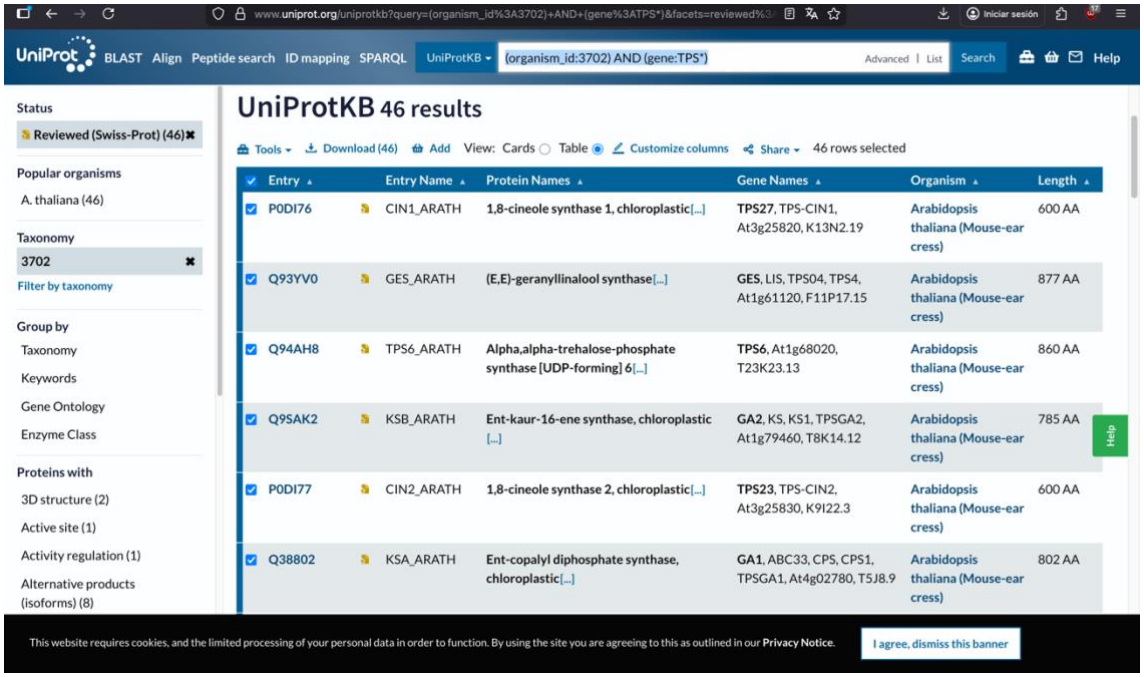
Análisis filogenético y clasificación funcional de la familia de las terpeno sintasas (TPS) en *Cannabis sativa*

Nota: *Los ficheros mencionados a lo largo del informe se encuentran disponibles en el repositorio de GitHub: https://github.com/ix-lopemateo/Comparative_Analysis_TPS_Cannabis_Arabidopsis*

1. Preparación de Secuencias



Selección de la especie modelo y secuencias de referencia

Como especie modelo se seleccionó *Arabidopsis thaliana*, una planta ampliamente utilizada en genética y biología molecular debido a la alta calidad de su anotación genómica, la disponibilidad de un genoma completamente secuenciado y la exhaustiva curación manual de sus proteínas en bases de datos públicas, lo que la convierten en una buena para la identificación por homología de miembros de la familia de las terpeno sintasas (TPS).

Para obtener un conjunto lo más completo posible de proteínas TPS anotadas en esta especie, se realizó una búsqueda en UniProtKB utilizando la siguiente consulta avanzada:

```
(organism_id:3702) AND (gene:TPS*)
```

El uso del identificador taxonómico (3702) garantiza la recuperación exclusiva de entradas pertenecientes a *Arabidopsis thaliana*, mientras que el comodín (*) permite incluir todas las variantes génicas asociadas a la familia TPS (TPS1, TPS2, TPS10, etc.).

Posteriormente, se aplicó el filtro **Reviewed (Swiss-Prot)** con el objetivo de seleccionar únicamente entradas curadas manualmente, ya que las proteínas incluidas en Swiss-Prot cuentan con anotación validada y, en muchos casos, respaldo experimental.

La búsqueda arrojó un total de **46 entradas revisadas**, correspondientes a diferentes miembros de la familia de las terpeno sintasas en *Arabidopsis thaliana*. Estas secuencias presentan longitudes comprendidas aproximadamente entre 550 y 900 aminoácidos, rango compatible con el tamaño esperado de las TPS vegetales, incluyendo tanto monoterpeno sintasas como di- y sesquiterpeno sintasas.

Este conjunto completo de TPS revisadas (*tps_arabidopsisthaliana.fasta*) se utilizará como referencia para los análisis posteriores de identificación por homología, para la construcción del alineamiento múltiple y el análisis filogenético.

Selección de la especie problema

Como especie problema se seleccionó ***Cannabis sativa***, una planta que presenta una notable diversidad química asociada a la expansión de familias génicas implicadas en el metabolismo secundario. Esto la convierte en un modelo adecuado para estudiar la diversificación evolutiva de la familia de las terpeno sintasas (TPS).

Selección de secuencias bona fide

Con el objetivo de anclar el análisis filogenético a enzimas funcionalmente caracterizadas, se seleccionaron tres terpeno sintasas revisadas en UniProtKB/Swiss-Prot y con evidencia experimental a nivel proteico. La inclusión de estas secuencias bona fide permite asignar con mayor fiabilidad los candidatos identificados en *Cannabis sativa* a subfamilias funcionales dentro de la familia TPS.

Las secuencias seleccionadas fueron:

- **Q84LB2 (AFS1_MALDO)** – *(E,E)-alpha-farnesene synthase* de *Malus domestica*. Enzima sesquiterpeno sintasa responsable de la síntesis de α-farneseno.
- **R4YXW8 (TPS2_COFAR)** – *Linalool synthase* de *Coffea arabica*. Monoterpeno sintasa implicada en la biosíntesis de linalool.
- **A0A1V0E492 (TPS1_PIPNI)** – *Terpene synthase 1* de *Piper nigrum*. Sesquiterpeno sintasa con actividad descrita en la conversión de farnesil difosfato en compuestos como β-cariofileno y α-humuleno.

2. Identificación de terpeno sintasas por homología

Con el objetivo de identificar los miembros de la familia de las terpeno sintasas (TPS) en *Cannabis sativa*, se realizó un análisis de homología utilizando la herramienta **NCBI Batch BLAST (BLASTP)**. En lugar de realizar búsquedas individuales secuencia por secuencia, se empleó una estrategia de búsqueda múltiple utilizando como queries el conjunto completo de proteínas TPS previamente descargadas de *Arabidopsis thaliana* (especie modelo).

Las búsquedas se realizaron contra la base de datos **RefSeq protein**, restringiendo el organismo a *Cannabis sativa*, de modo que la comparación se efectuó exclusivamente contra el proteoma anotado de esta especie.

Los parámetros empleados fueron los siguientes:

- Tipo de búsqueda: **BLASTP (proteína vs proteína)**
- Base de datos: **Reference proteins (RefSeq)**
- Organismo restringido: *Cannabis sativa*
- Umbral de significación (E-value): 1e-10
- Matriz de sustitución: BLOSUM62

El resultado inicial generó múltiples alineamientos redundantes, dado que cada proteína de *Arabidopsis* recuperaba frecuentemente los mismos candidatos en *Cannabis*. Para depurar el conjunto de resultados, se descargó la tabla de hits en formato CSV y se procesó mediante un script en Python (*Processing_and_Filtering_BLASTP_results_Cannabis_TPS.ipynb*) utilizando la librería pandas.

El procedimiento de filtrado incluyó:

- Eliminación de duplicados basándose exclusivamente en la accesión del sujeto (subject accession).
- Criterios de calidad:
  - Longitud de alineamiento ≥ 500 aminoácidos.
  - Identidad ≥ 25%.
  - e-value ≤ 1e-5.

El umbral de longitud se estableció en función del tamaño típico de las TPS vegetales completas (~550–650 aminoácidos), con el fin de descartar fragmentos parciales o alineamientos correspondientes a dominios incompletos.

Tras este proceso de depuración, se obtuvo un conjunto final de 61 proteínas candidatas a TPS en *Cannabis sativa* (*TPS_filtradas.txt*), número coherente con la expansión conocida de esta familia génica en plantas superiores.

Estas secuencias se utilizaron posteriormente para el alineamiento múltiple y el análisis filogenético.

3. Alineamiento múltiple

Una vez identificados los candidatos a terpeno sintasas en *Cannabis sativa* mediante búsqueda por homología, se procedió a la recuperación de sus secuencias proteicas completas desde la base de datos NCBI Protein. Para ello, se utilizaron los identificadores de acceso (accessions) obtenidos tras el filtrado de duplicados en la tabla de resultados del BLASTP.

Los identificadores únicos fueron introducidos en el buscador de NCBI Protein utilizando operadores booleanos (OR), lo que permitió recuperar de forma simultánea todas las

proteínas candidatas. Tras confirmar el número total de resultados (61 secuencias), estas fueron descargadas en formato FASTA (*Cannabis_TPS_filtradas.fasta*).

A continuación, se generó un archivo FASTA (*TPS_all_sequences.fasta*) combinado que incluía:

- Las terpeno sintasas de *Arabidopsis thaliana* (especie modelo).
- Las secuencias bona fide.
- Las 61 proteínas candidatas identificadas en *Cannabis sativa*.

Este archivo unificado constituyó el conjunto definitivo de secuencias para el análisis de alineamiento múltiple.

Resultados alineamiento

El alineamiento múltiple de las secuencias TPS de *Arabidopsis thaliana* junto con las candidatas identificadas en *Cannabis sativa* (*tps_alignment_results*) muestra un patrón claro de conservación estructural en la región central de las proteínas, mientras que los extremos N-terminales presentan mayor variabilidad y presencia de inserciones/deleciones.

En el bloque central del alineamiento se observa una región altamente conservada entre la mayoría de las TPS de *Arabidopsis* y las secuencias XP_ de *Cannabis*.

En esta zona, la identidad de aminoácidos es elevada y las sustituciones observadas son mayoritariamente conservativas, lo que indica que se trata de una región funcionalmente relevante y estructuralmente restringida. La homogeneidad de este segmento sugiere que las proteínas candidatas comparten la arquitectura catalítica típica de las terpeno sintasas vegetales.

Por el contrario, el extremo N-terminal muestra una mayor diversidad en longitud y composición. En varias secuencias se aprecian gaps y desplazamientos en el alineamiento, reflejando diferencias en péptidos señal o regiones reguladoras. Esta variabilidad es esperable en enzimas que pueden presentar localización subcelular diferencial o que han experimentado eventos de duplicación.

También se identifican pequeños subgrupos de secuencias de *Cannabis* que presentan patrones casi idénticos entre sí (por ejemplo, pares o tríos con secuencias prácticamente superponibles en el alineamiento), lo que sugiere la existencia de duplicaciones recientes o isoformas muy cercanas evolutivamente dentro del genoma de *Cannabis sativa*.

La longitud global de la mayoría de las secuencias candidatas se mantiene dentro del rango esperado (~550–650 aminoácidos según los alineamientos BLAST asociados).

Identificación del motivo catalítico DDXXD

Con el objetivo de evaluar la conservación del motivo catalítico característico de las terpeno sintasas, se desarrolló un script en Python (*detect_DDXXD.ipynb*) que analiza

automáticamente el archivo TPS_all_sequences.fasta y detecta todas las ocurrencias del patrón tipo **DDXXD** (dos residuos de ácido aspártico seguidos de dos aminoácidos variables y un tercer ácido aspártico).

El análisis identificó:

- **114 ocurrencias totales** del motivo tipo DDXXD en el conjunto de secuencias.
- **26 variantes distintas** del motivo.

Entre las variantes más frecuentes se encuentran:

- **DDIYD** (40 ocurrencias)
- **DDRSD** (15 ocurrencias)
- **DDTCD** (7 ocurrencias)
- **DDDDD** (8 ocurrencias)

Otras variantes menos frecuentes incluyen DDACD, DDFFD, DDIFD, DDDHD, DDEAD, entre otras.

La elevada frecuencia de ciertos patrones, especialmente DDIYD, sugiere una fuerte conservación estructural en una proporción significativa de las secuencias analizadas. No obstante, la existencia de múltiples variantes refleja la plasticidad evolutiva de esta región dentro de la familia TPS, donde sustituciones conservativas pueden mantener la funcionalidad catalítica sin comprometer la arquitectura global de la proteína.

El número total de ocurrencias (114) es superior al número de secuencias analizadas, lo que indica que algunas proteínas contienen más de un motivo tipo DDXXD, fenómeno descrito previamente en miembros de esta familia en plantas, donde pueden coexistir motivos relacionados dentro del dominio catalítico.

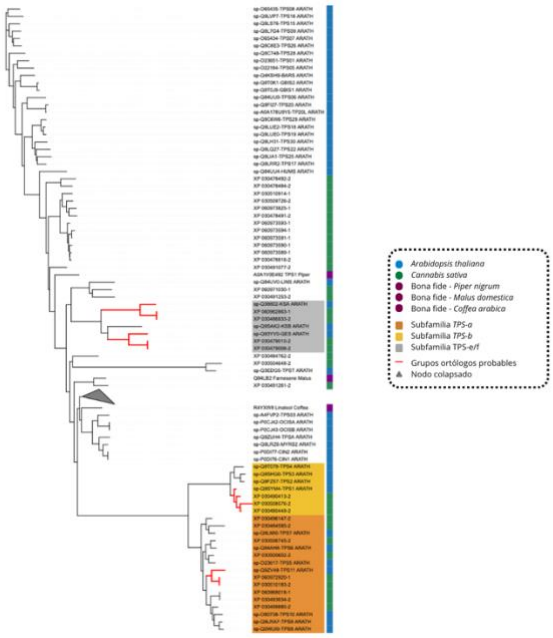
En conjunto, la presencia generalizada del motivo en sus distintas variantes respalda la correcta anotación funcional de las proteínas candidatas como miembros de la familia de las terpeno sintasas.

4. Filogenia rápida y clasificación en subfamilias TPS

Con el objetivo de clasificar funcionalmente los candidatos identificados en *Cannabis sativa*, se llevó a cabo un análisis filogenético utilizando la plataforma NGPhylogeny.fr en modo “à la carte” con el *fichero TPS_all_sequences.fasta*. El pipeline empleado incluyó alineamiento múltiple con MAFFT, curado del alineamiento mediante trimAl (modo automated1) y reconstrucción filogenética con FastTree bajo modelo proteico, siguiendo las recomendaciones del enunciado.

Anotación y mejora visual del árbol filogenético en iTOL

El árbol filogenético obtenido mediante NGPhylogeny.fr fue importado en la plataforma iTOL (Interactive Tree of Life) con el objetivo de mejorar su legibilidad y adaptarlo a un formato publicable.



En primer lugar, se aplicó un código de colores para distinguir el origen taxonómico de las secuencias. Las secuencias de *Arabidopsis thaliana* se representaron en azul, las de *Cannabis sativa* en verde y las secuencias bona fide de otras especies vegetales (*Piper nigrum*, *Malus domestica* y *Coffea arabica*) en morado. Este esquema permite identificar visualmente la distribución de los miembros de la Especie B en relación con las secuencias de referencia.

Posteriormente, se anotaron las subfamilias TPS mediante la herramienta "Colored ranges", aplicando un sombreado de fondo a los clados correspondientes. Se distinguieron tres subfamilias principales:

- Subfamilia TPS-b (fondo amarillo), identificada por la agrupación de TPS1–TPS4 de *Arabidopsis*.
- Subfamilia TPS-a (fondo naranja), correspondiente al clado que incluye TPS5–TPS11 y otras TPS relacionadas.
- Subfamilia TPS-e/f (fondo gris), asociada al clado que contiene KSA y KSB.

Con el fin de resaltar las relaciones evolutivas específicas entre especies, se marcaron en rojo las ramas correspondientes a clados que contienen secuencias de *Arabidopsis* y una o pocas secuencias de *Cannabis*, interpretados como grupos ortólogos probables. Además, los subclados formados exclusivamente por múltiples secuencias de *Cannabis* fueron colapsados y representados mediante un triángulo gris, indicando expansiones génicas específicas del linaje.

Finalmente, se incorporó una leyenda explicativa que resume el código visual utilizado (colores por especie, sombreado por subfamilia, ramas rojas para ortología y triángulo para expansiones), garantizando la interpretación autónoma de la figura.

5. Interpretación filogenética y asignación de subfamilias en *Cannabis sativa*

El análisis filogenético revela que los nuevos miembros identificados en *Cannabis sativa* se distribuyen en varias subfamilias de la familia TPS, determinadas a partir de su agrupación con secuencias bona fide y TPS caracterizadas de *Arabidopsis thaliana*.

Asignación a subfamilias

El clado sombreado en amarillo (TPS-b), que incluye TPS1, TPS2, TPS3 y TPS4 de *Arabidopsis*, agrupa múltiples secuencias de *Cannabis*, lo que indica que una parte importante de los miembros de la Especie B pertenece a esta subfamilia. Dado que TPS-b está asociada principalmente a monoterpeno sintasas plastidiales, estos miembros de *Cannabis* podrían desempeñar funciones similares en la biosíntesis de monoterpenos volátiles.

El clado sombreado en naranja (TPS-a), que contiene TPS5–TPS11 y otras TPS relacionadas de *Arabidopsis*, también incluye numerosas secuencias de *Cannabis*. Esto indica que la Especie B posee múltiples representantes de la subfamilia TPS-a, generalmente asociada a sesquiterpeno sintasas citosólicas.

Asimismo, se identificó un clado correspondiente a TPS-e/f (fondo gris), asociado a enzimas como KSA y KSB, en el que también se detectan secuencias de *Cannabis*, aunque en menor número. Esto sugiere que la Especie B conserva representantes de subfamilias más divergentes dentro de la familia TPS.

Identificación de ortólogos probables

El árbol anotado permite identificar varios grupos ortólogos probables entre *Cannabis sativa* y *Arabidopsis thaliana*. Estos se reconocen como subclados bien definidos que contienen una TPS específica de *Arabidopsis* junto con una o pocas secuencias de *Cannabis*, marcados en rojo en la figura.

Por ejemplo, se observan grupos ortólogos asociados a TPS individuales dentro de las subfamilias TPS-a y TPS-b, lo que indica relaciones de ortología directa o casi directa entre genes de ambas especies. En algunos casos, una TPS de *Arabidopsis* se agrupa con dos secuencias de *Cannabis*, lo que sugiere un evento de duplicación génica específico del linaje *Cannabis* posterior a la divergencia evolutiva.

Por otro lado, la presencia de subclados colapsados formados exclusivamente por secuencias de *Cannabis* indica expansiones génicas específicas del linaje, probablemente debidas a duplicaciones recientes.

6. Intereses personales

La elección de *Cannabis sativa* como especie B se basa tanto en su relevancia biológica como en su interés científico y terapéutico actual. Esta especie se caracteriza por producir una amplia variedad de compuestos terpénicos que

contribuyen de manera significativa a su perfil aromático y a sus funciones ecológicas, como la defensa frente a herbívoros y patógenos. La elevada diversidad de terpenos sugiere una posible expansión y diversificación de la familia de las terpeno sintasas (TPS), lo que la convierte en un modelo interesante para estudiar la evolución funcional de esta familia génica en plantas.

Además de su importancia biológica, *Cannabis sativa* posee un notable interés aplicado. En los últimos años, sus metabolitos secundarios, incluidos terpenos y cannabinoides, han sido ampliamente estudiados por su potencial terapéutico en distintos contextos clínicos, como el manejo del dolor, trastornos neurológicos y enfermedades inflamatorias. Desde una perspectiva personal, el estudio de esta especie resulta especialmente atractivo debido al creciente interés en su uso medicinal y a la necesidad de comprender mejor la base genética que sustenta su diversidad química.

Por tanto, analizar la familia TPS en *Cannabis sativa* permite no solo abordar cuestiones evolutivas comparativas con especies modelo como *Arabidopsis thaliana*, sino también contribuir al conocimiento de una especie con gran relevancia biomédica y biotecnológica.