	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015




UNIVERSIDAD AUTÓNOMA DE MANIZALES

VICERRECTORÍA ACADÉMICA

UNIDAD DE INVESTIGACIÓN

UNIDAD DE PREGRADO

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

TÓPICOS PARA LA PRESENTACIÓN DE INFORMES FINALES¹

UNIVERSIDAD AUTÓNOMA DE MANIZALES

PROYECTO: Diseño y construcción de un conjunto de datos de referencia de LTR retrotransposones presentes en plantas

GRUPO DE INVESTIGACIÓN: Ingeniería de Software y Automática


ESTUDIANTE: Paula Andrea Jaimes Buitrón

TUTOR DE TESIS: Simón Orozco Arias

DATOS DE IDENTIFICACIÓN:

AÑO: 2020


¹

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

RESUMEN

Los elementos transponibles (ET) son segmentos cortos del ADN que pueden moverse de un lugar a otro dentro del genoma. En diversos estudios se ha comprado sus roles importantes en la estructura del cromosoma, la expresión y regulación génica, así como también en la adaptación y evolución de las especies. La identificación y clasificación de estos elementos es un reto puesto que tienen una naturaleza repetitiva y una gran diversidad estructural. Sin embargo, la realización de estos procesos es crucial para entender de mejor manera las diferentes funciones del genoma y su evolución. En la actualidad se han desarrollado softwares bioinformáticos que han permitido la identificación y clasificación de los ET presentes en especies como las de plantas, los cuales requieren en su mayoría de una librería de ET conocidos. No obstante, existe una carencia de conjuntos de datos o librerías que contenga una cantidad adecuada de especies de plantas. Es por ello que en este proyecto se pretende realizar un conjunto de datos de ET de referencia, específicamente de LTR retrotransposones presentes en especies de plantas de diferentes familias, utilizando softwares bioinformáticos para la identificación y clasificación, para su posterior uso en futuros estudios.

PALABRAS CLAVES: Bioinformática, Elementos transponibles, conjunto de datos, LTR retrotransposones, genómica.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

ABSTRACT

Transposable elements (ET) are short segments of DNA that can move from one place to another within the genome. In various studies, their important roles in the structure of the chromosome, gene expression and regulation, as well as in the adaptation and evolution of species, have been confirmed. The identification and classification of these elements is a challenge since they have a repetitive nature and a great structural diversity. However, the performance of these processes is crucial to better understand the different functions of the genome and its evolution. At present, bioinformatics software has been developed that have allowed the identification and classification of the ETs present in species such as plants, most of which require a library of known ETs. However, there is a lack of data sets or libraries that contain an adequate number of plant species. That is why this project aims to make a reference ET data set, specifically of LTR retrotransposons present in plant species of different families, using bioinformatic software for identification and classification, for later use in future studies.

KEY WORDS: Bioinformatics, Transposable elements, dataset, LTR retrotransposons, genomics.



	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

TABLA DE CONTENIDO

1. PRESENTACIÓN	6
2. INTRODUCCIÓN	7
3. ÁREA PROBLEMÁTICA Y JUSTIFICACIÓN	8
4. REFERENTE TEÓRICO	9
5. LOS OBJETIVOS	13
6. METODOLOGÍA	14
7. RESULTADOS	15
8. DISCUSIÓN DE RESULTADOS	20
9. CONCLUSIONES	21
10. RECOMENDACIONES	22
11. EVIDENCIA DE RESULTADOS EN GENERACIÓN DE CONOCIMIENTO, FORTALECIMIENTO DE LA CAPACIDAD CIENTÍFICA Y APROPIACIÓN SOCIAL DEL CONOCIMIENTO, FORMACIÓN	23
12. IMPACTOS LOGRADOS	24
13. BIBLIOGRAFÍA	25
14. ANEXOS	30

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

1. PRESENTACIÓN


El equipo de investigación consta de dos ramas Bioinformática e Inteligencia Artificial, el presente proyecto trabaja bajo la línea de la bioinformática, la cual está conformada por:

Simón Orozco Arias, tutor principal del semillero. Es candidato a doctorado en Ingeniería en la Universidad de Caldas, ha desarrollado diversos proyectos, entre los que se encuentran: aplicación de técnicas de HPC para acelerar procesos bioinformáticos, análisis de las dinámicas y estructuras de elementos transponibles en diferentes especies de plantas (café robusta, café arábica, caña de azúcar, entre otros), en la roya del café y en la mosca de la fruta y últimamente está interesado en la aplicación de técnicas de aprendizaje de máquina para anotar elementos transponibles.

Romain Guyot, doctor en biología de la Universidad de Zürich, Suiza, investigador senior de Minciencias y director de investigación en genómica y evolución en bioinformática en Institut De Recherche Pour Le Développement (IRD) en Montpellier, Francia. Ha desarrollado múltiples proyectos internacionales de secuenciación y anotación de genomas como el café robusta, café arábigo, piña, lupín blanco, entre otros. Ha desarrollado múltiples estudios de las dinámicas de elementos transponibles y tiene más de 15 años de experiencia en genética, genómica, biotecnología y bioinformática.

Reinel Tabares Soto, coordinador de Ingeniería Electrónica de la Universidad Autónoma de Manizales, magister en Ingeniería – Automatización Industrial, ha participado en proyectos de supercomputación, minería de datos, bioinformática y aprendizaje profundo. Actualmente está haciendo su doctorado en la aplicación de redes neuronales convolucionales en estagooanálisis.


El Semillero de Bioinformática e Inteligencia Artificial inicia a finales del año 2018 y cuenta con cerca de 15 estudiantes de Ingeniería Biomédica, Ingeniería Electrónica e Ingeniería de Sistemas. Se ha divulgado los resultados obtenidos en diversos eventos como: Congreso Internacional de Ingeniería Biomédica y Bioingeniería llevado a cabo en la ciudad de Manizales en noviembre de 2019, participación en el XII Encuentro Departamental de Semilleros de Investigación –RREDSI, llevado a cabo en Manizales en abril de 2020, participación en el primer Congreso Latinoamericano de Mujeres en Bioinformática y Ciencias de Datos (1st-WBDS), llevado a cabo en Buenos Aires, Argentina en septiembre de 2020.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

2. INTRODUCCIÓN

En los últimos años, la bioinformática ha tenido un crecimiento progresivo, gracias a la aparición de nuevas tecnologías y descubrimientos informáticos. Esta rama de la biología, permite aportar herramientas necesarias para la elaboración de estudios y proyectos relacionados con diversos ámbitos biológicos como la genómica. Actualmente, debido a la aparición de la secuenciación de siguiente generación (NGS por sus siglas en inglés), se han finalizado proyectos de secuenciación importantes que permiten estudiar en profundidad los componentes del genoma, como lo son los elementos transponibles (ETs), implicados en mutaciones genómicas. La identificación de los ETs permite desarrollar estudios hacia el entendimiento de las características propias de una especie, y su posible impacto en la funcionalidad de otras estructuras genómicas.

En el presente informe se expondrá el proyecto relacionado con la creación de un conjunto de datos de elementos transponibles LTR-RTs en diversas especies de plantas, exponiendo primeramente el área problemática y justificación, donde se explica deficiencia que existe entorno a los conjuntos de datos y la importancia de realizar el proyecto. Seguidamente, se encuentra el marco teórico, en el cual se explican los diversos conceptos relacionados con el tema de investigación. Luego, se proponen los objetivos tanto general como específicos que son la guía para la realización de las actividades investigativas de este proyecto. Posteriormente, en la metodología se describe las actividades que se llevaron a cabo para cumplir cada uno de los objetivos planteados con anterioridad; después, se encuentra los resultados, en donde se presenta los elementos hallados de acuerdo a lo abordado; el análisis e interpretación de estos se llevó a cabo en la discusión de resultados. Finalmente, en la sección de conclusiones se expusieron los aportes finales, dando las recomendaciones para la realización de estudios posteriores en identificación y clasificación de ET, y en posibles cambios a la metodología aquí planteada.


	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

3. ÁREA PROBLEMÁTICA Y JUSTIFICACIÓN

Los ET se ha visto implicados en eventos génicos importantes, considerándose una parte del genoma de relevancia en la adaptación y evolución de las especies. Existe una orden de ET, llamado los retrotransposones de repetición terminal larga o LTR-RT (por sus siglas en inglés Long Terminal Repeat retrotransposons) los cuales son el principal componente de los genomas de plantas grandes [1], aportando considerablemente a la expansión genómica al multiplicarse y aumentar las copias de este [2]. En algunos estudios realizados estos genomas de plantas como lo es el del maíz (2500 Mb) se ha visto que contiene alrededor de 50% de LTR-RT [3].

Se ha observado que son agentes regulatorios de genes cuando se insertan dentro de uno de ellos o en una vecindad, provocando cambios en su función e incluso el silenciamiento de estos [4]. En plantas, además, se estima que la inserción de los ET puede regular ciertos genes para facilitar la adaptación rápida del organismo a cambios climáticos globales [5]. Actualmente, gracias a la aparición de secuenciación de nueva generación (NGS por sus siglas en inglés), los estudios genómicos, en especial la identificación y clasificación de elementos específicos como los ET ha crecido, teniendo como resultado herramientas bioinformáticas como EDTA [6], RED [7], Inpactor [8], entre otras. Los avances tecnológicos e informáticos han permitido la producción de diversas bases de datos enfocadas en secuencias repetitivas presentes en plantas que contienen los ET, como lo son REXdb [1], la cual contiene los dominios de las poliproteínas de los LTR-RT en plantas; PGSB-REdat [9], una recopilación de secuencias repetitivas y retrOryza [10], una base de datos específica de LTR-RT en el arroz. A pesar de esto, aún existe carencia de un conjunto de datos de LTR-RT intactos en plantas con una mayor cantidad de especies de plantas, que permita la realización de estudios relacionados con la expresión génica de estos y los diversos impactos que presentan en las especies de plantas.

A partir de la creación de un conjunto de datos de LTR-RT en plantas se pretende ofrecer un acceso libre a estos, con el fin aportar recursos para el estudio de interacciones potenciales entre los LTR-RT y los genes de las plantas a estudiar, así como también estudios evolutivos, relación entre especies y relación con el tamaño del genoma. Así mismo, debido al auge en el uso del aprendizaje de máquina (Machine learning) se han construido modelos enfocados en la clasificación de los ET presentes en

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

especies de plantas [11], en donde el conjunto de datos podría emplearse para el entrenamiento y validación de los modelos en cuestión.

4. REFERENTE TEÓRICO

Los elementos transponibles (ET) son segmentos cortos de ADN que pueden moverse e insertarse dentro del genoma. Los ET están presentes tanto en eucariotas como procariotas, llegando a representar hasta el 80% del ADN nuclear en plantas, 3-20% en hongos, y 3-52% en metazoos [12]. Por otra parte, los ET se clasifican según su mecanismo de transposición en Clase I o Retrotransposones y Clase II o transposones. Los Clase I se transponen a través de un intermediario de ARN, el cual se transcribe de una copia genómica, para luego ser transcrita en reversa en ADN por una transcriptasa inversa codificada por el ET, produciéndose una nueva copia [13] (Ver Figura 1). A diferencia de los Clase I, los Clase II se mueven a partir de un mecanismo de cortar y pegar, donde el intermediario es directamente el ADN [14] (Ver Figura 2). Al igual que los Clase I, los Clase II son antiguos y están presentes en casi todos los eucariotas.

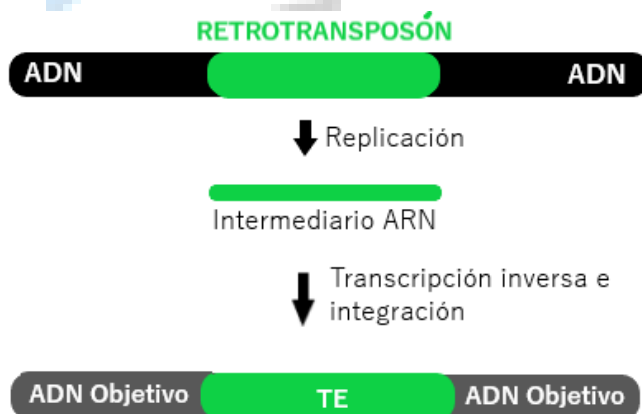


Figura 1. Transposición de los ET Clase 1

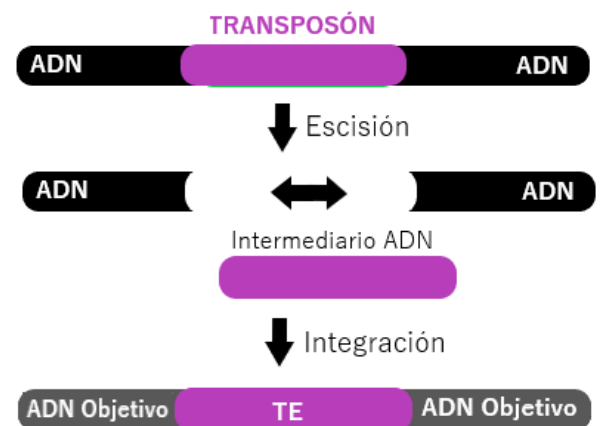




Figura 2. Transposición de los ET Clase 2

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

Los ET fueron descubiertos por primera vez en el maíz por Barbara McClintock en 1950 [14], proponiendo la idea de que la transposición de estos puede activarse bajo exposición a estrés, contribuyendo a la reestructuración del genoma [14]. Se ha observado que la inserción de los ET dentro de un gen o en su vecindad puede producir efectos negativos al interrumpir las secuencias reguladoras de los genes, eliminando o modificando la expresión génica como la funcionalidad de estos, e incluso generando nuevas funciones [15,16], sin embargo, también pueden causar cambios regulatorios, expansión genómica y generar nuevas variantes cromosómicas debido a las inversiones generadas [17]. Así mismo, la selección natural, junto con la deriva genética, son los responsables de mantener una distribución estable de ET en el genoma. Cuando las inserciones de los ET producen diversos efectos nocivos, estos casos son rápidamente removidos de la población, por otro lado, si las inserciones causan muy pocos efectos nocivos, o casi nulos, estas inserciones pueden fijarse dependiendo de la eficiencia de la selección natural y la deriva para eliminar estas inserciones [18].

El silenciamiento de los genes puede ocurrir cuando un ET se inserta dentro o cerca de estos en orientación opuesta creando transcripciones antisentido [19] que prevalecerán o se desintegrarán dependiendo de los procesos de selección y control de la especie durante períodos evolutivos [18]. A pesar de la desactivación de los ET, los promotores motif pertenecientes al ET pueden permanecer conservados, influyendo en los genes cercanos incluso si el ET se encuentra incompleto, otorgando la idea de que estas secuencias se incorporaron como promotores reguladores o en funciones de genes gracias a cambios adaptativos de la especie [20]. Además, en estudios previos se ha encontrado presencia de secuencias derivadas de ET en regiones codificantes de proteínas, lo que demuestra la evolución guiada por los ET [21]. De la misma forma, los polimorfismos creados por los ET se encuentran en individuos dentro de una especie, indicando que estos se han activado recientemente; alguno de estos casos está implicados en la obtención de características singulares como la resistencia del arroz a enfermedades [22] y el color de la piel de las uvas [23].


Los retrotransposones se dividen a su vez en dos grupos, los retrotransposones de repetición terminal larga o LTR-RT (*Long Terminal Repeat Retrotransposon*), y los no LTR-RT. Los LTR-RT son los que contribuyen de manera significativa a la expansión del tamaño del genoma, debido a la gran cantidad de copias que genera [24]. Los LTR-RT presentan un mecanismo de transposición muy similar al utilizado para los retrovirus, llevado a cabo por un sitio de unión del cebador (PBS por sus

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

siglas en inglés), un tracto de polipurina (PPT por sus siglas en inglés), un gen *gag*, el cual codifica proteínas estructurales para la transcripción inversa, un gen *pol*, que funciona como proteasa, transcriptasa reversa e integrasa, y algunos LTR-RT presenta un fragmento similar a *env*, que codifica una proteína de unión al receptor transmembrana que permite la transmisión del virus [25], los cuales se encuentran en la región interna del retrotransposón; los LTR-RT que codifican a *env*, son de hecho, retrovirus [25]. El proceso de transcripción reversa genera dos LTRs idénticos los cuales se encuentran a los extremos; con el paso del tiempo, las mutaciones se acumulan en estos, por lo tanto, la tasa de divergencia calculada entre ambos LTRs es usualmente usada como reloj molecular para encontrar el tiempo de inserción del LTR-RT [26].

El ciclo de amplificación de los LTR-RT es similar al ciclo retroviral realizado por los retrovirus y comienza con la transcripción, teniendo inicio en el LTR 5' y termina en el LTR 3', en el que se produce una plantilla de ARN la cual será utilizada para la transcripción inversa, y los ARNm que traducirán las proteínas luego de que la proteasa realice la escisión de la poliproteína *gag-pol*; la proteína *gag* se une a la plantilla de ARN, en donde se forma una partícula similar al virus citoplasmático (VLP), la cual se encuentra estrechamente relacionada con el núcleo del virión retroviral. Una vez formado, el VLP encapsula la transcriptasa inversa codificada por el gen *gag*, con el cual se producirá la copia de ADN y la integrasa realizará la transferencia de la copia de ADN lineal al núcleo y su posterior inserción en el genoma [27]; es por tanto que la amplificación de un LTR-RT en particular se realiza bajo una regulación transcripcional.

Los LTR-RT se clasifican a su vez en dos superfamilias más importantes: Gypsy y Copia, dependiendo del orden de los dominios proteicos en el gen *pol* (Ver Figura 3). En muchas especies de plantas, Gypsy, también conocido como *Metaviridae*, se encuentra insertado en regiones heterocromáticas, alejados de los genes, a diferencia de Copia o *Pseudoviridae*, los cuales se encuentran mayormente insertados en regiones cercanas a genes [27,20]. Se ha observado en estudios previos que existe un patrón en donde un número bajo de copias y las familias de LTR-RT jóvenes, están mayormente asociadas a genes que las familias con un número de copias alto [20].

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

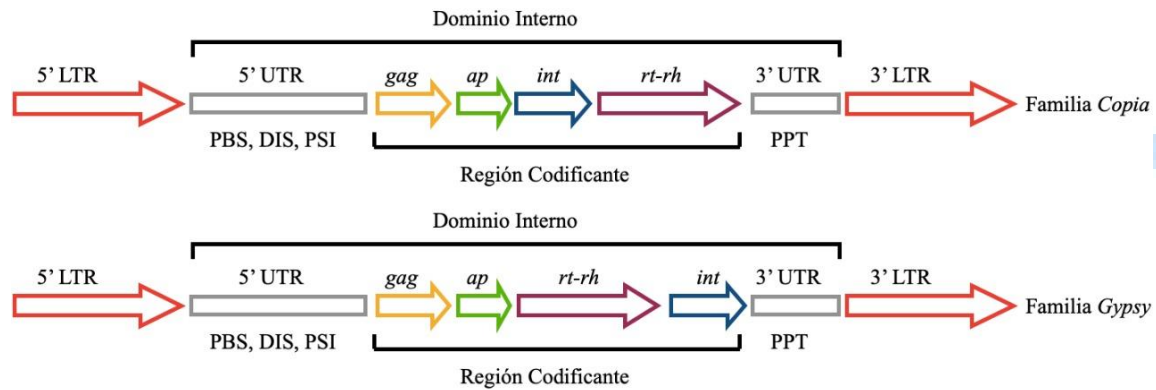



Figura 3. Estructura interna de *Copia* y *Gypsy*

Se ha especulado que la evolución de genes de resistencia (genes R) mediada por ET han sido beneficiosos para las plantas. Estos genes responden a factores virulentos del patógeno, de los cuales reconoce los efectores y la planta actúa en defensa a estos. Sin embargo, dado que los patógenos a su vez también presentan mutaciones en sus efectores estos pueden ingresar a la planta y pasar desapercibidos para realizar la infección, hasta que la planta desarrolla nuevos genes R encargados de detener esos nuevos efectores. Se estima que esta evolución de los genes de resistencia podría beneficiarse y mejorar gracias a la presencia de ET en estos [20].

Debido a la gran cantidad de ET presentes en plantas (más del 85%) [24], se hace necesario realizar una identificación y clasificación de estos en las diferentes especies de plantas con el fin de encontrar su relación con funcionalidad de los genes, importancia en la evolución y tamaño del genoma. En la actualidad existen diversos softwares bioinformáticos que permiten la identificación de los ET como RED [7], EDTA [6] así como también softwares encargados de la clasificación como LTRClassifier [28], e Inpactor [8].

Así mismo, se han creado diversas bases de datos en donde se encuentran secuencias repetitivas en algunas especies tanto de plantas, como de otro tipo de organismos, otorgando una herramienta para futuras investigaciones, como lo son Repbase [29], un base de datos que contiene secuencias consenso completas de familias de elementos transponibles y otro tipo de secuencias repetitivas en genomas eucariotas, la cual tiene gran variedad sin embargo, solo tiene libre acceso para usos

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

académicos e investigadores sin ánimo de lucro; Gypsy DataBase (GyDB) [30], en donde se encuentran los LTR-retrotransposones, principalmente los contenidos en las superfamilias Gypsy/Copia y los elementos similares a Retroviridae; RepetDB [31], en donde se presenta una anotación completa de ET en secuencias genómicas casi completa, no obstante, no contiene una cantidad significativa de especies a estudiar; y PGSB-REdat [9], o también conocida como base de datos de repeticiones MIPS, junto con PGSC-REcat son parte de la plataforma Plant Genome and Systems Biology platform (PGSB) incluida en PlantsDB, y contiene ET recuperados de TREP, repeticiones de TIGR [32], Repbase y ET detectados de novo en los genomas que se encuentran en PlantsDB. Sin embargo, aún existe una carencia de un conjunto de datos completo de ET, en específico de LTR-RT en plantas, que contenga copias significativas de cada familia, y una clasificación adecuada de cada una de estas.


5. LOS OBJETIVOS

General:

Crear un conjunto de datos de referencia de LTR retrotransposones presentes en especies de plantas.

Específicos:

- Seleccionar una lista de especies de plantas de libre acceso de acuerdo a la familia, tamaño y calidad del ensamblaje.
- Diseñar un *pipeline* para la detección y clasificación de los LTR retrotransposones
- Identificar los LTR retrotransposones encontrados por el *pipeline* construido.
- Examinar los resultados obtenidos del *pipeline* para dar paso a la construcción del conjunto de datos.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

6. METODOLOGÍA


El enfoque de este proyecto es mixto, pues se pretende encontrar la cantidad de LTR-RT en diferentes especies de plantas, así como también a qué familia pertenece cada uno de los elementos encontrados, con el fin de realizar un conjunto de datos completo

El tipo de investigación utilizado para este proyecto es experimental, puesto que se pretende realizar una herramienta que permita construir futuros estudios relacionados con las funciones de los genes y la influencia que tienen los LTR-RT en estos, así como su utilización en otros modelos de identificación y clasificación como los que se llevan a cabo con aprendizaje de máquina.

La construcción de este conjunto de datos consta de diferentes fases. En la primera fase se realiza la recolección de la información, siendo esta los genomas de las especies de plantas a utilizar. En esta fase, se eligen según su familia, puesto que no se deben repetir demasiadas especies por familia, con el fin de obtener una buena representatividad de cada una, así como también por el tamaño del genoma, calidad del ensamblaje, siguiendo medidas como el N50, y nivel de ensamblaje. Es de destacar que se pretende abordar la mayor cantidad de familias de plantas que posean las características óptimas anteriormente descritas.

Posteriormente, se diseñará un pipeline para detectar los ET, en específico de los LTR-RT, haciendo uso de softwares como EDTA [6], el cual entrega las secuencias de los elementos que hacen parte de dicho orden, y con esto se hace un conteo de cuantas ET se encontraron por cada especie de plantas. En esta sección también se realiza el cálculo del N50 con el software QUAST [33], para observar si los genomas presentan una calidad adecuada. De acuerdo a la información obtenida en esta fase, se procede a ajustar algunos parámetros como el tamaño del genoma de las especies, con el fin de encontrar una mayor cantidad de LTR-RT.

Seguidamente, se realizará la clasificación de los LTR-RT encontrados haciendo uso de un pipeline bajo Inpactor [8], con el fin de encontrar las familias y linajes de los elementos presentes en las especies de plantas. El último paso del pipeline consistirá en la realización de una serie de filtros para eliminar aquellas secuencias que no correspondan a LTR-RT intactos o que tengan inserciones anidadas de otros elementos. Se descartarán elementos que: a) tengan dominios de dos diferentes superfamilias, b) tengan menos de 3 dominios, c) con longitudes menores o mayores del linaje


	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

reportadas en Gypsy Database [30] (con una tolerancia del 20%) y d) que tengan inserciones de ET clase 2. Como resultado final se obtiene un archivo plano en formato FASTA que contendrá los LTR-RT encontrados en las 84 especies de plantas, con un ID que contendrá información importante como lo es la superfamilia, linaje, especie de planta a la que pertenece el LTR-RT, así como también la longitud de este. Así mismo, se divulgará por medio de un artículo de investigación en revista indexada, en el que se detallará el proceso metodológico y los resultados obtenidos.

7. RESULTADOS

Para la identificación de los ET en EDTA, se obtuvieron 84 especies de plantas pertenecientes a 81 familias, así mismo, se hizo uso de 4 conjuntos de datos de ET para realizar una recopilación: Repbase, conjunto de datos construida con LTR_STRUC, RepetDB y PGSB, obteniéndose un total de 64, 69, 13 y 20 especies respectivamente. En el Anexo 1 se encuentra el nombre de las especies y la familia a la que pertenecen cada uno de los conjuntos de datos anteriores. De igual forma, en el Anexo 2 se presentan información más detallada de las especies utilizadas para EDTA, así como también de LTR_STRUC, y sus respectivos enlaces de descarga.

Con el fin de observar la relación existente entre los conjuntos de datos anteriores, se realizó dos diagramas de Venn con información acerca de la cantidad de especies y familias dentro de estas. En la Figura 4 se encuentra el diagrama realizado por especies para los cinco conjuntos de datos, encontrándose una cantidad total de 195 especies entre los conjuntos de datos. Se observa que no hay especies compartidas entre estas, generando una diversidad en el conjunto de datos, así como también, se destaca el aporte significativo de especies nuevas utilizadas para EDTA, aportando variedad e importancia al conjunto de datos construido.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

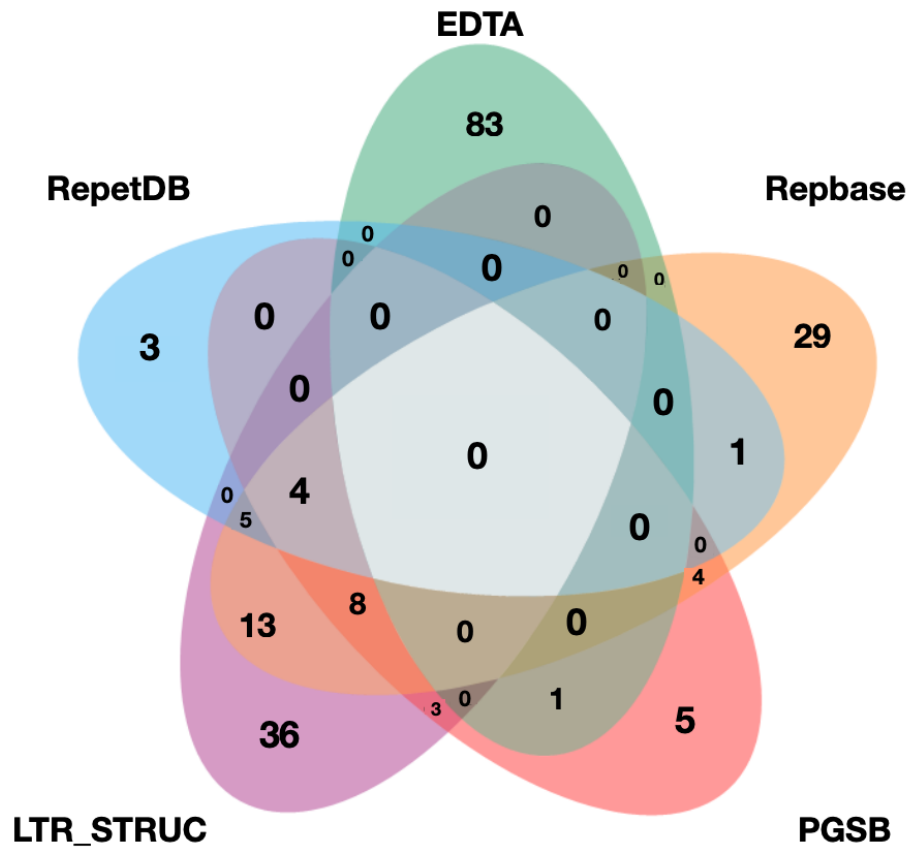



Figura 4. Diagrama de Venn para los 5 conjuntos de datos por especies

Por otra parte, en la Figura 5 están las familias de las respectivas especies para cada conjunto de datos. Así como se observa en la Figura 4, existe una contribución importante de las familias utilizadas para EDTA, viéndose unas cuantas compartidas entre otros conjuntos. Se obtuvo en total 108 familias entre los cinco conjuntos de datos, donde solo dos familias se comparten en todos los conjuntos propuestos.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

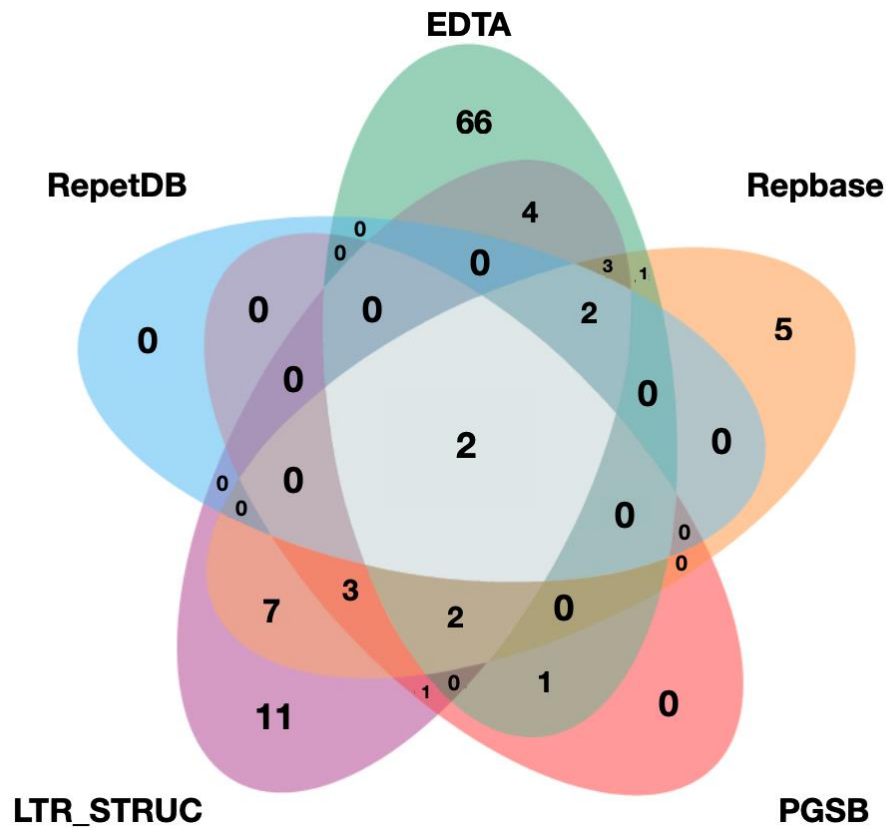



Figura 5. Diagrama de Venn para los 5 conjuntos de datos por familias.

Con las especies iniciales se realizó la identificación de los LTR-RTs usando EDTA, el cual entrega una librería de ET tanto intactos como fragmentados. En este proyecto, se trabajó con un total de 106129 LTR-RTs intactos identificados con EDTA, 49896 identificados con LTR_STRUC, 9278 de Repbase, 61730 y 16127 de PGSB y RepetDB respectivamente. Es de destacar que, a partir de la identificación con EDTA, se encontraron 8 especies que no presentaron ET, siendo *Calotropis procera*, *Spergula arvensis*, *Diospyros lotus*, *Magnolia ashei*, *Moringa oleifera*, *Passiflora edulis*, *Rafflesia leonardi* y *Aristotelia chilensis*.

Para realizar la clasificación a nivel de linaje, se diseñó e implementó una metodología similar a Inpactor, a través de 4 importantes pasos (Ver Figura 6). En el primer paso, se realizó la extracción


	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

de características importantes de la entrada del proceso, es decir, la librería de LTR-RTs, como longitud del elemento, similitud entre los elementos, y los dominios a partir de homología [34], utilizando un base de datos llamada REXdb, que contiene los dominios de referencia de los LTR-RTs. A través de BLASTX [35], se identifican los dominios encontrados, y gracias a la referencia, se extrae información acerca de la superfamilia y linaje al cual pertenecen.

En el segundo paso, se llevó a cabo diversos filtros. Inicialmente, se contabilizaron los dominios encontrados y se descartaron aquellos elementos que contuvieran menos de tres dominios, con el fin de eliminar los ET no autónomos. Posteriormente, se verificó la inexistencia de dominios de RLC como de RLG combinados, puesto que, si los hay, indicaría mutaciones complejas entre elementos pertenecientes a *Copia* y *Gypsy* que no se tratarán en la librería a crear.

En el tercer paso, se realizó la clasificación de los LTR-RTs en superfamilias, a partir de los dominios homólogos encontrados en la base de datos RexDB. Para la clasificación de estos elementos en linajes, se procedió a cuantificar los linajes que se encuentran en los respectivos dominios, así, el LTR-RT pertenecerá al que obtenga mayor puntaje. En el caso en el que la cantidad fue igualitaria en dos o más diferentes linajes, el elemento quedó descartado.

En el cuarto y último paso, se aplicó un filtro de tamaño con una tolerancia del 20% en los LTR-RTs teniendo como referencia los elementos de los linajes reportados en la base de datos GyDB, en donde se tiene en cuenta la suma del tamaño interno y los LTR flanqueantes de los ET. Seguidamente, se eliminaron las secuencias de los ET pertenecientes a Clase 2 o Transposones, a través de un alineamiento utilizando BLAST, con la librería creada de LTR-RTs y los ET clase 2 encontrados en la base de datos Repbase. Finalmente, para eliminar la redundancia en los LTR-RTs en los tres conjuntos de datos (PGSB, LTR_STRUC y EDTA), se procedió a seguir la metodología implementada en REPET. Primero, se realizó un alineamiento con BLASTn de todos los elementos consigo mismos, luego se realizó una clusterización con SiLiX [36]. Después, se efectúa un alineamiento múltiple para cada conjunto de datos con MAFT [37], para finalmente construir la secuencia consenso usando cons, una herramienta de EMBOS [38].

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

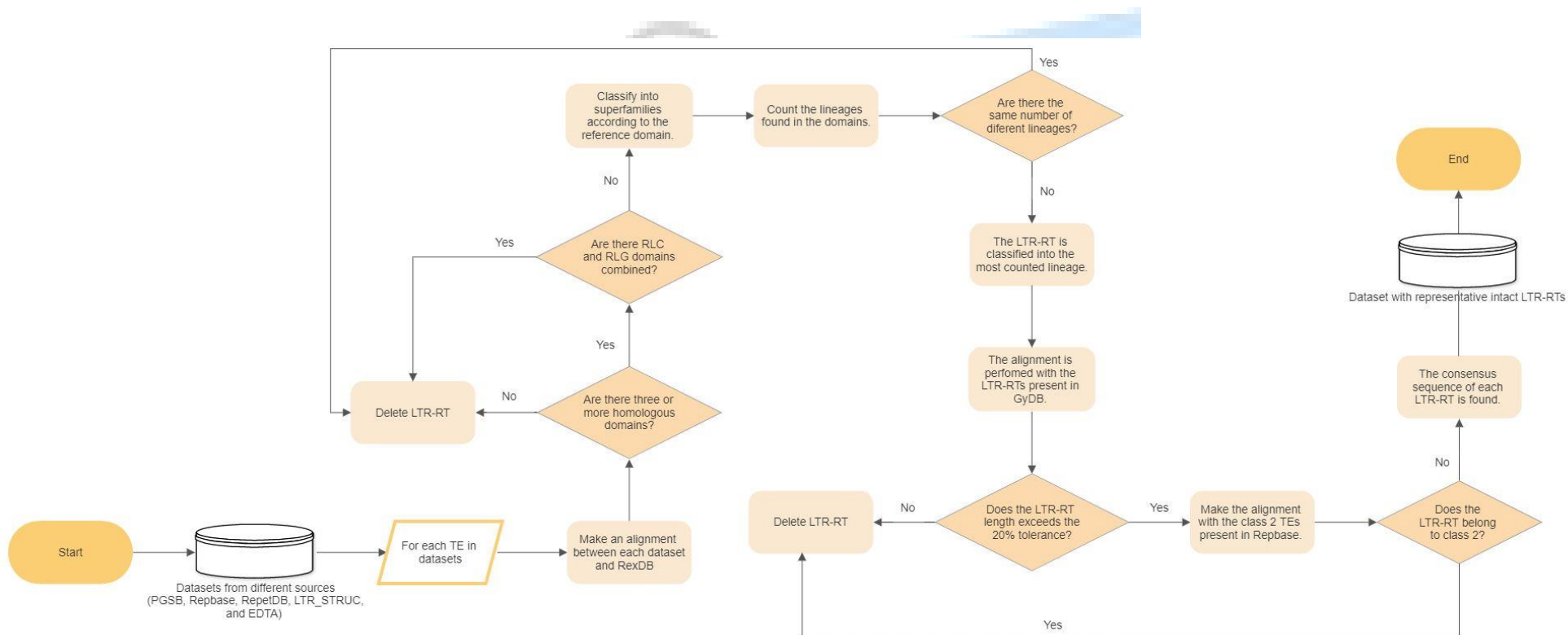



Figura 6. Diagrama de flujo del pipeline para clasificación.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


En el Anexo 3 se encuentra el script realizado para poner en marcha el diseño del pipeline anteriormente descrito, con el que se construyó finalmente la librería llamada InpactorDB.

8. DISCUSIÓN DE RESULTADOS

Actualmente, la biología ha tenido un crecimiento en la cantidad de datos debido a los avances significativos de tecnologías que permiten una extracción más sencilla de información. Los conjuntos de datos pasan por tres actividades como son la captura, curación y análisis de la información, con el fin de obtener datos de mayor complejidad y abrir puertas para futuras investigaciones [39]. Así mismo, estos presentan una importancia en la biología computacional, principalmente en la construcción e implementación de modelos bajo aprendizaje de máquina y aprendizaje profundo. Para el correcto entrenamiento de estos modelos, en especial los utilizados bajo aprendizaje profundo, se requieren de datos grandes que contengan variedad y representatividad con el fin de que estos trabajen de manera correcta [40]. La selección de la información que irá contenida dentro del conjunto de datos no debe presentar sesgos, con el fin de evitar posibles errores en los análisis posteriores [41].

Los ET se pueden identificar a partir de diversas metodologías como lo son: de novo, basados en estructura, genómica comparativa, basados en homología [42], siendo este último el comúnmente utilizado. La estrategia utilizada en la identificación por homología detecta ET en base a la similitud que presenta este con respecto a unas secuencias ET de referencia [43]. Cuando existe una librería de ET ya existente, el proceso puede ser mucho más sencillo [11]. Sin embargo, algunas dificultades que presenta esta metodología son la complejidad de la creación de la librería de ET debido a la gran diversidad que presentan estos elementos, debido a que estos no se asemejan a una estructura universal, es decir, mucho de estos presentan inserciones de otros elementos generando copias fragmentadas que son difíciles de tratar [11,44] o se evidencian delecciones internas por efectos de recombinación que producen un cambio en su estructura interna [45]. Así mismo, algunos elementos pertenecientes a ciertas familias tienen una cantidad de copias muy pequeña, lo que dificulta su caracterización [46].

A día de hoy, existen diversos conjuntos de datos sobre ET en diversas especies de plantas y otros organismos, como GyDB, Repbase, RepetDB, Plant Genome and System Biology (PGSB) Repeat

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


Database, SoyTEdb [46], entre otras. Algunos de estos conjuntos se especializan en el almacenamiento de tipo de ET, especialmente los LTR-RT en especies de plantas, y otras contienen ET en general en diversos organismos. Sin embargo, algunas presentan deficiencia en cantidad de especies relevantes que permiten eliminar la redundancia y aumentar la representatividad de las familias a las que pertenecen. Así mismo, es de destacar que ciertos conjuntos de datos como SoyTEdb son específicos a una sola especie de plantas, aumentando la cantidad de información para estudios relacionados con dicha especie y el entendimiento de su genoma, no obstante, no se puede lograr una generalización de los datos y observar relaciones de la especie con otras familias de plantas, limitando el uso de estas.

Finalmente, la utilidad del conjunto de datos InpactorDB genera la expansión de estudios relacionados con los LTR-RTs presentes en una gran cantidad de familias de plantas, generando investigaciones relacionadas con el impacto de los elementos al tamaño del genoma, la adaptabilidad y evolución genómica de estas, así como también generar estudios con un enfoque genómico comparativo. Por otra parte, en el diseño e implementación de modelos de aprendizaje de máquina y aprendizaje profundo se hace necesario un conjunto de datos robusto con alta representatividad como lo es InpactorDB, que permita el entrenamiento del modelo para la identificación y/o clasificación de los LTR-RTs en especies de plantas siguiendo una metodología de homología, permitiendo obtener unos resultados más precisos.

9. CONCLUSIONES

Inicialmente, la selección de las especies de plantas elegidas para realizar la identificación de los LTR-RTs con EDTA se llevó a cabo descartando las familias que ya se encontraban en los otros cuatro conjuntos de datos a analizar, con el fin de obtener las familias más representativas, y aportar de manera significativa a la creación del conjunto de datos InpactorDB.

En el diseño e implementación del *pipeline* para la clasificación de los LTR-RTs encontrados se tuvo en cuenta algoritmos de bioinformática más comunes y que han tenido impactos importantes en este ámbito, siendo este la identificación por homología, metodología que utiliza una gran parte de los softwares bioinformáticos existentes.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


Para la identificación de los LTR-RTs, se presencié una variabilidad grande en la cantidad de elementos encontrados en cada especie de planta donde algunos de estos se relacionaban con la medida de desempeño N50 del genoma de las especies. Sin embargo, existen casos donde no existe una relación entre estas dos medidas.

Finalmente, la reducción de los LTR-RTs a partir del concepto de consenso, se llevó a cabo para aumentar la calidad del conjunto de datos construido y disminuir la redundancia, sin perder la representatividad que tiene cada uno de los elementos en la superfamilia y linaje en el que se encuentran.

10.RECOMENDACIONES

En estudios posteriores, se recomienda tener en cuenta los LTR-RTs con inserciones anidadas y demás cambios estructurales en estos elementos, con el fin de realizar un conjunto de datos mucho más robusto y complejo que contenga información tanto de LTR-RTs intactos como LTR-RTs fragmentados. Es de destacar que es un proceso más complejo como el que aquí se plantea, sin embargo, abriría los caminos para la realización de proyectos relacionados con la relación entre los LTR-RTs fragmentados que han estado presentes en la escala evolutiva y su impacto en la funcionalidad de ciertos genes.

Por otra parte, se recomienda utilizar el conjunto de datos InpactorDB para posteriores estudios relacionados con la genómica comparativa debido a la cantidad de especies de plantas de diversas familias que contiene. Así mismo, su uso en los softwares bioinformáticos construidos bajo un modelo de aprendizaje profundo o aprendizaje de máquina es recomendable, puesto que la representatividad de los LTR-RTs encontrados para cada linaje y superfamilia de ET es alta, así como la gran diversidad que este presenta para obtener resultados generalizables.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


11.EVIDENCIA DE RESULTADOS EN GENERACIÓN DE CONOCIMIENTO, FORTALECIMIENTO DE LA CAPACIDAD CIENTÍFICA Y APROPIACIÓN SOCIAL DEL CONOCIMIENTO, FORMACIÓN

Relacionados con la generación de conocimiento y/o nuevos desarrollos tecnológicos

Resultado/Producto esperado	Indicador	Beneficiario
Conjunto de datos que contiene la identificación y clasificación de los LTR-RT de diferentes especies de plantas como recurso para el estudio de interacciones de los LTR-RT con el genoma de las plantas (expresión génica, evolución, aporte al tamaño del genoma, etc.)	1 conjunto de datos de acceso libre	Comunidad investigativa Nacional e Internacional relacionada en el campo de la bioinformática, inteligencia artificial, biología y áreas relacionadas
Artículo de revisión y comparación	1 Artículo en revista indexada	Comunidad Académica Nacional e Internacional

Dirigido a la formación de recurso humano

Resultado/Producto esperado	Indicador	Beneficiario
Formación de pregrado	Vinculación de estudiantes de pregrado	Estudiantes de la Universidad Autónoma de Manizales

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


Dirigido a la apropiación social del conocimiento

Resultado/Producto esperado	Indicador	Beneficiario
Ponencia	Participación en Encuentro Departamental de Semilleros de Investigación	Estudiantes UAM
Divulgación	Presentación de resultados en el foro de investigación UAM	Comunidad UAM

12.IMPACTOS LOGRADOS

Impacto esperado	Plazo (años) después de finalizado el proyecto: corto (1-4), mediano (5-9), largo (10 o más)	Indicador verificable	Supuestos²
Contribuir a la formación de estudiantes	Corto (1-4 años)	Número de estudiantes vinculados al semillero	Divulgación de la investigación en la comunidad UAM
Contribuir a la construcción de modelos de aprendizaje de máquina en el semillero de	Corto (1-4 años)	Resultados obtenidos tras la construcción del	Divulgación del conjunto de datos en el


² Los supuestos indican los acontecimientos, las condiciones o las decisiones, necesarios para que se logre el impacto esperado.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


investigación		modelo	semillero de investigación
Aportar herramientas para posteriores estudios relacionados con las interacciones de los LTR-RT en funciones de genes, relaciones entre especies, contribuciones el genoma, etc.	Mediano (5-9 años)	Cantidad de proyectos investigativos relacionados con los LTR-RT en plantas	Publicación del conjunto de datos para acceso libre

13. BIBLIOGRAFÍA


1. Neumann P., Novák P., Hoštáková N., and Macas J., "Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification," *Mobile DNA*, vol. 10, no. 1, 2019.
2. Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novák P, Neumann P, et al. "Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size". *New Phytol*, 208:596–607, 2015.
3. San Miguel P., Vitte C. *Handbook of Maize: "The LTR-Retrotransposons of Maize"*. Springer, New York, NY, 2019. https://doi.org/10.1007/978-0-387-77863-1_15
4. Feschotte C, Pritham EJ. "DNA transposons and the evolution of eukaryotic genomes". *Annu. Rev. Genet.* 41:331–68, 2007.
5. Li Z., Hou X., Chen J., Xu Y., et al. "Transposable Elements Contribute to the Adaptation of *Arabidopsis thaliana*", *Genome Biology and Evolution*, Volume 10, Issue 8:2140-2150, 2018. <https://doi.org/10.1093/gbe/evy171>
6. Ou S, Su W. The Extensive de-novo TE Annotator. GitHub. Available from: <https://github.com/oushujun/EDTA>

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


7. Everitt R., Minnema S. E., Wride M. A., Koster C. S., Hance J. E., Mansergh F. C., Rancourt D. E., RED: the analysis, management and dissemination of expressed sequence tags, *Bioinformatics*, Volume 18, Issue 12, December 2002, Pages 1692–1693, <https://doi.org/10.1093/bioinformatics/18.12.1692>
8. Orozco-Arias S., Liu J., Tabares-Soto R., Ceballos D., Domingues D. S., Garavito A., Ming R., and Guyot R., “Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics,” *Biology*, vol. 7, no. 2, p. 32, 2018.
9. Spannagl M., Nussbaumer T., Bader K. C., Martis M. M., Seidel M., Kugler K. G., H. Gundlach, and Mayer K. F., “PGSB PlantsDB: updates to the database framework for comparative plant genome research,” *Nucleic Acids Research*, vol. 44, no. D1, 2015.
10. Chaparro C., Guyot R., Zuccolo A., Piegu B., and Panaud O., “RetrOryza: a database of the rice LTR-retrotransposons,” *Nucleic Acids Research*, vol. 35, no. Database, 2007.
11. Orozco-Arias S., Isaza G., and Guyot R., “Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning,” *International Journal of Molecular Sciences*, vol. 20, no. 15, p. 3837, 2019.
12. Hua-Van A., Rouzic A. L., Maisonhaute C., and Capy P., “Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences,” *Cytogenetic and Genome Research*, vol. 110, no. 1-4, pp. 426–440, 2005.
13. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 2007, 8, 973–982.
14. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. USA* 1950, 36, 344–355.
15. Leprince, A.S., Grandbastien, M.A., Meyer, C., 2001. Retrotransposons of the Tnt1B family are mobile in *Nicotiana glauca* and can induce alternative splicing of the host gene upon insertion. *Plant Mol. Biol.* 47, 533–541. doi:10.1023/A:1011846910918
16. Varagona, M.J., Purugganan, M., Wessler, S.R., 1992. Alternative splicing induced by insertion of retrotransposons into the maize waxy gene. *Plant Cell* 4, 811–20. doi:10.1105/tpc.4.7.811

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

17. Serrato-Capuchina A. and Matute D., "The Role of Transposable Elements in Speciation," *Genes*, vol. 9, no. 5, p. 254, 2018.
18. 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. "Ten things you should know about transposable elements". *Genome Biology* 19:199, 2018.
19. S. D. A. Britto-Kido, J. R. C. F. Neto, V. Pandolfi, F. C. Marcelino-Guimarães, A. L. Nepomuceno, R. V. Abdelnoor, A. M. Benko-Iseppon, and E. A. Kido, "Natural Antisense Transcripts in Plants: A Review and Identification in Soybean Infected with *Phakopsora pachyrhizi* SuperSAGE Library," *The Scientific World Journal*, vol. 2013, pp. 1–14, 2013.
20. Galindo-González L., Mhiri C., Deyholos M. K., and Grandbastien M.-A., "LTR-retrotransposons in plants: Engines of evolution," *Gene*, vol. 626, pp. 14–25, 2017.
21. Lockton, S., Gaut, B.S. "The Contribution of transposable elements to expressed coding sequence in *Arabidopsis thaliana*". *J. Mol. Evol.* 68, 80–89, 2009. doi:[10.1007/s00239-008-91905](https://doi.org/10.1007/s00239-008-91905)
22. Hayashi K. and Yoshida H., "Refunctionalization of the ancient rice blast disease resistance gene *Pitby* by the recruitment of a retrotransposon as a promoter," *The Plant Journal*, vol. 57, no. 3, pp. 413–425, 2009.
23. Kobayashi S., "Retrotransposon-Induced Mutations in Grape Skin Color," *Science*, vol. 304, no. 5673, pp. 982–982, 2004.
24. Bonchev, G.N. Useful parasites: The evolutionary biology and biotechnology applications of transposable elements. *J. Genet.* 2016, 95, 1039–1052.
25. Tu Z., "Comprehensive Molecular Insect Science: Volume 4: Biochemistry and Molecular Biology", *Elsevier*, 395-436, 2005.
26. Kijima T. and Innan H., "On the Estimation of the Insertion Time of LTR Retrotransposable Elements," *Molecular Biology and Evolution*, vol. 27, no. 4, pp. 896–904, 2009.
27. Grandbastien M.-A., "Encyclopedia of Virology: Retrotransposons of Plants", *Elsevier*, 428-436, 2008.
28. Monat C., Tando N., Tranchant-Dubreuil C., and Sabot F., "LTRclassifier: A website for fast structural LTR retrotransposons classification in plants," *Mobile Genetic Elements*, vol. 6, no. 6, 2016.


	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

29. Bao W., Kojima K. K., and Kohany O., "Repbase Update, a database of repetitive elements in eukaryotic genomes," *Mobile DNA*, vol. 6, no. 1, 2015
30. Llorens C., Futami R., Bezemer D., and Moya A., "The Gypsy Database (GyDB) of mobile genetic elements," *Nucleic Acids Research*, vol. 36, no. Database, 2007.
31. Amselem J., Cornut G., Choisine N., Alaux M., Alfama-Depauw F., et al, "RepetDB: a unified resource for transposable element references," *Mobile DNA*, vol. 10, no. 1, 2019.
32. Ouyang S, Buell CR. The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.* 2004;32(Database issue):D360–D363. doi: [10.1093/nar/gkh099](https://doi.org/10.1093/nar/gkh099).
33. Gurevich A., Saveliev V., Vyahhi N., and Tesler G., "QUAST: quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, 2013.
34. Bergman C. M., Quesneville H., "Discovering and detecting transposable elements in genome sequences," *Briefings in Bioinformatics*, vol. 8, no. 6, pp. 382–392, 2007.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-410. doi:[10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
36. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 12, 116, 2011. <https://doi.org/10.1186/1471-2105-12-116>
37. Rozewicki J., Li S., Amada K.M., Standley D.M., Katoh , MAFFT-DASH: integrated protein sequence and structural alignment, *Nucleic Acids Research*, Volume 47, Issue W1, 02 July 2019, Pages W5–W10, <https://doi.org/10.1093/nar/gkz342>
38. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* 2000, 16, 276–277.
39. Md. Altaf-UI-Amin, Farit Mochamad Afendi, Samuel Kuria Kiboi, Shigehiko Kanaya, "Systems Biology in the Context of Big Data and Networks", *BioMed Research International*, vol. 2014, Article ID 428570, 11 pages, 2014. <https://doi.org/10.1155/2014/428570>
40. Arnal Barbedo J.G. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and Electronics in Agriculture*, vol. 153, pp. 46-53, 2018. <https://doi.org/10.1016/j.compag.2018.08.013>
41. Hakes, L., Robertson, D.L. & Oliver, S.G. Effect of dataset selection on the topological interpretation of protein interaction networks. *BMC Genomics* 6, 131, 2005.

	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

<https://doi.org/10.1186/1471-2164-6-131>

42. Loureiro, T.; Fonseca, N.; Camacho, R. Application of Machine Learning Techniques on the Discovery and Annotation of Transposons in Genomes. Master's Thesis, Faculdade de Engenharia, Universidade Do Porto, Porto, Portugal, 2012.
43. Pearson W. R. "An introduction to sequence similarity ("homology") searching." Current protocols in bioinformatics vol. Chapter 3 (2013): Unit3.1. doi:[10.1002/0471250953.bi0301s42](https://doi.org/10.1002/0471250953.bi0301s42)
44. Kennedy, R.C., Unger, M.F., Christley, S. et al. An automated homology-based approach for identifying transposable elements. BMC Bioinformatics 12, 130, 2011. <https://doi.org/10.1186/1471-2105-12-130>
45. Devos KM, Brown JK, Bennetzen JL: Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 2002, 12: 1075-1079. 10.1101/gr.132102.
46. Du, J., Grant, D., Tian, Z. et al. SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 11, 113 (2010). <https://doi.org/10.1186/1471-2164-11-113>


	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

14.ANEXOS

ANEXO 1. ESPECIES Y FAMILIAS DE LOS CINCO CONJUNTO DE DATOS

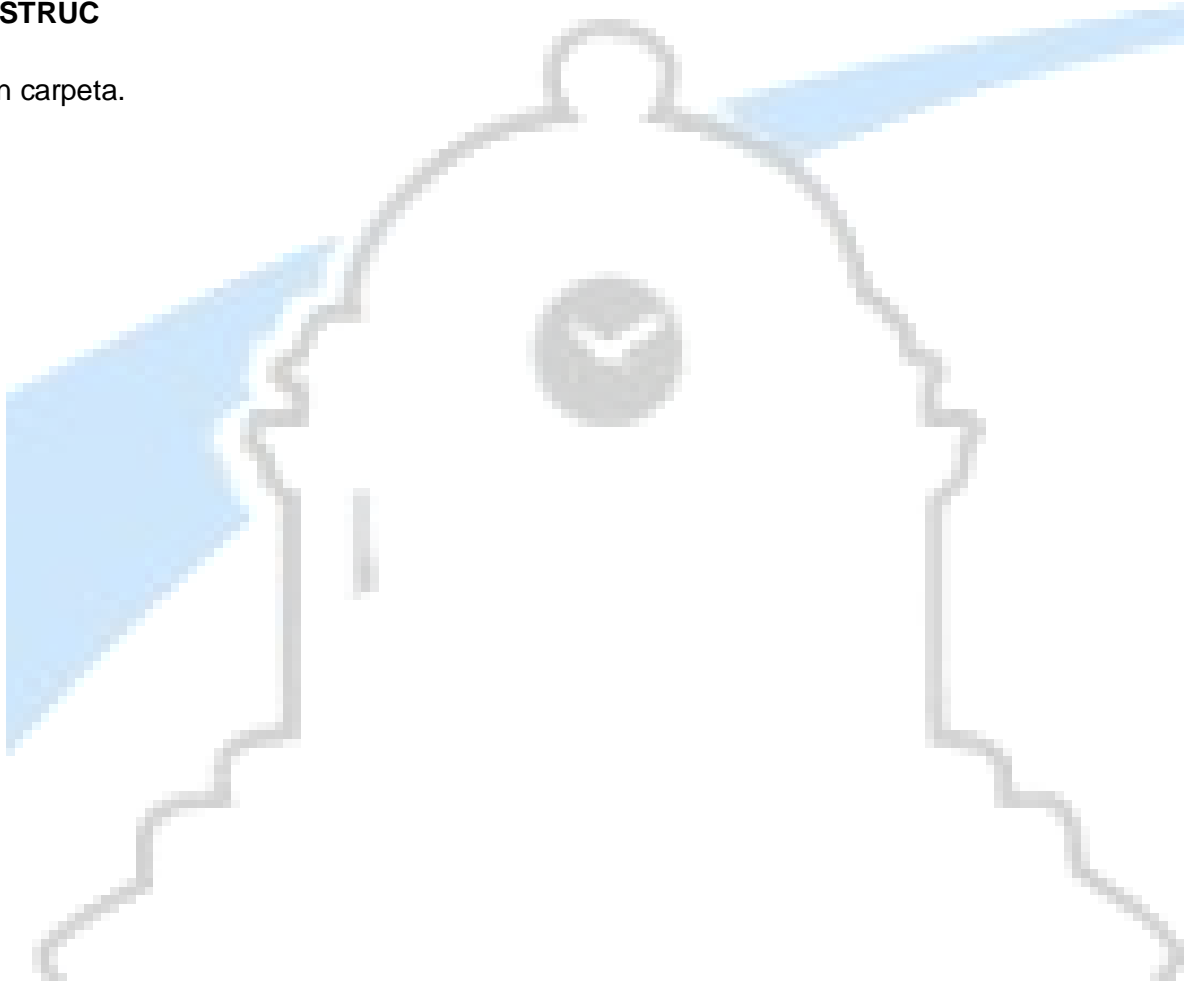
Ver en carpeta




	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

ANEXO 2. INFORMACIÓN ACERCA DE LAS ESPECIES UTILIZADAS PARA EDTA Y LTR_STRUC

Ver en carpeta.



	PRESENTACIÓN INFORME FINAL UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

ANEXO 3. SCRIPT PARA LA IMPLEMENTACIÓN DEL PIPELINE CONSTRUIDO

Ver en carpeta.

