).

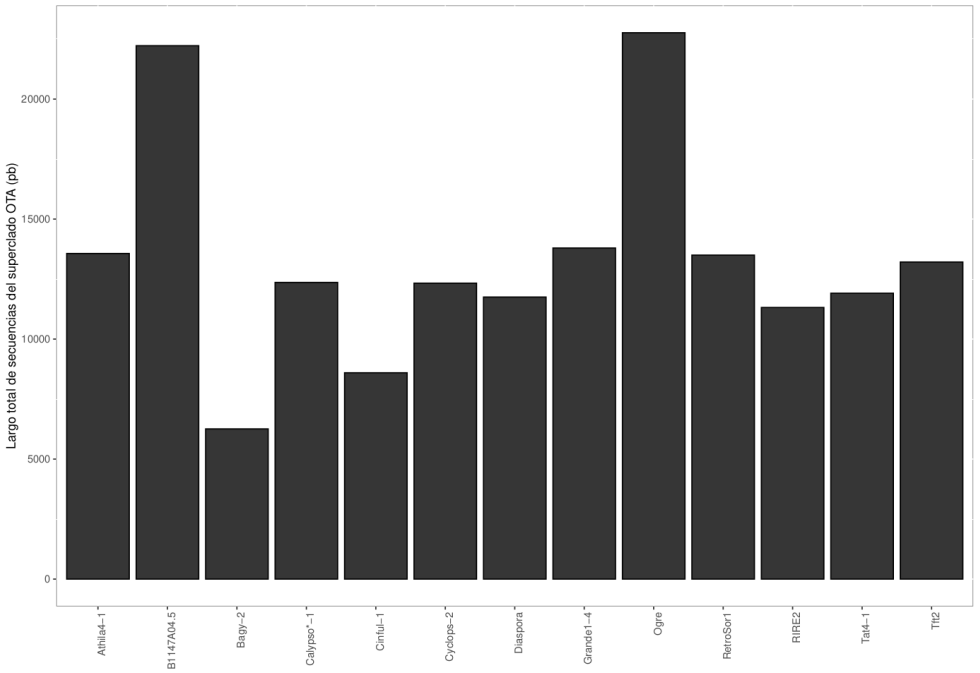

Figura 8.1. Largo del dominio RT LTR del superclado OTA completos (pb) pertenecientes a la base de datos GyDB 2.0

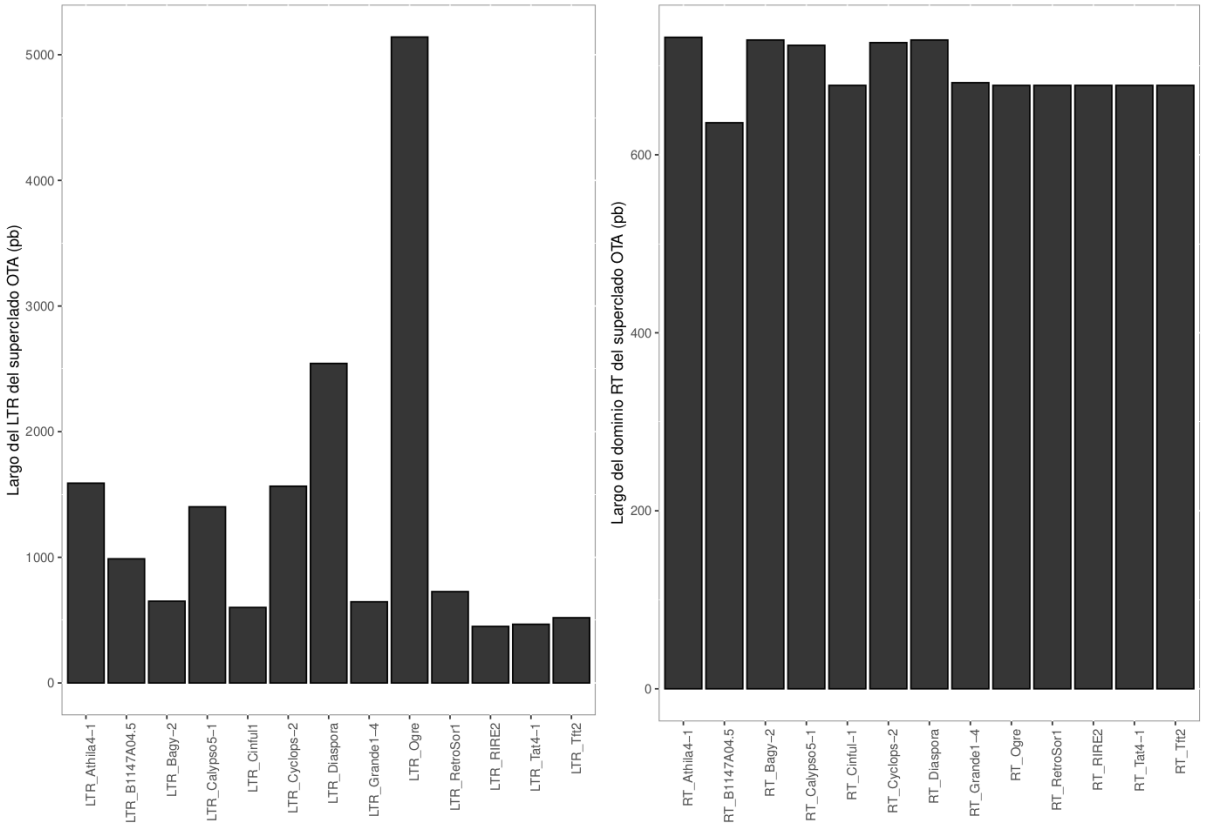

Figura 8.2. Largo (pb) de la región LTR (izq.) y del dominio transcriptasa inversa (RT; der.) de los retrotransposones pertenecientes al del superclado OTA de la base de datos GyDB 2.0.

Candidatos de RT LTR Athila/Tat

Para la identificación de secuencias candidatas de RT LTR del superclado OTA en el genoma de *A. sellowiana* se procedió a extraer del ensamblado genómico de *A. sellowiana* las secuencias asociadas a los 432 dominios RT identificados en dicho genoma. Además, se incluyeron 13 RT pertenecientes al superclado OTA de la base de datos, y se añadió una secuencia RT de un Cromovirus CRM (Ty3/Gypsy) como grupo externo (outgroup).

Se realizaron estadísticos descriptivos para las RT candidatas extraídas. Se estudiaron las distribuciones de la longitud y %GC, teniendo sólo el %GC una distribución normal con un valor de media de %GC 42.58 (Fig. 9). El largo promedio obtenido para RT candidatas es 570 pb, con un desvío estándar de 152 pb siendo un 15% menor al promedio del largo RT (691 pb) de Ty3/Gypsy/OTA de GyDB y una distribución bimodal.

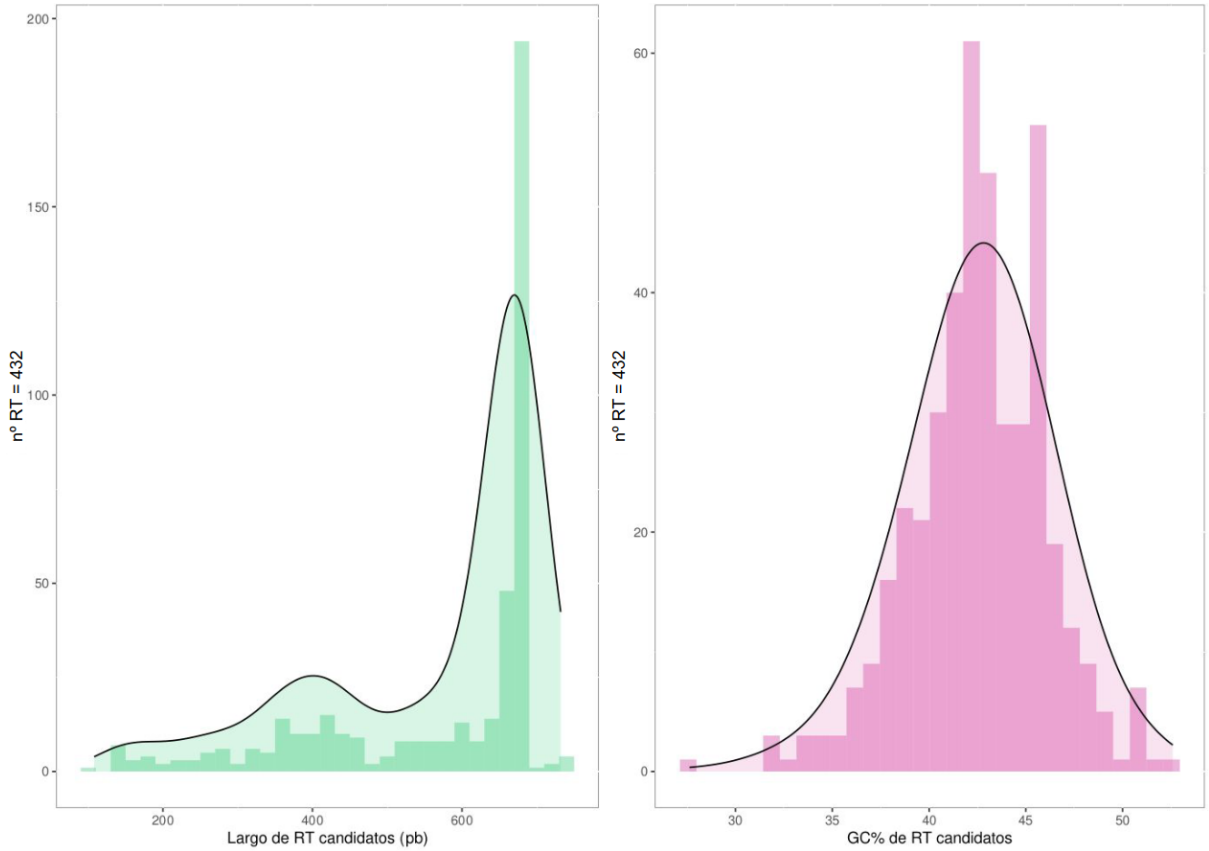

Figura 9. Histogramas de largo y composición. En verde: Distribución del largo de 432 secuencias para la transcriptasa inversa (RT) de Athila/Tat candidatos. En violeta: Distribución del porcentaje de guanina y citosina (%GC) de 432 secuencias RT candidatas en *Acca sellowiana.*

Las 432 secuencias RT candidatas se alinearon con MAFFT junto con las 13 RT del clado OTA de la base de datos y el outgroup. A partir del alineamiento se realizó la filogenia con IQ-TREE generando bootstrap de 1000 réplicas. En la filogenia se observan 4 clados principales (Fig. 10.1). Los dominios RT de Athila, Tat y Ogre de la base de datos GyDB se agruparon con uno de los cuatro clados obtenidos. Indicando que 289 secuencias candidatas Athila/Tat en el genoma de guayabo están relacionadas al superclado OTA (clado 1), con más fuerte relación en particular con Ogre/Tat. Ningún elemento de *A. sellowiana* se agrupó en los mismos grupos que los OTA de referencia. Las restantes 143 secuencias de *A. sellowiana* formaron otros agrupamientos más alejados, indicando que no están evolutivamente relacionadas con el clado OTA por tratarse de otro tipo de secuencia, mutaciones o sesgos en el proceso de extracción. Con el objetivo de retener los candidatos más confiables a pertenecer al superclado OTA se eliminaron posibles falsos positivos. Así, se descartaron las secuencias correspondientes a los dos clados más lejanos a los elementos de las bases de datos. De esta forma se redujo el número de secuencias candidatas a 289 correspondientes al clado 1 (que se agrupa con elementos de GyDB) y el clado 2 (hermano del anterior).

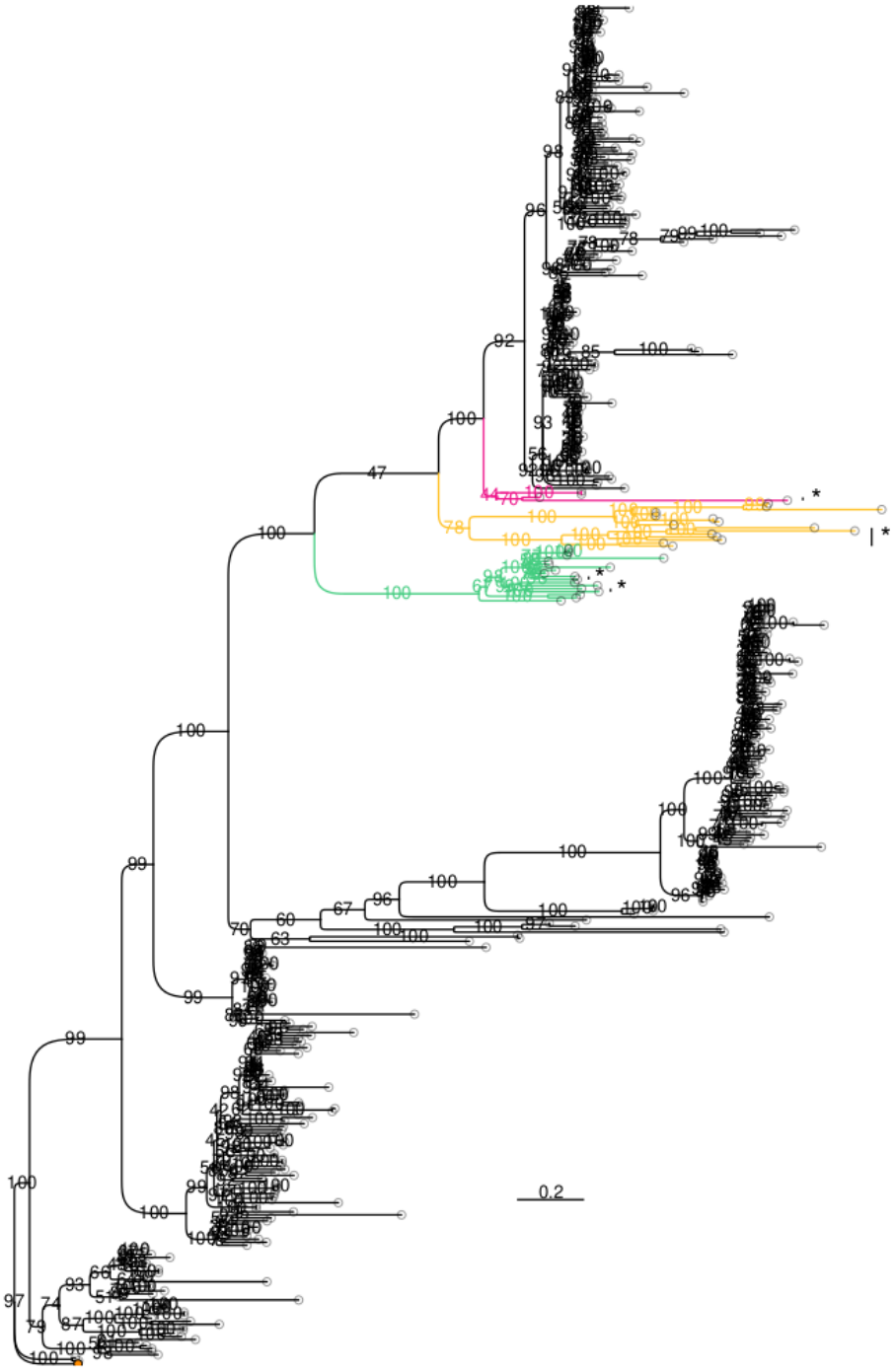

Figura 10.1. Árbol filogenético de retrotransposones LTR (RT LTR) del linaje Athila/Tat. La filogenia se realizó con 432 secuencias correspondientes a la transcriptasa inversa (RT) de Athila/Tat candidatas en *Acca sellowiana.* Las ramas de los candidatos RT se encuentran en negro con círculos negros. Las ramas en verde y marcadas con asterisco corresponden a RT LTR de Athila de la base de datos de referencia. Las ramas rosadas y marcadas con asterisco corresponden a RT LTR de elementos Ogre de referencia. Las ramas amarillas y marcadas con asterisco corresponden a RT LTR de elementos Tat de referencia. En naranja (punto) se encuentra el outgroup utilizado, que corresponde a la secuencia correspondiente a la RT de LTR gypsy perteneciente a chromovirus CRM.

Para identificar los elementos completos, se obtuvieron las secuencias contiguas (corriente arriba y corriente abajo) de los dominios RT. Estas secuencias pueden contener los otros dominios codificantes y las regiones LTR. Estas secuencias se utilizaron para identificar los repetidos terminales realizando blast contra sí mismas. El resultado se filtró en busca de los repetidos directos, usando como largo de alineamiento mínimo 450 pb, este valor se seleccionó considerando el valor mínimo del largo de las regiones LTR de la base de datos GyDB.

Visualización de RT LTR

Para corroborar la existencia de estos repetidos directos se visualizaron algunos de los resultados obtenidos en la búsqueda de las regiones LTR y se graficaron dotplot mediante el programa YASS (Fig. 11)

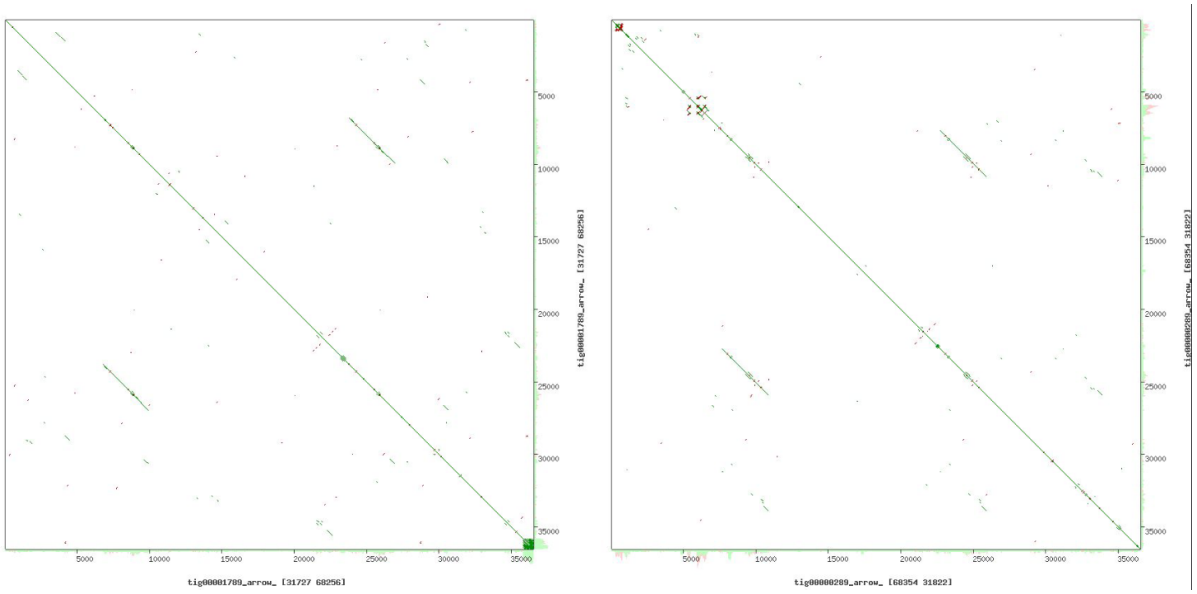

Figura 11. Dotplot de dominios transcriptasa inversa (RT) candidatos extendidos de *Acca sellowiana*. Se visualiza la homología de secuencia de la RT contra sí misma (línea diagonal central). Las regiones de los LTR se pueden observar cómo líneas paralelas fuera de la diagonal central.

Caracterización de elementos candidatos de RT LTR Athila/Tat

A partir de los 289 elementos candidatos RT LTR con ambas secuencias LTR , se realizó la búsqueda de cada dominio codificante, RT, GAG, RNaseH, INT, AP y ENV. El dominio ENV no fue encontrado en ninguno de los elementos analizados. Se consideraron elementos completos de la familia RT LTR Ty3/Gypsy aquellos cuya organización de los dominios correspondiera a la reportada para esta familia. Se filtraron los datos obtenidos mediante el orden LTR’5, GAG, AP, RT, RNaseH, INT y

LTR'3. Mediante este filtro, se obtuvieron un total de 103 secuencias. Estas secuencias se consideran elementos Athila/Tat completas y presentan un largo promedio 20476 pb, con un desvío estándar de 4505 pb.

Se realizaron diagramas visuales utilizando la ubicación de cada dominio para los 103 elementos del superclado OTA de *Acca sellowiana* identificados. Estos esquemas permiten visualizar la distribución de los dominios a lo largo de los elementos y pueden proporcionar información sobre su organización y patrones de inserción en la especie estudiada. Como ejemplo se muestra la visualización de 10 elementos al azar (Fig. 12).

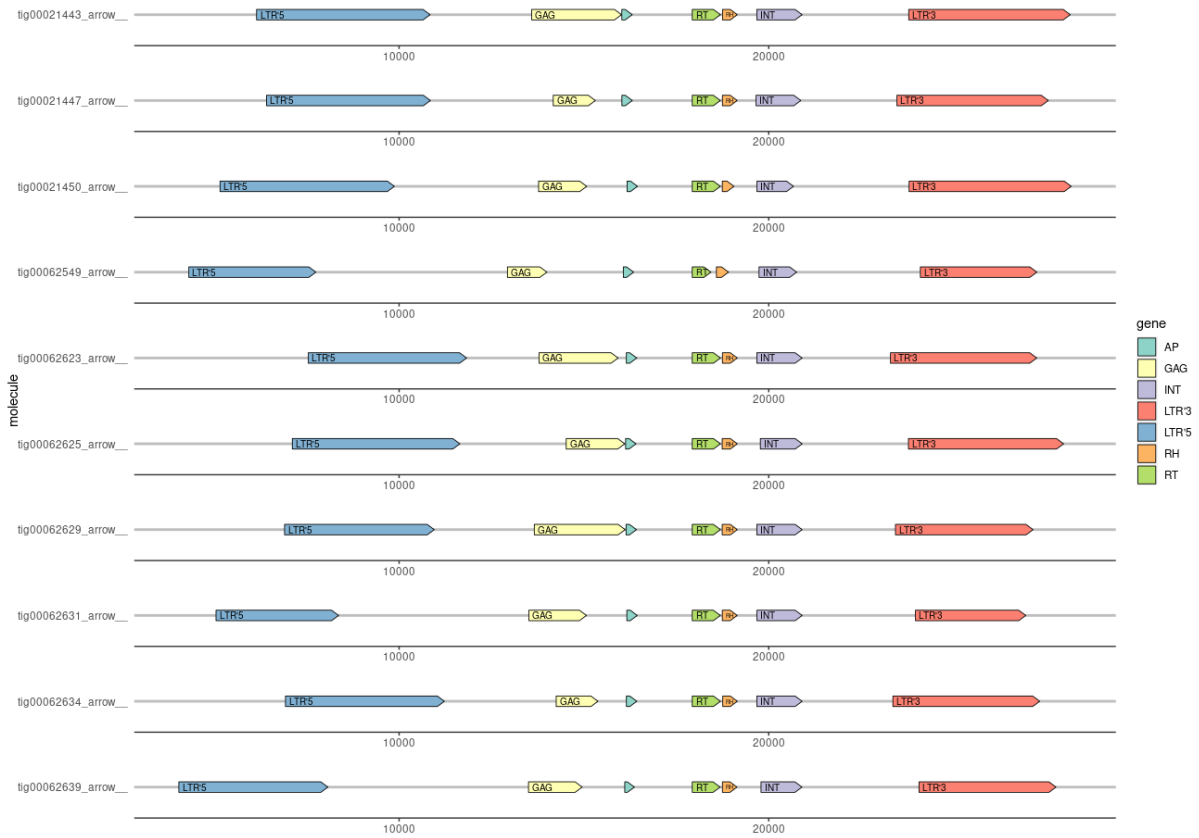

Figura 12. Representación esquemática de 10 secuencias candidatas retrotransposones LTR (RT LTR) OTA, de un total de 103. Las representaciones esquemáticas se crearon utilizando inicio y final de cada dominio. En la parte inferior de cada secuencia se muestra una escala en pares de bases (pb). Los distintos dominios se representan con diferentes colores: LTR = repetición terminal larga (azul y rosado); GAG = mordaza (amarillo); AP = proteasa aspartica (azul claro); RT = transcriptasa inversa (verde claro); RH = Ribonucleasa H (naranja); INT = Integrasa (púrpura).

Para los 103 elementos identificados, se realizó un análisis de la distribución de los largos de cada dominio incluyendo la región LTR. En particular para el gen Gag se distinguen 3 largos distintos, aproximadamente de 500, 1000 y 1800 pb. Para el resto de los dominios la distribución del largo es más uniforme, observándose un sólo pico para cada dominio (Fig. 12). Además se realizó un análisis filogenético utilizando solamente las secuencias RT (Fig. 14). Para los dominios INT, RT,

RNaseH y LTR se obtuvieron longitudes esperadas. No obstante, se observaron tres variaciones en la longitud del gen GAG. Finalmente, las longitudes totales del RT LTR cumplieron con lo esperado.

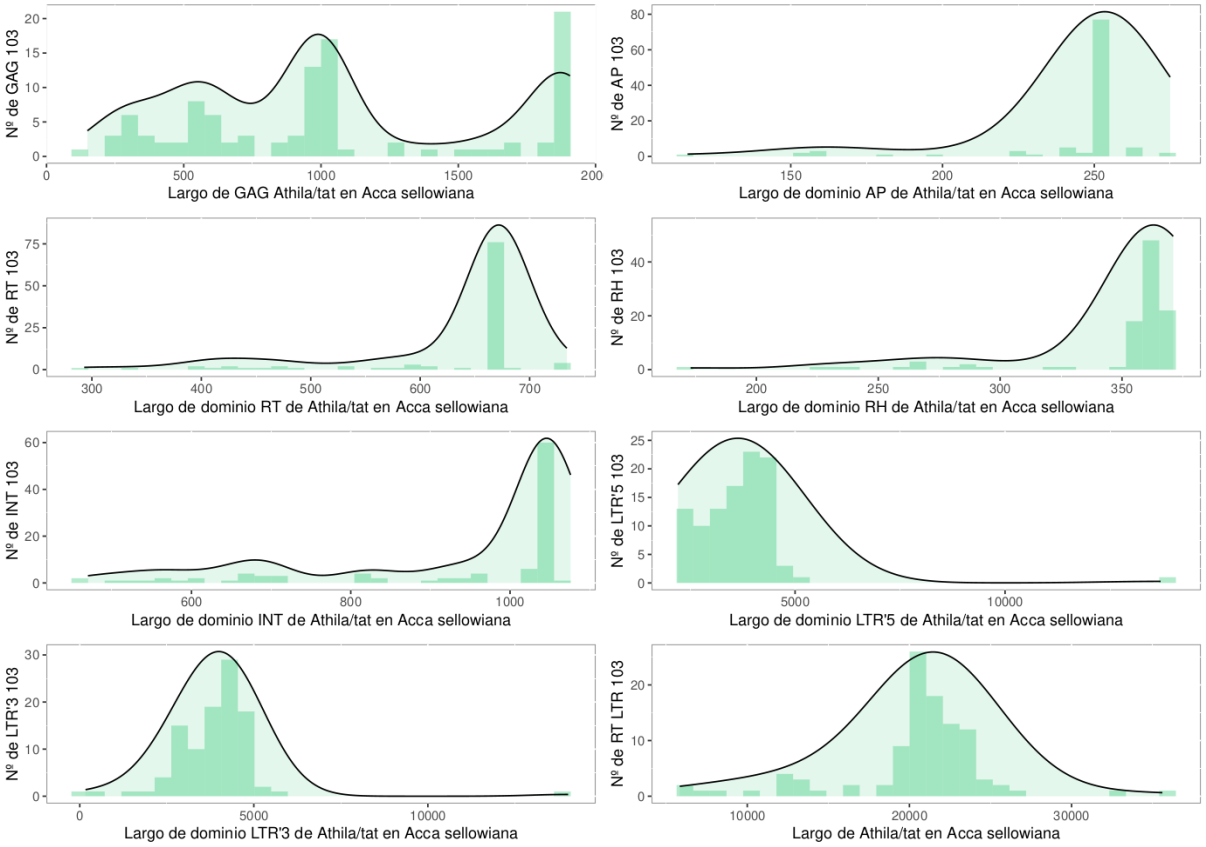

Figura 13. Histogramas del largo de los elementos Athila/Tat en *Acca sellowiana* y de los diferentes dominios que conforman los RT LTR (RT LTR) de linaje Athila/Tat en *Acca sellowiana*. LTR = repetición terminal larga; GAG = mordaza; AP = proteasa aspártica; RT = transcriptasa inversa; RH = Ribonucleasa H; INT = Integrasa.

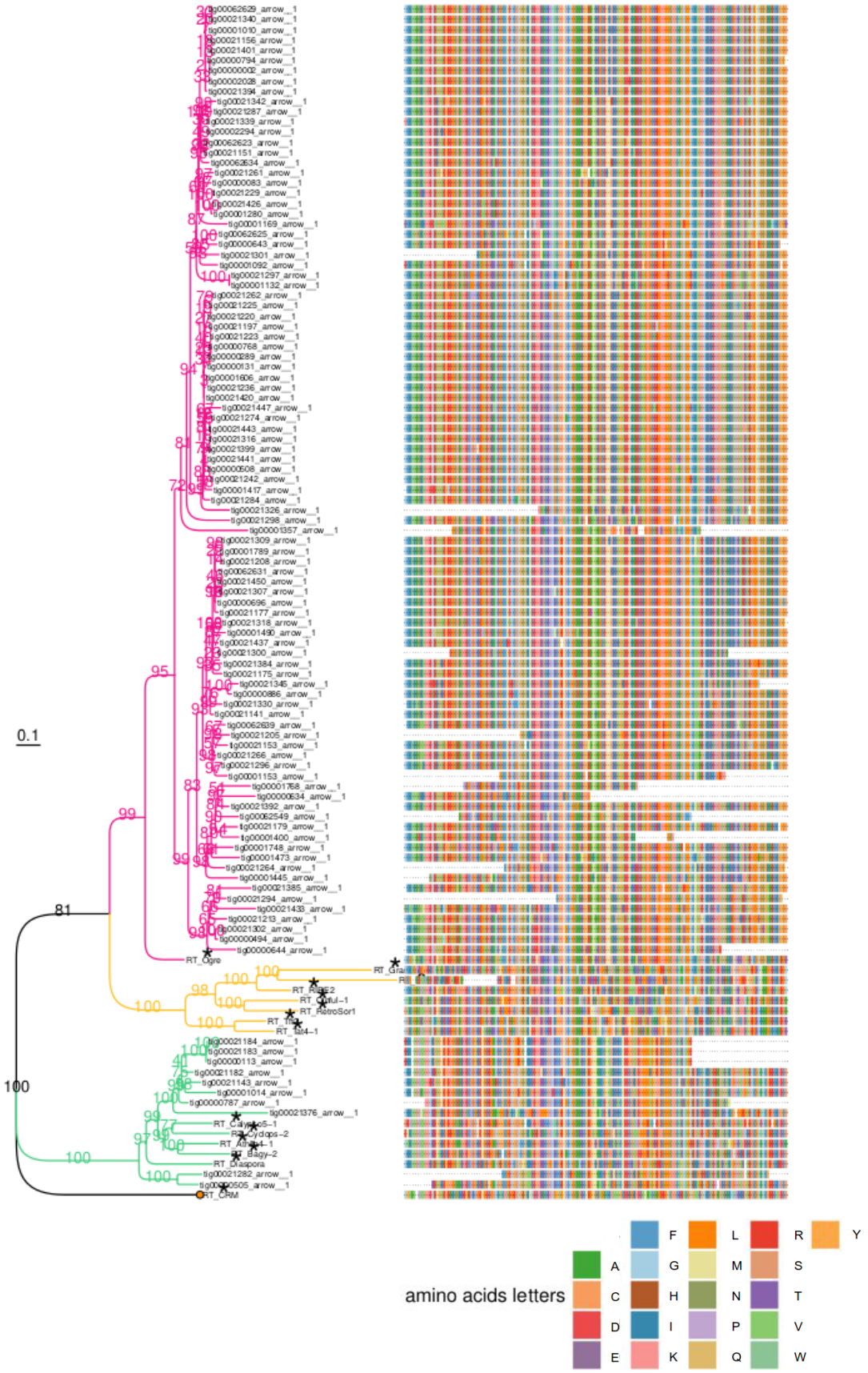

Figura 14. (derecha) Árbol filogenético de retrotransposones LTR (RT LTR) del linaje Athila/Tat, a partir del curado de transcriptasas reversas (RT) y su respectivo alineamiento. Los árboles se calcularon utilizando máxima verosimilitud a partir de los alineamientos de las secuencias en

aminoácidos de los dominios RT. Es un árbol filogenético con raíz, de secuencias RT de Athila/Tat en *Acca sellowiana*. Las rama en verde son RT de *Acca sellowiana* agrupadas con RT Athila de Gypsy DataBase (GyDB), en amarillo y marcadas con asterisco corresponden a las RT Tat de GyDB y las ramas en rosado son RT de *Acca sellowiana* agrupada con las RT Ogre de GyDB. El círculo en naranja representa el outgroup que es un RT LTR, gypsy, chromoviruse CRM. Las terminales marcadas con asterisco corresponden a RT LTR de GyDB. (izquierda) Alineamiento en aminoácidos del dominio RT de RT LTR Gypsy/Ty3 en *Acca sellowiana* recortados con trimAl. Los colores fueron generados por la libraria ggtree con la función msaplot.

El análisis filogenético de 103 elementos completos candidatos a Athila/Tat en base al dominio RT mostró que los elementos se agrupan en dos clados principales, particularmente uno corresponde a la rama verde, compuesta por los elementos del linaje Athila de referencia que agrupa a 10 elementos de *A. sellowiana* y otro corresponde a una rama amarilla, compuesta por elementos Tat solo con RT obtenidas de la base de datos. Los elementos del linaje Ogre de referencia (fucsia) que agrupa a los restantes elementos OTA de *A. sellowiana*, ningún elemento de *A. sellowiana* se relacionó fuertemente con el clado de Tat de referencia que corresponden con la la rama marcada de amarillo se encuentran los elementos de la base de datos que pertenecen exclusivamente al linaje Tat de la base de datos. La rama amarilla contiene los elementos Tat de referencia; en este caso, ningún elemento de *A. sellowiana* se vinculó con este grupo. Por otro lado, la rama verde alberga tanto los elementos del cluster Athila de la base de datos como de *A. sellowiana*. En este caso, 10 elementos de *A. sellowiana* fueron asignados a este grupo. Se puede observar que la rama que contiene el mayor número de elementos es aquella que contiene los elementos Ogre (en rosado) y 93 elementos de *A. sellowiana*. Por último, la rama con un punto de color naranja corresponde al grupo de referencia externo, CRM. Todos los elementos de *A. sellowiana* fueron incluidos en el árbol.

Búsqueda recursiva y clasificación de RT LTR Ty3/Gypsy

Se realizó una búsqueda recursiva, a partir de los 103 elementos ya identificados, en el genoma *A. sellowiana*, utilizando la misma metodología anteriormente utilizada. En resumen se realizó una búsqueda por blast en el genoma de *A. sellowiana* usando como subject los 103 elementos, se filtraron los hits por largo, se extrajeron, se identificaron los repetidos LTR y los dominios génicos y se verificó que el orden fuera el establecido para RT LTR. Así, se obtuvieron un total de 21 nuevos elementos putativos RT LTR pertenecientes al superclado OTA.

En total, se identificaron 124 nuevas RT LTR Ty3/Gypsy en el superclado OTA en *A. selowiana*. Se realizó un análisis detallado de las características estructurales y filogenéticas con todos elementos identificados incluyendo las referencias de las

bases de datos. En particular, se realizó un análisis filogenético para cada dominio INT, RT y RNaseH y para la concatenación de los mismos.

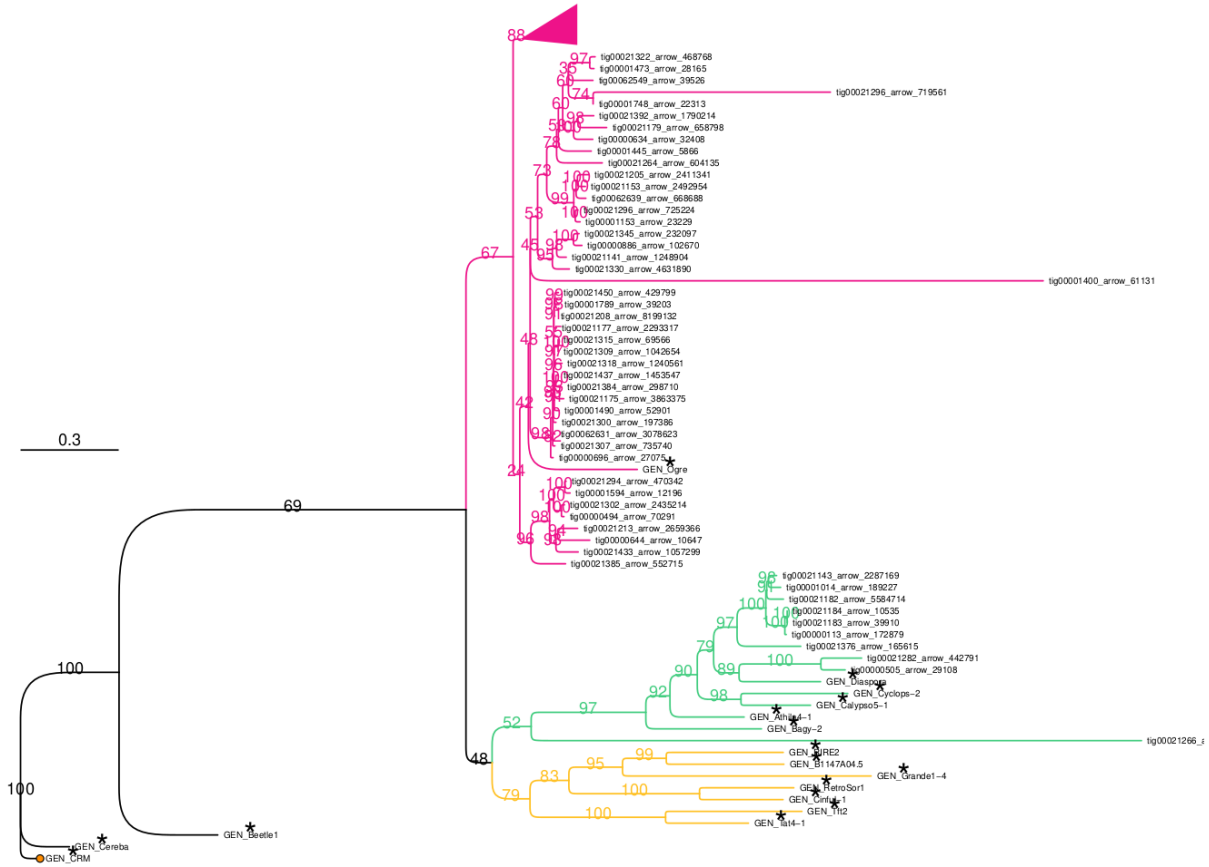

Figura 14.1. Árbol filogenético del dominio integrasa (INT) utilizando todos los retrotransposones LTR (RT LTR) pertenecientes a los linajes Athila, Tat y Ogre (OTA) de *Acca sellowiana*. La filogenia se construyó utilizando máxima verosimilitud a partir del alineamiento de las secuencias nucleotídicas de los dominios INT recortado con trimAl. En verde elementos Athila, en amarillo elementos Tat y en rosado elementos Ogre. El outgroup, RT LTR Ty3/Gypsy de chromoviruse CRM corresponde al círculo naranja. En cada elemento se encuentra el nombre del contig al que pertenecen, inicio y fin de su ubicación, aquellos marcados con asterisco corresponden a la Gypsy database 2.0 (GyDB).

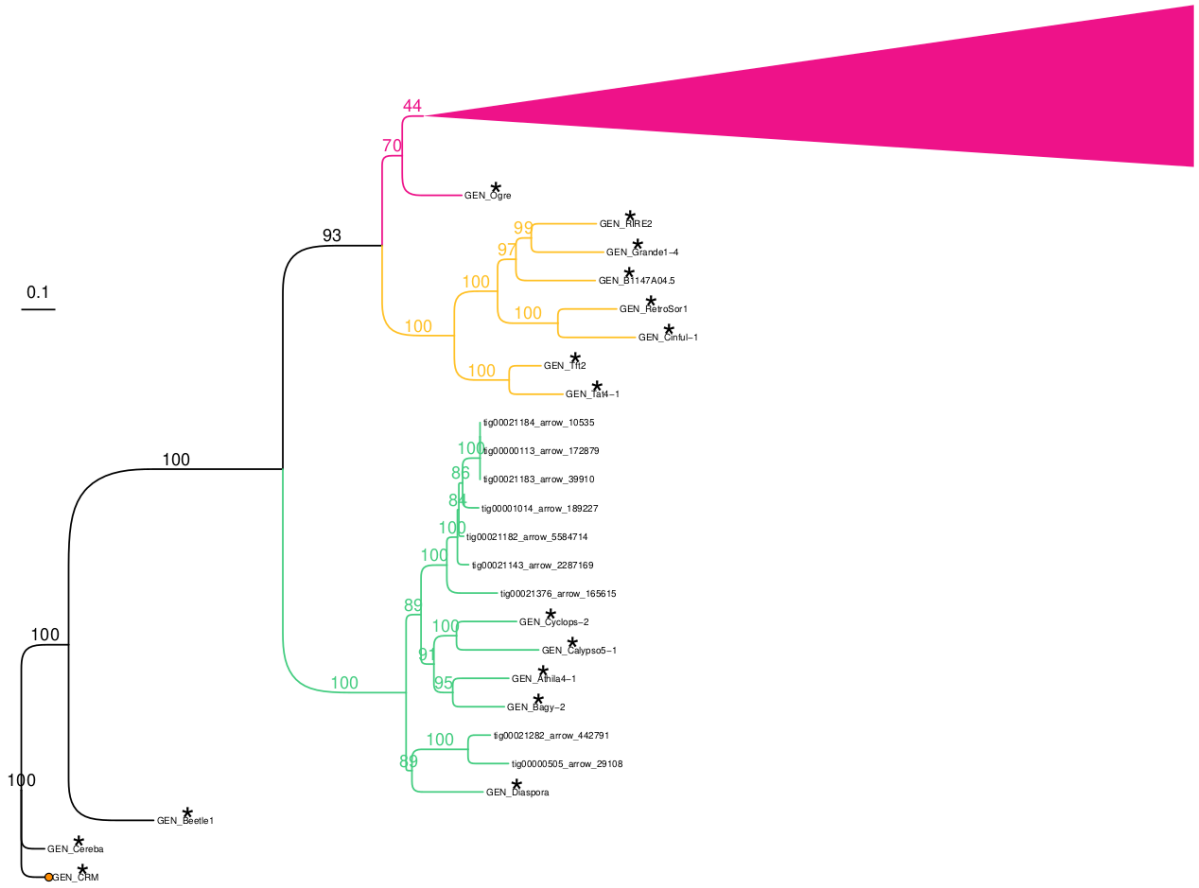

Figura 14.2. Árbol filogenético del dominio transcriptasa reversa (RT) utilizando todos los retrotransposones LTR (RT LTR) pertenecientes a los linajes Athila, Tat y Ogre (OTA) de *Acca sellowiana*. La filogenia se construyó utilizando máxima verosimilitud a partir del alineamiento de las secuencias nucleotídicas de los dominios RT recortado con trimAl. En verde elementos Athila, en amarillo elementos Tat y en rosado elementos Ogre. El outgroup, RT LTR Ty3/Gypsy de chromoviruse CRM corresponde al círculo naranja. En cada elemento se encuentra el nombre del contig al que pertenecen, aquellos marcados con asterisco corresponden a la Gypsy database 2.0 (GyDB).

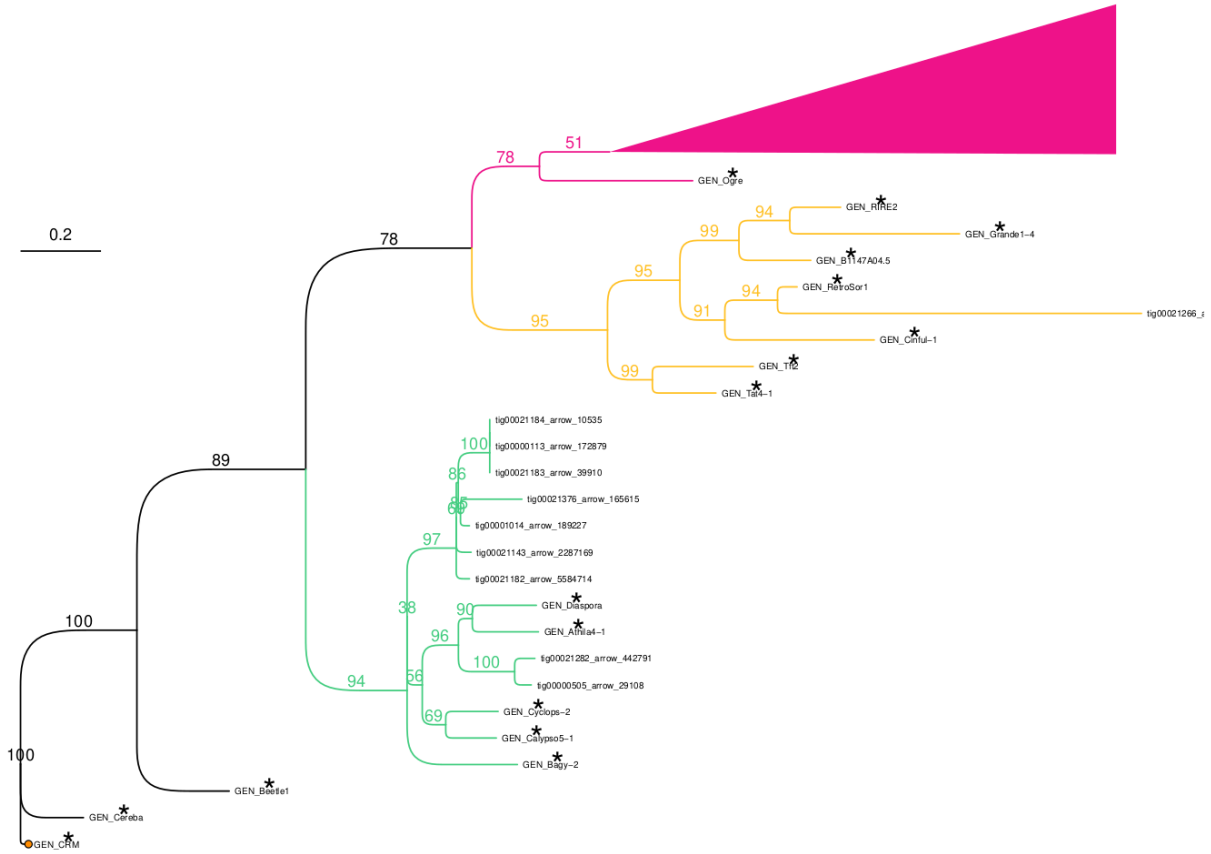

Figura 14.3. Árbol filogenético del dominio ribonucleasa H (RNaseH) utilizando todos los retrotransposones LTR (RT LTR) pertenecientes a los linajes Athila, Tat y Ogre (OTA) de *Acca sellowiana*. La filogenia se construyó utilizando máxima verosimilitud a partir del alineamiento de las secuencias nucleotídicas de los dominios RNaseH recortado con trimAl. En verde elementos Athila, en amarillo elementos Tat y en rosado elementos Ogre. El outgroup, RT LTR Ty3/Gypsy de chromoviruse CRM corresponde al círculo naranja. En cada elemento se encuentra el nombre del contig al que pertenecen, inicio y fin de su ubicación, aquellos marcados con asterisco corresponden a la Gypsy database 2.0 (GyDB).

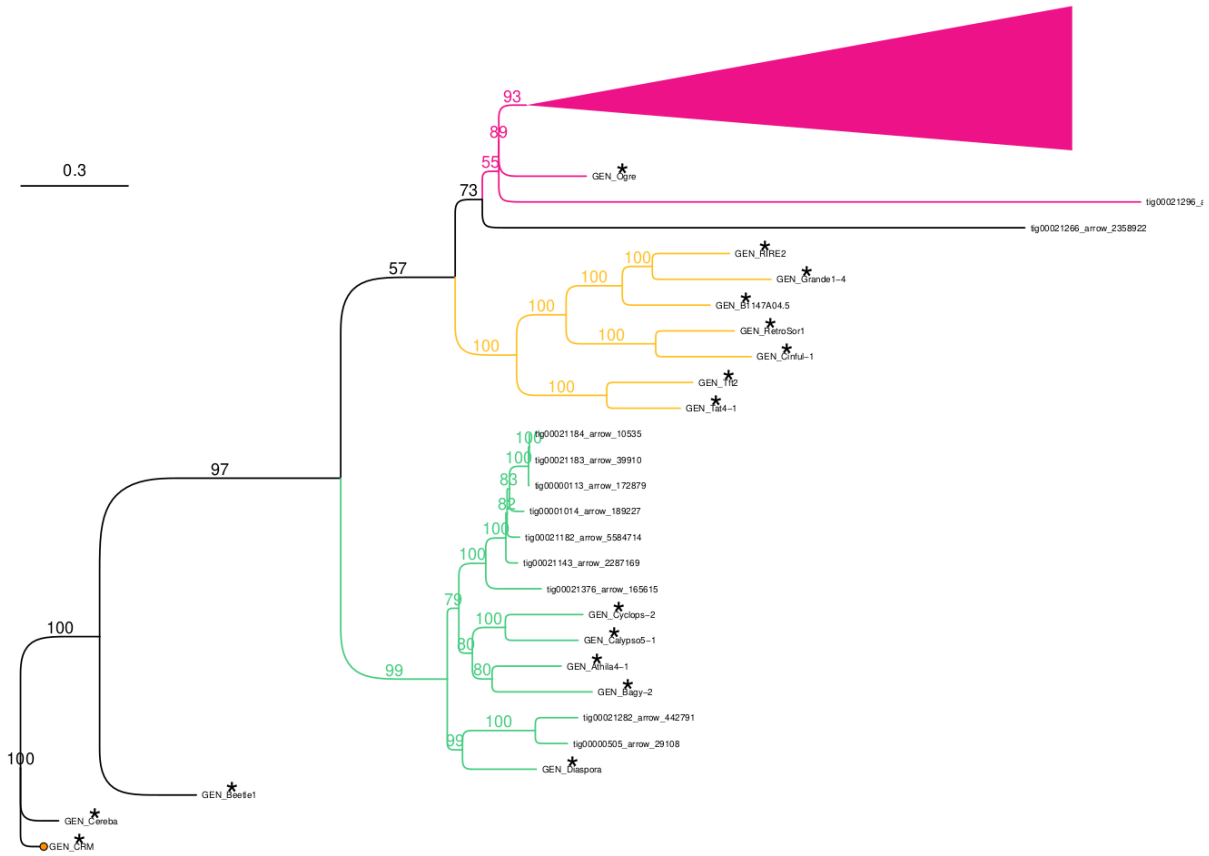

Figura 15. Árbol filogenético de los dominios concatenados en orden transcriptasa reversa (RT), ribonucleasa H (RNaseH) e integrasa (INT) utilizando todos los retrotransposones LTR (RT LTR) pertenecientes a los linajes Athila, Tat y Ogre de *Acca sellowiana*. La filogenia se construyó utilizando máxima verosimilitud a partir del alineamiento de las secuencias nucleotídicas de los dominios concatenados se recortaron con trimAl del superclado OTA en *Acca sellowiana*. En verde elementos Athila, en amarillo elementos Tat y en rosado elementos Ogre. El outgroup, RT LTR Ty3/Gypsy de chromoviruse CRM corresponde al círculo naranja. En cada elemento se encuentra el nombre del contig al que pertenecen, inicio y fin de su ubicación, aquellos marcados con asterisco corresponden a la Gypsy database 2.0 (GyDB).

En los análisis filogenéticos de las Fig. 14.1, 14.2 y 14.3, el dominio INT muestra 10 dominios agrupados con el linaje Athila y 114 dominios que se agruparon con el linaje Ogre. El dominio RT exhibe 9 dominios formando grupo con Athila y 115 dominios agrupados con Ogre. En cuanto al dominio RNaseH, se observan 9 dominios formando grupo con Athila, 1 dominio agrupado con el linaje Tat y 114 dominios agrupados con el linaje Ogre. En el caso de análisis de dominios concatenados (Fig. 15) , se observa que 9 dominios concatenados forman un grupo con el linaje Athila, mientras que 113 forman un grupo con el linaje Ogre. Además, 2 dominios concatenados no formaban ningún grupo, situándose cerca del clúster del linaje Ogre de referencia.

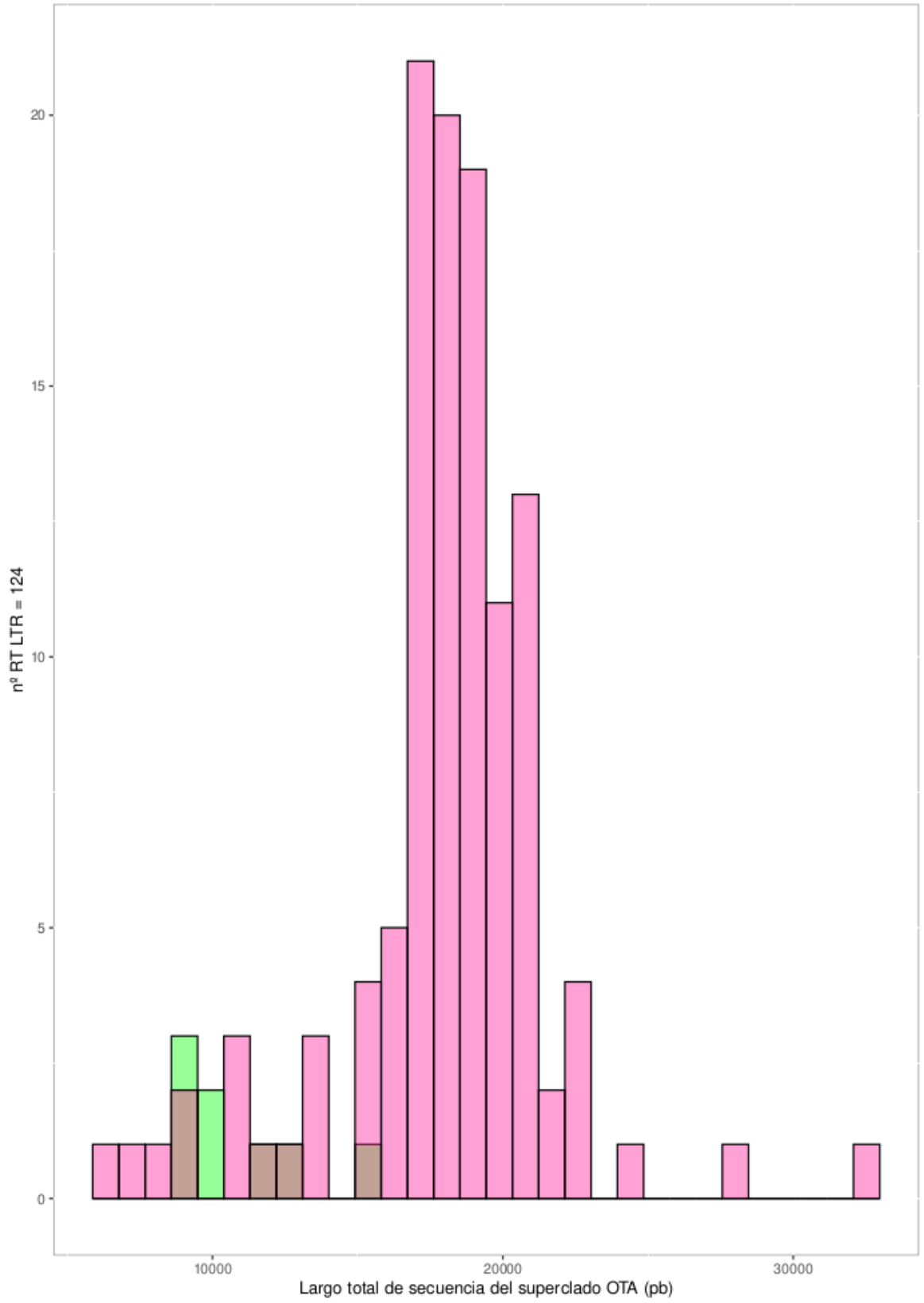

Figura 16. Largo de RT LTR completos del superclado OTA obtenido a partir del flujo de trabajo. En verde elementos Athila y en rosado elementos Ogre.

Se obtuvo la distribución del largo de los 124 elementos identificados (Fig. 16). Además se determinó si los elementos identificados se encontraban distribuidos en

todo el genoma o estaban en regiones particulares. Se identificó que los 124 están presentes en 108 contigs de 2075 que componen el genoma. 87 se encontraron en contigs de forma individual (1 por contig), estos contigs presentan una media de largo de 1,548,851 pb. Los demás elementos están presentes en más de una copia por contig. La distancia media entre distintos elementos en un mismo contig es de de 204,710 pb. Sin embargo, hay casos donde los elementos se encuentran agrupados. En especial, un contig de 941,164 pb contiene 4 elementos, a una distancia media de 10,790 pb.

Otro flujo de trabajo para abordar el problema

Se empleó una estrategia adicional para abordar el problema de identificar y analizar la distribución de los retrotransposones Athila, Tat y Ogre en *A. sellowiana*. Esta estrategia implicó realizar una búsqueda con RepeatMasker y luego procesar los resultados utilizando REannotate. Para este enfoque, se utilizó la misma base de datos, con RT del superclado OTA.

Los resultados obtenidos mediante el uso de RepeatMasker arrojaron un total de 2,339 identificaciones pertenecientes al superclado OTA. Se observó que muchos de estos resultados eran redundantes, variando ligeramente las posiciones de un elemento a otro en el genoma en unos pocos pb. Estos resultados fueron procesados con REannotate, y se redujeron a un total de 1,345 elementos. A pesar de la reducción en el número de elementos, los resultados obtenidos con REannotate mantienen cierta redundancia en la anotación.

Se obtuvo la distribución de longitudes para los 1345 elementos identificados mediante RepeatMasker/REannotate (RM/RE) (Fig. 18). Se puede observar que la mayoría de los elementos identificados con RM/RE (1288 elementos) son más cortos que lo esperado y solo 55 tienen una longitud mayor a 6 kb (mínimo para el superclado OTA). La clasificación se realiza solo por medio del algoritmo Smith-Waterman, lo que permite identificar que 51 corresponden al linaje Ogre y 4 al linaje Athila. Se encontraron además varios elementos largos que corresponden a elementos anidados dentro de otros elementos. En los 55 RT LTR analizados, se identificó la presencia de los dominios RT, RNaseH, INT, AP y GAG; no obstante, la mayoría de estos dominios se encontraban de manera fragmentada.

Se realizó una comparación mediante blast, de estos elementos mayores a 6 Kb con los elementos identificados en el flujo de trabajo, y sólo se pudieron identificar 9 elementos comunes, de los cuales 7 pertenecen al linaje Ogre y 2 al linaje Athila. En el anexo I se describen las características de los 124 elementos detectados a través

de búsqueda basada en homología y los 55 detectados por RepeatMasker y REannoate mediante de secuencias consenso y en modelos de Markov.

Los análisis filogenéticos de los dominios RT, RNaseH, INT y la concatenación de dominios de nuestro flujo de trabajo, junto con el flujo de trabajo RMRE, revelaron algunas secuencias que no se agruparon con los clústeres principales (material suplementario Fig. S1.1, S1.2, S1.3 y S2). En el caso del dominio RT, tres secuencias obtenidas de RMRE no se agruparon con ninguno de los clústeres. Similarmente, en el dominio INT se observó una situación similar. En cuanto al dominio RNaseH, tres secuencias de nuestro flujo de trabajo y tres secuencias del flujo de trabajo RMRE no se agruparon con otros clústeres, pero sí se agruparon entre sí para formar un grupo más pequeño. Finalmente, el análisis filogenético de los dominios concatenados proporcionó un resumen representativo, donde dos secuencias concatenadas del flujo de trabajo RMRE y una de nuestro flujo de trabajo no se agruparon con ninguno de los clústeres de linajes de la base de datos GyDB.

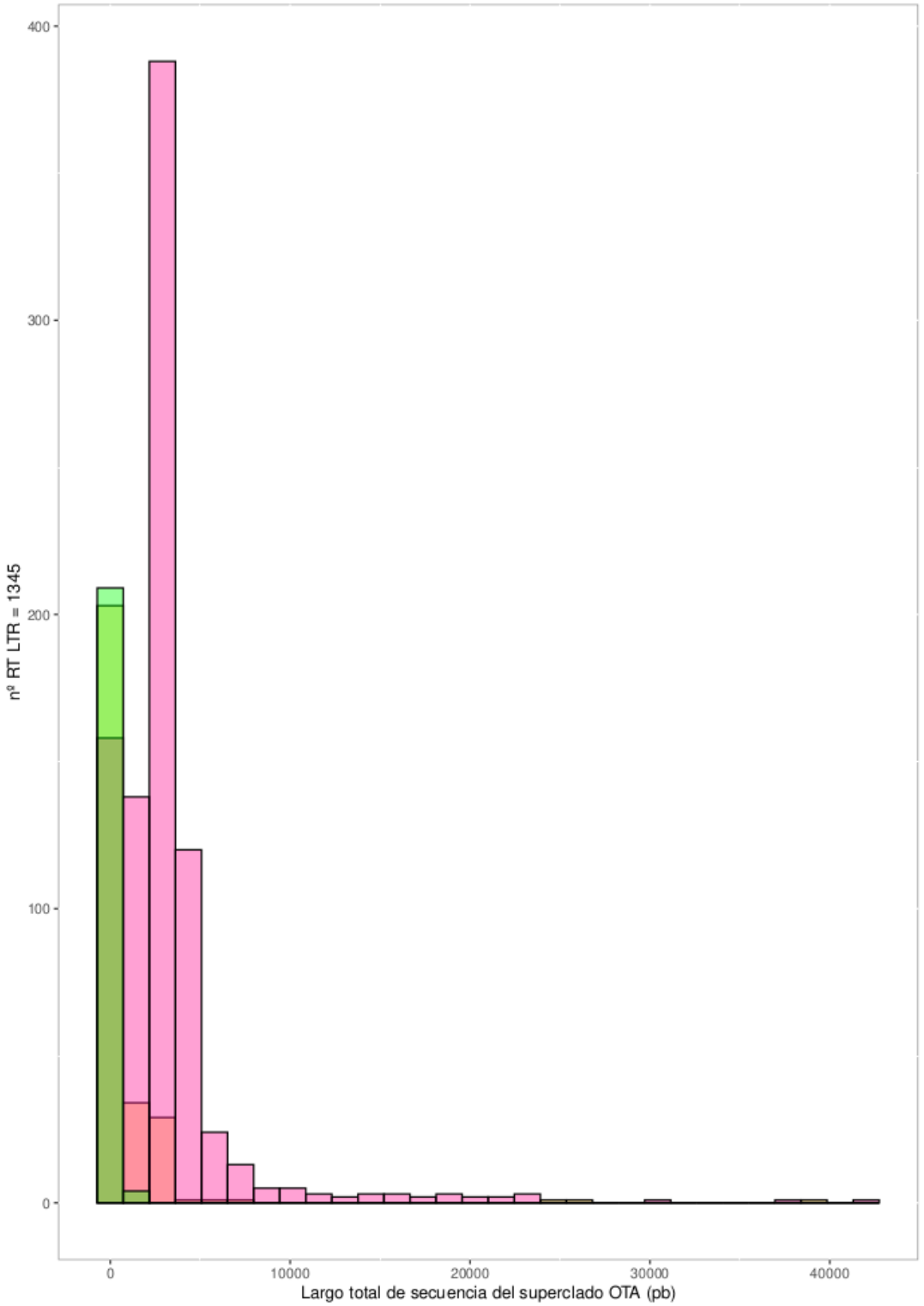

Figura 18. Largo de RT LTR completos y fragmentados del superclado OTA obtenido a partir de RepeatMasker/REannotate. En verde elementos Athila, en amarillo Tat y en rosado elementos Ogre.

DISCUSIÓN

Flujo de trabajo

En el presente trabajo se aplicaron distintas aproximaciones metodológicas heurísticas para la identificación y el análisis del repertorio de retrotransposones tipo LTR (RT LTR) de la familia Ty3/Gypsy, linajes Athila/Tat y Ogre en el genoma de *Acca sellowiana,* el guayabo.

Existen dos enfoques comunes para buscar y caracterizar RT LTR. Una de estas aproximaciones se fundamenta en la comparación de secuencias de ADN y dominios proteicos característicos de estos ET y la otra se basa en la identificación de elementos mediante búsqueda de *novo*. El primero, requiere dominios de RT LTR conocidos los cuales serán utilizados para identificar secuencias similares en el genoma o secuencias que se quieran interrogar. Las secuencias y dominios de RT LTR conocidos se obtienen de diferentes bases de datos. Este enfoque es útil para identificar RT LTR que presentan similitud con secuencias conocidas, obteniendo una identificación de forma precisa en una secuencia problema y además, permite la identificación de elementos completos (Kennedy et al., 2011). La búsqueda de similitud es un método computacionalmente rápido que puede realizarse en grandes bases de datos de genomas, y se caracteriza por ser una búsqueda eficiente y robusta. Como desventaja puede pasar por alto RT LTR que tienen baja similitud con secuencias conocidas de RT LTR, por lo que el método tiene una sensibilidad limitada a secuencias no conocidas (Kennedy et al., 2011). Por su parte, el método de búsqueda de *novo* se basa en la identificación de patrones comunes y repetidos en el genoma o basado en estructura. A través de este enfoque, se busca la identificación de retrotransposones. Algunos de los patrones de búsqueda que se utilizan son repeticiones terminales largas, la presencia de dominios y TSD. Este enfoque puede identificar RT LTR no relacionados con secuencias conocidas de RT LTR y es capaz de identificar retrotransposones no conocidos.

En el marco de esta investigación se desarrolló un flujo de trabajo basado en la similitud de secuencia. Se utilizó un conjunto de herramientas bioinformáticas para identificar secuencias candidatas de RT LTR mediante la comparación de secuencias con una base de datos de referencia de RT LTR. El proceso implica la identificación de regiones con una alta similitud de secuencia con los elementos de la base de datos de referencia. Para la clasificación de los elementos identificados se utilizó la comparación filogenética con secuencias de referencia. Además se identificó en cada elemento los distintos dominios y su posición absoluta y relativa, permitiendo clasificar adecuadamente los elementos.

Selección de base de datos

Dado el impacto que puede tener la elección de una base de datos en la calidad de los resultados de búsqueda de ET se evaluaron dos bases de datos de RT LTR para ser usadas como base en la búsqueda y anotación de RT LTR en Guayabo. La elección de la base de datos puede llevar a la identificación de diferentes ET, lo que resulta en una mayor variabilidad y cantidad de resultados, ya sea en forma de aumentos o disminuciones.

Al comparar los resultados de la búsqueda de RT, se observó que GyDB proporcionó una mayor cantidad y variabilidad de resultados siendo ligeramente superior a REXdb (Fig. 8). A pesar de esta ligera diferencia entre GyDB y REXdb, es probable que conduzcan a resultados finales similares. Sin embargo, es importante destacar que la búsqueda de ambas bases de datos concatenadas no arrojó los mismos resultados, siendo probable que conduzca a resultados finales diferentes. Por lo tanto se utilizó GyDB para el desarrollo del flujo de trabajo. Cabe mencionar que existen otras bases de datos de RT LTR que no fueron utilizadas en este estudio. En particular PlantLTRdb sólo contiene elementos Athila, y no Tat u Ogre y por lo tanto resultaba menos abarcativa, Por su parte Repbase no posee las estructuras internas de los elementos, lo que se consideró necesario para identificar y clasificar los elementos (Zhou et al., 2021).

Recopilación de datos preliminares para el flujo de trabajo

Durante el desarrollo del flujo de trabajo, se utilizaron los datos de GyDB de RT LTR de los linajes Athila, Tat y Ogre. Para asegurar la precisión del análisis, se calcularon y utilizaron los valores mínimos y la media de la longitud total, la región LTR y el dominio RT del superclado OTA. En la búsqueda de los linajes de Ogre, Athila y Tat, se emplearon los parámetros de longitud de secuencia como filtro de búsqueda. Es importante tener en cuenta que estos parámetros son específicos de GyDB y podrían no ser totalmente representativos de estos linajes. La presencia de los linajes Athila, Tat y Ogre puede afectar la eficacia de su búsqueda debido a sus notables diferencias en cuanto a longitud. Esto podría conducir a una posible disminución en la eficiencia del proceso de búsqueda (Macas & Neumann, 2007; Pelissier et al., 1995). Ademas, se pueden ajustar otros parámetros de búsqueda, como la identidad, o modificar globalmente el valor de e-value para explorar cómo estos afectan la restricción o flexibilidad de la búsqueda.

En la base de datos, se encontraron diferencias en la longitud entre los linajes. Mientras que la longitud de Ogre es de 20 kb, Athila y Tat tienen alrededor de 13 kb. Como era esperable, los resultados obtenidos en este trabajo poseen una relación

estrecha entre la longitud de los ET y las referencias utilizadas. Específicamente, la longitud promedio estimada de los ET del supuerclado OTA en *A. sellowiana* es de 20 kb, coincidiendo con la longitud de los elementos Ogre y Athila en nuestra base de datos. Asimismo, se observa una cantidad considerable de ET con longitudes superiores a los 25 kb, reportado para el linaje Ogre en *Pisum sativum y Vicia pannonica* (Macas & Neumann, 2007).

La longitud de la región LTR también es variable y dependiente del linaje. De acuerdo a la base GyDB, los LTR en el linaje Orge son mayores a 5 kb mientras que para Tat y Athila tienen una media de 1.3 kb en GyDB. Los elementos superclado OTA encontrados *A. sellowiana* presentan LTR cercanos a los 4 kb y sutilmente diferentes para los extremos 5' y 3' estos resultados son consistentes con los que se reportan en trabajos previos del superclado OTA (Macas & Neumann, 2007; Marín & Lloréns, 2000).

Dominio RT

La retrotranscriptasa (RT) o transcriptasa reversa es una enzima crucial en el ciclo de replicación de los RT LTR gypsy y otros retrovirus. Esta enzima tiene la capacidad de sintetizar una copia de ADN a partir del ARN (Hughes, 2015). Debido al papel central que desempeña el dominio RT en los RT LTR, su secuencia es altamente conservada (Eickbush & Jamburuthugoda, 2008). Esto ha permitido su utilización en investigación de eventos evolutivos y en la caracterización y clasificación de diversidad de estos elementos en diferentes organismos (Eickbush & Jamburuthugoda, 2008; Wang et al., 2021). Por ejemplo, el análisis del dominio RT ha demostrado que los RT LTR gypsy están presentes en una amplia gama de organismos, incluyendo plantas, animales y hongos (Llorens et al., 2009; Neumann et al., 2019; Wang et al., 2021).

Por lo tanto, el análisis del dominio RT es una herramienta clave en la detección de RT LTR en esta estrategia. Usando solo un filtro de similitud, un E-value de 1e-5, en primera instancia se identificaron 432 secuencias de *A. sellowiana* con dominios RT putativo de acuerdo a las base de datos las RT de GyDB en la cual se integran secuencias de especies vegetales muy divergentes en comparación con *A. sellowiana*. Esto sin duda muestra la alta conservación de este dominio en los RT LTR. El largo promedio de las secuencias RT identificadas, aún siendo sutilmente menor, se corresponde al largo de las RT de la base de datos utilizada, al igual que su contenido de %GC.

La visualización de los RT LTR con sus secuencias completas es una manera de verificar los resultados obtenidos. Durante este trabajo se analizaron visualmente los elementos encontrados, identificando además la ubicación y el orden de cada uno de los dominios. El orden de los distintos dominios es crucial para separar las

distintas familias de RT LTR y funciona como un filtro para su clasificación (Llorens et al., 2009; Neumann et al., 2019). De forma esperable, al aplicar un filtro basado en el orden específico de los elementos del superclado OTA, se observó una reducción considerable en el número de ET anotados en esta familia, pasando de un total inicial a 289 a solo 103 elementos. Aplicando filtros adicionales aparte del mencionado, entre ellos se encuentra el filtro basado en la ubicación y la longitud mínima. Estas dos características específicas son las que generan la mayor reducción en el conjunto de datos.

Búsqueda recursiva de RT LTR

Los filtros y los requisitos utilizados en la búsqueda inicial de elementos resultan en la identificación certera de RT LTR del superclado OTA en *A. sellowiana*. Sin embargo, al ser muy exigente, la misma puede excluir RT LTR presentes en el Guayabo pero más variables a los presentes en las bases de datos. Por lo tanto, la búsqueda recursiva permite capturar nuevos RT LTR de los linajes Athila, Tat y Ogre que presenten similitud a los ya identificados en la misma especie. De esta forma se pudo ampliar la identificación de RT LTR del superclado OTA, de 103 identificados en primera instancia, a un total de 124. En esta etapa, se vuelven a utilizar los mismos valores mínimos de longitud total de secuencia del superclado OTA y la región LTR. Sin embargo, se podría realizar un estudio más asertivo utilizando bases de datos adicionales o basándose en investigaciones previamente publicadas para las especies relacionadas de la familia mirtacea.

Clasificación de Ty3/Gypsy basado en filogenias

Los dominios INT, RNaseH y RT son secuencias cruciales para la actividad de los RT LTR, además de presentar una alta conservación (Eickbush & Jamburuthugoda, 2008; Hughes, 2015; Malik & Eickbush, 1999, 2001). Estas secuencias se utilizan ampliamente para identificar nuevos RT LTR y clasificar linajes de RT LTR (Neumann, 2019; Malik & Eickbush, 2001). Las filogenias inferidas con estos dominios, utilizados de forma independiente, muestran resultados similares y con altos valores de bootstrap. Sin embargo, existen algunas pequeñas variaciones que corresponden a nodos con bajos valores de bootstrap y que no contienen dominios en la base de datos, por lo que pueden considerarse como variaciones menores. Considerando que las filogenias inferidas de forma independiente no presentan grandes diferencias en las agrupaciones, estás representan la misma historia evolutiva, por lo que podemos asumir que los dominios enzimáticos no han sufrido grandes intercambios a lo largo de la evolución (Malik & Eickbush, 1999).

Además, debido a las características de cada dominio, su análisis por separado puede proporcionar información específica. El dominio INT, altamente conservado dentro de un linaje, puede ser divergente entre linajes diferentes (Neumann et al., 2019). La mayoría de las secuencias INT identificadas en *A. sellowiana* se agrupan con las INT de los linajes Athila y Ogre, mientras que ninguna INT se agruparon con el linaje Tat (Fig. 14.1). También se utilizó el dominio RNaseH que es conservado entre los elementos LTR y no LTR y por lo tanto también se ha utilizado para identificar linajes (Malik & Eickbush, 2001). En la filogenia, todas las RNaseH de *A. sellowiana* se agrupan mayoritariamente con los linajes Athila y Ogre, destacando su presencia principalmente en este último (Fig. 14.3). En el dominio RT se distribuye de forma similar a la filogenia de las RNaseH (Fig. 14.2).

En este trabajo se ha observado que el gen GAG en *Acca sellowiana* del superclado OTA presenta una alta variabilidad en su longitud, con tamaños de 500, 1000 y 1800 pb. Esta variabilidad en la longitud del gen GAG refleja, en parte, la variabilidad de sus secuencias. De hecho el dominio GAG y AP han sido identificados como regiones heterogéneas (de Marco et al., 2010; Neumann et al., 2019). Esta heterogeneidad dificulta la identificación de patrones de homología entre las secuencias, además pueden generar distorsiones en los resultados de la inferencia filogenética ya que podrían causar alineamientos incorrectos y la aparición de artefactos como atracción de ramas largas (Bergsten, 2005). Por lo tanto, la exclusión de estas secuencias se consideró necesaria para garantizar la precisión de los resultados de la inferencia filogenética. Esta estrategia de exclusión también fue utilizada por Neumann y colaboradores, quienes además encontraron mutaciones en el codón de terminación y cambios de marco en la región codificante, pudiendo afectar la función de los elementos transponibles en cuestión (Neumann et al., 2019).

Existe una limitación al generar filogenias que contienen diferentes dominios, incluso cuando estos son conservados. Esta limitación se debe a que los resultados obtenidos en cada filogenia con cada dominio genera filogenias ligeramente diferentes (Fig. 14.1, 14.2 y 14.3). Esta ligera diferencia en los resultados puede deberse a factores tales como presión selectiva diferente en cada dominio. Para superar estas limitaciones, se ha propuesto una estrategia de análisis filogenético basada en la comparación de múltiples genes o secuencias concatenadas (Llorens et al., 2009; Neumann et al., 2019; Wicker et al., 2007). Esta estrategia ha permitido lograr una resolución y precisión mayor en la clasificación de linajes, como se evidencia al comparar las figuras 14.1, 14.2 y 14.3 con respecto a la figura 15.

En las cuatro filogenias inferidas de forma independiente (Fig. 14.1, 14.2, 14.3 y 15) para los dominios, así como en la filogenia que combina los dominios concatenados, se destaca la presencia significativa de dominios que se agrupan con el linaje Ogre. Esto sugiere que estos elementos transponibles en la evolución de *A. sellowiana* podrían haber experimentado eventos de duplicación o olas de expansión (El

Baidouri & Panaud, 2013; Ouadi et al., 2022). Esta gran cantidad de elementos del linaje Ogre (113), con relación estrecha entre sí puede indicar una mayor actividad transposicional, de manera reciente (El Baidouri & Panaud, 2013). Además, se identificó un número menor (9) dominios concatenados pertenecientes a Athila. Esto sugiere que Athila está experimentando actualmente un proceso de eliminación (El Baidouri & Panaud, 2013).

En la filogenia con los dominios concatenados, y en aquellas basadas en dominios RT e INT, no se observó que ningún RT LTR Ty3/Gypsy de *A. sellowiana* se agrupara con el linaje Tat. Sin embargo, cuando la filogenia se realizó con el dominio RNaseH, el contig 21266 se agrupó con Tat, siendo este entonces el único elemento Tat detectado en *A. sellowiana*. La inconsistencia y frecuente no detección de Tat podría estar explicado por limitaciones en la estrategia utilizada, o bien a la ausencia de elementos Tat. En el caso mencionado por último puede deberse a una eliminación total o casi total del linaje debido a eventos de recombinación o deleciones (Ma et al., 2019). El linaje Tat se distingue por la presencia de dos dominios de RNaseH. Uno de estos dominios está presente en todos los RT LTR, y su posición varía según la familia. Sin embargo, hay otro dominio de RNaseH adicional que caracteriza al linaje Tat. Este dominio adicional es conocido como ribonucleasa H arqueal (RNaseHa) y se encuentra en el gen Pol en tres posibles ubicaciones distintas. El dominio de RNaseHa se ha atribuido un origen polifilético, ya que se ha adquirido de forma independiente al menos en tres ocasiones a lo largo de la historia evolutiva del linaje Tat. El núcleo catalítico de este dominio ha experimentado cambios que han afectado su capacidad para llevar a cabo su actividad catalítica (Ustyantsev et al., 2015). Para obtener una clasificación más acertada de este elemento el cual no se agrupa con ningún elemento de la base de datos se puede llevar a cabo una búsqueda del dominio RNaseHa y verificar la posición.

Otro flujo de trabajo para abordar el problema

La metodología de RM/RE presenta algunas limitaciones en su uso con el genoma de *A. sellowiana*. Por un lado, tiende a generar elementos redundantes y no logra clasificar adecuadamente los elementos identificados. En este sentido, no es capaz de distinguir si un elemento es un RT LTR completo o fragmentado, ni puede identificar específicamente si se trata de un RT LTR. En particular, los elementos identificados mediante este flujo de trabajo como OTA corresponden a secuencias más cortas de que la que presenta este tipo de elementos (Macas & Neumann, 2007; Marín & Lloréns, 2000; Pelissier et al., 1995). Solo 55 elementos tienen el largo esperado. Se confirmó la presencia de los dominios en estos 55 elementos. No obstante, se encontraron altamente fragmentados y mostraron diferencias significativas respecto a la longitud de los alineamientos obtenidos a través de

nuestro flujo de trabajo. La presencia de dominios funcionales en los retrotransposones LTR es crucial para su actividad y capacidad de movilización en el genoma.

Resulta notable que de estos 55 elementos con largo esperado, sólo 9 son comunes a los encontrados con nuestro flujo de trabajo, siendo 7 Ogre y 2 Athila. Cabe resaltar que algunos de los elementos identificados por ambos flujos de trabajo se encuentran anidados, y su comparación mediante blast para identificar uno a uno es compleja. La anidación de elementos es un proceso observado en los RT LTR que muchas veces suelen retrotransponerse en los mismos sitios (Bennetzen & Wang, 2014).

A pesar de esto, los análisis filogenéticos, tanto por dominios individuales como por dominios concatenados, realizados en nuestro flujo de trabajo y en RMRE, han clasificado a la totalidad de los elementos del superclado OTA, con tan solo 3 excepciones. En concreto, de los 168 elementos del superclado OTA, nuestro flujo de trabajo identificó 124, mientras que RMRE detectó 55. Únicamente hubo 9 elementos que se encontraron de manera coincidente en ambas estrategias. Las diferencias observadas pueden atribuirse a varios factores, destacando especialmente dos de ellos. La diferencia en los resultados puede atribuirse a dos estrategias diferentes. En nuestra estrategia, buscamos el dominio RT y otros dominios en los RT LTR. En contraste, la estrategia de RepeatMasker consiste en buscar una semilla y expandir hacia derecha e izquierda, sin importar cuál sea la semilla encontrada. Otro factor que podría estar influyendo en los resultados son las posibles variaciones en la implementación del algoritmo Smith-Waterman entre BLAST y RepeatMasker (Altschul, 2005; Tarailo-Graovac & Chen, 2009).

CONCLUSIONES

El desarrollo de un programa personalizado en R, bash, perl, python y programas bioinformáticos para buscar y caracterizar RT LTR es una estrategia valiosa para la investigación de la estructura y evolución del genoma de las plantas. La estrategia se centró en la búsqueda de RT LTR mediante la comparación de secuencias basadas en la homología de dominios e inferencias filogenéticas.

Se ha desarrollado un flujo de trabajo *in silico* para la detección y clasificación de RT LTR, el cual puede ser aplicable a cualquier especie, variando la base de datos utilizada. A diferencia de otros enfoques, la innovación de este flujo de trabajo radica en la búsqueda de homología con diferentes dominios proteicos del gen pol. Además, se llevó a cabo la búsqueda del gen gag y de las regiones LTR para complementar esta estrategia. Sin embargo, se encontraron limitaciones para su fácil aplicación debido a que presenta numerosas tareas que requieren supervisión y ajustes manuales para optimizar la detección de los RT LTR orientado a búsqueda de linaje.

Se detectó la presencia de elementos de los linajes Athila y Ogre en *Acca sellowiana* mediante una búsqueda basada en la similitud de secuencia. Se encontraron un total de 124 RT LTR del superclado OTA. Todos los elementos identificados fueron completos, conteniendo los dominios fundamentales necesarios para la actividad de los RT LTR, incluyendo los dominios repetidos LTR, AP, RT, RNaseH e INT.

El flujo de trabajo utilizado es sensible a la divergencia entre los elementos de la base de datos y el genoma en dónde se realiza la búsqueda, por tanto la búsqueda recursiva utilizando elementos propios del genoma mejora la detección.

El flujo de trabajo desarrollado que incluye una búsqueda por homología de dominios con inferencia filogenética, es una mejor estrategia para la identificación de elementos completos que herramientas automáticas como RM/RE basada en homología de secuencias. Encontrando más del doble de elementos, en este caso en particular.

Del análisis de los patrones de integración los RT LTR Ty3/Gypsy no surgen evidencias que apoyen la ocurrencia de inserciones aleatorias; por el contrario, fueron observados sitios preferenciales de inserción. Es necesario un análisis más detallado del entorno de inserción para identificar si hay dominios o secuencias de inserción específicas.

REFERENCIA BIBLIOGRÁFICA

Allshire, R. C., & Madhani, H. D. (2018). Ten principles of heterochromatin formation and function. *Nature Reviews Molecular Cell Biology*, *19*(4), 229-244. https://doi.org/10.1038/nrm.2017.119

Altschul, S. F. (2005). BLAST Algorithm. En John Wiley & Sons, Ltd (Ed.), *ELS* (1.ª ed.). Wiley. https://doi.org/10.1038/npg.els.0005253

Bennetzen, J. L., & Wang, H. (2014). The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology*, *65*(1), 505-530. https://doi.org/10.1146/annurev-arplant-050213-035811

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*(2), 163-193. https://doi.org/10.1111/j.1096-0031.2005.00059.x

Bonchev, G., & Parisod, C. (2013). Transposable elements and microevolutionary changes in natural populations. *Molecular Ecology Resources*, *13*(5), 765-775. https://doi.org/10.1111/1755-0998.12133

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

Capy, P. (2005). Classification and nomenclature of retrotransposable elements. *Cytogenetic and Genome Research*, *110*(1-4), 457-461. https://doi.org/10.1159/000084978

Cintra, L. A., Souza, T. B. de, Parteka, L. M., Barreto, L. M., Pereira, L. F. P., Gaeta, M. L., Guyot, R., & Vanzela, A. L. L. (2021). An 82 bp tandem repeat family typical of 3' non-coding end of Gypsy/TAT LTR retrotransposons is conserved in Coffea spp. Pericentromeres. *Genome*, *65*(3), 137-151. https://doi.org/10.1139/gen-2021-0045

Corradi, N., Pombert, J.-F., Farinelli, L., Didier, E. S., & Keeling, P. J. (2010). The complete sequence of the smallest known nuclear genome from the microsporidian

Encephalitozoon intestinalis. *Nature Communications*, *1*(1), Article 1. https://doi.org/10.1038/ncomms1082

da Costa, I. R., Dornelas, M. C., & Forni-Martins, E. R. (2008). Nuclear genome size variation in fleshy-fruited Neotropical Myrtaceae. *Plant Systematics and Evolution*, *276*(3-4), 209-217. https://doi.org/10.1007/s00606-008-0088-x

de Marco, A., Davey, N. E., Ulbrich, P., Phillips, J. M., Lux, V., Riches, J. D., Fuzik, T., Ruml, T., Kräusslich, H.-G., Vogt, V. M., & Briggs, J. A. G. (2010). Conserved and Variable Features of Gag Structure and Arrangement in Immature Retrovirus Particles. *Journal of Virology*, *84*(22), 11729-11736. https://doi.org/10.1128/JVI.01423-10

de Setta, N., Monteiro-Vitorello, C. B., Metcalfe, C. J., Cruz, G. M. Q., Del Bem, L. E., Vicentini, R., Nogueira, F. T. S., Campos, R. A., Nunes, S. L., Turrini, P. C. G., Vieira, A. P., Ochoa Cruz, E. A., Corrêa, T. C. S., Hotta, C. T., de Mello Varani, A., Vautrin, S., da Trindade, A. S., de Mendonça Vilela, M., Lembke, C. G., … Van Sluys, M.-A. (2014). Building the sugarcane genome for biotechnology and identifying evolutionary trends. *BMC Genomics*, *15*(1), 540. https://doi.org/10.1186/1471-2164-15-540

DeBarry, J. D., & Kissinger, J. C. (2011). Jumbled Genomes: Missing Apicomplexan Synteny. *Molecular Biology and Evolution*, *28*(10), 2855-2871. https://doi.org/10.1093/molbev/msr103

Eickbush, T. H., & Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Research*, *134*(1), 221-234. https://doi.org/10.1016/j.virusres.2007.12.010

El Baidouri, M., & Panaud, O. (2013). Comparative Genomic Paleontology across Plant Kingdom Reveals the Dynamics of TE-Driven Genome Evolution. *Genome Biology and Evolution*, *5*(5), 954-965. https://doi.org/10.1093/gbe/evt025

Elliott, T. A., & Gregory, T. R. (2015). Do larger genomes contain more diverse transposable elements? *BMC Evolutionary Biology*, *15*(1), 69. https://doi.org/10.1186/s12862-015-0339-8

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: Where genetics meets genomics. *Nature Reviews Genetics*, *3*(5), 329-341. https://doi.org/10.1038/nrg793

Filée, J., Siguier, P., & Chandler, M. (2007). Insertion Sequence Diversity in Archaea. *Microbiology and Molecular Biology Reviews*, *71*(1), 121-157. https://doi.org/10.1128/MMBR.00031-06

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, *5*, 103-107. https://doi.org/10.1016/0168-9525(89)90039-5

Finnegan, D. J. (2012). Retrotransposons. *Current Biology*, *22*(11), R432-R437. https://doi.org/10.1016/j.cub.2012.04.025

Fukagawa, T., & Kakutani, T. (2023). Transgenerational epigenetic control of constitutive heterochromatin, transposons, and centromeres. *Current Opinion in Genetics & Development*, *78*, 102021. https://doi.org/10.1016/j.gde.2023.102021

Gao, D., Jiang, N., Wing, R. A., Jiang, J., & Jackson, S. A. (2015). Transposons play an important role in the evolution and diversification of centromeres among closely related species. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00216

Grewal, S. I. S., & Jia, S. (2007). Heterochromatin revisited. *Nature Reviews Genetics*, *8*(1), 35-46. https://doi.org/10.1038/nrg2008

Havecker, E. R., Gao, X., & Voytas, D. F. (2004). The diversity of LTR retrotransposons. *Genome Biology*.

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801-811. https://doi.org/10.1016/j.humimm.2021.02.012

Hughes, S. H. (2015). Reverse Transcription of Retroviruses and LTR Retrotransposons. *Microbiology spectrum*, *3*(2), 1-48. https://doi.org/10.1128/microbiolspec.MDNA3-0027-2014

Kennedy, R. C., Unger, M. F., Christley, S., Collins, F. H., & Madey, G. R. (2011). An automated homology-based approach for identifying transposable elements. *BMC*

Bioinformatics, *12*(1), 130. https://doi.org/10.1186/1471-2105-12-130

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49-63. https://doi.org/10.1023/A:1016072014259

Li, L., Maher, K., Navarre-Sitchler, A., Druhan, J., Meile, C., Lawrence, C., Moore, J., Perdrial, J., Sullivan, P., Thompson, A., Jin, L., Bolton, E. W., Brantley, S. L., Dietrich, W. E., Mayer, K. U., Steefel, C. I., Valocchi, A., Zachara, J., Kocar, B., … Beisman, J. (2017). Expanding the role of reactive transport models in critical zone processes. *Earth-Science Reviews*, *165*, 280-301. https://doi.org/10.1016/j.earscirev.2016.09.001

Lippman, Z., Gendrel, A.-V., Black, M., Vaughn, M. W., Dedhia, N., Richard McCombie, W., Lavine, K., Mittal, V., May, B., Kasschau, K. D., Carrington, J. C., Doerge, R. W., Colot, V., & Martienssen, R. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature*, *430*(6998), 471-476. https://doi.org/10.1038/nature02651

Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics*, *14*(1), 49-61. https://doi.org/10.1038/nrg3374

Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H., & Moya, A. (2009). Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biology Direct*, *4*(1), 41. https://doi.org/10.1186/1745-6150-4-41

Ma, B., Kuang, L., Xin, Y., & He, N. (2019). New Insights into Long Terminal Repeat Retrotransposons in Mulberry Species. *Genes*, *10*(4), 285. https://doi.org/10.3390/genes10040285

Ma, Devos, K. M., & Bennetzen, J. L. (2004). Analyses of LTR-Retrotransposon Structures Reveal Recent and Rapid Genomic DNA Loss in Rice. *Genome Research*, *14*(5), 860-869. https://doi.org/10.1101/gr.1466204

Macas, J., & Neumann, P. (2007). Ogre elements—A distinct group of plant Ty3/gypsy-like retrotransposons. *Gene*, *390*(1-2), 108-116. https://doi.org/10.1016/j.gene.2006.08.007

Malik, H. S., & Eickbush, T. H. (1999). Modular Evolution of the Integrase Domain in the Ty3/Gypsy Class of LTR Retrotransposons. *Journal of Virology*, *73*(6), 5186-5190. https://doi.org/10.1128/jvi.73.6.5186-5190.1999

Malik, H. S., & Eickbush, T. H. (2001). Phylogenetic Analysis of Ribonuclease H Domains Suggests a Late, Chimeric Origin of LTR Retrotransposable Elements and Retroviruses. *Genome Research*, *11*(7), 1187-1197. https://doi.org/10.1101/gr.185101

Marín, I., & Lloréns, C. (2000). Ty3/Gypsy Retrotransposons: Description of New Arabidopsis thaliana Elements and Evolutionary Perspectives Derived from Comparative Genomic Data. *Molecular Biology and Evolution*, *17*(7), 1040-1049. https://doi.org/10.1093/oxfordjournals.molbev.a026385

Maumus, F., & Quesneville, H. (2016). Impact and insights from ancient repetitive elements in plant genomes. *Current Opinion in Plant Biology*, *30*, 41-46. https://doi.org/10.1016/j.pbi.2016.01.003

McClintock. (1948). Mutable Loci in Maize. *Carnegie Inst. Wash. Yearb.*, *47*, 155-169.

McClintock, B. (1956). Controlling Elements and the Gene. *Cold Spring Harbor Symposia on Quantitative Biology*, *21*(0), 197-216. https://doi.org/10.1101/SQB.1956.021.01.017

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D., Goodstein, D. M., Dubchak, I., Poliakov, A., Mizrachi, E., Kullan, A. R. K., Hussey, S. G., Pinard, D., van der Merwe, K., Singh, P., … Schmutz, J. (2014). The genome of Eucalyptus grandis. *Nature*, *510*(7505), Article 7505. https://doi.org/10.1038/nature13308

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Schmücker, A., Mandáková, T., Jamge, B., Lambing, C., Kuo, P., Yelina, N., Hartwick, N., Colt, K., Smith, L. M., Ton, J., Kakutani, T., Martienssen, R. A., Schneeberger, K., Lysak, M. A., … Henderson, I. R. (2021). The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science*, *374*(6569), eabi7489. https://doi.org/10.1126/science.abi7489

Neumann, P., Novák, P., Hoštáková, N., & Macas, J. (2019). Systematic survey of plant

LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA*, *10*(1), 1. https://doi.org/10.1186/s13100-018-0144-1

Neumann, P., Požárková, D., & Macas, J. (2003). Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. *Plant Molecular Biology*, *53*(3), 399-410. https://doi.org/10.1023/B:PLAN.0000006945.77043.ce

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268-274. https://doi.org/10.1093/molbev/msu300

Ouadi, S., Sierro, N., Goepfert, S., Bovet, L., Glauser, G., Vallat, A., Peitsch, M. C., Kessler, F., & Ivanov, N. V. (2022). The clove (Syzygium aromaticum) genome provides insights into the eugenol biosynthesis pathway. *Communications Biology*, *5*, 684. https://doi.org/10.1038/s42003-022-03618-z

Pelissier, T., Tutois, S., Deragon, J. M., Tourmente, S., Genestier, S., & Picard, G. (1995). Athila, a new retroelement from Arabidopsis thaliana. *Plant Molecular Biology*, *29*(3), 441-452. https://doi.org/10.1007/BF00020976

Pereira, V. (2004). Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. *Genome Biology*.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D. S., Jackson, S., Wing, R. A., & Panaud, O. (2006). Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis* , a wild relative of rice. *Genome Research*, *16*(10), 1262-1269. https://doi.org/10.1101/gr.5290206

Praça, M. M., Carvalho, C. R., & Novaes, C. R. D. B. (2009). Nuclear DNA content of three Eucalyptus species estimated by flow and image cytometry. *Australian Journal of Botany*, *57*(6), 524. https://doi.org/10.1071/BT09114

Pritsch C., Quezada M., Vignale B., Franco J. (2008). Estudio de la diversidad genética de una colección de *Acca sellowiana* Berg Burret con alto potencial agronómico mediante el uso de marcadores moleculares RAPD. En: IV Simposio Nacional do Morango III Encontro sobre pequenas frutas e frutas nativas do Mercosul. Pelotas, EMBRAPA Clima Temperado.- Brasil, 679, 16-35.

Puppo et al., M. P. (2009). *PROSPECCIÓN Y CARACTERIZACIÓN DE POBLACIONES SILVESTRES DE Acca sellowiana (Berg) Burret. (GUAYABO DEL PAÍS)*. 1-141.

Quezada, M., Amadeu, R. R., Vignale, B., Cabrera, D., Pritsch, C., & Garcia, A. A. F. (2022). Construction of a High-Density Genetic Map of Acca sellowiana (Berg.) Burret, an Outcrossing Species, Based on Two Connected Mapping Populations. *Frontiers in Plant Science*, *12*(626811), 1-15.

Quezada, M., Pastina, M. M., Ravest, G., Silva, P., Vignale, B., Cabrera, D., Hinrichsen, P., Garcia, A. A. F., & Pritsch, C. (2014). A first genetic map of Acca sellowiana based on ISSR, AFLP and SSR markers. *Scientia Horticulturae*, *169*, 138-146. https://doi.org/10.1016/j.scienta.2014.02.009

Ross, S., Pechi, E., Speroni, G., Vignale, B., Speranza, P., Castillo, A., & Cabrera, D. (2017). In vitro rooting of *Acca sellowiana* microshoots. *Acta Horticulturae*, *1155*, 537-542. https://doi.org/10.17660/ActaHortic.2017.1155.79

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., & Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nature Genetics*, *20*(1), 43-45. https://doi.org/10.1038/1695

Servant, G., & Deininger, P. L. (2016). Insertion of Retrotransposons at Chromosome Ends: Adaptive Response to Chromosome Maintenance. *Frontiers in Genetics*, *6*. https://doi.org/10.3389/fgene.2015.00358

Sharma, A., Wolfgruber, T. K., & Presting, G. G. (2013). Tandem repeats derived from centromeric retrotransposons. *BMC Genomics*, *14*(1), 142. https://doi.org/10.1186/1471-2164-14-142

Shimada, A., Cahn, J., Ernst, E., Lynn, J., Grimanelli, D., Henderson, I., Kakutani, T., &

Martienssen, R. A. (2023). Retrotransposon addiction promotes centromere function via epigenetically activated small RNAs. *bioRxiv*, 2023.08.02.551486. https://doi.org/10.1101/2023.08.02.551486

Slotkin, R. K. (2010). The epigenetic control of the Athila family of retrotransposons in Arabidopsis. *Epigenetics*, *5*(6), 483-490. https://doi.org/10.4161/epi.5.6.12119

Tarailo‐Graovac, M., & Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, *25*(1). https://doi.org/10.1002/0471250953.bi0410s25

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), Article 6814. https://doi.org/10.1038/35048692

Tsukahara, S., Kawabe, A., Kobayashi, A., Ito, T., Aizu, T., Shin-i, T., Toyoda, A., Fujiyama, A., Tarutani, Y., & Kakutani, T. (2012). Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata. *Genes & Development*, *26*(7), 705-713. https://doi.org/10.1101/gad.183871.111

Ustyantsev, K., Novikova, O., Blinov, A., & Smyshlyaev, G. (2015). Convergent Evolution of Ribonuclease H in LTR Retrotransposons and Retroviruses. *Molecular Biology and Evolution*, *32*(5), 1197-1207. https://doi.org/10.1093/molbev/msv008

Vergara, Z., Sequeira-Mendes, J., Morata, J., Peiró, R., Hénaff, E., Costas, C., Casacuberta, J. M., & Gutierrez, C. (2017). Retrotransposons are specified as DNA replication origins in the gene-poor regions of Arabidopsis heterochromatin. *Nucleic Acids Research*, *45*(14), 8358-8368. https://doi.org/10.1093/nar/gkx524

Vicient, C. M., & Casacuberta, J. M. (2017). Impact of transposable elements on polyploid plant genomes. *Annals of Botany*, *120*(2), 195-207. https://doi.org/10.1093/aob/mcx078

Wang, Q., Wang, Y., Wang, J., Gong, Z., & Han, G. (2021). Plants acquired a major retrotransposon horizontally from fungi during the conquest of land. *New Phytologist*, *232*(1), 11-16. https://doi.org/10.1111/nph.17568

Weber, B., & Schmidt, T. (2009). Nested Ty3-gypsy retrotransposons of a single Beta procumbens centromere contain a putative chromodomain. *Chromosome Research*, *17*(3), 379-396. https://doi.org/10.1007/s10577-009-9029-y

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

Wilhelm & Wilhelm, M. (2001). Reverse transcription of retroviruses and LTR retrotransposons. *Cellular and Molecular Life Sciences*, *58*, 1246-1262.

Zhang, L., Yan, L., Jiang, J., Wang, Y., Jiang, Y., Yan, T., & Cao, Y. (2014). The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast *Candida albicans*. *Virulence*, *5*(6), 655-664. https://doi.org/10.4161/viru.32180

Zhou, S.-S., Yan, X.-M., Zhang, K.-F., Liu, H., Xu, J., Nie, S., Jia, K.-H., Jiao, S.-Q., Zhao, W., Zhao, Y.-J., Porth, I., El Kassaby, Y. A., Wang, T., & Mao, J.-F. (2021). A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes. *Scientific Data*, *8*, 174. https://doi.org/10.1038/s41597-021-00968-x

MATERIAL SUPLEMENTARIO

Tabla S1. Total de 124 retrotransposones LTR del tipo Ty3/Gypsy, superclase OTA, como resultado de nuestro flujo de trabajo.

contig	start	end	type	length	GC	sequence
tig00000113_arrow	172879	182508	Athila	9630	34.7	tig00000113_arrow_172879_Athila
tig00021183_arrow	39910	49539	Athila	9630	33.87	tig00021183_arrow_39910_Athila
tig00021184_arrow	10535	20165	Athila	9631	33.88	tig00021184_arrow_10535_Athila
tig00001014_arrow	189227	198453	Athila	9227	34.59	tig00001014_arrow_189227_Athila
tig00021182_arrow	5584714	5594144	Athila	9431	34.96	tig00021182_arrow_5584714_Athila
tig00021143_arrow	2287169	2299617	Athila	12449	35.81	tig00021143_arrow_2287169_Athila
tig00021376_arrow	165615	177477	Athila	11863	38.96	tig00021376_arrow_165615_Athila
tig00000505_arrow	29108	44561	Athila	15454	38.39	tig00000505_arrow_29108_Athila
tig00021282_arrow	442791	451729	Athila	8939	38.84	tig00021282_arrow_442791_Athila
tig00000002_arrow	53867	71980	Ogre	18114	40.66	tig00000002_arrow_53867_Ogre
tig00000083_arrow	9838	18643	Ogre	8806	42.68	tig00000083_arrow_9838_Ogre
tig00021151_arrow	1218604	1209793	Ogre	8812	42.83	tig00021151_arrow_1218604_Ogre
tig00021261_arrow	2631734	2649720	Ogre	17987	40.1	tig00021261_arrow_2631734_Ogre
tig00001280_arrow	8205	25699	Ogre	17495	40.53	tig00001280_arrow_8205_Ogre
tig00001280_arrow	25975	8481	Ogre	17495	40.69	tig00001280_arrow_25975_Ogre
tig00021426_arrow	759148	776644	Ogre	17497	40.73	tig00021426_arrow_759148_Ogre
tig00021229_arrow	4218734	4236782	Ogre	18049	40.5	tig00021229_arrow_4218734_Ogre
tig00000294_arrow	6850037	6867337	Ogre	17301	41.59	tig00000294_arrow_6850037_Ogre
tig00062634_arrow	3320184	3337834	Ogre	17651	41.32	tig00062634_arrow_3320184_Ogre
tig00021151_arrow	517356	535417	Ogre	18062	42.61	tig00021151_arrow_517356_Ogre
tig00002294_arrow	16502	33759	Ogre	17258	41.74	tig00002294_arrow_16502_Ogre
tig00021339_arrow	4287344	4304661	Ogre	17318	41.8	tig00021339_arrow_4287344_Ogre
tig00062623_arrow	352850	369803	Ogre	16954	41.68	tig00062623_arrow_352850_Ogre
tig00021287_arrow	3284948	3266855	Ogre	18094	42.68	tig00021287_arrow_3284948_Ogre
tig00021287_arrow	158913	175573	Ogre	16661	41.32	tig00021287_arrow_158913_Ogre
tig00021342_arrow	1973097	1990395	Ogre	17299	42.01	tig00021342_arrow_1973097_Ogre
tig00021296_arrow	719561	731907	Ogre	12347	40.33	tig00021296_arrow_719561_Ogre
tig00000494_arrow	70291	86688	Ogre	16398	47.06	tig00000494_arrow_70291_Ogre
tig00021302_arrow	2435214	2451640	Ogre	16427	47	tig00021302_arrow_2435214_Ogre
tig00001594_arrow	12196	1015	Ogre	11182	47.4	tig00001594_arrow_12196_Ogre
tig00021294_arrow	470342	488994	Ogre	18653	48.25	tig00021294_arrow_470342_Ogre
tig00000644_arrow	10647	28783	Ogre	18137	45.21	tig00000644_arrow_10647_Ogre
tig00021213_arrow	2659366	2680224	Ogre	20859	45.26	tig00021213_arrow_2659366_Ogre
tig00021433_arrow	1057299	1077691	Ogre	20393	44.07	tig00021433_arrow_1057299_Ogre

tig00021385_arrow	552715	572568	Ogre	19854	42.51	tig00021385_arrow_552715_Ogre
tig00001400_arrow	61131	67772	Ogre	6642	38.35	tig00001400_arrow_61131_Ogre
tig00000634_arrow	32408	51360	Ogre	18953	41.41	tig00000634_arrow_32408_Ogre
tig00021179_arrow	658798	681657	Ogre	22860	42.03	tig00021179_arrow_658798_Ogre
tig00021392_arrow	1790214	1809613	Ogre	19400	41.68	tig00021392_arrow_1790214_Ogre
tig00062549_arrow	39526	59714	Ogre	20189	40.25	tig00062549_arrow_39526_Ogre
tig00001473_arrow	28165	49385	Ogre	21221	41.12	tig00001473_arrow_28165_Ogre
tig00021322_arrow	468768	487572	Ogre	18805	41.18	tig00021322_arrow_468768_Ogre
tig00001748_arrow	22313	39537	Ogre	17225	40.5	tig00001748_arrow_22313_Ogre
tig00001445_arrow	5866	27130	Ogre	21265	41.86	tig00001445_arrow_5866_Ogre
tig00021264_arrow	604135	624226	Ogre	20092	40.33	tig00021264_arrow_604135_Ogre
tig00001153_arrow	23229	43946	Ogre	20718	43.73	tig00001153_arrow_23229_Ogre
tig00021296_arrow	725224	746004	Ogre	20781	46.42	tig00021296_arrow_725224_Ogre
tig00021153_arrow	2492954	2513116	Ogre	20163	46.33	tig00021153_arrow_2492954_Ogre
tig00021205_arrow	2411341	2431273	Ogre	19933	45.76	tig00021205_arrow_2411341_Ogre
tig00062639_arrow	668688	689664	Ogre	20977	46.81	tig00062639_arrow_668688_Ogre
tig00000886_arrow	102670	125033	Ogre	22364	41.67	tig00000886_arrow_102670_Ogre
tig00021345_arrow	232097	254401	Ogre	22305	43.23	tig00021345_arrow_232097_Ogre
tig00021141_arrow	1248904	1269853	Ogre	20950	46.02	tig00021141_arrow_1248904_Ogre
tig00021330_arrow	4631890	4652096	Ogre	20207	42.98	tig00021330_arrow_4631890_Ogre
tig00000696_arrow	27075	47398	Ogre	20324	42.59	tig00000696_arrow_27075_Ogre
tig00021307_arrow	735740	754919	Ogre	19180	43.59	tig00021307_arrow_735740_Ogre
tig00062631_arrow	3078623	3097774	Ogre	19152	43.43	tig00062631_arrow_3078623_Ogre
tig00001490_arrow	52901	72863	Ogre	19963	42.29	tig00001490_arrow_52901_Ogre
tig00021175_arrow	3863375	3883704	Ogre	20330	42.5	tig00021175_arrow_3863375_Ogre
tig00021384_arrow	298710	321188	Ogre	22479	43.28	tig00021384_arrow_298710_Ogre
tig00021437_arrow	1453547	1473829	Ogre	20283	42.38	tig00021437_arrow_1453547_Ogre
tig00021318_arrow	1240561	1260958	Ogre	20398	42.95	tig00021318_arrow_1240561_Ogre
tig00001789_arrow	39203	58321	Ogre	19119	43.4	tig00001789_arrow_39203_Ogre
tig00021309_arrow	1042654	1062990	Ogre	20337	43.11	tig00021309_arrow_1042654_Ogre
tig00021315_arrow	69566	51023	Ogre	18544	43.04	tig00021315_arrow_69566_Ogre
tig00021208_arrow	8199132	8223367	Ogre	24236	42	tig00021208_arrow_8199132_Ogre
tig00021450_arrow	429799	450073	Ogre	20275	43.24	tig00021450_arrow_429799_Ogre
tig00021177_arrow	2293317	2305476	Ogre	12160	41.8	tig00021177_arrow_2293317_Ogre
tig00021300_arrow	197386	216557	Ogre	19172	43.59	tig00021300_arrow_197386_Ogre
tig00001357_arrow	100209	122082	Ogre	21874	40.9	tig00001357_arrow_100209_Ogre
tig00000131_arrow	71495	90485	Ogre	18991	41.49	tig00000131_arrow_71495_Ogre
tig00000289_arrow	39515	57691	Ogre	18177	41.88	tig00000289_arrow_39515_Ogre
tig00021151_arrow	759588	776313	Ogre	16726	41.89	tig00021151_arrow_759588_Ogre
tig00021441_arrow	129033	148118	Ogre	19086	41.85	tig00021441_arrow_129033_Ogre
tig00021316_arrow	171946	191115	Ogre	19170	41.17	tig00021316_arrow_171946_Ogre
tig00021443_arrow	482135	501393	Ogre	19259	41.33	tig00021443_arrow_482135_Ogre

tig00021420_arrow	1335586	1354399	Ogre	18814	41.86	tig00021420_arrow_1335586_Ogre
tig00000508_arrow	74190	93408	Ogre	19219	40.81	tig00000508_arrow_74190_Ogre
tig00021182_arrow	8271479	8284855	Ogre	13377	43.52	tig00021182_arrow_8271479_Ogre
tig00001606_arrow	32390	49704	Ogre	17315	40.28	tig00001606_arrow_32390_Ogre
tig00021220_arrow	1195138	1213786	Ogre	18649	41.83	tig00021220_arrow_1195138_Ogre
tig00021242_arrow	1008646	1026074	Ogre	17429	41.89	tig00021242_arrow_1008646_Ogre
tig00021262_arrow	232789	249722	Ogre	16934	43.04	tig00021262_arrow_232789_Ogre
tig00021399_arrow	124342	152539	Ogre	28198	42.88	tig00021399_arrow_124342_Ogre
tig00001234_arrow	53543	67210	Ogre	13668	43.46	tig00001234_arrow_53543_Ogre
tig00000768_arrow	31343	48987	Ogre	17645	41.46	tig00000768_arrow_31343_Ogre
tig00021236_arrow	654089	672643	Ogre	18555	41.45	tig00021236_arrow_654089_Ogre
tig00021223_arrow	2571631	2588154	Ogre	16524	42.53	tig00021223_arrow_2571631_Ogre
tig00021274_arrow	153006	172043	Ogre	19038	41.44	tig00021274_arrow_153006_Ogre
tig00021225_arrow	752533	769065	Ogre	16533	41.37	tig00021225_arrow_752533_Ogre
tig00021197_arrow	2728118	2746212	Ogre	18095	42.1	tig00021197_arrow_2728118_Ogre
tig00021141_arrow	1695705	1684789	Ogre	10917	44.24	tig00021141_arrow_1695705_Ogre
tig00001417_arrow	90450	108993	Ogre	18544	42.07	tig00001417_arrow_90450_Ogre
tig00021284_arrow	467957	486451	Ogre	18495	41.05	tig00021284_arrow_467957_Ogre
tig00021447_arrow	290451	308844	Ogre	18394	40.5	tig00021447_arrow_290451_Ogre
tig00021326_arrow	1318714	1335945	Ogre	17232	40.99	tig00021326_arrow_1318714_Ogre
tig00021298_arrow	18430	38250	Ogre	19821	39.3	tig00021298_arrow_18430_Ogre
tig00000294_arrow	7332919	7340531	Ogre	7613	40.47	tig00000294_arrow_7332919_Ogre
tig00021179_arrow	1188827	1175683	Ogre	13145	42.2	tig00021179_arrow_1188827_Ogre
tig00021197_arrow	1515548	1504701	Ogre	10848	42.73	tig00021197_arrow_1515548_Ogre
tig00062625_arrow	888540	906651	Ogre	18112	41.04	tig00062625_arrow_888540_Ogre
tig00000643_arrow	14842	32154	Ogre	17313	40.83	tig00000643_arrow_14842_Ogre
tig00000505_arrow	137984	122709	Ogre	15276	40.18	tig00000505_arrow_137984_Ogre
tig00000505_arrow	203665	187983	Ogre	15683	40.25	tig00000505_arrow_203665_Ogre
tig00021300_arrow	418371	434049	Ogre	15679	40.1	tig00021300_arrow_418371_Ogre
tig00021301_arrow	12193	27872	Ogre	15680	40.22	tig00021301_arrow_12193_Ogre
tig00001092_arrow	22931	41046	Ogre	18116	40.36	tig00001092_arrow_22931_Ogre
tig00021411_arrow	814139	796875	Ogre	17265	40.31	tig00021411_arrow_814139_Ogre
tig00001132_arrow	101454	121859	Ogre	20406	40.52	tig00001132_arrow_101454_Ogre
tig00021296_arrow	937698	917300	Ogre	20399	40.53	tig00021296_arrow_937698_Ogre
tig00021297_arrow	6686	24017	Ogre	17332	39.7	tig00021297_arrow_6686_Ogre
tig00021297_arrow	34819	14415	Ogre	20405	40.51	tig00021297_arrow_34819_Ogre
tig00021296_arrow	909567	926904	Ogre	17338	39.73	tig00021296_arrow_909567_Ogre
tig00001169_arrow	20918	39309	Ogre	18392	40.3	tig00001169_arrow_20918_Ogre
tig00000794_arrow	26875	44298	Ogre	17424	41.21	tig00000794_arrow_26875_Ogre
tig00021394_arrow	59241	77198	Ogre	17958	41.79	tig00021394_arrow_59241_Ogre
tig00002028_arrow	89765	122608	Ogre	32844	40.84	tig00002028_arrow_89765_Ogre
tig00001010_arrow	8438	25928	Ogre	17491	42.2	tig00001010_arrow_8438_Ogre

tig00021401_arrow	1744109	1762135	Ogre	18027	41.42	tig00021401_arrow_1744109_Ogre
tig00021455_arrow	30162	48180	Ogre	18019	41.41	tig00021455_arrow_30162_Ogre
tig00062629_arrow	342520	360015	Ogre	17496	42.17	tig00062629_arrow_342520_Ogre
tig00021340_arrow	452371	470293	Ogre	17923	42.14	tig00021340_arrow_452371_Ogre
tig00062623_arrow	831691	813916	Ogre	17776	41.16	tig00062623_arrow_831691_Ogre
tig00021266_arrow	2358922	2367107	Ogre?	8186	45.74	tig00021266_arrow_2358922_Ogre?

Tabla S2. Total de 55 retrotransposones LTR del tipo Ty3/Gypsy, superclase OTA, como resultado del flujo de trabajo RepeatMasker y REannotate.

contig	start	end	type	length	GC	sequence
tig00000122_arrow	3381368	3387729	Ogre	6362	38.98	tig00000122_arrow_3381368_Ogre
tig00000122_arrow	3685094	3698986	Ogre	13893	37.34	tig00000122_arrow_3685094_Ogre
tig00000122_arrow	3793698	3816124	Ogre	22427	41.86	tig00000122_arrow_3793698_Ogre
tig00000122_arrow	4752677	4762590	Ogre	9914	40.24	tig00000122_arrow_4752677_Ogre
tig00000122_arrow	4900523	4907403	Ogre	6881	43.02	tig00000122_arrow_4900523_Ogre
tig00000122_arrow	5185976	5228026	Ogre	42051	42.46	tig00000122_arrow_5185976_Ogre
tig00000512_arrow	49871	60145	Ogre	10275	43.87	tig00000512_arrow_49871_Ogre
tig00000888_arrow	159528	168123	Ogre	8596	43.22	tig00000888_arrow_159528_Ogre
tig00000888_arrow	184647	196107	Ogre	11461	40.31	tig00000888_arrow_184647_Ogre
tig00001764_arrow	34496	53842	Ogre	19347	41.94	tig00001764_arrow_34496_Ogre
tig00021141_arrow	202476	225004	Ogre	22529	41.08	tig00021141_arrow_202476_Ogre
tig00021141_arrow	559926	590737	Ogre	30812	40.09	tig00021141_arrow_559926_Ogre
tig00021143_arrow	2292583	2299214	Athila	6632	37.26	tig00021143_arrow_2292583_Athila
tig00021153_arrow	2741947	2748620	Ogre	6674	40.7	tig00021153_arrow_2741947_Ogre
tig00021153_arrow	3423832	3449837	Athila	26006	41.53	tig00021153_arrow_3423832_Athila
tig00021153_arrow	3696165	3708647	Ogre	12483	41.46	tig00021153_arrow_3696165_Ogre
tig00021153_arrow	3797446	3818885	Ogre	21440	40.33	tig00021153_arrow_3797446_Ogre
tig00021179_arrow	770010	776749	Ogre	6740	43.86	tig00021179_arrow_770010_Ogre
tig00021194_arrow	592182	599817	Ogre	7636	40.66	tig00021194_arrow_592182_Ogre
tig00021208_arrow	5404924	5423088	Ogre	18165	45.39	tig00021208_arrow_5404924_Ogre
tig00021208_arrow	6548413	6559201	Ogre	10789	40.52	tig00021208_arrow_6548413_Ogre
tig00021208_arrow	7101710	7112791	Ogre	11082	43.97	tig00021208_arrow_7101710_Ogre
tig00021208_arrow	7527299	7548062	Ogre	20764	41.29	tig00021208_arrow_7527299_Ogre
tig00021208_arrow	8332396	8371066	Athila	38671	43.22	tig00021208_arrow_8332396_Athila
tig00021208_arrow	8401145	8408888	Ogre	7744	43.36	tig00021208_arrow_8401145_Ogre
tig00021208_arrow	8542159	8566991	Athila	24833	41.18	tig00021208_arrow_8542159_Athila
tig00021208_arrow	8923328	8931606	Ogre	8279	41.04	tig00021208_arrow_8923328_Ogre
tig00021213_arrow	1469867	1478433	Ogre	8567	42.8	tig00021213_arrow_1469867_Ogre
tig00021229_arrow	4940137	4954949	Ogre	14813	42.71	tig00021229_arrow_4940137_Ogre

tig00021242_arrow	1797653	1813661	Ogre	16009	38.73	tig00021242_arrow_1797653_Ogre
tig00021262_arrow	504475	510962	Ogre	6488	40.46	tig00021262_arrow_504475_Ogre
tig00021266_arrow	2823552	2833760	Ogre	10209	42.62	tig00021266_arrow_2823552_Ogre
tig00021276_arrow	1453761	1469162	Ogre	15402	43.35	tig00021276_arrow_1453761_Ogre
tig00021278_arrow	408960	424939	Ogre	15980	41.3	tig00021278_arrow_408960_Ogre
tig00021302_arrow	654474	661709	Ogre	7236	41.33	tig00021302_arrow_654474_Ogre
tig00021302_arrow	804566	827430	Ogre	22865	41.23	tig00021302_arrow_804566_Ogre
tig00021302_arrow	1216923	1225916	Ogre	8994	40.56	tig00021302_arrow_1216923_Ogre
tig00021302_arrow	1319742	1356713	Ogre	36972	43.43	tig00021302_arrow_1319742_Ogre
tig00021302_arrow	1540533	1554118	Ogre	13586	43.47	tig00021302_arrow_1540533_Ogre
tig00021302_arrow	1734855	1742840	Ogre	7986	38.12	tig00021302_arrow_1734855_Ogre
tig00021305_arrow	68174	76115	Ogre	7942	37.79	tig00021305_arrow_68174_Ogre
tig00021307_arrow	1281290	1288245	Ogre	6956	44.01	tig00021307_arrow_1281290_Ogre
tig00021307_arrow	1489344	1509737	Ogre	20394	41.73	tig00021307_arrow_1489344_Ogre
tig00021307_arrow	2927861	2935230	Ogre	7370	42.65	tig00021307_arrow_2927861_Ogre
tig00021307_arrow	2943814	2950446	Ogre	6633	43.78	tig00021307_arrow_2943814_Ogre
tig00021330_arrow	2262659	2276646	Ogre	13988	38.81	tig00021330_arrow_2262659_Ogre
tig00021330_arrow	4502799	4520141	Ogre	17343	40.29	tig00021330_arrow_4502799_Ogre
tig00021330_arrow	4509816	4516545	Ogre	6730	40.68	tig00021330_arrow_4509816_Ogre
tig00021392_arrow	1945429	1952111	Ogre	6683	42.03	tig00021392_arrow_1945429_Ogre
tig00021415_arrow	394872	412428	Ogre	17557	41.9	tig00021415_arrow_394872_Ogre
tig00021433_arrow	1139293	1146511	Ogre	7219	41.82	tig00021433_arrow_1139293_Ogre
tig00021447_arrow	646757	656528	Ogre	9772	47.35	tig00021447_arrow_646757_Ogre
tig00021450_arrow	438725	461319	Ogre	22595	46.12	tig00021450_arrow_438725_Ogre
tig00021450_arrow	522914	542068	Ogre	19155	42.27	tig00021450_arrow_522914_Ogre
tig00062639_arrow	513902	525264	Ogre	11363	42.84	tig00062639_arrow_513902_Ogre

Figura S1.1. Árbol filogenético del dominio integrasa (INT) utilizando todos los retrotransposones LTR (RT LTR) pertenecientes a los linajes Athila, Tat y Ogre de *Acca sellowiana* de nuestro flujo de trabajo y el de RepeatMasker con REannotate (RMRE). La filogenia se construyó utilizando máxima verosimilitud a partir del alineamiento de las secuencias nucleotídicas de los dominios INT recortado con trimAl del superclado OTA en *Acca sellowiana*. En verde elementos Athila, en amarillo elementos Tat y en rosado elementos Ogre. El outgroup, RT LTR Ty3/Gypsy de chromoviruse CRM corresponde al círculo naranja. En cada elemento se encuentra el nombre del contig al que pertenecen, inicio y fin de su ubicación, aquellos que no presentan el sufijo “tig” pertenecen a Gypsy database 2.0 (GyDB). Los dominios INT pertenecientes a la base de datos se marcan con asteriscos para cada linaje.

Figura S1.2. Árbol filogenético del dominio retrotranscriptasa (RT) utilizando todos los retrotransposones LTR (RT LTR) pertenecientes a los linajes Athila, Tat y Ogre de *Acca sellowiana* de nuestro flujo de trabajo y el de RepeatMasker con REannotate (RMRE). La filogenia se construyó utilizando máxima verosimilitud a partir del alineamiento de las secuencias nucleotídicas de los dominios RT recortado con trimAl del superclado OTA en *Acca sellowiana*. En verde elementos Athila, en amarillo elementos Tat y en rosado elementos Ogre. El outgroup, RT LTR Ty3/Gypsy de chromoviruse CRM corresponde al círculo naranja. En cada elemento se encuentra el nombre del contig al que pertenecen, inicio y fin de su ubicación, aquellos que no presentan el sufijo “tig” pertenecen a Gypsy database 2.0 (GyDB). Los dominios RT pertenecientes a la base de datos se marcan con asteriscos para cada linaje.

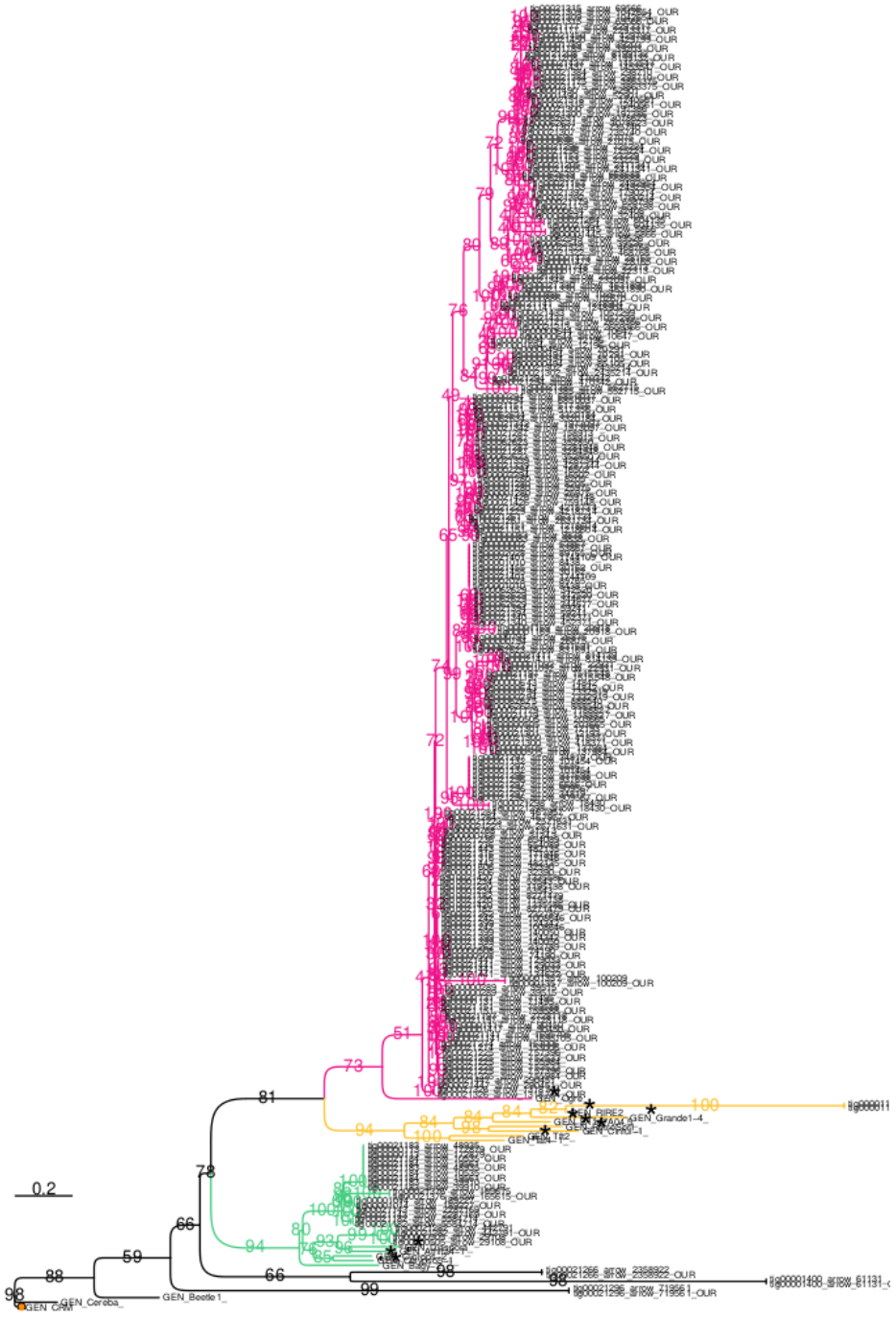

Figura S1.3. Árbol filogenético del dominio ribonucleasa H (RNaseH) utilizando todos los retrotransposones LTR (RT LTR) pertenecientes a los linajes Athila, Tat y Ogre de *Acca sellowiana* de nuestro flujo de trabajo y el de RepeatMasker con REannotate (RMRE). La filogenia se construyó utilizando máxima verosimilitud a partir del alineamiento de las secuencias nucleotídicas de los dominios RNaseH recortado con trimAl del superclado OTA en *Acca sellowiana*. En verde elementos Athila, en amarillo elementos Tat y en rosado elementos Ogre. El outgroup, RT LTR Ty3/Gypsy de chromoviruse CRM corresponde al círculo naranja. En cada elemento se encuentra el nombre del contig al que pertenecen, inicio y fin de su ubicación, aquellos que no presentan el sufijo “tig” pertenecen a Gypsy database 2.0 (GyDB). Los dominios RNaseH pertenecientes a la base de datos se marcan con asteriscos para cada linaje.

Figura S2. Árbol filogenético del dominios concatenados retrotranscriptasa, ribonucleasa H e integrasa (RT-RNaseH-INT) utilizando todos los retrotransposones LTR (RT LTR) pertenecientes a los linajes Athila, Tat y Ogre de *Acca sellowiana* de nuestro flujo de trabajo y el de RepeatMasker con REannotate (RMRE). La filogenia se construyó utilizando máxima verosimilitud a partir del alineamiento de las secuencias nucleotídicas de los dominios RT-RNaseH-INT recortado con trimAl del superclado OTA en *Acca sellowiana*. En verde elementos Athila, en amarillo elementos Tat y en rosado elementos Ogre. El outgroup, RT LTR Ty3/Gypsy de chromoviruse CRM corresponde al círculo naranja. En cada elemento se encuentra el nombre del contig al que pertenecen, inicio y fin de su ubicación, aquellos que no presentan el sufijo "tig" pertenecen a Gypsy database 2.0 (GyDB). Los dominios RT-RNaseH-INT pertenecientes a la base de datos se marcan con asteriscos para cada linaje.