	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	<p>CÓDIGO: GIN-GUI-001</p>
		<p>VERSIÓN: 01</p>
		<p>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</p>



UNIVERSIDAD AUTÓNOMA DE MANIZALES

VICERRECTORÍA ACADÉMICA

UNIDAD DE INVESTIGACIÓN

UNIDAD DE POSGRADOS

	PRESENTACIÓN DE INFORMES FINALES 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: Análisis comparativo de técnicas basadas en aprendizaje de máquina para detectar y clasificar LTR retrotransposones en plantas

GRUPO DE INVESTIGACIÓN: Ingeniería de Software y Automática

ESTUDIANTE: Maradey Mercedes Arias Mendoza

TUTOR DE TESIS: Simón Orozco Arias

DATOS DE IDENTIFICACIÓN:

C.C. 1.192.769.418

Correo: maradey.ariasm@autonoma.edu.co

AÑO: 2021


1

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

RESUMEN

Los elementos transponibles (ET), descubiertos por McClintock en 1944, son segmentos cortos móviles del ADN, con capacidad de integrarse en diferentes regiones del genoma, representando hasta el 80% del ADN en eucariotas, lo que influye en la expresión de los genes y sus repercusiones directas en la adaptación y evolución de las especies. A su vez, los ET se encuentran clasificados según su estructura y modo de propagación, en clases, órdenes, superfamilias, linajes y familias, siendo la clase I, *Long Terminal Repeat* (LTR)-retrotransposones, los más relevantes en el estudio de genomas de plantas debido a su abundancia. Actualmente, gracias a la aparición de tecnologías de secuenciación de nueva generación (NGS), se han desarrollado softwares bioinformáticos y algunas herramientas basadas en técnicas de aprendizaje de máquina para la identificación y clasificación de ET. En ese sentido, este proyecto analiza el rendimiento de un algoritmo propuesto basado en aprendizaje de máquina frente a herramientas bioinformáticas convencionales (conocidas y gratuitas) para detección y clasificación de LTR retrotransposones en plantas, mediante la extracción de métricas de rendimiento como lo son sensibilidad, especificidad, precisión, exactitud, tasa de descubrimientos falsos y F1-score, según líneas de trabajo previamente establecidas, para las especies *Oryza sativa* y *Arabidopsis thaliana*, obteniendo como resultado precisión y especificidad del 97% y 99% respectivamente, para el algoritmo propuesto en la evaluación de la especie *Oryza Sativa*, y la más baja tasa de descubrimientos falsos, equivalente a un 3%, demostrando así las ventajas del aprendizaje de máquina frente a herramientas bioinformáticas convencionales.

PALABRAS CLAVES: bioinformática, aprendizaje de máquina, redes neuronales, elementos transponibles, LTR retrotransposones.

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

ABSTRACT

Transposable elements (TEs), discovered by McClintock in 1944, are short mobile segments of DNA, with the capacity to integrate into different regions of the genome, representing up to 80% of the DNA in eukaryotes, which influences gene expression and its direct repercussions on the adaptation and evolution of species. In turn, TEs are classified according to their structure and mode of propagation, into classes, orders, superfamilies, lineages and families, being class I, Long Terminal Repeat (LTR)-retrotransposons, the most relevant in the study of plant genomes due to their abundance. Currently, thanks to the emergence of next generation sequencing (NGS) technologies, bioinformatics software and some tools based on machine learning techniques have been developed for the identification and classification of TEs. In that sense, this project analyzes the performance of a proposed machine learning-based algorithm against conventional bioinformatics tools (known and free) for detection and classification of retrotransposon LTRs in plants, by extracting performance metrics such as sensitivity, specificity, precision, accuracy, false discovery rate and F1-score, according to previously established lines of work, for the species *Oryza sativa* and *Arabidopsis thaliana*, obtaining as a result accuracy and specificity of 97% and 99% respectively, for the proposed algorithm in the evaluation of the species *Oryza sativa*, and the lowest false discovery rate, equivalent to 3%, thus demonstrating the advantages of machine learning over conventional bioinformatics tools.

KEY WORDS

Bioinformatics, Machine Learning, Neural Networks, Transposable Elements, LTR Retrotransposons.

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

TABLA DE CONTENIDO

1.	6
2.	7
3.	8
4.	10
5.	12
6.1	12
6.2	12
7.	13
6.	14
7.	23
8.	25
9.	26
10.	27
10.1	27
10.2	27
10.3	28
11.	28
12.	29
13.	37


	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

1. PRESENTACIÓN

El proyecto que se expone a continuación se desarrolló en asociación con los grupos de investigación de ingeniería de software y automática de la Universidad Autónoma de Manizales-UAM, bajo la línea de Bioinformática de la rama de Bioinformática e Inteligencia Artificial de este último. Cabe destacar que este proyecto forma parte de un proyecto macro, denominado “Hacia el entendimiento de genomas de plantas de interés productivo”.

Desde la línea de Bioinformática, el equipo de investigación está conformado por:

- Simón Orozco Arias, tutor principal del semillero. Es candidato a doctor 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 arábigo, caña de azúcar, entre otros), en la roya del café y en la mosca de la fruta, últimamente interesado en la aplicación de técnicas de aprendizaje de máquina para anotar elementos transponibles en distintas especies de plantas.
- 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. Con más de 15 años de experiencia en genética, genómica, biotecnología y bioinformática ha desarrollado múltiples proyectos internacionales de secuenciación y anotación de genomas como el café robusto, café arábigo, piña, lupín blanco, entre otros, además de múltiples estudios de las dinámicas de elementos transponibles.
- 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 estegoanálisis.

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

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, Ingeniería de Sistemas y Biología.

2. INTRODUCCIÓN

Hablar del genoma de una especie, es referirse a la información genética contenida en los cromosomas que constituyen los genes, lo que se traduce en instrucciones genéticas que se encuentran en una célula como un conjunto completo [1]. En otras palabras, el genoma constituye el manual de instrucciones para el funcionamiento y sostenimiento de los seres vivos, ya que en su conformación se determinan los niveles más fundamentales de los procesos celulares [2]. Por tanto, la expresión génica en organismos eucariotas frente a organismos procariotas, es distinta. Los genomas eucariotas, por ejemplo, presentan una variedad de regiones genómicas con proporciones distintas según especie; las regiones que se han descrito son: genes, secuencias codificantes (CDS), secuencias no codificantes, secuencias regulatorias, elementos repetitivos, elementos transponibles, entre otros [3].

En la actualidad, el estudio de la variabilidad genética en genomas eucariotas, concentra grandes investigaciones en los elementos transponibles (ET), descubiertos por McClintock en 1944 como segmentos cortos móviles del ADN los cuales tienen la capacidad de integrarse en diferentes regiones del genoma y que en la actualidad 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 [4]. Algunos de los estudios de ET han registrado su estructura, funcionalidad e impacto en las especies [5], la influencia de condiciones de estrés en el medio en que se desarrolla el organismo eucariota [6] y de manera particular, la clasificación de ET en diferentes niveles según estructura, mecanismo de transposición y proliferación [7]. De la clasificación propuesta por Wicker, según el mecanismo de transposición se pueden encontrar ET Clase I o Retrotransposones y ET Clase II o Transposones, y a su vez, subclasificaciones. Por ejemplo, los ET Clase I, se encuentran clasificados en cuatro órdenes, siendo estos los retrotransposones de repetición terminal larga o LTR-RT (*Long Terminal Repeat Retrotransposon*), los no LTR-RT, los PLEs y los DIRs, donde los LTR-RT son los que contribuyen de

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

manera significativa a la expansión del tamaño del genoma en plantas, debido a la gran cantidad de copias que generan lo que deriva en la reprogramación de múltiples procesos en la funcionalidad del genoma de estos organismos. Los LTR-RT a su vez, se encuentran subclasificados por superfamilias, siendo las más importantes en plantas, *Gypsy* y *Copia* [8], con sus distintos linajes.


En ese orden de ideas, el estudio de LTR-RT es importante, por su influencia en la expresión de los genes y su relación con los sistemas regulatorios debido a su activación por estrés [9]. Algunos de esos estudios refieren el uso de técnicas basadas en aprendizaje de máquina para detectar y clasificar ET a fin de reducir los tiempos de ejecución y aumentar la precisión en sus hallazgos usando la aceleración por GPU, además de las diversas herramientas bioinformáticas que han sido desarrolladas para al fin.

Así, en este documento, se expone el desarrollo de un proyecto para el análisis comparativo de la propuesta de un nuevo algoritmo que usa técnicas de aprendizaje de máquina para la detección y clasificación eficiente de LTR-RT en genomas de plantas, frente a herramientas y/o softwares que se basan en técnicas bioinformáticas convencionales y realizan las actividades anteriormente mencionadas; a fin de demostrar las ventajas del aprendizaje de máquina frente a herramientas bioinformáticas convencionales.

3. ÁREA PROBLEMÁTICA Y JUSTIFICACIÓN

De la clasificación propuesta por Wicker [10] para los ET, los LTR-RT aportan considerablemente a la expansión genómica debido a la capacidad de moverse de una ubicación cromosómica a otra, aumentando sus copias [11] como es el caso del maíz, que, con magnitud genómica de 2500 Mb, contiene alrededor del 50% de LTR-RT [12]. La movilización de dichos elementos es causante de múltiples mutaciones en el genoma, que pueden resultar beneficiosas o negativas, teniendo un impacto en la evolución del genoma de la especie.


Actualmente, gracias a la aparición de secuenciación de nueva generación (NGS por sus siglas en inglés), se han incrementado los estudios genómicos de identificación y clasificación de ET, como lo

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

evidencian los proyectos de secuenciación masiva que han surgido durante los últimos años [13] [14], lográndose el desarrollo de diversas técnicas y metodologías, con el fin de identificar los ET en los distintos genomas, ellas son las basadas en estructura, basadas en homología, de novo las de genómica comparativa y las basadas en aprendizaje de máquina, surgidas recientemente [15]; todas a excepción de la última, poseen una tasa relativamente alta de detecciones de falsos positivos [16]. En ese sentido, algunos estudios refieren el uso de técnicas basadas en aprendizaje de máquina para detectar y clasificar ET a fin de reducir los tiempos de ejecución y aumentar la precisión en sus hallazgos usando la aceleración por GPU; por ejemplo, las estrategias para la clasificación jerárquica de ET usando redes neuronales [17], el pipeline del aprendiz de representación de elementos transponibles, TERL, que usa redes neuronales convolucionales para clasificar ET [18] y la medición de métricas de rendimiento de algoritmos de aprendizaje automático para detectar y clasificar ET [19]; esto sumado a las existentes herramientas bioinformáticas desarrolladas para la detección de ET como EDTA [20], RED [21], LTR_Finder [22], y las desarrolladas para clasificación de ET como TEsorter [23], PASTEC [24], Inpactor [25], o TransposonUltimate [26]. Sin embargo, aún persiste la necesidad de poseer herramientas para la detección y clasificación eficiente de LTR-RT en genomas eucariotas, que presenten el mejor rendimiento en términos de precisión y costo computacional, y que en sus tareas de clasificación puedan alcanzar niveles aún más profundos.

En ese sentido, al disponer de un análisis comparativo entre herramientas para la detección y clasificación eficiente de LTR-RT en genomas eucariotas, investigadores que deseen realizar estudios relacionados con la expresión génica de LTR-RT y los diversos impactos que representan estos en las especies de plantas, pueden optar por una herramienta u otra de acuerdo al objetivo del proyecto y en conformidad a sus necesidades y oportunidades, con el fin de seguir desarrollando investigaciones en mejoramiento genómico de especies de plantas, estudios evolutivos y de adaptabilidad entre especies, dada la enorme cantidad de genomas de plantas que se liberan en bases de datos cada año.

Así, considerando la problemática identificada y el apogeo de las técnicas de aprendizaje de máquina enfocado a diversos aspectos de la genética y genómica [27] se propone analizar la implementación de algunas de estas técnicas frente a las empleadas por algunas herramientas bioinformáticas convencionales, en respuesta a la pregunta de investigación, *¿cuál de las herramientas bioinformáticas*

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

seleccionadas, incluyendo la propuesta de un nuevo algoritmo, es más eficiente para detectar y clasificar LTR-RT hasta el nivel de superfamilias en genomas de plantas?

4. REFERENTE TEÓRICO

Los elementos transponibles (ET) son segmentos cortos de ADN que pueden moverse e insertarse dentro del genoma, de una ubicación a otra. Estos están presentes tanto en eucariotas como procariotas, llegando a representar hasta el 80% del ADN en plantas, 3-20% en hongos, y 3-52% en metazoos [28], contribuyendo de manera significativa en la expansión genómica, la diversidad de este entre especies y promoviendo la generación de mutaciones y cambios en el mismo [29]. Los ET fueron descubiertos por primera vez en el maíz por Barbara McClintock en 1944 [30], proponiendo la idea de que la transposición de estos puede activarse bajo exposición a estrés, contribuyendo a la reestructuración del genoma, causando cambios regulatorios, expansión genómica o generando nuevas variantes cromosómicas [31]. A su vez, los ET se clasifican según su mecanismo de transposición en Clase I o Retrotransposones y Clase II o transposones. Los Clase I utilizan la vía de la molécula de ARN para moverse, produciéndose una nueva copia [32]. mientras que los Clase II se mueven usando la molécula de ADN, partir de un mecanismo de cortar y pegar [33]; al igual que los Clase I, los Clase II son antiguos y están presentes en casi todos los eucariotas. Así mismo, los Retrotransposones (Clase I), pueden clasificarse en cuatro órdenes: LTR (*Long Terminal Repeat*), no-LTR que pueden ser LINEs (*Long Interspersed Nuclear Elements*) o SINEs (*Short Interspersed nuclear elements*), PLEs (*Penelope-like elements*), y elementos DIRs, siendo los LTR retrotransposones los más abundantes en los genomas de plantas [34].

Los LTR están conformados por dos secuencias idénticas al inicio y al final, dominios U5, R y U3 y proteínas *gag*, *env* y *pol*, además de los sitios PBS (*primer binding site*) y PPT (*Poly-Purine Tract*), regiones que funcionan como primer e inician replicación, tal como se muestra en la Figura 1 [35].




	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

Figura 1. Estructura de los LTR retrotransposones. Recuperado de Orozco-Arias et al. 2019

De igual forma, los LTR-RT se clasifican a su vez en dos superfamilias más importantes, bastante relevantes en el dominio eucariota: *Gypsy* y *Copia*, con sus respectivos linajes y familias [36]. Así, dada la importancia de los LTR-RT en el genoma de las plantas, se han desarrollado varios métodos bioinformáticos para detectar ET en secuencias del genoma, incluidos los basados en homología, de novo, basados en estructuras y genómicos comparativos, pero ninguna combinación de ellos puede proporcionar una detección fiable en un tiempo relativamente corto [37] ya que carecen de sensibilidad y especificidad debido a las estructuras polimórficas de los TE.

Aunque la técnica basada en estructura, es la más usada para la identificación de LTR-RT, porque hace uso de información de características previamente conocidas sobre la estructura del elemento, sin necesidad de tener una base de datos de los mismos [38], esto mismo deriva en la no correcta identificación de los ET [39]. Existen diversos softwares basados en esta técnica, como: LTR_FINDER [40], LTR_retriever [41], LTR_STRUC [42] y LTRharvest [43]. Por su parte, la técnica basada en homología es la más usada para la clasificación, y necesita una base de datos completa de los ET para relizar correctamente la identificación requiere una referencia de elementos transponibles para realizar la comparación mediante alineamientos y detectar si la secuencia es un ET [44]. Algunos softwares basados en esta metodología son: Inpactor [45] y LTRclassifier [46]. Por otro lado, la técnica de novo busca secuencias similares que se encuentren en múltiples posiciones del genoma, utilizando únicamente la repetición de los ET en el genoma inicial [47]; sin embargo, cuando un elemento transponible no se encuentra repetido múltiples veces, puede que no lo identifique. En cuanto a la estrategia de genómica comparativa, las secuencias de un genoma bien anotado y completo son comparadas para detectar inserciones o deleciones debidas a elementos transponibles [48].

De esta manera, las desventajas de las herramientas convencionales han llegado a representar el campo de estudio de investigadores que han planteado y ejecutado el uso de técnicas de aprendizaje de máquina para mejorar la precisión de la detección de ET. La principal función de los modelos derivados de dichas técnicas es lograr optimizar el rendimiento para que funcione, no únicamente con los datos de entrenamiento, sino, también para datos ingresados posteriormente [49].

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

En ese sentido, investigaciones recientes [50] [51] destacan diversos algoritmos para realizar la clasificación de los ET, como lo son KNN (*K-nearest neighbors*), SVM (*support vector machine*), modelos lineales como, LR (*logistic regression*), LDA (*linear discriminant analysis*), NB (*naive Bayesian classifier*), MLP (*multi-layer perceptron*) y modelos basados en árboles de decisión (DT) como RF (*random forest*). Así, cualquiera sea la o las técnicas empleadas, el porcentaje de efectividad del modelo, siempre puede medirse mediante el *Recall* (sensibilidad), la medida F1, la exactitud, la precisión, la especificidad, el coeficiente de rendimiento, la tasa de falsos positivos y la curva ROC [52], para garantizar un buen entrenamiento del modelo y generalizando su comportamiento.


5. LOS OBJETIVOS

6.1 GENERAL

Analizar el rendimiento de un algoritmo propuesto basado en aprendizaje de máquina frente a herramientas bioinformáticas convencionales para detección y clasificación de LTR retrotransposones en plantas.

6.2 ESPECÍFICOS

- Seleccionar una lista de softwares bioinformáticos basados en técnicas bioinformáticas o en aprendizaje de máquina, diferenciada según detecten o clasifiquen LTR retrotransposones.
- Extraer métricas de rendimiento para cada software seleccionado, teniendo en cuenta líneas de trabajo.
- Implementar y extraer métricas de rendimiento de la propuesta de un pipeline para la detección y clasificación de LTR retrotransposones.
- Analizar los resultados obtenidos tanto del pipeline como de los diferentes softwares seleccionados, en términos de rendimiento para detectar y clasificar LTR-RT en genomas de plantas.

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


7. METODOLOGÍA

La presente investigación se define como aplicada con un enfoque mixto, pues además de realizar la selección de los genomas de las especies con alta calidad en el ensamblaje y los softwares bioinformáticos objeto de estudio, cuando éstos sean ejecutados bajo líneas de trabajo, se hará un análisis cuantitativo de resultados, producto de la identificación de la eficiencia de los mismos para la detección y clasificación de LTR retrotransposones en los genomas elegidos. Por lo cual, el presente proyecto cubre investigación de tipo experimental, porque teniendo en cuenta la comparación entre diferentes técnicas basadas en aprendizaje de máquina frente al pipeline para la detección y clasificación de LTR retrotransposones, se analizará cuál de ellos es mejor en términos de precisión, exactitud, especificidad y sensibilidad, tasa de descubrimientos falsos y F1-score, proponiendo así insumos para el desarrollo de proyectos futuros que estén relacionados con la influencia que tienen los LTR-RT en la funcionalidad y desarrollo de algunas especies de plantas.

Así, el desarrollo del proyecto se llevará a cabo en diferentes etapas, usando recursos de los clústeres a los cuales brinda acceso el *Institut de Recherche pour le Développement, IRD*. En la primera etapa, se recolectará información sobre las especies de plantas con alta calidad en el ensamblaje y la anotación, según información extraída del NCBI (*National Center for Biotechnology Information*); además se identificarán los softwares basados en técnicas bioinformáticas más conocidos y de acceso/revisión gratuita en la detección o clasificación de LTR retrotransposones, y los softwares basados en técnicas de aprendizaje de máquina que también realicen dichas tareas. Para la selección de los softwares se deberá tener en cuenta que puedan clasificar LTR-RT mínimamente hasta superfamilias, sin importar la metodología o metodologías empleada(s).

Una vez identificadas las herramientas existentes para contrastar con nuevos desarrollos en el área, se ejecutarán cada uno de los softwares teniendo en cuenta los flujos de trabajo para clasificación de LTR-RT hasta nivel de superfamilias.

Seguidamente, en la tercera etapa, se extraerán las métricas de rendimiento para cada software, teniendo en cuenta las líneas de trabajo establecidas, las superfamilias y el desempeño en general. Las métricas a estudiar serán: exactitud, precisión, especificidad, sensibilidad, tasa de descubrimientos

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

falsos y F1-score; en la Figura 2 se explica gráficamente la correspondencia de la extracción de métricas según el genoma de referencia.

Finalmente, en la cuarta etapa se realizará el análisis comparativo del rendimiento entre softwares y líneas de trabajo en relación a las métricas obtenidas, los tiempos de ejecución y el costo computacional, haciendo especial énfasis en la sensibilidad, precisión, tasa de descubrimientos falsos y F1-score.

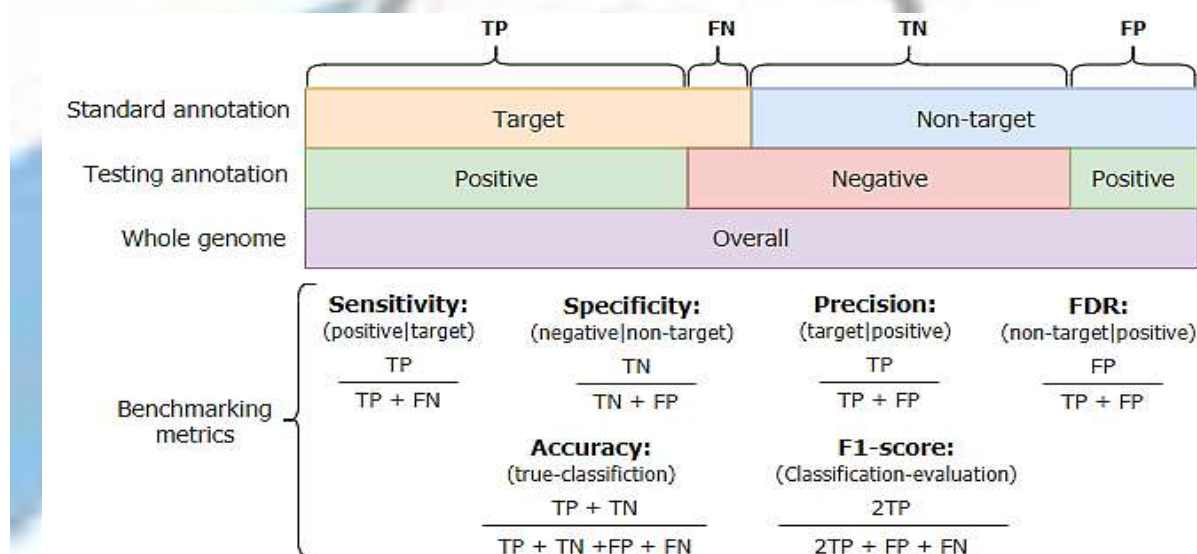



Figura 2. Representación esquemática de las métricas de evaluación comparativa

6. RESULTADOS

Se seleccionaron los genomas de referencia de las especies *Oryza sativa* [53],[54],[55],[56],[57], y *Arabidopsis thaliana* [58], gracias a la alta calidad de su ensamblaje, sus largas historias de descubrimientos y anotaciones. El tamaño de los genomas es 389 Mb y 120 Mb, respectivamente. El genoma de *Arabidopsis thaliana* fue descargado desde el *National Center for Biotechnology Information* (NCBI) y para el caso de *Oryza sativa*, su descarga se realizó desde el material suplementario de EDTA [59], un Anotador TE extenso de-novo, en el cual especifican que utilizaron una biblioteca de alta calidad, curada manualmente y no redundante, denominándola la biblioteca estándar v6.9.5, y con la


	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

cual hicieron la anotación curada de ET del genoma o anotación de referencia para la evaluación comparativa, esto sin detectar elementos de inserción bacterianos, pequeños (pseudo) genes de ARN y ADN de baja complejidad. En el Anexo 1 se presentan los genomas de referencia y librerías estándar.

Para la curación manual de LTR retrotransposones de *Oryza sativa*, los creadores de EDTA primero recopilaron elementos LTR conocidos, los cuales fueron usados para enmascarar candidatos LTR. Los candidatos desenmascarados se comprobaron manualmente buscando repeticiones terminales, secuencias de TSD y secuencias codificantes conservadas y luego estas primeras se alinearon con secuencias extendidas, descartando los candidatos que se extendían más allá de sus límites. De esa manera, para la creación de la biblioteca curada no redundante, cada nuevo candidato de ET fue enmascarado primero por la biblioteca que se tenía, y adicionalmente se verificaba la integridad estructural y dominios conservados de los candidatos desenmascarados. En el caso de los candidatos parcialmente enmascarados y presentados como elementos verdaderos, los creadores de EDTA aplicaban la regla “80-80-80” ($\geq 80\%$ de la consulta alineada con $\geq 80\%$ de identidad y la alineación es ≥ 80 pb de largo) para determinar si el elemento se mantendría. Por otro lado, a los elementos que contenían inserciones anidadas conocidas detectables, se le eliminaban las porciones anidadas se eliminaron y las regiones restantes se unían como una secuencia. En el Anexo 1 se encuentra la librería curada y la anotación estándar.

De esa manera, para obtener la librería estándar para *Arabidopsis thaliana*, la referencia metodológica fue EDTA, y para ello se extrajeron de InpactorDB [60] (un conjunto de datos de referencia de LTR retrotransposones de 195 especies de plantas), las secuencias curadas no redundantes correspondientes a *Arabidopsis thaliana*, para posteriormente hacer su anotación y usar dicho resultado como anotación seleccionada o de referencia y así poder realizar comparaciones con la anotación de prueba, que es la que se obtiene con RepeatMasker de la salida de cada flujo de trabajo. Desde InpactorDB fueron 612 los LTR-RTs correspondientes a *Arabidopsis thaliana* que fueron extraídos como se indica en el Anexo 2.

Para la segunda parte de este proyecto, la detección de ET, se determinó en base de los resultados obtenidos en EDTA al comparar los dos métodos generales de identificación de repetición con clasificación RepeatModeler [61] y Repbase [62], 7 softwares con métodos basados en estructura

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

diseñados específicamente para la identificación de novo LTR (LTR_STRUC [63], LTR_FINDER [64], LTRharvest [65], MGEScan3 [66], LTR_retriever [67], LtrDetector [68] y GRF [69] y las comparaciones de LTR_retriever con LTR_STRUC, LTR_FINDER, LTRharvest y MGEScan_LTR. Los resultados de la evaluación comparativa entre métodos generales y los siete softwares, se pueden ver en la Figura 3, mientras que la de las comparaciones de LTR_FINDER con otros softwares se puede apreciar en la Figura 4.

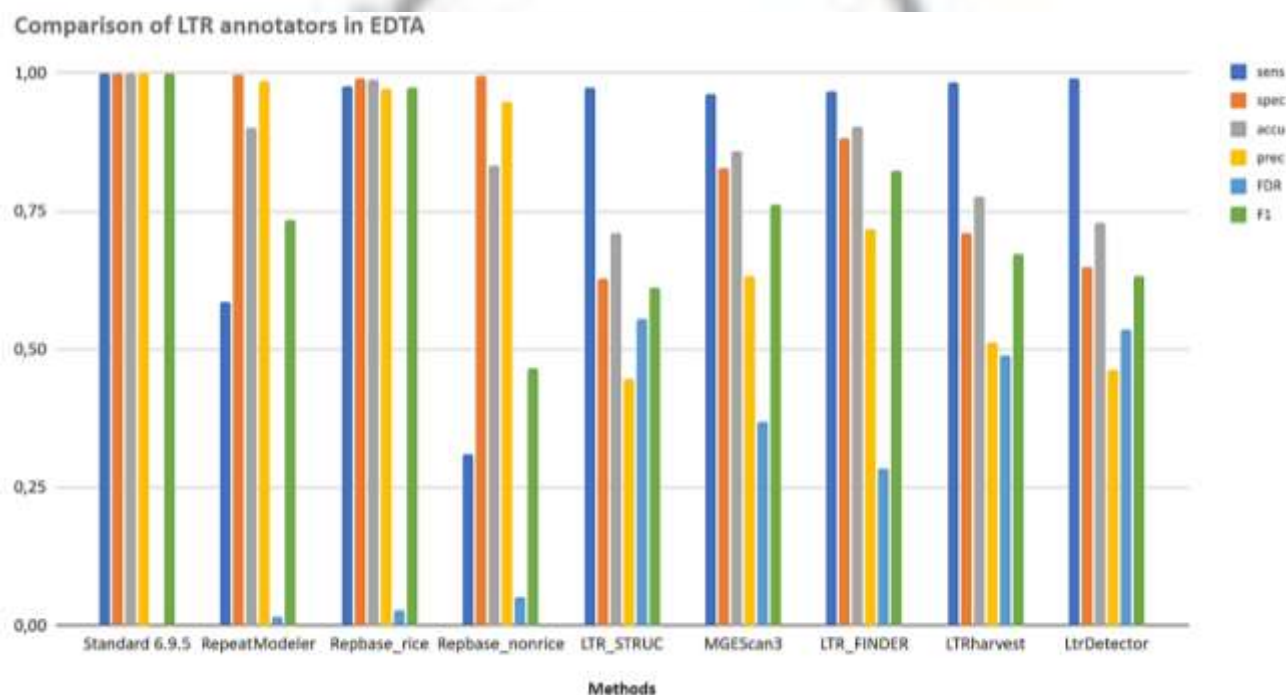



Figura 3. Evaluación comparativa entre dos métodos generales de identificación de repetición con clasificación (RepeatModeler y Repbase) y siete softwares con métodos basados en estructura diseñados específicamente para la identificación de novo LTR.

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

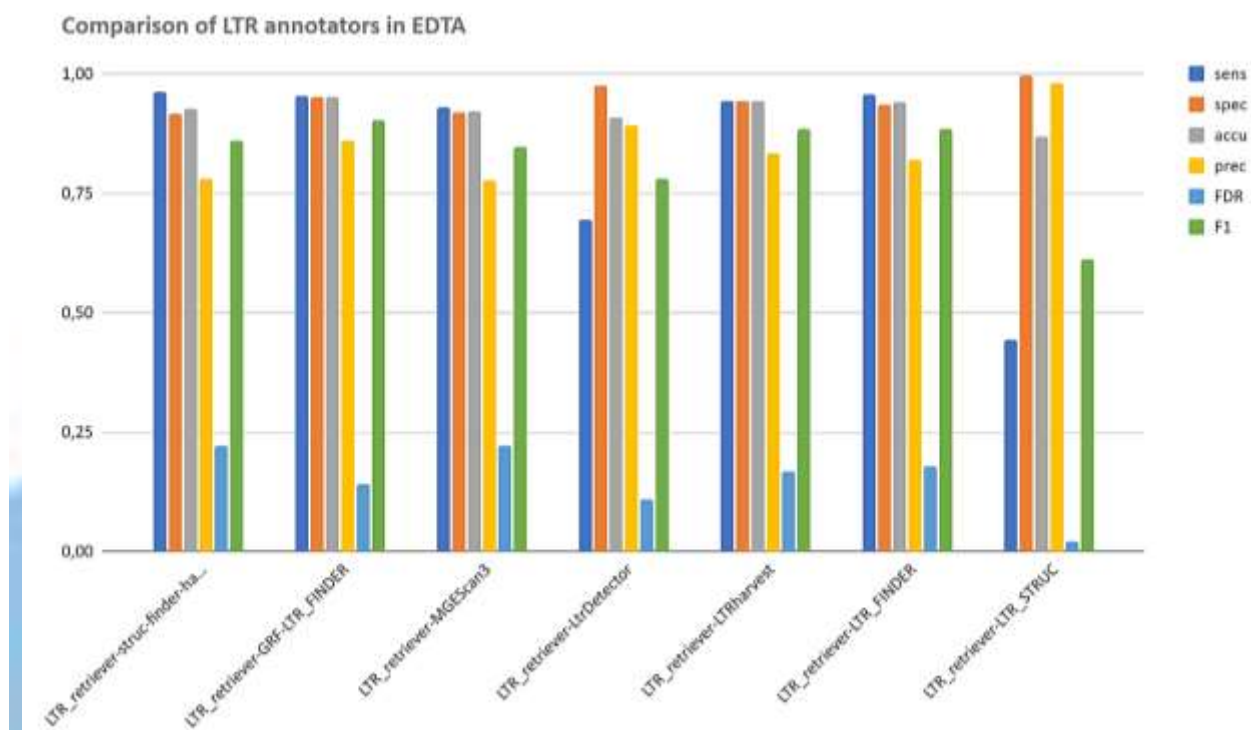



Figura 4. Evaluación comparativa entre LTR-retriever y siete softwares con métodos basados en estructura diseñados específicamente para la identificación de novo LTR

Como se puede apreciar en la Figura 3, LTR_FINDER demostró el mejor equilibrio de rendimiento en todas las métricas seguido por MGEScan3, aunque se registra en el artículo una ejecución lenta por emplear un solo subproceso. No obstante, como los resultados son óptimos, se elige este software para la identificación de LTR-RT en los genomas seleccionados, a fin de que la salida sea la entrada a los distintos softwares de clasificación y poder así realizar la evaluación comparativa entre softwares.

Por otro lado, en el artículo se menciona que, aunque LTR_retriever es un método estricto de filtrado de resultados sin procesar de otros programas LTR, este fue combinado con los otros seis métodos antes mencionados, para comparar su desempeño, como se muestra en la Figura 4, mostrando su alta especificidad ($94,8\% \pm 3\%$), precisión ($92,2\% \pm 3\%$), precisión ($84,9\% \pm 7\%$), medida F1 ($82,4\% \pm 10\%$) y FDR relativamente bajo ($15,1\% \pm 7\%$), lo que en resumen le convierte en el que representa el mejor compromiso entre sensibilidad y especificidad. De esa manera, para nuestro proyecto, teniendo en

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

cuenta el buen desempeño de LTR_retriever anteriormente descrito, y su realización de tareas de clasificación hasta nivel de superfamilias, se decide usar este método como parte de un flujo de clasificación, como se evidencia en la Figura 5.

En ese orden de ideas, para la clasificación de LTR-RT, como otra actividad de la segunda etapa y resultado de la primera etapa de la metodología propuesta, se seleccionaron los siguientes softwares: Inpactor [70], TEsorter [71], TransposonUltimate [72] y el método LTR_retriever [73] para realizar la comparación en cuanto a la clasificación de LTR-RT mínimamente hasta el nivel de superfamilias, frente al desarrollo de un nuevo algoritmo basado en aprendizaje de máquina, denominado de momento, Inpactor v1.3. Los flujos de trabajo propuestos para lograr dicho objetivo son descritos en la Figura 5.

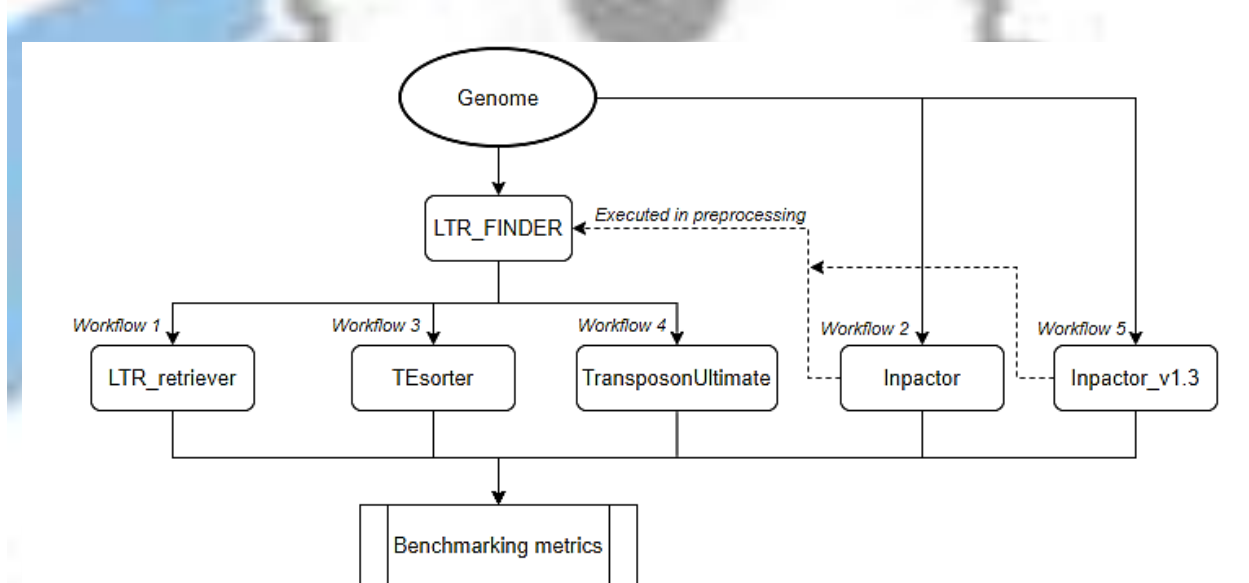


Figura 5. Flujos de trabajo para la clasificación de LTR-RT a nivel de superfamilias

En la tercera etapa, para evaluar cada software se usó la anotación de referencia que contenía la anotación ET de los genomas completos, y con ella se hizo el cálculo de las métricas de sensibilidad, especificidad, exactitud, precisión, FDR y medida F1, utilizando el script respectivo y que forma parte del kit de herramientas de EDTA. A fines prácticos, dicho script fue modificado para ajustarlo a la necesidad de obtener métricas a nivel de superfamilias, como se evidencia en el Anexo 3, pero sin perder la metodología, que es inicialmente etiquetar como “objetivo” las secuencias predichas correspondientes a una superfamilia según coincidan con la librería de referencia (librería curada no

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


redundante), o como “no objetivo” si no corresponden. Posterior a ello, la anotación de cada flujo de trabajo se comparaba con la anotación de referencia de la librería curada, de la siguiente manera:

- Verdaderos positivos (TP): Aquellas secuencias que estuvieren incluidas en el subconjunto “objetivo” (es decir, secuencias predichas que estaban en la librería curada)
- Falsos Positivos (FP): Aquellas secuencias que estuvieren incluidas en el subconjunto “no objetivo” (es decir, secuencias predichas que no estaban en la librería curada)
- Falsos Negativos (FN): Aquellas secuencias que deberían estar incluidas en el subconjunto “objetivo”, pero no estaban, es decir “objetivos perdidos”
- Verdadero Negativo (TN): Resto del genoma

Con dicha información, se procedió a realizar los cálculos de las distintas métricas, en función del número total de bases de ADN genómico. En ese orden de ideas, las métricas se eligieron por su significancia, como se expone a continuación:

- Sensibilidad: Indicará qué tan bien la biblioteca de pruebas puede anotar correctamente las secuencias de TE objetivo.
- Especificidad: describe qué tan bien la biblioteca de prueba puede excluir correctamente secuencias que no son el objetivo.
- Exactitud: Tasa de descubrimiento verdadero.
- Precisión: Tasa real en la discriminación de secuencias objetivo y no objetivo.
- FDR: Tasa de descubrimiento falso.
- Medida F1: Evaluación del modelo de clasificación.

Cabe resaltar que para poder anotar la salida de los softwares Inpactor y TransposonUltimate, para calcular las métricas señaladas, fue necesario hacer un procesamiento a la salida, tal como se muestra en


	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

los Anexos 4 y 5, respectivamente. Adicionalmente, en el Anexo 6 se encuentra un paso a paso de la correcta ejecución para anotar la salida de cada software para hacer los cálculos de las matrices.

En la Tabla 1 se pueden apreciar las métricas obtenidas por línea de trabajo para la clasificación de LTR-RT y según la superfamilia. Allí, en primera instancia es evidente que en general los 4 softwares seleccionados frente a la nueva propuesta del nuevo algoritmo basado en aprendizaje de máquina, tienen un rendimiento óptimo, aunque por métrica existan diferencias notables, como se muestra gráficamente en la Figura 6.

Tabla 1. Métricas obtenidas por línea de trabajo para la clasificación de LTR-RT y según la superfamilia

Category	Methods	sens	spec	accu	prec	FDR	F1	TP	TN	FP	FN
Copia	LTR_FINDER-										
	LTR_retriever	0,894	0,993	0,989	0,840	0,160	0,866	12873348	357973583	2456401	1519022
Gypsy	LTR_FINDER-										
	LTR_retriever	0,862	0,972	0,950	0,888	0,112	0,875	66028613	292959865	8309416	10556671
Total	LTR_FINDER-										
	LTR_retriever	0,867	0,984	0,970	0,880	0,120	0,874	78901961	650933448	10765817	12075693
Copia	FASTA-Inpactor	0,894	0,933	0,932	0,349	0,651	0,502	12848852	336434218	24007655	1529849
Gypsy	FASTA-Inpactor	0,908	0,951	0,943	0,823	0,177	0,864	68514208	286559606	14715405	6932701
Total	FASTA-Inpactor1	0,906	0,941	0,937	0,678	0,322	0,775	81363060	622993824	38723060	8462550
Copia	LTR_FINDER-										
	TEsorter	0,907	0,975	0,972	0,591	0,409	0,716	13003651	351446199	8990421	1327220
Gypsy	LTR_FINDER-										
	TEsorter	0,846	0,960	0,937	0,843	0,157	0,844	65176962	289159281	12110489	11908609
Total	LTR_FINDER-										
	TEsorter	0,855	0,968	0,954	0,787	0,213	0,820	78180613	640605480	21100910	13235829
Copia	LTR_FINDER-										
	TransposonUltimate	0,815	0,958	0,952	0,440	0,560	0,571	12011005	345156202	15280744	2733798
Gypsy	LTR_FINDER-										
	TransposonUltimate	0,866	0,941	0,926	0,789	0,211	0,826	66358105	283531919	17739463	10307023
Total	LTR_FINDER-										
	TransposonUltimate	0,857	0,950	0,939	0,704	0,296	0,773	78369110	628688121	33020207	13040821
Copia	FASTA-										
	Inpactor_V1.3	0,248	1,000	0,963	1,000	0,000	0,398	4581522	360429112	0	13878525
Gypsy	FASTA-										
	Inpactor_V1.3	0,364	0,996	0,847	0,966	0,034	0,528	33645892	300068713	1196178	58886269

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

	FASTA-
Total	Inpactor_V1.3 0,344 0,998 0,904 0,970 0,030 0,508 38227414 660497825 1196178 72764794

Comparison of Classifiers LTR

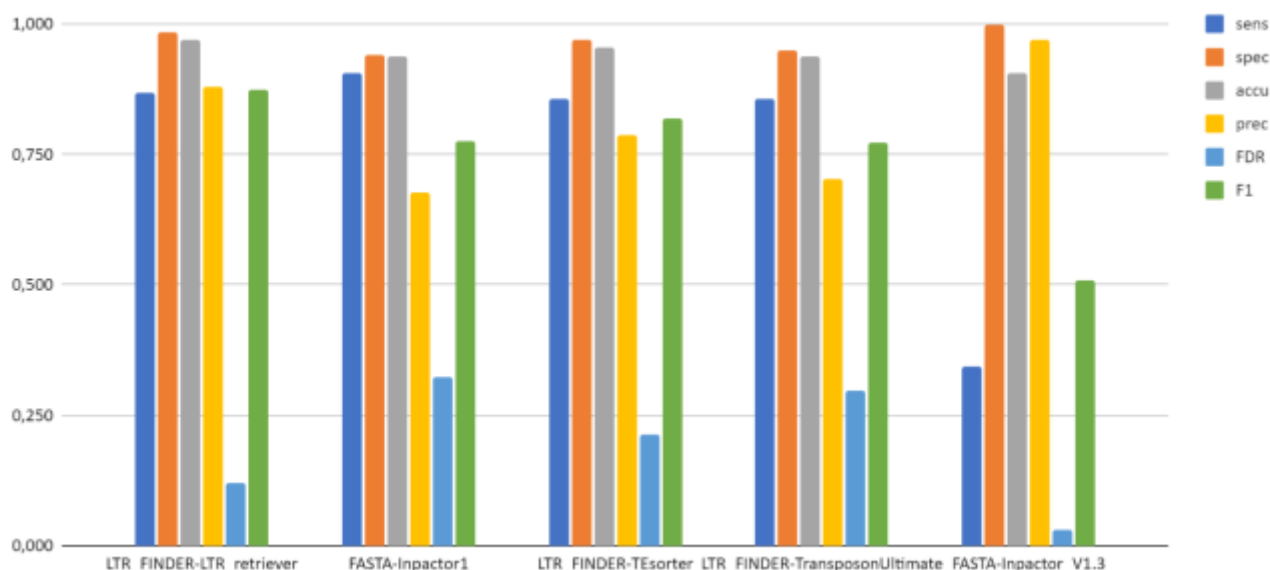



Figura 6. Resultados gráficos de las métricas obtenidas por línea de trabajo para la clasificación de LTR-RT y según la superfamilia.

Como se evidencia en la Figura 6, del flujo de LTR_FINDER con LTR_retriever es donde se obtienen resultados con el mejor equilibrio de rendimiento en todas las métricas, debido a que LTR_retriever es una potente herramienta de filtrado de resultados sin procesar de otros programas LTR. Para el caso de Inpactor v1.3, se evidencian los mejores resultados en especificidad y precisión, además de tener la tasa más baja de descubrimientos falsos, con sustancial diferencia frente a los demás softwares. A manera específica según métricas, los resultados por líneas de trabajo para *Oryza sativa* pueden evidenciarse en las Figuras 7, 8, 9, 10, 11 y 12.

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

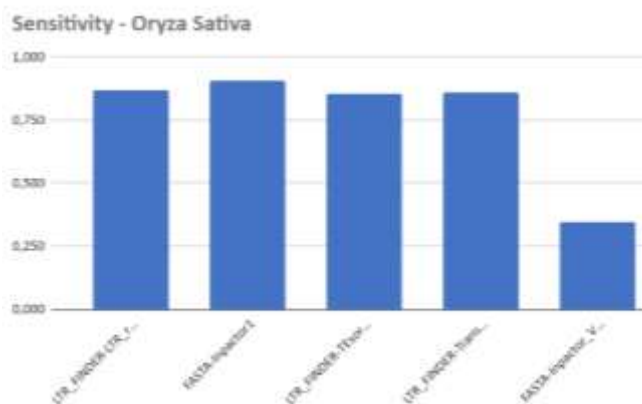


Figura 7. Sensibilidad por líneas de trabajo

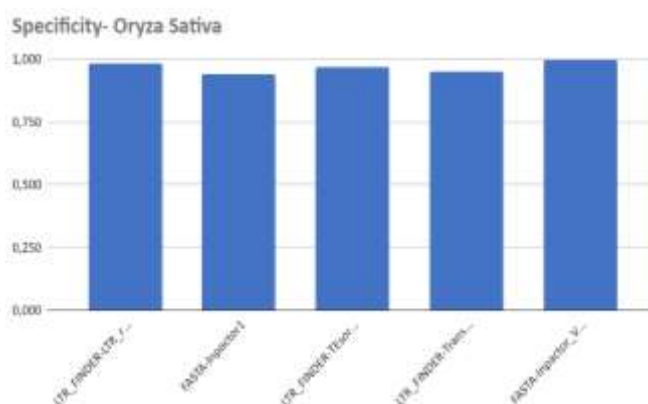


Figura 8. Especificidad por líneas de trabajo

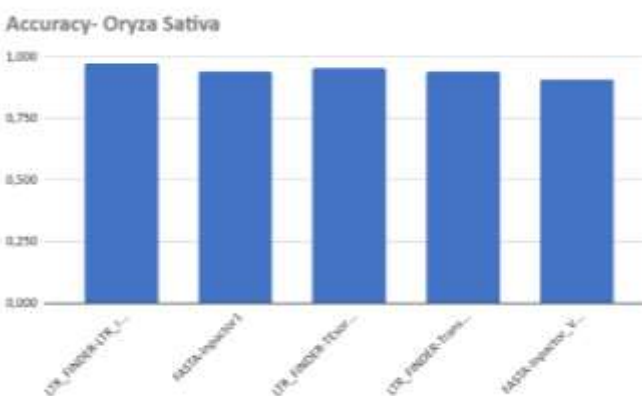


Figura 9. Exactitud por líneas de trabajo

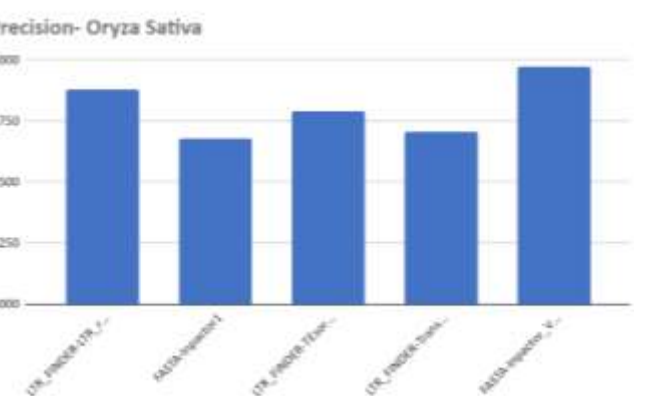


Figura 10. Precisión por líneas de trabajo

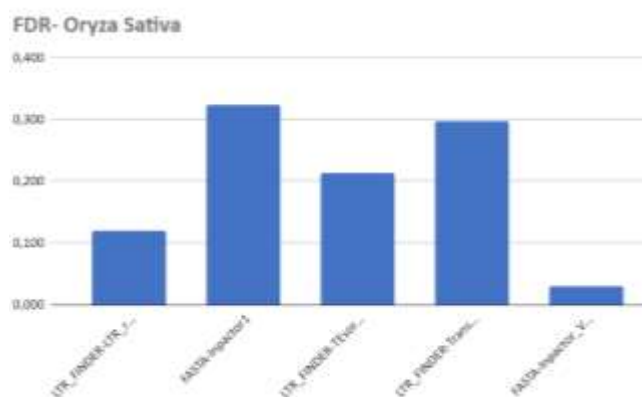


Figura 11. Tasa de descubrimientos falsos por líneas de trabajo

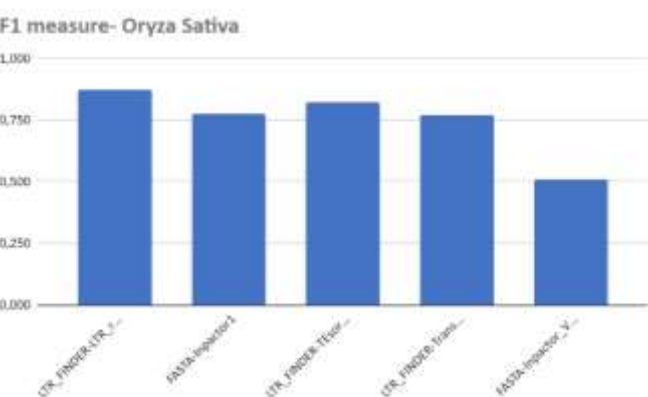


Figura 12. Medida F1 por líneas de trabajo

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

Test Sensibility for Inpactor_v1.3

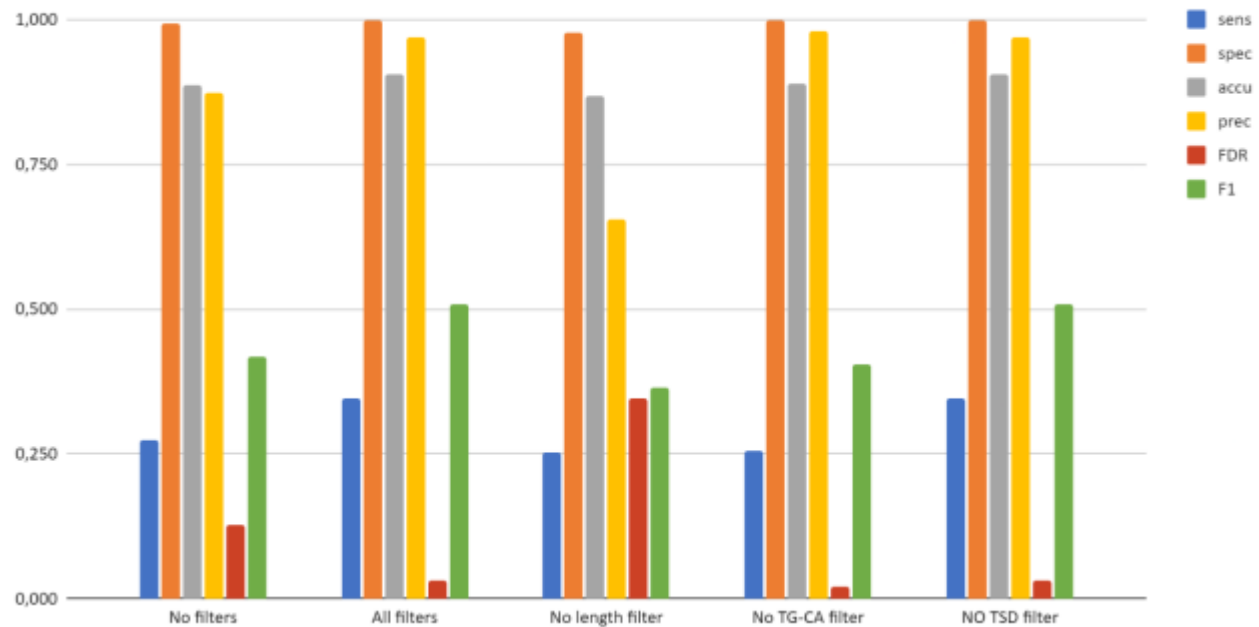



Figura 13. Resultados gráficos de las métricas obtenidas para Inpactor v1.3 en la variación de parámetros para mejorar el rendimiento

En relación a los resultados obtenidos en la ejecución de todos los softwares y reconociendo las posibilidad de mejora para el algoritmo propuesto basado en técnicas de aprendizaje de máquina, en la Figura 13 se presentan los resultados de las algunas de las pruebas de sensibilidad, ilustrando claramente la influencia de los parámetros de filtrado y de curación en el mejoramiento del rendimiento de la herramienta, como parámetros determinantes en la astringencia de los modelos, con el fin de adaptarlos para lograr un rendimiento óptimo en la clasificación de ET al nivel más profundo posible.

7. DISCUSIÓN DE RESULTADOS

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

Debido a la naturaleza repetitiva de los ET, una detección y clasificación bien curada de estos es importante, pues su anotación podría proporcionar información sobre la dinámica genómica. En los últimos años, aunque se han desarrollado muchas herramientas para detectar y clasificar ET basadas principalmente en estructura, en homología, de novo y usando genómica comparativa, la complejidad de estos elementos debido a sus estructuras polimórficas, no permite resultados precisos, confiables y reproducibles para todos los tipos de ET, evidenciado en la carencia de sensibilidad y especificidad, sumado al hecho de que la mayoría de ellos solo pueden identificar clases u órdenes específicos. Algunos de los softwares bioinformáticos que emplean dichas técnicas bioinformáticas convencionales encontrados en la literatura son LTR_FINDER, LTR_retriever, LTRHarvest, PASTEC e Inpactor.

Por lo tanto, varios estudios han propuesto y evaluado el uso de aprendizaje de máquina para el análisis de ET que, aprovechando los miles de secuencias disponibles en conjuntos de datos, como Repbase [74], RepetDB [75], PGSB [76] e InpactorDB [77] y la posibilidad de emplear esquemas de codificación para la representación correcta de los nucleótidos siguiendo diferentes enfoques, mejoran la precisión y rendimiento de la clasificación de ET (incluso hasta el nivel de linajes). Softwares como RED [78], TEClass [79], TransposonUltimate e Inpactor v1.3 emplean técnicas actuales de aprendizaje automático como Máquinas de vectores de soporte (SVM), bosques aleatorios (RF), modelos ocultos de Markov (HMM), K-vecinos más cercanos (KNN), redes neuronales (NN) y modelos gráficos, que representan grandes ventajas en la clasificación de LTR-RT, la extracción de características, la automatización de procesos y ejecuciones más rápidas de los algoritmos, reduciendo así el costo computacional, valiéndose de GPU y pudiendo analizar grandes genomas en un tiempo reducido. Además, hasta ahora se han publicado varias arquitecturas de redes neuronales profundas (DNN) que realizan la clasificación de ET, como la red neuronal completamente conectada (FNN) de Nakano [80] la red neuronal convolucional (CNN) con representación 2D de las secuencias de da Cruz [81] y la red neuronal convolucional (CNN) en 1D para clasificar a los ET en superfamilias de Yan [82].

Aquí, hemos demostrado desde la evaluación comparativa entre las técnicas basadas en herramientas bioinformáticas convencionales y las basadas en aprendizaje de máquina, que estas últimas ofrecen mejores resultados siendo más específicas y precisas en la clasificación de LTR-RT, con la menor tasa de descubrimientos falsos a mayor velocidad y realizando sus tareas en varios procesos a la vez. Al

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

realizar los análisis de los resultados, encontramos que, con todos los filtros, la herramienta propuesta alcanza una especificidad de 99% y una precisión de 97%, además de una tasa de descubrimiento falso de apenas el 3%, en cuestión de minutos, automatizado y acelerado debido a su ejecución a varios procesos. Todo esto, para el genoma de *Oryza sativa*, con tamaño de 389 Mb. En relación a esto, los resultados de las pruebas de sensibilidad para la propuesta del algoritmo basado en aprendizaje de máquina, ilustran la influencia de los parámetros de filtrado y de curación en el mejoramiento del rendimiento de la herramienta, como parámetros determinantes en la astringencia de los modelos para lograr la clasificación de ET incluso a nivel de linajes. Sin embargo, cabe destacar que el filtrado de falsos positivos (como hace LTR_retriever), perfeccionaría los resultados, y proporcionaría un rendimiento óptimo a la herramienta.

Gracias a los resultados obtenidos anteriormente, en relación a la metodología seguida y planteada, y teniendo en cuenta la literatura, la herramienta que está en proceso de construcción, la cual fue usada en su versión 1.3 para este análisis comparativo, puede adaptar sus modelos de clasificación para lograr un rendimiento óptimo entre la sensibilidad y la medida F1 (por la cual se evalúa el modelo), dado que en sus demás métricas su rendimiento es excelente.

8. CONCLUSIONES

La selección de las especies a estudiar, es relevante puesto que se deben contar con secuencias curadas de alta calidad y en la medida de las posibilidades no redundantes, para obtener un rendimiento más óptimo de las herramientas en general.

Este análisis comparativo permite decidir entre una herramienta frente a la otra según aplique herramientas bioinformáticas convencionales o utilice técnicas de aprendizaje de máquina para detectar y clasificar LTR-RT hasta superfamilias, según los recursos y necesidades particulares del proyecto.

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

Emplear técnicas basadas en aprendizaje de máquina para detectar y clasificar ET, y LTR-RT específicamente, es novedoso, se vale del aprendizaje automático y de una serie de potentes herramientas que permiten generalizar los datos y entre otras cosas, analizar gran cantidad de información en poco tiempo, gracias a la ejecución multiprocesos y las técnicas de alta computación.


La metodología empleada en el desarrollo del proyecto permite continuar las evaluaciones comparativas entre especies de plantas, más softwares y la variación de parámetros en el algoritmo propuesto.

El análisis aquí evidenciado permite denotar la importancia de identificar los ET, específicamente los LTR retrotransposones, debido al impacto que presentan para el desarrollo evolutivo de las especies, lo cual permite profundizar en el conocimiento del genoma de las mismas. Además, sirve de insumo a investigaciones futuras, en las que sea necesario la detección y clasificación de estas estructuras, garantizando un proceso automático y óptimo.

Para finalizar, es posible afirmar que el problema de identificación y clasificación de LTR retrotransposones, puede ser unificado en una herramienta computacional, que tenga un rendimiento óptimo en todas sus métricas, mediante la utilización de un modelo automático para la resolución del problema multiclase, en el cual, un LTR-RT, puede ser clasificado correctamente según el linaje al que pertenece.

9. RECOMENDACIONES

En estudios posteriores, se recomienda tener en cuenta más especies con alta calidad en el ensamblaje, con el fin de analizar mejor la dinámica y comportamiento de los softwares según superfamilias, para tener resultados más robustos y significativos de distintas especies. En ese sentido, las secuencias deberían formar parte de librerías de LTR-RTs completamente curadas no redundantes. Es de destacar que es un proceso complejo; sin embargo, abriría los caminos para la realización de

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

proyectos relacionados con la identificación y clasificación hasta nivel de linajes de ET para distintas especies.

Po otro lado, se recomienda utilizar la mayor cantidad de procesadores para el algoritmo propuesto, puesto que este ejecuta procesos en paralelo, y así reduce ampliamente los tiempos en la obtención de los datos.

Además, se recomienda usar los resultados aquí mostrados para argumentar la elección de una herramienta u otra según su rendimiento por evaluación comparativa.

Y, en definitiva, se recomienda continuar la evaluación comparativa a medida se vayan desarrollando herramientas para la detección y clasificación de LTR-RT, sin importar la metodología empleada, pues la idea es evaluar entre todas, la mejor o más óptima según las necesidades y recursos.


10. EVIDENCIA DE RESULTADOS

10.1 Generación de conocimiento y/o nuevos desarrollos tecnológicos

Resultado/Producto esperado	Indicador	Beneficiario
Artículo de revisión y comparación	Artículo en revista indexada	Comunidad académica nacional e internacional

10.2 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 DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

10.3 Apropiación social del conocimiento

Resultado/Producto esperado	Indicador	Beneficiario
Ponencia	Participación en Encuentro departamental de Semilleros de investigación	Estudiantes UAM
Ponencia	Participación en el 2do Congreso Latinoamericano de Mujeres en Bioinformática y Ciencia de Datos	Comunidad bioinformática nacional e internacional
Divulgación	Presentación de resultados en el foro de investigación UAM	Comunidad UAM

11. IMPACTOS LOGRADOS

Impacto esperado	Plazo (años) después de finalizado el proyecto: corto (1-4), mediano (5-9), largo (10 o más)	Indicador verificable	Supuestos ²
Contribuir a la formación de estudiantes	Corto (1 - 4 años)	Número de estudiantes vinculados al semillero	Divulgación en foros de investigación y páginas web en la comunidad UAM.
Contribuir a mejoras en los modelos de aprendizaje de máquina para	Corto (1 - 4 años)	Resultados cuantitativos y	Divulgación de resultados entre se semillero y los


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

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

identificación y clasificación de LTR-RT, desde el semillero de investigación		cualitativos de los proyectos ejecutados	grupos de investigación aliados
Contribuir al desarrollo eficiente de anotación de genomas de plantas, gracias a la identificación de las mejores herramientas dado su rendimiento	Corto (5 - 9 años)	Cantidad de proyectos investigativos relacionados con la anotación de los LTR-RT en plantas, haciéndolo eficientemente.	Divulgación de resultados en revista indexada

12. BIBLIOGRAFÍA


1. 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.
2. McClintock, B. The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. USA 1950, 36, 344–355.
3. 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.
4. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, et al., Correction: Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. PLOS Genetics 11(10): e1005566, 2015. <https://doi.org/10.1371/journal.pgen.1005566>
5. Serrato-Capuchina A. and Matute D., "The Role of Transposable Elements in Speciation," Genes, vol. 9, no. 5, p. 254, 2018.
6. Feschotte C, Pritham EJ. "DNA transposons and the evolution of eukaryotic genomes". Annu. Rev. Genet. 41:331–68, 2007.

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


7. 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.
8. 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.
9. Feschotte C, Pritham EJ. "DNA transposons and the evolution of eukaryotic genomes". *Annu. Rev. Genet.* 41:331–68, 2007.
10. 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
11. 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.
12. György Abrusán, Norbert Grundmann, Luc DeMester, Wojciech Makalowski, TEclass—a tool for automated classification of unknown eukaryotic transposable elements, *Bioinformatics*, Volume 25, Issue 10, 15 May 2009, Pages 1329–1330, <https://doi.org/10.1093/bioinformatics/btp084>
13. "Earth BioGenome Project", Earth BioGenome Project. [Online]. Available: <https://www.earthbiogenome.org/>.
14. "10KP: 10,000 Plant Genomes Project", Db.cngb.org. [Online]. Available: <https://db.cngb.org/10kp/>.
15. Orozco-arias, S., Isaza, G., Guyot, R., & Tabares-soto, R. (2019). A systematic review of the application of machine learning in the detection and classification of transposable elements. *Peerj*, 7, 18311. <https://doi.org/10.7717/peerj.8311>
16. Loureiro, T., Fonseca, N., & Camacho, R. (2012). Application of Machine Learning techniques on the Discovery and annotation of Transposons in genomes. 1, 1–3. <http://paginas.fe.up.pt/~ei07087/dokuwiki/files/Abstract.pdf>
17. F. K. Nakano, W. J. Pinto, G. L. Pappa and R. Cerri, "Top-down strategies for hierarchical classification of transposable elements with neural networks," 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2539-2546, doi: 10.1109/IJCNN.2017.7966165.

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

18. M. da Cruz, D. Domingues, P. Saito, A. Paschoal and P. Bugatti, "TERL: classification of transposable elements by convolutional neural networks", Briefings in Bioinformatics, 2020. doi: 10.1093/bib/bbaa185
19. Orozco-Arias, S.; Piña, J.S.; Tabares-Soto, R.; Castillo-Ossa, L.F.; Guyot, R.; Isaza, G. Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements. Processes 2020, 8, 638. <https://doi.org/10.3390/pr8060638>
20. Ou S, Su W. The Extensive de-novo TE Annotator. GitHub. Available from: <https://github.com/oushujun/EDTA>
21. 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>
22. Zhao Xu, Hao Wang, LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, Nucleic Acids Research, Volume 35, Issue suppl_2, 1 July 2007, Pages W265–W268, <https://doi.org/10.1093/nar/gkm286>
23. Ren-Gang Zhang, Zhao-Xuan Wang, Shujun Ou, Guang-Yuan Li, TESorter: lineage-level classification of transposable elements using conserved protein domains, doi: <https://doi.org/10.1101/800177>
24. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, et al. (2014) PASTEC: An Automatic Transposable Element Classification Tool. PLOS ONE 9(5): e91929. <https://doi.org/10.1371/journal.pone.0091929>
25. 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.
26. Makołowski W., Gotea V., Pande A., Makołowska I. (2019) Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. In: Anisimova M. (eds) Evolutionary Genomics. Methods in Molecular Biology, vol 1910. Humana, New York, NY. https://doi.org/10.1007/978-1-4939-9074-0_6
27. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning in genetics and genomics. Nature Review Genetics, 16(6), 321–332. <https://doi.org/10.1038/nrg3920>

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


28. Orozco-Arias, S.; Piña, J.S.; Tabares-Soto, R.; Castillo-Ossa, L.F.; Guyot, R.; Isaza, G. Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements. *Processes* 2020, 8, 638. <https://doi.org/10.3390/pr8060638>
29. Arango-López, J., Orozco-Arias, S., Salazar, J. A., Guyot, R., Arango-Lopez, J., Orozco-Arias, S., Salazar, J. A., & Guyot, R. (2017). Application of Data Mining Algorithms to Classify Biological Data: The *Coffea canephora* Genome Case. In *Communications in Computer and Information Science* (Vol. 735, pp. 156–170). https://doi.org/10.1007/978-3-319-66562-7_12
30. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. USA* 1950, 36, 344–355.
31. McClintock, B. (1953). Induction of Instability at Selected Loci in Maize. *Genetics*, 38(6), 579–599. <http://www.ncbi.nlm.nih.gov/pubmed/17247459> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1209627>
32. Orozco-arias, S., Isaza, G., Guyot, R., & Tabares-soto, R. (2019). A systematic review of the application of machine learning in the detection and classification of transposable elements. *Peerj*, 7, 18311. <https://doi.org/10.7717/peerj.8311>
33. Gao, D., Jimenez-Lopez, J. C., Iwata, A., Gill, N., & Jackson, S. A. (2012). Functional and Structural Divergence of an Unusual LTR Retrotransposon Family in Plants. *PLoS ONE*, 7(10), 1–12. <https://doi.org/10.1371/journal.pone.0048595>
34. Orozco-arias, S., Isaza, G., Guyot, R., & Tabares-soto, R. (2019). A systematic review of the application of machine learning in the detection and classification of transposable elements. *Peerj*, 7, 18311. <https://doi.org/10.7717/peerj.8311>
35. Orozco-Arias, S., Isaza, G., & Guyot, R. (2019). Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning. *International Journal of Molecular Sciences*, 20(15), 1–29. <https://doi.org/10.3390/ijms20153837>
36. Orozco-arias, S., Isaza, G., Guyot, R., & Tabares-soto, R. (2019). A systematic review of the application of machine learning in the detection and classification of transposable elements. *Peerj*, 7, 18311. <https://doi.org/10.7717/peerj.8311>
37. Loureiro, T., Fonseca, N., & Camacho, R. (2012). Application of Machine Learning techniques on the Discovery and annotation of Transposons in genomes. 1, 1–3. <http://paginas.fe.up.pt/~ei07087/dokuwiki/files/Abstract.pdf>

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


38. Loureiro, T., Camacho, R., Vieira, J., & Fonseca, N. A. (2013). Improving the performance of Transposable Elements detection tools. *Journal of Integrative Bioinformatics*, 10(3), 231. <https://doi.org/10.2390/biecoll-jib-2013-231>
39. Orozco-Arias, S., Isaza, G., & Guyot, R. (2019). Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning. *International Journal of Molecular Sciences*, 20(15), 1–29. <https://doi.org/10.3390/ijms20153837>
40. Xu, Z., & Wang, H. (2007). LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, 35(SUPPL.2), 265–268. <https://doi.org/10.1093/nar/gkm286>
41. Ou, S., & Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology*, 176(2), 1410–1422. <https://doi.org/10.1104/pp.17.01310>
42. McCarthy, E. M., & McDonald, J. F. (2003). LTR STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics*, 19(3), 362–367. <https://doi.org/10.1093/bioinformatics/btf878>
43. Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9. <https://doi.org/10.1186/1471-2105-9-18>
44. Orozco-arias, S., Isaza, G., Guyot, R., & Tabares-soto, R. (2019). A systematic review of the application of machine learning in the detection and classification of transposable elements. *Peerj*, 7, 18311. <https://doi.org/10.7717/peerj.8311>
45. Orozco-arias, S., Liu, J., Id, R. T., Ceballos, D., Silva, D., Id, D., Ming, R., & Guyot, R. (2018). Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics. *Biology*. <https://doi.org/10.3390/biology7020032>
46. Monat, C., Tando, N., Tranchant-Dubreuil, C., & Sabot, F. (2016). LTRclassifier: A website for fast structural LTR retrotransposons classification in plants. *Mobile Genetic Elements*, 6(6), e1241050. <https://doi.org/10.1080/2159256x.2016.1241050>
47. Loureiro, T., Camacho, R., Vieira, J., & Fonseca, N. A. (2013). Improving the performance of Transposable Elements detection tools. *Journal of Integrative Bioinformatics*, 10(3), 231. <https://doi.org/10.2390/biecoll-jib-2013-231>

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

48. Orozco-arias, S., Isaza, G., Guyot, R., & Tabares-soto, R. (2019). A systematic review of the application of machine learning in the detection and classification of transposable elements. *Peerj*, 7, 18311. <https://doi.org/10.7717/peerj.8311>
49. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, 51(1), 12–18. <https://doi.org/10.1038/s41588-018-0295-5>
50. Orozco-arias, S., Isaza, G., Guyot, R., & Tabares-soto, R. (2019). A systematic review of the application of machine learning in the detection and classification of transposable elements. *Peerj*, 7, 18311. <https://doi.org/10.7717/peerj.8311>
51. Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Bucheli, V. S., Rodríguez-Sotelo, J. L., & Jiménez-Varón, C. F. (2020). A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Computer Science*, 2020(4), 1–22. <https://doi.org/10.7717/peerj-cs.270>
52. Orozco-arias, S., Liu, J., Id, R. T., Ceballos, D., Silva, D., Id, D., Ming, R., & Guyot, R. (2018). Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics. *Biology*. <https://doi.org/10.3390/biology7020032>
53. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, et al. An active DNA transposon family in rice. *Nature*. 2003;421:163–7.
54. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. *Nature*. 2004;431:569–73.
55. Feschotte C, Swamy L, Wessler SR. Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics*. 2003;163:747–58.
56. Xie Y, Wang Y, Wu R. A rice DNA sequence that resembles the maize Mu 1 transposable element. *Rice Genetics Collect*. 2008;2:377–87.
57. Barret P, Brinkman M, Beckert M. A sequence related to rice Pong transposable element displays transcriptional activation by in vitro culture and reveals somaclonal variations in maize. *Genome*. 2006;49:1399–407.D
58. F"TAIR - Home Page", *Arabidopsis.org*. [Online]. Available: <https://www.arabidopsis.org/index.jsp>. [Accessed: 01- Jun- 2021].

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

59. Ou S, Su W. The Extensive de-novo TE Annotator. GitHub. Available from: <https://github.com/oushujun/EDTA>
60. Orozco-Arias, S.; Jaimes, P.A.; Candamil, M.S.; Jiménez-Varón, C.F.; Tabares-Soto, R.; Isaza, G.; Guyot, R. InpactorDB: A Classified Lineage-Level Plant LTR Retrotransposon Reference Library for Free-Alignment Methods Based on Machine Learning. *Genes* 2021, 12, 190. <https://doi.org/10.3390/genes12020190>
61. Smit AFA, Hubley R. RepeatModeler Open-1.0. 2008—2015. 2015. Available from: www.repeatmasker.org
62. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6:11.
63. McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*. 2003;19:362–7.
64. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. 2007;35:W265–8.
65. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*. 2008;9:18.
66. Lee H, Lee M, Mohammed Ismail W, Rho M, Fox GC, Oh S, et al. MGEScan: a Galaxy-based system for identifying retrotransposons in genomes. *Bioinformatics*. 2016;32:2502–4.
67. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol*. 2018;176:1410–22
68. Valencia JD, Girgis HZ. LtrDetector: a tool-suite for detecting long terminal repeat retrotransposons de-novo. *BMC Genomics*. 2019;20:450.
69. Shi J, Liang C. Generic Repeat Finder: a high-sensitivity tool for genome-wide de novo repeat detection. *Plant Physiology*. 2019;00386. <https://doi.org/10.1104/pp.19.00386>.
70. Orozco-arias, S., Liu, J., Id, R. T., Ceballos, D., Silva, D., Id, D., Ming, R., & Guyot, R. (2018). Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics. *Biology*. <https://doi.org/10.3390/biology7020032>

	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


71. Ren-Gang Zhang, Zhao-Xuan Wang, Shujun Ou, Guang-Yuan Li, TEsorter: lineage-level classification of transposable elements using conserved protein domains, doi: <https://doi.org/10.1101/800177>
72. Riehl, K, Riccio,C, Miska, E, Hemberg, M. bioRxiv 2021.04.30.442214; doi: <https://doi.org/10.1101/2021.04.30.442214>
73. Shujun Ou, Ning Jiang, LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons ,Retrotransposons , Plant Physiology, Volume 176, Issue 2, February 2018, Pages 1410–1422, <https://doi.org/10.1104/pp.17.01310>
74. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
75. Cornut, G.; Choisne, N.; Alaux, M.; Alfama-Depauw, F.; Jamilloux, V.; Maumus, F.; Letellier, T.; Luyten, I.; Pommier, C.; Adam-Blondon, A.-F.; et al. RepetDB: A unified resource for transposable element references. Mob. DNA 2019, 10, 6.
76. Spannagl, M.; Nussbaumer, T.; Bader, K.C.; Martis, M.M.; Seidel, M.; Kugler, K.G.; Gundlach, H.; Mayer, K.F.X. PGSB PlantsDB: Updates to the database framework for comparative plant genome research. Nucleic Acids Res. 2015, 44, D1141–D1147.
77. Orozco-Arias, S.; Jaimes, P.A.; Candamil, M.S.; Jiménez-Varón, C.F.; Tabares-Soto, R.; Isaza, G.; Guyot, R. InpactorDB: A Classified Lineage-Level Plant LTR Retrotransposon Reference Library for Free-Alignment Methods Based on Machine Learning. Genes 2021, 12, 190. <https://doi.org/10.3390/genes12020190>
78. Girgis HZ. 2015. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 16(1):227
79. György Abrusán, Norbert Grundmann, Luc DeMester, Wojciech Makalowski, TEclass—a tool for automated classification of unknown eukaryotic transposable elements, Bioinformatics, Volume 25, Issue 10, 15 May 2009, Pages 1329–1330, <https://doi.org/10.1093/bioinformatics/btp084>
80. Nakano, F.K.; Mastelini, S.M.; Barbon, S.; Cerri, R. Improving Hierarchical Classification of Transposable Elements Using Deep Neural Networks. In Proceedings of the International Joint Conference on Neural Networks, Rio de Janeiro, Brazil, 8–13 July 2018.
81. Da Cruz, M.H.P.; Domingues, D.S.; Saito, P.T.M.; Paschoal, A.R.; Bugatti, P.H. TERL: Classification of Transposable Elements by Convolutional Neural Networks. bioRxiv 2020.

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

82. Yan, H.; Bombarely, A.; Li, S. DeepTE: A computational method for de novo classification of transposons with convolutional neural network. Bioinformatics 2020.

13. ANEXOS

Anexo 1. Información acerca de las especies utilizadas en el estudio (Ver en Carpeta)

	<p align="center">PRESENTACIÓN DE INFORMES FINALES UAM</p>	<p>CÓDIGO: GIN-GUI-001</p>
		<p>VERSIÓN: 01</p>
		<p>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</p>

Anexo 2. Extracción de secuencias de InpactorDB para anotar y crear librería estándar de Arabidopsis thaliana (Ver en Carpeta)



	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015


Anexo 3. Paso a paso para ejecutar las líneas de trabajo de clasificación hasta la extracción de métricas



	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

**Anexo 4. Script para la conversión de formato de salida de Inpactor versión 1 a formato RM
(Ver en Carpeta)**



	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

Anexo 5. Script para la conversión de formato de salida de TransposonUltimate a formato RM (Ver en Carpeta)



	PRESENTACIÓN DE INFORMES FINALES UAM	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

Anexo 6. Script para el cálculo de métricas según superfamilias (Ver en Carpeta)

