UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Agronòmica i del Medi Natural

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural

Desarrollo de herramientas en Python para el estudio de la diversidad de elementos transponibles: análisis en el género Nicotiana

Trabajo Fin de Grado

Grado en Biotecnología

AUTOR/A: Amata, Agustín

Tutor/a: Forment Millet, José Javier

Director/a Experimental: Bombarely Gomez, Aureliano

CURSO ACADÉMICO: 2023/2024

Development of Tools in Python for the Study of the Diversity of Transposable Elements: Analysis in the *Nicotiana* Genus

Keywords: Transposon; *Nicotiana*; Diversity; TE burst; Python; Evolution

Abstract: Transposable elements (TEs) or transposons are DNA mobile elements with a great ability to move across the genome and intervene in its structure and evolution, representing a source of genetic variability. Several environmental and genetic factors regulate their diversification and activity, so these structures can be useful to study evolutionary relationships between species. There exist different types of software to identify and classify TEs, such as RepeatModeler or TESorter. Programs such as RepeatMasker can recognize the location of the TEs in the genome. Nonetheless, there is an important lack of tools that allow to analyze the diversity of TEs between species, compare their transposon profiles and analyze the divergence of different copies of a given TE in different genomes. The present Bachelor's Degree Final Project has developed a package tools in Python, named Repeattools, designed to perform statistical analysis and visualization on the output of the programs RepeatMasker and TESorter. Its performance has been evaluated in this work, as well as it has been applied to a user case: the study of the TE evolution on the *Nicotiana* genus. Performance results show that the RECollector program can process large files coming from RepeatMasker and TESorter. Furthermore, the use of this program on the study on the *Nicotiana* genus has allowed us to identify a specific footprint on the *Suaveolentes* section associated with the TE *Copia*, *hAT* and SINE elements. The hypothesis that the expansion of these elements during the evolution of the section may have been involved in the diversification of the section across the Australian territory has been proposed. In conclusion, Repeattools has been shown to prove a useful package of tools for the study of transposable elements in the field of evolutionary and population genomics.

Student: Agustín Amata

Tutor: José Javier Forment Millet

Experimental tutor: Aureliano Bombarely Gómez

Academic year: 2023/2024

Desarrollo de herramientas en Python para el estudio de la diversidad de elementos transponibles: análisis en el género *Nicotiana*

Palabras clave: Transposón; *Nicotiana*; Diversidad; Estallido de ET; Python; Evolución

Resumen: Los elementos transponibles (ET) o transposones son elementos móviles del ADN con una gran capacidad para desplazarse a lo largo del genoma y que intervienen tanto en la estructura como en la evolución de este último, por lo que representan una fuente de variación genética. Diversos factores ambientales y genéticos regulan su diversificación y su actividad, de modo que son estructuras interesantes para estudiar relaciones evolutivas entre distintas especies. Existen diferentes softwares para identificar y clasificar ET, tales como RepeatModeler o TESorter. Por otro lado, programas como RepeatMasker permiten la anotación de dichos ET en el genoma. Sin embargo, no existen herramientas que permitan analizar la diversidad de ET entre especies, ni comparar sus perfiles transposónicos o la divergencia de las distintas copias de un determinado ET en diferentes genomas. El presente TFG desarrolla un paquete de herramientas en Python, llamado Repeattools, que permite combinar datos procedentes de RepeatMasker y TESorter para su posterior análisis estadístico y su visualización. Además, se evalúa su rendimiento en diferentes soportes y se analiza un grupo de especies de diferentes secciones del género *Nicotiana* para obtener más información acerca de sus relaciones evolutivas, representando un ejemplo de funcionamiento del paquete. Los resultados obtenidos del rendimiento del programa RECollector indican que puede procesar archivos de gran tamaño procedentes de RepeatMasker y TESorter. Asimismo, el estudio de especies del género *Nicotiana* evidencia que las especies australianas de la sección *Suaveolentes* poseen un perfil de elementos transponibles característico, formado principalmente por elementos *Copia*, *hAT* y SINE. Además, se propone la hipótesis de que la expansión de dichos elementos transponibles durante la evolución de la sección haya podido estar involucrada en la diversificación de esta por el territorio australiano. De este modo, se concluye que Repeattools resulta un paquete de herramientas útil para el estudio de elementos transponibles en el ámbito de la genómica evolutiva y poblacional.

Alumno: D. Agustín Amata

Tutor: Prof. D. José Javier Forment Millet

Tutor experimental: D. Aureliano Bombarely Gómez

Curso académico: 2023/2024

Agradecimientos

Primero de todo, quiero agradecer a mis padres y a mi hermana por todo el apoyo que me han dado durante estos cuatro años de grado. Gracias de corazón por el cariño con el que siempre me habéis tratado, tanto en los malos como en los buenos momentos.

Gracias también a mis amigos de siempre (Vicent, Mario, Marcos, Cristian, Manuel), con los que he compartido tantos buenos momentos y con los que he compartido mis experiencias en el grado, así como ellos también han compartido las suyas conmigo.

Quiero agradecer también al profesor Javier Forment y al Doctor Aureliano Bombarely por la oportunidad que me han ofrecido de trabajar en su laboratorio de bioinformática con este proyecto con el que he descubierto el camino que quiero seguir profesionalmente. Agradezco también a Aureliano toda su guía y apoyo a lo largo del proyecto, con el que no habría sido capaz de completarlo de forma tan detallada.

Muchas gracias también a Víctor por sus consejos y ayuda a la hora de redactar el código sobre el que se basa todo este proyecto. También le agradezco a él y al resto de integrantes que han pasado por el laboratorio de bioinformática (Martín, Sabela, Mar, Luana, Lluna, Chiara, Paulo, Olivia, María, Alberto y Koko) por hacer mucho más agradable y amena mi estancia.

Por último, quiero agradecer a Javi y a Adam por haber estado conmigo durante estos cuatro años de grado. Sois gente trabajadora y que se esfuerza y espero que os siga yendo igual de bien o mejor en vuestro futuro.

Gracias a todos por haber formado parte de mi vida, haciéndome quien soy ahora.

Índice

Índice de figuras

Índice de tablas

	Alto	Medio	Bajo	No procede
ODS 1. Fin de la pobreza				**X**
ODS 2. Hambre cero				**X**
ODS 3. Salud y bienestar				**X**
ODS 4. Educación de calidad				**X**
ODS 5. Igualdad de género				**X**
ODS 6. Agua limpia y saneamiento				**X**
ODS 7. Energía asequible y no contaminante				**X**
ODS 8. Trabajo decente y crecimiento económico				**X**
ODS 9. Industria, innovación e infraestructuras				**X**
ODS 10. Reducción de las desigualdades				**X**
ODS 11. Ciudades y comunidades sostenibles				**X**
ODS 12. Producción y consumo responsables				**X**
ODS 13. Acción por el clima	**X**			**X**
ODS 14. Vida submarina				**X**
ODS 15. Vida de ecosistemas terrestres		**X**		**X**
ODS 16. Paz, justicia e instituciones sólidas				**X**
ODS 17. Alianzas para lograr objetivos.				**X**

El trabajo se alinea en un grado alto con el Objetivo de Desarrollo Sostenible (ODS) número 13, acción por el clima, debido al análisis de los elementos transponibles de las plantas del género *Nicotiana*. En concreto, tienen mayor relevancia los resultados obtenidos en relación con la evolución de la sección *Suaveolentes* en Australia. Se conoce que las especies australianas de *Suaveolentes* experimentaron un periodo de rápida diversificación mucho tiempo después del origen de la sección. Este hecho pudo venir condicionado por la expansión de ciertos elementos transponibles antes y durante ese periodo de diversificación. Por otro lado, estas especies se han adaptado a las condiciones áridas y hostiles del interior del territorio australiano. Continuar con la investigación sobre la diversificación de esta sección de *Nicotiana* y su relación con los elementos transponibles contribuirá a entender la proliferación de estas plantas en entornos secos y de condiciones extremas. Esto resulta de utilidad en un contexto de desertificación en varias partes del mundo debido al aumento de las temperaturas, en el que se buscan soluciones para generar plantas más resistentes al calor y la sequía.

1. INTRODUCCIÓN

1.1. El Papel de los Elementos Transponibles en la Evolución

1.1.1. *Elementos transponibles: definición, clasificación y diversidad entre especies*

Los elementos transponibles (ET) o transposones son elementos de ADN móviles capaces de cambiar su posición en el genoma huésped. Fueron descubiertos por Barbara McClintock en sus estudios sobre la herencia de la coloración de los granos de maíz durante el siglo pasado. Se caracterizan por poseer regiones codificantes de proteínas con diferentes actividades moleculares y secuencias que promueven su transcripción (Wells y Feschotte, 2020). Desde el descubrimiento de los primeros ET (*Ac* y *Ds*) en maíz, se han identificado una gran variedad de transposones, dando a conocer la diversidad de estas estructuras en lo que respecta a su organización y formas de desplazamiento por el genoma. Estos pueden encontrase en todos los tipos de organismos a excepción de los virus y algunas bacterias (Zhao *et al.*, 2016).

En este sentido, un primer sistema de clasificación de ET fue propuesto por Finnegan (1989), en el que separaba los transposones en dos clases. La clase I consistía en elementos capaces de transponerse mediante transcripción reversa con intermediarios de ARN (mecanismo de "copiar y pegar") mientras que la clase II agrupaba elementos capaces de transponerse directamente sin la necesidad de intermediarios (mecanismo de "cortar y pegar"). Sin embargo, el descubrimiento de ET que "copian y pegan" sin requerir intermediarios de ARN y de los elementos transponibles de repetición invertida en miniatura (*Minitature Inverted-repeat Transposable Elements* o MITEs) dio lugar a otros sistemas de clasificación que bien introducían una tercera clase o bien apostaban por una clasificación basada en otros criterios (Wicker *et al.*, 2007).

De este modo, Wicker *et al.* (2007) propusieron un nuevo sistema jerárquico que combinaba la división en dos clases de Finnegan (1989) con criterios enzimáticos y mecanicistas. Como antes, las clases se dividen según la presencia o ausencia de intermediario de ADN. La clase I (los llamados retrotransposones) se subdivide en órdenes (tales como LTR-RT (*Long-Terminal RetroTransposon*), LINE (*Long-INterspersed Elements*), PLE (*Penelope-Like Elements*), SINE (*Short-INterspersed Elements*) o DIRS (*Dictyostelium Intermediate Repeat Sequences*)) en función del mecanismo de inserción. Dentro de la clase II (conocidos como transposones de ADN) encontramos distintas subclases (por ejemplo, TIR (*Terminal-Inverted Repeats*), Crypton, Helitron o Maverick) en base a si el movimiento de los ET por el genoma se basa en insertar copias en otras regiones o bien abandonan el sitio original para reintegrarse en otra región. Órdenes y subclases se dividen a su vez en superfamilias y estas en familias (y estas, en ocasiones, en subfamilias) en base a características estructurales y de conservación de su secuencia de ADN, respectivamente. Dicho esto, este sistema de clasificación presenta sus inconvenientes, bien sea porque fueron explicados por los mismos autores o bien porque otros (Seberg y Petersen, 2009) repararon en dichas inconsistencias.

Los transposones son especialmente abundantes en los genomas vegetales. La variación del tamaño de su genoma, que puede llegar a ser de varios órdenes de magnitud, se explica en parte por la proliferación y supervivencia de diferentes tipos de ET (Wendel et al., 2016). Destaca también la disparidad entre distintos taxones en cuanto a qué grupos de ET dominan sus genomas. En el caso de las angiospermas, los retrotransposones del orden LTR-RT son los más abundantes; por ejemplo, en el caso del maíz (*Zea mays*) llegan a ocupar cerca del 80% de su genoma. Las superfamilias *Copia* y *Gypsy* son las más frecuentes (Mhiri et al., 2022). En cambio, en mamíferos los elementos predominantes pertenecen a retrotransposones no-LTR tales como LINE y SINE (Richardson et al., 2015).

Por otro lado, no sabemos todavía con total exactitud los factores que gobiernan esta acumulación y diversificación en los genomas de distintas especies. Además, todavía queda la incógnita de por qué ciertas especies poseen genomas dominados por una o muy pocas familias de ET, mientras otros genomas se caracterizan por su enorme diversidad en cuanto a estos elementos se refiere. En este sentido, una de las hipótesis que se baraja está relacionada con la genética de poblaciones, en la que se propone que estas diferencias vienen dadas por el tamaño de población efectiva, debido a que la eliminación de mutaciones deletéreas es proporcional a dicha población. Si bien esta hipótesis puede explicar la mayor frecuencia de fijación de ET en vertebrados que en moscas de la fruta, no puede justificar, por ejemplo, las variaciones en la diversidad de ET entre especies con poblaciones similares o entre ciertos grupos taxonómicos (Wells & Feschotte, 2020).

1.1.2. *Ciclo de vida de un elemento transponible, domesticación y estallidos*

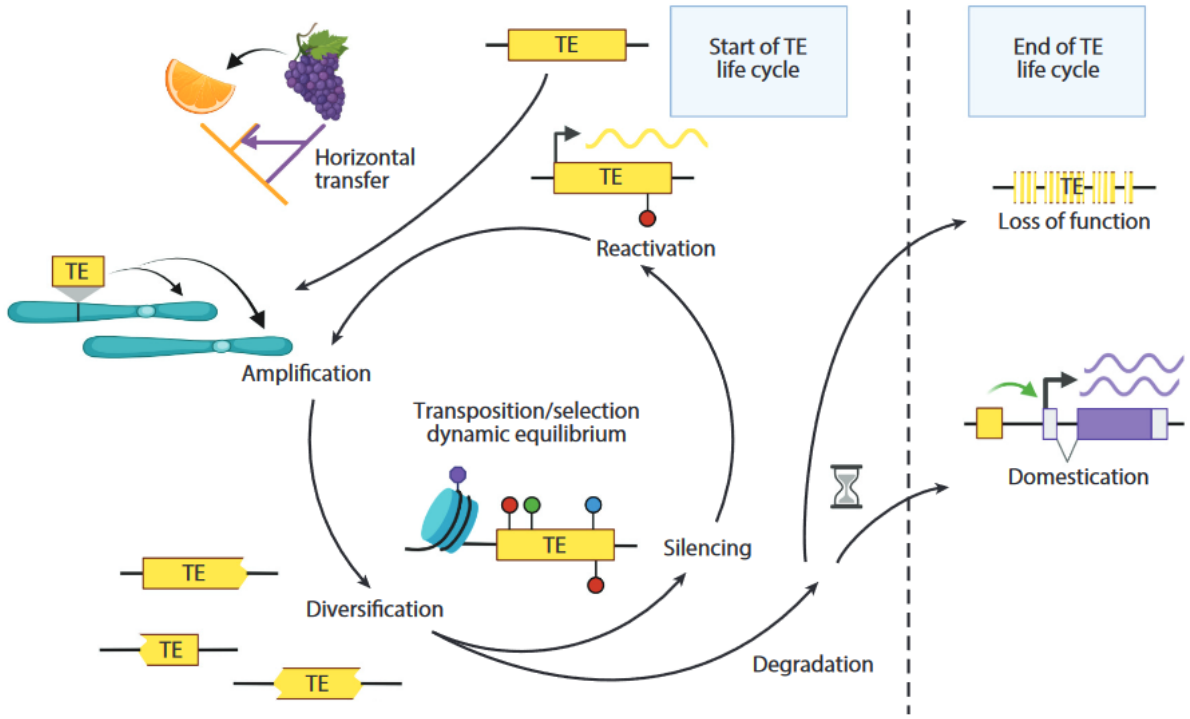

Figura 1.1. Diagrama-resumen del ciclo de vida de los elementos transponibles. En resumen, el ciclo comienza con la inserción del ET en el genoma. A ello le sigue un periodo de rápida amplificación y diversificación. El ciclo continúa con el silenciamiento de los ET, que se mantienen inactivos hasta que surja una oportunidad para reactivarse. Por otro lado, los ET pueden escapar del ciclo bien por la pérdida de su función o bien por ser domesticados por el organismo huésped para su beneficio. (Obtenido de Liu et al. (2022))

Los ET siguen un ciclo de vida (Fig. 1.1) en el genoma en el que existen que va a condicionar su abundancia y diversidad en el mismo. El ciclo comienza bien con la transferencia de un ET en un nuevo genoma mediante transferencia horizontal o bien con la reactivación de un ET inactivo que ya estaba insertado en el genoma (Liu et al., 2022). A partir de este punto, el ET tiene un proceso de amplificación (aumento en el número de copias del elemento) y de actividad desenfrenada en un genoma que todavía no ha conseguido regularlo, siempre y cuando el elemento sea capaz de transponerse de forma autónoma, el proceso se realice correctamente y la maquinaria requerida para la transposición esté presente. Esta etapa de gran actividad da lugar a un periodo de diversificación de los ET y de mutagénesis (Liu et al., 2022). Es también en esta diversificación donde se forman nuevas familias de ET, que incluyen elementos que contienen fragmentos de genes, ET no autónomos y MITE (Zhao et al., 2016).

En este punto del ciclo, los ET pueden tomar dos destinos distintos: continuar el ciclo en el caso de ser silenciados, con la posibilidad de reactivarse en un periodo futuro, o bien salir del ciclo, en el caso de que vayan degradándose por mecanismos discutidos más adelante. La continuación del ciclo implica que los ET van a pasar por una etapa de represión de su actividad. En el caso de ser elementos ya establecidos en el genoma, estos serán simplemente re-silenciados mientras que los elementos introducidos por transferencia horizontal van a ser reconocidos por la maquinaria de silenciamiento. Sea como fuese, la célula recurre al

silenciamiento de los ET por medio de modificaciones epigenéticas. El silenciamiento epigenético permite la represión de todos los elementos de una misma familia (Liu et al., 2022).

Principalmente, el mecanismo por el que se lleva a cabo este silenciamiento en plantas es conocido como silenciamiento basado en identidad, en el que interviene un mecanismo de formación de ARN pequeño de interferencia (*Small Interference RNA* o siRNA). De este modo, el silenciamiento de los ET viene mediado por siRNA que proceden de la misma secuencia de dichos elementos (Fultz & Slotkin, 2017). Este mecanismo se inicia con la formación de un ARN de doble cadena procedente de transcritos de ET y causado por diferentes posibilidades. El fragmento de ARN bicatenario es procesado por la maquinaria de silenciamiento del organismo y da lugar al silenciamiento del resto de elementos de la familia por homología con el ARN de interferencia, que actúa como un elemento de regulación genética en *trans*. Posteriormente, la alteración de la expresión se mantiene gracias a la metilación de las secuencias basada en siRNA de los propios ET (creados por las polimerasas de plantas Pol IV y Pol V), en la que median diferentes metiltransferasas, tales como MET1, CMT3 y CMT2 (Lisch & Slotkin, 2011).

Sin embargo, los ET tienen varias formas de escapar el silenciamiento epigenético y completar el ciclo volviendo a la etapa de amplificación. Algunas formas ocurren durante eventos de transferencia horizontal del elemento o en eventos de hibridación, gracias a la ausencia o baja presencia de mecanismos de silenciamiento, mientras que otras formas consisten en aprovechar la regulación de genes cercanos, adquirir mutaciones que inhiban la maquinaria de silenciamiento, generar productos que combatan dicha maquinaria o aprovechar ciertos momentos o cambios en el organismo huésped, como son la presencia de estreses, mutaciones en la maquinaria de silenciamiento o periodos de relajación en ciertas etapas del desarrollo (Liu et al., 2022).

Finalmente, el fin del ciclo (y su salida de él) viene marcado por la pérdida de autonomía del elemento y su lenta deriva hacia su degradación. Liu et al. (2022) resumen en tres los principales mecanismos que conducen a la pérdida de función y "muerte" de los ET. En primer lugar, los ET corren el mismo riesgo de sufrir mutaciones que el resto de las secuencias en el genoma. En segundo lugar, encontramos eventos de recombinación ectópica entre ET que pueden dar lugar a modificaciones cromosómicas considerables. Por último, pueden ocurrir eventos de transposición incompletos que dan lugar a copias no autónomas o aberrantes.

Adicionalmente, el fin del ciclo no solo se da con la "muerte" de los ET, sino que también se da con su domesticación por parte del organismo huésped. Por domesticación se entiende el proceso que da lugar a la conversión del ET, considerado por el organismo un elemento foráneo que debe ser silenciado, por un elemento con funciones beneficiosas para este (Almeida et al., 2022).

En general, la domesticación de los ET puede devenir en distintos escenarios, en función de las características finales del ET reprogramado. Aquí, veremos brevemente tres escenarios distintos.

- En el primer escenario, se reprograma el mecanismo de transposición del ET, es decir, se conserva la capacidad de transposición, pero esta queda baja regulación del organismo huésped para su propio beneficio. Un ejemplo destacado de este escenario en vertebrados es la reacción de recombinación V(D)J durante el desarrollo de los linfocitos, basado en la domesticación de la familia *Transib*. Por otro lado, un ejemplo claro en

plantas lo constituyen los genes de resistencia de plantas, especializados en reconocer moléculas asociadas a patógenos y actuar en respuesta (Liu et al., 2022).

- El siguiente escenario consiste en la adquisición de una función nueva y diferente por parte de los ET. En este caso, los genes codificantes de proteínas de los ET sufren procesos de neofuncionalización, como es el caso de la transposasa del ET *Mutator*, que dio lugar a factores de transcripción vitales para la respuesta a la luz de las plantas superiores (Lin et al., 2007). Sin embargo, también existe la posibilidad de que la domesticación del ET se dé por la formación de secuencias quiméricas entre fragmentos del ET y genes propios del huésped (Liu et al., 2022).
- El último escenario consiste en la reprogramación de los ET como elementos reguladores del ADN, tanto elementos en *cis* como en *trans*. Los ET domesticados como elementos reguladores en *cis* suelen actuar como *enhancers* (potenciadores de la expresión) o como sitios de unión de factores de transcripción. Se ha observado que en el caso de las plantas, la formación de *enhancers* a partir de ET ha apoyado ciertas características como la resistencia al aluminio y al estrés térmico y salino (Liu et al., 2022). Por el contrario, dentro de los elementos reguladores en *trans* procedentes de ET podemos encontrar que actúan, por ejemplo, como siRNA (como se ha mencionado previamente) o como miRNA (*MicroRNA*). En este último caso, Li et al. (2011) sugirieron un modelo en plantas para la domesticación de los ET como miRNA que incluía modificaciones en estos tales como la formación de repeticiones invertidas, la especiación de secuencias y la adaptación de las secuencias a la maquinaria de miRNA del organismo.

1.1.3. *Herramientas para la identificación de elementos transponibles*

La detección y búsqueda de elementos transponibles resulta una tarea compleja debido a la naturaleza repetitiva de estos y a la diversidad estructural que presentan. Este hecho ha dado lugar a una variedad de técnicas y *software* basados en distintos enfoques, si bien todos presentan deficiencias que impiden la completa identificación de los ET en diversas especies. Podemos dividir los *software* de detección y anotación de ET según la estrategia que sigan en dos principales grupos (Goerner-Potvin y Bourque, 2018):

1. Detección y anotación basada en repositorios, mediante la aportación de un genoma ensamblado.
2. Detección y anotación *de novo*, que bien puede utilizar un genoma ensamblado o bien se pueden emplear directamente las lecturas obtenidas de la secuenciación.

Por una parte, la detección y anotación basada en repositorios tiene como objetivo principal la búsqueda a lo largo de todo el genoma proporcionado de secuencias consenso (secuencias representativas de un conjunto particular de secuencias distintas) de ET o de motivos asociados a estos elementos. Para ello, las herramientas basadas en este enfoque normalmente realizan búsquedas de homología de secuencias mediante diferentes algoritmos (Ramakrishnan et al., 2022). Dentro de este tipo de anotación destaca RepeatMasker (https://www.repeatmasker.org/) por ser la más utilizada, y que utiliza las bases de datos RepBase (https://www.girinst.org/) y Dfam (https://dfam.org/) para detectar y anotar ET mediante el uso de algoritmos de búsqueda como RMBlast y HMMER. Otro ejemplo lo podemos encontrar en TESorter (Zhang et al., 2022), que permite clasificar LTR-RT hasta el nivel de clado, subdivisiones especiales para este orden destinadas a una mejor clasificación de este (Zhang et al., 2022), gracias al uso de las bases de datos GyDB (https://gydb.org/index.php/Main_Page) y REXdb (http://repeatexplorer.org/?page_id=918) y el algoritmo HMMScan.

Por otra parte, la detección y anotación *de novo* permite tanto la identificación de ET en nuevos genomas como la detección de secuencias no reconocidas por otros programas como RepeatMasker y de nuevas familias de ET (Goerner-Potvin & Bourque, 2018). Para ello, existen dos enfoques principales utilizados por los programas de esta categoría. El primero consiste en el uso de k-meros, es decir, subsecuencias (en este caso del genoma a anotar) de longitud *k*, empleando una amplia variedad de métodos para analizarlos. Algunos ejemplos de este enfoque son RepeatScout, RED, phRAIDER y P-Clouds (Storer et al., 2022). En cambio, el segundo enfoque consiste en utilizar directamente las lecturas procedentes de los distintos métodos de secuenciación de nueva generación que, previo paso de filtrado de lecturas, permiten evitar sesgos introducidos por los programas de ensamblado de genomas (Storer et al., 2022).

Finalmente, es útil mencionar que muchas de estas herramientas se encuentran integradas en paquetes o *pipelines*, que usan de forma paralela y/o secuenciales programas dirigidos a un paso específico del análisis. Un buen ejemplo de ello es RepeatModeler2, el cual usa un primer paso de identificación de repeticiones con RepeatScout basado en abundancia de k-meros. Tras este, usa cinco pasos adicionales donde las repeticiones son refinadas y clasificadas con BLAST y RECON, para finalmente mejorar las anotaciones de los elementos LTR con LTRdetector y LTR_retriever (Flynn et al., 2020). Otras *pipelines* tan populares como RepeatModeler2 son REPET y EDTA.

1.2. Elementos Transponibles en el Género *Nicotiana*

Nicotiana representa el quinto género de plantas más grande de la familia *Solanaceae*, entre la que se incluyen otros géneros como *Solanum* y *Cestrum*. *Nicotiana* está formado actualmente por más de 86 especies, la mayor parte de las cuales habitan las Américas y Australia (Bally et al., 2021). Destacan *Nicotiana benthamiana* y *Nicotiana tabacum* (tabaco) por sus usos en investigación y consumo humano, respectivamente. Asimismo, el género se encuentra dividido en 13 secciones diferentes, de las cuales 5 corresponden a especies alotetraploides (cuatro conjuntos de cromosomas procedentes de dos progenitores diploides), constituyendo casi la mitad de las especies del género (Knapp et al., 2004).

De entre las secciones alotetraploides destaca *Suaveolentes* (en la que se incluye *N. benthamiana*) por ser la más diversificada. Esta sección está compuesta por especies que se encuentran mayoritariamente en Australia, si bien también son nativas de África y de varias islas del Pacífico (Bally et al., 2021). Las especies australianas de la sección se encuentran ampliamente distribuidas por todo el territorio, hecho que hace surgir la cuestión de su diversificación. La hipótesis que más fuerza parece cobrar actualmente sitúa la aparición de la sección *Suaveolentes circa* 6 millones de años, gracias a análisis filogenéticos que apuntan a un único evento de poliploidización (Clarkson et al., 2017; D'Andrea et al., 2023). En este sentido, los orígenes de la sección parecen encontrarse en secciones sudamericanas de *Nicotiana*, desde donde pudo surgir la sección *Suaveolentes* para luego dispersarse hacia África, el Pacífico y Australia. Si bien no se ha logrado esclarecer completamente, la sección *Sylvestres* parece ser el progenitor paterno mientras que la sección *Noctiflorae* pudo contribuir como progenitor materno, a lo que se suman contribuciones de las secciones *Alatae* (respecto del progenitor paterno) y *Petunioides* (respecto del progenitor materno) (Kelly et al., 2013). D'Andrea et al. (2023) propusieron que el surgimiento de *Suaveolentes* se produjo hace unos 6 millones de años mediante de un evento de poliploidización entre ancestros comunes de *Sylvestres* y *Alatae*, que actuó como progenitor paterno, y de *Noctiflorae* y *Petunioides*, que actuó como progenitor materno, previo a su diversificación en las cuatro secciones contemporáneas. Por otro lado, Wang et al. (2023) sitúan este evento casi un millón de años antes y teniendo a *N. sylvestris* (sec. *Sylvestres*) como progenitor paterno y a *N. glauca* (sec. *Noctiflorae*) como ejemplo de progenitor materno. Sea como fuere, la introducción del ancestro de la sección *Suaveolentes* en Australia se habría producido tras un periodo de aridificación (*ca*. 10-7 millones de años, (Clarkson et al., 2017)) y habría logrado adaptarse y diversificarse en un territorio de condiciones extremas.

Además de esclarecer el origen concreto de *Suaveolentes*, también queda la pregunta de la rápida especiación y diversificación de la subsección australiana. Tradicionalmente se ha asociado diversificación con poliploidización, pero en el caso del género *Nicotiana*, solo la sección poliploide *Suaveolentes* presenta un alto índice de diversificación. Otras secciones tetraploides como *Repandae* y *Polydicliae* presentan un número mucho menor de especies (4 y 2 respectivamente). Una posible explicación del éxito de la subsección en la conquista y adaptación a las distintas condiciones extremas de Australia puede hallarse en los elementos transponibles. Recientemente, Ranawaka et al. (2023) identificaron en *N. benthamiana* un estallido de elementos Copia que empezó hace unos 2 millones de años, teniendo su pico hace 750.000 años, todo ello en un periodo coincidente con la rápida diversificación de la subsección australiana. En contraste, *N. africana* (la única especie de *Nicotiana* nativa de Africa, localizada concretamente en Namibia) y *N. forsteri* (una especie de *Nicotiana* australiana que se encuentra en la costa oriental) (D'Andrea et al., 2023), no experimentaron un fenómeno similar al de sus parientes australianos más recientes (Clarkson et al., 2017).

2. OBJETIVOS

El objetivo principal de este Trabajo Final de Grado consiste en el **desarrollo de una herramienta que permita analizar y visualizar los resultados de dos programas comúnmente usados para la identificación de elementos transponibles, RepeatMasker y TESorter, así como integrar los resultados de varias especies de forma simultánea para su comparación**. Para ello, el proyecto se ha separado en subobjetivos:

1. Crear un programa que una los datos de RepeatMasker y TESorter (si existen de este último) para cada especie, asignarle una identificación y crear una matriz con el número en cada especie de cada ET único, así como almacenar los datos de divergencia para cada ET.
2. Con los archivos generados por el programa anterior, obtener gráficas que permitan visualizar la diversidad de ET en cada especie y compararlos entre sí.
3. Analizar el panorama de ET de varias especies de *Nicotiana*, centrándose en especies pertenecientes a la sección *Suaveolentes*, como ejemplo práctico del uso de los programas desarrollados. A su vez, se buscarán indicios sobre expansiones de ET en la sección *Suaveolentes* con tal de aportar más información a la hipótesis planteada por Ranawaka et al. (2023).

3. MATERIALES Y MÉTODOS

3.1. Desarrollo de Repeattools

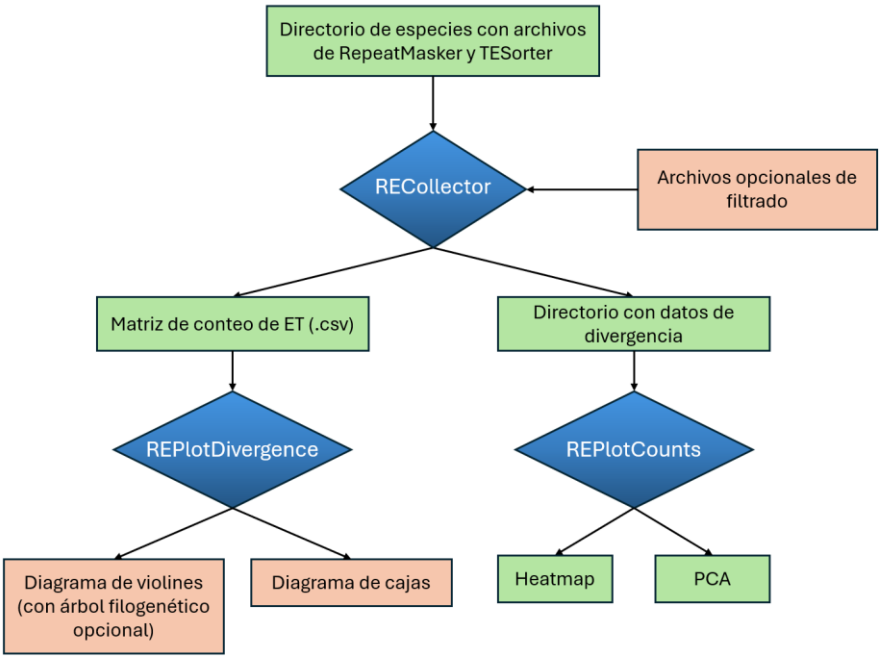

Figura 3.1. Flujo de trabajo del paquete de herramientas Repeattools con las principales entradas y salidas.

Repeattools es un paquete de herramientas desarrollado en Python 3.10.11 y que depende principalmente de las librerías Pandas, Numpy, SciPy, Scikit-learn, Matplotlib y Seaborn. El paquete se encuentra distribuido en tres programas diferentes (Fig. 3.1). RECollector integra los datos de los dos *softwares* de detección y anotación de ET a la par que crea un sistema de clasificación básico y unificado para poder llevar a cabo un análisis más homogéneo de los ET. Por otro lado, REPlotCounts y REPlotDivergence permiten llevar a cabo el análisis y comparación entre especies de la diversidad de ET. El código está a libre disposición en el siguiente repositorio de Github: https://github.com/AgustinAmata/Repeattools.

3.1.1. *RECollector: integración de los datos*

RECollector constituye el programa principal del paquete que integra los datos de RepeatMasker y TESorter para ser analizados por los siguientes programas del paquete. RECollector emplea la librería Pandas (versión 2.0.3) para llevar a cabo tal propósito. RECollector genera *dataframes* (marcos de datos) de los archivos de RepeatMasker y TESorter. Estos *dataframes* son estructuras compuestas de columnas y filas sobre las que se pueden realizar todo tipo de operaciones (operaciones aritméticas, modificación de datos en puntos concretos, obtención de subconjuntos según una serie de condiciones, etc.).

En líneas generales, RECollector toma como archivos de entrada los archivos *results.out* de RepeatMasker y *results.cls.tsv* de TESorter de cada especie, que son procesados por el programa para obtener varios archivos y directorios en su salida (Fig. 3.1). Por un lado, se genera una matriz de conteo de ET en formato CSV (*Comma-Separated Value*, esto es, Valores Separados por Comas), que contiene el perfil de ET (el número de copias en el genoma de cada ET) para cada especie analizada. Por otro lado, se genera un directorio que contiene los datos de divergencia de cada ET detectado. Este directorio se encuentra separado en distintos archivos en formato CSV para cada ET; cada archivo incluye los porcentajes de divergencia de todas las copias detectadas de ese determinado ET separadas por especie. La información contenida en los archivos de salida depende del nivel jerárquico del sistema de clasificación (explicado más adelante) seleccionado por el usuario.

Adicionalmente, RECollector requiere de un archivo de entrada con el que etiquetar a las especies analizadas. Además, el programa también presenta la posibilidad de filtrar los datos a procesar según una longitud determinada, el porcentaje de divergencia de cada copia individual y la presencia (o ausencia) de datos de dominios proteicos, aportados por TESorter, en cada copia.

Por último, RECollector utiliza un sistema de clasificación jerárquica compuesto de cuatro niveles. Esta clasificación surge de la necesidad de unificar de forma sencilla y homogénea los sistemas de clasificación utilizados por RepeatMasker y TESorter. El sistema se basa en los niveles jerárquicos del sistema propuesto por Wicker et al. (2007), junto con la clasificación de los distintos ET incluidos en la base de datos Dfam (https://dfam.org/classification), así como la clasificación más detallada para el orden LTR empleada por la base de datos REXdb (http://repeatexplorer.org/?page_id=918). De este modo, los elementos parten del nivel "clase" (distinguiéndose Clase I y Clase II, así como otros grupos para otros elementos no clasificados como ET), pasando por el nivel "subclase", seguido del nivel "superfamilia" y finalizando en el nivel "elemento".

El uso de RECollector como líneas de comandos queda resumido en el siguiente cuadro de texto.

```
## Uso de RECollector
python RECollector.py [-h] --input INPUT --names NAMES [--length LENGTH]
[--domains] [-D D] [--per] [-t T] [-m {lower_than,higher_than,equal}]
[--depth                    {class,subclass,superfamily,element,tes_order,
tes_superfamily,clade}] [--override] --output OUTPUT
```

3.1.2. *REPlotCounts: PCA y mapas de calor*

El programa REPlotCounts genera figuras que permiten observar las semejanzas y diferencias entre los perfiles de ET de cada especie. Para cumplir con este objetivo, REPlotCounts hace uso de las librerías Pandas (versión 2.0.3), NumPy (versión 1.25.0), SciPy (versión 1.11.1), Scikit-learn (versión 1.3.0), Matplotlib (versión 3.7.3) y Seaborn (versión 0.12.2).

Brevemente, REPlotCounts utiliza como archivo de entrada el archivo en formato CSV con contiene la matriz de conteo de ET de las distintas especies (Fig. 3.1). Los datos son cargados en un *dataframe* de Pandas y a partir de aquí se presentan dos procesos. Por un lado, se realiza un análisis de componentes principales (PCA en inglés) mediante las funciones que SciPy y Scikit-learn tienen dedicadas a ello. Los resultados del análisis son visualizados en un gráfico de dispersión. Por otro lado, se genera un mapa de calor con el *dataframe* original, previa normalización de los datos para permitir la comparación de datos entre las distintas especies.

REPlotCounts también requiere de un archivo en el que se especifique la distribución de las distintas especies en grupos. Asimismo, el programa presenta distintas opciones para modificar estéticamente las figuras generadas, tales como la selección de tamaño del mapa de calor, la adición de un dendrograma en el mapa de calor o la visibilidad del nombre de la especie que corresponde a cada punto en el gráfico de dispersión del PCA. Finalmente, REPlotCounts también permite excluir del análisis y la representación gráfica datos pertenecientes a elementos no relacionados con ET, así como datos que hayan sido clasificados como desconocidos.

El uso de REPlotCounts como líneas de comandos queda resumido en el siguiente cuadro de texto.

```
## Uso de REPlotCounts
python REPlotCounts.py [-h] --input INPUT [--exclude] [--dendro] [--names]
--gfile GFILE [--hsize HSIZE HSIZE] --output OUTPUT
```

3.1.3. *REPlotDivergence: análisis de los datos de divergencia*

Por último, el programa REPlotDivergence permite la representación gráfica de los datos de divergencia de cada ET para su comparación entre especies. El programa requiere del uso de Pandas (versión 2.0.3), ETE3 (versión 3.1.3), PyQt5 (versión 5.15.9), Matplotlib (versión 3.7.3) y Seaborn (versión 0.12.2).

REPlotDivergence emplea los archivos contenidos en el directorio generado por RECollector (Fig. 3.1). De esta forma, el programa precisa del directorio completo o bien de un único archivo de este si se quieren generar diagramas de violín o diagramas de cajas, respectivamente. De forma similar a REPlotCounts, REPlotDivergence genera *dataframes* de Pandas para producir los archivos de salida. En el caso de los diagramas de violín, REPlotDivergence elimina de la imagen final aquellos ET que no estén presentes en al menos el 75% de las especies analizadas.

Además, REPlotDivergence admite la entrada de un archivo en formato Newick con los datos filogenéticos de las especies analizadas para la creación de un árbol filogenético adicional en el diagrama de violines. El programa también permite excluir del diagrama de violines datos no relacionados con los ET y datos desconocidos, además de admitir la entrada de un archivo extra para agrupar las distintas especies en el diagrama de cajas.

El uso de REPlotDivergence como líneas de comandos queda resumido en el siguiente cuadro de texto.

```
## Uso de REPlotDivergence
python REPlotDivergence.py [-h] [--violin VIOLIN] [--names NAMES] [--
exclude] [--tree TREE] [--box BOX] [--groups GROUPS] --output OUTPUT
```

3.2. Rendimiento de las Funciones de Lectura de RECollector

Durante el desarrollo de RECollector y del resto de herramientas del paquete, se hizo evidente que RECollector era el programa que más tiempo tardaba en ejecutarse completa y correctamente. En concreto, la sección que mayor tiempo requería era la de la lectura de los archivos de entrada de RepeatMasker y TESorter y la de integración de sus datos en un mismo conjunto de datos. Este hecho planteó la posibilidad de que esta sección tuviese un comportamiento no lineal que dificultase o bien imposibilitase el análisis de archivos de gran tamaño. De esta forma, se llevó a cabo una prueba para evaluar el rendimiento de las funciones destinadas a realizar dichas tareas.

La prueba se realizó en *scripts* separados del resto del paquete e individuales para cada función en un ordenador portátil con el sistema operativo Microsoft Windows 11 Home, una memoria RAM de 8GB y un procesador AMD Ryzen 5 3500U con Radeon Vega Mobile Gfx, 2100 Mhz, 4 procesadores principales y 8 procesadores lógicos. Dicha prueba consistió en contabilizar el tiempo desde que se ejecutaba la función hasta que producía la salida deseada, todo ello con archivos de entrada de diferente tamaño. Esto se realizó mediante el módulo integrado de Python *time* (versión de Python 3.10.11). Asimismo, las pruebas se repitieron dos veces más para su posterior análisis estadístico.

Por un lado, para la función de lectura y creación del *dataframe* del archivo procedente de RepeatMasker, *read_repeatmasker_out()*, se utilizaron archivos de 1, 10, 100, 200, 500 y 1000MB. Los archivos de menor tamaño consistieron en versiones reducidas del archivo original de 1000MB.

Por otro lado, la prueba de la función de lectura y creación del *dataframe* del archivo procedente de TESorter se realizó con archivos de 1, 10, 20, 50, 100 y 300MB. De nuevo, se redujo el tamaño del archivo original de 300MB para generar los archivos de menor tamaño.

Por último, la prueba para la función que integra los *dataframes* producidos por las dos anteriores funciones en un único *dataframe* (llamada *merge_inputs()*) fue más compleja. En cuanto a la entrada procedente de RepeatMasker (previo uso de *read_repeatmasker_out()* para crear el *dataframe*), se emplearon dos archivos de 500 y 1000MB, procedentes de dos especies distintas. Cada una de estas entradas fue probada con sus respectivos archivos de TESorter (previa creación del *dataframe* con *read_tesorter_cls_tsv()*). En cuanto a estos últimos, se emplearon archivos de 1, 10, 25, 50 y 100MB de tamaño. De forma similar a las dos otras pruebas, los archivos de menor tamaño fueron creados mediante la reducción del archivo original correspondiente.

3.3. Caso Práctico: Análisis de Elementos Transponibles de especies del género *Nicotiana*

3.3.1. *Especies analizadas y obtención de datos* Todas las especies analizadas son pertenecientes al género *Nicotiana*. En total, fueron analizadas 26 especies que en conjunto representaban 10 de las 13 secciones en las que se divide el género (*Nicotiana, Noctiflorae, Paniculatae, Petunioides, Rusticae, Suaveolentes, Sylvestres, Tomentosae, Trigonophyllae* y *Undulatae*).

Todos los datos genómicos fueron extraídos del proyecto relacionado con el artículo publicado por D'Andrea et al. (2023). La información es accesible a través de: https://www.ncbi.nlm.nih.gov/, código de accesión PRJNA853913. Los ensamblajes de las secuencias de los genomas no publicados (todas las especies de Suaveolentes a excepción de *N. benthamiana*) se realizaron con Minia v3.2.5 con tres K-meros distintos 31, 63, y 95, tras lo cual se realizó un *scaffolding* con SOAPdenovo2 v2.40. Los huecos se rellenaron con GapCloser v1.12.r6. Todos los ensamblajes se realizaron por el Dr. Aureliano Bombarely. Para *N. africana* y *N. cavicola* se usaron dos ensamblajes distintos (v005 y v015) dependiendo de si se había realizado una ronda adicional de *scaffolding* y rellenado de huecos con Ragoo v1.1 y GapCloser respectivamente. El genoma de *N. benthamiana* versión LAB330 fue obtenido de la página web https://www.nbenth.com/. El resto de los ensamblajes se obtuvieron del NCBI, con los siguientes números de referencia: *N. attenuata*, GCA_001879085.1; *N. sylvestris*, GCA_000393655.2; *N. tomentosiformis*, GCA_000390325.3; *N. tabacum* accesión TN90, GCA_000715135.1; *N. rustica*, GCA_005239535.1; *N. glauca*, GCA_026770625.1; *N. obtusifolia* GCA_002018475.1; *N. undulata*, GCA_005239495.1; *N. otophora*, GCA_000715115.1; *N. paniculata*, GCA_005239505.1; y *N. knightiana*, GCA_005239525.1.

Una vez obtenidos los datos genómicos de las distintas especies se procedió a la detección y anotación de los ET. Esta tarea se llevó a cabo mediante el procedimiento desarrollado por Yujie Zhu para su Trabajo Final de Máster (Zhu, 2024). Brevemente, cada genoma fue procesado por RepeatModeler2 (versión 2.0.4) (Flynn et al., 2020) para detectar a las secuencias pertenecientes a ET, generando una librería con todos los ET identificados. Luego, los genomas fueron anotados mediante RepeatMasker (versión 4.1.5) y empleando la base de datos Dfam y la librería obtenida por RepeatModeler2. De este proceso se obtuvo un archivo (formato *.out*), en el que se detallan las posiciones en el genoma de cada ET, así como información adicional sobre su clasificación y su porcentaje de divergencia (número que describe cómo de idéntico es una copia de un elemento en posición dada del genoma con respecto a la secuencia con la que se ha alineado para su identificación y clasificación). Por último, se empleó TESorter (versión 1.4.6) para obtener mayor información acerca de las secuencias de ET identificadas, generando en el proceso un archivo de clasificación (formato *.cls.tsv*). Los archivos de salida obtenidos por RepeatMasker y TESorter (formatos *.out* y *.cls.tsv*, respectivamente) fueron utilizados para su análisis posterior con Repeattools.

3.3.2. *Análisis mediante Repeattools* Los archivos de RepeatMasker y TESorter relativos a cada especie fueron introducidos en un directorio, separados en subdirectorios, uno para cada especie, siguiendo con las especificaciones de RECollector. Se utilizó el argumento opcional –*override* para completar con los datos proporcionados por TESorter aquellas secuencias que no pudieron ser totalmente reconocidas por RepeatMasker. Fueron necesarias 4 rondas de análisis con RECollector para obtener todos los datos pertenecientes a los cuatro niveles jerárquicos del sistema de clasificación utilizado por el programa. Los archivos de salida obtenidos para cada nivel fueron procesados de forma diferente por REPlotCounts y REPlotDivergence.

En el caso del primer nivel, "clase", únicamente se crearon los diagramas de caja para las clases I y II mediante REPlotDivergence. Para "subclase", se creó el diagrama de violines del nivel y los diagramas de caja para LINE, LTR y SINE mediante REPlotDivergence. De forma similar al anterior nivel, los datos de "superfamilia" fueron procesados por REPlotDivergence para generar el diagrama de violines del nivel y los diagramas de cajas de los grupos *Copia*, *Gypsy* y *hAT*. En todos los niveles, los datos no relacionados con ET fueron excluidos

Finalmente, se generaron los PCA y mapas de calor de "subclase", "superfamilia" y "elemento" mediante REPlotCounts utilizando como archivos de entrada los archivos en formato CSV producidos por RECollector para cada nivel. Los datos no relacionados con los ET fueron excluidos del análisis.

Todos los comandos utilizados para el análisis se encuentran detallados en el Anexo I

4. RESULTADOS Y DISCUSIÓN

4.1. RECollector es capaz de manejar Archivos Grandes sin impedimento

Repeattools se desarrolló con una serie de archivos de pequeño tamaño ya que la optimización de la herramienta requiere de docenas de pruebas donde el tiempo de ejecución depende del tamaño de este. De esta manera, el primer análisis consistió en evaluar su ejecución con archivos reales de diferentes tamaños.

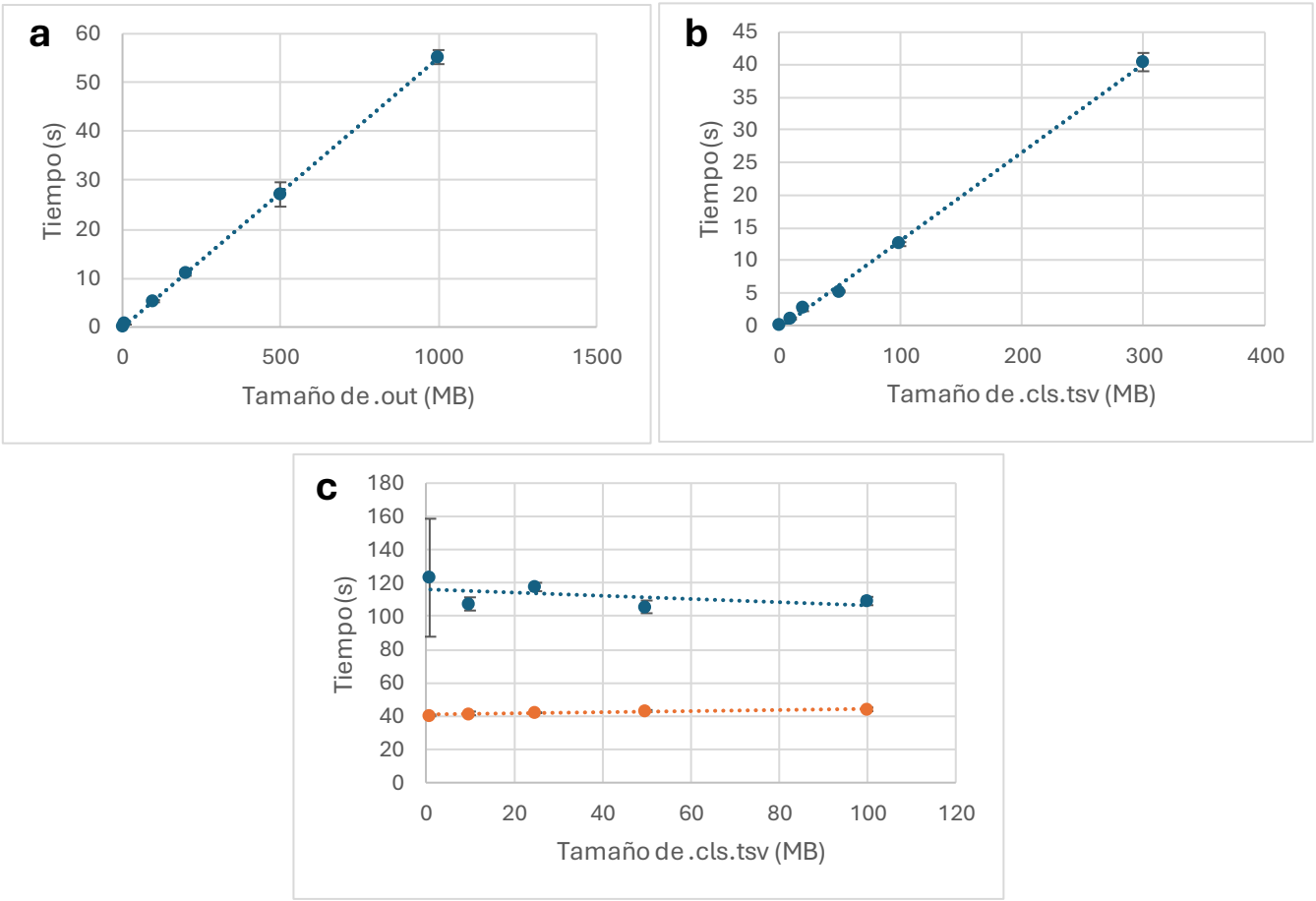

Figura 4.1. Rendimientos de las funciones de la sección de lectura de RECollector. Cada punto corresponde a un tamaño de archivo junto con el tiempo que tarda en procesarse. **a** Rendimiento de *read_repeatmasker_out()* (R^2=0,9999). **b** Rendimiento de *read_tesorter_cls_tsv()* (R^2=0,9984). **c** Rendimiento de *merge_inputs()*. El conjunto de datos azul corresponde a los resultados para el archivo de RepeatMaskser de 1000MB, mientras que el naranja corresponde a los resultados para el archivo de 500MB (R^2=0,2616 para el archivo de 1000MB y R^2=0,9052 para el archivo de 500MB).

En cuanto a las funciones de lectura para los archivos de RepeatMasker y TESorter, *read_repeatmasker_out()* y *read_tesorter_cls_tsv()*, estas presentan un crecimiento lineal en el rango analizado (Fig. 4.1a,b). Destaca que la función de lectura del archivo de entrada de TESorter presenta un crecimiento más pronunciado con respecto a la función de lectura del archivo de entrada de RepeatMasker. Este hecho puede ser debido al mayor número de operaciones internas que contiene la función *read_tesorter_cls_tsv()*.

Por otro lado, también destaca que la función para unir los datos de los dos archivos de entrada, *merge_inputs()* (Fig. 4.1c), presenta un tiempo constante con respecto a la variación de tamaño del archivo de entrada de TESorter. Sin embargo, se observa que el tiempo de ejecución de la función aumenta conforme mayor es el tamaño del archivo de entrada de RepeatMasker, aunque con estos datos se desconoce si el crecimiento es o no lineal.

Con todo, se puede afirmar que la sección de lectura de los archivos de entrada y de creación del *dataframe* final de RECollector no suponen un impedimento para el análisis de archivos de gran tamaño.

4.2. Las especies australianas de *Suaveolentes* tienen un perfil de Elementos Transponibles diferente al resto de la sección

A fin de evaluar la utilidad del paquete de herramientas de Repeattools, éste se aplicó sobre los datos reales obtenidos de RepeatMasker y TESorter sobre distintos genomas de *Nicotiana*. El objetivo era identificar patrones de evolución de elementos repetitivos, especialmente en la sección Suaveolentes. Los resultados de RepeatMasker se resumen en la Tabla 4.1.

Tabla 4.1. Resumen del análisis de elementos repetitivos obtenidos con RepeatMasker en cada uno de los ensamblajes utilizados. Las especies indicadas con un asterisco (*) no fueron tomadas en cuenta para el análisis con Repeattools al fallar el *pipeline* descrito en el apartado 3.3.1.

Species	Assembly Size (Gb)	Repeats (%)	SINE (%)	LINE (%)	LTRCopia (%)	LTRGypsy (%)	DNA TE (%)
N_africana	4.37	81.50	0.03	2.54	6.69	32.53	1.29
N_amplexicaulis	2.20	79.79	0.13	3.75	7.63	25.32	1.53
N_attenuatta	2.37	73.87	0.07	1.76	14.96	35.49	1.19
N_benthamiana	2.84	79.26	0.10	4.14	16.66	23.58	2.40
N_cavicola	1.89	78.27	0.11	3.41	11.14	22.47	2.30
N_excelsior	3.21	79.75	0.17	3.34	11.60	22.41	1.90
N_glauca	3.22	80.23	0.03	1.38	8.26	26.23	1.67
N_goodspeedii	2.07	80.68	0.12	3.30	7.93	25.80	1.60
N_gossei	2.36	80.85	0.12	4.39	10.18	23.91	1.22
N_ingulba	2.42	76.77	0.05	3.89	7.46	21.60	1.64
N_knightiana	2.30	79.80	0.12	3.72	8.33	26.76	1.36
N_maritima	2.66	84.32	0.02	1.56	10.91	39.19	2.29
N_megalosiphon	2.29	80.09	0.04	3.56	15.58	19.33	2.03
N_obtusifolia	1.22	78.29	0.14	3.91	15.03	18.53	2.17
N_occidentalis	2.22	76.78	0.04	2.55	5.77	34.00	1.88
N_otophora	2.69	75.28	0.04	3.02	9.71	21.85	1.13
N_paniculata	2.19	78.24	0.03	1.32	4.30	35.05	1.36
*N_rosulata**	2.48	85.68	0.00	1.50	7.81	41.91	1.67
N_rotundifolia	2.11	80.08	0.12	4.34	9.61	23.79	2.37
N_rustica	4.23	77.17	0.17	4.28	14.63	19.96	1.11
N_simulans	1.84	80.10	0.01	1.95	7.64	35.72	1.65
N_sylvestris	2.22	80.65	0.16	3.42	10.44	24.50	1.33
N_tabacum	4.69	76.68	0.03	1.49	10.56	34.63	1.09
N_tomentosiformis	1.69	70.50	0.02	1.14	8.54	30.78	1.38
*N_umbratica**	2.46	79.91	0.02	1.09	8.70	39.38	1.03
N_undulata	1.91	85.51	0.19	3.82	9.12	29.10	1.51

REPlotCounts permitió visualizar la comparación de los perfiles de ET entre las distintas especies y evaluar la semejanza entre ellos, así como evidenciar cómo el análisis estadístico se ve afectado por la complejidad de los datos aportados. Además, se observan otros hechos de cierta relevancia.

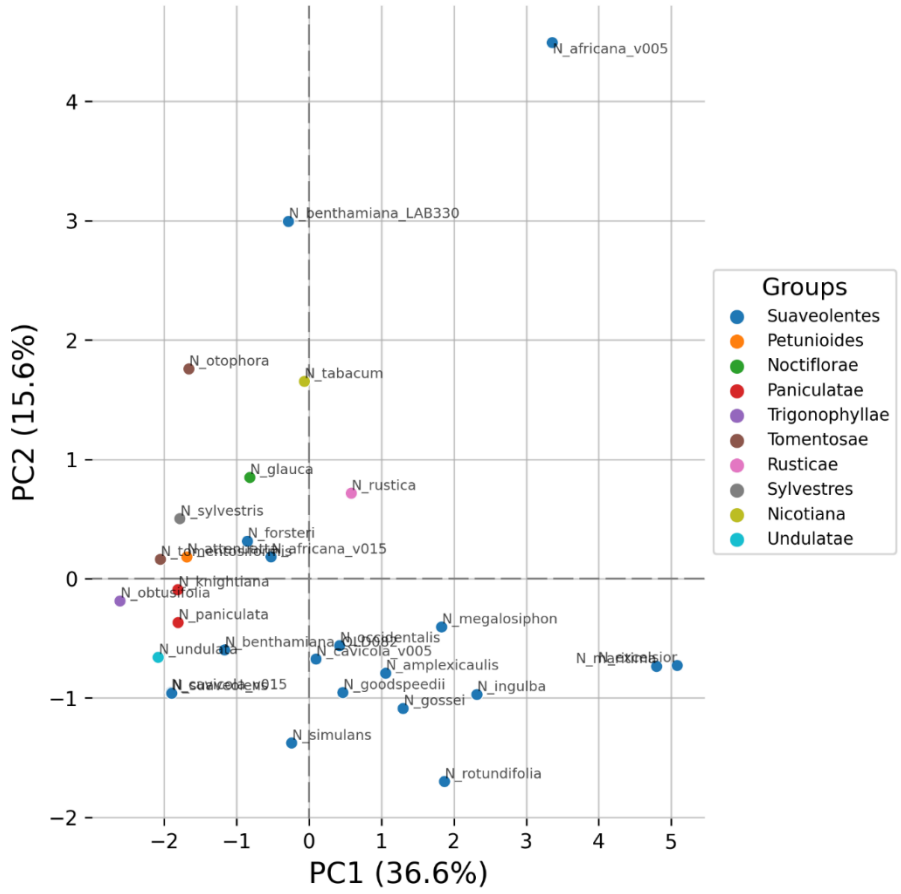

Figura 4.2. Gráfico de dispersión para el PCA del nivel de clasificación de RECollector "subclase". Una leyenda acompaña la gráfica en la que se detalla la separación por colores de las secciones de *Nicotiana*. Cada punto tiene el nombre de la especie a la que representa. Cada componente principal (PC) tiene asociado su porcentaje de varianza.

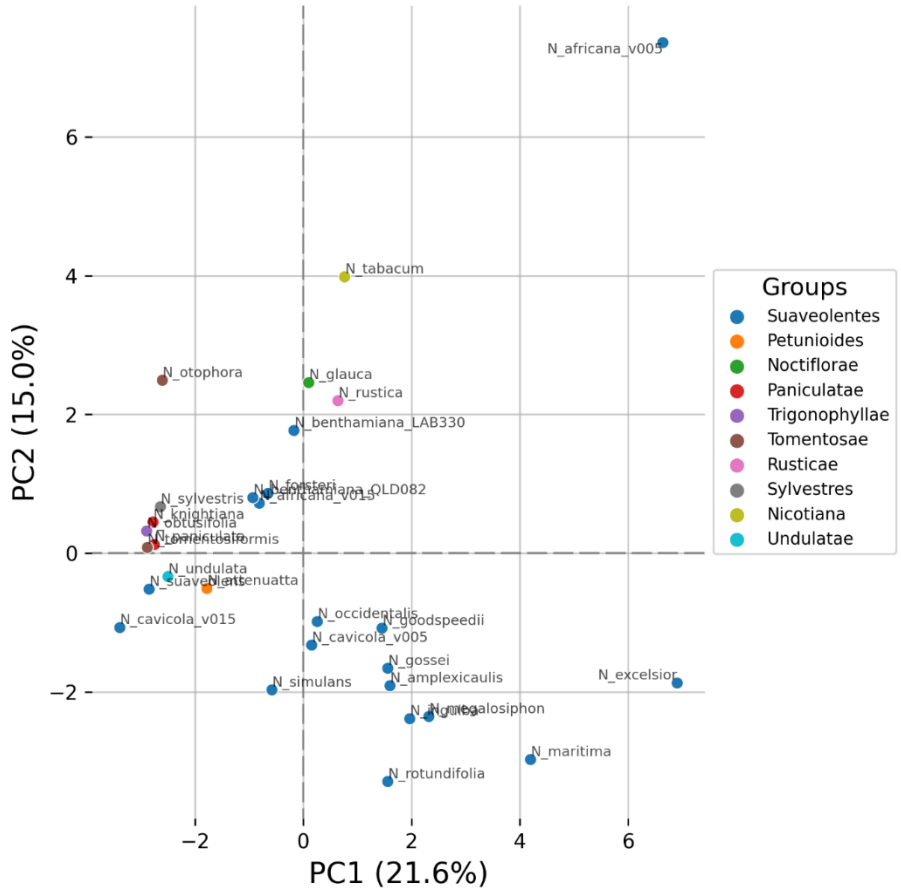

Figura 4.3. Gráfico de dispersión para el PCA del nivel de clasificación de RECollector "superfamilia". Una leyenda acompaña la gráfica en la que se detalla la separación por colores de las secciones de *Nicotiana*. Cada punto tiene el nombre de la especie a la que representa. Cada componente principal (PC) tiene asociado su porcentaje de varianza.

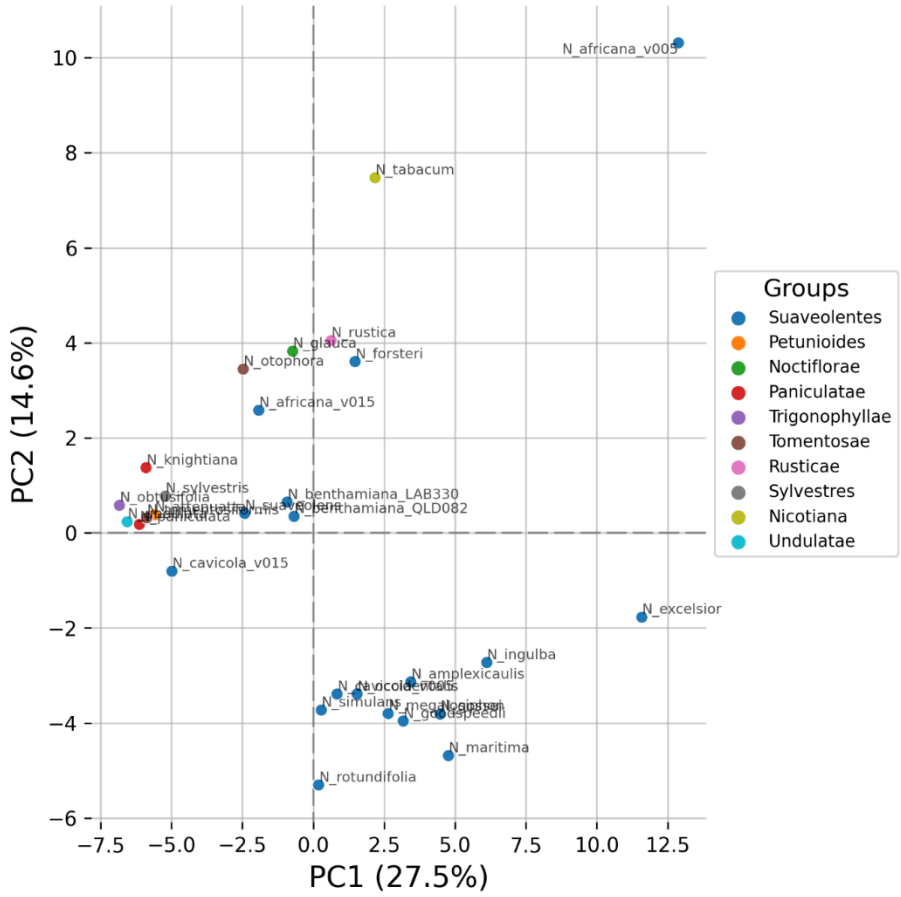

Figura 4.4. Gráfico de dispersión para el PCA del nivel de clasificación de RECollector "elemento". Una leyenda acompaña la gráfica en la que se detalla la separación por colores de las secciones de *Nicotiana*. Cada punto tiene el nombre de la especie a la que representa. Cada componente principal (PC) tiene asociado su porcentaje de varianza.

En general, se observa que las especies de la sección *Suaveolentes* se encuentran ampliamente distribuidas por la componente principal 1 (PC1) de las representaciones gráficas de los PCA para los niveles "subclase" (Fig. 4.2), "superfamilia" (Fig. 4.3) y "elemento" (Fig. 4.4). Por otro lado, la componente principal 2 (PC2) también presenta este rasgo, aunque en este se distinguen dos grupos separados entre la parte positiva y negativa de la componente, teniendo en cuenta que estas agrupaciones se diferencian mejor en las representaciones de "superfamilia" (Fig. 4.3) y "elemento" (Fig. 4.4).

De este modo, teniendo en cuenta las dos componentes principales, dentro de la sección *Suaveolentes* parece apreciarse un subconjunto de especies con perfiles de ET similares en el cuadrante inferior derecho de las representaciones gráficas de los PCA, sobre todo en aquellas a nivel de "superfamilia" y "elemento". Este subconjunto está integrado por *N. gossei, N. goodspeedii, N. megalosiphon, N. cavicola* (v005), *N. amplexicaulis, N. occidentalis, N. ingulba, N. rotundifolia, N. maritima* y *N. simulans*. Todas las especies son nativas del territorio australiano. No obstante, como ambas componentes principales representan menos del 50% de la variabilidad de los datos, únicamente se puede establecer este subconjunto de forma general. De este modo, el PCA no permite determinar relaciones más estrechas.

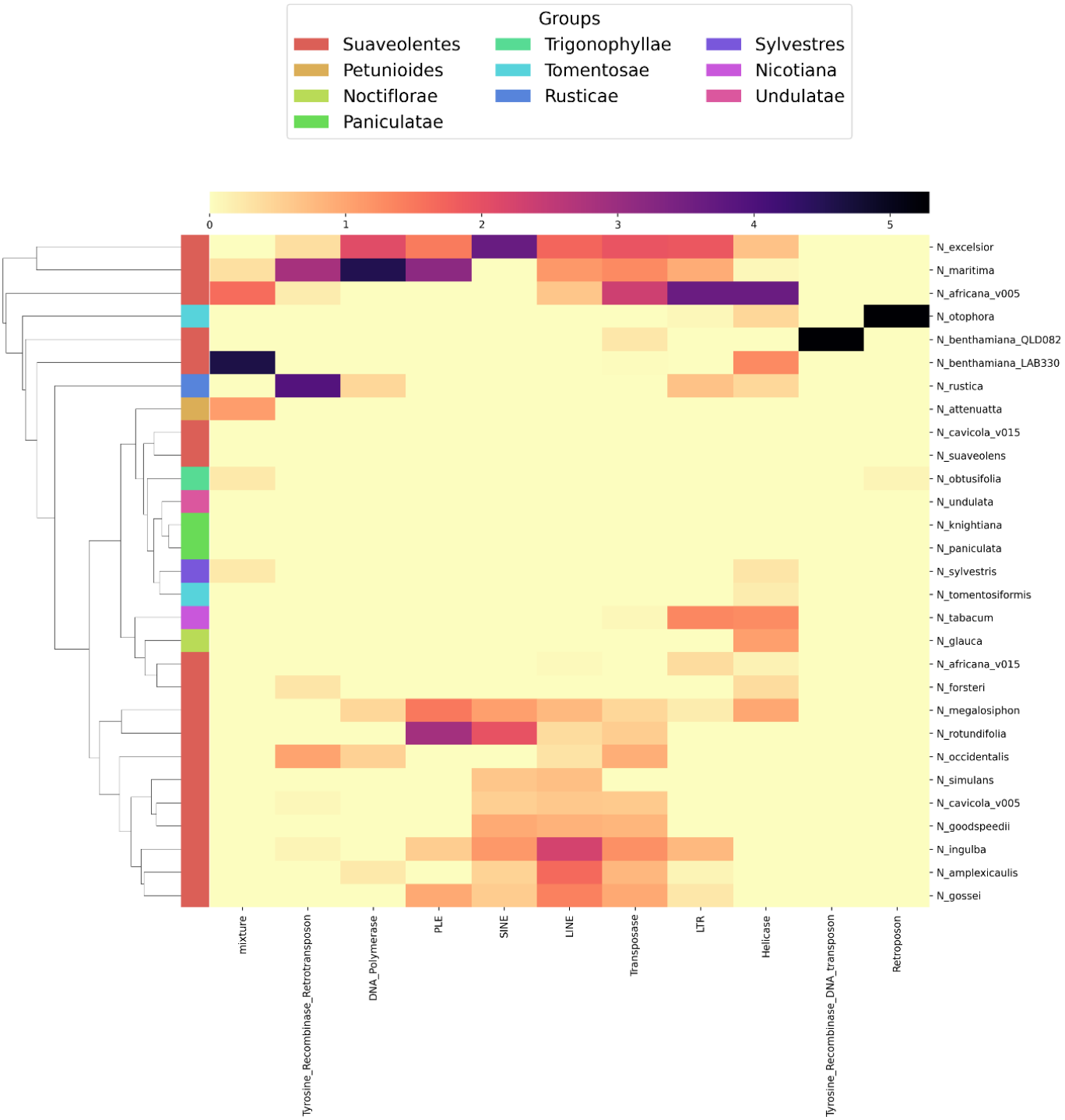

Figura 4.5. Mapa de calor normalizado del nivel de clasificación de RECollector "subclase". Una leyenda acompaña la gráfica en la que se detalla la separación por colores de las secciones de *Nicotiana*. Un dendrograma indica las agrupaciones entre especies. Colores claros indican baja o nula presencia de un ET en la especie dada; colores oscuros indican alta presencia.

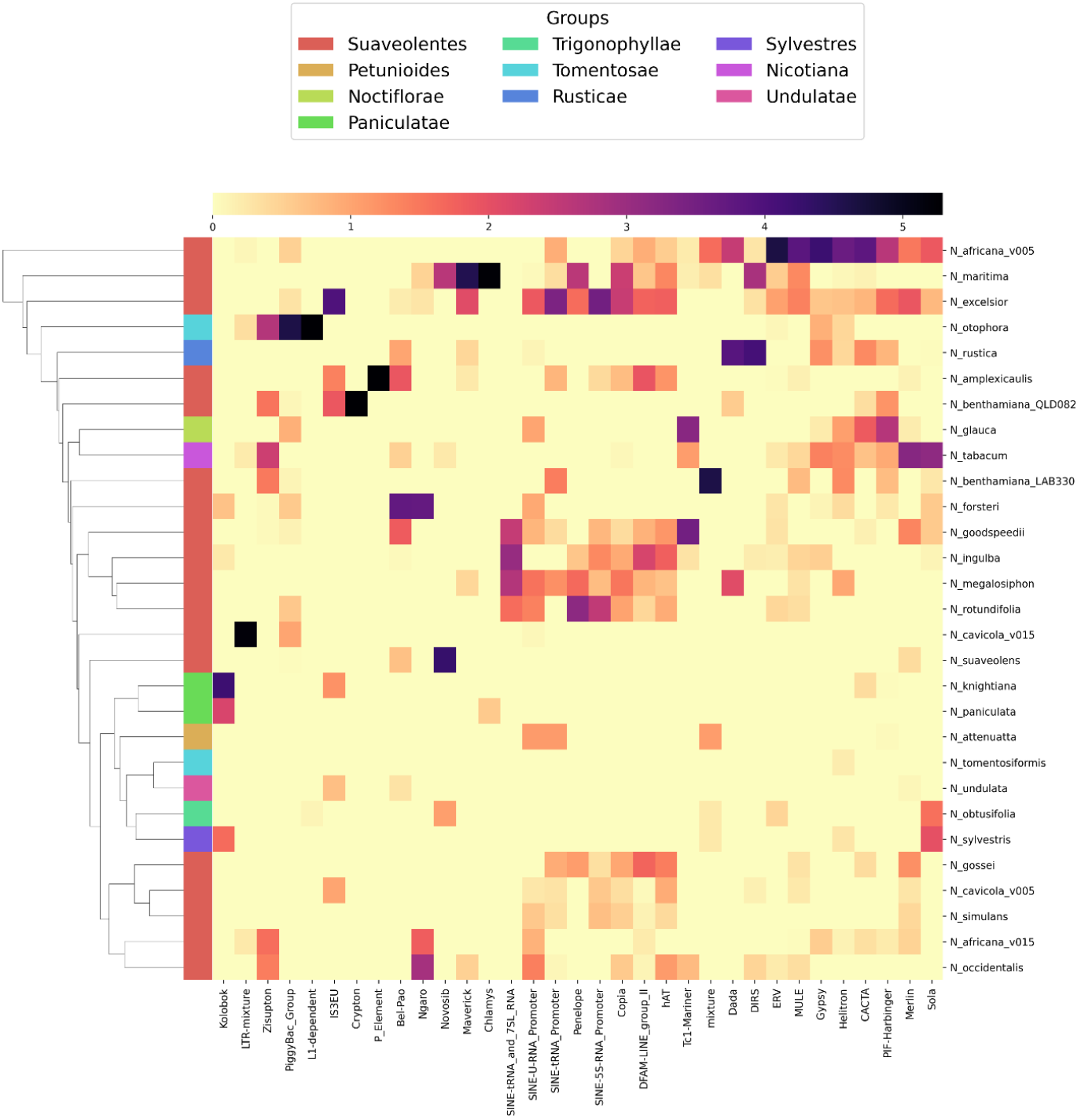

Figura 4.6. Mapa de calor normalizado del nivel de clasificación de RECollector "superfamilia". Una leyenda acompaña la gráfica en la que se detalla la separación por colores de las secciones de *Nicotiana*. Un dendrograma indica las agrupaciones entre especies. Colores claros indican baja o nula presencia de un ET en la especie dada; colores oscuros indican alta presencia.

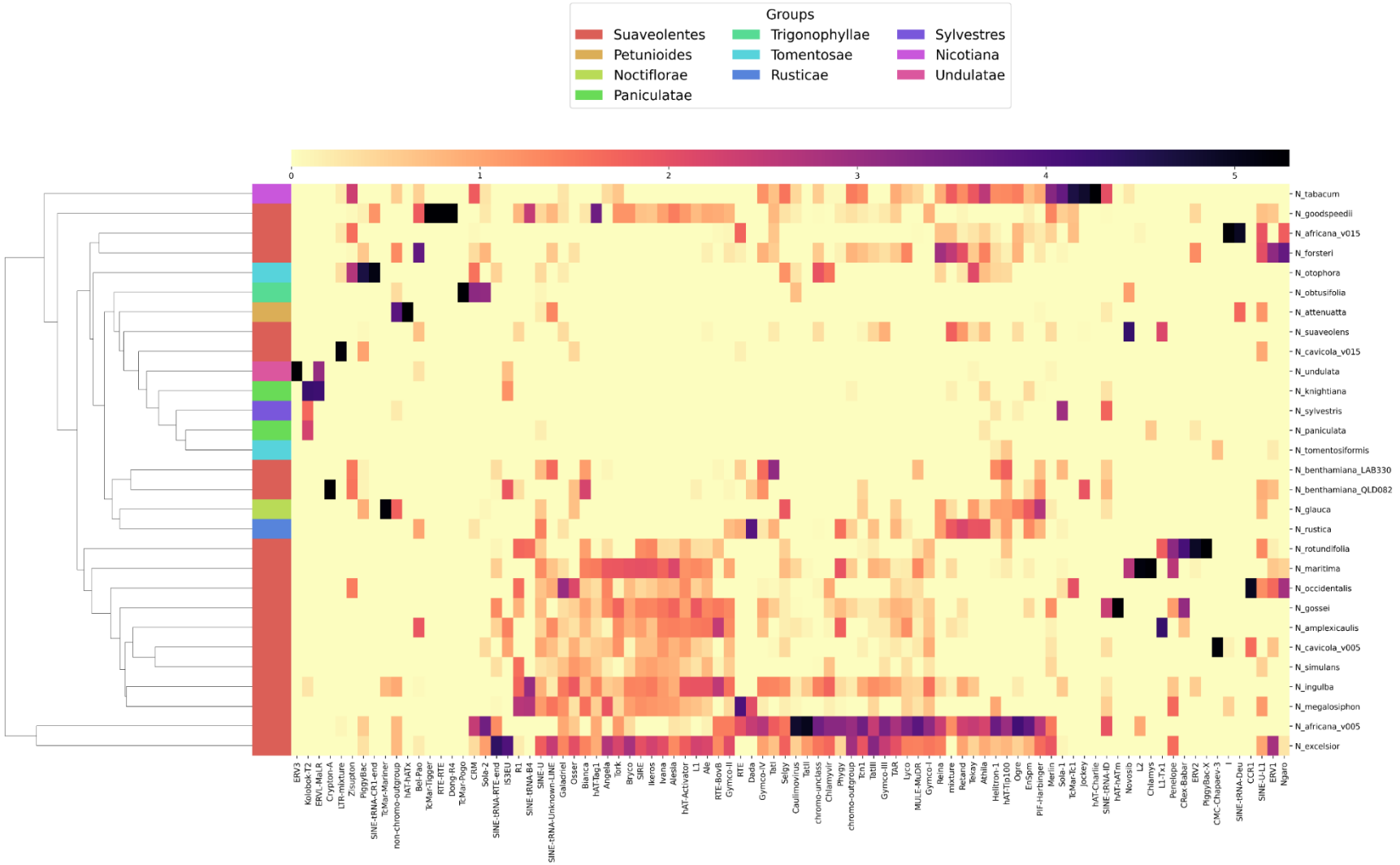

Figura 4.7. Mapa de calor normalizado del nivel de clasificación de RECollector "elemento". Una leyenda acompaña la gráfica en la que se detalla la separación por colores de las secciones de *Nicotiana*. Un dendrograma indica las agrupaciones entre especies. Colores claros indican baja o nula presencia de un ET en la especie dada; colores oscuros indican alta presencia

Los mapas de calor refuerzan este razonamiento del subconjunto de *Suaveolentes*. El mapa de calor para el perfil de ET sobre el nivel de "subclase" (Fig. 4.5) agrupa las especies mencionadas anteriormente (a excepción de *N. maritima*), señalando que sus perfiles de ET tienen mayoritariamente en común elementos SINE, LINE y *Tranposase*. Los dos primeros grupos corresponden a elementos de clase I mientras que el tercero corresponde a ciertos elementos de clase II. También puede observarse como el mapa de calor sobre el nivel "superfamilia" (Fig. 4.6) separa a los miembros del subconjunto, dejando únicamente agrupados a *N. goodspeedii, N. megalosiphon, N. ingulba* y *N. rotundifolia*. En este grupo destacan los elementos *Copia, hAT*, *Penelope* y ciertas familias de SINE; hecho en el que coinciden el resto de las especies del subconjunto original. Finalmente, el mapa de calor del nivel "elemento" (Fig. 4.7) vuelve a agrupar a todos los integrantes del subconjunto, a excepción de *N. goodspeedii*, si bien presenta un perfil similar en ciertos elementos. De nuevo, la huella propia de este grupo se centra en los elementos mencionados anteriormente, con una amplia presencia de diversos clados de la superfamilia *Copia*. *N. excelsior* resulta interesante al no tener un perfil de ET similar al del resto de integrantes de la sección *Suaveolentes*. Sin embargo, *N. excelsior* también presenta una mayor abundancia de elementos *Copia* y SINE, así como de elementos *Gypsy* y otros de clase II.

El resto de las especies de *Suaveolentes*, *N. africana*, *N. benthamiana* y *N. forsteri*, presentan una huella distinta. *N. africana* es totalmente diferente al resto de especies tanto de *Suaveolentes* como de las otras secciones. Esto evidencia que la huella del subconjunto australiano es única a este con respecto del integrante africano de la sección. *N. benthamiana* y *N. forsteri*, especies nativas de Australia, poseen una huella semejante a la de los descendientes de los progenitores de *Suaveolentes*, *N. glauca* y *N. sylvestris*. Este hecho deja claro que la huella del subconjunto excluye a especies también australianas. Una explicación puede ser que la amplificación de los ET identificativos se produjese después de o durante la especiación de estas dos especies.

Se puede destacar también la disparidad entre las dos versiones de *N. africana* y *N. cavicola*. Las versiones v015 de estas especies fueron sometidas a una ronda de *scaffolding* con el *software* Ragoo v1.1, que utilizó como genoma de referencia el de *N. benthamiana*. De este modo, este proyecto también pone de manifiesto que el uso de un genoma de referencia para mejorar la calidad de un ensamblaje genera sesgos importantes que pueden conducir a resultados ambiguos e incluso erróneos. Por ello, se decidió excluir del análisis posterior con REPlotDivergence las versiones 015 de *N. africana* y *N. africana*.

Con todo ello, la evidencia apunta a una diferencia interna en cuanto a perfiles de ET en la sección *Suaveolentes*. Así, encontraríamos un grupo de especies (todas ellas pertenecientes a la subsección australiana) caracterizado por una mayor abundancia de elementos *Copia*, *hAT* y SINE, entre otros, y otro grupo con una menor presencia de estos ET.

4.3. La sección australiana de *Suaveolentes* ha sufrido recientemente crecimientos explosivos de Elementos Transponibles

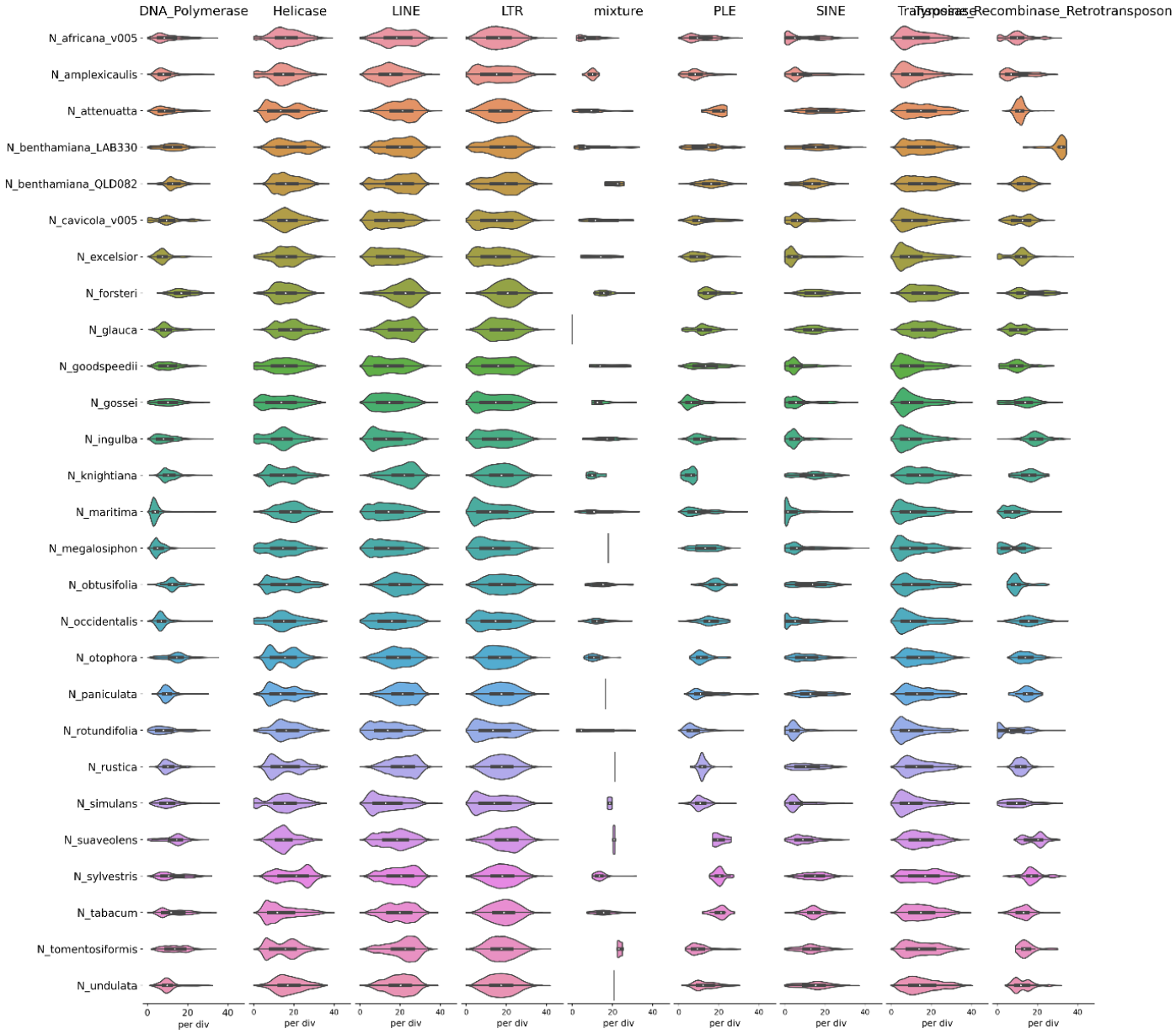

Figura 4.8. Diagramas de violín de REPlotDivergence para el nivel "subclase". Las especies están ordenadas por orden alfabético. El eje de las abscisas indica el porcentaje de divergencia.

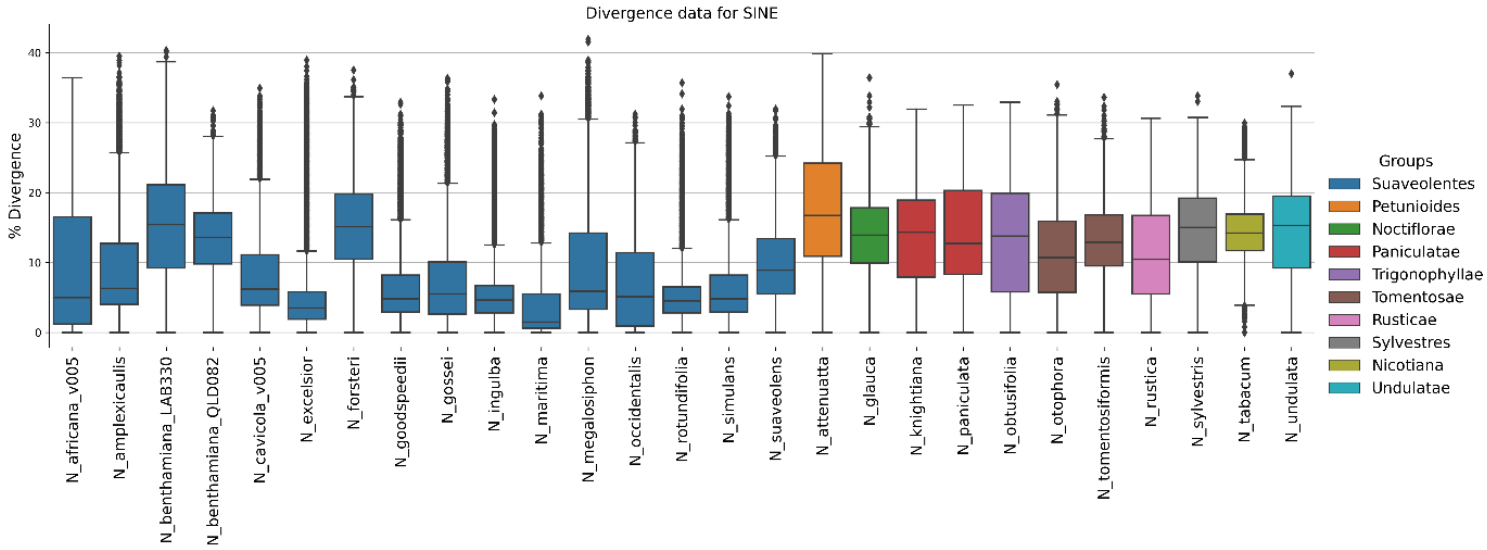

Figura 4.9. Diagrama de cajas para el elemento SINE; las especies se encuentran separadas por sección. El eje de las ordenadas indica el porcentaje de divergencia

Otro de los análisis que queríamos conducir sobre los elementos repetitivos de *Nicotiana* era la caracterización temporal de cuándo se han expandido las diferentes familias de elementos de manera que podamos relacionarlas con los distintos eventos de especiación y diversificación. Para ello se aplicó REPlotDivergence.

Los diagramas de violín ofrecen una visión general de posibles expansiones y contracciones estudiando el grado de divergencia entre los elementos. Esto permite que el usuario sea capaz de detectar ET que cumplan ciertos criterios con mayor facilidad, mientras que los diagramas de cajas aportan una visión individual y detallada de cada ET. En nuestro caso, continuando con la hipótesis formulada al final del apartado 4.2, intentamos buscar diferencias internas entre las especies de la sección *Suaveolentes* en lo que respecta a porcentajes de divergencia de las copias de los diferentes ET.

Observando el análisis a nivel de "subclase" (Fig. 4.8), se descubre que las especies del subconjunto de *Suaveolentes* (incluyendo también a *N. excelsior*) presentan un mayor número de elementos SINE de reciente formación en comparación con el resto de las especies de la sección y con el resto de las secciones, que presentan distribuciones con porcentajes de divergencia más dispersos. Este hecho queda mejor representado en el diagrama de cajas (Fig. 4.9), con el que se comprende mejor la evidencia gracias a la separación de las especies por secciones. En conjunto, ambos diagramas muestran cómo 11 de las 15 especies de *Suaveolentes* analizadas presentan lo que podría considerarse una expansión reciente de elementos SINE.

Una observación adicional respecto del nivel "subclase" (Fig. 4.8) es la presencia de dos picos pronunciados en los diagramas de violín para *Helicase* de *N. otophora*, *N. tomentosiformis*, *N. attenuata*, *N. paniculata, N. knightiana, N. obtusifolia y N. rustica*. Esto sugiere que estas especies pueden haber sufrido dos eventos de expansión de ET. *Helicase* corresponde a un subgrupo de retrotransposones que se replican por replicación de círculo rodado. Dentro de este subgrupo de ET figuran los elementos *Helitron*. En el caso de *N. otophora* y *N. tomentosiformis*

(pertenecientes a la sección Tomentosae), existe la posibilidad de que las copias de ET identificadas como *Helicase* correspondan a secuencias relacionadas con geminivirus (Murad et al., 2004), que estructuralmente se asemejan a elementos *Helitron*. Sin embargo, este hecho contradiría lo observado por Murad et al. (2004), ya que la inserción de estas secuencias se produjo después de la especiación de *N. otophora*, por lo que solo deberían encontrarse en *N. tomentosiformis*. Curiosamente, *N. tabacum*, descendiente de los ancestros de las actuales especies *N. tomentosiformis* y *N. sylestris*, únicamente presenta un posible evento de expansión. Posiblemente, estos ET procedan del genoma del ancestro de *N.* tomentosiformis. Además, parte de estos ET, en concreto los pertenecientes al evento de expansión más antiguo de *N. Tomentosiformis*, han sido eliminados del genoma de *N. tabacum* debido a la eliminación preferencial de secuencias del ancestro de *N. tomentosiformis* (Skalická et al., 2005). Se desconoce el origen de los supuestos eventos de expansión en el resto de las especies.

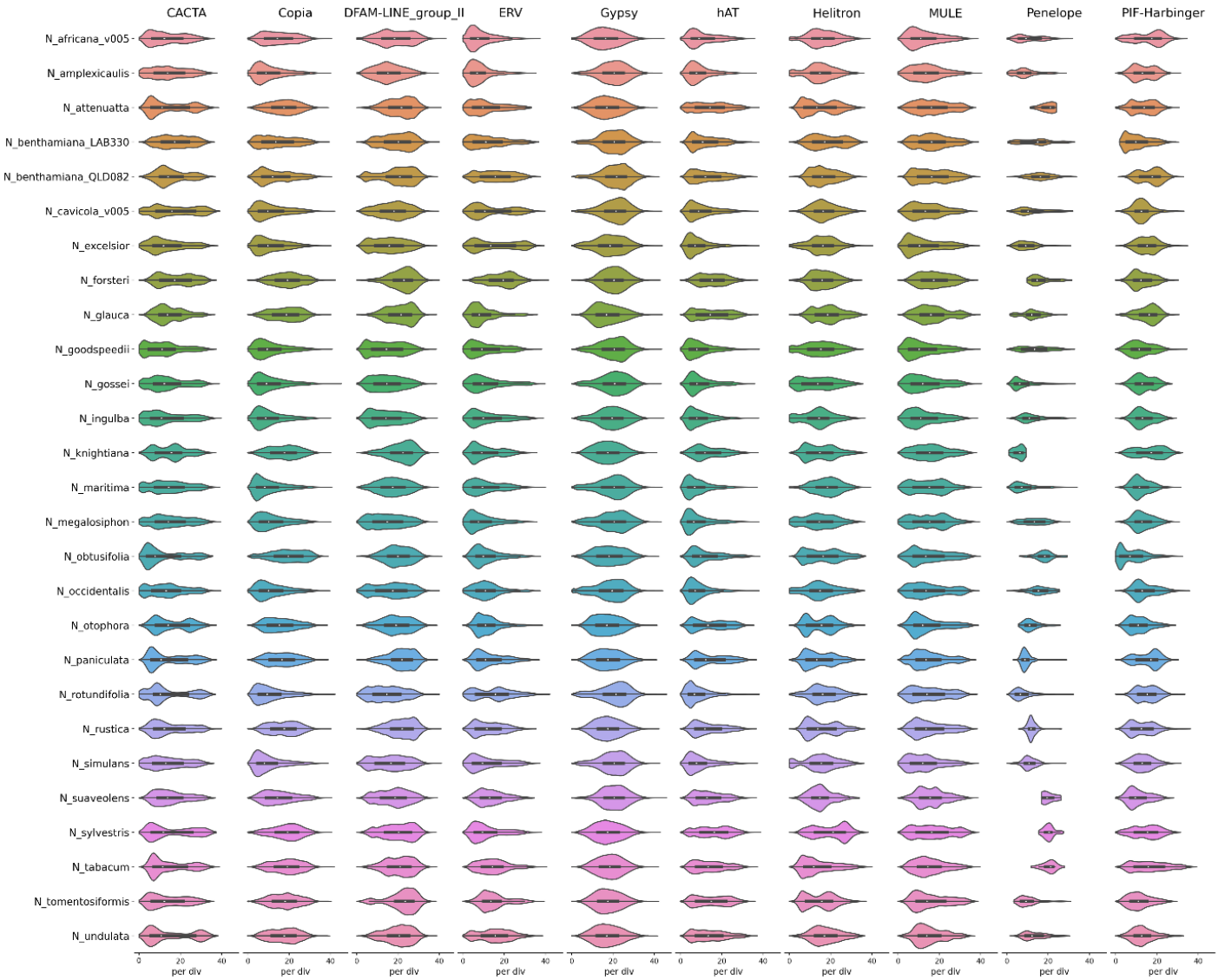

Figura 4.10. Diagramas de violín de REPlotDivergence para el nivel "superfamilia". Las especies están ordenadas por orden alfabético. El eje de las abscisas indica el porcentaje de divergencia.

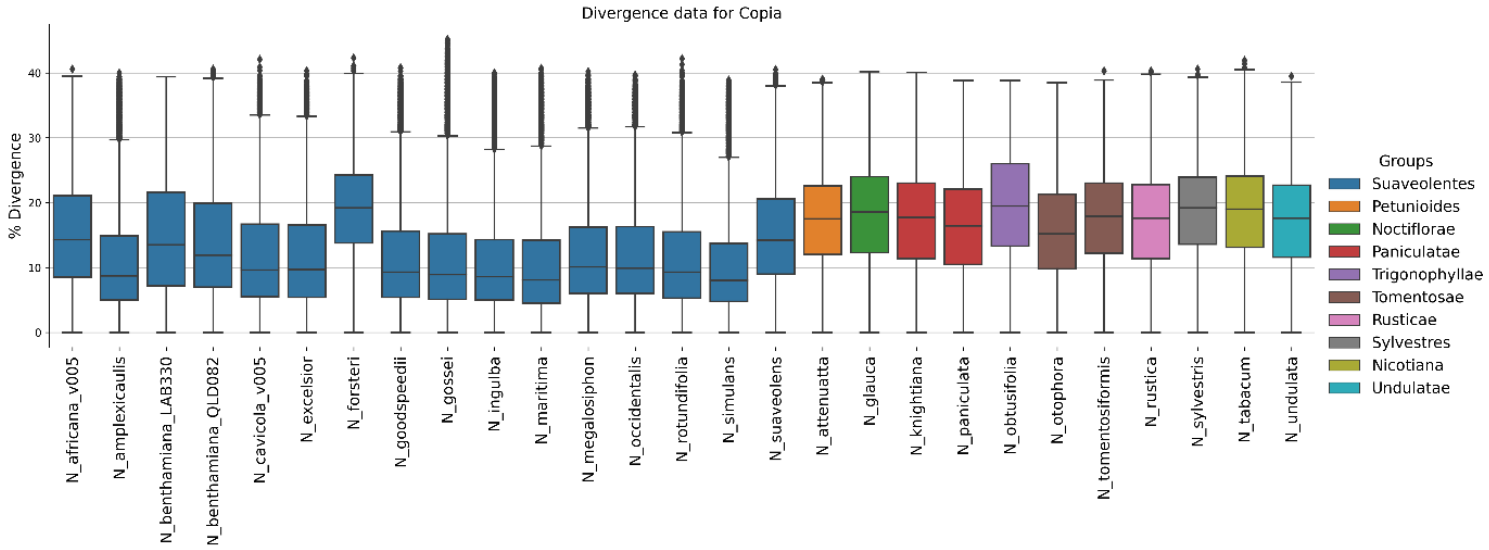

Figura 4.11. Diagrama de cajas para el elemento *Copia*; las especies se encuentran separadas por sección. El eje de las ordenadas indica el porcentaje de divergencia.

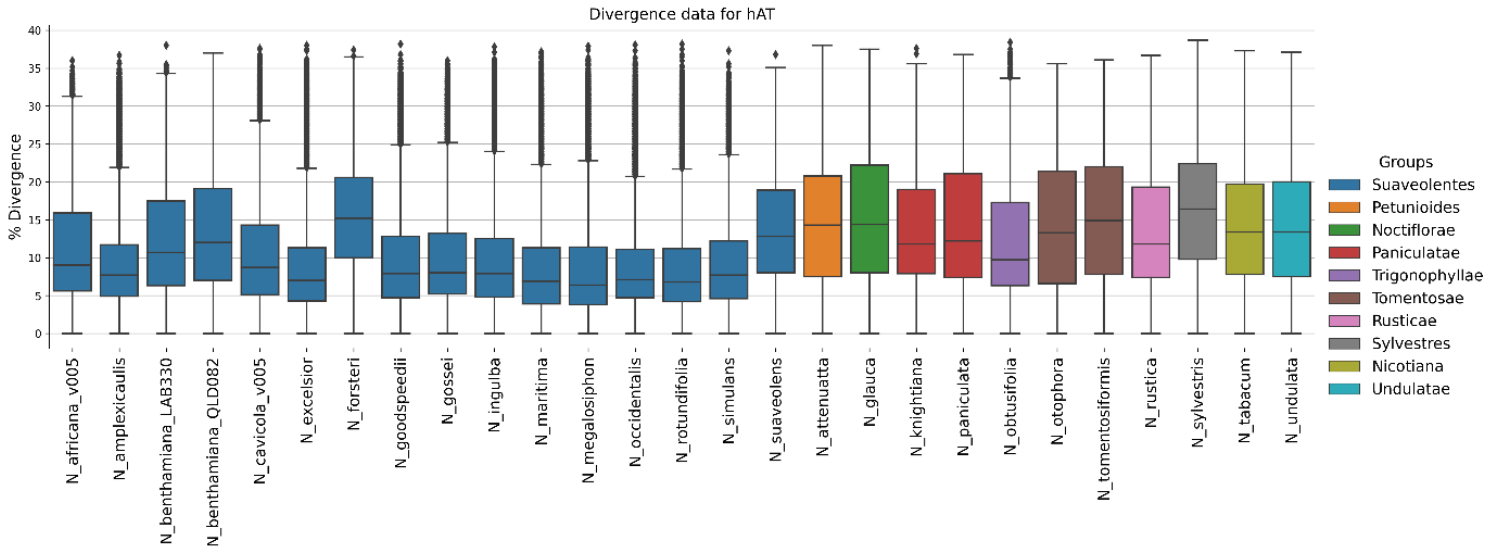

Figura 4.12. Diagrama de cajas para el elemento *hAT*; las especies se encuentran separadas por sección. El eje de las ordenadas indica el porcentaje de divergencia.

Continuando con el análisis del nivel "superfamilia" (Fig. 4.10), se pueden observar diferencias entre las especies de *Suaveolentes* y el resto de las secciones para los elementos *Copia* y *hAT*, principalmente. Los diagramas de cajas (Figs. 4.11 y 4.12) permiten ahondar en esta observación y afirmar que existen variaciones en las distribuciones de divergencia dentro de *Suaveolentes*. Además, se puede también sugerir que ha existido una explosión en el número de copias de los elementos *Copia* y *hAT* para las especies del subconjunto mencionado en el

subapartado anterior. También se puede confirmar parcialmente la hipótesis de Ranawaka et al. (2023) sobre la expansión reciente de elementos *Copia* en las especies australianas de la sección *Suaveolentes*, pero no se percibe una expansión tan aparente de los mismos elementos en *N. benthamiana*. Una posible explicación podría ser que esta especie comenzó su especiación más tempranamente, cuando los elementos *Copia* todavía continuaban en proceso de expansión.

Además, es interesante la ligera diferencia entre las dos accesiones analizadas de *N. benthamiana*, QLD (Queensland) y LAB (*Laboratory*), tanto en su perfil de ET como en la distribución de los datos de divergencia de los distintos ET. Esto puede responder al hecho de que ambas accesiones divergieron hace 880000 años (Bally et al., 2015). La evolución de las diferentes accesiones en sus respectivos territorios ha dado lugar a estas sutiles diferencias.

En general, considerando todos los hechos y evidencias, existen indicios que demuestran que las especies australianas de *Suaveolentes* han experimentado eventos de expansión de ciertos ET, eventos no presentes en otras especies de la sección, que se presume divergieron más temprano (D'Andrea et al., 2023). La principal hipótesis que se puede plantear es, pues, que la expansión de ciertos ET pudo haber ayudado a la rápida diversificación de la sección *Suaveolentes* por el territorio australiano. Esta hipótesis también cobra fuerza si tenemos en cuenta que la especie africana de la sección, *N. africana*, no presenta estos eventos de expansión. Sin embargo, su validez como argumento a favor se ve reducido por el simple hecho de existir en otro entorno distinto en el que se desconoce si dichos crecimientos explosivos de ET hubiesen tenido las mismas consecuencias.

Existen evidencias y ejemplos de cómo los crecimientos explosivos de ET pueden estar relacionados con eventos de diversificación de especies. Además, los ET pueden tener efectos funcionales en los huéspedes, tal y como fue discutido en la introducción. Por ejemplo, Kozlowski et al. (2021) sugirieron que la actividad de ET pudo estar involucrada en la evolución del nematodo parasítico *Meloidogyne incognita*. También observaron que ciertas inserciones de ET se encontraban en genes específicos de plantas, por lo que estos elementos han podido estar involucrados en las capacidades parasitarias de la especie. Latzel et al. (2023) observaron que diversidad poblacional de los retrotransposones ONSEN en *Arabidopsis thaliana* contribuye a la mejora de aspectos funcionales (entre otros, biomasa y longitud de las raíces) de los individuos de la especie y a una disminución en el rendimiento de otras especies competidoras. Los eventos de poliploidía e hibridación posibilitan los eventos de expansión de ET y con ello la probabilidad de inserciones en genes del organismo huésped que tengan consecuencias a nivel de fenotipo (Vicient & Casacuberta, 2017). Con ello, es posible que la expansión de ciertos ET (*Copia*, *hAT* y SINE) en el ancestro común a las especies australianas de *Suaveolentes* (previa especiación de *N. forsteri*) pudiera haber afectado a genes clave para la supervivencia del organismo. Dichas inserciones pudieron tener efectos funcionales potencialmente favorecedores en el entorno árido de Australia hace varios millones de años, que promovieron la expansión y diversificación de la sección por todo el territorio.

5. CONCLUSIÓN

Este proyecto ha dado lugar a la creación de un paquete de herramientas capaz de analizar los datos procedentes de varios *softwares* de detección y anotación de elementos transponibles con el objetivo de analizar la diversidad de estos entre varias especies y contribuir a caracterizar diferentes especies en base a sus perfiles de elementos transponibles, teniendo un potencial uso para el campo de la genómica y, en concreto, el de la genómica evolutiva y poblacional.

Sin embargo, el paquete Repeattools todavía tiene margen de mejora. La primera sugerencia sería crear todos los archivos de salida de RECollector para los cuatro niveles del sistema de clasificación en una única operación. Además, el paquete se podría integrar en herramientas como Docker, lo que facilitaría su uso al hacer más simple la instalación. Asimismo, resultaría conveniente plantear otros análisis estadísticos que complementen o proporcionen más información que los actuales sin perder de vista el objetivo del paquete de herramientas. Por último, se podrían integrar de una forma más eficiente los datos filogenéticos. Se mencionó anteriormente que REPlotDivergence admitía archivos filogenéticos en formato Newick, pero esta integración con el resto de la imagen generada requiere de más pruebas.

Con respecto al análisis de las especies del género *Nicotiana*, se deben mencionar varios hechos de cierta importancia. El análisis ha permitido dejar claro que utilizar genomas de referencia para mejorar el ensamblaje de los genomas de otras especies es perjudicial para la búsqueda y análisis de elementos transponibles. Este hecho debería ser seriamente considerado para futuros análisis con tal de minimizar sesgos y conclusiones erróneas. Además, todavía queda por resolver el periodo de tiempo de los estallidos de ET para confirmar verdaderamente si estos elementos del genoma estuvieron involucrados en la diversificación de la subsección australiana de *Suaveolentes*.

En resumen, Repeattools constituye una herramienta útil para el análisis de elementos transponibles en el contexto de la genómica evolutiva y poblacional, tal y como se ha demostrado con el ejemplo del género *Nicotiana*, ayudando a la comprensión de cómo los elementos transponibles pueden tener implicaciones a nivel evolutivo.

6. BIBLIOGRAFÍA

Almeida, M. V., Vernaz, G., Putman, A. L. K., & Miska, E. A. (2022). Taming transposable elements in vertebrates: From epigenetic silencing to domestication. *Trends in Genetics*, *38*(6), 529-553. https://doi.org/10.1016/j.tig.2022.02.009

Bally, J., Marks, C. E., Jung, H., Jia, F., Roden, S., Cooper, T., Newbigin, E., & Waterhouse, P. M. (2021). Nicotiana paulineana, a new Australian species in Nicotiana section Suaveolentes. *Australian Systematic Botany*, *34*(5), 477-484. https://doi.org/10.1071/SB20025

Bally, J., Nakasugi, K., Jia, F., Jung, H., Ho, S. Y. W., Wong, M., Paul, C. M., Naim, F., Wood, C. C., Crowhurst, R. N., Hellens, R. P., Dale, J. L., & Waterhouse, P. M. (2015). The extremophile Nicotiana benthamiana has traded viral defence for early vigour. *Nature Plants*, *1*(11), 1-6. https://doi.org/10.1038/nplants.2015.165

Clarkson, J. J., Dodsworth, S., & Chase, M. W. (2017). Time-calibrated phylogenetic trees establish a lag between polyploidisation and diversification in Nicotiana (Solanaceae). *Plant Systematics and Evolution*, *303*(8), 1001-1012. https://doi.org/10.1007/s00606-017-1416-9

D'Andrea, L., Sierro, N., Ouadi, S., Hasing, T., Rinaldi, E., Ivanov, N. V., & Bombarely, A. (2023). Polyploid Nicotiana section Suaveolentes originated by hybridization of two ancestral Nicotiana clades. *Frontiers in Plant Science*, *14*, 999887. https://doi.org/10.3389/fpls.2023.999887

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

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451-9457. https://doi.org/10.1073/pnas.1921046117

Fultz, D., & Slotkin, R. K. (2017). Exogenous Transposable Elements Circumvent Identity-Based Silencing, Permitting the Dissection of Expression-Dependent Silencing. *The Plant Cell*, *29*(2), 360-376. https://doi.org/10.1105/tpc.16.00718

Goerner-Potvin, P., & Bourque, G. (2018). Computational tools to unmask transposable elements. *Nature Reviews Genetics*, *19*(11), 688-704. https://doi.org/10.1038/s41576-018-0050-x

Kelly, L. J., Leitch, A. R., Clarkson, J. J., Knapp, S., & Chase, M. W. (2013). RECONSTRUCTING THE COMPLEX EVOLUTIONARY ORIGIN OF WILD ALLOPOLYPLOID TOBACCOS (NICOTIANA SECTION SUAVEOLENTES). *Evolution*, *67*(1), 80-94. https://doi.org/10.1111/j.1558-5646.2012.01748.x

Knapp, S., Chase, M. W., & Clarkson, J. J. (2004). Nomenclatural changes and a new sectional classification in Nicotiana (Solanaceae). *TAXON*, *53*(1), 73-82. https://doi.org/10.2307/4135490

Kozlowski, D. K. L., Hassanaly-Goulamhoussen, R., Da Rocha, M., Koutsovoulos, G. D., Bailly-Bechet, M., & Danchin, E. G. J. (2021). Movements of transposable elements contribute to the genomic plasticity and species diversification in an asexually reproducing nematode pest. *Evolutionary Applications*, *14*(7), 1844-1866. https://doi.org/10.1111/eva.13246

Latzel, V., Puy, J., Thieme, M., Bucher, E., Götzenberger, L., & de Bello, F. (2023). Phenotypic diversity influenced by a transposable element increases productivity and resistance to competitors in plant populations. *Journal of Ecology*, *111*(11), 2376-2387. https://doi.org/10.1111/1365-2745.14185

Li, Y., Li, C., Xia, J., & Jin, Y. (2011). Domestication of Transposable Elements into MicroRNA Genes in Plants. *PLOS ONE*, *6*(5), e19212. https://doi.org/10.1371/journal.pone.0019212

Lin, R., Ding, L., Casola, C., Ripoll, D. R., Feschotte, C., & Wang, H. (2007). Transposase-Derived Transcription Factors Regulate Light Signaling in Arabidopsis. *Science*, *318*(5854), 1302-1305. https://doi.org/10.1126/science.1146281

Lisch, D., & Slotkin, R. K. (2011). Chapter Three - Strategies for Silencing and Escape: The Ancient Struggle Between Transposable Elements and Their Hosts. En K. W. Jeon (Ed.), *International Review of Cell and Molecular Biology* (Vol. 292, pp. 119-152). Academic Press. https://doi.org/10.1016/B978-0-12-386033-0.00003-7

Liu, P., Cuerda-Gil, D., Shahid, S., & Slotkin, R. K. (2022). The Epigenetic Control of the Transposable Element Life Cycle in Plant Genomes and Beyond. *Annual Review of Genetics*, *56*(1), 63-87. https://doi.org/10.1146/annurev-genet-072920-015534

Mhiri, C., Borges, F., & Grandbastien, M.-A. (2022). Specificities and Dynamics of Transposable Elements in Land Plants. *Biology*, *11*(4), 488. https://doi.org/10.3390/biology11040488

Murad, L., Bielawski, J. P., Matyasek, R., Kovarík, A., Nichols, R. A., Leitch, A. R., & Lichtenstein, C. P. (2004). The origin and evolution of geminivirus-related DNA sequences in Nicotiana. *Heredity*, *92*(4), 352-358. https://doi.org/10.1038/sj.hdy.6800431

Ramakrishnan, M., Satish, L., Sharma, A., Kurungara Vinod, K., Emamverdian, A., Zhou, M., & Wei, Q. (2022). Transposable elements in plants: Recent advancements, tools and prospects. *Plant Molecular Biology Reporter*, *40*(4), 628-645. https://doi.org/10.1007/s11105-022-01342-w

Ranawaka, B., An, J., Lorenc, M. T., Jung, H., Sulli, M., Aprea, G., Roden, S., Llaca, V., Hayashi, S., Asadyar, L., LeBlanc, Z., Ahmed, Z., Naim, F., de Campos, S. B., Cooper, T., de Felippes, F. F., Dong, P., Zhong, S., Garcia-Carpintero, V., … Waterhouse, P. M. (2023). A multi-omic Nicotiana benthamiana resource for fundamental research and biotechnology. *Nature Plants*, *9*(9), 1558-1571. https://doi.org/10.1038/s41477-023-01489-8

Richardson, S. R., Doucet, A. J., Kopera, H. C., Moldovan, J. B., Garcia-Perez, J. L., & Moran, J. V. (2015). The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology Spectrum*, *3*(2), 10.1128/microbiolspec.mdna3-0061-2014. https://doi.org/10.1128/microbiolspec.mdna3-0061-2014

Seberg, O., & Petersen, G. (2009). A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews Genetics*, *10*(4), 276-276. https://doi.org/10.1038/nrg2165-c3

Skalická, K., Lim, K. Y., Matyasek, R., Matzke, M., Leitch, A. R., & Kovarik, A. (2005). Preferential elimination of repeated DNA sequences from the paternal, Nicotiana tomentosiformis genome donor of a synthetic, allotetraploid tobacco. *New Phytologist*, *166*(1), 291-303. https://doi.org/10.1111/j.1469-8137.2004.01297.x

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

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, S., Gao, J., Li, Z., Chen, K., Pu, W., & Feng, C. (2023). Phylotranscriptomics supports numerous polyploidization events and phylogenetic relationships in Nicotiana. *Frontiers in Plant Science*, *14*. https://doi.org/10.3389/fpls.2023.1205683

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

Wendel, J. F., Jackson, S. A., Meyers, B. C., & Wing, R. A. (2016). Evolution of plant genome architecture. *Genome Biology*, *17*(1), 37. https://doi.org/10.1186/s13059-016-0908-1

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), Article 12. https://doi.org/10.1038/nrg2165

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., & Ma, Y. (2022). TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research*, *9*, uhac017. https://doi.org/10.1093/hr/uhac017

Zhao, D., Ferguson, A. A., & Jiang, N. (2016). What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1859*(2), 366-380. https://doi.org/10.1016/j.bbagrm.2015.12.005

Zhu, Y. (2024). *Study of the evolution of the transposable element space in plant genomes* [Trabajo Final de Máster]. Universitat Politècnica de València.