PROCESOS DE INVESTIGACIÓN II

UNIVERSIDAD AUTÓNOMA DE MANIZALES
UAM®
VIGILADA MINEDUCACIÓN

Análisis comparativo de técnicas basadas en aprendizaje de máquina para detectar y clasificar LTR-retrotransposones en plantas

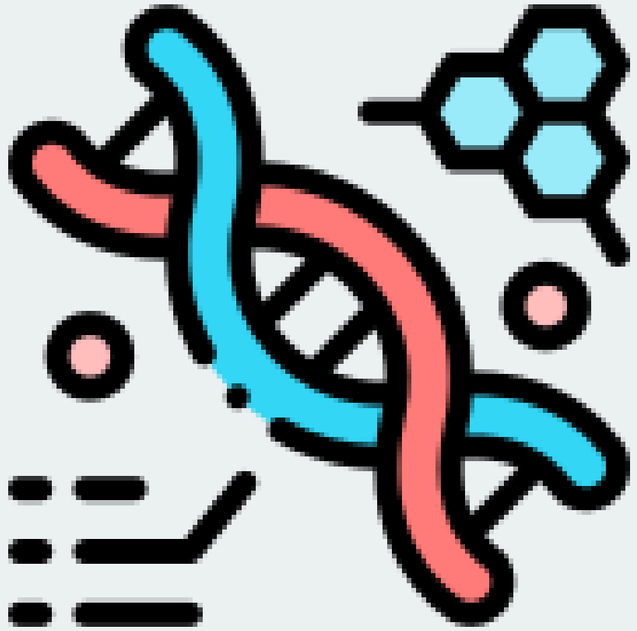
Estudiante:
Maradey Mercedes Arias Mendoza
Co-autor:
Paula Andrea Jaimes Buitrón

Tutor:
Simón Orozco Arias
Co-tutor:
Reinel Tabares Soto

Semillero de Investigación:
Bioinformática e Inteligencia Artificial
Grupos de Investigación:
Automática - Ingeniería de Software

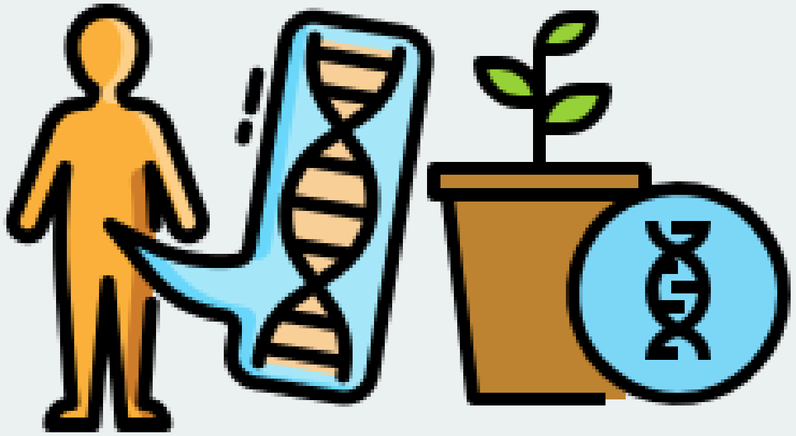
Contenido

INTRODUCCIÓN



Genoma

Procariota

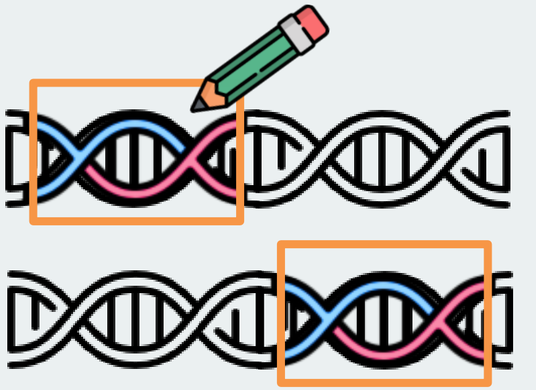


Eucariota

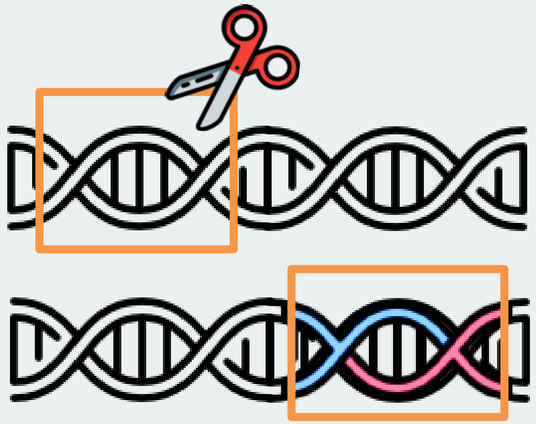
INTRODUCCIÓN

Elementos Transponibles (ET)

Mecanismos



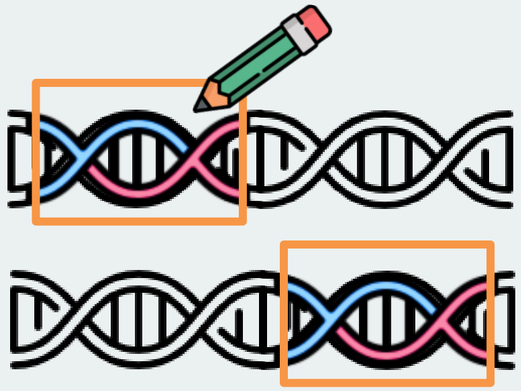
Clase I (Retrotransposones)



Clase II (Transposones)

INTRODUCCIÓN

Clase I (Retrotransposones)



Órdenes

- Long Terminal Repeat (LTR)
- DIRS
- PLE
- LINE
- SINE

Superfamilias

- *Copia*
- *Gypsy*
- *Bel-Pao*

Linajes

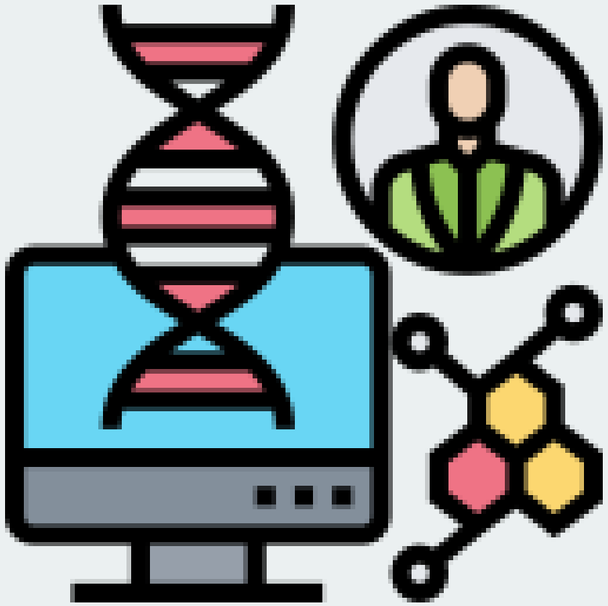
- *ANGELA*
- *BIANCA*
- *TORK*
- *SIRE*
- *GALADRIEL*
- *REINA*
- *ATHILA*

Figura 1. Estructura de los LTR retrotransposones.

Recuperado de Orozco-Arias et al. 2019

INTRODUCCIÓN

Bioinformática

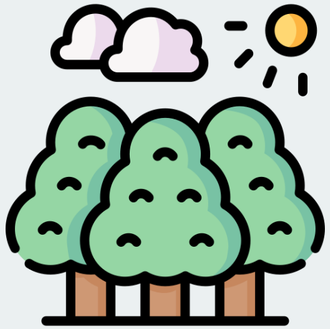


Metodologías/arquitecturas/algoritmos

- Basada en estructura
- Basada en homología
- De novo
- Genómica comparativa
- Basada en ML

INTRODUCCIÓN

Repercusiones - influencia



Adaptabilidad

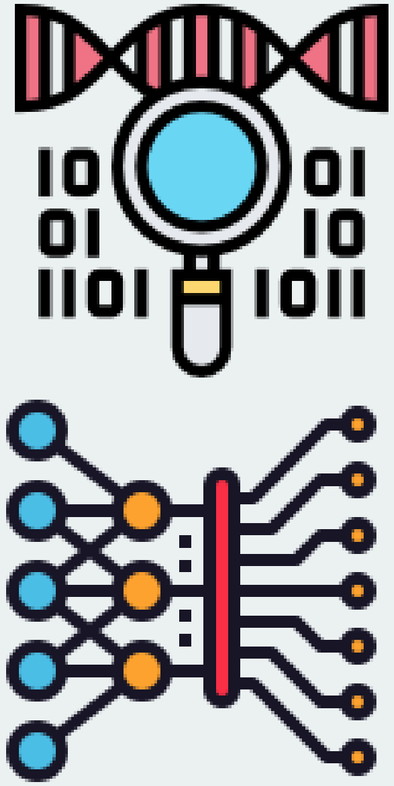
Resistencia a enfermedades

Diversidad

PLANTEAMIENTO DEL PROBLEMA

NGS

(Secuenciación de nueva generación)

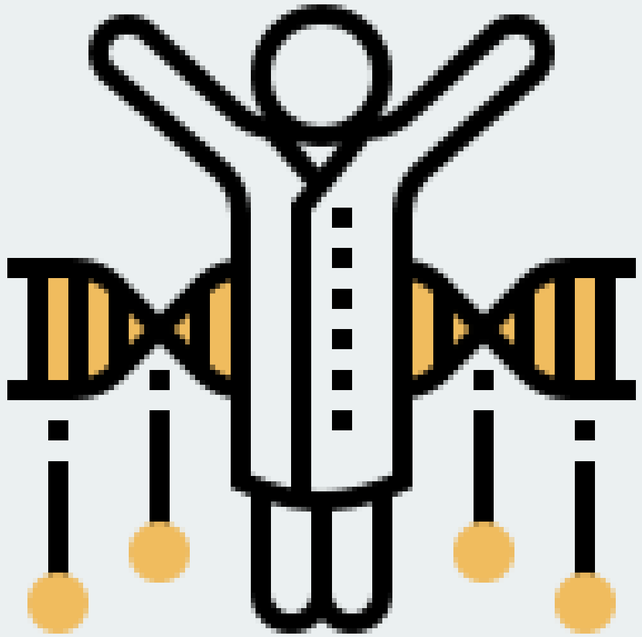


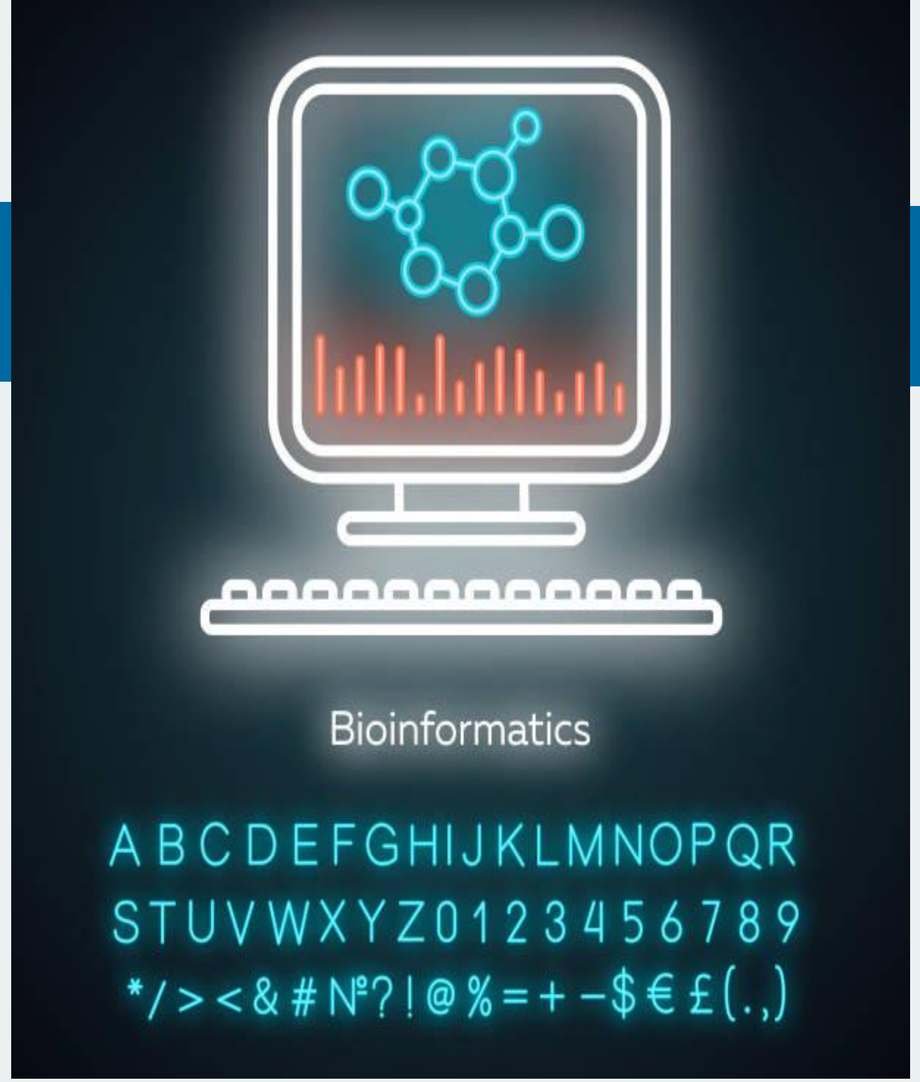
Identificación y clasificación de ET

- TransposonUltimate
- Inpactor
- LTR_FINDER
- LTR_retriever
- EDTA
- TEsorter
- RED
- DeepTE
- Nakano
- TERL

PLANTEAMIENTO DEL PROBLEMA

¿Cuál de las herramientas bioinformáticas 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?





OBJETIVOS

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.

OBJETIVOS

ESPECÍFICOS

- Seleccionar una lista de softwares bioinformáticos basados en aprendizaje de máquina, diferenciada según detecten o clasifiquen LTR retrotransposones.
- Instalar y ejecutar cada software bioinformático con mínimo 2 genomas de plantas, ampliamente reconocidos.
- Extraer métricas de rendimiento para cada software, 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.

METODOLOGÍA

Recolectar información:

*Especies de plantas con alta calidad en el ensamblaje

*Softwares de identificación y clasificación de LTR-RT

Identificación de los LTR-RT en los genomas:

*Resultados previos

*Flujos de trabajo

Extracción de métricas de rendimiento:

*Anotación del genoma

*Métricas dado flujos de trabajo y superfamilias

Análisis comparativo de los resultados:

*Softwares bioinformáticos vs nueva propuesta de algoritmo

METODOLOGÍA

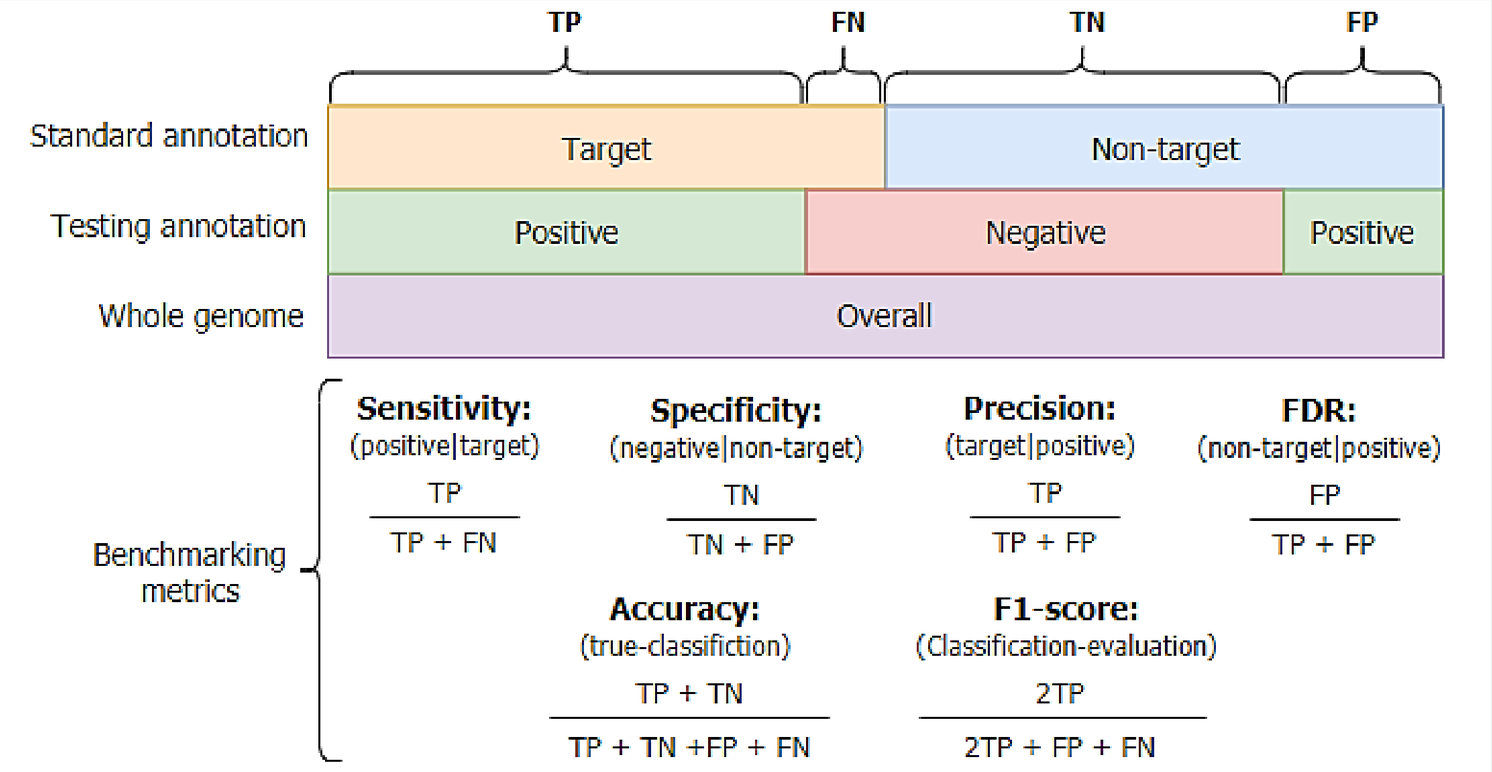


Figura 2. Representación esquemática de las métricas de evaluación comparativa

RESULTADOS

Genomas

Oryza sativa

Arabidopsis thaliana

Extensive de-novo TE Annotator (EDTA) (Ou S, Su W. et al)

InpactorDB (S. Orozco-Arias et al, 2021)

- ❑Conjunto de datos de LTR-RT en plantas
- ❑Alta calidad
- ❑Intactos
- ❑Alrededor de 130400 LTR-RT de 195 especies de plantas

- ❑**612 corresponden a *Arabidopsis thaliana***

- ❑Proyecto del semillero

Librería estándar

RESULTADOS

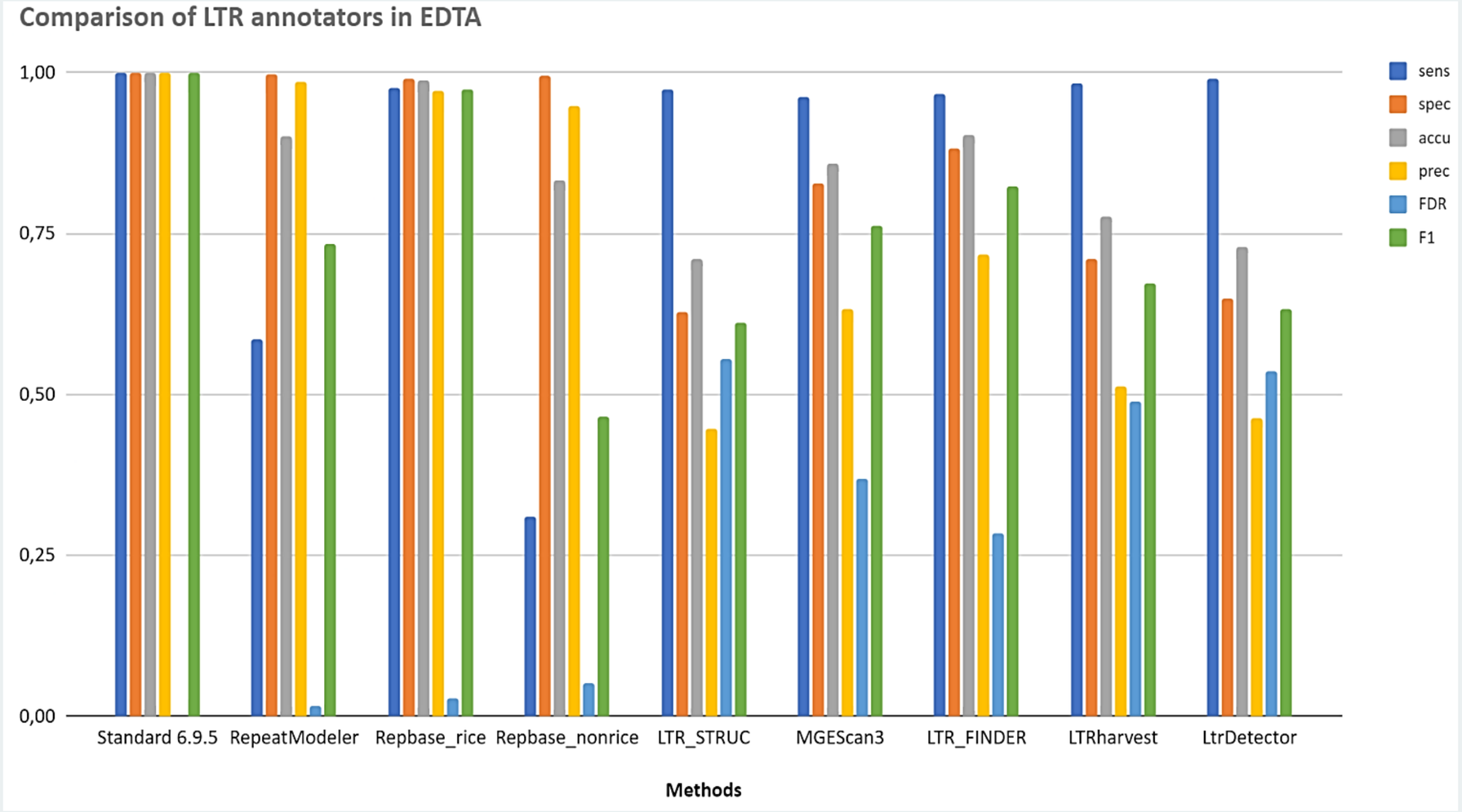


Figura 3. Evaluación comparativa entre dos métodos generales de identificación de repetición con clasificación (RepeatModeler y Repbase) y siete softwares con métodos basados en estructura diseñados específicamente para la identificación de novo LTR.

RESULTADOS

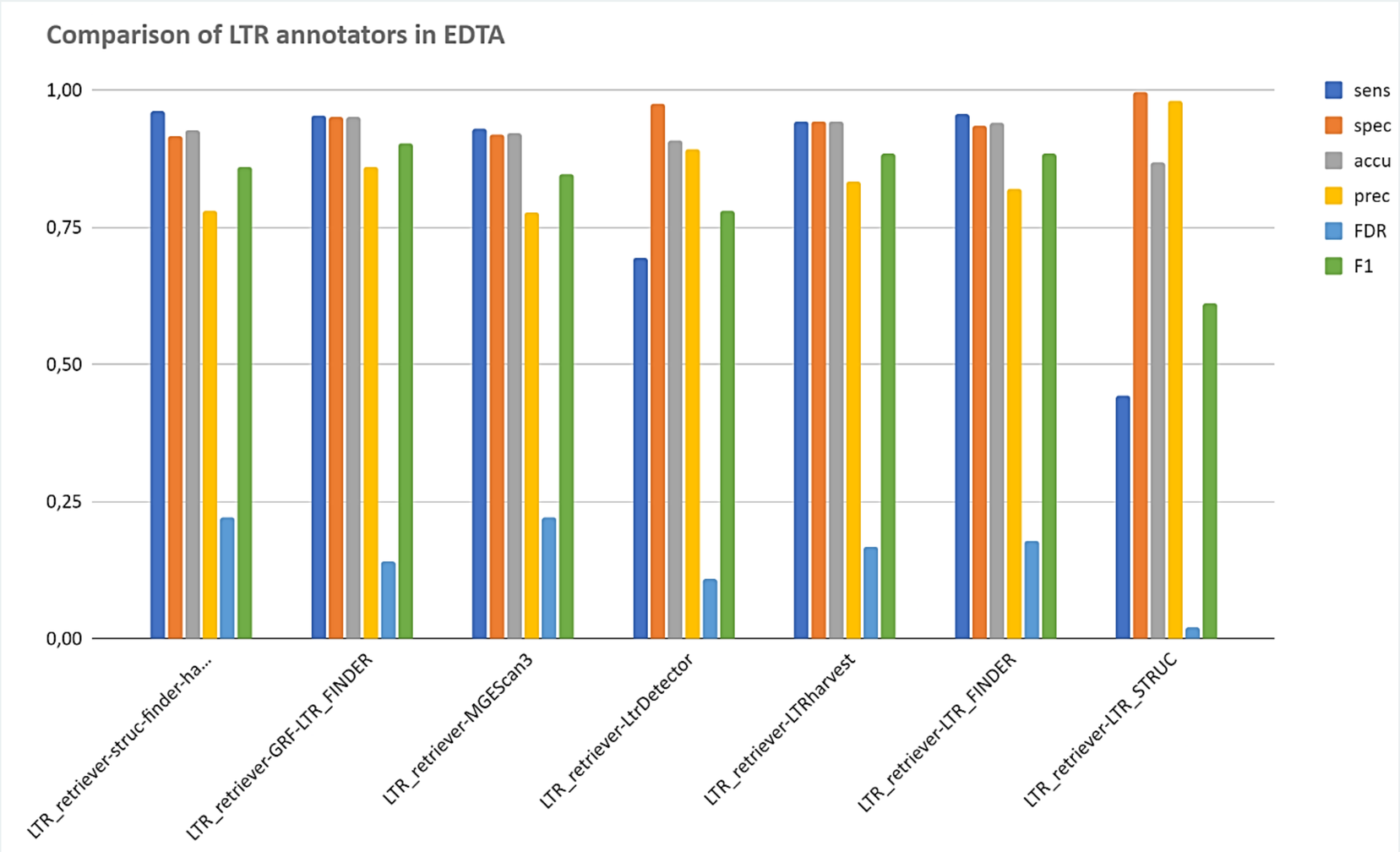


Figura 4. Evaluación comparativa entre LTR-retriever y siete softwares con métodos basados en estructura diseñados específicamente para la identificación de novo LTR

RESULTADOS

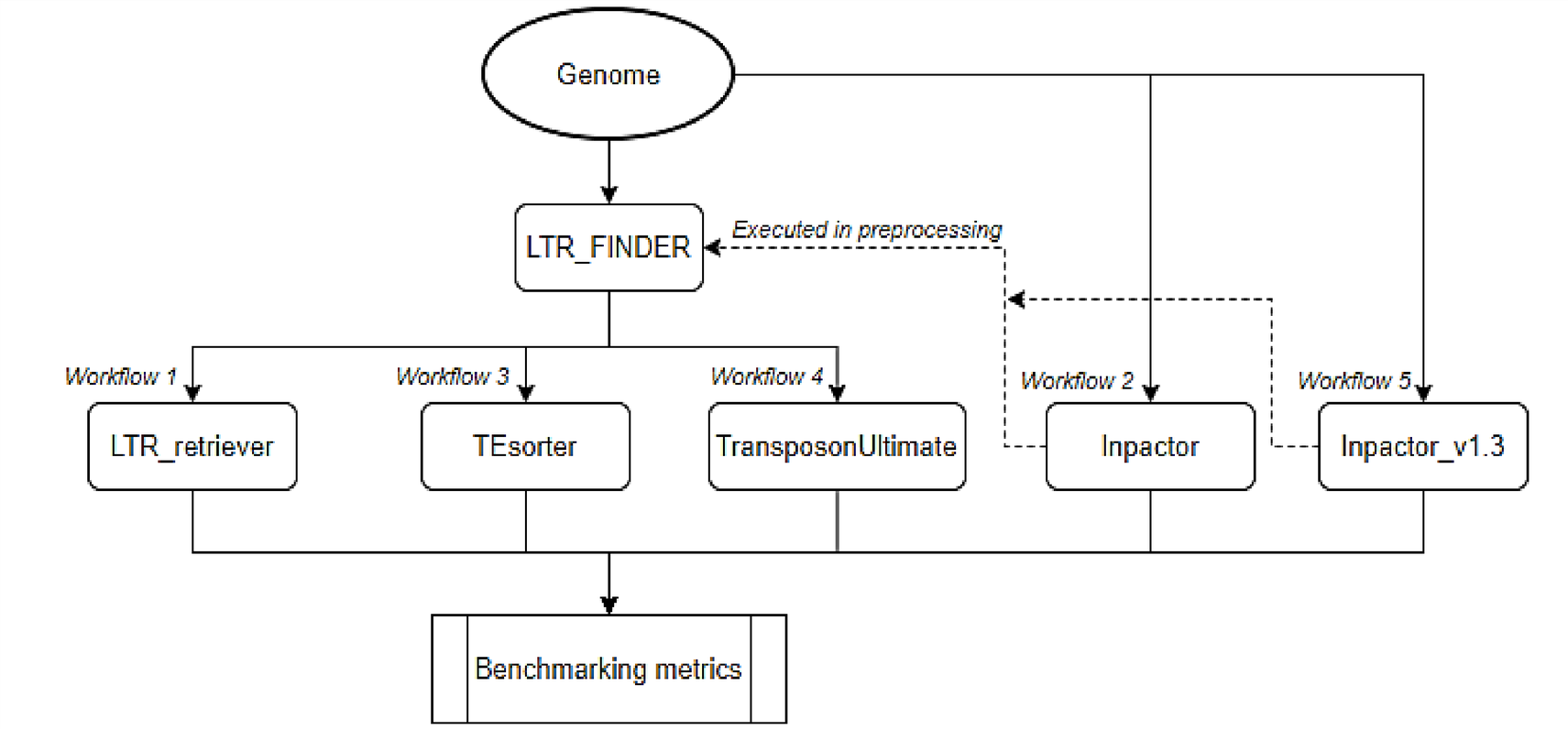


Figura 5. Flujos de trabajo para la clasificación de LTR-RT a nivel de superfamilias

RESULTADOS

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

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

RESULTADOS

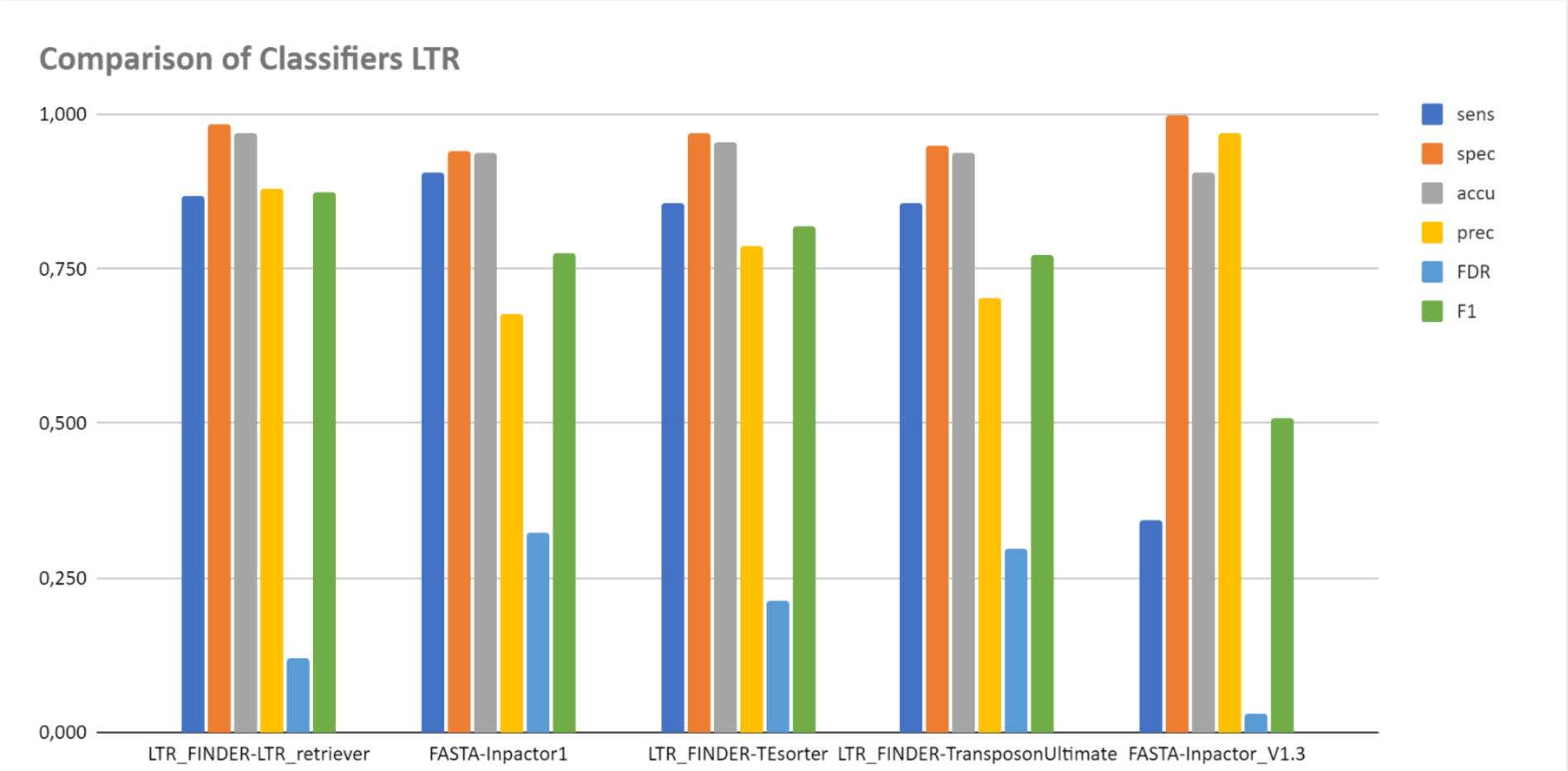


Figura 6. Resultados gráficos de las métricas obtenidas por línea de trabajo para la clasificación de LTR-RT y según la superfamilia.

RESULTADOS

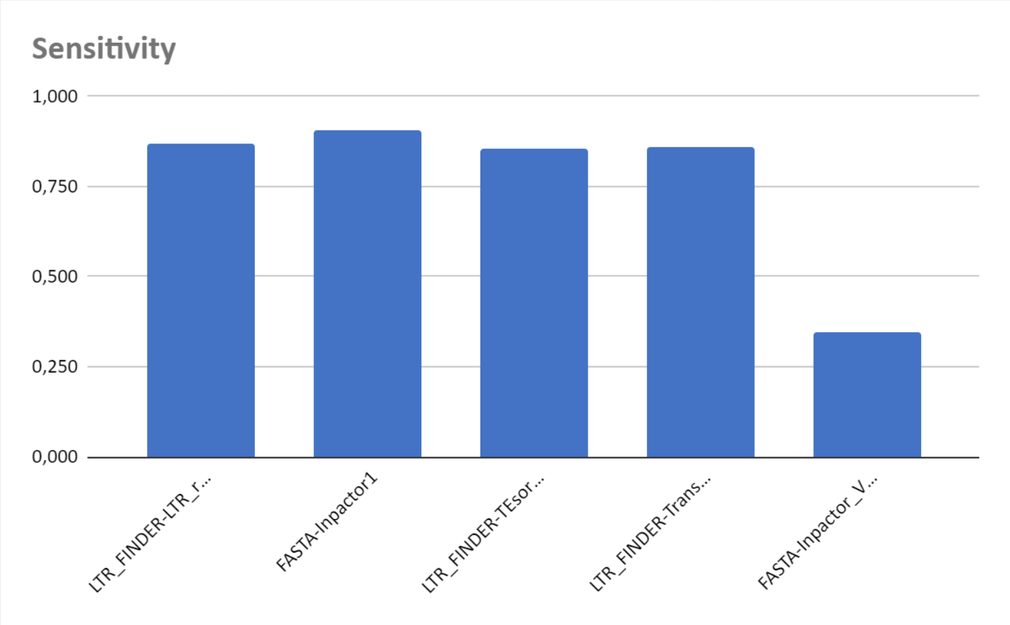


Figura 7. Sensibilidad por líneas de trabajo, *Oryza sativa*

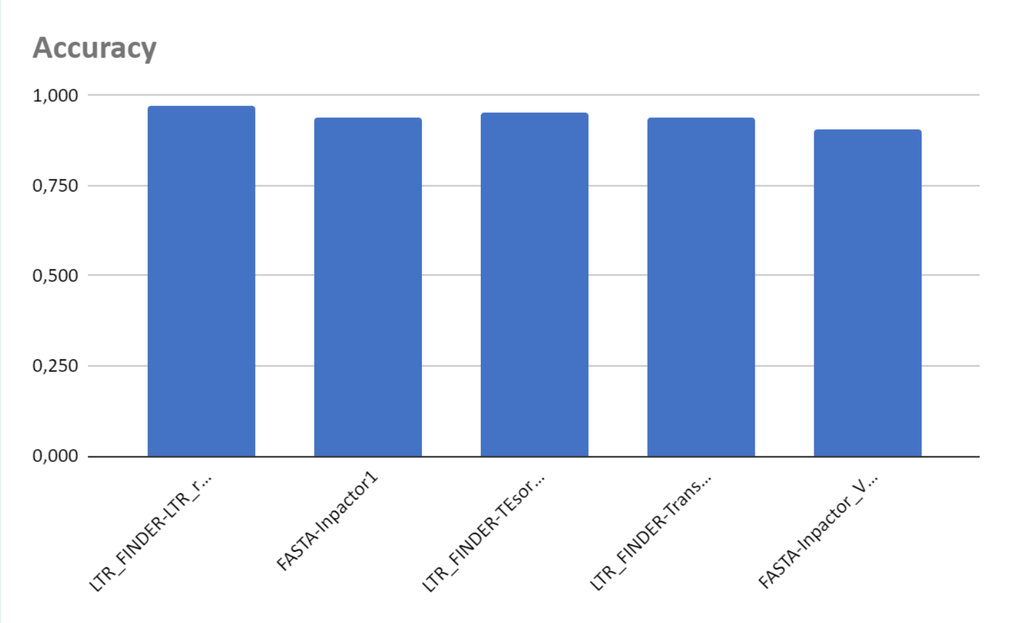


Figura 8. Exactitudpor líneas de trabajo, Oryza sativa

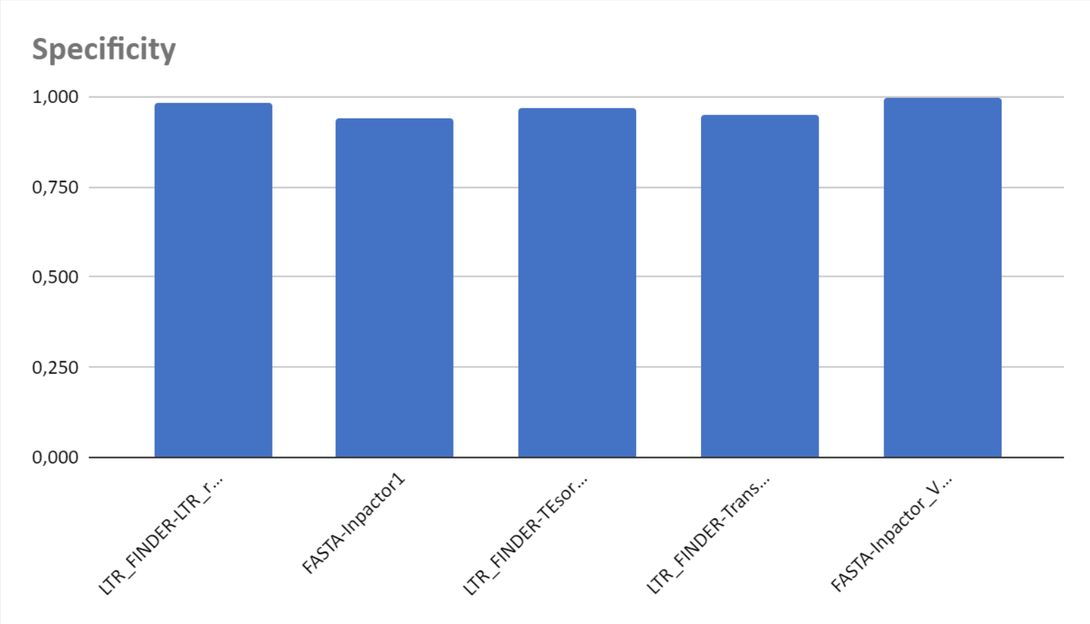


Figura 9. Especificidad por líneas de trabajo, Oryza sativa

UNIVERSIDAD AUTÓNOMA DE MANIZALES
UAM®
VIGILADA MINEDUCACIÓN

RESULTADOS

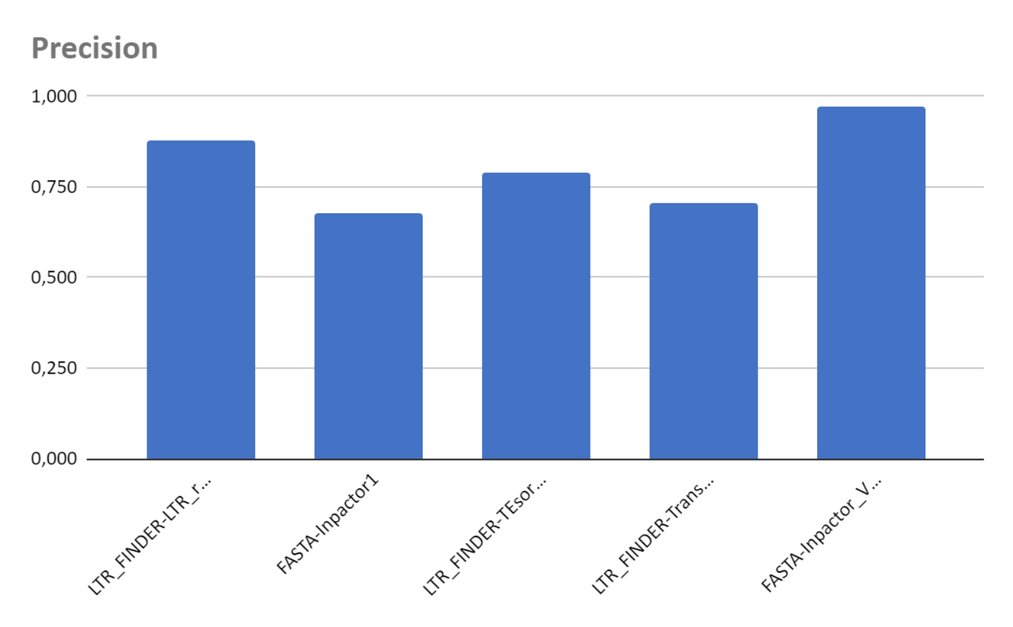


Figura 10. Precisiónpor líneas de trabajo, Oryza sativa

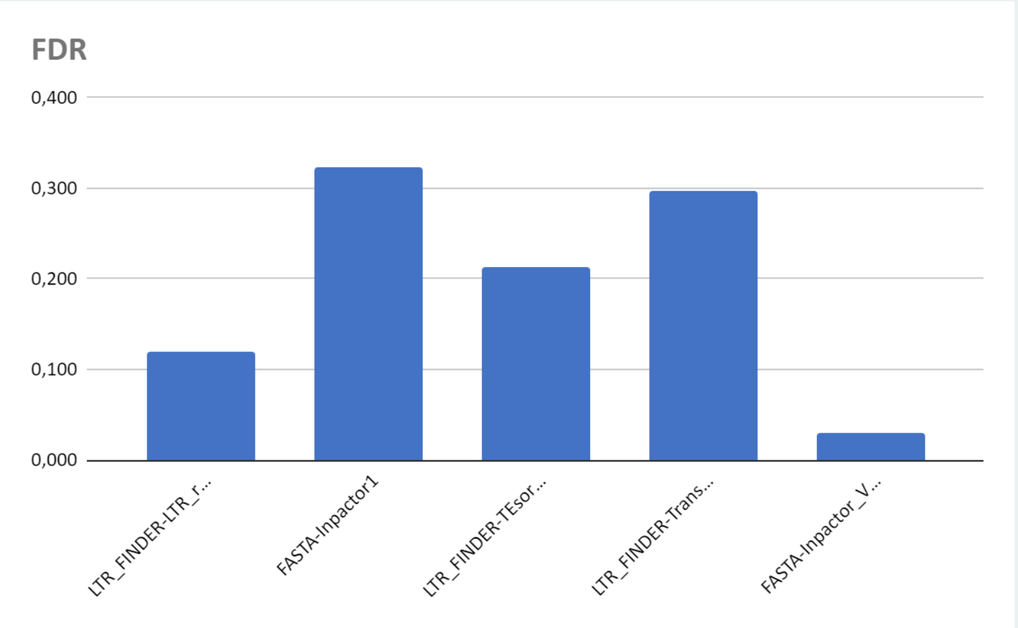


Figura 11.Tasa de descubrimientos falsos por líneas de trabajo, Oryza sativa

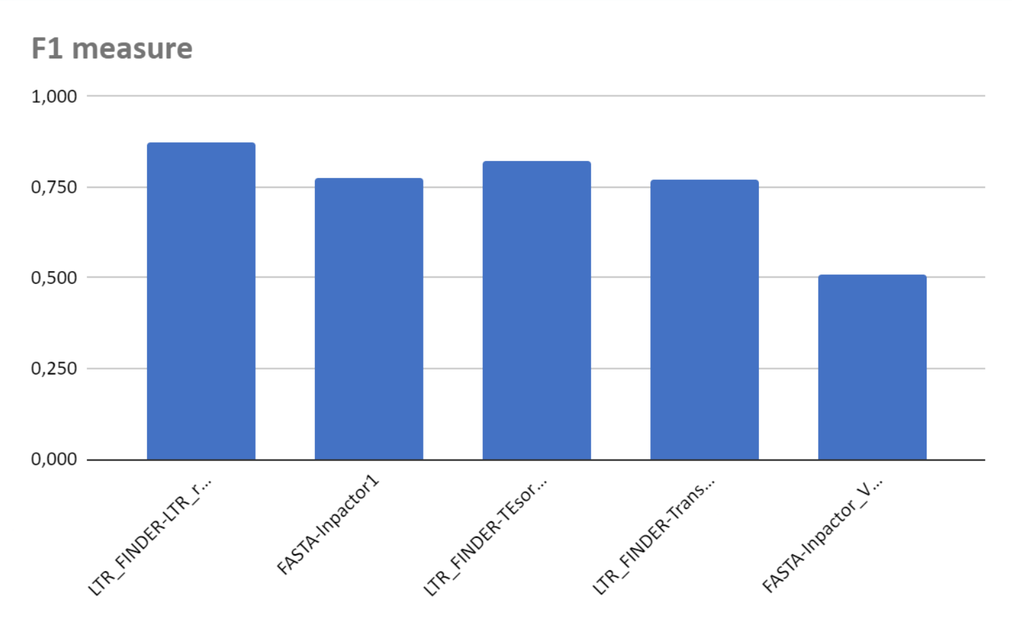


Figura 12. Medida F1 por líneas de trabajo, Oryza sativa

UNIVERSIDAD AUTÓNOMA DE MANIZALES
UAM®
VIGILADA MINEDUCACIÓN

RESULTADOS

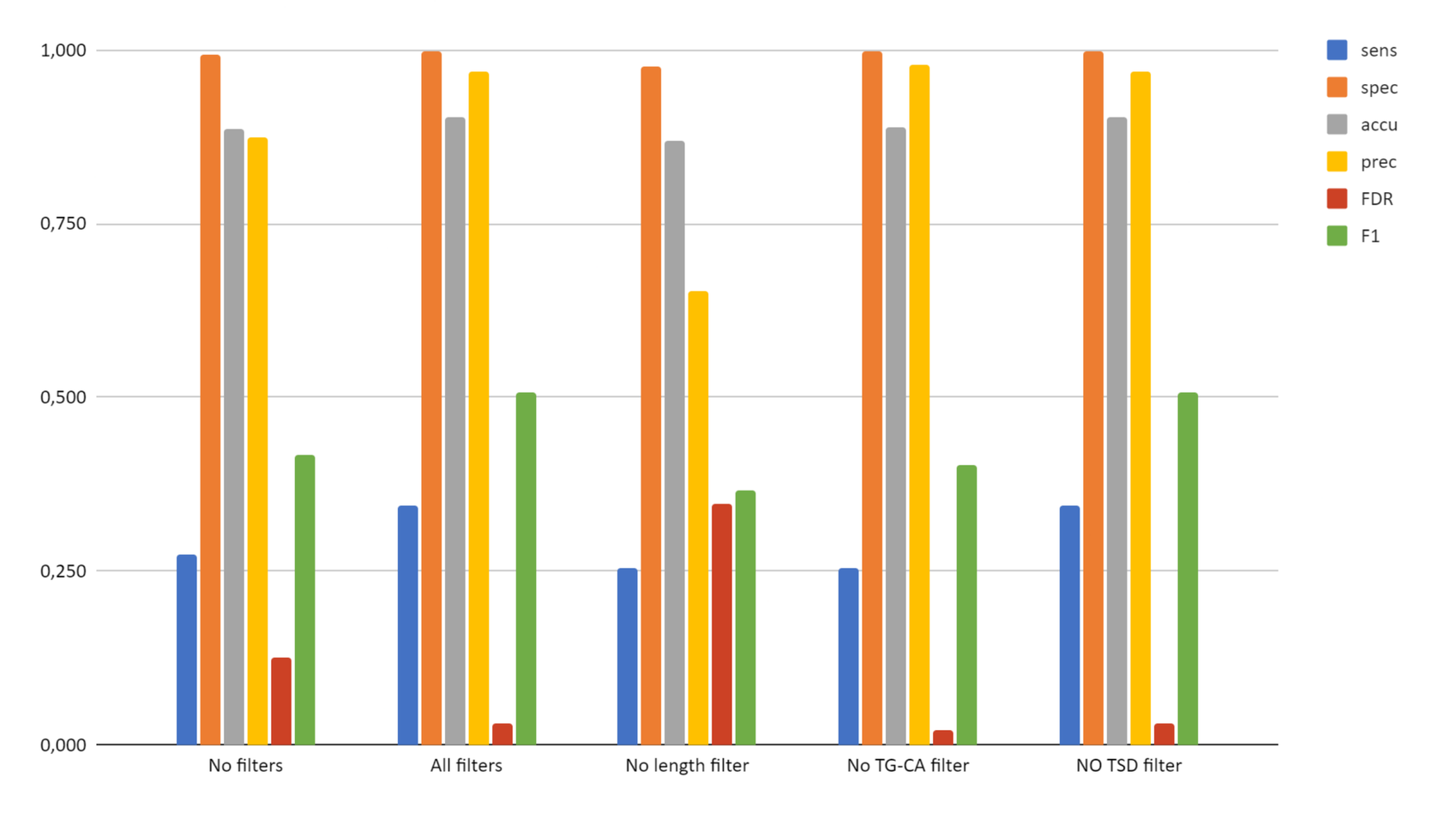


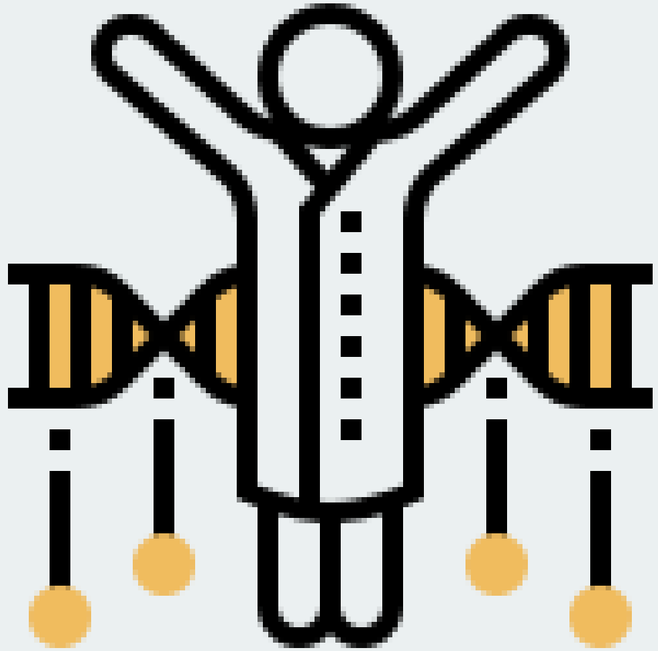
Figura 13. Resultados gráficos de las métricas obtenidas para Inpactor v1.3 en la variación de parámetros para mejorar el rendimiento

CONCLUSIONES

- Ventajas de los algoritmos de ML:

  - Mejoras en los tiempos de ejecución, rendimiento, costo computacional

- Influencia de los parámetros de filtrado y curación para mejoramiento de la herramienta
- Integración de variadas metodologías
- Evaluaciones comparativas entre especies de plantas y distintas herramientas



REFERENCIAS

Cui, X., & Cao, X. (2014). Epigenetic regulation and functional exaptation of transposable elements in higher plants. *Current Opinion in Plant Biology*, *21*(Figure 1), 83–88. https://doi.org/10.1016/j.pbi.2014.07.001

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

McClintock, B. (1953). Induction of Instability at Selected Loci in Maize. *Genetics*, *38*(6), 579–599. http://www.ncbi.nlm.nih.gov/pubmed/17247459%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1209627

Neumann, P., Novák, P., Hoštáková, N., & MacAs, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mobile DNA, 10(1), 1–17. https://doi.org/10.1186/s13100-018-0144-1

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V., & Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Research, 42(D1), 26–31. https://doi.org/10.1093/nar/gkt1069

Orozco-arias, S., Piña, J. S., Tabares-soto, R., & Castillo-ossa, L. F. (2020). Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements. Processes, 8(638), 1–20. https://doi.org/10.3390/pr8060638

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.

REFERENCIAS

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.

S. Orozco-Arias et al., “Inpactordb: A classified lineage-level plant LTR retrotransposon reference library for free-alignment methods based on machine learning,” Genes (Basel)., vol. 12, no. 2, pp. 1–17, 2021, doi: 10.3390/genes12020190.

Ou S, Su W. The Extensive de-novo TE Annotator. GitHub. Available from: https://github.com/oushujun/EDTA

Spannagl, M., Nussbaumer, T., Bader, K. C., Martis, M. M., Seidel, M., Kugler, K. G., Gundlach, H., & Mayer, K. F. X. (2016). PGSB plantsDB: Updates to the database framework for comparative plant genome research. Nucleic Acids Research, 44(D1), D1141–D1147. https://doi.org/10.1093/nar/gkv1130

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

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.

GRACIAS POR SU ATENCIÓN

Maradey Mercedes Arias Mendoza
Estudiante de Ingeniería Biomédica
maradey.ariasm@autonoma.edu.co