Pontificia Universidad Católica del Ecuador

Facultad de Ingeniería

Maestría en Bilogía Computacional

IDENTIFICACIÓN DE GENES DE LA RUTA METABÓLICA DE LA CAFEÍNA EN ENSAMBLAJES DE GENOMAS EN DIFERENTES ESPECIES DE CAFÉ SILVESTRES NATIVAS DE MADAGASCAR Y ÁFRICA

Proyecto de Titulación

Jordan Daniel Black Guevara

Asesor:

Dr. Romain Guyot

Lectores:

Dr. Abraham Avelar

Dr. Francisco Flores

Quito – Ecuador

Enero del 2024

Derechos de autor

Título del trabajo de titulación: Identificación de genes de la ruta metabólica de la cafeína en ensamblajes de genomas en diferentes especies de café silvestres nativas de Madagascar y África.

Autor: Jordan Black

ÍNDICE GENERAL

INDICE DE FIGURAS

INDICE DE TABLAS

DEDICATORIA

Este trabajo va dedicado para mi familia, sin ellos no pudiese estar donde me encuentro actualmente, gracias totales por su esfuerzo y dedicación en este transcurso de vida. Especialmente le dedico a mi hermano Juan a quien admiro con mi vida por ser un excelente papá de mis pequeños sobrinos, además por su valentía y determinación, gracias por todo, te amo.

AGRADECIMIENTO

Quiero comenzar con un sincero agradecimiento hacia mi tutor y asesor Dr. Romain Guyot en este arduo trabajo, que a pesar de que reside en Francia el me abrió y brindó un lugar en su agenda para reunirnos semanalmente por Zoom para lograr concluir con esta investigación. De igual manera muy agradecido con mi familia en especial mi madre Guadalupe y mi novia Gislayne que siempre estuvieron brindándome apoyo para culminar con este trabajo. Y sin dejar a un lado a Dios por su ayuda infinita.

RESUMEN

Cuenta la leyenda que el consumo del café empieza por parte de un pastor el cual veía comer a sus cabras "cerezas rojas" y estas comenzaban a actuar de una forma más animada con respecto a lo habitual, también por el consumo de estas pepas por parte de esclavos, los cuales tenían energía suficiente para continuar con sus arduas actividades diarias. Es así que el café ha venido consumiéndose desde años atrás hasta convertirse hoy en día en la bebida más consumida alrededor del mundo, por lo tanto, el alcaloide psicoactivo legal más ingerido globalmente. Se busca identificar los genes de la ruta metabólica de la cafeína en ensamblajes de 17 genomas de especies de café silvestres nativas de Madagascar y África, donde se llevó en marcha un flujo de trabajo en Lubuntu (v20.04) usando grandes paquetes de análisis como algoritmos para mapear proteínas contra genomas (exonerate v2.4.0), algoritmos para gestionar el formato gff (gffread v.0.12.7 y EMBOSS v6.6.0), algoritmos para crear alineamientos múltiples (muscle v5.1.0) y algoritmos para inferencia filogenética (FastTree v2.1.11), usando los genes de referencia de *C. canephora*: *XMT* (*xantosina 7 N-metiltransferasa*) (AFV60437.1), *MXMT* (*7-metilxantina metiltransferasa*) (AFV60435.1) y *DXMT* (*3,7-dimetilxantina metiltransferasa* o cafeína sintasa) (AFV60434.1) y los controles positivos y negativos: *C. humblotiana* y *C. canephora*. Se llegó a la conclusión que la ruta metabólica de la cafeína necesita obligatoriamente los aceptores de metilo *XMT* y *MXMT* para que así entre en acción la *DXMT* y sintetice este alcaloide, es decir para que exista la presencia de cafeína deben estar presentes los 3 genes mencionados, con estos resultados se pudo validar la metodología empleada ya que se pudo detectar la presencia de estos.

Palabras clave: café, genes, ruta metabólica, flujo de trabajo, *XMT, DXMT, MXMT*

ABSTRACT

Legend has it that coffee consumption began with a shepherd who noticed that his goats became livelier after eating "red cherries." Similarly, slaves consumed these seeds, gaining enough energy to carry out their strenuous daily activities. Over the years, coffee has continued to be consumed and has become the most widely consumed beverage worldwide, making it the most globally ingested psychoactive alkaloid. The study aimed to identify the genes in the caffeine metabolic pathway in 17 genomes assemblies of wild coffee species native to Madagascar and Africa. The workflow was implemented in Lubuntu (v20.04) using large analysis packages such as algorithms for mapping proteins against genomes (exonerate v2.4.0), algorithms for managing the gff format (gffread

v.0.12.7 and EMBOSS v6.6.0), algorithms for creating multiple alignments (muscle v5.1.0) and algorithms for phylogenetic inference (FastTree v2.1.11)). Reference genes from *C. canephora* were used: *XMT* (*xanthosine 7 N-methyltransferase*) (AFV60437.1), *MXMT (7-methylxanthine methyltransferase*) (AFV60435.1), and *DXMT (3,7-dimethylxanthine methyltransferase or caffeine synthase)* (AFV60434.1), and the positive and negative controls: *C. humblotiana and C. canephora*. It was concluded that the caffeine metabolic pathway necessarily requires the methyl acceptors *XMT* and *MXMT* for *DXMT* to come in action and synthesize this alkaloid, with these results it was possible to validate the methodology used since the presence of these genes could be detected.

Keywords: coffee, genes, metabolic pathway, workflow, *XMT, DXMT, MXMT*

1. Introducción

El consumo y el cultivo del café es una de las historias que más causa atracción a miles de personas, empieza exactamente en el Cuerno de África en el país de Etiopía en la provincia de Kaffa. Existe una leyenda de su descubrimiento, la cual se da cuando un pastor veía a sus cabras con un comportamiento mucho más animado de lo habitual después de consumir cerezas rojas de café y otro relato es que esta pepa era consumida por esclavos que eran enviados desde Sudan a Yemen por el puerto de Moca; y es así que se fue viendo la acción de esta pepa y los árabes la empezaron a distribuir quitando las capas exteriores para que no sea fértil. Esto duro hasta el año de 1616 donde los holandeses llevaron este grano a su tierra y lo empezaron a cultivar en invernaderos. Y es así que el café empezó a consumirse en unos inicios en la Meca y después se extendió por todo el mundo persa donde esta bebida era acompañada de amistades, chismes, risas, juegos, bailes, cantos y se empezaron a abrir estos nuevos establecimientos donde el café era su titular (International Coffee Organization, 2022).

El café se ha convertido en un alcaloide psicoactivo legal y el más consumido todo el mundo el cual es cultivado en al menos 80 países en 4 continentes del planeta, tras varios estudios se llegó a determinar que 3 especies de plantas de café son las principales en ser utilizadas con fines comerciales, estamos hablando de: *Coffea arabica* L. *(*alotetraploide)*, Coffea canephora* Pierre ex A. Froehner (conocido como café Robusta) y *Coffea liberica* Bull. ex Hiern (conocido como café Excelsa, liberiano o Liberica), donde C. arabica es el más importante comercialmente con el 95% de la producción del café (Mannino, 2023). Datos han revelado que el amargor del café no depende solamente de la cantidad de cafeína que este tenga, sino de los cambios químicos que se generan durante el tueste del mismo, entre las principales reacciones se encuentra la reacción de Maillard, la degradación de Strecker y la caramelización de los azucares (Verónica Belchior, 2019).

La 1, 3, 7 – trimetilxantina o más conocida como cafeína es un tipo de alcaloide de xantina que es producto de las plantas, descubierta por los alemanes Von Giese y Runge de semillas de café en el año de 1820 (Lin et al., 2023). Se presumen que las dos principales funciones ecológicas de la cafeína son la inhibición del crecimiento de plantas vecinas y la protección hacia animales y patógenos, es decir alelopatía y defensa química correspondientemente (Lin et al., 2023). Hollingsworth et al. (2022) llegó a la conclusión

que al suministrar concentraciones bajas de cafeína (1-2%) sirve como un repelente hacia insectos, el cual es muy eficaz contra babosas y caracoles, con la gran ventaja que el área foliar de la planta queda intacta.

Este alcaloide gracias a diversos estudios ha sido caracterizado como un biomarcador muy potente para detectar enfermedades, como la de Parkinson; en el ser humano la cafeína juega un rol muy importante, por ejemplo alivia la fatiga, mejora la atención sostenida, la memoria, eleva los efectos anticancerígenos, disminuye la presión arterial, hasta puede retrasar el envejecimiento; en la medicina es utilizada para ayudar con la diuresis, como estimulante y además como un adyuvante analgésico, cabe mencionar que es una muy buena fuente de antioxidantes (Lin et al., 2023).

A la cafeína se la puede encontrar en varias fuentes, las que más destacan son plantas de café, y su dosis varía dependiendo la genética de los granos y también la forma de preparación del mismo; en segundo lugar, está el té, que básicamente es la hoja desecada de los arbustos de *Camellia sinensis, Thea bohea*, o *Thea viridis*; la concentración de este alcaloide así mismo varía dependiendo del método de elaboración, tiempo de extracción y variedad del mismo (fermentado (té verde), semifermentado (té rojo), fermentado (té negro) y blanco). En tercer lugar, está el cacao (*Theobroma cacao L.*), en esta semilla la teobromina gana en porcentaje con un 2.5, frente a 0.4 de la cafeína; también la cafeína está presente en otras plantas, es el caso de la *Paullinia cupana (*guaraná), *Ilex paraguariensis* (mate), *Cola nitida* (cola) y *Paullinia yoco* (yoco) con un aproximado entre 2 a 4% (Pardo et al., 2007).

El café pertenece al género *Coffea*, con 139 especies diferentes nativas de Madagascar, África y Asia (Guyot et al., 2020). Madagascar contiene la mayor diversidad de especies de *Coffea*, con más de 66 especies reconocidas (Davis & Rakotonasolo, 2021). Gracias a estudios realizados por Romain Guyot del IRD de Francia se sabe que existen variedades de café con y sin cafeína. En la siguiente Figura (Figura 1) se puede ver la filogenia de 52 especies de café con una alta calidad con 28800 marcadores de tipo SNP y un soporte máximo (100%) con su posición geográfica y el contenido de cafeína en los granos (Hamon et al., 2017). Las especies autóctonas de Madagascar no suelen contener cafeína (azul oscuro en el gráfico, excepto dos especies), mientras que las especies de África oriental tienen poca o ninguna cafeína y las de África occidental contienen altos niveles de cafeína (rojo). La especie con mayor contenido de cafeína es *C. canephora* (o Robusta) con más del 3% de la materia seca. Gracias a esto se sabe que

existen especies de café, pero no se sabe el porqué de la expresión de cafeína en ciertas plantas, por este motivo este estudio va a ser enfocado en saber que genes son los causantes de la ausencia o no de cafeína en diferentes tipos de cafés ya secuenciados, y despejar dudas sobre la ruta metabólica de la N-metil transferasa y su acción en la producción de cafeína en estas especies.

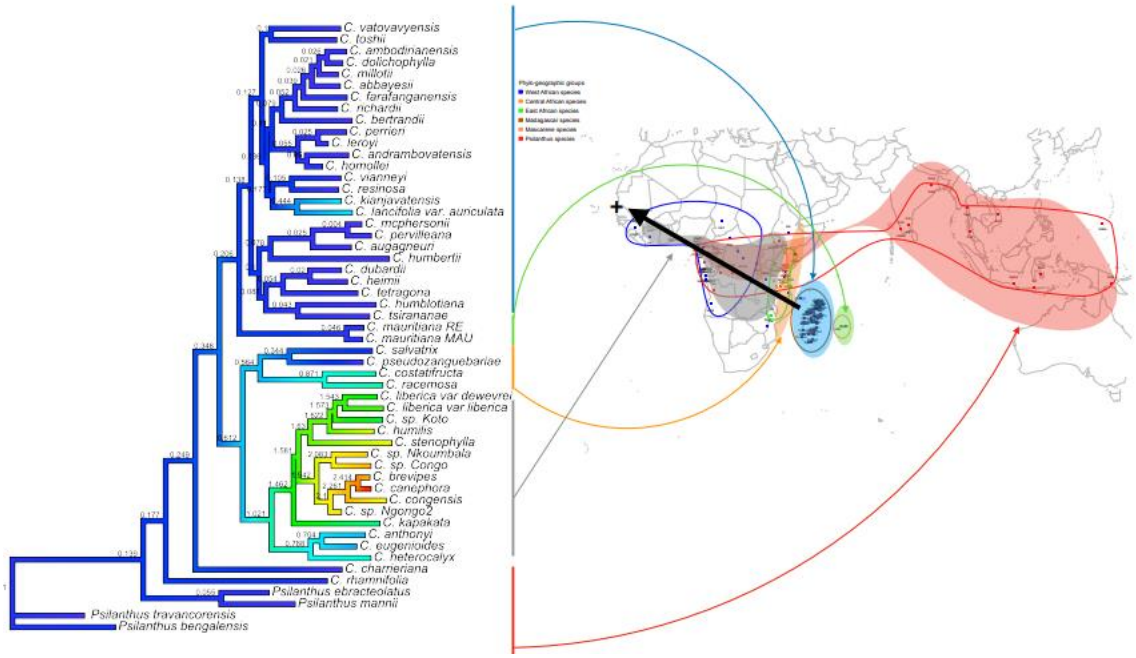


Figura 1. Filogenia de 52 especies de Café. *Con la concentración de cafeína en los granos (Izquierda) las especies se agrupan en grupos filogeográficos y su origen geográfico se indica a la derecha* (Hamon et al., 2017). *(Azul: ausencia de la cafeína; rojo: concentración máxima en cafeína).*

Gracias a estudios realizados se sabe que la vía biosintética principal de la cafeína lleva 3 pasos de metilación, los cuales son necesarios para la síntesis de cafeína a partir de xantosina, la cual lleva a una acción en cadena de 3 N-Metiltransferasas. La primera en actuar es la xantosina metiltransferasa (*XMT*), que de xantosina transforma a 7-metil-xantocina, se pierde una ribosa y termina con el nombre de 7-metil-xantina, después entra la segunda denominada teobromina sintasa (*MXMT*) y da como resultado a la teobromina y la última que lleva el nombre de cafeína sintasa (*DXMT*) que como su nombre mismo lo dice da resultado a la cafeína. Entre cada proceso de metilación interviene como intermediarios SAM (S-adenosilmetionina) y SAH (S- adenosilhomocisteína) Denoeud et al., 2014).

En esta investigación se va a trabajar con:

Tabla 1. Especies de café a ser estudiadas.

Se tiene las 18 especies de café a ser estudiadas, cada una con su porcentaje de cafeína presente en el grano y su ubicación correspondiente (agropolis fondation, 2024) (Global Core Biodata Resource, 2024).

ESPECIE DE CAFÉ	**% CAFEÍNA M.S.**	**UBICACIÓN**
Coffea canephora Pierre ex A. Froehner.	1.51 - 3.33	- África Occidental (Liberia y Guinea) hasta Uganda y el sur de Sudán. - Desde República Centroafricana hasta Angola.
Coffea arabica L.	0.96 - 1.62	- Sur Oeste de Etiopía (provincias de Illubador y Kaffa). - Monte Marsabit (Norte de Kenia). - Sur de Sudán
Coffea humblotiana Baill.	0.00 - 0.01	- Archipiélago de las Comoras - Anjouan hasta Mayotte
Coffea homollei J.-F. Leroy	0.00 – 0.06	Este de Madagascar
Coffea myrtifolia (A. Rich. ex DC.) J.-F. Leroy	-	Bosque seco en Trois Mamelles, Mauricio
Coffea pseudozanguebariae Bridson.	0.00 – 0.03	Desde Tanzania (incluido Zanzíbar) hasta el sudeste de Kenia (altitud baja)
Coffea salvatrix Swynn. & Philipson	0.01 – 0.19	- Desde Mozambique hasta suroeste de Tanzania (altitud baja). - Zimbabue y Malawi.
Coffea farafanganensis J.-F. Leroy	0.00 – 0.09	- Cerca de Farafangana. - Sureste de Madagascar
Coffea charrieriana Stoff. & F. Anthony.	0.00 – 0.03	- Bosque de Bakossi región de Kumba-Loum. - Desde el oeste hasta el sur oeste de Camerún
Coffea tetragona Jum. & H. Perrier.	0.00 – 0.03	- norte de Madagascar - Sambirano
Coffea mauritiana Lam.	0.00 – 0.07	Plaine Champagne Islas Mascareñas, Mauricio.
Coffea racemosa Lour.	0.86 - 1.25	Desde Sudáfrica hasta Zimbabwe and Mozambique (altitudes bajas)
Coffea stenophylla G. Don.	1.74 – 2.43	- Costa de Marfil-Sierra Leona. - Senegal y Costa de Marfil. - Guinea.

		- Desde Liberia hasta Ghana
Coffea humilis A. Chev.	1.67 – 2.27	- Liberia. - Guinea. - Costa de Marfil
Coffea eugenioides S. Moore.	0.44 – 0.60	- Kenia (altitudes altas), República Democrática del Congo (parte oriental) - Ruanda. - Uganda - Tanzania
Coffea kianjavatensis J.-F. Leroy.	0.70	Este de Madagascar
Coffea lancifolia A. Chev.	0.70	Este de Madagascar

Se tiene 8 especies de café con cafeína (*C. stenophylla* (2.05-2.43%)*, C. humilis* (1.67-2.27%), *C. racemosa* (0.86-1.25%), *C. canephora* (1.51-3.33%), *C. eugenioides* (0.44-0.60%), *C. arabica* (1.42-1.62%) de África*; C. kianjavatensis* (0.7%) y *C. lancifolia* (0.7%)*,* de Madagascar. Y 8 especies de café sin cafeína (*C. salvatrix, C. humblotiana, C. pseudozanguebariae* y *C. charrieriana* de África*, C. homollei, C. farafanganensis,* C. *tetragona* de Madagascar y *C. mauritiana* de Islas Mascareñas). Y *C. myrtifolia* que se desconoce el porcentaje de cafeína (agropolis fondation, 2024)(Global Core Biodata Resource, 2024). De los cuales se obtuvo borradores provenientes de ensamblaje usando MaSuRCA ("Maryland Super Read Cabog Assembler") de genomas de referencia del IRD de Francia, no fue posible utilizar lecturas crudas ya que no existen aún algoritmos fiables y además se presentan limitaciones en el uso del clúster CEDIA ya que se necesita estar en la raíz del sistema para poder instalar algoritmos necesarios para el análisis de las muestras. Se utilizará a *C. humblotiana* para tener una referencia de estos genomas a ser ensamblados ya que este genoma tiene un ensamblaje del 88.7% (Raharimalala et al., 2021)

Se tiene a *Coffea humblotiana* como el control positivo en este estudio, esta especie de café fue elegida debido a que se tiene datos de su genoma bien caracterizados, el cual sirve para comparar con los resultados que se van a obtener, además no tiene el gen implicado en la producción de la cafeína, es decir la cafeína sintetasa o *DXMT* (sintetiza teobromina en cafeína), por ende, es una especie libre de cafeína, su origen es en el archipiélago de las Comoras y es una especie silvestre, catalogada en peligro de extinción. Como se mencionó anteriormente *C. humblotiana* presenta la ausencia en su 100% de cafeína, ya sea en semilla como en hojas, la cual está presente en especies nativas

de Madagascar, Islas Mascareñas y más especies africanas centrales y orientales (Raharimalala et al., 2021).

Determinación del problema

El problema del presente estudio se basa en el conocimiento de la ruta de biosíntesis de la cafeína en especies de café productoras de cafeína y especies que no la producen. La ruta de biosíntesis de la cafeína es bien conocida para *C. canephora*, mejor conocida como Robusta. Esta ruta biosintética implica tres pasos de metilación que es catalizada por diferentes genes de N-metiltransferasa (*NMT*); el gen *XMT* (xantosina 7 N-metiltransferasa), el gen *MXMT* (7-metilxantina metiltransferasa) y el gen *DXMT* (3,7-dimetilxantina metiltransferasa o cafeína sintasa). Los tres genes *NMT* están ubicados en dos regiones distintas en *C. canephora*, en el cromosoma 1 para el gen *DXMT* y en el cromosoma 9 para los genes *XMT* y *MXMT* (Denoeud et al., 2014). Los 3 genes están muy conservados, ya que pertenecen a la misma familia en el genoma del café. Existen otros genes *NMT* en el genoma de *C. canephora*, pero con sustratos y productos que aún no se conocen.

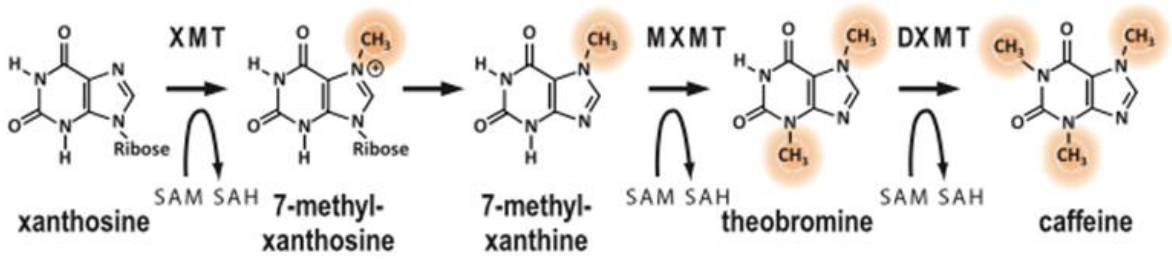


Figura 2. Vía biosintética de la cafeína. *Empieza con la xantosina, y tiene 3 pasos de metilación, donde se dan acciones consecutivas de 3 NMT: XMT, MXMT, DXMT* (Denoeud et al., 2014).

En las especies de *Coffea*, existen grandes variaciones en el contenido de cafeína según el origen filogeográfico de la especie (Hamon et al., 2017) (Guyot et al., 2020). Las especies silvestres de África occidental generalmente muestran un mayor contenido de cafeína, mientras que las especies de África oriental muestran un contenido bajo o nulo de cafeína en las semillas. En Madagascar, de las 66 especies silvestres, solo 2 muestran trazas de cafeína en las semillas (*C. kianjavatensis* y *C. lancifolia*) pero no en las hojas (Ranarivelo & ND, 2011)

Para comprender la variación del contenido de cafeína en las especies de café silvestre y considerar estrategias para controlar el contenido de cafeína, parece interesante estudiar

la diversidad molecular de los genes *NMT* implicados en la vía de la cafeína en especies que producen un bajo contenido de cafeína y en especies que no producen cafeína. Recientemente, la secuenciación del genoma de *Coffea humblotiana*, una especie silvestre sin cafeína en peligro de extinción del archipiélago de las Comoras ha mostrado la pérdida del gen de la cafeína sintasa (*DXMT*) que convierte la teobromina en cafeína. Se cree que la pérdida se debe a la deleción de un segmento de 70 kbp en el cromosoma 1. El resultado es la pérdida de producción de cafeína en la semilla de esta especie, pero sin una sobreproducción de teobromina (otro compuesto muy amargo).

Este resultado plantea varias cuestiones en las especies de café. La primera es si la pérdida del gen de la cafeína sintasa (*DXMT*) es común a todas las especies silvestres sin cafeína o si otros mecanismos pueden explicar esta característica. La segunda es conocer el mecanismo evolutivo de la ausencia / presencia de cafeína en las plantas de café. ¿Apareció la cafeína muy pronto en la evolución de los cafetos africanos y luego desapareció en los cafetos malgaches o la cafeína sólo apareció en las especies africanas?

En el pasado, los enfoques de identificación molecular (por ejemplo, la amplificación por PCR) han fracasado debido a la gran conservación de la secuencia de los genes de la familia, lo que indica que se necesitan nuevos enfoques.

Para investigar las relaciones entre los genes *NMT* y la cafeína, proponemos identificar y anotar los genes de la familia *NMT* en diferentes recursos genómicos disponibles para este proyecto. Se han secuenciado 8 genomas de *Coffea* usando lecturas largas, dónde 3 producen un alto contenido de cafeína, 2 tienen un bajo contenido de cafeína y 3 son especies sin cafeína. Además, se secuenciaron 72 especies silvestres de *Coffea* con diferentes niveles de contenido de cafeína mediante lecturas cortas (Illumina) y se ensamblaron.

Estos recursos genómicos representan una buena base para estudiar la relación entre las especies silvestres, el contenido de cafeína y los genes *NMT*. Sin embargo, es necesario desarrollar una metodología bioinformática para la identificación y anotación rápida de genes *NMT* teniendo en cuenta la cantidad de datos disponibles. En particular, para las lecturas cortas (72 genomas), deben investigarse diferentes enfoques para la identificación y anotación de los genes *NMT*.

Para este enfoque se puede estudiar el ensamblaje local de las lecturas, guiado por una secuencia de referencia, utilizando Exonerate, que básicamente es una herramienta que

permite alinear proteína a genoma y señala el sitio donde se dio el empalme con gran exactitud ya que usa una matriz de peso específica (Li, 2023). Por último, en general, el desarrollo de una metodología de anotación de genes dirigida en secuencias brutas o genomas ensamblados no anotados podría ser muy útil para buscar genes de interés, analizar su diversidad y evolución de forma muy rápida y con un gran volumen de datos.

2. Objetivos

General

Identificar y clasificar los genes de la N-metiltransferasa implicados en la vía biosintética de la cafeína a partir de diferentes recursos genómicos producidos a partir de especies silvestres de café de África y Madagascar y de especies con y sin cafeína, para establecer si la presencia o ausencia de estos es crucial para la producción de cafeína en estas plantas.

Específicos

- Identificar y anotar genes *NMT* a partir de genomas de referencia completamente ensamblados a partir de lecturas largas (PacBio) de especies productoras de cafeína.
- Diseñar y probar un flujo de trabajo eficaz y estrategias de anotación rápida.
- Probar la metodología con genomas conocidos y completos como control positivo y negativo (*C. humblotiana, C. canephora*).
- Aplicar el flujo más relevante con secuencias o ensamblaje de una selección de 17 genomas de especies de café ensamblados a partir de lecturas largas o cortas (Illumina).
- Aplicar la filogenia molecular para separar la familia de genes de interés.

3. Revisión de Literatura

3.1. CDS.

Son secuencias de nucleótidos que corresponden con otras secuencias de aminoácidos (a.a) de proteínas. Estas secuencias comienzan por lo general con el codón ATG y termina con uno de parada, los CDS también pueden ser subconjuntos de ORF (marco de lectura abierto), en eucariotas la predicción de estas secuencias se complica debido a interrupciones por parte de intrones (Furuno et al., 2003).

3.2. Árbol filogenético

La filogenia molecular es una herramienta utilizada en investigaciones comparativas dentro de la genética, además de aplicaciones como Mr. Bayes, PAUP y Beast. La mayoría de estimadores filogenéticos son basados en modelos explícitos de la evolución de nucleótidos para así poder estimar parámetros de la evolución como puede ser longitud de ramas y topologías de árboles, por lo tanto, los árboles filogenéticos son creados para poder analizar las relaciones evolutivas que se observan y así poder tener información a través de estas, logrando encontrar divergencia de linajes o la relación que estos pueden tener (Madrigal-Valverde, 2017).

3.3. Herramientas bioinformáticas usadas

3.3.1. Oracle VM VirtualBox (versión 7.0).

Es software o máquina virtual que permite la virtualización de varias plataformas con un código abierto, el cual da la posibilidad de ejecutar varios sistemas operativos a la vez en un solo dispositivo, es utilizado para entregar código abierto de una manera más veloz probando y ejecutando varios sistemas operativos en una computadora, puede ser ejecutado en macOS, Windows y Oracle Solaris. Es muy famoso ya que es fácil de usar, además de ser potente, rápido y con una cobertura amplia de plataforma. (Oracle, 2020).

3.3.2. Bash.

Es el Shell del proyecto GNU, sus siglas significan “Bourne-again SHell”, es compatible con sh que tenga funciones de ksh (Korn Shell) y de csh (C Shell), cumple con los estándares “IEEE POSIX P1003.2/ISO 9945.2 Shell and Tools”. Este sirve para programar y para uso en general, además tiene la capacidad de ejecutar scripts sh sin más modificaciones, aquí se puede editar líneas de comando, controlar trabajo, brinda el historial de comandos ilimitadamente, capacidad de función Shell y alias, matrices con tamaño ilimitado y entre otros (GNU Operating System, 2020).

3.3.3. Lubuntu (versión 20.04).

Es una versión formal de Ubuntu el cual usa un entorno ligero de escritorio LXQt con el objetivo de brindar una distribución liviana de Linux, pero a la vez funcional, este brinda una interfaz gráfica sencilla al usuario, pero a la vez potente y moderna, además que tiene una variedad amplia de aplicaciones para

el uso cotidiano. La versión 20.04 es la cuarta de Lubuntu con el entorno LXQt (lubuntu, 2023).

3.3.4. Miniconda (versión 23.9.0).

Es in mini instalador suministrado por Anaconda, es usado para instalar la mayoría de paquetes por uno mismo, además que no se necesita permisos de administrador o root (CONDA, 2017b). Es una herramienta muy potente de línea de comandos para gestionar entornos y paquetes, ejecutable en Windows, Linux y macOS, conda da la posibilidad de crear ambientes o entornos independientes, los cuales tendrán archivos propios y dependencia de paquetes, aquí los ambientes o entornos no interactuarán entre si (CONDA, 2017a).

3.3.5. Exonerate (versión 2.4.0).

Es una herramienta común utilizada para la alineación de secuencias que usa la biblioteca de programación dinámica C4, la cual compara secuencias por pares, tiene varios modelos de alineación que pueden ser utilizados, usando programación heurística o dinámica exhaustiva. Viene con varias utilidades las cuales son simples y pueden ser aplicadas en archivos FASTA grandes o pequeños. Exonerate permite alinear secuencias de la siguiente manera, a estos se los conoce como modelos: cDNA – secuencia genómica (est2genome), proteína – secuencia genómica (protein2genome), proteína – DNA (protein2dna), genoma – genoma (genome2genome) (EMBL's European Bioinformatics Institute, 2024).

3.3.6. Gffread (versión 0.12.7).

Permite convertir formatos GTF/GFF, extraer secuencias FASTA, filtrado, y más, además este algoritmo puede ser utilizado para validar y más operaciones en archivos con formato GFF, gffread puede ser verificado con gffcompare, Stringtie y Cufflinks ya que comparten el mismo código (pkgsrc, n.d.)

3.3.7. EMBOSS (versión 6.6.0).

Sus siglas significan "Suite Europea de Software Abierto de Biología Molecular", este es un paquete para análisis con un software gratis y código abierto desarrollado para su mayoría para biólogos moleculares, EMBOSS maneja con autonomía datos en una variedad de formatos, incluso permite recuperar secuencias de la web. Este paquete tiene bibliotecas muy amplias que

permiten el desarrollo de software con código abierto, unas de las ventajas de este paquete es que es una recopilación de herramientas para construir flujos de trabajo de biología computacional muy sólidos, además de manejar secuencias en varios formatos (Rice et al., 2000).

3.3.8. MUSCLE (versión 5.1.0).

Es un Software que es utilizado para realizar alineamientos múltiples de secuencias biológicas, esta versión de MUSCLE logra obtener puntuaciones muy altas en múltiples puntos de referencia de alineación donde se incluye Balifam, Balibase, Bralibase y todo esto en una computadora básica de escritorio, permite la generación de grupos de alineamientos alternativos con una precisión alta usando parámetros predeterminados. Este software es hasta 30% más preciso que Clustal-Omega y MAFFT (Center for Quantitative Life Sciences, 2023).

3.3.9. FastTree (versión 2.1.11).

Este algoritmo infiere arboles filogenéticos con una probabilidad aproximadamente máxima desde alineamientos de proteínas o secuencias de nucleótidos, FastTree es capaz de trabajar con alineaciones de hasta 1.000.000 de secuencias en un tiempo razonable y uso de memoria, se ha calculado que FastTree puede ser entre cien a mil veces más rápido que RAxML 7 o PhyML 3.0. El formato de salida de FastTree es NEWICK. (ILRI Research Computing, 2020).

3.3.10. iTOL (versión 6.0.0).

Sus siglas significan “Interactive Tree Of Life”, la cual es una herramienta online que sirve para la visualización, gestión, administración y anotación de árboles filogenéticos, la ventaja de iTOL es que se puede tener conjunto de datos en un número ilimitado, además que se puede crear cifras de calidad para publicaciones (iTOL Interactive Tree of Life, 2024).

3.4. Formatos principales usados en la investigación

3.4.1. GFF.

Sus siglas significan “Gene Feature Format” o en español “Formato de características generales”, este es un archivo muy popular utilizado en bioinformática el cual intercambia y representa información de varias

características genómicas, como por ejemplo su estructura, ubicación y genes de transcripción, los archivos GFF son característicos ya que se implementan en C++ en OS X y Linux con una licencia MIT. (Pertea & Pertea, 2020).

3.4.2. FASTA.

Son formatos, donde su texto debe iniciar con un quilate (">") acompañado de un identificador de secuencia que debe ser único y no contener espacios ("SeqID"), este será cambiado por un número de accesión al momento de ser enviado a revisión, esto es conocido como línea de definición FASTA. Después de esta línea empieza la secuencia de nucleótidos o proteínas (NCBI, 2021).

3.4.3. NEWICK.

Este formato es diseñado para representar arboles con raíz, con nodos con su respectiva etiqueta y longitud especifica entre "padre e hijo". Este formato termina en un punto y coma, los árboles son representados como lista anidada anotada con nombres de los nodos, donde estos tienen una letra seguida con o sin más letras, también pueden ser dígitos que pueden o no estar subrayados, entonces, los nodos tendrán un nombre que ira seguido de su longitud, separada por dos puntos, por ejemplo, (perro:15, (gato:30, caballo:21, (venado:18, vaca:12):50) (Idaho University, n.d.).

4. Metodología

La presente investigación se basó en un *pipeline* de 12 principales etapas, como se puede observar en la figura 3. Los genes de referencia utilizados fueron provenientes de *C. canephora*, presentes en la ruta biosintética de producción de cafeína en plantas de café, estas son: *XMT* (*xantosina 7 N-metiltransferasa*) (número de accesión nucleotídica: JX978509 y proteica: AFV60437.1), el gen *MXMT* (*7-metilxantina metiltransferasa*) (número de accesión nucleotídica: JX978507 y proteica: AFV60435.1) y *DXMT* (*3,7-dimetilxantina metiltransferasa* o cafeína sintasa) (número de accesión nucleotídica: JX978506 y proteica: AFV60434.1) (Denoeud et al., 2014) (Perrois et al., 2015).

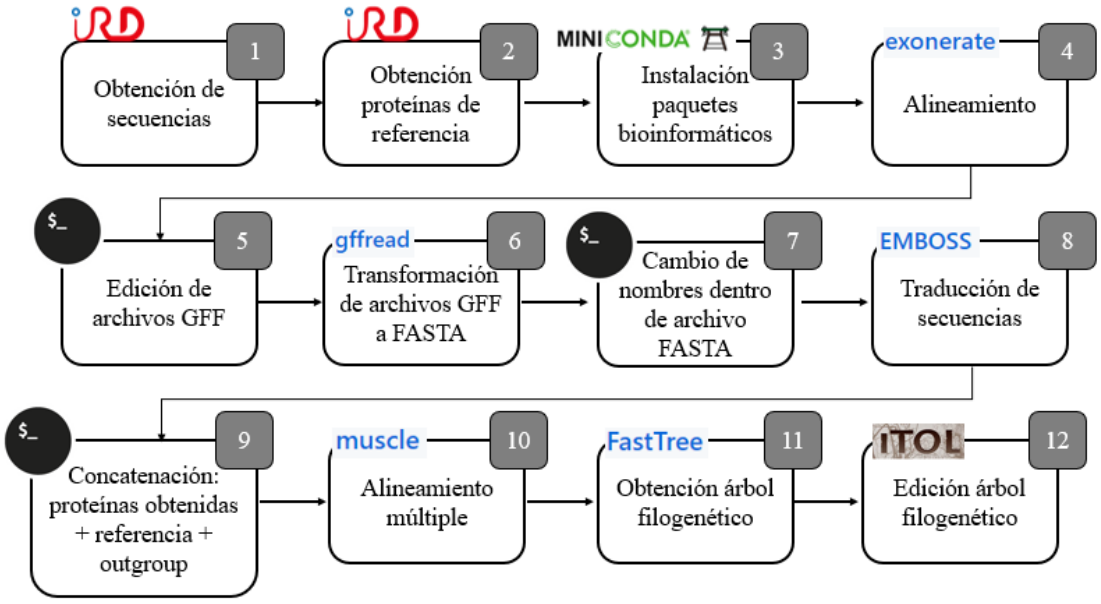


Figura 3. Flujo de trabajo de la investigación. *Flujo de trabajo utilizado para la identificación de genes de la ruta metabólica de la cafeína en borradores de genomas de diferentes especies de café silvestres nativas de Madagascar y África.*

Los borradores de los genomas de las especies de café utilizados en la investigación fueron suministrados por mi tutor Romain Guyot del IRD de Francia, de igual manera los 3 genes de referencia implicados en la ruta biosintética de la producción de cafeína en *C. canephora*, todos estos en formato FASTA. Se instalaron los algoritmos necesarios (exonerate (versión 2.4.0), gffread (versión 0.12.7), EMBOSS (versión 6.6.0), muscle (versión 5.1) y FastTree (versión 2.1.11)) en la máquina virtual (Oracle VM VirtualBox (versión 7.0)) utilizando MINICONDA (versión 23.9.0) y se empezó con el alineamiento utilizando exonerate, donde se obtiene un archivo en formato GFF2, el cual debe ser modificado a GFF3 con varios comandos en el Bash de Linux y manualmente, después se utiliza el algoritmo GFFREAD para pasar de formato GFF hacia formato FASTA, ya con el archivo en formato FASTA se utiliza el algoritmo EMBOSS para traducir a proteínas nuestro alineamiento. Se concatenan los archivos (genes de referencia, alineamientos obtenidos y grupo externo (*K. floribunda)*) y se procede a hacer el alineamiento múltiple usando el algoritmo MUSCLE, con esto se tiene ya el archivo definitivo. Se activa FASTTREE y se tiene un archivo en formato Newick, el cual se lo guarda y se lo carga en el programa iTOL y se genera el árbol filogenético, al cual hay que aditarlo hasta tener el deseado.

4.1. Obtención de secuencias

En total fueron 18 borradores comprimidos en archivos .zip y .gz, donde fueron descomprimidos usando los comandos unzip y gunzip, correspondientemente. Cada archivo contaba con un "código" y pertenecía a cada una de las especies a ser estudiadas, estos fueron los siguientes:

Tabla 2. Archivos con borradores de secuencias de café.

Al lado izquierdo el nombre del archivo y en el lado derecho el nombre de la especie de cada uno, para evitar confusiones durante la investigación. Además del tipo de secuencia del ensamblaje, donde la calidad del ensamblaje puede variar dependiendo del tipo de secuenciación, cobertura y nivel de correcciones, para estar seguro de los resultados se debe secuencias con lecturas largas los genomas de interés.

Archivo	**Tipo de secuencias del ensamblaje**	**Especie**
coffea_arabica_v0.6_06.25.19.fasta	Largas (PacBio)	*C. arabica*
coffea_homollei_22Feb2020_pj8sL.fasta	Largas (PacBio)	*C. homollei*
coffea_humblotiana_22Feb2020_j2waN.fasta	Largas (PacBio)	*C. humblotiana*
coffea_pseudozanguebariae_24Feb2020_SqZk9.fasta	Largas (PacBio)	C. *pseudozanguebariae*
Masurca_C033_genome.scf.ragtag.fasta	Cortas (Illumina)	*C. canephora*
Masurca_C314_IRD_S6_S25_genome.scf.ragtag.fassta	Cortas (Illumina)	*C. farafanganensis*
Masurca_C320_IRD_S12_S14_genome.scf.ragtag.fasta	Cortas (Illumina)	*C. lancifolia*
Masurca_C330_IRD_S25_S27_genome.scf.ragtag.fasta	Cortas (Illumina)	*C. kianjavatensis*
Masurca_C408_salvatrix-final.genome.scf.ragtag.fasta	Cortas (Illumina)	*C. salvatrix*
Masurca_C414_myrtifolia_final.genome.scf.ragtag.fasta	Cortas (Illumina)	*C. myrtifolia*
Masurca_DA-final.genome.scf.kaiju.ragtag.fasta	Cortas (Illumina)	*C. eugenoides*
Masurca_FB_genome.scf.ragtag.fasta	Cortas (Illumina)	*C. stenophylla*
Masurca_GH_genome.scf.ragtag.fasta	Cortas (Illumina)	*C. humillis*
Masurca_IB_genome.scf.ragtag.fasta	Cortas (Illumina)	*C. racemosa*
Masurca_MAUR_mau_C354_IRD_S49_S30_genome.scf.ragtag.fasta	Cortas (Illumina)	*C. mauritiana*
Masurca_OA_genome.scf.ragtag.fasta	Cortas (Illumina)	*C. charrieriana*
Masurca_TET_genome.scf.ragtag.fasta	Cortas (Illumina)	*C. tetragona*

Cada archivo fue descomprimido y guardado en una sola carpeta para tener una sola dirección a ser trabajada en el terminal de Linux.

4.2. Obtención de genes de referencia (proteínas)

Las proteínas de referencia fueron enviadas por correo de parte de mi tutor Romain Guyot del IRD de Francia, y fueron descargadas en mi máquina virtual.

Figura 4. Proteínas de referencia de C. canephora. *Se tienen los 3 genes de referencia provenientes de C. canefora: XMT_AFV60437.1, DXMT_AFV60434.1 y MXMT_AFV60435.1 en formato FASTA, traducidas a proteínas*(Perrois et al., 2015).

4.3. Instalación de paquetes bioinformáticos.

Lo primero que se hizo fue instalar MINICONDA en la máquina virtual, para ayudarnos con la descarga de los algoritmos a ser utilizados, estos fueron exonerate, gffread, EMBOSS, muscle y FastTree, para cada uno de estos se creó un ambiente con el comando -n, para tener un espacio aislado con el algoritmo a ser utilizado donde se podrá gozar de sus bibliotecas o paquetes por completo. A cada algoritmo se lo llama escribiendo en el terminal “conda actívate” y seguido del mismo.

4.4. Alineamiento.

Activar el entorno o ambiente con el comando conda actívate como se puede ver en la figura 5 y dirigirse a la carpeta donde se encuentran los borradores de las secuencias y genes de referencia (Assembly) con el comando cd.

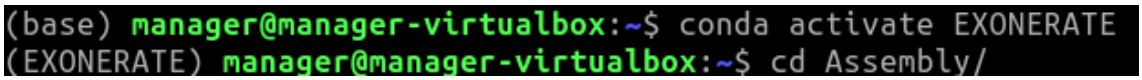


Figura 5. Ambiente EXONERATE activado*. En la primera línea de la imagen se observa el antes a activar el ambiente con el que se va trabajar, y en la segunda línea se observa cómo se activa el ambiente EXONERATE (versión 2.4.0), el cual esta antes del nombre de la máquina virtual, y seguido de este se encuentra el nombre de la carpeta donde se va a trabajar.*

Exonerate goza de varios modelos y opciones que se pueden utilizar combinados en la línea de comando, los que se utilizaron fueron: *--useaatla* el cual muestra la abreviatura de los AA en tres letras, y se le pone FALSE para desactivar esta opción y que salga la "M" en lugar de "Met" en la alineación; *model* para elegir el modelo de alineación a utilizar, en este caso se usó protein2genome, el cual da la opción de alinear ADN genómico con una secuencia de proteínas; *percent 50* para mostrar el porcentaje de identidad, donde se quiere aquellos que tengan al menos el 50% de este entre las secuencias a comparar; *-q* que indica la secuencia query o la secuencia a consultar, que vienen a ser las proteínas de referencias en formato FASTA; *-t* el cual vienen a ser las secuencias objetivo, en este caso los borradores de los genomas de café, siempre en formato FASTA; *showalignment* que muestra los resultados del alineamiento en forma leíble para todas las personas, pero se pone FALSE ya que se quiere mantener la integridad del formato GFF; *showcigar* que indica el resultado en formato CIGAR, es decir un informe compacto de alineación de espacios idiosincrásicos, pero en este caso va acompañado de FALSE ya que no se requieren las diferencias entre las secuencias en el alineamiento resultante, es decir deleciones, inserciones, etc.; *showvulgar* que muestra el alineamiento en formato "vulgar" pero ya que no se quiere las etiquetas <label, query_length, target_length> y el alineamiento con espacio se pone FALSE; *showtargetgff* para obtener un archivo resultante en formato GFF y por último se crea un nuevo archivo (se acaba la línea de comando con el signo mayor que >) con el nombre deseado, en este caso: especie de café_ gen de referencia .gff (EMBL's European Bioinformatics Institute, 2024). Por lo tanto, la línea de comando utilizado con exonerate fue:

exonerate --useaatla false --model protein2genome --percent 50 -q XMT_AFV60437.1.aa -t coffea_arabica_v0.6_06.25.19.fasta --showalignment false --showcigar false --showvulgar false --showtargetgff > arabica_XMT.gff

De este modo tenemos un archivo GFF2 con el resultado del alineamiento entre el borrador del genoma de café con la proteína de referencia.

4.5. Procesamiento archivo GFF

Como resultado del alineamiento se tiene un archivo GFF2 (Figura 6) el cual debe ser modificado a GFF3 para poder transformarlo a archivo FASTA.

```
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  gene    7379904 7381770 1400    -       .       gene_id 1 ; sequence XMT_AFV60
437.1 ; gene_orientation + ; identity 78.17 ; similarity 87.06
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  cds     7381696 7381770 .       -       .
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  exon    7381696 7381770 .       -       .       insertions 0 ; deletions 0 ; i
dentity 76.92 ; similarity 88.46
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  splice5 7381694 7381695 .       -       .       intron_id 1 ; splice_site "GT"
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  intron  7381494 7381695 .       -       .       intron_id 1
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  splice3 7381494 7381495 .       -       .       intron_id 0 ; splice_site "AG"
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  cds     7381074 7381493 .       -       .
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  exon    7381074 7381493 .       -       .       insertions 0 ; deletions 0 ; i
dentity 78.01 ; similarity 84.40
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  splice5 7381072 7381073 .       -       .       intron_id 2 ; splice_site "GT"
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  intron  7380784 7381073 .       -       .       intron_id 2
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  splice3 7380784 7380785 .       -       .       intron_id 1 ; splice_site "AG"
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  cds     7380523 7380783 .       -       .
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  exon    7380523 7380783 .       -       .       insertions 0 ; deletions 0 ; i
dentity 80.68 ; similarity 87.50
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  splice5 7380521 7380522 .       -       .       intron_id 3 ; splice_site "GT"
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  intron  7380303 7380522 .       -       .       intron_id 3
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  splice3 7380303 7380304 .       -       .       intron_id 2 ; splice_site "AG"
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  cds     7380146 7380302 .       -       .
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  exon    7380146 7380302 .       -       .       insertions 0 ; deletions 0 ; i
dentity 75.00 ; similarity 88.46
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  splice5 7380144 7380145 .       -       .       intron_id 4 ; splice_site "AT"
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  intron  7380107 7380145 .       -       .       intron_id 4
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  splice3 7380107 7380108 .       -       .       intron_id 3 ; splice_site "CG"
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  cds     7379904 7380106 .       -       .
CC1.8.Chr02_RagTag      exonerate:protein2genome:local  exon    7379904 7380106 .       -       .       insertions 0 ; deletions 0 ; i
dentity 77.61 ; similarity 91.04
```

Figura 6. Archivo GFF2 resultado del alineamiento entre un gen de referencia y un borrador de genoma de café.

Se tienen varios datos que no son de importancia para la transformación al formato objetivo, en este caso solo se necesitan los CDS, o más conocidos como secuencias codificantes, para esto se aplica la siguiente línea de comando:

sed '/splice/d' arabica_XMT.gff | sed '/intron/d' | sed 's/exonerate://' | sed '/exon/d' | sed 's/; similarity.//' | sed '/similarity/d' > arabica_XMTlisto*

Dónde se eliminarán las filas no deseadas con el comando *sed* y se crea un nuevo archivo el cual estará más limpio, como se observa en la figura 7

```
CC1.8.Chr02_RagTag      protein2genome:local    gene    7379904 7381770 1400    -       .       gene_id 1 ; sequence XMT_AFV60437.1 ; gene_orientation + ; identity 78.
17
CC1.8.Chr02_RagTag      protein2genome:local    cds     7381696 7381770 .       -       .
CC1.8.Chr02_RagTag      protein2genome:local    cds     7381074 7381493 .       -       .
CC1.8.Chr02_RagTag      protein2genome:local    cds     7380523 7380783 .       -       .
CC1.8.Chr02_RagTag      protein2genome:local    cds     7380146 7380302 .       -       .
CC1.8.Chr02_RagTag      protein2genome:local    cds     7379904 7380106 .       -       .
```

Figura 7. Archivo GFF2 únicamente con secuencias codificantes.

En este nuevo archivo editado ya no se tiene la presencia de líneas con nombres de intrones, exones, splices y los datos de similitud como se muestran en la figura 6 ya que

interfieren en el análisis y ya están dentro de las líneas “cds”, enseguida a esto se sigue editando el archivo con la siguiente línea de comando:

sed -i 's/gene_id /ID=gene /' arabica_XMTlisto

Esto debido a que los archivos GFF3 deben tener el formato de clave=valor, en este caso ID=gene, mas no gene_id 1 como se tiene en el resultado del alineamiento.

Después de este paso entra la parte manual, donde hay que llenar todo el campo de atributos con la clave y el valor correspondiente, como se puede observar en la figura 8, para esto se pone el comando *vi* seguido del nombre del archivo, aquí se entra a la edición de texto, se aplasta la tecla *a* para escribir y para salir se aplasta la tecla *esc*, dos puntos (:) y se digita *wq* para guardar, los cambios a realizarse se los efectúa en la columna de atributos, donde se completa con ID=exon1; Parent=gene1 en cada fila, solo varia el número de exón, todo esto separado por punto y coma, lo cual quiere decir que todo este bloque pertenece al gen 1, pero tiene 5 agrupaciones de exones en transcripciones (Ensembl release 111, 2024).

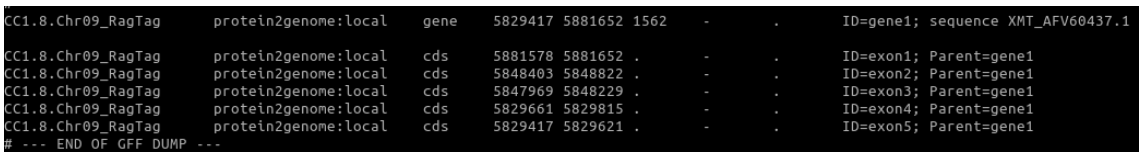


Figura 8. Archivo GFF3 listo. *Aquí se tienen todos los campos llenos, con una representación detallada y además correctamente estructurada de la información genómica obtenida.*

4.6. Transformación GFF a FASTA – GFFread

Una vez listo los archivos GFF3, se continua con la transformación a formato FASTA, para esto se activó el ambiente GFFREAD (versión 0.12.7) y se pone la siguiente línea de comando:

gffread -x arabica_fastaXMT -g coffea_arabica_v0.6_06.25.19.fasta arabica_XMTlisto

Dónde *-x* es el archivo nuevo que se generara, en este caso el archivo FASTA, *-g* el borrador del genoma del café que se está utilizando, seguido del archivo GFF3.

Figura 9. Archivo FASTA transformado desde archivo GFF.

Finalmente se tiene el archivo en formato FASTA, donde se observa los genes obtenidos después del alineamiento realizado en los pasos anteriores.

4.7. Traducción de secuencias

Ya con el archivo en formato FASTA, hay que traducirlo a proteína, para esto se activa el ambiente EMBOSS (versión 6.6.0) y se digita la siguiente línea de comando:

transeq arabica_fastaXMT

Con esto se genera un mensaje (Translate nucleic acid sequences) el cual indica que el proceso se está llevando a cabo, donde hay que digitar el nuevo nombre del archivo ya traducido.

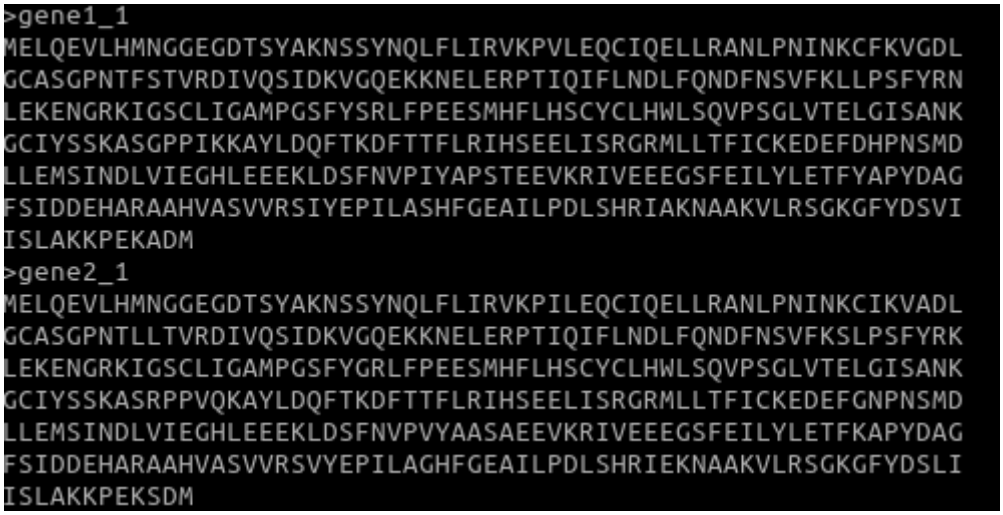


Figura 10. Archivo FASTA traducido.

Es así que se obtuvo el archivo traducido (figura 10) al cual hay que cambiarlo de nombre para evitar confusiones al momento del alineamiento múltiple.

4.8. Cambio de nombres de secuencias FASTA

Figura 11. Archivo FASTA con sus nombres ya editados.

En la figura 11 se tiene los diferentes genes ya traducidos obtenidos del alineamiento realizado con los genes de referencia (*XMT, DXMT, MXMT*) y el borrador del genoma, en este caso de *C. canephora*. Cada uno con su nombre propio, debido que en el alineamiento

múltiple y en la elaboración de los árboles filogenéticos no se permiten nombre duplicados.

4.9. Concatenar archivos producidos con proteínas de referencia + grupo externo

Para el alineamiento múltiple se debe tener un solo archivo con todas las proteínas de referencia, proteínas obtenidas después del alineamiento y grupo externo, para que el algoritmo pueda comparar y analizar las muestras de manera simultánea, para esto se usa el comando *cat* seguido de los archivos a unir y se crea un nuevo archivo, como se observa en la línea de comando:

cat arabica_fastaXMT arabica_fastaDXMT arabica_MXMT PROTEINASdeREFERENCIA OUTGROUP > arabica_final

4.10. Alineamiento múltiple – MUSCLE

Ya con los nuevos archivos con toda la información unida, se procede a activar el ambiente de MUSCLE (versión 5.1) para realizar el alineamiento múltiple y se digita la siguiente línea de comando:

muscle -align final_arabica_KraussiaXDM -output arabica_KraussiaXDM_alineado.afa

Dónde *-align* es el archivo a ser alineado y *-output* el nuevo archivo a ser generado después del alineamiento múltiple

Figura 12. Alineamiento múltiple.

4.11. Creación de árbol filogenético – FASTTREE

Con el alineamiento múltiple listo, se sigue con la creación de todos los árboles filogenéticos, esto activando el ambiente de FASTTREE (versión 2.1.11), y poniendo la siguiente línea de comando:

fasttree arabica_KraussiaXDM_alineado.afa > arabica_arbol

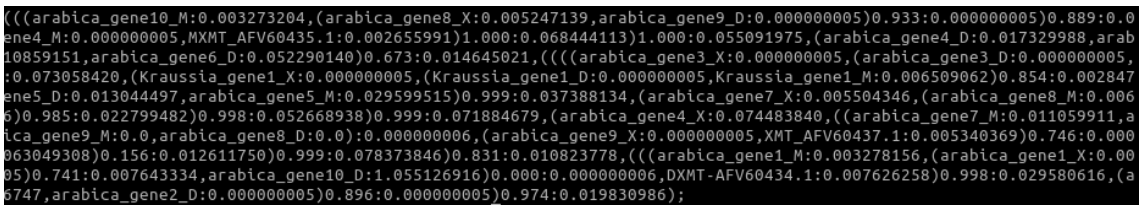


Figura 13. Archivo de salida de FastTree en formato Newick.

Se tiene un nuevo archivo listo para ser ingresado en editores de árboles, el cual tiene toda la información para la representación de árboles filogenéticos.

4.12. Edición y visualización de los árboles

En el programa iTOL (https://itol.embl.de/), hay que crearse una cuenta, con esto se pueden cargar los archivos en formato Newick, para tener un buen árbol hay que activar la opción "*midpointroot*" (para evitar sesgos en el análisis, tener neutralidad, simplicidad

y una mejor comparación de resultados) que está en control panel, en la pestaña de avanzado, además activar la visualización de *Bootstrap* con su display en texto. Con esto listo, se pueden poner colores a las ramas, texto y muchas más funciones de acuerdo a lo que usuario desee.

5. Resultados

Se obtuvieron 17 árboles filogenéticos, dónde el más relevante con cafeína fue *C. canephora* y sin cafeína *C. humblotiana* y *C. pseudozanguebariae*, esto ya que el primero tienen la presencia de los 3 genes implicados en la ruta de producción de cafeína, el segundo no tiene la cafeína sintasa y el ultimo no tiene ninguno de los 3 genes, los cuales fueron obtenidos tras el alineamiento múltiple entre los genes de referencia (*XMT, DXMT* y *MXMT*) obtenidos de *C. canephora* y los borradores de los genomas de las especies mencionadas anteriormente. Después del nombre de cada gen, se puede observar la letra X (*XDM*), D (*DXMT*) y M (*MXMT*) que se puso para tener referencia de que alineamiento fueron obtenidos, y por último el outgroup de *K. floribunda* (Rubiaceae) para el correcto enraizamiento de los árboles. Se puede observar de 1 a 3 genes iguales (gene1, gene2, gene3) pero con un distintivo al final (X, D, M) ya que son provenientes de alineamientos individuales con cada gen de referencia *DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437*.1. Cabe señalar que la detección de homólogos de los genes *XMT*, *DXMT* y *MXMT* puede ser redundante. De hecho, utilizamos los genes *C. canephora. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1* y éstos pueden detectar los mismos genes. Así, en estos árboles se detectó un gen en *K. floribunda* gen*1M*, gene*1X* y *gene1D* por sus homologías en *C. canephora (XMT, MXMT, DXMT*). Se trata probablemente de un único gen en *K. floribunda*. De la misma manera, se tiene el mismo caso en la detección de otros genes que pueden ser redundantes en cada alineamiento realizado.

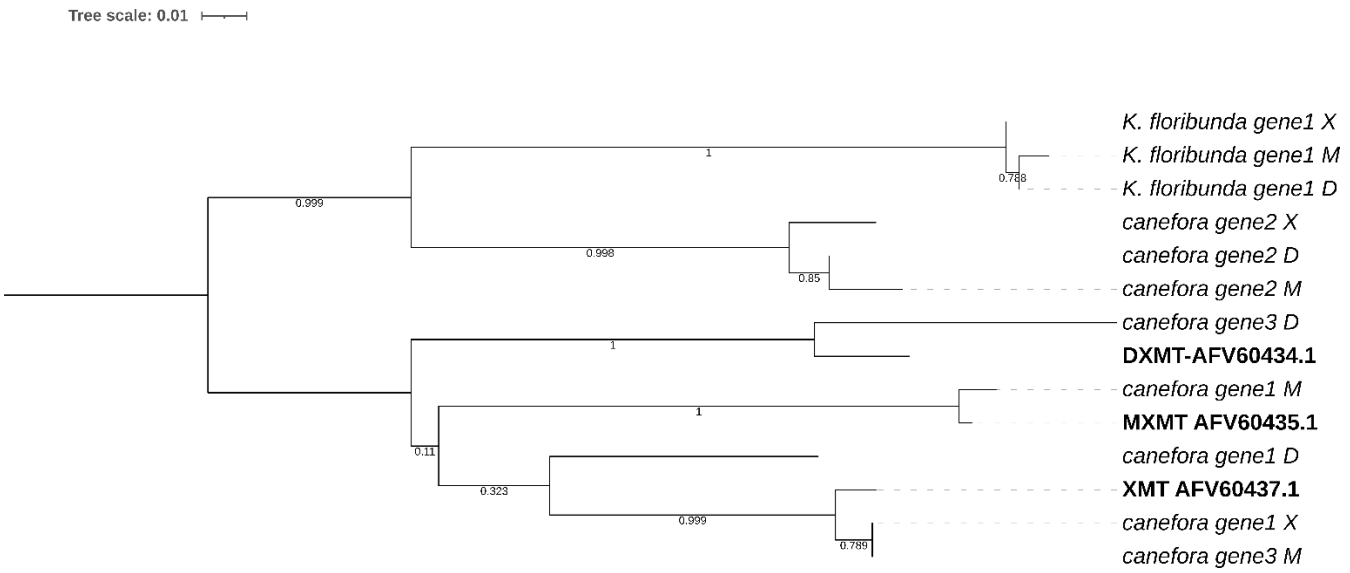


Figura 14. Árbol filogenético de genes NMT de C. canephora. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia *en C. canephora* (Perrois et al., 2015).

Presencia de genes *XMT, DXMT* y *MXMT*, implicados en la producción de cafeína en plantas de café, se agrupan en la segunda rama del árbol, *DXMT, MXMT y XMT* con un Bootstrap igual a 1, 1 y 0.99 correspondientemente, mientras que en la otra rama se encuentran otros genes encontrados en el alineamiento y con función aún desconocida con un Bootstrap igual a 0.999. Dónde los 3 genes están super conservados y son homólogos (vienen de duplicaciones ancestrales), tienen entre 90-95% de identidad a nivel de las proteínas.

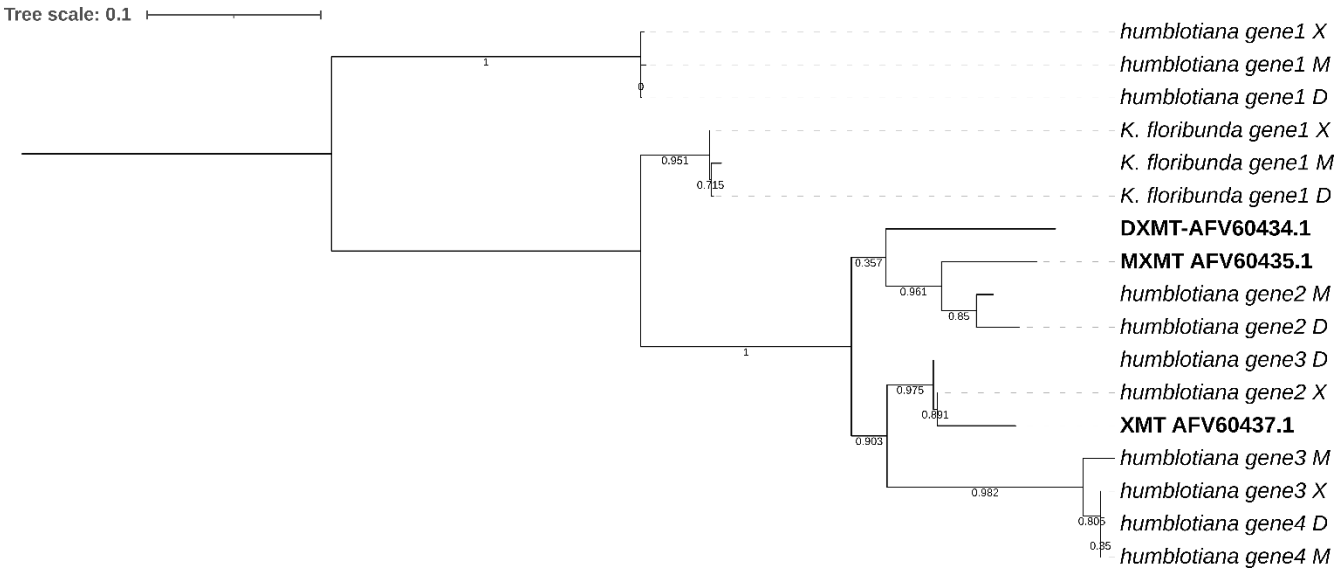


Figura 15. Árbol filogenético de genes NMT de C. humblotiana*. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en C. canephora* (Perrois et al., 2015)

Presencia de genes de *XMT* y M*XMT*, pero ausencia de genes en la rama del gene de D*XMT* el cual es un indicador de la ausencia de cafeína en esta especie de café, todos los genes de interés se encuentran agrupados en una rama, con un Bootstrap igual a 1. Se puede observar de 1 a 3 genes iguales (gene1, gene2, gene3, gene4) pero con un distintivo al final (X, D, M) ya que son provenientes de alineamientos individuales con cada gen de referencia *DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1*

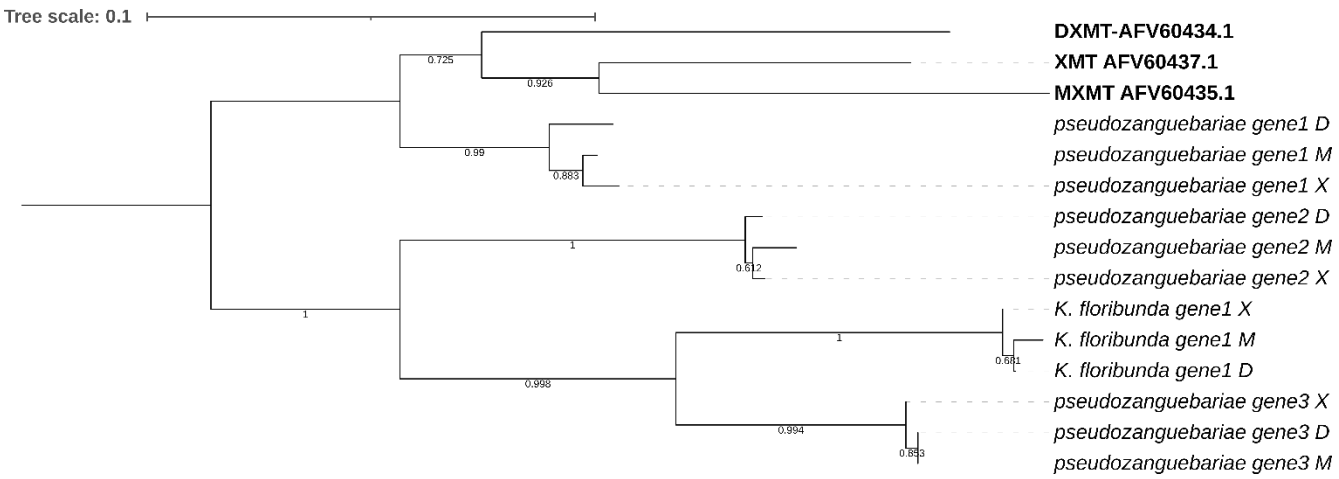


Figura 16. Árbol filogenético de genes NMT de C. pseudozanguebariae. *DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1* son los genes de referencia en *C. canephora* (Perrois et al., 2015).

Ausencia de genes *MXMT, XMT y DXMT*. Esto debido a que es una especie sin rastro de cafeína en su interior, razón por la cual la posición de estos genes en el árbol, es decir en lo más alto del mismo y en rama independiente con un Bootstrap igual a 0.725. Se puede observar de 1 a 3 genes iguales (gene1, gene2, gene3) pero con un distintivo al final (X, D, M) ya que son provenientes de alineamientos individuales con cada gen de referencia *DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437*.1.

Tabla 3. Resumen del número de genes XMT, DXMT y MXMT de acuerdo a cada especie estudiada.

Se observa cuantos genes implicados en la ruta de la cafeína están presentes por especie, después de aplicar la filogenia, esto ya que hay 3 genes involucrados en la producción de cafeína donde se puede ver si tenemos cada uno de estos agrupados con la referencia de *C. canephora*.

ESPECIE	**GEN**			**Cafeína**
	XMT	**MXMT**	**DXMT**	
C. canephora	3	1	1	Con
C. stenophylla.	2	3	3	Con
C. humilis.	4	1	6	Con
C. eugenoides.	2	3	3	Con
C. kianjavatensis.	9	3	3	Con
C. salvatrix.	3	3	3	Con
C. arabica	6	5	0	Sin

C. lancifolia.	0	3	6	Sin
C. humblotiana	6	2	0	Sin
C. homollei.	0	0	0	Sin
C. myrtifolia.	0	0	0	¿?
C. *pseudozanguebariae*	0	0	0	Sin
C. farafanganensis.	0	0	0	Sin
C. charrieriana.	4	0	0	Sin
C. tetragona.	3	3	0	Sin
C. mauritiana.	0	0	0	Sin
C. racemosa.	4	0	3	Sin

Cada resultado obtenido depende de los genes que estén más cercanos y en una misma rama de los genes de referencia en *C. canephora DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437* (Perrois et al., 2015).

6. Discusión

Actualmente es muy rápido y barato secuenciar genomas en Illumina (secuencias cortas), por ejemplo, secuencias de ADN o ARN tienen un precio de $150 por muestra (NC State University, 2024). Tomando en cuenta épocas pasadas, donde esta era una tecnología nueva y sus costos eran muy elevados, en comparación a esta época, los costos están reducidos en su totalidad, en el año 2001 el valor por megabase cruda de ADN a secuenciar tenía un valor estimado de $10.000.000, mientras que en datos tomados en el 2022 muestra como este precio baja a $10 (National Human Genome Research Institute, 2023). Lo cual permite avanzar con investigaciones y descubrir nuevas cosas que antes no se podía por limitantes como lo es el dinero, actualmente, por ejemplo, una secuenciación de genoma de café tiene un valor de aproximadamente $75 por muestra con una cobertura de 10x con la tecnología Illumina. Esto significa que se puede explorar la diversidad de las variedades de especies de café a bajo costo. Lo cual tiene muchas ventajas ya que se pueden continuar con estudios que tendrán aplicaciones en genética, agronomía y generalmente en ciencias.

Los instrumentos de secuenciación dan como resultado numerosos terabytes de información que son analizados con bioinformática, estos datos generados se han convertido en un cuello de botella debido a que se necesita el desarrollo de metodologías de rápido análisis y el acceso a supercomputadores como clusters. Como se sabe, los

estudios en biología entraron en una nueva época, la era digital, donde se tienen alternativas para salir de este cuello de botella que se ha venido generando a lo largo de los tiempos, esta alternativa es el uso de máquinas virtuales, que brindan la posibilidad de encapsular varios servidores informáticos muy completos con el propio sistema operativo y con paquetes de software existentes (Wultsch, 2015). Una ventaja muy visible en el manejo de máquinas virtuales en relación a Clusters, es que no se necesita ser administrador o Root, no se necesitan permisos para instalaciones ni más, con el uso de Conda las instalaciones son más rápidas y eficientes, y además que se tiene una infinidad de paquetes (ANACONDA.ORG, 2024).

Ya con estas herramientas (MV) para poder trabajar en otra interfaz para el manejo de datos de secuenciación sin tener que acceder a Clusters pagados, el paso siguiente es eliminar otra barrera, que es el desarrollar metodologías de análisis rápidas, en el actual trabajo se trabajó con un flujo de trabajo (figura 3) que funcionó correctamente, fue desarrollado, probado y efectuado sin problema, existen pasos (5, 7, 9) donde es necesario usar el terminal para agregar datos faltantes en los archivos generados, pero con nuevos algoritmos se puede automatizar este flujo y se puede tener resultados de una manera más rápida con la misma eficacia.

La metodología puesta a prueba se puede corroborar como exitosa ya que se tiene a *C. canephora* como control positivo, el cual proporciona un efecto de referencia para poder juzgar otras intervenciones (Steckler et al., 2020). Al tener los genes de referencia (*DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437*) de *C. canephora* presentes en el árbol filogenético (figura 14) de *C. canephora* se comprueba que el flujo de trabajo es el correcto, y se puede efectuar esta metodología con más especies con una total certeza, ya que los resultados serán los esperados. Así mismo, se utilizó a *C. humblotiana* como control negativo para poder descartar posibles interpretaciones no causales en los resultados (Lipsitch et al., 2010), se utilizó esta especie de café debido a que se sabe por qué no tiene cafeína, en este caso no tiene el gen *DXMT* (Raharimalala et al., 2021), y se corrobora en el árbol filogenético obtenido (figura 15). De igual manera se tiene un árbol extra (Figura 16) que corresponde a *C. pseudozanguebariae,* que es una especie sin rasgos de cafeína, y de igual manera se lo corrobora ya que en el árbol no se tiene la presencia de ninguno de los 3 genes implicados en la síntesis de cafeína.

Ya con el flujo de trabajo con resultados positivos, se continuó con el análisis la presencia o ausencia de los genes *NMT* de la cafeína, en la tabla 3 se observan los resultados

obtenidos en este análisis, donde se tienen a 6 especies (*C. canephora, C. salvatrix, C. stenophylla, C. humillis, C. eugenoides, C. kianjavatensis*) con todos los genes (*XMT, MXMT, DXMT*) involucrados en la producción de cafeína, mientras que se tiene a 10 especies sin 1, 2 o 3 genes involucrados en la síntesis de cafeína (*C. arabica, C. humblotiana, C. homollei, C. myrtifolia, C. pseudozanguebariae, C. farafanganensis, C. charrieriana, C. tetragona, C. mauritiana, C. racemosa, C. lancifolia)*, por lo que son catalogadas especies sin cafeína, pero se sabe según la bibliografía que *C. racemosa, C. arabica y C. lancifolia* si tienen todos los genes productores del alcaloide. Además, cabe indicar que tenemos una especie extra la cual se desconocía si tenía o no cafeína (C. *myrtifolia*) y que después de los análisis muestra que no tiene la presencia de los 3 genes involucrados en la producción de este alcaloide.

C. humblotiana, café sin cafeína debido a la ausencia de un segmento con peso igual a 76kb dentro del cromosoma #1 que contiene al gen *DXMT*, que es encargado de la síntesis de la teobromina en cafeína (Raharimalala et al., 2021), es el ejemplo base para saber que una planta de café necesita la presencia de los 3 genes (*XMT, MXMT, DXMT*) para poder sintetizar este alcaloide. Según Raharimalala et al., la razón para la ausencia de este segmento puede ser debido a una recombinación ilegitima durante este proceso; se tienen a los 2 primeros genes de la ruta de la cafeína (*XMT* y *MXMT*) funcionales, ya que la xantosina es convertida en teobromina, por lo que esta ruta biosintética de cafeína termina en la producción de teobromina. Como se observa dentro de la tabla 3, la especie de café *C. arabica (*especie tetraploide)*,* tiene 6 genes *XMT*, 5 *MXMT* y ningún gen *DXMT*, caso similar a *C. humblotiana*, esto quiere predecir que este café no tiene cafeína, pero en la bibliografía se tiene en claro que *C. arabica* tiene de 0.96 a 1.62% de cafeína (Wild Coffea Species Database, n.d.-a), en este caso para esta especie de café hay que observar que se tenga el genoma completo, a su vez puede existir una mutación natural de la planta o el alineamiento con Exonerate resulto deficiente. Se tiene otra hipótesis, la cual dice que esta planta sea proveniente de Etiopía, donde se realizó el descubrimiento de esta especie de café, pero con diferencia que ésta es descafeinada naturalmente, en este caso se encontró que estas plantas descafeinadas naturalmente, tenían acumulaciones de teobromina (precursor de cafeína) con lo que esta planta no podría tener la enzima cafeína sintasa que es la que sintetiza la teobromina en cafeína (Silvarolla et al., 2004). De igual manera los resultados de *C. charrieriana* y *C. tetragona* indican que no cuentan con el

ultimo gen (*DXMT*) lo que hace estén en el grupo de plantas de café sin cafeína lo cual se corrobora con la bibliografía (http://publish.plantnet-project.org/project/wildcofdb_en).

C. lancifolia es establecida como una especie de café con cafeína (0.70%, (Wild Coffea Species Database, n.d.-b)), en la tabla 3 se puede observar que no tiene la presencia de genes *XMT*, que tiene 3 genes *MXMT* y 6 genes *DXMT*. Al no tener el primer gen (*XMT*) de la ruta de la cafeína esta planta no sería capaz de producir cafeína, existen hipótesis para la producción de cafeína de esta especie, la primera es la presencia de rutas menores que empiezan con la 7-metilxantina (*MXMT*) y de igual manera tiene como resultado cafeína, hay que mencionar que tiene a la paraxantina como intermediario en esta ruta de síntesis de la cafeína (Ashihara & Suzuki, 2004). La segunda hipótesis es que el genoma no está completo, y justamente la región faltante es donde este gen *XMT* podría estar presente.

Para Ashihara y Suzuki para que la vía de la cafeína tenga como resultado cafeína debe existir aceptores de metilo, que en este caso son los genes *XMT* o *MXMT*, donde *C. racemosa* tiene 4 genes *XMT*, ningún gen MXMT y 3 genes DXMT (tabla 3) el cual puede producir cafeína sin ningún problema, ya que tiene a *XMT* como aceptor de metilo, pero como se sabe desde un inicio, se necesitan los 3 genes para la producción de este alcaloide, por lo que no hay que descartar la idea que pueden existir otros genes que tengan la función de *MXMT*. Adicionalmente, existen casos en el que la función del gen *DXMT* es doble, es decir puede agregar grupos metilos en la teobromina (cafeína sintasas) como en 7-metilxantosina (teobromina sintasa) (Mizuno et al., 2003). Hay que mencionar que se tiene una actividad más elevada de N-metiltransferasas en hojas jóvenes en comparación a hojas ya desarrolladas y que la cafeína sintasa se encuentra en los cloroplastos de las hojas que están en desarrollo(Ashihara & Suzuki, 2004).

Pueden existir también versiones que indican que la presencia o ausencia de cafeína depende de la ubicación geográfica de la planta, es el caso de cafés silvestres ubicados en islas del Océano Índico (Madagascar) como *C. homollei, C. farafanganensis,* C. *tetragona* y África Oriental (Islas Mascareñas) como *C. mauritiana y C. myrtifolia* las cuales sugieren que su ancestro común no tendría rastros de cafeína en su genoma (Raharimalala et al., 2021). Sin embargo, puede presentarse excepciones, es el caso de *C. charrieriana* que se encuentra en África central y no tiene cafeína y *C. kianjavatensis* (0.7%) y *C. lancifolia* (0.7%), de Madagascar con presencia de cafeína razón por la cual se cuestiona la hipótesis del ancestro sin cafeína, lo que da apertura a nuevas hipótesis

sobre los mecanismos para poder explicar la ausencia o presencia de cafeína en estas plantas. Se necesitan más estudios en genética para determinar la funcionalidad de los genes *NMT* con lo que se podrá tener una hipótesis más fuerte de los mecanismos de evolución de estos genes en café y poder realizar un modelo evolutivo; estudios realizados muestran una microcolinealidad en los locus *XMT/MXMT* dentro del cromosoma #9 que indica un patrón de evolución que está ligado a los genes *NMT*, este patrón tiene pseudogenización y duplicaciones en tándem en un entorno con un 60% de elementos transponibles, lo que indica cruzamientos intergénicos desiguales, que tienen como resultado duplicaciones, perdidas y mutaciones que pueden explicar los cambios en la genética de las regiones *NMT* en el cromosoma #9 (Raharimalala et al., 2021). Lo que indica que la presencia o ausencia de cafeína no depende de la región en donde estos cafés se encuentren.

7. Conclusión

Gracias a las máquinas virtuales que tienen un propio sistema operativo y que sus elementos de software están encapsulados dentro de este, se eliminan la mayoría de obstáculos técnicos presentes para instalaciones de paquetes bioinformáticos y para el manejo de los mismos, con lo que se pudo poner en desarrollo, probar y ejecutar un fujo de trabajo que sirvió totalmente para la identificación de genes, a pesar de no tener una supercomputadora todos los algoritmos corrieron sin problema, no hubo daños en la computadora y los resultados fueron muy buenos, en conclusión la metodología implementada cumple con el objetivo. Cabe mencionar que este es un trabajo piloto donde se trabajó solo con 17 especies, donde se pudo explicar la ausencia o presencia de estos genes, en un futuro puede ser utilizado para trabajo con más cantidad de especies, ser usado en estadística para ver su correlación, y poder determinar su posición geográfica.

Gracias a los controles positivos y negativos de referencia se pudo validar la metodología puesta en marcha, donde los resultados obtenidos tenían similitud con los resultados encontrados en bibliografía. El ensamblaje de *C. humblotiana* es un genoma de referencia con una altísima calidad dentro del género *Coffea*, se pudo obtener valiosa información junto a los genes de referencia provenientes de *C. canephora*, donde se quedó en claro que se necesitan obligatoriamente los 3 genes de la ruta de síntesis de cafeína para la producción de esta, donde los 2 primeros son los aceptores de metilo y el tercero que va

a ser el único que pueda seguir con la ruta al sintetizar la teobromina para tener como resultado final, la cafeína.

Una aplicación importante sobre este trabajo es la posibilidad de identificar genes o familias de genes en las lecturas o ensamblaje de secuencias, además de que proporciona una valiosa información para poder promover la preservación del café y puede ser un recurso para estudios evolutivos y genómicos de tanto género como familia de estas especies. Puede ser también tomado como plantilla para generar nuevos pipelines y automatizarlos para poder procesar muestras más grandes y en un menor tiempo. No hay que dejar a un lado las aplicaciones biotecnológicas, donde se puede usar CRISPR CAS-9 o a su vez buscar el conocido *Decaffito* (*C. arabica* sin cafeína naturalmente) para hibridar con otras especies y seguir con las mismas cualidades organolépticas de un *C. arabica* normal, lo que no se obtiene en un café proveniente de *C. humblotiana* que tienen otros compuestos amargos como la cafamarina.

8. Referencias

agropolis fondation. (2024). *Wild Coffea Species Database*. http://publish.plantnet-project.org/project/wildcofdb_en

ANACONDA.ORG. (2024). *anaconda / packages*. https://anaconda.org/anaconda/repo

Ashihara, H., & Suzuki, T. (2004). DISTRIBUTION AND BIOSYNTHESIS OF CAFFEINE IN PLANTS. In *Frontiers in Bioscience* (Vol. 9).

Center for Quantitative Life Sciences. (2023). *muscle 5.1.0*. Oregon University. https://software.cqls.oregonstate.edu/updates/muscle-5.1.0/

CONDA. (2017a). *Getting started with conda*. https://conda.io/projects/conda/en/latest/user-guide/getting-started.html

CONDA. (2017b). *Installing conda* . https://conda.io/projects/conda/en/latest/user-guide/install/index.html

Davis, A. P., & Rakotonasolo, F. (2021). Six new species of coffee (Coffea) from northern Madagascar. *Kew Bulletin*, *76*(3), 497–511. https://doi.org/10.1007/s12225-021-09952-5

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., Aury, J.-M., Bento, P., Bernard, M., Bocs, S., Campa,

C., Cenci, A., Combes, M.-C., Crouzillat, D., Da Silva, C., … Lashermes, P. (2014). *The coffee genome provides insight into the convergent evolution of caffeine biosynthesis*. www.mobot.

EMBL's European Bioinformatics Institute. (2024). *Exonerate: a generic tool for sequence comparison*. https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate-manual

Ensembl release 111. (2024). *GFF3 File Format - Definition and supported options*. http://www.ensembl.org/info/website/upload/gff3.html

Furuno, M., Kasukawa, T., Saito, R., Adachi, J., Suzuki, H., Baldarelli, R., Hayashizaki, Y., & Okazaki, Y. (2003). CDS annotation in full-length cDNA sequence. *Genome Research*, *13*(6 B), 1478–1487. https://doi.org/10.1101/gr.1060303

Global Core Biodata Resource. (2024). *Global Biodiversity Information Facility*. https://www.gbif.org/

GNU Operating System. (2020). *GNU Bash*. GNU. https://www.gnu.org/software/bash/

Guyot, R., Hamon, P., Couturon, E., Raharimalala, N., Rakotomalala, J. J., Lakkanna, S., Sabatier, S., Affouard, A., & Bonnet, P. (2020). WCSdb: A database of wild Coffea species. *Database*, *2020*. https://doi.org/10.1093/database/baaa069

Hamon, P., Grover, C. E., Davis, A. P., Rakotomalala, J. J., Raharimalala, N. E., Albert, V. A., Sreenath, H. L., Stoffelen, P., Mitchell, S. E., Couturon, E., Hamon, S., de Kochko, A., Crouzillat, D., Rigoreau, M., Sumirat, U., Akaffou, S., & Guyot, R. (2017). Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (Coffea) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content. *Molecular Phylogenetics and Evolution*, *109*, 351–361. https://doi.org/10.1016/j.ympev.2017.02.009

Hollingsworth, R. G., Armstrong, J. W., & Campbell, E. (2002). Caffeine as a repellent for slugs and snails. *Nature*, *417*(6892), 915–916. https://doi.org/10.1038/417915a

Idaho University. (n.d.). *Newick Tree Formats*. Retrieved February 24, 2024, from http://marvin.cs.uidaho.edu/Teaching/CS515/newickFormat.html

ILRI Research Computing. (2020). *FastTree*. https://hpc.ilri.cgiar.org/fasttree-software

International Coffee Organization. (2022). *Historia del Café*. https://www.ico.org/ES/coffee_storyc.asp#:~:text=El%20caf%C3%A9%20lleg%C3%B3%20primero%20a,mundo%2C%20en%20las%20Blue%20Mountains

iTOL Interactive Tree of Life. (2024). *Welcome to iTOL v6*. https://itol.embl.de/

Li, H. (2023). Protein-to-genome alignment with miniprot. *Bioinformatics (Oxford, England)*, *39*(1). https://doi.org/10.1093/bioinformatics/btad014

Lin, Z., Wei, J., Hu, Y., Pi, D., Jiang, M., & Lang, T. (2023). Caffeine Synthesis and Its Mechanism and Application by Microbial Degradation, A Review. In *Foods* (Vol. 12, Issue 14). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/foods12142721

Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative Controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, *21*(3), 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb

lubuntu. (2023). *Lubuntu 20.04 LTS (Focal Fossa) Released!* https://lubuntu.me/focal-released/

Madrigal-Valverde, K. A. (2017). Uso de herramientas para alineación de secuencias y creación de árboles filogenéticos para la determinación de especies. *Revista Tecnología En Marcha*, *30*(5), 30. https://doi.org/10.18845/tm.v30i5.3218

Mannino, G. (2023). Discrimination of Green Coffee (Coffea arabica and Coffea canephora) of Different Geographical Origin Based on Antioxidant Activity, High-Throughput Metabolomics, and DNA RFLP Fingerprinting. *Antioxidants (Basel)*.

Mizuno, K., Okuda, A., Kato, M., Yoneyama, N., Tanaka, H., Ashihara, H., & Fujimura, T. (2003). Isolation of a new dual-functional caffeine synthase gene encoding an enzyme for the conversion of 7-methylxanthine to caffeine from coffee (Coffea arabica L.). *FEBS Letters*, *534*(1–3), 75–81. https://doi.org/10.1016/S0014-5793(02)03781-X

National Human Genome Research Institute. (2023). *DNA Sequencing Costs: Data*. NIH. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

NC State University. (2024). *Genomic Sciences Laboratory*. https://research.ncsu.edu/gsl/pricing/

NCBI. (2021). *FASTA Format for Nucleotide Sequences*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/genbank/fastaformat/

Oracle. (2020). *Oracle VM VirtualBox Datasheet*.

Pardo, R., Alvarez, Y., Barral, D., & Farré, M. (2007). Cafeína: un nutriente, un fármaco, o una droga de abuso. *Revista Adicciones*.

Perrois, C., Strickler, S. R., Mathieu, G., Lepelley, M., Bedon, L., Michaux, S., Husson, J., Mueller, L., & Privat, I. (2015). Differential regulation of caffeine metabolism in Coffea arabica (Arabica) and Coffea canephora (Robusta). *Planta*, *241*(1), 179–191. https://doi.org/10.1007/s00425-014-2170-7

Pertea, M., & Pertea, G. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*. https://doi.org/10.12688/f1000research.23297.1

pkgsrc. (n.d.). *biology/gffread - The NetBSD Packages Collection*. Retrieved February 24, 2024, from https://ftp.netbsd.org/pub/pkgsrc/current/pkgsrc/biology/gffread/index.html

Raharimalala, N., Rombauts, S., McCarthy, A., Garavito, A., Orozco-Arias, S., Bellanger, L., Morales-Correa, A. Y., Froger, S., Michaux, S., Berry, V., Metairon, S., Fournier, C., Lepelley, M., Mueller, L., Couturon, E., Hamon, P., Rakotomalala, J. J., Descombes, P., Guyot, R., & Crouzillat, D. (2021). The absence of the caffeine synthase gene is involved in the naturally decaffeinated status of Coffea humblotiana, a wild species from Comoro archipelago. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-87419-0

Ranarivelo, & ND. (2011). *Variabilité de la composition chimique des caféiers spontanés de la région malgache (Mascarocoffea Chev.) : cas de Coffea homollei, C. kianjavatensis, C. lancifolia de la série Verae*.

Rice, P., Longden, I., & Bleasby, A. (2000). The European Molecular Biology Open Software Suite. *Trends in Genetics 16*, *pp276--277*.

Silvarolla, M., Mazzafera, P., & Fazuoli Luis. (2004). A naturally decaffeinated arabica coffee. *NATURE*.

Steckler, T., Bespalo, A., & Michel, M. (2020). *Handbook of Experimental Pharmacology 257*. 2004. http://www.springer.com/series/164

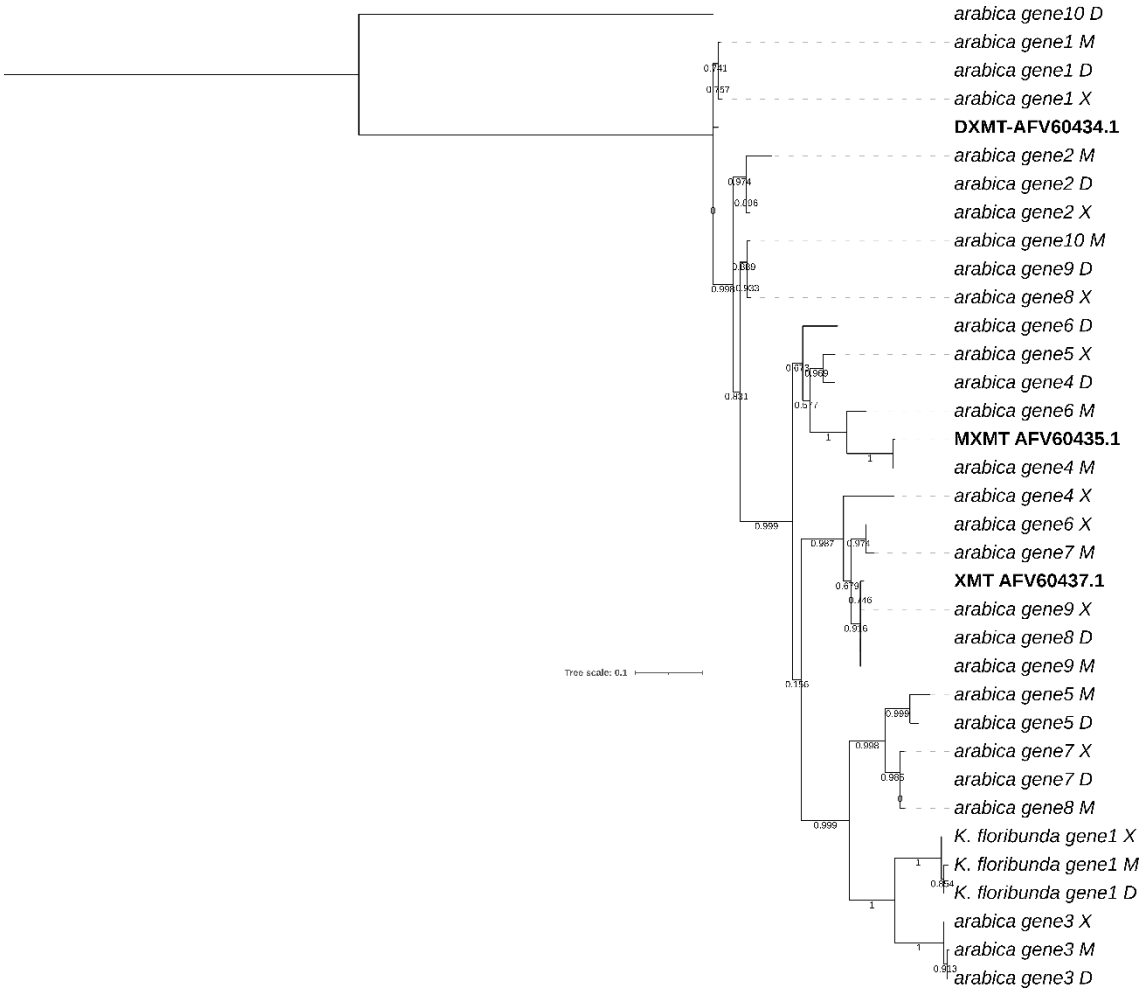
Verónica Belchior. (2019). *¿Qué Ocurre Durante El Tueste Del Café? Los Cambios Químicos*. Perfect Daily Grind. https://perfectdailygrind.com/es/2019/04/05/que-ocurre-durante-el-tueste-del-cafe-los-cambios-quimicos/

Wild Coffea Species Database. (n.d.-a). *C. arabica L. - Eucoffea - ENDANGERED*. Retrieved March 1, 2024, from http://publish.plantnet-project.org/project/wildcofdb_en/collection/coffee-species/species/details/C.%20arabica%20L.

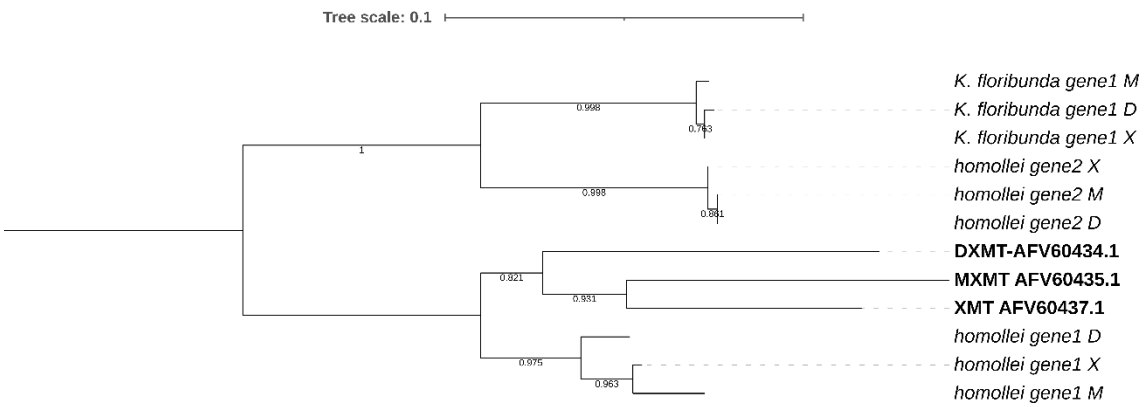
Wild Coffea Species Database. (n.d.-b). *C. lancifolia A.Chev. var. auriculata J.-F.Leroy*. Retrieved March 1, 2024, from http://publish.plantnet-project.org/project/wildcofdb_en/collection/coffee-species/species/details/C.%20lancifolia%20A.Chev.%20var.%20auriculata%20J.-F.Leroy

Wultsch, C. (2015). *A Review of Cloud Computing Bioinformatics Solutions for Next-Gen Sequencing Data Analysis and Research*. https://doi.org/10.13140/RG.2.1.4842.9848

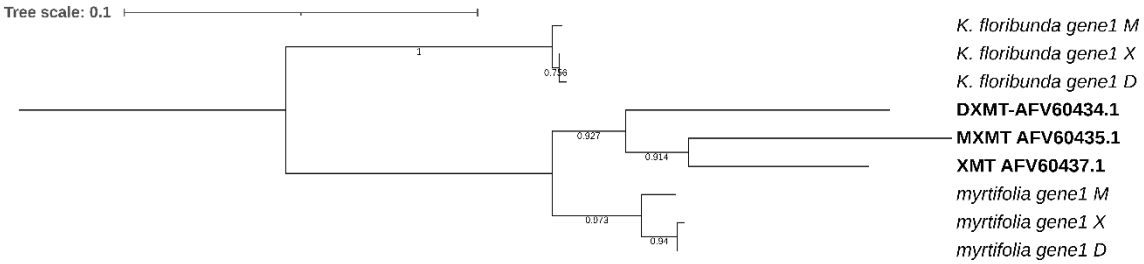
9. Anexos



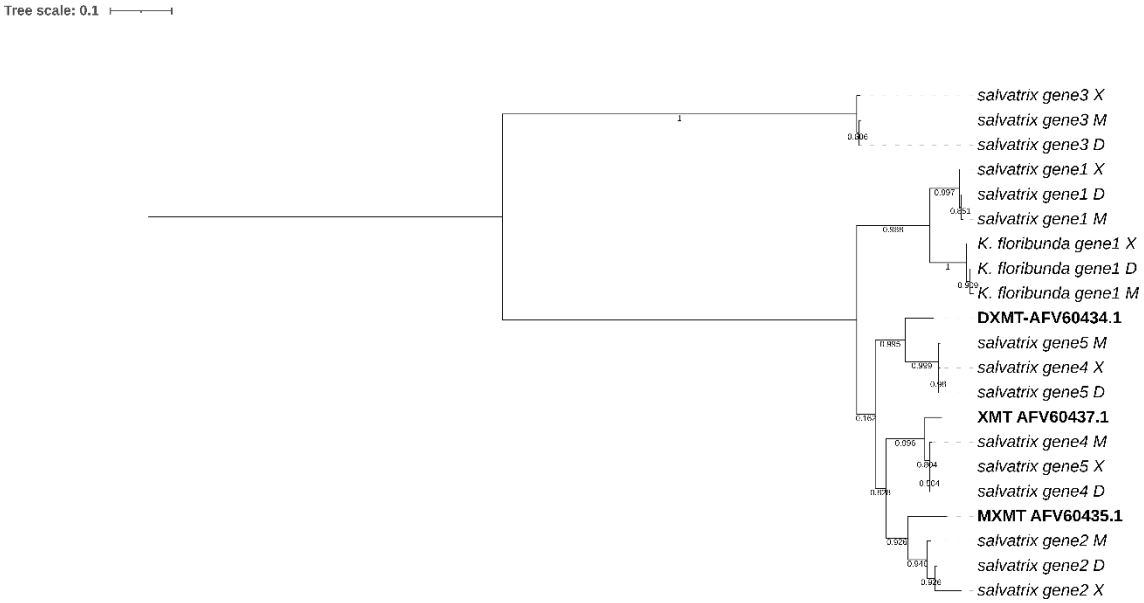
Anexo 1. Árbol filogenético de genes NMT de C. arabica. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



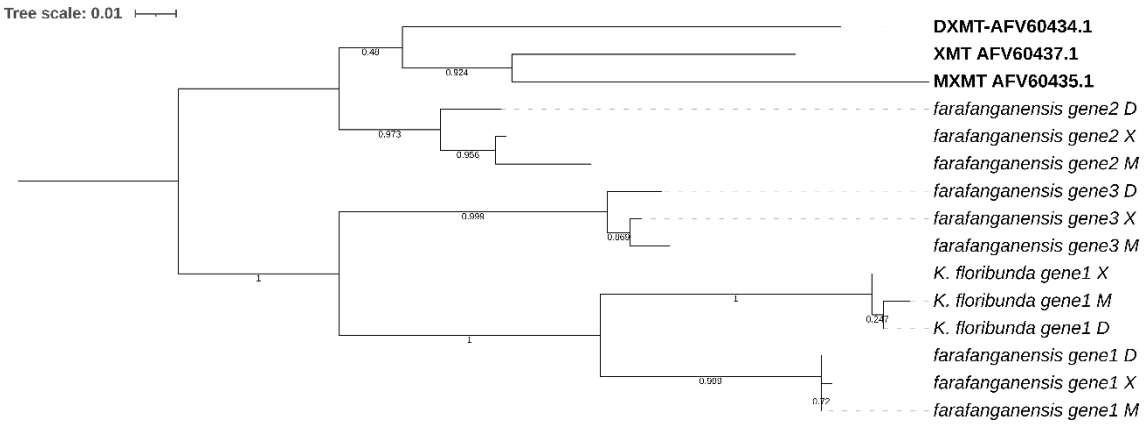
Anexo 2. Árbol filogenético de genes NMT de C. homollei. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



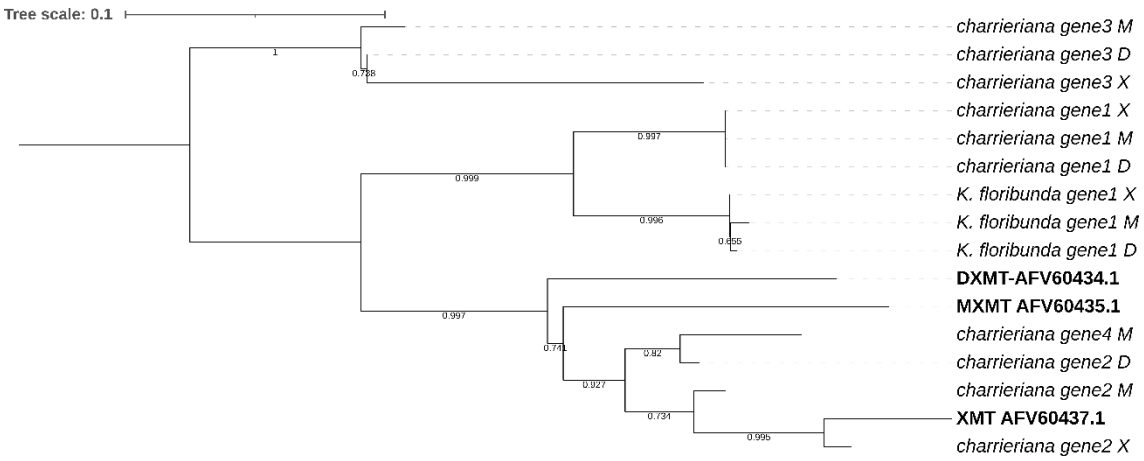
Anexo 3. Árbol filogenético de genes NMT de C. myrtifolia. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



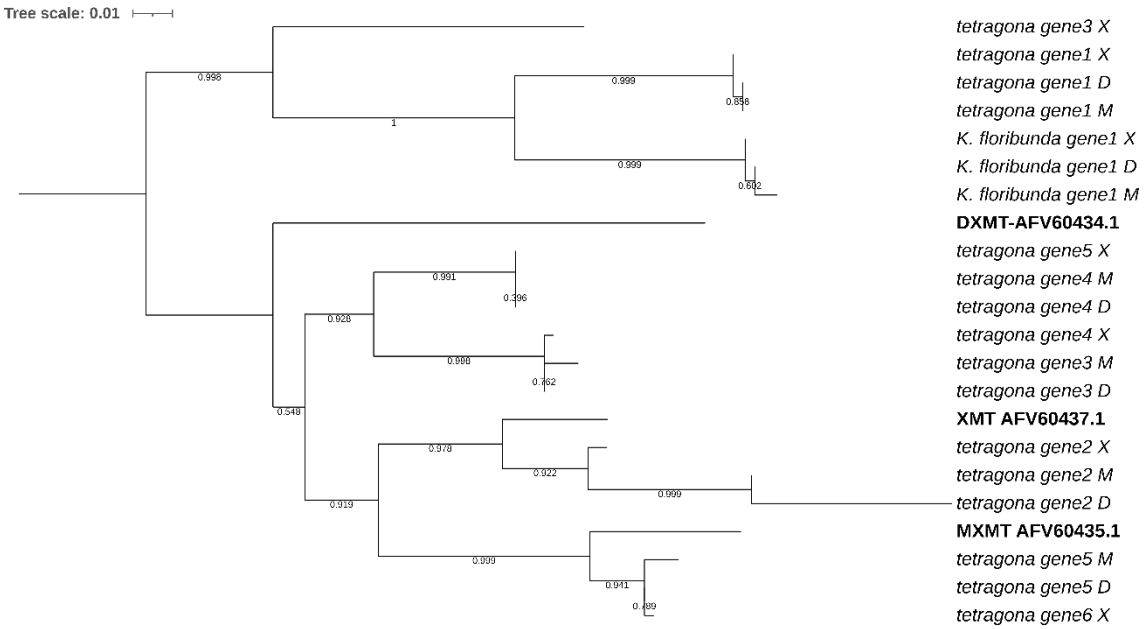
Anexo 4. Árbol filogenético de genes NMT de C. salvatrix. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



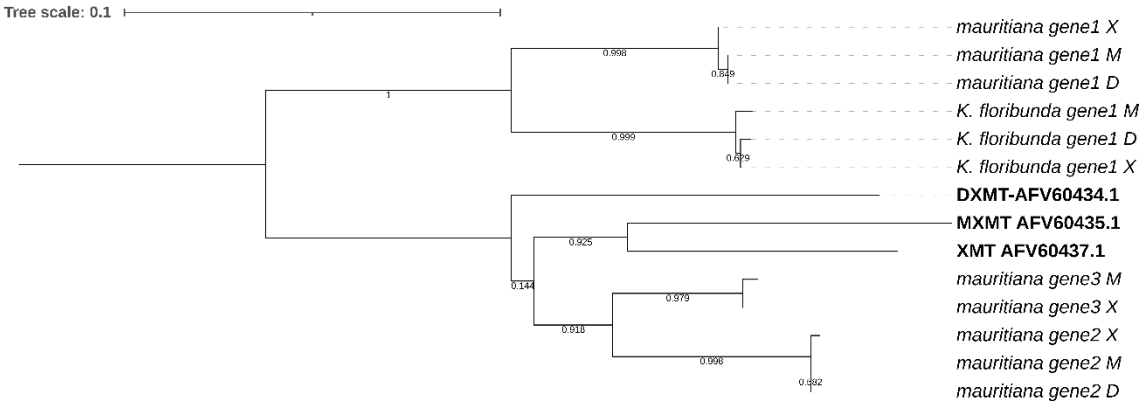
Anexo 5. Árbol filogenético de genes NMT de C. farafanganensis. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



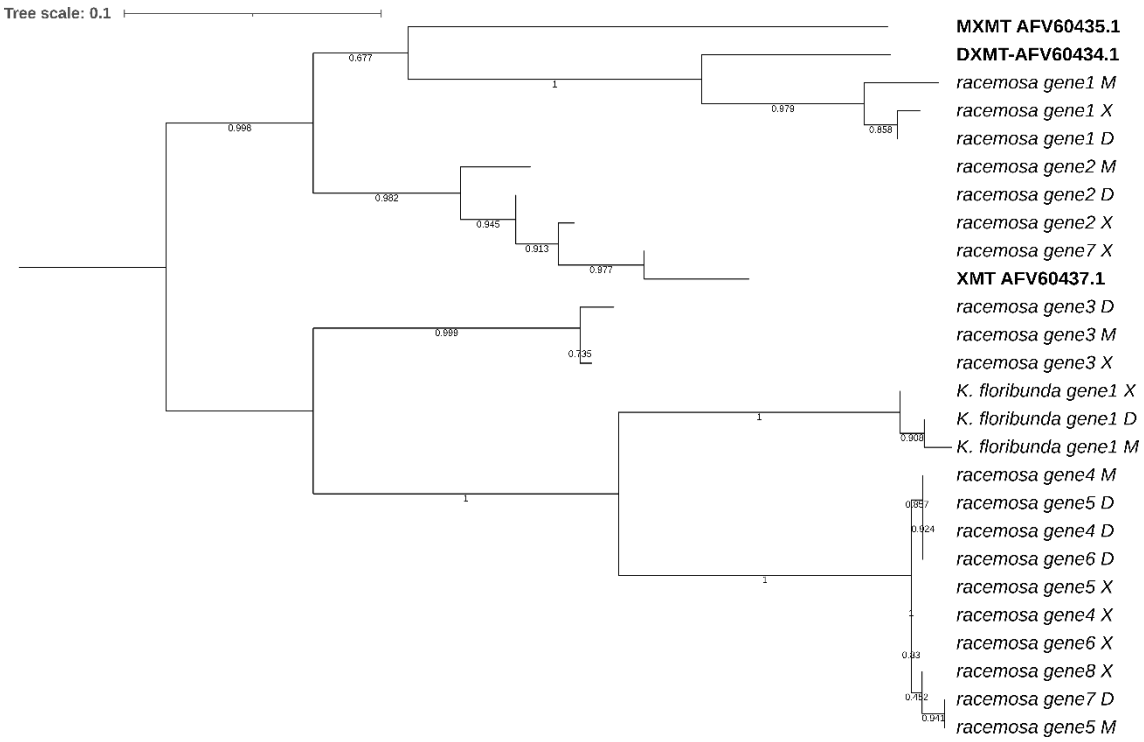
Anexo 6. Árbol filogenético de genes NMT de C. charrieriana. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



Anexo 7. Árbol filogenético de genes NMT de C. tetragona. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).

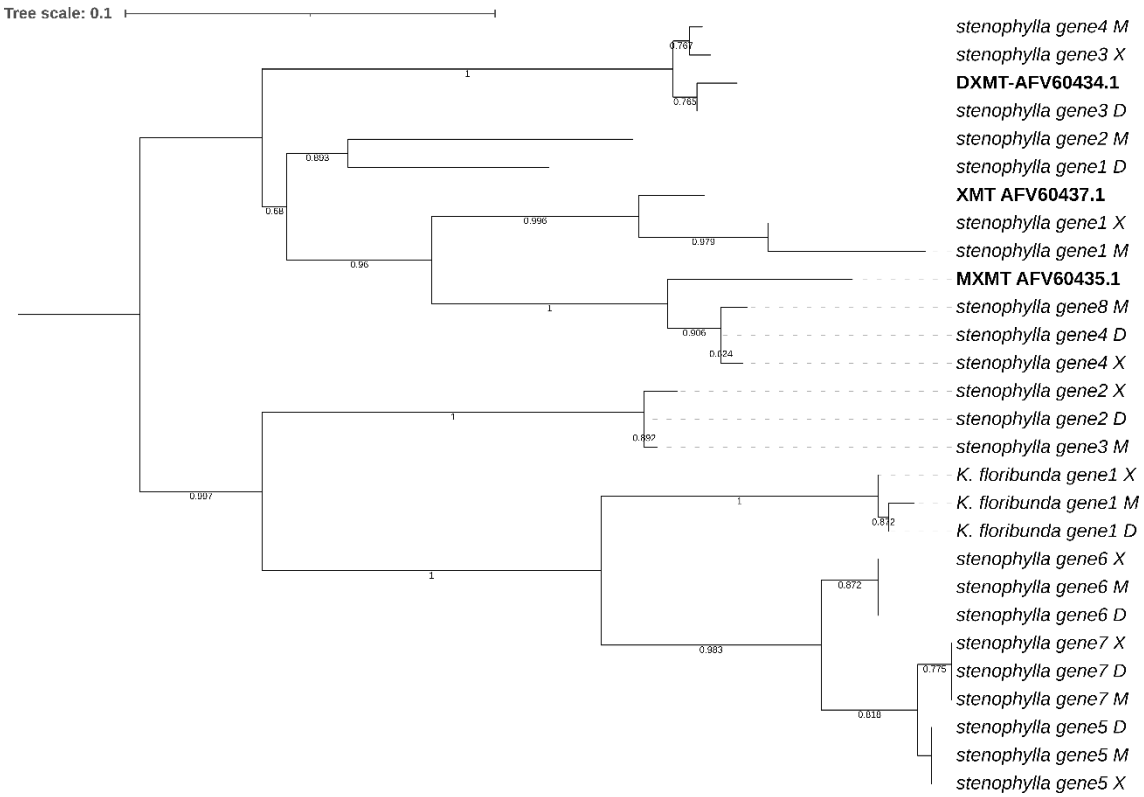


Anexo 8. Árbol filogenético de genes NMT de C. mauritiana. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).

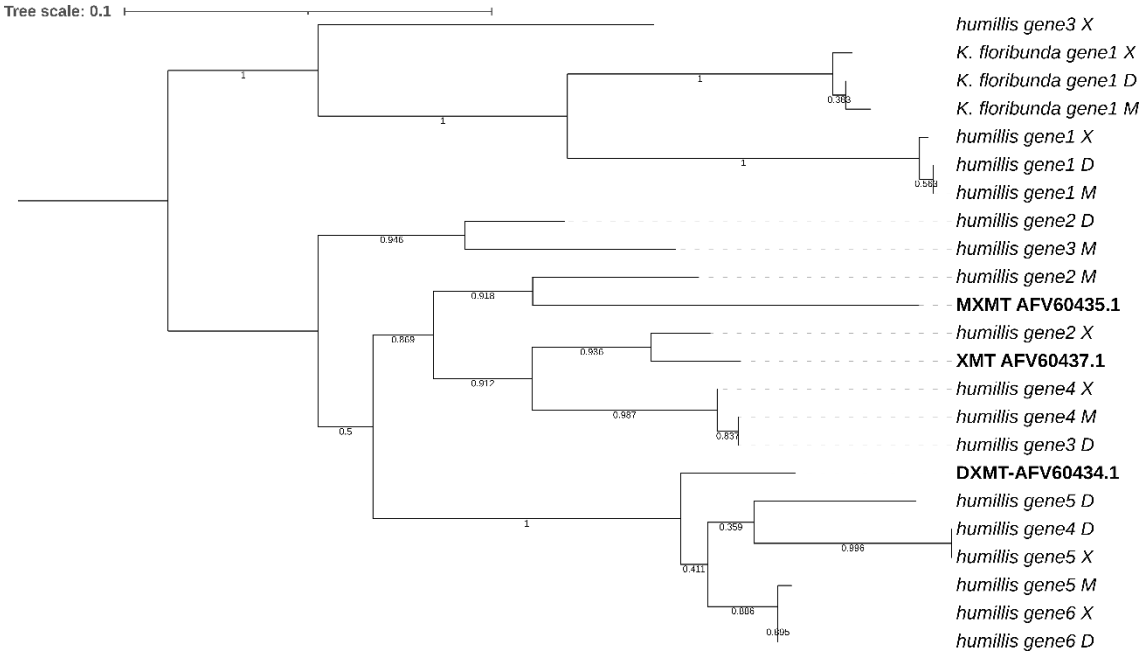


Anexo 9. Árbol filogenético de genes NMT de C. racemosa. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).

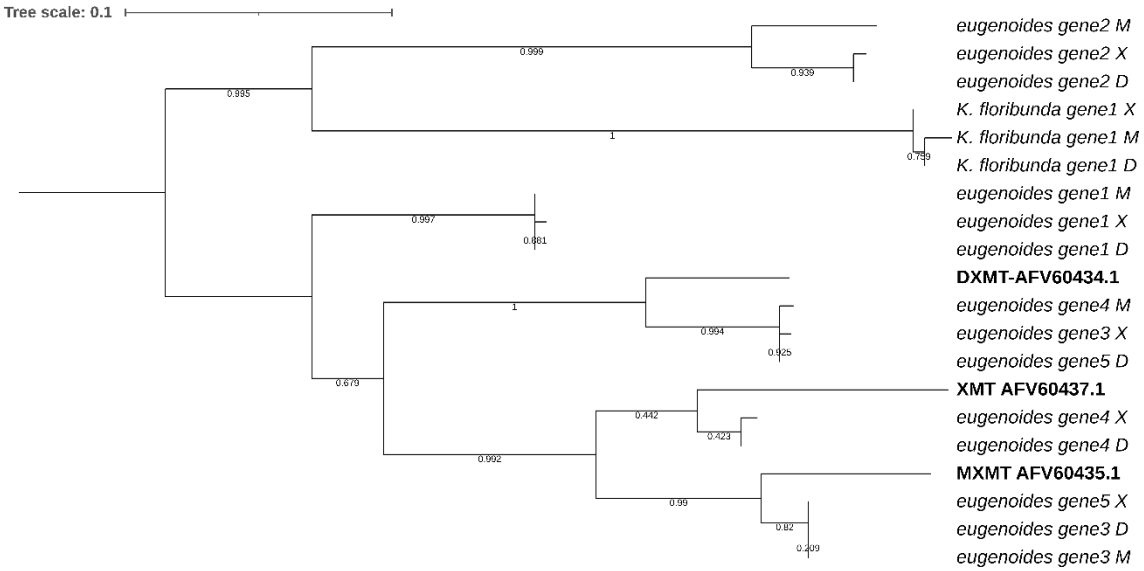
Árbol no es tan bueno no se puede interpretar, problema durante el ensamblaje del genoma, baja cobertura, calidad baja del ensamblaje, en lancifolia es raro porque no tiene xmt, ver calidad del ensamblaje, muy raro aquí en racemosa, ver si no hay otro nmt que tiene la función del que falta, DXMT puede tener poca actividad de teobromin sintasa, sil falta el 2do gen el 3ro puede reemplazarle, pero no de manera eficaz



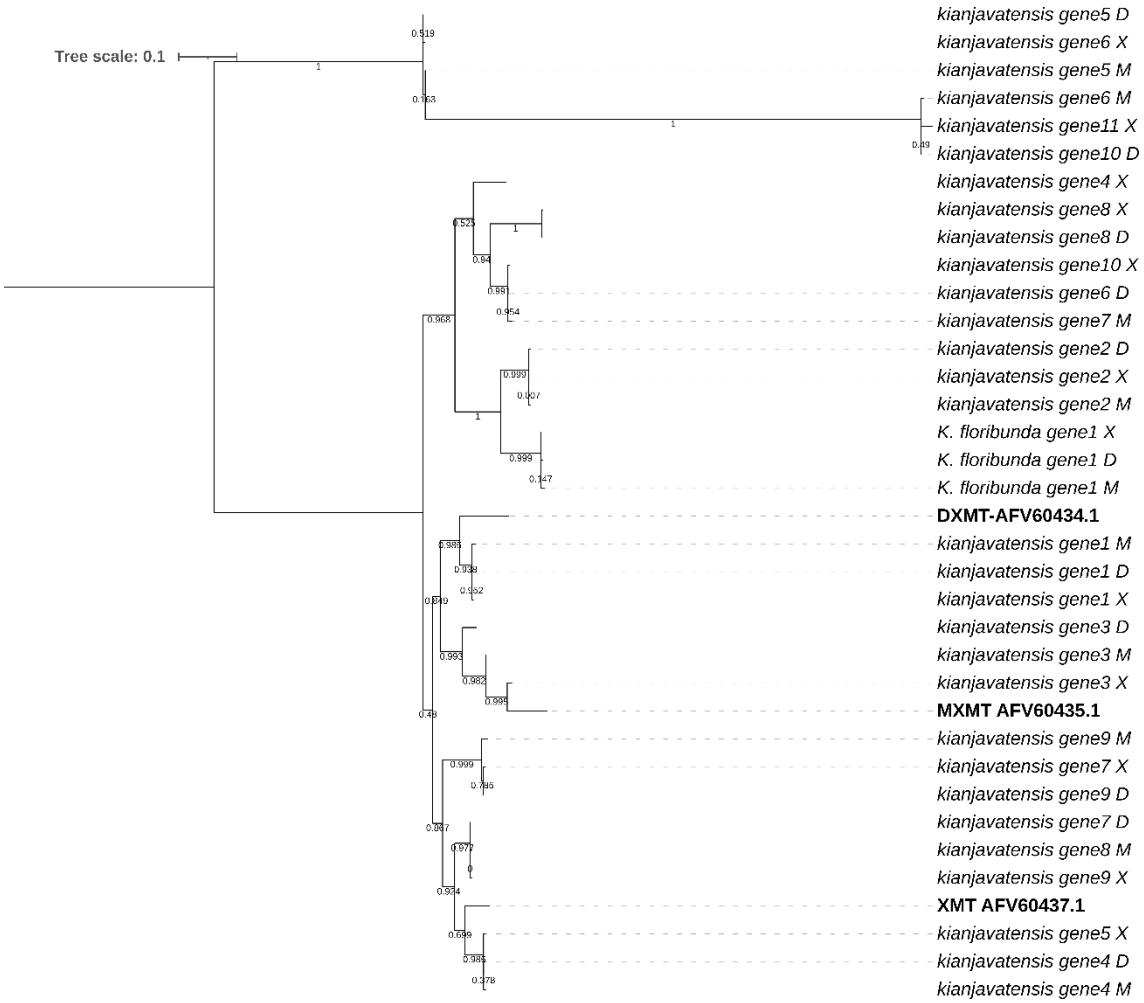
Anexo 10. Árbol filogenético de genes NMT de C. stenophylla. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



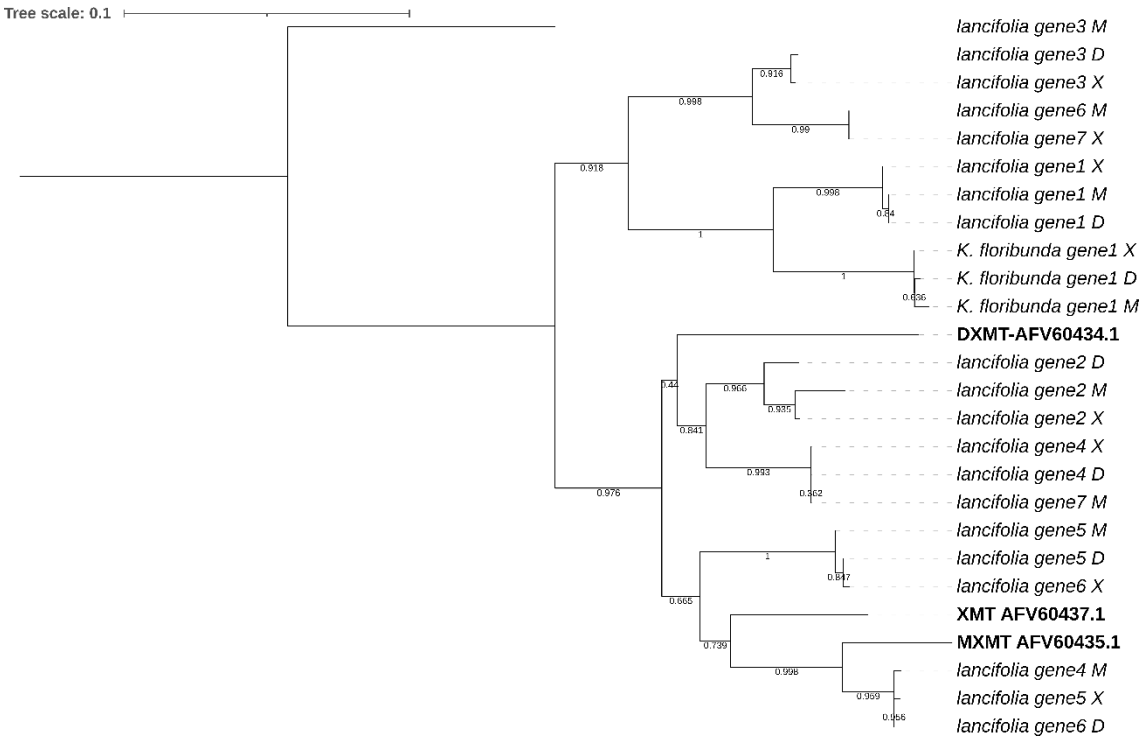
Anexo 11. Árbol filogenético de genes NMT de C. humillis. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



Anexo 12. Árbol filogenético de genes NMT de C. eugenoides. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



Anexo 13. Árbol filogenético de genes NMT de C. kianjavatensis. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).



Anexo 14. Árbol filogenético de genes NMT de C. lancifolia. DXMT-AFV60434.1, MXMT-AFV60435.1 y XMT-AF60437.1 son los genes de referencia en *C. canephora* (Perrois et al., 2015).