UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Agronòmica i del Medi Natural

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural

Re-anotación del genoma de la Petunia axillaris

Trabajo Fin de Grado

Grado en Biotecnología

AUTOR/A: Gadea Martínez, María

Tutor/a: Forment Millet, José Javier

Director/a Experimental: Bombarely Gomez, Aureliano

CURSO ACADÉMICO: 2023/2024

Re-anotación del genoma de la *Petunia axillaris*

Universitat Politècnica de València

Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural (ETSIAMN)

Trabajo de Fin de Grado

Grado en Biotecnología

Curso académico 2023/2024

Autora: Gadea Martínez, María

Tutor: Forment Millet, José

Cotutor: Bombarely Gomez, Aureliano

VALENCIA, julio de 2024

RESUMEN

La anotación estructural de un genoma es el proceso de identificar los elementos genómicos, tales como genes y elementos transponibles, presentes en su secuencia, mientras que la anotación funcional es el proceso por el cual se asigna una función biológica y/o molecular a un gen basado en la homología de secuencia de la proteína que codifica con proteínas de función conocida. Este trabajo consiste en la re-anotación del genoma de la especie vegetal *Petunia axillaris*, tanto estructural como funcionalmente usando datos públicos, así como en la evaluación de la calidad de esta anotación.

El proceso de anotación de los modelos genéticos se realiza mediante dos enfoques: mediante el alineamiento de una secuencia proveniente de un experimento (evidencia experimental) y mediante el uso de modelos de predicción *ab-initio*. Para el primer caso, se alinea una secuencia de una proteína y/o un transcriptoma con el genoma de interés; y en el segundo caso, se escanea la secuencia del genoma buscando patrones asociados a secuencias codificantes a través de modelos ocultos de Markov (HMM) que han sido previamente entrenados con datos experimentales. Una vez los modelos génicos se han obtenido, estos se evaluaron mediante tres metodologías: 1- métricas sobre su longitud, número de exones por gen y número de genes totales; 2- métricas sobre el porcentaje de genes conservados que han sido anotados con la herramienta BUSCO; 3- métricas sobre la distancia de cada modelo génico a su evidencia a través del parámetro AED. En este trabajo se han considerado diferentes tipos de evidencias (proteínas de distintas especies y datos transcriptómicos de distintos tejidos), así como programas de predicción como es BRAKER3.

Como resultado de este trabajo se identificaron 1.562.383 elementos repetitivos (67,58 % del genoma), siendo su mayoría elementos transponibles de tipo LTR los cuales representaban el 35, 21% del genoma. Más de 30,000 modelos génicos fueron identificados con cada una de las metodologías, aunque fue el uso del programa BRAKER3 el que produjo los mejores resultados, habiendo identificado 34.402 genes. Estos resultados también demuestran que el uso de una combinación de distintos sets de datos de proteínas de genomas bien anotados como *Arabidopsis* y tomate, junto a datos transcriptómicos de una gran variedad de tejidos y un filtrado de las predicciones, es esencial para producir una anotación de calidad.

Palabras clave: Anotación; Ensamblaje; Genoma; Modelos génicos; *Petunia axillaris*

Alumna: María Gadea Martínez

Tutor: José Forment Millet

Cotutor: Aureliano Bombarely Gomez

Valencia, mes de julio 2024

SUMMARY

The structural annotation of a genome is the process of identifying genomic elements, such as genes and transposable elements, present in its sequence. Functional annotation, on the other hand, is the process by which a biological and/or molecular function is assigned to a gene based on the sequence homology of the protein it encodes with proteins of known function. This work involves the re-annotation of the genome of the plant species *Petunia axillaris*, both structurally and functionally, using public data, as well as the evaluation of the quality of this annotation.

The annotation process of genetic models is carried out using two approaches: alignment of a sequence from an experiment (experimental evidence) and the use of ab-initio prediction models. For the first case, a sequence of a protein and/or a transcriptome is aligned with the genome of interest; in the second case, the genome sequence is scanned for patterns associated with coding sequences using Hidden Markov Models (HMM) that have been previously trained with experimental data. Once the gene models have been obtained, they were evaluated using three methodologies: 1- metrics on their length, number of exons per gene, and total number of genes; 2- metrics on the percentage of conserved genes that have been annotated with the BUSCO tool; 3- metrics on the distance of each gene model to its evidence through the AED parameter. This work considered different types of evidence (proteins from various species and transcriptomic data from different tissues), as well as prediction programs such as BRAKER3.

As a result of this work, 1,562,383 repetitive elements (67.58% of the genome) were identified, the majority being LTR-type transposable elements, which represented 35.21% of the genome. More than 30,000 gene models were identified with each of the methodologies, although the use of the BRAKER3 program produced the best results, identifying 34,402 genes. These results also demonstrate that the use of a combination of different sets of protein data from well-annotated genomes such as Arabidopsis and tomato, along with transcriptomic data from a wide variety of tissues and filtering of predictions, is essential to produce a quality annotation.

Keywords: Annotation; Assembly; Gene models; Genome; *Petunia axillaris*

Student: María Gadea Martínez

Tutor: José Forment Millet

Co-tutor: Aureliano Bombarely Gomez

Valencia, month of July 2024

Agradecimientos

En primer lugar quiero agradecer a mi tutor Aureliano por haberme ayudado y enseñado tanto durante la realización de este trabajo, y a todos mis compañeros: Víctor, Carmen, Olivia, Paulo, Coco, Miguel, Agustín, Alberto…, por haber hecho mi día a día en el laboratorio tan entretenido y agradable.

También agradecer a mis amigos residencia y a los cuencheadores por haberme acompañado durante estos cuatro años; sobre todo a mis compañeras de piso Ángela, Marta, Luz y Olga.

Por último quiero agradecer a mi padre, a mi hermano y al resto de mi familia por siempre haber confiado en mí y haberme brindado las oportunidades para llegar hasta aquí.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Nomenclaturas y abreviaturas

AED- *Annotation Edit Distance* (Distancia de edición de anotación)

BUSCO- *Benchmarking Universal Single-Copy Ortholog* (Evaluación comparativa del ortólogo universal de copia única)

CDS- *Coding sequence* (Región de codificación)

EST- *Expressed sequence tag* (Marcadores de secuencia expresada)

ER- Elemento repetitivo

ET- Elemento transponible

GeMoMa- *Gene Model Mapper (*Mapeador de modelos genéticos)

Hisat2 - *Hierarchical indexing for spliced alignment of transcripts* (Indexación jerárquica para alineación empalmada de transcripciones)

HMM – *Hidden Markov Models* (Modelos Ocultos de Markov)

LCR- *Low Copy Repeats* (Repeticiones de bajas copias)

LINE- *Long interspersed nuclear elements* (Elemento nuclear intercalado largo)

LTR- *Long Terminal Repeat* (Repetición terminal larga)

MMP- *Maximal Mapeable Prefix* (Prefijo Máximo Mapeable)

NCBI- *National Center for Biotechnology Information* (Centro Nacional de Información Biotecnológica)

NGS- *Next Generation Sequencing* (Secuenciación de próxima generación)

ORF- *Open Reading Frame* (Marco abierto de lectura)

PSSM- *Position-specific scoring matrix* (Matriz de puntuación específica de la posición)

SINE- *Short interspersed nuclear elements* (Elemento nuclear intercalado corto)

SNAP- *Semi-HMM-based Nucleic Acid Parser* (Analizador de ácidos nucleicos basado en semi-HMM)

SNP- *Single Nucleotide Polymorphism* (Polimorfismo de único nucleótido)

SRA- *Short Reads Archive* (Archivo de lecturas cortas)

STAR- *Spliced Transcripts Alignment to a Reference* (Alineación de transcripciones empalmadas con una referencia)

WGS- *Whole Genome Sequencing* (Secuenciación del genoma completo)

1. INTRODUCCIÓN

1.1. Secuenciación y Ensamblaje de Genomas

Los genomas son el conjunto de información genética asociada a la secuencia completa de ADN de un organismo vivo y representan el punto de partida para estudios genéticos y de biología molecular. Para ello, desde el descubrimiento del ADN y su estructura, se ha dedicado grandes esfuerzos en determinar su secuencia de una manera precisa y fiable, para así mejorar el entendimiento de la estructura del genoma y conocer el rol de los genes presentes en este. Así, las primeras moléculas de ADN provenientes de un fago λ fueron secuenciadas en 1968 (Wu y Kaiser, 1968); mientras que en 1976, Fiers secuenció de manera completa por primera vez un genoma, perteneciente al bacteriófago de ARN MS2 (Minjou *et al*.,1972; Giani *et al*., 2019). Más tarde, en 1977, Sanger y Gilbert secuenciaron el genoma de ADN del bacteriófago φX174 (Sanger *et al*.,1977), y en 1982 se publicó la secuenciación del genoma completo de uno de los modelos más importantes en la biología molecular, el bacteriófago λ (Sanger *et al*., 1982; Koonin *et al*., 2003).

No obstante, para poder determinar las secuencias genómicas de organismos de mayor tamaño y más complejos se comenzaron a desarrollar tecnologías más eficientes y veloces. Estos secuenciadores son capaces de generar lecturas definidas por su longitud. El ensamblaje del genoma se basa en la búsqueda de lecturas solapantes y la determinación de la secuencia consenso. Una vez ensamblado, el genoma se anota estructural y funcionalmente, para así poder realizar posteriores estudios como análisis funcionales, estudios evolutivos o medicina de precisión entre otros. Debido a todas estas aplicaciones, es fundamental que las secuencias obtenidas sean lo más completas y libres de errores posible (Giani *et al*., 2019).

A mitades del siglo XX, las tecnologías de secuenciación disponibles permitían sólo analizar genomas relativamente pequeños, debido a los altos costes de los reactivos, la complejidad de las técnicas y los largos tiempos de realización. No obstante, desde el nuevo milenio se han desarrollado plataformas novedosas, conocidas como secuenciación de próxima generación (NGS), capaces de abordar genomas más grandes, en un proceso llamado secuenciación del genoma completo (WGS). Esta nueva tecnología ha probado reducir el coste de secuenciación y presentar mejor eficiencia, permitiendo generar una amplia gama de genomas secuenciados de especies modelo eucariotas. Las tecnologías NGS también permitieron el desarrollo de aplicaciones extremadamente novedosas, incluida la secuenciación del exoma completo, la secuenciación de ARN de alto rendimiento (ARN-seq), y la determinación del paisaje epigenético de todo el genoma mediante la inmunoprecipitación de cromatina acoplada a secuenciación (ChIP-seq). En cuanto al genotipado humano, fue posible la identificación de millones de polimorfismos de un solo nucleótido (SNP) por proyectos como el Proyecto Internacional HapMap (Giani *et al*., 2019; McCombie *et al*., 2019).

Con el desarrollo de las últimas plataformas de secuenciación de tercera generación se planea revolucionar el campo de la genómica, generando genomas de incluso mayor calidad. Desde un punto de vista técnico, la nueva generación de enfoques de secuenciación fue el resultado de una combinación de avances en microfabricación, imágenes de alta resolución y potencia computacional (Giani *et al*., 2019). Estas últimas plataformas permiten ensamblajes de genoma independientes de referencia y generación de haplotipos de largo alcance. La secuenciación rápida de ADN y ARN es actualmente una práctica habitual y seguirá teniendo un impacto cada vez mayor en la biología y la medicina (McCombie *et al*., 2019).

Actualmente, las tecnologías de secuenciación y ensamblaje de genomas han logrado generar genomas sin huecos (*gapless genomes*), denominados '*telomere-to-telomere'* o T2T. Buenos ejemplos de estos genomas T2T son el genoma de humano y el de maíz (Nurk et al., 2022; Chen et al., 2023).

1.2. Anotación de Genomas

La anotación de genomas es el proceso de identificar elementos genómicos dentro de una secuencia de ADN, como genes, exones, intrones, elementos transponibles o regiones reguladoras. Los elementos genómicos más destacados son los modelos génicos, los cuales son regiones del genoma que se consideran transcribibles a ARN. Este ARN, a su vez, puede traducirse en proteínas o pertenecer a una de varias categorías definidas de genes de ARN no codificantes. (Schnable J.C., 2020).

Su proceso de anotación se divide en dos metodologías principales: *Ab-initio* y basado en evidencias. En el primer caso, un programa predictor, escanea toda la secuencia del genoma buscando patrones asociados a secuencias codificantes a través de Modelos Ocultos de Markov (HMM). Los HMM son un tipo de modelo estadístico que se utiliza para representar secuencias de eventos observados que están influenciados por factores internos, que no son directamente observables. En este modelo, los eventos observables denominan símbolos y los factores de influencia ocultos se denominan estados. Un HMM se compone de dos procesos aleatorios interrelacionados: uno es un proceso invisible de estados ocultos y otro es un proceso visible de símbolos observables. Los estados ocultos siguen una cadena de Markov, es decir, cada estado depende sólo del anterior; y la probabilidad de observar un símbolo particular está determinada por el estado oculto actual (Yoon, 2019). En el segundo caso, una evidencia como una secuencia de un transcrito o secuencia proteica se alinea con la secuencia del genoma para identificar una secuencia genética. Aunque la segunda metodología es más precisa, pueden existir limitaciones a la hora obtener evidencias de todos los genes de un genoma completo. Además, la alineación de secuencias puede ser susceptible de errores en genomas y familias de genes complejos. Las herramientas modernas combinan ambos enfoques para mejorar la anotación de los modelos genéticos (Armstrong *et al*., 2019).

1.2.1. Elementos repetitivos

Los genomas de las plantas difieren considerablemente en sus tamaños. No obstante, estas diferencias se deben predominantemente al número de secuencias repetidas en lugar de a la cantidad de genes. Por lo tanto, el análisis y estudio detallado del paisaje repetitivo de un genoma es crucial para una comprensión profunda del genoma en cuestión (Kumekawa *et al*., 1999). Los elementos repetitivos son secuencias de ADN que se repiten múltiples veces a lo largo del genoma. Según su distribución ilustrada en la Figura 1, se encuentran dos grandes grupos: las repeticiones en tándem y las repeticiones dispersas. Dentro del primer grupo se encuentran el rARN, los genes en tándem parálogos y los ADN satélite, encontrado en la región de los telómeros y los centrómeros. Por otro lado, dentro de las repeticiones dispersas se encuentran los genes parálogos, el tADN, los motivos promotores y los elementos transponibles (Richard *et al*., 2018).

Los elementos transponibles o ETs son secuencias de ADN capaces de cambiar su posición dentro del genoma (Bourque *et al*., 2018), que desempeñan papeles importantes en la evolución de los organismos y la dinámica genómica, y su importancia en el proceso evolutivo se puede atribuir a varias propiedades clave. Según los mecanismos utilizados para su transposición, se clasifican en dos grandes grupos: los de clase I y los de clase II.

Dentro de los ETs de clase I encontramos los retrotransposones, los cuales son transposones de ARN capaces de moverse mediante un mecanismo de "copia y pega". Asimismo, estos se encuentran divididos en diferentes subclases, que incluyen los *Long Terminal Repeats* (LTR) y los retrotransposones no LTR, que incluyen, entre otro, a los elementos nucleares intercalados largos (LINEs) y los elementos nucleares intercalados cortos (SINEs) (Han *et al*., 2010). Los retrotransposones LTR son capaces de integrarse en el genoma mediante un proceso similar al utilizado por los retrovirus, basado en una rotura catalizada por una enzima y una transferencia de hebra (Hughes *et al*., 2015).

Además, en cada subclase los elementos repetitivos se encuentran divididos en diferentes superfamilias. Así, dentro de las superfamilias de LTR encontramos dos que se encuentran en la mayoría de los linajes de organismos eucariotas: Copia y Gypsy (Paz *et al*., 2017). Por otro lado, los

ETs de clase II se mueven por un mecanismo de "corta y pega" mediante un intermediario de ADN, sin necesidad de ARN (Muñoz-López y García-Pérez, 2010).

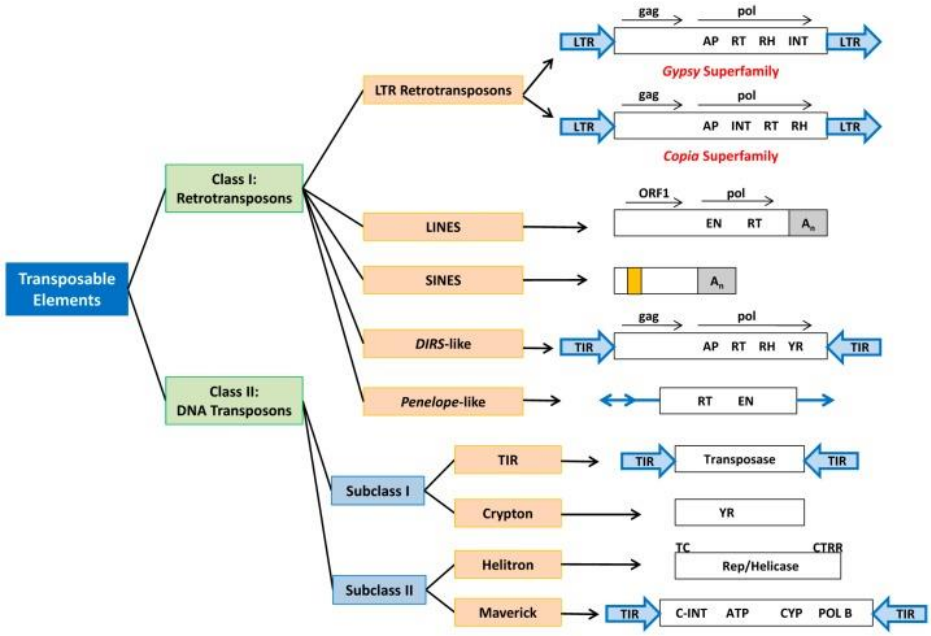


Figura 1. Esquema y representación visual de la organización y clasificación de los elementos repetitivos (Mat Razali *et al*., 2019).

Los genomas del género *Petunia* están conformados en aproximadamente un 60% por ADN repetitivo, con un elevado número de secuencias de copias bajas (LCR) y transposones de ADN (Alisawi *et al*., 2023). Con el fin de anotar y enmascarar estos elementos repetitivos, se han de identificar mediante búsqueda de homología de secuencia con una colección de elementos conocidos y/o el análisis del número de apariciones de una secuencia dentro de una secuencia del genoma.

Existen dos tipos de aproximaciones para identificar elementos repetitivos. En la primera, la secuencia de ADN es analizada para buscar patrones repetitivos. Este tipo de metodologías se considera *de-novo*, y puede realizarse mediante K-meros, unas subsecuencias de una longitud fija contenidas dentro de una secuencia biológica. Un buen ejemplo de este tipo de programas es RepeatScout (https://github.com/mmcco/RepeatScout). En la segunda estrategia, se realiza una búsqueda por homología de secuencias en una base de datos de elementos conocidos, siendo un ejemplo de este tipo de herramientas RepeatMasker (Smit et al., 2013).

Uno de los programas utilizados en la primera aproximación mencionada es RepeatModeler2, el cual es un paquete de modelado e identificación de familias de elementos transponibles que se utiliza para encontrar elementos repetitivos desconocidos e identificar el LTR. Al tratarse de una herramienta que utiliza un enfoque *de novo*, por lo que es capaz de encontrar motivos repetitivos sin necesidad de librerías pre-existentes. Para ello, incluye tres programas de búsqueda de repeticiones (RECON, RepeatScout y LtrHarvest/Ltr_retriever) que utilizan métodos computacionales complementarios para identificar los límites de los elementos repetitivos y sus relaciones familiares a partir de los datos de la secuencia. RepeatModeler facilita la automatización de las ejecuciones de los diferentes algoritmos, dada una base de datos genómica. Este programa agrupa resultados redundantes, refina y clasifica las familias, y genera una biblioteca de familias de ET de alta calidad, adecuada para su uso posterior con RepeatMasker y, finalmente, para su envío a la base de datos Dfam (Flynn *et al*., 2020).

Con el fin de enriquecer la anotación *de novo* del repetitivo realizada por RepearModeler2, se utiliza el programa RepeatMasker. Este es un programa que analiza secuencias de ADN en busca de repeticiones intercaladas y secuencias de ADN de baja complejidad, que se utiliza para anotar los elementos en la secuencia FASTA del genoma utilizando la base de datos DFAM y la librería de elementos repetitivos encontrados por RepeatModeler2. El resultado del programa es una anotación detallada de las repeticiones que están presentes en la secuencia de interés, así como una versión modificada en la que se han anotado todas las repeticiones encontradas. Las comparaciones de secuencias en RepeatMasker se realizan mediante el programa cross_match, una implementación del algoritmo Smith-Waterman-Gotoh o por WU-Blast (Gish, W. (1996-2003) http://blast.wustl.edu).
Una vez realizada la anotación del repetitivo con RepeatModeler y RepeatMasker, TEsorter se utiliza para clasificar más refinadamente los elementos repetitivos individuales en superfamilias, permitiendo una visión más categorizada del paisaje repetitivo. Este programa es capaz de clasificar retrotransposones de repetición terminal larga (LTR-RT) y cualquier otro ET, incluidos los elementos de Clase I y Clase II (Zhang *et al*., 2022).

1.2.2.Modelos de genes basados en evidencias experimentales

La anotación basada en evidencia se refiere al uso de datos experimentales, como secuencias de ADNc o proteínas conocidas, para identificar y anotar genes en un genoma. Una de las metodologías más usadas consiste en el uso de datos experimentales transcriptómicos de ARN-Seq. En un experimento de este tipo se generan miles de millones de lecturas cortas (generalmente mediante la tecnología de secuenciación de Illumina), en la que cada una de ellas representa un fragmento de un transcrito. Al alinear estas lecturas con un genoma de referencia, es posible reconstruir la estructura de los genes, así como cuantificar su expresión en función del número de fragmentos secuenciados. No obstante, primero se ha de filtrar las lecturas obtenidas del ARN-Seq por su calidad y longitud, y eliminar los adaptadores utilizados para la secuenciación por Illumina mediante la herramienta Fastmcf (Aronesty, 2013).

Una vez filtradas, las lecturas se mapean al genoma de referencia. Este paso se puede realizar con diferentes programas de alineamiento, siendo tanto Hisat2 como STAR útiles.

HISAT2 (*Hierarchical indexing for spliced alignment of transcripts*) es un programa de alineamiento rápido y sensible para el mapeado de lecturas tanto de ADN como de ARN a un genoma. Se basa en el uso de dos tipos de índices FM obtenidos a partir de la transformación de Burrows–Wheeler (BWT), uno global el cual representa el genoma y un grupo de pequeños índices FM solapantes para regiones que de manera colectiva cubren el genoma entero. Estos últimos índices locales facilitan alinear las lecturas secuenciadas en el genoma cubierto por los dos tipos de índices de manera más rápida y efectiva (Keel y Snelling, 2018). Este programa es particularmente útil para estudios con límite reducido de memoria o en los que la variabilidad genética sea un factor clave (Kim *et al*., 2015).

Otro programa ampliamente utilizado para el mapeado de lecturas a un genoma es STAR (*Spliced Transcripts Alignment to a Reference*), puesto que realiza un análisis sensible a la identificación de los sitios de empalme. STAR es un alineador que consta de dos fases principales: un primer paso de 'búsqueda de semillas (*seeds*)' y otro paso de 'unión, agrupamiento y puntuación (*clustering/stitching/scoring*). Para obtener las semillas, STAR realiza una búsqueda secuencial de un Prefijo Máximo Mapeable (MMP), el cual se define como la subsecuencia de la lectura más larga encontrada que coincide de manera exacta con una o más subsecuencias del genoma de referencia (Figura 2). Así, las semillas son las diferentes partes de la lectura que se asignan por separado (Dobin *et al*., 2013).

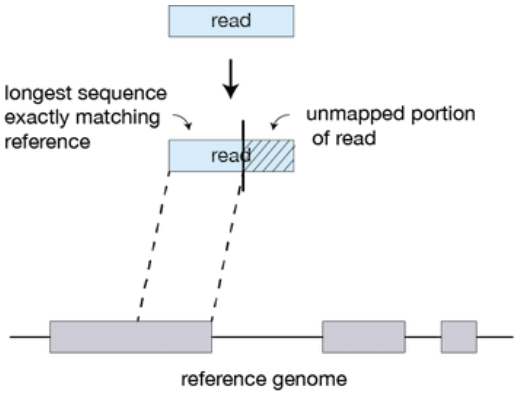


Figura 2. Representación esquemática de la búsqueda del Prefijo Máximo Mapeable en el algoritmo STAR (Dobin *et al*., 2013).

Una vez se han mapeado las lecturas con STAR o HISAT2, se procede a generar modelos de transcritos a partir de los alineamientos de ARN-Seq. Para ello se hace uso de programas de ensamblaje como StringTie, el cual es un eficiente y rápido programa capaz de generar posibles transcritos. Utiliza un método computacional que aplica un algoritmo de flujo en red además de un paso opcional de ensamblaje *de novo* para ensamblar y cuantificar la información en transcritos completos que representan múltiples variantes de empalme/splicing para cada locus genético (Pertea *et al*., 2015). Esta herramienta es capaz de producir reconstrucciones de genes y estimaciones de niveles de expresión más precisos que otros utilizados programas de ensamblaje de transcritos como Cufflinks o Scripture.

Además, con el fin de obtener las posibles regiones codificantes de genes (CDS) y secuencias de proteínas a partir de los transcritos, es posible combinar StringTie con el programa Transdecoder (https://github.com/TransDecoder/TransDecoder). Este programa identifica las CDS basándose en varios factores como son la longitud mínima del marco abierto de lectura (ORF), composición de las secuencias, identificación de codones de inicio y codones de parada. En el caso de que un único transcrito genere más de una ORF, el programa crea y entrena una matriz de pesos posicionales (PPSM), la cual es una forma de representación y predicción de motivos o patrones en secuencias biológicas que tiene como objetivo describir las variaciones intrínsecas en sus patrones, para la predicción refinada del codón de inicio.

Las herramientas de predicción de modelos de genes basados en evidencias de secuencias de proteínas funcionan de una manera parecida. Una secuencia de proteína es alineada con un genoma de referencia, y los datos del alineamiento son procesados para generar un modelo de intrones y exones, con la diferencia de que los exones son equivalentes a las regiones codificantes. Existen varios programas para modelizar estructura de genes basados en proteínas como Exonerate (Slater y Birney, 2005) o Spaln (https://github.com/ogotoh/spaln). Uno de los más usados en los últimos cinco años es GeMoMa (Gene Model Mapper), un programa de búsqueda de homología entre genes con distintos tipos de evidencias, incluido proteínas. Para esto, extrae los modelos génicos, ARNms y proteínas de un genoma de una especia de referencia relacionada evolutivamente para predecir modelos génicos y sitios de corte en el genoma objetivo. De esta manera, GeMoMa se basa en la conservación de las posiciones de intrones, secuencia de aminoácidos y del ARN (Keilwagen et al., 2019).

1.2.3.*Ab-initio*

La anotación *ab-initio* se basa en algoritmos computacionales entrenados que predicen genes y otras características genómicas únicamente a partir de la secuencia de ADN, sin información experimental previa. Los programas *ab-initio* usan generalmente modelos ocultos de Markov (HMM) (ver

apartado 1.2. para más detalles), aunque en los últimos cinco años se han empezado a usar modelos de aprendizaje profundo. Helixer (Stiehler *et al*., 2023) es un buen ejemplo de este otro tipo de herramientas.

Uno de los programas basados en HMM más utilizados para la anotación *ab-initio* es AUGUSTUS (Stanke y Morgenstern, 2005), un software capaz de predecir las distribuciones de probabilidad para las diferentes secuencias del genoma eucariota basándose en los modelos ocultos de Markov generalizados (GHMM). Para ello se modela probabilísticamente la secuencia alrededor de los sitios de empalme, la secuencia de la región del punto de ramificación, las bases antes del inicio de la traducción, las regiones codificantes y no codificantes, incluyendo 5′UTR y 3′UTR, las primeras bases codificantes de un gen, la distribución de longitud de exones individuales, inicial exones, exones internos, exones terminales, regiones intergénicas, la distribución del número de exones por gen y la distribución de longitud de los intrones (Ejigu *et al*., 2020). AUGUSTUS también puede incorporar información sobre la estructura genética provenientes de fuentes extrínsecas como marcadores de secuencia expresada (EST), MS/MS, alineamientos de proteínas y alineamientos genómicos sinténicos.

1.2.4.*Pipelines*

Los *pipelines* son flujos de trabajo automatizados que integran múltiples herramientas y métodos en un orden predefinido para realizar la anotación de genomas de manera eficiente. Estos son capaces de procesar cantidades masivas de datos de secuencias y los metadatos asociados utilizando múltiples componentes de software, bases de datos y entornos. Para ello, se entrenan utilizando la información relevante proporcionada por predictores genéticos *ab initio* y basados en similitudes, como los anteriormente mencionados AUGUSTUS y SNAP (Ejigu y Jung, 2020). Las *´pipelines´* de anotación publicas más usadas son MAKER (Holt y Yandell, 2011), BRAKER (Brůna *et al.,* 2021) y EVidenceModeler (Haas *et al.,* 2008). Bases de datos como NCBI y ENSEMBL tienen sus propias *´pipelines´* de anotación.

BRAKER es un sistema para la predicción totalmente automatizada, no supervisada y altamente precisa de estructuras genéticas codificantes de proteínas para genomas eucariotas que incluye los programas GeneMark-ETP y AUGUSTUS. Esta herramienta es capaz de realizar anotaciones en todo el genoma basándose en información de ARN-seq de manera no supervisada pero altamente precisa (Hoff *et al*., 2019). Para ello, combina GeneMark- ETP (Brůna *et al*., 2024), el cual genera predicciones *ab initio* de la estructura genética mediante un entrenamiento iterativo no supervisado de lecturas de secuencias de ARN sin ensamblar, con AUGUSTUS, que utiliza los genes predichos como un conjunto de entrenamiento, para predecir genes, utilizando información de lectura de ARN-seq sin ensamblar mapeada (Stanke *et al*., 2008; Stanke *et al*., 2006).

Una vez realizado la predicción mediante BRAKER, se añaden las UTRs a las anotaciones obtenidas. Esto se puede realizar por el programa GeMoMa anteriormente mencionado (siendo una parte experimental de BRAKER), o por Ingenannot. Este se trata de es un conjunto de utilidades y herramientas útiles para anotar las UTRs, inspeccionar y generar estadísticas de anotaciones genéticas, anotar isoformas y comparar anotaciones (https://forgemia.inra.fr/bioger/ingenannot).

1.2.5.Evaluación de la calidad de la anotación y análisis de resultados

Una vez completada la anotación estructural mediante diversas metodologías, se realiza un estudio utilizando distintas herramientas para evaluar la exactitud y calidad de la anotación realizada.

De esta manera, es posible evaluar la integridad de los espacios génicos estudiando el número de genes identificados en el genoma, además del ensamblaje del genoma y la anotación de proteínas mediante BUSCO (*Benchmarking Universal Single-Copy Ortholog).* Este programa es capaz de estimar la integridad y redundancia de los datos genómicos y proteicos procesados basándose en datos sobre el contenido genético de ortólogos de copia única universales (Manni *et al*., 2021). Además, los resultados se simplifican en categorías de BUSCO completo y de copia única, completo y duplicado, fragmentado o faltante, donde BUSCO es la abreviatura de "genes marcadores BUSCO".

Estos genes marcadores se seleccionan como aquellos que están presentes en al menos el 90% de las especies de un linaje determinado y presentes en una sola copia en el 90% de esas especies. No obstante, es esencial filtrar los datos de transcriptoma y proteínas antes de realizar la evaluación, pues las distintas isoformas podrían dar lugar a una alta proporción de duplicados (Seppey *et al*., 2019).

Por otra parte, la herramienta agat_sp_statistics.pl de AGAT Toolkit (Dainat J.) permite extraer métricas sobre el número de genes, exones, transcritos y la longitud de estos contextualizado en un marco filogenético respecto a una especie cercana.

Otra manera de evaluar la calidad de una anotación es medir cuanto se parece un modelo a datos experimentales. Para ello se usa la Distancia de Edición de Anotación (AED) es una métrica utilizada para evaluar la precisión de las anotaciones genómicas comparando la anotación de un gen contra su correspondiente secuencia de referencia. Esta puntuación mide la disimilitud entre la anotación predicha y la anotación de referencia basada en alineamientos de secuencias, considerando tanto la estructura del gen (como exones e intrones) como la precisión de las predicciones de inicio y fin de los genes. Un valor de AED cercano a 0 indica una alta concordancia entre la anotación y la referencia, reflejando una anotación precisa y confiable, mientras que un valor cercano a 1 sugiere discrepancias significativas, señalando posibles errores o áreas de mejora en la anotación del gen. (Holt y Yandell, 2011; Eilbeck *et al*., 2009).

Finalmente, otra aproximación es la de comparar los genes del genoma con genes de genomas de especies cercanas y/o especies modelos. Además, la herramienta en línea Orthovenn3 es capaz de correr análisis comparativos para la identificación y visualización de ortólogos y grupos de genes parálogos entre múltiples conjuntos de datos genómicos o proteómicos. Este programa funciona utilizando algoritmos de comparación de secuencias para identificar genes ortólogos y parálogos, y presenta los resultados en forma de diagramas de Venn y redes de ortología, lo que facilita la interpretación visual de las relaciones evolutivas y funcionales entre los genes de los diferentes organismos analizados. OrthoVenn3 también ofrece funcionalidades para el análisis de enriquecimiento de funciones y anotaciones, proporcionando una comprensión profunda de las funciones biológicas compartidas y específicas de los conjuntos de datos comparados (Sun *et al*., 2023).

1.2.6.Anotación funcional

Por último, se realiza la anotación funcional del genoma, la cual se basa en la asignación de información biológica a elementos genómicos, como son la función bioquímica, proceso biológico, regiones regulatorias y compartimento celular. Esta asignación de funciones se lleva a cabo mediante búsquedas de homología de secuencia con proteínas conocidas de bases de datos como NCBI NR y EMBL SwissProt. Actualmente, para realizar la búsqueda de homología de secuencia, se emplea el programa Diamond, que ha reemplazado el uso de BLASTX debido a su mejorada eficiencia y velocidad (Reeves *et al*., 2009).

Diamond es una herramienta de alineación de secuencias de proteínas con bases de datos conocidas para identificar similitudes y posibles funciones biológicas. El software utiliza algoritmos optimizados y estructuras de datos avanzadas para acelerar el proceso de comparación de secuencias, manteniendo una precisión alta en la identificación de homologías de proteínas. Su capacidad para manejar grandes conjuntos de datos lo hace ideal para su uso en análisis masivos, como proyectos de investigación genómica y metagenómica, donde los volúmenes de datos pueden ser extremadamente grandes y se requiere un procesamiento rápido para analizar y anotar secuencias de manera eficiente (Buchfink *et al*., 2021; Buchfink *et al*., 2015).

Una vez realizada la búsqueda de homología, el siguiente paso es añadir las funciones a las homologías encontradas mediante AHRD (*Automated Assignment of Human Readable Descriptions*), la cual asigna diferente peso según el nivel de curación de la base de datos. Esta herramienta es capaz de proporcionar descripciones más accesibles y comprensibles para las anotaciones de proteínas, facilitando la interpretación de los datos (https://github.com/groupschoof/AHRD).

1.3. *Petunia axillaris*

1.3.1. Taxonomía

La *Petunia axillaris* es una especie nativa de América del Sur perteneciente a la familia *Solenaceae*, la cual incluye a unas 3000-4000 especies entre las que destaca el tomate (*Solanum lycopersicum*), la patata *(Solanum tuberosum*), el tabaco (*Nicotiana tabacum*), el pimiento (*Capsicum annuum*) y la berenjena (*Solanum melongena*) entre otros (Gerats y Vandenbussche, 2005). Esta a su vez incluye tres tipos de subespecies alopátricas: *axillaris, parodii* y *subandina* (Turchetto *et al*., 2014).

Mientras que la mayoría de las especies pertenecientes a Solanaceae como el tomate, la patata, el tabaco, el pimiento y la berenjena presentan un número de cromosomas base de x = 12, la petunia presenta sólo x = 7 (Bombarely et al., 2016). Además, el genoma de las especies de petunia presenta un tamaño entre 1,3 Gb a 1,57 Gb, siendo mayor al de la mayoría de otras especies de la familia de las solanáceas (Alisawi et al., 2023).

1.3.2. Modelo

La petunia es considerada la primera planta cultivada de jardín y se ha mantenido como una de las favoritas para el desarrollo de nuevas variedades (Gerats y Vandenbussche, 2005). Así, esta planta constituye el 6% en Europa y 11% en los Estados Unidos del valor al por mayor del total de plantas de jardín. No obstante, fue a partir de 1950 cuando los genetistas empezaron a realizar análisis genéticos y bioquímicos en petunia con el fin de predecir nuevas clases de color (Druege y Franken, 2019). Así, en el primer boletín PMB publicado por la asociación de biología molecular de plantas (*Plant Molecular Biology Association*) en junio de 1980, la petunia, junto con el tomate, se mencionó como sistema modelo sobresaliente.

Algunas de las características que presenta como una buena especie modelo es su fácil cultivo y corto ciclo de vida, la facilidad para estudios bioquímicos y citogenéticos (Gerats y Vandenbussche, 2005), su alta variabilidad genética, fácil propagación asexual, y procedimientos de transformación sencillos.

Así, a lo largo de los años la petunia ha contribuido a estudios sobre la inflorescencia (Castel *et al*., 2010), el desarrollo floral (Frost, 1915), su interacción simbiótica con hongos (Druege *et al*., 2019), la prevención de autofertilización por autoincompatibilidad gametofítica o GSI (Bombarely *et al*., 2016), la composición del ADN repetitivo, organización cromosómica (Alisawi *et al*., 2023), la transformación del ADN desnudo y la caracterización de genes de síntesis de flavonoides (Gerats y Vandenbussche, 2005). No obstante, la petunia es ampliamente conocida por ser el modelo base para el descubrimiento del ARN de interferencia en 1990 por Napoli y Jorgensen, el cual es un mecanismo molecular conservado el cual juega un rol esencial en el silenciamento post-transcripcional de genes en varios organismos (Chaudhary *et al*., 2024).

2. OBJETIVOS

El **objetivo general** de este proyecto es realizar una **re-anotación del genoma de la *Petunia axillaris*,** anotado previamente por Bombarely *et al*. (2016). Sin embargo, debido a los rápidos avances en las tecnologías de secuenciación y anotación y al aumento significativo de datos disponibles en las bases de datos genómicas, se ha llevado a cabo una nueva anotación con el propósito de mejorar la precisión y la calidad de la información genética disponible, siendo posible la identificación de genes y elementos genéticos que no se detectaron en la anotación anterior. Para ello se plantearon los **objetivos parciales** detallados a continuación:

- Identificación y enmascaramiento de los **elementos repetitivos.**
- **Analizar** qué **herramientas bioinformáticas** eran las **más adecuadas** para trabajar con el genoma de interés.
- Realizar la **anotación** tanto **basada en evidencias** como ***ab initio.***
- **Evaluar la calidad** de la nueva anotación y la **comparación** con la previamente realizada.
- Realizar la **anotación funcional** a partir de los resultados obtenidos a través de la mejor metodología de anotación.

3. MATERIALES Y MÉTODOS

3.1. Bases de Datos

Para este estudio se recopilaron datos de diferentes bases de datos públicas. La versión PaxPHifi2020 del genoma de la *Petunia axillaris* se descargó de la base de datos del Centro Nacional de Información Biotecnológica (NCBI) mediante la herramienta curl ((https://api.ncbi.nlm.nih.gov/datasets/v2alpha/genome/accession/GCA_029990575.1/download?include_annotation_type=GENOME_FASTA,GENOME_GFF,ARN_FASTA,CDS_FASTA, PROT_FASTA,SECUENCIA_REPORT), mientras que las lecturas del transcriptoma de *Petunia axillaris* se recolectaron utilizando la base de datos Archivo de lecturas cortas (SRA) con la herramienta Fastq-dump v2.11.3 (Proyectos: PRJNA863259, PRJNA858035, PRJNA797226, PRJNA750419, PRJEB27162, PRJNA261953 and PRJDB6807). Las lecturas del ARN-Seq con un mínimo de longitud de 50 bases se dividieron en dos grupos: extremo único (*single end reads)* y de extremo pareado (*paired end reads*); y por último se analizó la integridad del genoma mediante la herramienta BUSCO, utilizando la base de datos de Solanales.

Los datos de ensamblaje y anotación de genomas de especies cercanas se recopilaron de diferentes fuentes:

- El genoma de *Arabidopsis thaliana* versión ARAPORT11 se descargó del Phytozome (https://phytozome-next.jgi.doe.gov/info/Athaliana_Araport11).
- El genoma de *Solanum lycopersicum* versión 5.0 se descargó del Phytozome (https://phytozome-next.jgi.doe.gov/info/Slycopersicum_ITAG5_0).
- El genoma de *Nicotiana sylvestris* versión 1.0.0. fue descargado de Zenodo (https://zenodo.org/records/8256252).
- El genoma previo de *Petunia axillaris* versión 1.6.2 de Solgenomics (https://solgenomics.net/ftp/genomes/Petunia_axillaris/).

Los perfiles de modelos ocultos de Markov (HMMs) de la base de datos Pfam-A, que representan familias de dominios proteicos conservados, se descargaron de la página web de InterPro (https://www.ebi.ac.uk/interpro/download/pfam/).

3.2. Programas y Herramientas Utilizados

Se puede encontrar una representación esquemática del flujo de trabajo seguido para la anotación del genoma de la *Petunia axillaris* en el Anexo II.

3.2.1. Procesado de lecturas y evaluación del genoma

El primer paso para la anotación estructural es la preparación de las lecturas crudas del transcriptoma. Para ello, se eliminaron los adaptadores de la secuenciación por Illumina, y se filtraron aquellas lecturas más cortas de 50 pb y con baja calidad mediante Fastq-mcf v1.04.676. Las lecturas finales obtenidas se pueden revisar en el Anexo III. Además. de manera paralela, se obtuvieron estadísticas sobre la calidad de secuenciación y procesado de las lecturas mediante la herramienta Fastq-stats v1.01.

En cuanto al genoma, se evaluó la integridad de los espacios génicos y la calidad del ensamblaje mediante la herramienta BUSCO v5.6.1.

3.2.2. Enmascaramiento del repetitivo

Para realizar la anotación de los elementos repetitivos en el genoma de interés, se sigue un procedimiento estructurado. Primero, se activó un contenedor Docker que contiene todas las herramientas y programas necesarios para el enmascaramiento del contenido repetitivo siguiendo las instrucciones de TETools (https://github.com/Dfam-consortium/TETools). Luego, se utilizó la herramienta BuildDatabase para crear una base de datos de archivos en formatos fáciles de evaluar, ya que el análisis de archivos en formato FASTA es extremadamente ineficiente.

Con estos archivos, se corre el programa de anotación de elementos repetitivos *de novo*, RepeatModeler2 v2.0.5. Este programa realiza varias rondas de análisis y genera varios archivos, uno de los cuales contiene las secuencias consenso para cada familia identificada. Estas secuencias consenso se utilizarán como una de las bibliotecas en el siguiente programa, RepeatMasker v4.1.6. Para obtener más detalles sobre las familias de elementos repetitivos encontradas, se utiliza el programa TESorter v1.4.6 con el archivo de secuencias consenso generado por RepeatModeler.

3.2.3. Anotación estructural

Como se ha mencionado anteriormente, para obtener las lecturas procesadas a nuestro genoma de interés, se pueden utilizar dos programas: HISAT2 v2.2.1 y STAR v2.7.11b. Para determinar cuál ofrecía mejores resultados, se mapeó un pequeño grupo de lecturas con cada software y se compararon los porcentajes de lecturas mapeadas obtenidos por ambos métodos. Una vez seleccionada la herramienta de mapeo basada en estos resultados, se procedió a mapear todas las lecturas utilizando el software seleccionado.

Posteriormente, se utilizó StringTie v2.2.1 para generar modelos de transcritos a partir de las alineaciones realizadas. Para crear un archivo único que contuviera el conjunto no redundante de modelos genéticos generados, se fusionaron todos los modelos utilizando nuevamente StringTie, empleando la opción específica para este propósito.

A continuación, se realizó la identificación de las regiones CDS utilizando TransDecoder v5.7.1. Para ello, primero se empleó un *script* en lenguaje Perl diseñado para convertir el archivo de modelos génicos de Stringtie en formato GTF (Gene Transfer Format) a secuencias de ADNc en formato FASTA. Una vez convertidas, se utilizó la herramienta TransDecoder.LongOrfs para identificar los ORFs en estas secuencias. Posteriormente, se realizó una búsqueda de evidencias proteicas utilizando Diamond blastp, comparando las secuencias ORF contra la base de datos de secuencias de proteínas de TrEMBL, la cual forma parte del conjunto de recursos mantenidos por UniProt. Además, se empleó la herramienta hmmscan para buscar evidencias proteicas por homología de secuencias utilizando la base de datos de dominios proteicos Pfam, que contiene perfiles de modelos ocultos de Markov (HMMs). Finalmente, utilizando los resultados obtenidos de los programas anteriores, se predijeron las CDS mediante TransDecoder.Predict, lo que permitió generar un genoma anotado con las regiones codificantes identificadas.

Para llevar a cabo la predicción de genes mediante homología de secuencia se utilizó la herramienta GeMoMa v1.9, la cual permite transferir anotaciones génicas de especies relacionadas al genoma de interés. Con el objetivo de obtener los mejores resultados posibles, se realizaron cuatro análisis utilizando diferentes especies descargadas de diversas bases de datos previamente mencionadas. Las especies seleccionadas fueron: *Arabidopsis thaliana*, un modelo vegetal ampliamente estudiado, aunque filogenéticamente distante de nuestra especie de interés; *Nicotiana sylvestris*, la cual es filogenéticamente cercana a nuestra especie, pero con ensamblaje y anotación de menor calidad; *Solanum lycopersicum*, la cual es filogenéticamente cercana y con ensamblaje y anotación de alta calidad; y la previa anotación y ensamblaje de *Petunia axillaris*. Además, se realizó un análisis adicional combinando los ensamblajes y anotaciones de todas las especies mencionadas. Para determinar cuál de las opciones de las especies proporciona las predicciones más precisas y

completas, se analizaron los resultados mediante el script Perl agat_sp_statistics.pl y el programa BUSCO, utilizando la base de datos de embriofitas como referencia.

Por otra parte, se empleó la *pipeline* BRAKER para realizar la predicción de modelos génicos *ab initio*, la cual incluye el proceso de entrenamiento de AUGUSTUS. Este enfoque integró el genoma enmascarado obtenido a partir de RepeatMasker, las lecturas mapeadas y las secuencias de proteínas de la especie relacionada que mostró mejores resultados en la homología de secuencia previamente descrita. Para la anotación de las regiones no traducidas (UTRs) se utilizó el programa Ingenannot, el cual permitió identificar las distintas posibles isoformas.

Por último, se evaluó la integridad del espacio génico sobre el conjunto de proteínas obtenidas mediante TransDecoder, BRAKER y GeMoMa con la herramienta BUSCO 5.6.1.

Además, se utilizó AGAT v0.7.0 sobre el ensamblaje del genoma para obtener métricas sobre el número de genes, exones y transcritos. También se calculó la puntuación AED mediante el programa Ingenannot usando los genes predichos por BRAKER, las lecturas mapeadas al genoma obtenidas mediante StringTie y las secuencias de proteínas predichas por GeMoMa.

Finalmente, se utilizó OrthoVenn3 para obtener representaciones visuales de las relaciones filogenéticas entre diversas especies y la de interés. Para ello se utilizó el genoma anotado por la metodología que mejor resultados haya presentado, junto con las especies *Arabidopsis thaliana*,y las solenáceas *Solanum tuberosum*¸ *Solanum lycopersicum*, y *Nicotiana attenuata* (siendo estas las que estaban disponibles en el programa de OrthoVenn3).

3.2.4.Anotación funcional

Una vez finalizada y evaluada la anotación estructural, se procedió a la anotación funcional. Para ello, se seleccionó el conjunto de proteínas predichas que presentó la mejor integridad del espacio génico, evaluada mediante el análisis de ortólogos de copia única utilizando BUSCO. Este análisis permitió determinar el conjunto de proteínas con mayor completitud y menor fragmentación y duplicación. Las secuencias proteicas seleccionadas se compararon con las bases de datos TrEMBL y Swiss-Prot para identificar homólogos mediante el programa Diamond Blastp.

Posteriormente, las funciones de las proteínas se anotaron basándose en las homologías encontradas utilizando AHRD (Automated Assigned Human Readable Descriptions). Este proceso automatizado asigna descripciones funcionales comprensibles a las proteínas, facilitando la interpretación biológica de los resultados. Los datos de anotación funcional obtenidos se integraron en el genoma anotado, proporcionando una visión completa y precisa de las capacidades funcionales del organismo estudiado.

4. RESULTADOS Y DISCUSIÓN

Con el objetivo de mejorar la previa versión de la anotación del genoma de la *Petunia axillaris*, se emplearon diferentes metodologías de anotación, las cuales se evaluaron para comprobar cuál es la que producía mejores resultados en términos de identificación más completa y precisa de los modelos génicos.

4.1. Realización y Evaluación de Anotación

4.1.1. Identificación del paisaje repetitivo

Los resultados de la anotación de los elementos repetitivos del genoma de *Petunia axillaris* se resumen en la Tabla 1. Más del 65% del genoma fue identificado como elementos repetitivos (ER). Los genomas de las especies pertenecientes a la familia *Solanaceae* están constituidos hasta en un 50-60% por ERs, siendo especialmente abundantes los retrotransposones (Mehra *et al*,.2015). Sin embargo, el paisaje repetitivo de los genomas de petunia muestra una proporción relativamente mayor en comparación con otras especies de esta familia, con un rango de entre el 60% y el 65% del genoma total. Estos hallazgos concuerdan con los resultados obtenidos, en los que se logró enmascarar 0,88 Gb de elementos repetitivos de un total de 1,31 Gb, lo cual representa un 67,58% de elementos repetitivos en el genoma de *Petunia axillaris*.

Se detectó una ausencia total de elementos Penelope, lo cual es esperable puesto que en tanto genomas de planta como hongos y protistas estos retroelementos se encuentran prácticamente ausentes, excepto en organismos específicos (Arkhipova, 2006). Además, se observó una escasa presencia de retrotransposones no-LTR, con un porcentaje nulo de SINE y un 1,63% de LINE, lo cual también era previsible dado que estos se encuentran raramente en plantas, estando presentes solo en genomas nucleares de algunas especies del reino vegetal (Schmidt, 1999).

Por otro lado, la cantidad de elementos LTR representa hasta un tercio del genoma total de *Petunia axillaris*. Esto es consistente con lo esperado, ya que estos retrotransposones constituyen el grupo más abundante de elementos transponibles en las plantas y son de gran importancia debido a su influencia en la evolución y expresión de genes (Galindo-González *et al*., 2017).

Se detectó una notable abundancia de retrotransposones de la superfamilia Gypsy, constituyendo el 25,49% del contenido genómico. Este hallazgo es previsible, ya que los retrotransposones Gypsy están ampliamente distribuidos en el reino vegetal, siendo comunes en numerosos genomas de plantas debido a su alta capacidad de proliferación y su integración preferencial en regiones heterocromáticas del genoma (de Assis *et al*., 2020).

Hasta un 27,26% de los elementos repetitivos no pudieron ser clasificados, lo cual puede explicarse por el hecho de que las clasificaciones de RepeatMasker se basan en parte en la similitud de secuencia con los elementos encontrados por RepeatModeler con familias de elementos transponibles ya conocidas en la base de datos como Dfam. Esto implica que la calidad de la clasificación mejora cuanto más estrechamente relacionado esté el genoma con una especie ya representada en la base de datos y el nivel de representación de dicha especie. Por lo tanto, es posible que la especie en estudio esté muy distanciada de las especies presentes en la base de datos, lo que resultaría en el descubrimiento de grupos de elementos transponibles que no pueden ser clasificados de manera confiable mediante métodos automatizados (Smit *et al*., 2013).

Tabla 1. Resultados de anotación de elementos repetitivos del genoma de la *Petunia axillaris*.

	Número elementos	Tamaño (Mb)	Porcentaje de secuencia
Retroelementos	310.118	482,25	36,84%
SINEs	0	0	0,00%
LINEs	31.444	21,31	1,63%
Penelope	0	0	0,00%
Elementos LTR	278.674	461	35,21%
Ty1/Copia	64.841	94,19	7,20%
Gypsy/DIRS1	189.087	333,67	25,49%
Transposones de DNA	42.058	34,71	2,65%
Sin clasificar	988.234	356,78	27,26%
Repeticiones simples y baja complej	50.326	10,42	0,80%
Total	1.562.383	884,16	67.58 %

4.1.2.Comparación entre herramientas de mapeo de lecturas

Para el mapeo de evidencia transcriptómica al genoma de interés, se llevaron a cabo alineamientos utilizando un conjunto reducido de lecturas y se compararon los resultados obtenidos por dos programas, STAR y HISAT2. A pesar de que los porcentajes de lecturas mapeadas variaron considerablemente entre los diferentes conjuntos, todos los análisis coincidieron en que STAR presentó un mayor porcentaje de alineaciones en comparación con HISAT2 (Tabla 2).

Esto coincide con lo esperado puesto que STAR suele superar a HISAT2 en el alineamiento de lecturas de ARN-seq debido a varios factores como su alta sensibilidad y precisión en la detección de uniones de empalme, y en la eficiencia para manejar genomas grandes y datos de alto rendimiento (Dobin *et al*., 2013). Así, en un análisis en el que se evaluaron múltiples métodos de alineamiento y cuantificación utilizando datos simulados y reales, los resultados mostraron que STAR, cuando se usaba en modo de alineación, proporcionaba una correlación más alta con los valores de expresión verdaderos en comparación con HISAT2, y mostró una mayor precisión en la cuantificación de genes expresados, especialmente en transcripciones complejas y largas (Srivastava *et al*., 2020). En otro estudio comparativo realizado con datos de ARN-seq de 48 muestras geográficamente distintas del hongo *Erysiphe necator*, los resultados mostraron que, aunque HISAT2 era significativamente más rápido, STAR y HISAT2 ofrecieron un rendimiento comparable en términos de tasa de alineación y cobertura de genes. Sin embargo, STAR tendió a mostrar un mejor rendimiento en la alineación de transcripciones más largas, mayores de 500 pares de bases (Musich *et al*., 2021).

Tabla 2. Resultados de porcentaje de mapeado de lecturas al genoma de referencia con STAR y con HISAT2, incluyendo el set de datos que falló (SRR17617394).

	STAR	HISAT2
DRR126327	86,34%	80,69%
SRR10416215	71,64%	64,61%
SRR1585635	94,22%	86,40%
SRR17617394	0,14%	0,10%

4.1.3.Comparación de especies para la anotación basada en evidencias proteicas

En cuanto a la anotación de proteínas basada en comparación con genomas de especies relacionadas mediante GeMoMa, se utilizaron los genomas anotados y ensamblados de las especies *Arabidopsis thaliana, Nicotiana sylvestris, Solanum lycopersicum* y también la versión anotada 1.6.2 de *Petunia axillaris*. Mediante este análisis se pretende identificar la base de datos más adecuada para encontrar el mayor número de homologías con las evidencias proteicas. Además, se realizó un análisis complementario que incluyó las secuencias de todas las especies mencionadas para ampliar el espectro de comparación y mejorar la precisión de la anotación proteica.

Las relaciones filogenéticas entre estas especies son esenciales para entender la eficiencia del análisis. *Arabidopsis thaliana* pertenece a la familia *Brassicaceae* y es un modelo ampliamente utilizado en

estudios de biología molecular y genética debido a su genoma completamente secuenciado y ampliamente caracterizado (Meinke *et al*., 1998). *Nicotiana sylvestris* y *Solanum lycopersicum*, por otro lado, pertenecen a la familia *Solanaceae*, al igual que *Petunia axillari*s, lo cual implica una mayor cercanía evolutiva entre ellas en comparación con *Arabidopsis* (Ghatak *et al*., 2017). Esta cercanía filogenética sugiere que los genomas de *Nicotiana sylvestris* y *Solanum lycopersicum* podrían proporcionar un mayor número de homologías y anotaciones de proteínas más precisas para *Petunia axillaris* debido a su similitud genética.

No obstante, al contrario de lo esperado por las relaciones filogenéticas, la predicción usando las evidencias de *Arabidopsis thaliana* mostró los mejores resultados según el análisis BUSCO, con un 93,7% de conservación de genes, a pesar de presentar el menor número de genes encontrados como se observa en la Tabla 3. Esta observación indica que, a pesar de ser una especie filogenéticamente más distante, la identificación de genes conservados ha sido más exitosa en *Arabidopsis thaliana*. Este fenómeno puede explicarse por el hecho de que esta especie es el organismo modelo vegetal por excelencia, por lo que se encuentra perfectamente anotada y estudiada cómo es posible comprobar en la tabla 4 y, por lo tanto, presentará un nivel de genes conservados anotados mucho más alto en comparación con otras especies filogenéticamente más cercanas pero menos estudiadas, gracias a la robustez de su base de datos de referencia.

Por otro lado, AGAT ha mostrado mejores métricas para las predicciones en las que se han usado las especies filogenéticamente cercanas como *Nicotiana sylve*stris y *Solanum lycopersicum*, tanto en número de genes como número de exones (Tabla 3). Esto es debido posiblemente a que la proximidad filogenética ha facilitado la identificación de genes y elementos genómicos homólogos, incluso cuando las bases de datos de referencia no están tan bien desarrolladas como en el caso de *Arabidopsis thaliana* (Tabla 4).

En el caso de la predicción basada en las evidencias de la versión anterior de la anotación de *Petunia axillaris*, se han obtenido peores métricas por AGAT y resultados con BUSCO que con el resto de las especies. Esto es esperable debido a que a pesar de ser la misma especie a la de interés, su anotación es la menos completa como se comprueba en la Tabla 4. Mientras tanto, los mejores resultados obtenidos por ambas herramientas se tratan de la predicción en las que se ha utilizado las evidencias proteicas de todas las especies en conjunto, lo cual concuerda con lo esperado, debido a que al usar evidencias de diferentes anotaciones y especies, se cubre un mayor número de posibilidades de proteínas. Además, GeMoMa ha sido capaz de reconocer las distintas isoformas únicamente utilizando el conjunto de especies, presentando un número diferente de genes y transcritos (Tabla 3).

Tabla 3. Resultados de evaluación de proteínas predichas por GeMoMa utilizando bases de datos de diferentes especies. Para la columna de BUSCO: C, significa completo; S, significa singular; D, significa duplicado; F, significa fragmentado; y M, significa ausente (del inglés *Missing*).

	Número de genes	Número de transcritos	Número de exones	Longitud total genes (Mb)	BUSCO
Arabidopsis thaliana	26.884	26.884	138.873	103,32	C:93.7%[S:90.9%,D:2.8%],F:0.3%,M:6.0%,n:1614
Nicotiana sylvestris	40.902	40.902	217.802	182,89	C:85.7%[S:82.5%,D:3.2%],F:3.0%,M:11.3%,n:1614
Solanum lycopersicum	39.289	39.289	182.913	146,27	C:84.6%[S:81.6%,D:3.0%],F:2.4%,M:13.0%,n:1614
Petunia axillaris	34.486	34.486	166.958	132,36	C:75.6%[S:71.8%,D:3.8%],F:4.3%,M:20.1%,n:1614
Conjunto especies	61.325	69.273	313.148	215,88	C:98.8%[S:94.2%,D:4.6%],F:0.7%,M:0.5%,n:1614

Tabla 4. Resultados de evaluación por BUSCO de cada anotación de genoma de las especies utilizadas.

	BUSCO
Arabidopsis thaliana	C:99.5%[S:55.3%,D:44.2%],F:0.1%,M:0.4%,n:1614
Nicotiana sylvestris	C:93.6%[S:91.5%,D:2.1%],F:3.3%,M:3.1%,n:1614
Solanum lycopersicum	C:94.1%[S:75.5%,D:18.6%],F:2.3%,M:3.6%,n:1614
Petunia axillaris	C:91.2%[S:88.2%,D:3.0%],F:6.5%,M:2.3%,n:1614

4.1.4.Comparación de las diferentes metodologías de anotación

Una vez completadas las anotaciones mediante las tres diferentes metodologías (*ab-initio* con BRAKER, con evidencias proteicas con GeMoMa y con evidencias transcriptómicas con StringTie), se procedió a evaluar cada metodología para determinar cuál ofrecía mejores resultados.

Como se detalla en la Tabla 5, las métricas obtenidas a partir de las herramientas BUSCO y AGAT muestran que StringTie presenta la menor calidad de anotación en comparación con las otras metodologías. La calidad deficiente de las anotaciones generadas por StringTie puede estar relacionada con la calidad de las evidencias transcriptómicas provenientes de las bases de datos (Kovala *et al*., 2019). Además, se observó una limitada variedad en cuanto al origen y estado de desarrollo de los tejidos en las evidencias transcriptómicas, siendo la mayoría de las muestras procedentes de diferentes partes de la flor, particularmente del pétalo. Esta escasa diversidad en la expresión génica pudo haber afectado negativamente la calidad de las anotaciones obtenidas mediante esta metodología (Ward *et al*., 2012).

Tabla 5. Resultados de las evaluaciones de tres diferentes metodologías de anotación estructural de genoma.

	Número de genes	Número de transcritos	Número de exones	Longitud total genes (Mb)	BUSCO
StringTie	50.134	149.133	186.172	301,85	C:79.4%[S:62.3%,D:17.1%],F:2.6%,M:18.0%,n:5950
BRAKER	82.065	95.492	346.612	187,24	C:97.0%[S:69.3%,D:27.7%],F:1.2%,M:1.8%,n:5950
GeMoMa	61.325	69.273	313.148	215,88	C:98.8%[S:94.2%,D:4.6%],F:0.7%,M:0.5%,n:1614

De la misma manera, BRAKER ha sido entrenado con la misma evidencia transcriptómica, por lo que a pesar de ser una herramienta de predicción *ab-initio*, enfrenta los mismos problemas que el anterior programa. Así, encontramos que a pesar de haber obtenido mayor número de genes y de exones que GeMoMa, este presenta una menor longitud de genes, lo cual encaja con el mayor porcentaje de fragmentación respecto a GeMoMa, además de encontrar un elevado nivel de duplicaciones según el análisis BUSCO (Tabla 5).

Para analizar en mayor profundidad los 95.492 transcritos asociados a los modelos génicos identificados por BRAKER, se realizó un análisis de los valores AED. Este análisis reveló que hasta 25.729 modelos génicos contaban con evidencias proteicas con un AED < 0,2 y evidencias transcriptómicas con un AED < 0,5. Sin embargo, se encontraron 54.230 modelos génicos sin ninguna evidencia de soporte, mientras que 5.182 estaban respaldados únicamente por transcritos y 10.348 por proteínas (ver Figura 3).

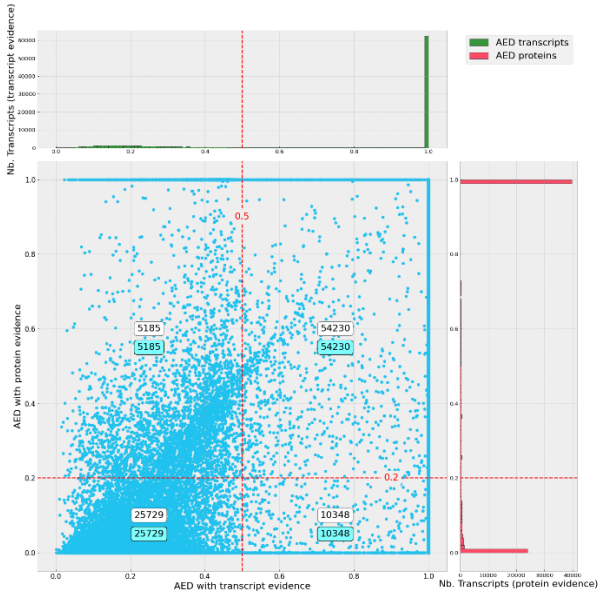


Figura 3. Análisis AED de los transcritos asociados modelos génicos obtenidos mediante BRAKER.

Con base en estos resultados, se decidió clasificar los modelos génicos en dos grupos: aquellos de alta calidad, que estaban respaldados por ambas evidencias, y aquellos de calidad media y alta, respaldados por al menos una de las evidencias. El análisis BUSCO de cada grupo, así como del conjunto total de modelos génicos sin filtrar, demostró que al filtrar los modelos de calidad media y bajase perdían modelos relevantes, evidenciado por una disminución en el porcentaje de integridad BUSCO (Tabla 6). Sin embargo, incluso después de filtrar solamente los modelos génicos de peor calidad, el porcentaje de integridad de la anotación resultante seguía siendo inferior al obtenido mediante GeMoMa (Tabla 5).

Tabla 6. Resultados de la evaluación por BUSCO de los tres grupos de transcritos asociados a los modelos génicos de BRAKER generados a partir del análisis de AED: todos, respaldados por datos transcriptómicos y proteómicos como calidad alta, y respaldados por datos transcriptómicos y/o proteómicos, como calidad media y alta.

	Número de modelos génicos	BUSCO
Modelos génicos encontrados	95.492	C:95.0%[S:62.7%,D:32.3%],F:3.9%,M:1.1%,n:1614
Modelos génicos de calidad alta	25.729	C:90.3%[S:60.3%,D:30.0%],F:1.9%,M:7.8%,n:1614
Modelos génicos de calidad media	41.263	C:95.1%[S:62.8%,D:32.3%],F:3.8%,M:1.1%,n:1614

Se puede observar que, basándonos en los sistemas de evaluación utilizados, inicialmente GeMoMa ha demostrado ser la metodología que ofrece la mejor anotación estructural. Como se mencionó anteriormente, GeMoMa se benefició de evidencias provenientes de múltiples organismos de referencia, lo que contribuye a una mejora en el rendimiento de la predicción. Además, estudios comparativos han demostrado que GeMoMa supera a herramientas como BRAKER, MAKER2 (Holt y Yandell, 2011) y CodingQuarry (Testa *et al*., 2015), así como a los *pipelines* basados exclusivamente en secuencias de ARN para la identificación de transcripciones (Keilwagen *et al*., 2018).

4.2. Comparación con Anotación Anterior

En relación con los resultados obtenidos sobre el contenido repetitivo, se observaron diferencias significativas entre la versión anterior realizada por Bombarely *et al*. (2016) y la versión actualizada, siendo la proporción de elementos repetitivos mucho menores en esta última. Así, mientras los elementos LTR siguen siendo el tipo de elemento repetitivo más abundante, como se muestra en la Tabla 7, la proporción de abundancia de sus familias varía considerablemente. En la versión anterior, la superfamilia Gypsy representaba el 51% de los retrotransposones LTR, y el resto correspondía principalmente a la superfamilia Copia. En la nueva anotación, del 35,21% de elementos LTR encontrados, el 72,39% pertenecían al grupo Gypsy y solo el 7,53% al grupo Copia, como se muestra en la Tabla 1. También destaca la drástica reducción de repeticiones simples y de baja complejidad, pasando de 15,13% en la versión 1.6.2 al 0,8% en la versión actualizada.

Estas marcadas diferencias en el contenido repetitivo entre ambas anotaciones podrían atribuirse a las distintas metodologías empleadas. En la anotación realizada en este proyecto se utilizaron los programas RepeatMasker, RepeatModeler y TESorter, mientras que en el proyecto anterior se emplearon Geneious versión 7.1.4 (http://www.geneious.com/), Jellyfish versión 2.1.3 (Marçais y Kingsford, 2011), LTR-STRUC (McCarthy y McDonald, 2003), LTR finder (http://tlife.fudan.edu.cn/ltr_finder) y RepeatExplorer (Novak *et al*., 2013).

Tabla 7. Resultados del análisis del repetitivo de la anotación anterior y actualizada.

	Petunia axillaris versión1.6.2		*Petunia axillaris* nueva versión	
	Longitud (Mb)	**Porcentaje de secuencia**	**Longitud (Mb)**	**Porcentaje de secuencia**
SINEs y LINEs	29,28	2,33%	21,31	1,63%
Elementos LTR y retrotransposones	508,79	40,41%	461	35,21%
Transposones de DNA	65,59	5,21%	34,71	2,65%
Repeticiones simples y baja complejidad	190,48	15,13%	10,42	0,80%

Por otro lado, en cuanto a la anotación obtenida mediante GeMoMa, tal como se muestra en la Tabla 8, se obtuvo casi el doble de genes, exones e introns en comparación con la anotación anterior. Esta cantidad de genes superior a lo esperado podría deberse a falsos positivos producidos por interferencias causadas por elementos transponibles, ya que como se ha comprobado anteriormente con la tabla 7, se ha enmascarado una menor proporción de estos en comparación con la versión anterior. No obstante, es importante señalar que el porcentaje de fragmentación se redujo del 6,5% al 0,7%, lo cual también podría contribuir al aumento en el número de genes detectados.

Así, al igual que en el análisis del contenido repetitivo, las diferencias observadas también pueden atribuirse a los distintos programas utilizados para realizar la anotación, al uso de diferentes evidencias proteicas y transcriptómicas, e incluso a la versión diferente del genoma. El genoma utilizado para la versión 1.6.2 se ensambló mediante SOAPdenovo3 (Li *et al*., 2015) y se anotó estructuralmente mediante el predictor *ab initio* MAKER-P. Esta anotación se integró con la obtenida mediante evidencias, en las cuales las bibliotecas de lecturas se secuenciaron tanto por PacBio como por Illumina, y estas fueron mapeadas mediante TopHat2 (Kim *et al*., 2013) y ensambladas mediante Cufflinks (Trapnell *et al*., 2012).

No obstante, las métricas obtenidas a partir de los modelos génicos de media calidad de BRAKER anteriormente obtenidos presentan un número de genes mucho menor que los obtenidos mediante GeMoMa y más cercano a la anotación anterior, con un total de 34,402 genes. Sin embargo, es extremadamente llamativo el elevado porcentaje de genes duplicados encontrados mediante el análisis BUSCO, alcanzando el 32,3%.

Debido a ello, se realizó un análisis BUSCO complementario a los modelos génicos de media calidad de BRAKER, utilizando solo los 34,402 genes con isoformas de mayor longitud, es decir, sin tener en cuenta las distintas isoformas producidas por el empalme alternativo. Esto resultó en un porcentaje de BUSCO completo del 94,8%, y reduciendo el porcentaje de genes duplicados al 3%. Esto sugiere que muchas de las proteínas predichas son diferentes isoformas de un mismo gen. Por lo tanto, aunque inicialmente GeMoMa parecía ofrecer los mejores resultados, una vez comparada con la versión anterior de la anotación de *Petunia axillaris*, se ha observado que BRAKER realizó la anotación de manera más precisa.

Tabla 8. Métricas y resultados de la anotación realizada por GeMoMa, por BRAKER y la de la versión 1.6.2.

	Petunia axillaris versión 1.6.2	*Petunia axillaris* GeMoMa	*Petunia axillaris* BRAKER de calidad media
Número de genes	32.928	61.325	34.402
Número de mRNAs/proteínas	32.928	69.273	42.027
Número de exones	173.712	313.148	218438
Número de introns	138.743	243.875	176411
Longitud total de mRNAs (Mb)	140,06	243,73	146,77
Longitud total de exones (Mb)	41,39	81,36	53,65
Longitud total de introns (Mb)	98,29	162,37	93,12
Media de longitud por gen (bp)	4.252	3.520	3.287
Media longitud mRNA (bp)	4.252	3.518	3492
Media de longitud de exon (bp)	238,2	259	245
Media de longitud de introns (bp)	708,4	665	527
BUSCO	C:91.2%[S:88.2%,D:3.0%],F:6.5%,M:2.3%,n:1614	C:98.8%[S:94.2%,D:4.6%],F:0.7%,M:0.5%,n:1614	C:95.1%[S:62.8%,D:32.3%],F:3.8%,M:1.1%,n:1614

4.3. Genómica Comparada con Especies Relacionadas

Con el fin de contextualizar a *Petunia axillaris* dentro de la familia de las solanáceas y analizar las relaciones filogenéticas, se compararon los resultados de su anotación con cuatro especies vegetales: *Solanum tuberosum* versión 6.1 (Pham *et al*., 2020), *Solanum lycopersicum* version 5.0 (Zhou *et al*., 2022), *Nicotiana attenuata* (Xu *et al*.,2017), y la especie modelo *Arabidopsis thaliana* versión ARAPORT11 (Cheng *et al*., 2017). En esta comparación se observó que el número de proteínas y genes encontrados mediante el análisis de GeMoMa en *Petunia axillaris*, al igual que en el apartado anterior, fue significativamente mayor al de las otras especies. Sin embargo, el número de modelos génicos de media y alta calidad filtrados por BRAKER mostró una mayor concordancia con los valores esperados (Tabla 9).

Tabla 9. Resultado de número de proteínas y genes totales de las cinco especies incluidas en el análisis de genómica comparada.

Especies	Versión	Número de genes	Número de transcritos	Tamaño genoma (Gb)
Petunia axillaris	**GeMoMa**	61.325	69.273	1,31
Petunia axillaris	**BRAKER calidad media y alta**	34.402	42.027	1,31
Solanum tuberosum	**6.1**	32.917	44.851	0,84
Solanum lycopersicum	**5.0**	36.648	43.748	0,9
Nicotiana attenuata	***NIATTr2***	34.094	44.491	2,57
Arabidopsis thaliana	**ARAPORT11**	27.655	48.359	0,135

El análisis con OrthoVenn3 permitió identificar 69.273 proteínas en el genoma anotado por GeMoMa, organizándolas en 23.661 clústeres, mientras que 9.145 proteínas no pudieron ser agrupadas. Aunque el número de clústeres y proteínas únicas (*singletons*) fue similar al de las otras especies, especialmente las de la familia *Solenaceae*, la cantidad total de proteínas encontradas en *Petunia axillaris* fue notablemente superior, casi duplicando la cantidad de proteínas en las demás especies incluidas en la comparación (Tabla 10). Esta diferencia se atribuye a las diferencias ya comentadas causadas por el uso de GeMoMa, como el uso de un genoma de referencia donde los elementos repetitivos no se han marcado con RepeatMasker.

Tabla 10. Estadísticas de resultados de agrupamientos de proteínas por OrthoVenn3 con genoma anotado por GeMoMa.

Especies	Proteínas	*Clusters*	*Singletons*
Solanum tuberosum	39.021	20.482	8.248
Solanum lycopersicum	34.429	20.322	9.624
Petunia axillaris	69.273	23.661	9.145
Nicotiana attenuata	33.320	20.508	4.443
Arabidopsis thaliana	27.628	14.229	4.714

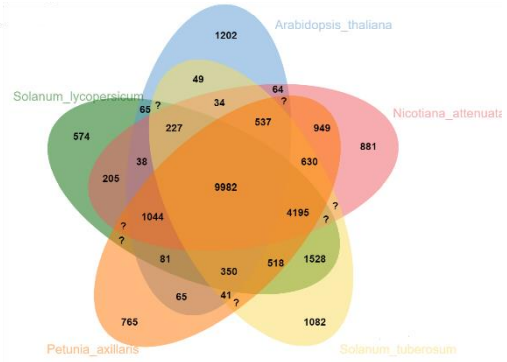
Por ello, ese realizó un análisis de OrthoVenn3 complementario con los modelos génicos de media y alta calidad de BRAKER, utilizando solo los 34.402 genes de isoformas de mayor longitud y, por lo tanto, sin tener en cuenta las distintas isoformas producidas por el splicing alternativo. En este caso se identificaron 34.402 proteínas en el genoma de la *Petunia axillaris*, las cuales se organizaron en 20.252 clústers y dejando 5.318 como *'singletons'*. De esta manera, y como se ha comprobado en apartados anteriores, el número de proteínas encontradas por BRAKER coincide en mayor grado con el resto de las especies relacionadas que en el caso de GeMoMa.

Tabla 11. Estadísticas de resultados de agrupamientos de proteínas por OthoVenn3 con modelos génicos de media y alta calidad filtrados por BRAKER.

Especies	Proteínas	*Clusters*	*Singletons*
Solanum tuberosum	39.021	20.342	8.268
Solanum lycopersicum	34.429	20.133	9.615
Petunia axillaris	34.402	20.252	5.318
Nicotiana attenuata	33.320	20.319	4.475
Arabidopsis thaliana	27.628	14.154	4.771

La comparación de los diferentes clústeres de proteínas permitió inferir relaciones evolutivas entre las especies estudiadas, representadas en la Figura 4.a, y b. Se identificaron 9.982 clústeres compartidos por todas las especies, con una proporción similar de proteínas aportadas por cada especie, lo cual era esperable dado que todas pertenecen al mismo reino y, por lo tanto, comparten un elevado número de proteínas. El siguiente grupo más significativo consistió en 4.195 clústeres compartidos por las especies de la familia *Solenaceae*, sin incluir a la especie *Arabidopsis thaliana*, la cual forma parte de la familia *Brassicaceae* y se estima que divergió del resto de especies incluidas hace alrededor 120 Ma, por lo que concuerda con los resultados (Cao *et al*., 2021). Además, esta es la especie que presenta mayor número de clústeres no compartidos con otras especies, por las mismas razones mencionadas anteriormente. En cuanto a *Solanum lycopersicum* y *Solanum tuberosum*, comparten una gran mayoría de sus clústeres, lo cual era previsible dado que pertenecen al mismo género y divergieron hace relativamente poco tiempo, unos 8 Ma (Cao *et al*., 2021). Por otro lado, *Petunia axillaris* presenta 765 clústeres únicos, en el que se incluyen 2.810 proteínas, siendo similar al número de clústeres y proteínas únicas del resto de especies y, por lo tanto, coincidiendo con lo esperado.

a.



b.

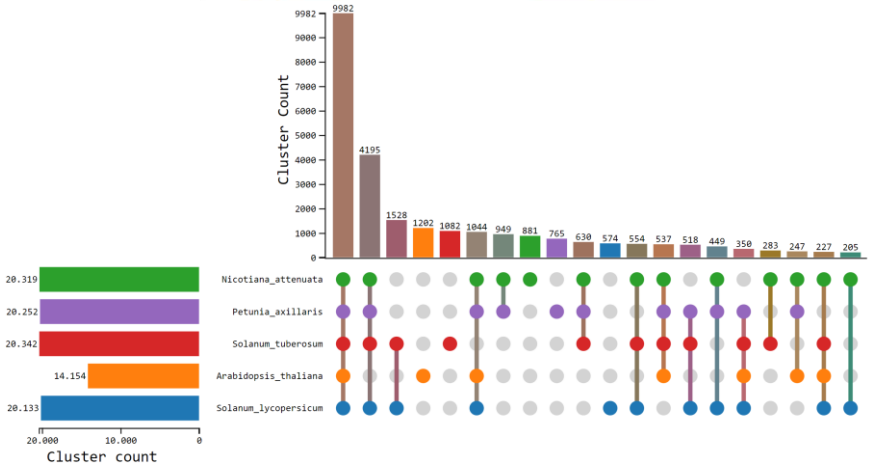


Figura 4. a, Diagrama de Venn basado en el análisis de grupos de familias de genes de las cinco especies analizadas. b, Tabla de ocurrencia basado en el análisis de grupos de familias de genes de las cinco especies analizadas.

Así, según el árbol filogenético obtenido a partir de los clústeres de proteínas entre las especies analizadas (Figura 5) coincide correctamente con el esperado (Figura 6), siendo *Arabidopsis thaliana* la especie más divergente del grupo y las dos especies de *Solanum* las que comparten un ancestro común más reciente.

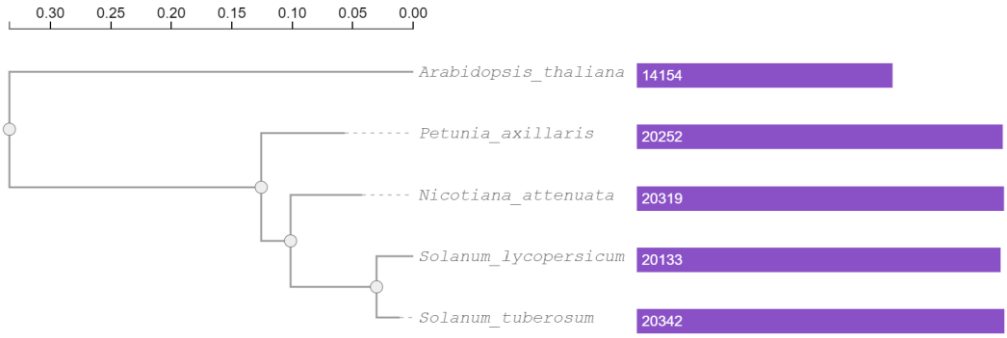


Figura 5. Árbol filogenético basado en la identificación de genes de copia única altamente conservados para describir las relaciones evolutivas entre especies.

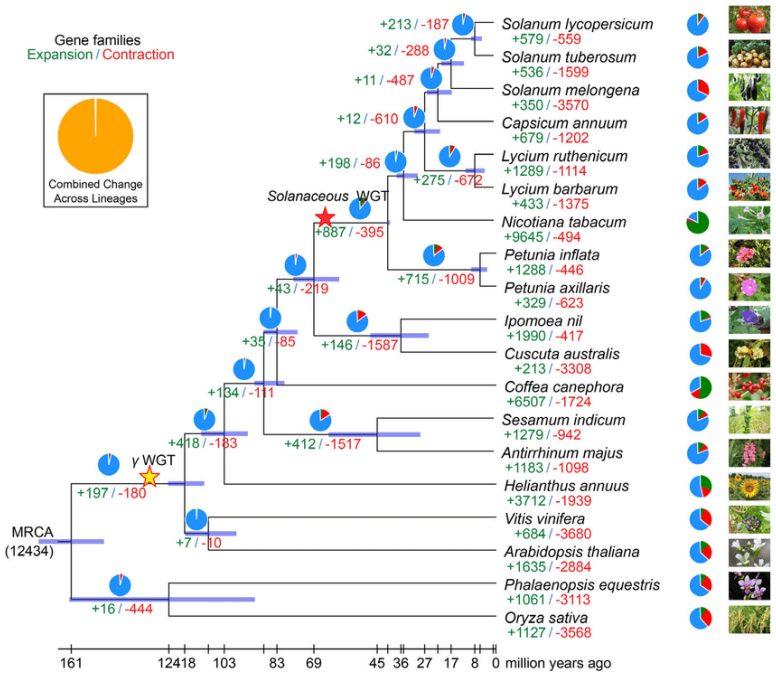


Figura 6. Árbol filogenético que muestra tiempos de divergencia y la evolución del tamaño de la familia de genes para 19 especies de plantas (Cao *et al*., 2021).

4.4. Genes Anotados Funcionalmente

Finalmente, tras comparar las diferentes anotaciones realizadas con la versión anterior del genoma de *Petunia axillaris* y otras especies vegetales, se procedió a la anotación de genes de alta y media calidad (utilizando únicamente las isoformas de mayor longitud) obtenidas mediante BRAKER para la anotación funcional.

De un total de 34.402 proteínas obtenidas, 33.638 fueron identificadas mediante comparación con la base de datos TrEMBL y 24.877 con la base de datos SwissProt. Estos resultados son acordes a lo

esperado, dado que la base de datos SwissProt contiene proteínas manualmente curadas, resultando en un menor número de secuencias en comparación con TrEMBL, aunque estas últimas puedan presentar menor calidad (https://www.ebi.ac.uk/training/online/courses/uniprot-quick-tour/the-uniprot-databases/uniprotkb/).

Además, la herramienta AHRD proporciona un código de calidad para cada una de las proteínas basado en tres criterios: un puntaje de bits en BLAST superior a 50 y un valor e inferior a e^{-10}, un solapamiento en BLAST superior al 60%, y un puntaje de token superior en HRD mayor a 0,5. La asignación funcional fue exitosa, ya que 28.130 proteínas cumplieron con los tres criterios de calidad, representando más del 80% del total, mientras que solo 14 proteínas no cumplieron con ninguno de los criterios. Un total de 4.846 proteínas cumplieron con un puntaje de bits BLAST mayor a 50 y un puntaje HRD mayor a 0,5, pero no alcanzaron un porcentaje de solapamiento en BLAST superior al 60%. Además, 193 proteínas cumplieron con una o dos combinaciones de criterios de calidad. Por otro lado, en 1,220 proteínas fue imposible asignar un código de calidad, representando solo el 4.3% de proteínas sin calidad evaluada.

5. CONCLUSIÓN

Así, del desarrollo del presente proyecto es posible extraer varias conclusiones.

1. A pesar de haber sido posible enmascarar un 67,58% de elementos repetitivos en el genoma de la *Petunia axillaris*, lo cual es similar al esperado, la distribución de las familias de elementos fue extremadamente discordante con la anterior anotación, siendo necesaria una revisión de los métodos utilizados y comprobación de los resultados.
2. Se demostró la importancia de una base de datos completa y de calidad para una correcta anotación, siendo necesaria una mayor cantidad y variedad de lecturas de ARN-Seq para mejorar la anotación basada en evidencias transcriptómicas del genoma de la *Petunia axillaris*.
3. Se comprobó que a pesar de que inicialmente la anotación basada en evidencias proteicas mediante la herramienta GeMoMa presentase mayor conservación de genes y métricas, la anotación *ab initio* realizada mediante BRAKER, una vez filtrados los modelos génicos de peor calidad, presentaba los resultados que corresponden mejor al resto de anotaciones de especies cercanas y a la anotación anterior.
4. En conclusión, tras realizar la re-anotación del genoma de la *Petunia axillaris*, fue posible identificar 34.403 genes y 42.027 transcritos.

6. REFERENCIAS BIBLIOGRÁFICAS

Alisawi, O., Richert-Pöggeler, K. R., Heslop-Harrison, J. S. P., & Schwarzacher, T. (2023). The nature and organization of satellite ADNs in Petunia hybrida, related, and ancestral genomes. Frontiers in plant science, 14, 1232588. https://doi.org/10.3389/fpls.2023.1232588

Armstrong, J., Fiddes, I. T., Diekhans, M., & Paten, B. (2019). Whole-Genome Alignment and Comparative Annotation. Annual review of animal biosciences, 7, 41–64. https://doi.org/10.1146/annurev-animal-020518-115005

Arkhipova I. R. (2006). Distribution and phylogeny of Penelope-like elements in eukaryotes. Systematic biology, 55(6), 875–885. https://doi.org/10.1080/10635150601077683

Aronesty, E. (2013) 'Comparison of Sequencing Utility Programs', The Open Bioinformatics Journal, 7(1), pp. 1-8. https://doi.org/10.2174/1875036201307010001

Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C. S., Bliek, M., Boersma, M. R., Borghi, L., Bruggmann, R., Bucher, M., D'Agostino, N., Davies, K., Druege, U., Dudareva, N., Egea-Cortines, M., Delledonne, M., Fernandez-Pozo, N., Franken, P., Grandont, L., … Kuhlemeier, C. (2016). Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida. Nature plants, 2(6), 16074. https://doi.org/10.1038/nplants.2016.74

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. Genome biology, 19(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Brůna, T., Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR genomics and bioinformatics, 3(1), lqaa108. https://doi.org/10.1093/nargab/lqaa108

Brůna, T., Lomsadze, A., & Borodovsky, M. (2024). A new gene finding tool GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. bioRxiv : the preprint server for biology, 2023.01.13.524024. https://doi.org/10.1101/2023.01.13.524024

Buchfink B, Reuter K, Drost HG, "Sensitive protein alignments at tree-of-life scale using DIAMOND", Nature Methods 18, 366–368 (2021). doi:10.1038/s41592-021-01101-x

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1), 59.

Cao, Y. L., Li, Y. L., Fan, Y. F., Li, Z., Yoshida, K., Wang, J. Y., Ma, X. K., Wang, N., Mitsuda, N., Kotake, T., Ishimizu, T., Tsai, K. C., Niu, S. C., Zhang, D., Sun, W. H., Luo, Q., Zhao, J. H., Yin, Y., Zhang, B., Wang, J. Y., … Liu, Z. J. (2021). Wolfberry genomes and the evolution of Lycium (Solanaceae). Communications biology, 4(1), 671. https://doi.org/10.1038/s42003-021-02152-8

Castel, R., Kusters, E., & Koes, R. (2010). Inflorescence development in petunia: through the maze of botanical terminology. Journal of experimental botany, 61(9), 2235–2246. https://doi.org/10.1093/jxb/erq061

Chaudhary, D., Jeena, A. S., Rohit, Gaur, S., Raj, R., Mishra, S., Kajal, Gupta, O. P., & Meena, M. R. (2024). Advances in ARN Interference for Plant Functional Genomics: Unveiling Traits, Mechanisms, and Future Directions. Applied biochemistry and biotechnology, 10.1007/s12010-023-04850-x. Advance online publication. https://doi.org/10.1007/s12010-023-04850-x

Chen, J., Wang, Z., Tan, K., Huang, W., Shi, J., Li, T., Hu, J., Wang, K., Wang, C., Xin, B., Zhao, H., Song, W., Hufford, M. B., Schnable, J. C., Jin, W., & Lai, J. (2023). A complete telomere-to-telomere assembly of the maize genome. Nature genetics, 55(7), 1221–1231. https://doi.org/10.1038/s41588-023-01419-6

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant journal : for cell and molecular biology, 89(4), 789–804. https://doi.org/10.1111/tpj.13415

Dainat J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. (Version v0.7.0). Zenodo. https://www.doi.org/10.5281/zenodo.3552717

de Assis, R., Baba, V. Y., Cintra, L. A., Gonçalves, L. S. A., Rodrigues, R., & Vanzela, A. L. L. (2020). Genome relationships and LTR-retrotransposon diversity in three cultivated Capsicum L. (Solanaceae) species. BMC genomics, 21(1), 237. https://doi.org/10.1186/s12864-020-6618-9

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal ARN-seq aligner. Bioinformatics (Oxford, England), 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Druege, U., & Franken, P. (2019). Petunia as model for elucidating adventitious root formation and mycorrhizal symbiosis: at the nexus of physiology, genetics, microbiology and horticulture. Physiologia plantarum, 165(1), 58–72. https://doi.org/10.1111/ppl.12762

Eilbeck, K., Moore, B., Holt, C., & Yandell, M. (2009). Quantitative measures for the management and comparison of annotated genomes. BMC bioinformatics, 10, 67. https://doi.org/10.1186/1471-2105-10-67

Ejigu, G. F., & Jung, J. (2020). Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Biology, 9(9), 295. https://doi.org/10.3390/biology9090295

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

Frost, H. B. (1915). The Inheritance of Doubleness in Matthiola and Petunia. I. The Hypotheses. The American Naturalist, 49(586), 623–636. http://www.jstor.org/stable/2456229

Galindo-González, L., Mhiri, C., Deyholos, M. K., & Grandbastien, M. A. (2017). LTR-retrotransposons in plants: Engines of evolution. Gene, 626, 14–25. https://doi.org/10.1016/j.gene.2017.04.051

Gilly, A., Etcheverry, M., Madoui, M. A., Guy, J., Quadrana, L., Alberti, A., Martin, S., Cruaud, C., Gavory, F., Valiere, S., Courtois, B., Descombes, P., Dartevelle, H., Nabholz, B., Aury, J. M., & Wincker, P. (2022). A comprehensive mutation catalog reveals common origins of structural variations in Arabidopsis thaliana. bioRxiv. https://doi.org/10.1101/2022.09.01.506176

Golicz, A. A., Bhalla, P. L., & Singh, M. B. (2020). lncARNs in plant and animal sexual reproduction. Trends in plant science, 25(3), 207–216. https://doi.org/10.1016/j.tplants.2019.12.015

González, M. N., Massa, G. A., Andersson, I., & Hajdukiewicz, P. T. J. (2019). Rational Engineering and Validation of a Novel Nitrogen-Fixing Chassis Based on Pseudomonas protegens. Frontiers in plant science, 10, 1834. https://doi.org/10.3389/fpls.2019.01834

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology, 9(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Heslop-Harrison, J. S. P., Schwarzacher, T., Anamthawat-Jónsson, K., & Shi, Y. (2023). In situ hybridization with specific ADN probes reveals the chromosome organization of repetitive ADN in barley (Hordeum vulgare) and related species. Heredity, 130, 108–118. https://doi.org/10.1038/s41437-023-00563-y

Heslop-Harrison, J. S. P., & Schwarzacher, T. (2011). Organisation of the plant genome in chromosomes. The Plant journal : for cell and molecular biology, 66(1), 18–33. https://doi.org/10.1111/j.1365-313X.2011.04544.x

Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., Negoro, S., Ohyama, A., Yamaguchi, H., Sato, S., Isobe, S., & Tabata, S. (2014). Draft genome sequence of eggplant (Solanum melongena L.): the representative Solanum species indigenous to the Old World. ADN research : an international journal for rapid publication of reports on genes and genomes, 21(6), 649–660. https://doi.org/10.1093/ADNres/dsu027

Hoen, D. R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A. S., Hua-Van, A., Hubley, R., Kapusta, A., & Rouzic, A. L. (2015). A call for benchmarking transposable element annotation methods. Mobile ADN, 6, 13. https://doi.org/10.1186/s13100-015-0044-6

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics, 12, 491. https://doi.org/10.1186/1471-2105-12-491

Jung, S., Ficklin, S. P., Lee, T., Cheng, C. H., Blenda, A., Zheng, P., Yu, J., Bombarely, A., Cho, I., Ru, S., Evans, K., Peace, C., Abbott, A., Mueller, L. A., Olmstead, M., Main, D. (2014). The Genome Database for Rosaceae (GDR): year 10 update. Nucleic acids research, 42(1), D1237–D1244. https://doi.org/10.1093/nar/gkt1012

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, 30(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Kersey, P. J., Allen, J. E., Allot, A., Barba, M., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C., Humphrey, J., Kerhornou, A., Khobova, J., Langridge, N., McDowall, M. D., Maheswari, U., Nightingale, A., Ong, C. K., Paulini, M., Pedro, H., … Staines, D. M. (2018). Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic acids research, 46(D1), D802–D808. https://doi.org/10.1093/nar/gkx1011

Kersey, P. J. (2019). Plant genome sequences: past, present, future. Current opinion in plant biology, 48, 1–8. https://doi.org/10.1016/j.pbi.2018.11.001

Koster, J., & Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. Bioinformatics (Oxford, England), 28(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Mandel, J. R., Dikow, R. B., Funk, V. A., Masalia, R. R., Staton, S. E., Kozik, A., Michelmore, R. W., Rieseberg, L. H., & Burke, J. M. (2014). A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. Applications in plant sciences, 2(2), 1300085. https://doi.org/10.3732/apps.1300085

Martínez-García, P. J., Crepeau, M. W., Puiu, D., Gonzalez-Ibeas, D., Whalen, J., Stevens, K. A., Paul, R., Butterfield, T. S., Britton, M. T., Reagan, R. L., Chakraborty, S., Walawage, S. L., Vasquez-Gross, H. A., Cardeno, C., Famula, R. A., Pratt, K., Kuruganti, S., Aradhya, M. K., Leslie, C. A., Dandekar, A. M., … Neale, D. B. (2016). The walnut (Juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. The Plant journal : for cell and molecular biology, 87(5), 507–532. https://doi.org/10.1111/tpj.13207

Mehra, M., Gangwar, I., & Shankar, R. (2015). A Deluge of Complex Repeats: The Solanum Genome. PloS one, 10(8), e0133962. https://doi.org/10.1371/journal.pone.0133962

Muñoz-Sanhueza, L., Pedro, N., Ventura, R., Carocha, V., Neves, L. G., Almeida, P., Segura, V., & Sederoff, R. (2021). A high-quality genome of Pinus pinaster provides insights into conifer evolution and adaptation. Plant biotechnology journal, 19(6), 1174–1186. https://doi.org/10.1111/pbi.13539

Murat, F., Van de Peer, Y., & Salse, J. (2012). Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. Genome biology and evolution, 4(9), 917–928. https://doi.org/10.1093/gbe/evs066

Niu, S. C., Cao, Y. L., Li, Z., Li, J., Yao, M. C., Hasi, A., Yin, Y., Zhang, B., Wang, J. Y., & Liu, Z. J. (2022). Genome sequencing and analysis of wolfberry, an important fruit crop from the family Solanaceae. Horticulture research, 9, uhac100. https://doi.org/10.1093/hr/uhac100

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., … Phillippy, A. M. (2022). The complete sequence of a human genome. Science (New York, N.Y.), 376(6588), 44–53. https://doi.org/10.1126/science.abj6987

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics (Oxford, England), 20(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., & Salzberg, S. L. (2021). CHESS 2.0: a database of sequences and annotations for human genes and transcripts. Genome biology, 22(1), 310. https://doi.org/10.1186/s13059-021-02563-7

Petroli, C. D., Sansaloni, C. P., Carling, J., Steane, D. A., Vaillancourt, R. E., Myburg, A. A., Kulheim, C., & Duran, C. (2012). Genomic characterization of DArT markers based on high-density linkage analysis and physical mapping to the Eucalyptus genome. PloS one, 7(9), e44684. https://doi.org/10.1371/journal.pone.0044684

Pham, G. M., Hamilton, J. P., Wood, J. C., Burke, J. T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., & Buell, C. R. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. GigaScience, 9(9), giaa100. https://doi.org/10.1093/gigascience/giaa100

Pyke, K. A. (2019). Plastid biogenesis and differentiation. The Plant cell, 31(11), 2699–2726. https://doi.org/10.1105/tpc.19.00442

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature, 592(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Robles, P., & Micol, J. L. (2001). Genome-wide linkage analysis of Arabidopsis seed longevity in recombinant inbred lines. Proceedings of the National Academy of Sciences of the United States of America, 98(9), 4710–4715. https://doi.org/10.1073/pnas.081626598

Rudd, S. (2003). Expressed sequence tags: alternative or complement to whole genome sequences?. Trends in plant science, 8(7), 321–329. https://doi.org/10.1016/S1360-1385(03)00162-6

Schaefer, H., & Renner, S. S. (2011). Phylogenetic relationships in the order Cucurbitales and a new classification of the gourd family (Cucurbitaceae). Taxon, 60(1), 122–138. https://doi.org/10.1002/tax.601011

Schrader, L., Kim, J., Ence, D., Zimin, A., Klein, A., Wyschetzki, K., Weichselgartner, T., Kemena, C., Stökl, J., Schultner, E., Wurm, Y., Smith, C. D., Yandell, M., Heinze, J., Gadau, J., Hultmark, D., & Reinberg, D. (2014). Transposable element islands facilitate adaptation to novel environments in an invasive species. Nature communications, 5, 5495. https://doi.org/10.1038/ncomms6495

Shumate, A., Wong, B., Pertea, G., Pertea, M., & Salzberg, S. L. (2020). Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS computational biology, 17(6), e1009730. https://doi.org/10.1371/journal.pcbi.1009730

Slater, G. S., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC bioinformatics, 6, 31. https://doi.org/10.1186/1471-2105-6-31

Smit, A. F. A., & Hubley, R. (2008). RepeatModeler Open-1.0. Available from: http://www.repeatmasker.org

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics (Oxford, England), 19 Suppl 2, ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080

Stiehler, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, N.., Saadat, N., Ebenhöh, O,M Usadel, B., Schwacke, R., Bolger, M., Weber, A. P. M., & Denton, A. K. (2023). Helixer—de novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model. bioRxiv 2023.02.06.527280; doi: https://doi.org/10.1101/2023.02.06.527280

Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., Zhang, J., Zhang, H., Gong, G., Jia, Z., Zhang, F., Tian, S., Lucas, W. J., Li, C., Fei, Z., Xu, Y., & Weng, Y. (2017). Karyotype stability and unbiased fractionation in the paleo-allotetraploid Cucurbita genomes. Molecular plant, 10(10), 1293–1306. https://doi.org/10.1016/j.molp.2017.09.003

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. Nature reviews. Genetics, 18(7), 411–424. https://doi.org/10.1038/nrg.2017.26

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

Xie, C., & Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC bioinformatics, 10, 80. https://doi.org/10.1186/1471-2105-10-80

Xu, S., Brockmöller, T., Navarro-Quezada, A., Kuhl, H., Gase, K., Ling, Z., Zhou, W., Kreitzer, C., Stanke, M., Tang, H., Lyons, E., Pandey, P., Pandey, S. P., Timmermann, B., Gaquerel, E., & Baldwin, I. T. (2017). Wild tobacco genomes reveal the evolution of nicotine biosynthesis. Proceedings of the National Academy of Sciences of the United States of America, 114(23), 6133–6138. https://doi.org/10.1073/pnas.1700073114

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., Zhang, Z., … Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. Nature, 606(7914), 527–534. https://doi.org/10.1038/s41586-022-04808-9