Anotación funcional de genes en kinetoplástidos mediante búsqueda de homología remota basada en estructura

Mag. Juan Manuel Trinidad Barnech

Doctorado en ciencias biológicas – PEDECIBA

Universidad de la República

Anotación funcional de genes en kinetoplástidos mediante búsqueda de homología remota basada en estructura

Mag. Juan Manuel Trinidad Barnech

Tesis presentada con el objetivo de obtener el título de Doctor en ciencias biológicas en el marco del Programa PEDECIBA.

Tutor: Dr. Pablo Smircich, Profesor Adjunto (FCien) e Investigador Nivel III (IIBCE).

Co-tutor: Dr. Darío Fernández do Porto, Investigador Asistente (UBA).

Resumen en español

Los kinetoplástidos pertenecen al supergrupo Discoba, un linaje eucariota de divergencia temprana. Aunque la cantidad de datos genómicos sobre estos parásitos ha aumentado de manera notable, asignar funciones génicas mediante métodos de homología basados únicamente en secuencias sigue siendo un reto. Recientemente, se han logrado avances importantes en la predicción *in silico* de estructuras proteicas y en el desarrollo de algoritmos capaces de comparar de forma rápida y precisa grandes volúmenes de estructuras. En este trabajo, diseñamos y aplicamos un flujo de trabajo de búsqueda de homología basado en la comparación de estructuras proteicas (ASC, Annotation by Structural Comparisons) para transferir información biológica a todas las proteínas de kinetoplástidos disponibles en TriTrypDB, la base de datos de referencia para este grupo. Gracias a nuestro pipeline, pudimos asignar similitudes estructurales a una fracción sustancial de las proteínas estudiadas, enriqueciendo las anotaciones existentes mediante la transferencia de información. Además, identificamos homólogos estructurales para representativos de 6 700 proteínas no caracterizadas en 33 especies de kinetoplástidos, proteínas que no podían anotarse usando las herramientas y bases de datos de homología de secuencia convencionales. De este modo, nuestro enfoque permitió inferir información biológica potencial para un gran número de proteínas. Entre ellas, detectamos homólogos estructurales de proteínas eucariotas omnipresentes que resultan difíciles de reconocer en los genomas de kinetoplástidos mediante los métodos de anotación genómica estándar. Los resultados de este trabajo, agrupados bajo el nombre KASC (Kinetoplastid Annotation by Structural Comparison), están disponibles de forma abierta para la comunidad en kasc.fcien.edu.uy, a través de una interfaz amigable que permite la inspección visual de los datos de manera individual por gen.

Resumen en inglés (Abstract)

Kinetoplastids belong to the Discoba supergroup, an early divergent eukaryotic clade. Although the amount of genomic information on these parasites has grown substantially, assigning gene functions through traditional sequence-based homology methods

remains challenging. Recently, significant advancements have been made in in-silico protein structure prediction and algorithms for rapid and precise large-scale protein structure comparisons. In this work, we developed a protein structure-based homology search pipeline (ASC, Annotation by Structural Comparisons) and applied it to transfer biological information to all kinetoplastid proteins available in TriTrypDB, the reference database for this lineage. Our pipeline enabled the assignment of structural similarity to a substantial portion of kinetoplastid proteins, improving current knowledge through annotation transfer. Additionally, we identified structural homologs for representatives of 6,700 uncharacterized proteins across 33 kinetoplastid species, proteins that could not be annotated using existing sequence-based tools and databases. As a result, this approach allowed us to infer potential biological information for a considerable number of kinetoplastid proteins. Among these, we identified structural homologs to ubiquitous eukaryotic proteins that are challenging to detect in kinetoplastid genomes through standard genome annotation pipelines. The results (KASC, Kinetoplastid Annotation by Structural Comparison) are openly accessible to the community at kasc.fcien.edu.uy through a user-friendly, gene-by-gene interface that enables visual inspection of the data.

Palabras clave en español y en inglés

anotación genómica, homología remota, Kinetoplástidos, Tripanosoma, Leishmania, estructura de proteína, AlphaFold.

Genomic annotation, remote homology, Kinetoplastids, Trypanosome, Leishmania, protein structure, AlphaFold

Tabla de contenido

Introducción

Clasificación taxonómica

Los kinetoplástidos son parásitos protozoarios flagelados pertenecientes al supergrupo Discoba, filo Euglenozoa, clase Kinetoplastea. Los Discoba comprenden un grupo de divergencia temprana en el árbol evolutivo de los eucariotas [1–3]. Al mismo tiempo, los organismos de este gran grupo tienen una alta tasa de divergencia y estilos de vida [4]. Consecuentemente, estos organismos presentan procesos moleculares distintivos que no están presentes en otros grupos eucariotas que describiremos más adelante [5,6]. Dentro de Discoba la clase Kinetoplastea es una de las más estudiadas, e incluye organismos que parasitan una amplia variedad de hospedadores, como vertebrados, artrópodos, sanguijuelas, plantas y ciliados [7]. Los datos de filogenias moleculares y de genómica comparativa muestran que los kinetoplástidos provienen de un ancestro común de vida libre y que el parasitismo en los tripanosomátidos tuvo un único origen. Los parientes vivos más cercanos como *Bodo saltans* o dilponemidos, son bacteriófagos de vida libre que habitan la microbiota terrestre y/o de agua dulce [8] (**Figura 1**). A los kinetoplástidos se los conoce especialmente porque algunos de sus miembros provocan enfermedades humanas, tales como la enfermedad del sueño (*Trypanosoma brucei*), la enfermedad de Chagas (*Trypanosoma cruzi*) y diversas formas de leishmaniasis (*Leishmania spp.*). Según su ciclo de vida, los kinetoplástidos se dividen en dos grupos no taxonómicos: monoxénicos (con un único hospedador) y dixénicos (también conocido como “ciclo heteroxeno”, que alternan entre dos hospedadores, uno de los cuales actúa como vector). Los insectos son los hospedadores predominantes de los tripanosomátidos monoxénicos y, a la vez, los vectores más frecuentes de los dixénicos [7] (**Figura 1**). En el siglo XXI, los avances en el conocimiento de la biología de los tripanosomátidos se han vinculado estrechamente con el auge de los estudios genómicos. Para este grupo de protistas, la era genómica comenzó con el análisis de sus tres especies de mayor relevancia médica *T. brucei* [9], *T. cruzi* [10] y *Leishmania major* [11] (denominados TriTryps) para luego extenderse a un gran número de especies [4]. Se estima que en cada especie hay aproximadamente entre 10.000 y 15.000 genes codificantes para proteína (**Figura 1**) donde los estudios seminales identificaron 6.200

genes homólogos entre los TriTryps [12]. En base a homología o por dominios funcionales, se pudo asignar una función tentativa a aproximadamente el 50% de los genes codificante para proteína predichos [13].

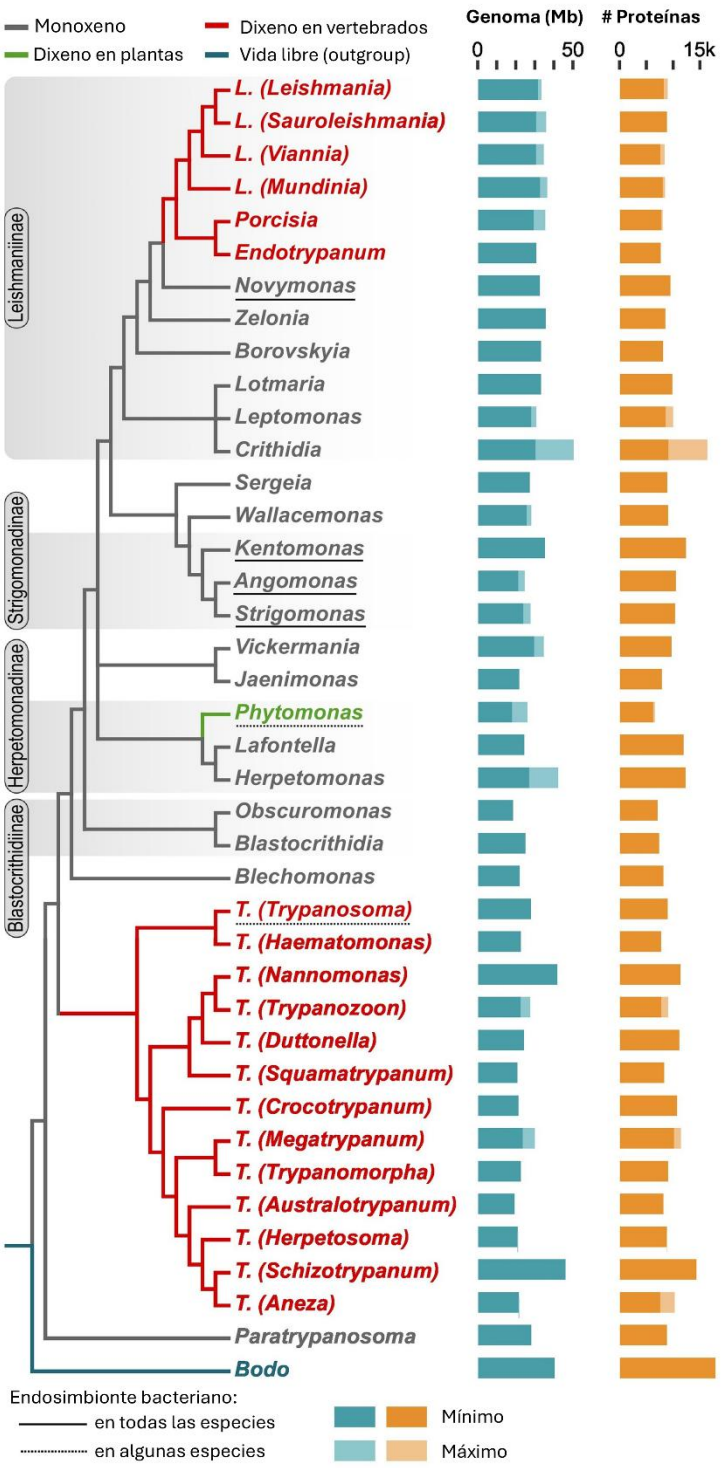

Figura 1. Árbol filogenómico de la familia Trypanosomatidae con información de características genómicas. El árbol se obtuvo combinando análisis independientes. Las ramas están coloreadas según su modo de vida. Las subfamilias se indican únicamente si contienen más de un género. Se indica el tamaño del genoma y el número de genes máximo y mínimo predicho para las especies a agrupadas. Tomado y modificado de Kostygov et. al. 2024.

Kinetoplástidos y sus particularidades genómicas

Los kinetoplástidos destacan por diferir significativamente de otros grupos de eucariotas, y por ello son un excelente ejemplo de la diversidad de este grupo. Poseen una serie de particularidades genómicas. Por ejemplo, la organización del ADN donde se observa una casi total ausencia de intrones [14] y la base J [15], una modificación única que desempeña un papel clave en la terminación de la transcripción. En cuanto a la expresión génica carecen de regulación transcripcional clásica, de modo que los genes se transcriben como policistrones; luego, el procesamiento de los pre-ARNm se realiza mediante *trans-splicing* generalizado, momento en que cada transcrito recibe su propio *splice-leader* y cola de poliA [16]. Por otro lado, la maquinaria mitocondrial presenta también características excepcionales. Sus ARNm mitocondriales sufren ediciones extensas, en las que se insertan y eliminan uridinas para producir moléculas funcionales [17,18]. Finalmente, en el plano bioquímico destacan otras innovaciones adaptativas: la localización de gran parte de la vía glucolítica dentro de los glicosomas y el uso de un sistema redox basado en tripanotiona, en lugar del glutatión convencional, lo que evidencia un metabolismo redox especializado [19].

Estas excepciones subrayan hasta qué punto los kinetoplástidos han desarrollado soluciones únicas a procesos celulares fundamentales, consolidándose como un modelo clave para comprender la amplitud de estrategias evolutivas en los eucariotas. A continuación, describiremos las características más relevantes del grupo.

Kinetoplasto

Una de las características más destacadas de Kinetoplastea es la posesión de una única pero muy desarrollada mitocondria que ocupa gran parte del citoplasma. Esta mitocondria conserva los compartimentos convencionales como membrana externa e interna, espacio intermembrana, matriz mitocondrial (entre otros) y su volumen, morfología y composición puede variar según el estadio y/o estado fisiológico del organismo [20].

El ADN mitocondrial es denominado kinetoplasto, del cual este grupo toma su nombre y representa la sinapomorfía del grupo. La cantidad de ADN mitocondrial varía entre especies representando generalmente el 25% del total del material genético celular, pero llegando a extremos como el 90% en Perkinsela [21]. Se organiza en una estructura distintiva, que alberga dos tipos de moléculas circulares denominadas en función de su tamaño como: maxicírculos y minicírculos. Estas moléculas de ADN circular altamente compactado se encuentran interconectadas formando una red y ancladas al cuerpo basal, ubicado en la base del flagelo y dispuesto perpendicularmente a su eje [22]. Los minicírculos, se encuentran en el orden de los miles de copias, miden cerca de 1 kb cada uno y codifican para ARNs guía, mientras que los maxicírculos, de los que sólo hay unas pocas decenas, alcanzan unos 25 kb de longitud [23,24]. Los maxicírculos, equivalentes al genoma mitocondrial de otros eucariotas, contienen genes mitocondriales típicos que codifican para ARNr y subunidades de la cadena de transporte de electrones [17,18].

Núcleo, organización génica y transcripción

El núcleo, a diferencia de la mitocondria, exhibe una arquitectura típica de eucariotas. Sin embargo, la arquitectura genómica de estos organismos presenta particularidades sin precedentes en eucariotas. La mayoría de los kinetoplástidos son parásitos y se encuentran clasificados dentro del Orden Trypanosomatida, siendo los géneros más conocidos *Trypanosoma* y *Leishmania* [4]. La transición al parasitismo, al igual que en otros eucariotas, suele asociarse a reducciones genómicas en este caso con una pérdida de aproximadamente el 50 % de los genes en comparación con *B. saltans*, un kinetoplástido de vida libre [25,26]. Sin embargo, hoy en día el tamaño de sus genomas es comparable con los kinetoplástidos de vida libre, lo que indica que los parásitos presentan menor densidad génica. Esto se explica en gran medida por expansiones de ADN no codificante en sus genomas. Del total de genes de los parásitos, alrededor del 90 % muestra homología con los de kinetoplástidos de vida libre, y el restante se atribuye a innovaciones vinculadas al parasitismo [26]. Estos resultados sugieren que, si bien ha habido una reducción constante en la complejidad de numerosas vías metabólicas (especialmente en catabolismo, degradación de macromoléculas y transporte de iones) la evolución del parasitismo no condujo a la pérdida generalizada de rutas metabólicas

en el Orden Trypanosomatida. Por tanto, rasgos eucariotas ausentes en estos parásitos, como la biosíntesis de purinas o un sistema redox basado en glutatión, representan características ancestrales de los kinetoplástidos y no pérdidas específicas en el orden Trypanosomatida [8,26]. La simplificación funcional de los genomas de los Trypanosomatida se explica también por la pérdida de diversidad en ciertas familias génicas. Entre los descensos más marcados se encuentran las proteasas de cisteína tipo catepsina, las lipasas, los canales iónicos dependientes de voltaje y los transportadores de membrana tipo ABC [26]. Además, se han observado pérdidas génicas independientes en géneros como *Trypanosoma* y *Leishmania*, lo que indica que la reducción genómica continuó durante su diversificación y dio lugar a un reparto asimétrico de repertorios génicos ancestrales entre las distintas líneas parasitarias [26].

Al igual que el reparto asimétrico del repertorio ancestral afecta a diversas familias génicas multicopia, es evidente que también afectan a elementos transponibles. El genoma de *B. saltans* contiene todos los elementos transponibles identificados previamente en los tripanosomátidos, como: ingi (retrotransposón presente en todas las especies), SLACS/CZAR (retrotransposón hallado en la mayoría de las especies), VIPER (retrotransposón exclusivo de *Trypanosoma*) y TATE (presente sólo en *Leishmania*). Por tanto, en varios aspectos, los genomas de *Trypanosoma spp.* y *Leishmania spp.* son muestras independientes de un repertorio génico ancestral más amplio [26].

No obstante, la abundancia de clústeres génicos exclusivos de uno o varios parásitos demuestra que la evolución del parasitismo implicó algo más que pérdidas: la ganancia de nuevos genes también desempeñó un papel clave. Estos clústeres específicos suelen incluir genes codificantes de proteína de superficie celular y confirman que la divergencia genómica de los trypanosomatidos estuvo dominada por la rápida expansión de familias génicas multicopia [12].

La comparación de los genomas de los TriTryps mostró que tanto el orden de los genes como su repertorio están, en gran medida, conservados [12]. A pesar de las divergencias antiguas entre las especies, esta colinealidad se interpreta como el resultado de fuertes restricciones selectivas sobre la estructura genómica [12]. El análisis comparativo de la conservación del orden génico en eucariotas revela que los pares de genes altamente conservados suelen mantenerse juntos por razones funcionales y de regulación

transcripcional [27]. Sin embargo, en los tripanosomátidos, no hay evidencia clara de que la proximidad conserve funciones compartidas [28]. Estos extensos tramos contiguos de genes codificantes situados sobre la misma hebra de ADN son denominados grupos direccionales de genes o DGCs (*Directional Gene Clusters*, por sus siglas en inglés) (**Figura 2**). Las DGCs se transcriben generalmente a partir de un único promotor pudiendo haber unos pocos promotores internos, generando largos transcritos policistrónicos, a pesar de carecer de las secuencias promotoras canónicas reconocidas por la ARN polimerasa II [29,30] . Cada DGC está flanqueado por regiones de cambio de hebra o SSRs (*Strand Switch Regions*, por sus siglas en inglés), que pueden ser divergentes (puntos desde los cuales comienzan dos DGCs en hebras opuestas) o convergentes (zonas donde terminan dos DGCs enfrentados). En los SSRs divergentes se localizan sitios de inicio de la transcripción o TSSs (*Transcription Start Sites,* por sus siglas en inglés), caracterizados por la acumulación de variantes y modificaciones de histonas y por señales intrínsecas de curvatura del ADN [31–33]. En los SSRs convergentes, además de marcar los puntos de terminación, es frecuente hallar genes de ARNt transcritos por la ARN polimerasa III [34].

El paso de estos policistrones a ARNm monocistrónicos maduros se lleva a cabo mediante un mecanismo único de los tripanosomátidos, que utiliza un pequeño transcrito de ~140 nt (el mini-exón o *splice-leader*) producido en tándem desde regiones con promotores reconocidos por la ARN polimerasa II [35]. Durante el procesamiento, cada gen del policistrón incorpora un fragmento de 38–40 nt de este mini-exón en su extremo 5’ mediante un corte y empalme entre moléculas distintas, un fenómeno descubierto en estos protozoos conocido como *trans-splicing* [36,37]. Puesto que los tripanosomátidos carecen prácticamente de intrones, el empalme *cis-splicing* es muy raro en ellos [38]. La adición del mini-exón aporta la “caperuza” 5’, o cap-4, con varios nucleósidos metilados en el carbono 2’ de la ribosa, exclusiva de estos organismos [39,40]. La poliadenilación del extremo 3’ del gen precedente se produce de manera coordinada con el *trans-splicing*, logrando así eliminar las secuencias intergénicas y generar ARNm monocistrónicos listos para exportarse al citoplasma (**Figura 2**) [41,42]. Las señales principales para dirigir el *trans-splicing* son un dinucleótido AG en el sitio 3’ de empalme y una región adyacente rica en pirimidinas [43], mientras que los sitios de

poliadenilación suelen encontrarse en motivos ricos en adeninas, separados en promedio por unos 40 nt en *T. cruzi* [44,45].

Debido a que la transcripción policistrónica depende de promotores de actividad esencialmente constitutiva (al menos para genes de copia única), los tripanosomátidos regulan la expresión génica principalmente a nivel post-transcripcional [46]. Entre los mecanismos de control se incluyen el procesamiento y maduración del transcrito, su exportación al citoplasma, la estabilidad y degradación diferencial, la movilización a los polisomas y la accesibilidad a la maquinaria de traducción.

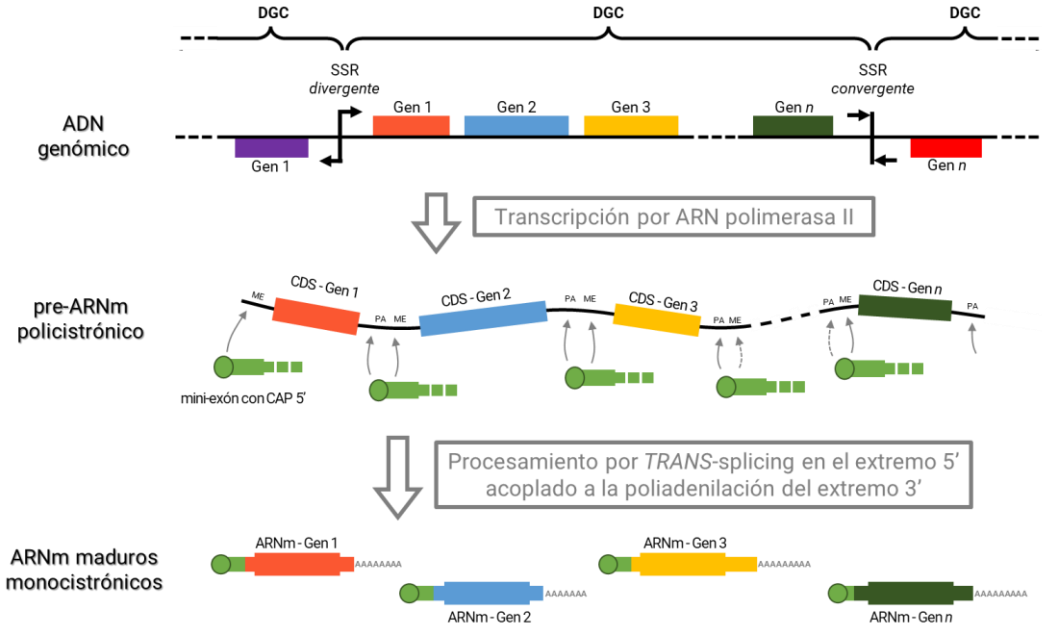

Figura 2. Particularidades de la expresión génica de los tripanosomátidos. Se esquematiza una región de ADN genómico y cómo actúan en la misma la transcripción y el posterior procesamiento de los transcritos primarios. En otra región genómica se expresa el ARN codificante para el mini-exón, que también es transcrito por la ARN polimerasa II. ME: mini-exón, PA: señal de poliadenilación, CDS: secuencia codificante, UTR: región no traducida, CAP: caperuza. Tomado de la tesina de doctorado de Santiago Chavez, 2022.

Anotación genómica

Hace ya unas décadas, la era de los grandes conjuntos de datos ha llegado a la biología molecular principalmente orquestada por las tecnologías de secuenciación masiva. Esto

ha resultado en una gran cantidad de datos en cuanto a secuencias de proteínas, expresión de genes, interacciones e incluso estructura de proteína. Debido a la aceleración de la secuenciación de genomas y metagenomas, hoy disponemos de millones de proteínas sin una anotación funcional. Entender la función de estas moléculas es esencial, pues las proteínas intervienen en prácticamente todos los procesos celulares desde el crecimiento y mantenimiento hasta la proliferación y la apoptosis, constituyendo un eje central de la biología celular [47]. Para cerrar la brecha entre la avalancha de datos y el conocimiento funcional, se han desarrollado dos grandes familias de herramientas.

En primer lugar, los métodos de laboratorio húmedo tradicionales como la deleción de genes (*gene knockout*), las mutaciones dirigidas y la inhibición de la expresión génica, entre otras [48]. Estos métodos hoy en día siguen siendo la referencia en cuanto a anotación y aunque proporcionan información directa sobre el papel de una proteína en la célula, son métodos de muy baja capacidad, pues requieren un esfuerzo considerable y sólo permiten afrontar generalmente un gen o proteína a la vez.

En segundo lugar, las aproximaciones computacionales para la predicción de función proteica han demostrado ser esenciales para escalar el análisis a miles o millones de proteínas. Dentro de estas, los métodos basados en la similitud de secuencias fueron pioneros y todavía constituyen la primera línea de anotación automática. Algoritmos como BLAST permiten comparar rápidamente dos secuencias completas y transferir anotaciones cuando existe elevado grado de identidad [49]. Para superar las limitaciones en la "*twilight zone*" [50] (generalmente por debajo del 25-30 % de identidad de secuencia) se han desarrollado enfoques como PSI-BLAST y modelos ocultos de Markov o HMM (*Hidden Markov Models*, por sus siglas en inglés) (y derivados), capaces de detectar homología remota más allá de los umbrales convencionales [51]. En nuestro grupo, se han utilizado abordajes basados en comparación de perfiles de HMM para la anotación genómica de kinetoplástidos cuyos resultados se encuentran documentados en la tesis de maestría de Santiago Radío (hdl.handle.net/20.500.12008/28615) y disponibles en bases de datos públicas (github.com/sradiouy/DARK). Todos estos abordajes tienen una región de certeza, basada en los parámetros estadísticos de los resultados, en los cuales la anotación es fácilmente transferible, pero cuando el

resultado no está dentro de este espacio se debe recurrir a inspección manual o alternativas [49].

Idealmente se busca que los resultados de alineamiento u otros métodos cubran un porcentaje considerable de la proteína de interés. Pero muchas veces este no es el caso y los resultados únicamente detectan homología en un subconjunto de los aminoácidos. Consecuentemente, algunas estrategias se basan en la identificación de dominios o motivos conservados. Estos se definen como unidades evolutiva y funcionalmente independientes de entre 30 y 150 aminoácidos que se pliega independientemente del resto de la proteína. Dado que al menos dos tercios de las proteínas de mamíferos contienen múltiples dominios, esta aproximación resulta muy eficaz para caracterizar proteínas multidominio que desempeñan funciones diversas [51]. Existen bases de datos especializadas como CATH, SCOP, CDD, Pfam, entre otras en las que se catalogan miles de dominios estructurales, facilitando su detección automática en nuevas secuencias [52].

Por último, gracias a los avances recientes en inteligencia artificial se han desarrollado herramientas basadas en estos principios para la anotación funcional de genes. Entre ellas, destaca ProtNLM (Protein Natural Language Model), desarrollado por Google Research e integrado en la base de datos UniProt. Este modelo, basado en la arquitectura T5, recibe como entrada la secuencia de aminoácidos de una proteína y genera descripciones funcionales en lenguaje natural, devolviendo varias opciones ordenadas según su grado de confianza. ProtNLM fue entrenado mediante el emparejamiento de secuencias provenientes de UniProt con frases en inglés usadas para describir sus funciones, lo que le permite predecir descripciones incluso para proteínas cuya función no ha sido determinada experimentalmente. Además, este sistema ya ha sido aplicado en la anotación automática en UniProt, generando millones de descripciones para proteínas previamente no caracterizadas [53].

Revolución en la biología estructural y relevancia en la detección de homología

El interés de predecir estructuras proteicas a partir de secuencias de aminoácidos surgió poco después de determinarse las primeras estructuras experimentales, pero hicieron falta décadas y la convergencia de numerosas herramientas y bases de datos para alcanzar métodos generales y precisos. El interés era tan intenso que incluso se crearon competencias como CASP, en las que participaban tanto académicos como investigadores de la industria privada. Fue en CASP14 donde la biología estructural vivió un punto de inflexión gracias al desempeño del algoritmo AlphaFold [54]. Este modelo predice la conformación tridimensional de una proteína a partir de su única secuencia de aminoácidos. En esencia, AlphaFold usa una base de datos masiva de secuencias para buscar homologos a la del gen de interés y construir un alineamiento múltiple de secuencias (MSA). A partir del MSA detecta correlaciones entre residuos que sugieren proximidad espacial, las transforma en predicciones de distancias y ángulos, y luego refina el modelo con una red neuronal que arroja, además de la estructura final, métricas de confianza como el pLDDT para evaluar su fiabilidad [54]. Esta innovación, galardonada con el Premio Nobel de Química en 2024, democratizó el acceso a estructuras proteicas y permitió la creación de bases de datos como AlphaFold Data Base (AFDB), que ya reúne más de 200 millones de modelos.

Otro algoritmo destacado es ESMFold, que adopta un enfoque radicalmente distinto: prescinde del MSA y se apoya en grandes modelos de lenguaje (LLM) entrenados sobre secuencias proteicas. Esta estrategia es especialmente útil para organismos muy divergentes o para aquellos de los que carecemos de especies cercanas de referencia, como sucede frecuentemente en metagenómica. Aunque ESMFold es capaz de generar una estructura solo a partir de una secuencia de aminoácidos (una característica única) sus modelos suelen ser algo menos fiables que los de AlphaFold. Por ello, se recurre a él cuando no existe un MSA informativo o cuando es necesario modelar un gran volumen de secuencias gracias a su notable rapidez [54].

El auge de estos algoritmos y la proliferación de bases de datos masivas planteó otro desafío, cómo interrogar con rapidez y eficiencia toda esa información estructural. En

sus inicios no había métodos suficientemente veloces para buscar homología a nivel de estructura tridimensional. En ese contexto nació Foldseek [55], un nuevo algoritmo que convierte la estructura terciaria de las proteínas en un nuevo alfabeto denominado 3Di (*3D interaction*) de 20 estados, transformando la búsqueda en un problema de alineamiento de secuencias unidimensionales. Cada estado describe para cada residuo *i* la conformación geométrica con su residuo espacialmente más cercano *j*. Para describir la geometría de interacción de los residuos *i* y *j*, se usan 20 estados discretos, que se generan a partir de descriptores de vecindad que contienen 10 características (distancias y ángulos entre Cα) que resumen la conformación del *backbone* de la proteína alrededor de los residuos *i* y *j*. De este modo, conserva las características esenciales de la geometría proteica, pero simplifica el proceso a una tarea de similitud de secuencias, aprovechando algoritmos de alineamiento ya optimizados para este fin [55].

En resumen, la estructura de proteínas se posiciona como una alternativa prometedora frente a los enfoques tradicionales basados únicamente en secuencia, especialmente para identificar homología remota en genes hipotéticos de kinetoplástidos. Estrategias análogas en otros organismos han demostrado capacidad para revelar funciones previamente desapercibidas, otorgando solidez a la viabilidad de nuestro enfoque. Dado que los datos estructurales predichos por inteligencia artificial están hoy disponibles de forma masiva en la AFDB con más de 200 millones de modelos estructurales de proteínas, contamos con una base sólida y actualizada para realizar comparaciones estructurales profundas. Aprovechando la disponibilidad tanto de modelos para nuestros organismos de interés como de proteomas de referencia, proponemos desarrollar un pipeline computacional innovador que integre alineamientos de estructuras para detectar relaciones evolutivas remotas, fortalecer las anotaciones existentes en TriTrypDB y transferir sistemáticamente nueva información funcional a los genomas de kinetoplástidos.

Objetivo general:

Anotar los genomas de kinetoplástidos disponibles en TriTrypDB mediante la comparación de estructura de proteínas.

Objetivos específicos

1. Desarrollar un software que permita la búsqueda de homólogos de proteínas mediante alineamientos de estructura.
2. Identificar homología remota entre proteínas hipotéticas de kinetoplástidos y la base de datos.
3. Trasladar datos de anotación a proteínas de kinetoplástidos.

Materiales y métodos

1. Desarrollo del flujo de trabajo

En nuestra implementación, los pasos que van desde la agrupación de secuencias con MMseq2 [56], hasta la detección de homólogos estructurales con Foldseek [55] y FATCAT [57], se orquestaron utilizando el organizador de flujo de trabajo Snakemake [58] (**Figura 4**). Cada regla declara claramente sus entradas, salidas y dependencias, lo que nos permitió paralelizar tareas y garantizar la integridad de los datos en cada paso. Snakemake facilita el uso de entornos con Conda integrados en el propio Snakefile asegura que todas las herramientas bioinformáticas utilizadas se instalen con versiones controladas, facilitando la portabilidad. El código completo y los resultados finales se encuentran disponibles en GitHub, donde se incluye un README detallado, ejemplos de ejecución y los perfiles de configuración para ejecutar el flujo de trabajo. Siendo que el desarrollo de un flujo de trabajo es el Objetivo específico 1 de esta tesis los detalles serán abordados en la sección "Resultados y discusión".

2. Obtención de secuencias y agrupamiento

Se descargaron manualmente todas las secuencias proteicas predichas en genomas de kinetoplástidos y sus anotaciones correspondientes de la base de datos TriTrypDB (versión 65) (fecha de acceso: agosto de 2024) (**Tabla S1**) [59]. Clasificamos como proteínas hipotéticas aquellas cuya descripción en TriTrypDB contenía los términos "hypothetical protein", "Hypothetical protein" o "unspecified product". Esto resulta en una clasificación binaria que a lo largo del texto referiremos como "proteínas hipotéticas" (PHip) y "proteínas anotadas" (PAnot).

Con el objetivo de reducir redundancia, las secuencias de proteínas se agruparon mediante MMseq2 [56] (modo sensible `mmseqs cluster`, parámetros `--cluster-mode 1 --similarity-type 2 --min-seq-id 0.5 -c 0.8 --cov-mode 0 -e 1e-5`), considerando homologías cuando la identidad de

secuencia es ≥ 50 % y cobertura en ambos sentidos es ≥ 80 %. Se descartaron grupos con menos de diez miembros para evitar artefactos de anotación.

3. Obtención de datos de estructura de proteínas

Se obtuvieron mediante File Transfer Protocol (FTP) todas las estructuras disponibles para los miembros de los agrupamientos descriptos anteriormente en AFDB utilizando como referencia la asignación de PDB por cada identificador de TriTrypDB disponible en TriTrypDB. Se verificó la concordancia entre la secuencia del PDB y la de TriTrypDB; discrepancias superiores al 20% de identidad de secuencia en un 80% de la cobertura de alineamiento para ambos lados llevaron a la exclusión del modelo. Este último paso surge de haber detectado errores en la asignación de PDB mediante el identificador de TriTrypDB. Cuando un grupo contiene múltiples predicciones estructurales, se calculó el promedio de pLDDT de cada modelo y se seleccionó como representante aquel con mayor valor promedio.

Se descargaron los proteomas de 45 organismos de referencia (OrgRef) representativos de diversos linajes eucariotas desde "AFDB model organism proteomes" (excluyendo kinetoplástidos) (**Tabla S2**). Cada uno de estos se preparó como base de datos independiente para las comparaciones posteriores.

4. Comparación estructural

La comparación estructural se llevó a cabo utilizando el software Foldseek (versión 8.ef4e960) [55]. Se generó una base de datos de consulta la cual está integrada por la estructura mejor predicha (promedio de pLDDT más alto) de cada clúster (generados por MMseqs2) de los kinetoplástidos. Por otro lado, se generó una base de datos por cada proteoma de OrgRef sumando un total de 45 bases de datos "target".

Todas las bases de datos fueron generadas empleando `foldseek createdb`. Posteriormente se ejecutaron las búsquedas de "reciprocal best hits" con `foldseek`

`rbh -s 9.5 -c 0 -a` la cual abreviaremos como SRBH (*structural reciprocal best hit,* por sus siglas en inglés).

Cada gen representante de cluster de kinetoplástido fue sometido a 45 SRBH de los cuales se eligieron los cinco mejores aciertos (menor e-value de foldseek). El archivo PDB del kinetoplástido y los cinco mejores aciertos provenientes de los OrgRefs se sometieron a alineamiento estructural flexible con FATCAT [57] (parámetros por defecto), que permite ajustar dominios con posibles reorientaciones. Del resultado se extrajeron los valores de TM-score mediante el algoritmo TM-align [60] de la comparación de la Cadena1 versus Cadena2 y viceversa. Siendo cadena 1 y 2 los archivos de PDB correspondientes a los genes comparados. Los alineamientos estructurales fueron visualizados con el software ChimeraX [61].

5. Obtención de datos de anotación de UniProt

Para obtener la información de anotación funcional de las proteínas, se implementaron reglas automatizadas en el pipeline que descargan datos directamente desde UniProt utilizando su servicio REST. Se obtuvo la anotación de todas las proteínas clasificadas bajo el taxón Kinetoplastea (taxonomy_id:5653), y en el caso de los organismos de referencia se usó el código del proteoma.

6. Validación *in silico* de anotaciones inferidas

Para evaluar la precisión de los resultados de SRBH, decidimos comparar el nivel de consistencia entre los dominios proteicos y clasificación a nivel de familia génica de la proteína "*query*" y "*target*". El objetivo es ver cuan consistente es la clasificación o la arquitectura proteica entre los SRBH.

Para esto obtuvimos de UniProt las anotaciones de las bases de datos Protein Families, PANTHER, InterPro y Pfam tanto para las proteínas de kinetoplástidos como para sus contrapartes de OrgRef. Cada par SRBH se clasificó en categorías de concordancia

("TODO", "K_in_RO", "RO_in_K", "PARCIAL", "CERO") y se sometió a prueba de independencia Chi-cuadrado con SciPy**.**

Categorías:

- TODO: todas las anotaciones de la proteína de kinetoplástidos coinciden con las anotaciones de la proteína de los OrgRef (K = OR).
- K_in_OR (kinetoplástidos en OrgRef): todas las anotaciones de la proteína de kinetoplástidos están contenidas en las anotaciones de la proteína del OrgRef, pero esta última tiene anotaciones adicionales (K ⊆ OrgRef).
- OR_in_K (OrgRef en kinetoplástidos): todas las anotaciones de la proteína del OrgRef están contenidas en las anotaciones de la proteína de kinetoplástidos, pero esta última tiene anotaciones adicionales.
- PARCIAL: algunas, pero no todas, las anotaciones coinciden entre ambas proteínas.
- CERO: no hay anotaciones compartidas entre las dos proteínas.

7. Validación con datos experimentales de localización subcelular con TrypTag

Los resultados de SRBH fueron también evaluados contrastando con datos experimentales. Con este objetivo comparamos las anotaciones de localización celular inferidas por homología estructural versus los obtenidos en TrypTag [62], una base de datos de imágenes de localización subcelular a nivel genómico de *T. brucei*.

Por un lado, extrajimos los datos de localización subcelular de TrypTag disponibles en TriTrypDB (versión 65). Consideramos nueve categorías ubicuas de localización subcelular para evaluar: citoplasma, nucleoplasma, nucléolo, cuerpo basal, retículo endoplásmico, mitocondria, envoltura nuclear, núcleo y aparato de Golgi. De los datos de TrypTag solo se incluyeron en el análisis anotaciones con una confianza de señal mayor al 75 % al mismo tiempo que se excluyeron etiquetas como "weak". Estos datos van a ser sometidos a dos evaluaciones. La primera, a modo de control, es evaluar cuan precisa es la anotación actual de los genes de *T. brucei* por métodos *in silico* con respecto

a los datos experimentales en *T. brucei*. El resultado permite establecer el estado de precisión de los métodos automáticos utilizados por las bases de datos, específicamente el de UniProt/InterPro. Para segunda evaluación, compararemos nuestros resultados *in silico* de SRBH contra los datos experimentales de TrypTag. Los resultados de ambas comparaciones cotejaron entre sí. Estos son los resultados obtenidos de SRBH de las cuales utilizaremos los datos de anotación de los OR los cuales provienen de UniProt/InterPro.

En ambos casos los datos de TrypTag serán comparados con las anotaciones de localización subcelular (términos GO en la categoría "Componente Celular", GO CC) de UniProt/InterPro. Dado que para un gen individual puede haber más de un GO CC se evalúan todos y se elige para incluir en la distribución el más cercano semánticamente. El valor numérico de la comparación es el resultado de calcular la similitud semántica entre los términos GO utilizando la herramienta GOGO. La similitud semántica cuantifica qué tan cercanos son los términos GO dentro del grafo GO.

La comparación entre las distribuciones fue evaluada mediante la prueba U de Mann–Whitney utilizando la librería SciPy [63] de Python.

8. Análisis BUSCO

Con el objetivo de identificar genes relevantes para la biología de un organismo eucariota que no han sido anotados en los genomas de kinetoplástidos, utilizamos BUSCO (*Benchmarking Universal Single-Copy Orthologs*) [64]. Ejecutamos BUSCO utilizando la base de datos `eukaryota_odb10`, que contiene ortólogos de copia única universales de eucariotas, sobre los conjuntos de proteínas de TriTrypDB (`-m prot`). Aquellos grupos BUSCO que no fueron encontrados en ningún genoma de kinetoplástidos fueron seleccionados para evaluar si nuestro flujo de trabajo lograba identificar alguno. A continuación, determinamos si cada una de estas proteínas ausentes tenía un homólogo estructural significativo en nuestros resultados de SRBH.

9. Anotación genómica con eggNOG e InterProScan

Se utilizo la versión eggNOG-mapper 2.1.12 [65] en el servidor web con parámetros por default e InterProScan versión 5.67-99.0 [66] corrido en servidor local.

Resultados y discusión

Desarrollo de flujo de trabajo utilizando Snakemake

En esta tesis adoptamos Snakemake como sistema de gestión de flujos de trabajo para garantizar que cada etapa del análisis fuera reproducible, escalable y fácil de distribuir. Snakemake permite describir los pasos del análisis mediante un lenguaje basado en Python que resulta legible para el usuario y que, al mismo tiempo, incorpora toda la potencia de un entorno de programación completo. Gracias a esta sintaxis declarativa, cada regla define explícitamente sus entradas, salidas y dependencias, lo que simplifica la incorporación de nuevos pasos o la modificación de parámetros sin afectar al resto del flujo de trabajo. Uno de los principales beneficios de emplear Snakemake radica en su capacidad para escalar automáticamente desde un equipo de escritorio hasta servidores, clústeres de cómputo o la nube, sin necesidad de alterar la definición inicial del flujo de trabajo. Este escalado se complementa con la instalación automática de dependencias de software mediante Conda o contenedores (ej. Docker), garantizando que cada regla se ejecute en un entorno controlado con versiones fijas de las herramientas empleadas. En este caso decantamos por la utilización de entornos de Conda. Además, Snakemake facilita la paralelización automática de tareas, calcula internamente qué reglas son independientes y las ejecuta simultáneamente, optimizando así el uso de recursos de CPU y acelerando los tiempos de cómputo en comparación con un enfoque secuencial tradicional. La modularidad inherente a las reglas, unida al uso de *wildcards* y archivos de configuración en formato YAML, permite configurar parámetros internos y rutas de entrada (entre otros) de forma centralizada, lo que favorece la reutilización y adaptación del flujo de trabajo a distintos proyectos.

Para fomentar la lectura, distribución y reproducibilidad de nuestro trabajo seguimos las *best practices* sugerida por los desarrolladores de Snakemake. A modo de resumen, dentro de la carpeta raíz del proyecto, conviene crear dos subdirectorios principales: uno llamado `workflow`, donde residirá todo el código del flujo de trabajo, y otro llamado `config`, dedicado a los archivos de configuración. Este Snakefile puede mantenerse compacto si, de manera opcional, se fragmenta en módulos: dentro de

`workflow/rules`, cada archivo de reglas debe llevar la extensión `.smk` para facilitar su reconocimiento, mientras que los scripts auxiliares (por ejemplo, en Python o R) se ubican en `workflow/scripts` y los cuadernos interactivos en `workflow/notebooks` (**Figura 3**).

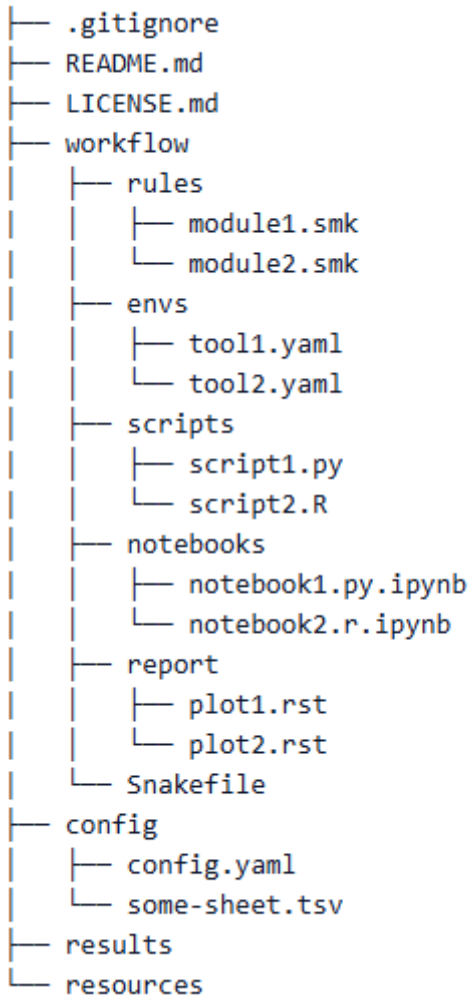

Figura 3. Estructura de directorios y archivos sugerida por los desarrolladores.

En el marco de este trabajo de tesis, se desarrolló un pipeline automatizado para la anotación de genomas basado en la estructura de proteínas. El pipeline fue diseñado para integrar diversas herramientas bioinformáticas, permitiendo ampliar la anotación de proteínas a partir de datos genómicos, con un enfoque centrado en la comparación estructural y la transferencia de información desde organismos de referencia. A

continuación, se describe conceptualmente la arquitectura del pipeline, las reglas que lo componen, los programas empleados en cada etapa y la lógica subyacente a su diseño. Los detalles algorítmicos de cada paso están documentados en Materiales y métodos (M&M) y serán referenciados a lo largo del texto. Los módulos más relevantes se esquematizan en la **Figura 4**.

El archivo principal del flujo de trabajo o `snakefile`, denominado `snakemake_ASC.smk`, constituye el núcleo de la automatización y orquestación de los diferentes pasos del análisis. El pipeline está estructurado en cinco reglas principales, cada una de las cuales se encuentra implementada en archivos independientes incluidos mediante la directiva `include` de Snakemake. Estas reglas tienen como objetivo:

1. Generar ageupamientos de secuencias con MMseqs2.
2. Descargar, seleccionar y filtrar estructuras PDB.
3. Comparar estructura de proteína para grandes conjuntos de datos con Foldseek.
4. Comparar estructura de proteína con FATCAT.
5. Descargar e incorporar anotaciones desde UniProt.

La primera regla, implementada en el archivo `00_MMseq2_sequence_clustering.smk`, está dedicada al agrupamiento de secuencias proteicas utilizando el programa MMseqs2 (M&M 2). Utilizamos MMseqs2 por ser una herramienta ampliamente reconocida por su eficiencia y sensibilidad en la comparación de grandes conjuntos de secuencias de proteínas. Actualmente esta herramienta se ha convertido en el *gold standard* a la hora de reducir redundancia o agrupar secuencias [56,67,68].

En esta etapa, el pipeline toma como entrada un archivo FASTA que contiene todas las secuencias proteicas de interés y genera la base de datos pertinente. El objetivo principal de este paso es agrupar las secuencias en clústeres, lo que permite reducir la redundancia, seleccionar representantes de cada grupo y consecuentemente agilizar los análisis posteriores. Este paso también permite analizar secuencias que no tengan

disponible una estructura, siempre y cuando en el clúster haya al menos un modelo estructural. Utilizamos esta herramienta con fines de optimización de recursos computacionales y para definir grupos de genes homólogos. Es por eso que para armar los grupos o clústeres usamos 50% de identidad y 80% de cobertura tanto en la secuencia *query* como en la *target* (entre otros parámetros ver M&M 2). Estos parámetros exigentes están basados en estudios anteriores y buscan estar dentro de un espacio de similitud confiable para establecer homología y así transferir anotación del representante al resto de las proteínas del clúster [49,68].

Los parámetros de MMseqs2 como la identidad de secuencia, el porcentaje de cobertura, e-value, entre otros son ajustables por el usuario a través del `config file`, lo que facilita el uso del pipeline y no requiere experiencia en el uso de snakemake.

La segunda regla, `01_Downloading_Selecting_and_Filtering_PDBs.smk` orquesta todo el proceso de obtención, filtrado y organización de las estructuras proteicas necesarias para la anotación basada en estructura (M&M 3). Este archivo contiene varias subreglas, cada una de las cuales automatiza una etapa específica del flujo de trabajo, utilizando principalmente scripts en Python desarrollados para este proyecto. La regla comienza con la descarga automatizada de todos los modelos estructurales de proteínas de kinetoplástidos desde AFDB. Utiliza como insumos un archivo obligatorio, el cual indica el código de UniProt/AFDB que corresponde a cada identificado de TriTrypDB del fasta inicial, este archivo se encuentra disponible en VEuPathDB para sus distintas bases de datos. La regla continúa con el cálculo del valor medio de pLDDT para cada estructura descargada. El pLDDT es una métrica de confianza en la predicción estructural generada por AlphaFold. Esta regla ejecuta un script Python multiproceso que recorre todos los modelos descargados, extrae los valores promedio de pLDDT y genera un reporte tabulado. Dado que encontramos que TriTrypDB no tiene bien asignados muchos de los códigos de UniProt, la regla compara las secuencias originales de las proteínas con las secuencias derivadas de los archivos PDB descargados, para detectar posibles inconsistencias (ver M&M). Este control de calidad es esencial para evitar el uso de modelos estructurales que no correspondan exactamente a las secuencias de interés. El proceso se realiza mediante un script Python que compara

ambas fuentes mediante alineamiento de secuencia y si no cumple con 80% de cobertura e identidad el PDB es descartado. Luego se integra la información de pLDDT, los encabezados de UniProt, los clústeres de MMseqs2 y los resultados de la comparación de secuencias. El objetivo es generar un reporte que asocie cada clúster de MMseqs2 con las estructuras modeladas disponibles y sus métricas de calidad. Este reporte se utiliza para la selección de representantes estructurales, quedándose con el de mayor valor promedio de pLDDT.

La regla procede con la organización de las estructuras representativas de cada clúster de MMseqs2 para la creación de la base de datos estructural que será utilizada en los pasos de comparación y anotación. Utiliza como insumo el reporte generado anteriormente y ejecuta un script propio que copia y renombra los archivos seleccionados, asegurando que cada grupo esté representado por la estructura de mayor calidad disponible. Finalmente, identifica aquellos grupos homólogos que no cuentan con ninguna estructura modelada en AFDB y genera un archivo FASTA con las secuencias correspondientes. Este paso permite mantener un registro de las proteínas que quedan fuera del análisis estructural y facilita su priorización para futuros estudios de modelado o anotación alternativa.

La tercera regla corresponde al archivo `02_Foldseek_rules-SingleOrgApproach.smk` que está dedicado a la comparación estructural de proteínas utilizando Foldseek (M&M 4). Foldseek ha revolucionado el campo de la bioinformática estructural al permitir la comparación eficiente de grandes bases de datos de estructuras, superando las limitaciones de velocidad de los métodos predecesores convirtiéndose en el método de referencia para la interrogación de bases de datos estructurales.

Esta regla automatiza la preparación de bases de datos estructurales, la ejecución de búsquedas recíprocas entre la base de datos de kinetoplástidos y organismos de referencia, y la consolidación de los resultados obtenidos. Esta sección comienza creando una base de datos para Foldseek (que utilizaremos como *query*) a partir de las estructuras representativas de kinetoplástidos, utilizando la mejor estructura para cada clúster de secuencia. Utiliza como insumo los reportes de agrupamiento y los archivos de estructuras seleccionadas, y ejecuta el comando para generar todos los archivos

necesarios para búsquedas estructurales posteriores. Al mismo tiempo, la regla automatiza la descarga de los archivos de estructuras de los organismos de referencia desde la base de datos AlphaFoldDB. Cada archivo correspondiente a un organismo es utilizado para generar una base de datos Foldseek (*target*) independiente. Esto permite realizar búsquedas estructurales específicas y comparaciones uno a uno entre el organismo de interés y cada organismo de referencia. Estas comparaciones son realizadas usando el método *reciprocal best hit* implementado en Foldseek. Utilizando parámetros configurables de sensibilidad y cobertura (entre otros), esta regla identifica pares de proteínas que son mutuamente los mejores alineamientos estructurales entre ambos conjuntos, lo que constituye una fuerte evidencia de homología. Finalmente, la regla concatena todos los archivos TSV generados para cada organismo de referencia en un único archivo resumen.

La cuarta regla correspondiente al archivo `03_FATCAT.smk` está dedicada a la comparación estructural avanzada de proteínas utilizando FATCAT, una herramienta reconocida por su capacidad para detectar similitudes estructurales incluso en presencia de flexibilidad conformacional y reordenamientos en las proteínas (M&M 4). Esta capacidad es especialmente relevante en este abordaje debido a que las estructuras de AlphaFold no suelen tener una buena representación de topologías/orientaciones entre dominios [69]. En el contexto del pipeline, FATCAT se utiliza para validar y refinar los cinco mejores (configurable por el usuario) alineamientos generados por Foldseek, proporcionando una segunda capa de análisis estructural que incorpora la flexibilidad conformacional de las proteínas utilizando los archivos PDB y computar TM-score de manera clásica sobre el alineamiento resultante. Los resultados de FATCAT incluyen puntuaciones de similitud, alineamientos detallados y medidas de significancia estadística, que son integrados en la tabla final de anotaciones.

La regla comienza tomando como insumo los resultados de los alineamientos estructurales previos (obtenidos con Foldseek) y genera una lista de pares de proteínas que serán alineadas con FATCAT. Además, esta regla se encarga de descargar y organizar los archivos PDB necesarios en un directorio temporal, asegurando que todos los modelos estructurales requeridos estén disponibles y correctamente nombrados para los pasos siguientes. El script utilizado en esta etapa permite seleccionar los mejores

alineamientos según criterios configurables, lo que optimiza el uso de recursos y enfoca el análisis en los casos más relevantes. A continuación, el pipeline automatiza la obtención e instalación de FATCAT desde su repositorio oficial de GitHub. Esto garantiza que la herramienta esté disponible en el entorno local y correctamente configurada, lo que es fundamental para la reproducibilidad y portabilidad del flujo de trabajo. Una vez instalado FATCAT, ejecuta los alineamientos estructurales entre los pares de proteínas seleccionados. El script Python asociado a esta regla automatiza la ejecución de FATCAT sobre todos los pares definidos, gestionando la paralelización y el manejo de los archivos de salida. Para evaluar cuantitativamente la similitud estructural de estas proteínas, la regla utiliza la herramienta TM-align, que calcula el TM-score, una métrica estándar para comparar la similitud global entre estructuras proteicas. Esta regla ejecuta TM-align sobre los pares identificados y genera archivos de salida con los resultados de cada alineamiento. Finalmente, la regla genera un reporte tabulado que resume la similitud estructural de los pares analizados.

La quinta regla implementada en el archivo `04_Downloading_annotation_from_uniprot.smk`, está dedicada a la descarga y procesamiento de anotaciones funcionales desde la base de datos UniProt (M&M 5). UniProt es la base de datos de referencia para la información funcional de proteínas, integrando datos experimentales y computacionales sobre función, localización subcelular, participación en rutas metabólicas, entre otros aspectos. En esta etapa, el pipeline recupera las anotaciones correspondientes a todas proteínas involucradas en la comparación. El proceso incluye la extracción de términos *Gene Ontology*, dominios identificados por InterProScan, descripciones funcionales, identificadores de rutas metabólicas y otra información relevante. Esta regla culmina con la creación del documento final del pipeline el cual incluye la información de homología estructural mediante SRBH, sus valores derivados de la comparación de Foldseek, para los cinco mejores también se incluye los valores reportados por FATCAT, además, se incluye la anotación derivada de UniProt, tanto para la proteína de kinetoplástidos como para la proteína del organismo de referencia.

La integración de estas cinco reglas principales se realiza mediante la regla “all”, que define como objetivo final la generación de una tabla resumen en formato TSV, ubicada en el directorio de resultados. Esta tabla constituye el producto final del pipeline, integrando la información de agrupamiento, alineamientos estructurales, validaciones y anotaciones funcionales en un formato fácilmente interpretable y reutilizable para análisis posteriores. Cabe destacar que el pipeline está diseñado para ser altamente configurable y reproducible. La configuración de los archivos de entrada, parámetros de los programas y opciones de filtrado se realiza a través de un archivo YAML centralizado, lo que facilita la adaptación del flujo de trabajo a diferentes organismos, conjuntos de datos y objetivos de análisis. Además, la modularidad del diseño, basada en la inclusión de reglas independientes, permite la actualización y mejora de cada etapa sin afectar la integridad del pipeline global.

En cuanto a la lógica de ejecución, Snakemake gestiona automáticamente las dependencias entre reglas, asegurando que cada paso se ejecute únicamente cuando sus archivos de entrada estén disponibles y actualizados. Esto optimiza el uso de recursos computacionales y permite la ejecución paralela de tareas independientes, acelerando significativamente el tiempo total de análisis. La integración de mensajes informativos y la impresión de los archivos y organismos modelo utilizados en cada ejecución contribuyen a la transparencia y trazabilidad del proceso.

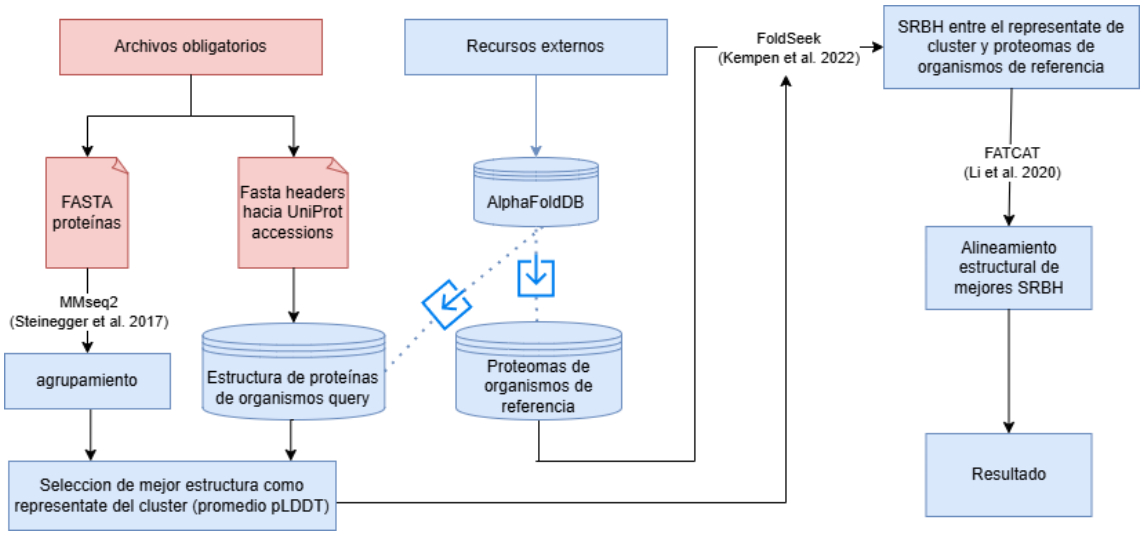

Figura 4. Flujo de trabajo de ASC. Esquema del proceso de búsqueda de homología de proteínas utilizando datos estructurales. El flujo de trabajo comienza con dos archivos obligatorios: un archivo FASTA con las secuencias de proteínas y un archivo TSV que vincula los encabezados FASTA con los códigos de UniProt correspondientes. Se emplea MMseqs2 para generar los clústeres de proteínas. Se descargarán todas las estructuras disponibles en AFDB para cada clúster y se seleccionará el mejor representante estructural basándose en las puntuaciones medias de pLDDT. Foldseek se utiliza para identificar SRBH entre los representantes de los clústeres y los proteomas de los organismos de referencia. Los hits principales se someten a alineamiento estructural con FATCAT y, tras el filtrado por TM-score, se obtiene la tabla final que resume los resultados.

Flujo de datos a lo largo del pipeline

En este trabajo comenzamos con 657.192 secuencias de proteínas de kinetoplástidos procedentes de todos los genomas disponibles en la base de datos pública TriTrypDB. El pipeline comienza con la aplicación de la primera regla (descrita en la sección anterior y en M&M 2). Este conjunto inicial de secuencias de proteínas se agruparon en clústeres y se filtraron según su tamaño (número de secuencias que contienen), seleccionando únicamente aquellos con al menos 10 secuencias con el objetivo de evitar errores de predicción además de otros casos particulares que se detallaran más adelante.

Este proceso resultó en 14.778 clústeres que abarcan 478.612 secuencias (**Figura 5**). Para evaluar la consistencia de los clústeres comparamos nuestros resultados con los grupos ortólogos definido por la base de datos OrthoMCL [70] utilizada por TriTrypDB como referencia de ortólogos. El 90% de nuestros clústeres se corresponden con un único grupo de OrthoMCL, demostrando que nuestra estrategia de agrupamientos es en gran medida un subconjunto de los grupos de OrthoMCL. Este resultado era de esperar siendo que nuestros parámetros de agrupamiento son más exigentes que los reportados en OrthoMCL y las estrategias de agrupamiento similares [70]. Hoy nuestro objetivo de agrupamiento, más allá de lo computacional, es obtener grupos de genes donde la transferencia de anotación y la representatividad por un único miembro sea confiable. La comparación con OrthoMCL demuestra que estamos dentro de límites seguros para este objetivo. Por lo tanto, la transferencia de anotaciones dentro de estos grupos parece un enfoque justificado.

De las 478.612 secuencias englobadas en 14.778 clústeres o grupos, 210.201 (44 %) estaban etiquetadas como PHip y 268.411 como PAnot (66%). El resultado primario de clasificación de secuencias en PHip o PAnot demuestra que prácticamente la mitad de la base de datos está clasificada como PHip. Es importante destacar que esto no se refleja necesariamente en que la mitad de todos los genomas este compuesto por PHip. Algunos organismos como es el caso de *T. brucei* y *T. cruzi* cuentan con porcentajes mayores al 50% de anotación, pero otros mucho menos.

Al mismo tiempo llama la atención como el paso inicial de agrupamiento y selección descartó 117.582 PHip y 61.198 PAnot. Esto se puede explicar en parte por genes con un

número insuficiente de genomas para cumplir con el umbral mínimo de homologías, genes específicos de especies, pseudogenes, secuencias altamente variables o artefactos generados durante la anotación del genoma. A modo de ejemplo, la reducción genómica de los trypanosomas asociadas al parasitismo hace que muchos de los presentes genes en *B. saltans* tengan pocos homólogos y queden descartados. Además, los métodos de anotación pueden introducir artefactos debido a la predicción errónea de regiones codificantes, particularmente en especies con ensamblajes genómicos fragmentados o incompletos [71]. La pérdida significativa de genes en estas etapas subraya los retos inherentes a la anotación genómica en kinetoplástidos.

Por otro lado, el umbral estricto de agrupamiento ayuda a garantizar análisis posteriores robustos, pero excluye inevitablemente predicciones génicas genuinas. Estas pérdidas reflejan las limitaciones actuales de los enfoques basados en la homología para capturar la diversidad completa de genes de kinetoplástidos. Para contrarrestar la perdida de genes genuinos nos proponemos para la futura actualización del pipeline el uso de conservación a nivel de especies. Por ejemplo, incluir aquellos clústeres que no cumplan con los 10 miembros pero que dentro de ellos haya al menos dos genes de especies o cepas diferentes. Este aspecto fue explorado y observamos que hay genes con relevancia biológica que quedarían incluidos en nuestro conjunto de datos inicial.

De los 14.778 clústeres, 3.565 consistían únicamente en PHip a los cuales denominamos como “Dark Clusters” [68] comprendiendo 86.679 secuencias. Las 123.522 PHip restantes (el 59 % del total) se agruparon junto con PAnot. Gracias a nuestros parámetros estrictos de agrupamiento, pudimos transferir la anotación a un número significativo de estas proteínas basándonos únicamente en homología de secuencia [49,68]. Este resultado también muestra como habiendo suficiente información a nivel de secuencia la estrategia de anotación elegida por la base de datos TriTrypDB no transfiere la misma, incluso entre genes del mismo grupo de homología de la propia base.

De los 14.778 clústeres retenidos, 14.267 tenían estructuras proteicas disponibles en AFDB (que incluye los proteomas de 20 *Leishmania*, 20 *Trypanosoma*, 2 *Leptomonas*, *B. saltans* y *Porcisia hertigi*), los cuales sirvieron como nuestra base de datos de consulta para los pasos posteriores de pipeline (M&M 3). Los 511 clústeres restantes, que carecían de estructuras disponibles, fueron excluidos del análisis. En trabajos futuros,

sería interesante incluir representantes de estos clústeres de proteínas conservadas en kinetoplástidos, presumiblemente relevantes desde el punto de vista biológico. Para ello, se podrían realizar predicciones estructurales de manera local e integrarlas posteriormente en el flujo de trabajo actual. Para los clústeres en los que se obtuvo una estructura representativa, realizamos el paso SRBH contra los 45 OrgRef, reteniendo 11.753 clústeres, incluidos 2.689 Dark Clusters (M&M 4). Al conjunto de datos resultante lo denominaremos "**Conjunto Bruto**", dado que no vamos a aplicar ningún tipo de filtro posterior. De este conjunto, seleccionamos los cinco mejores SRBH para cada clúster y llevamos a cabo alineamientos estructurales con FATCAT. Calculamos los valores de TM-score mediante TM-align. Aplicando un umbral estricto de TM-score de 0,5 para homología proteica [72] filtramos los resultados, obteniendo así nuestro "**Conjunto Final**" (M&M 4). Aunque el umbral de TM-score de 0,5 se definió originalmente para proteínas de un solo dominio [72], también se ha utilizado como referencia en alineamientos de proteínas multidominio [69]. Cabe señalar que el TM-score considera la longitud del alineamiento entre los aciertos ("hits"). Por lo tanto, nuestros parámetros priorizan alineamientos completos de la proteína, excluyendo alineamientos parciales que pudieran derivarse de dominios o subdominios similares. El Conjunto Final está compuesto por 7.486 clústeres, incluidos 942 Dark Clusters que comprenden 23.290 secuencias proteicas (**Figura 5**). Es importante señalar que estos clústeres se definieron como "Dark" en función de los nombres génicos actuales de TriTrypDB. Por tanto, el número de Dark Clusters y las mejoras alcanzables mediante búsquedas de homología basadas en estructura dependen de este campo específico de cada entrada génica y no tienen en cuenta información adicional que pudiera no estar incorporada en el nombre de la proteína en la base de datos.

Nuestra definición de proteína hipotética no implica que necesariamente estos genes no cuentan con ningún tipo de información en las bases de datos ni que no puedan ser anotadas por métodos basados en secuencia. Por lo tanto y con objetivo de profundizar en la relevancia de nuestros hallazgos exploramos estas posibilidades para los genes de los Dark Clusters, ya que como comentamos en párrafos anteriores el resto de las proteínas hipotéticas ya habían sido posible anotarlas por transferencia de anotación entre los miembros de los clústeres generados.

Comenzamos por explorar la información disponible en la base de datos. Dado que el nivel o profundidad de anotación es un problema semántico difícil de abordar sistemáticamente o para grandes conjuntos de datos decidimos hacer una clasificación binaria. En esta clasificación consideramos como información de anotación cualquier campo reportado en la base de datos TriTrypDB. Por lo que sin importar cuán informativo sea ese campo lo consideramos como anotación para encontrar aquellos genes dónde nuestro abordaje haya hecho un aporte significativo. Además, ejecutamos dos softwares de anotación independientes a TriTrypDB y con estrategias de búsqueda distintas: eggNOG e InterProScan. Los resultados de estos van a ser interpretados de la misma forma binaria, dónde por ejemplo en el caso de InterProScan que usa HMM para identificar dominios con que un solo perfil de HMM identifique una región de la proteína caerá en la clasificación de "anotada".

Encontramos que 3.944 proteínas presentes en nuestros Dark Clusters cuentan efectivamente con algún tipo de anotación en TriTrypDB, tales como términos GO asignados, identificadores de dominios proteicos, entre otros. Además, con el creciente número de organismos anotados y las mejoras en los métodos de anotación funcional, es posible que algunas de estas proteínas pudieran ahora tener homólogos de secuencia o dominios identificables. Por ello, evaluamos si la anotación de estas proteínas podía mejorarse utilizando enfoques actuales basados en secuencia. Para ello, ejecutamos eggNOG e InterProScan en las aproximadamente 23.000 proteínas de los Dark Clusters. Este análisis reveló que 1.700 proteínas podían anotarse usando eggNOG y que para aproximadamente 4.000 podía identificarse un dominio mediante InterProScan. Finalmente, la integración de información de todos los enfoques y su transferencia entre los miembros del clúster refuerza que una proporción significativa de las 23.000 proteínas en los Dark Clusters, aproximadamente 6.700, pudieron recibir una anotación putativa únicamente a través de comparaciones estructurales. Por lo tanto, este conjunto final de proteínas mediante ninguna estrategia de las antes mencionadas se le pudo asignar ni siquiera un GO o un dominio proteico identificado por InterProScan. Estas son proteínas que de no ser por la comparación estructural no tendrían ningún tipo de información. Mas allá de exactamente que anotación logramos transferir, que hayan

tenido resultado de SRBH habla de homología con organismos modelo y por lo tanto genes de interés para su estudio y posiblemente con relevancia biológica.

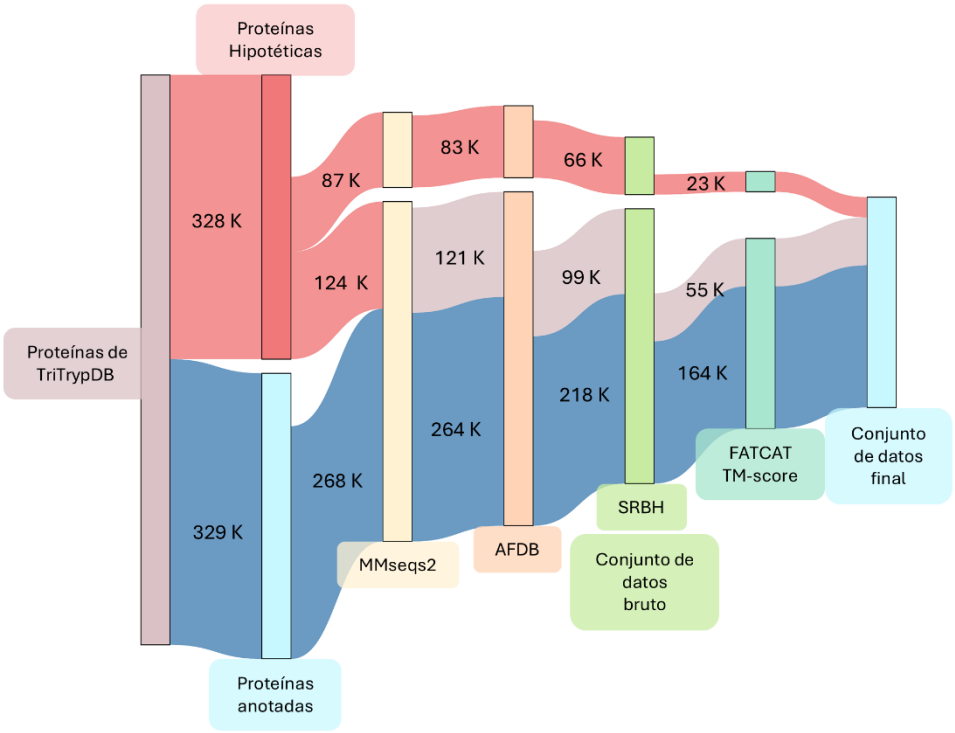

Figura 5. Número de secuencias de proteínas retenidas a lo largo de los pasos principales del pipeline. Diagrama de Sankey que muestra el número de genes retenidos tras cada paso principal. Los nodos del diagrama, representados cada uno con un color específico, indican las etapas del pipeline, mientras que las conexiones entre nodos reflejan el número de genes (en miles) que pasan de un paso al siguiente. El código de colores distingue, según TriTrypDB, entre genes hipotéticos y anotados: rojo para genes denominados "hipotéticos", azul para genes anotados y rojo claro para genes hipotéticos con homología de secuencia a genes anotados. Los pasos consisten en el agrupamiento de secuencias de proteínas (MMseqs2), obtención de estructuras (AFDB), *reciprocal-best-hit* estructural por Foldseek (SRBH) y el alineamiento estructural (FATCAT y TM-score), el flujo de trabajo se depura hasta obtener el "Conjunto Final".

Validación y anotación

En este trabajo abordamos por primera vez la transferencia de anotaciones basadas en comparación estructural en un grupo de organismos tan divergentes como los kinetoplástidos. Para ello, implementamos dos métodos independientes de evaluación: el primero se fundamenta en la identificación de dominios y la clasificación a nivel de familia génica mediante perfiles de HMM; el segundo consiste en una validación comparativa con datos experimentales a gran escala, lo cual pudo ser implementado en este trabajo gracias a la existencia de la base de datos TrypTag. A partir de estos dos enfoques, evaluaremos la precisión y el alcance de nuestra estrategia de transferencia de anotaciones.

Validación basada en HMM

Para evaluar la fiabilidad de la transferencia de anotaciones desde los genes de OrgRefs hacia los genes de kinetoplástidos, comparamos la presencia/ausencia de dominios identificados por HMM en distintas bases de datos incorporadas en UniProt. Es decir, para cada SRBH en los cuales ya se cuenta con dominios identificables por HMM tanto para kinetoplástido como para el OrgRef comparamos cuan consistentes son sus anotaciones en distintas bases de datos. Esta estrategia ya ha sido utilizada en estudios similares [68,73,74]. En este trabajo se decidió evaluar Protein Families, PANTHER y Pfam por separado de InterPro siendo que son bases de datos ampliamente reconocidas en el campo y fueron utilizadas en trabajos similares [68,73,74] lo que permite comparaciones directas con sus resultados.

La evaluación la realizamos empleando dos conjuntos de datos: el **Conjunto Bruto**, basado únicamente en SRBH obtenidos por Foldseek, y el **Conjunto Final**, en el que además se aplicó un filtro de 0,5 TM-score calculado mediante FATCAT y TM-align (M&M 4). En primer lugar, para cada par de SRBH obtuvimos las anotaciones de UniProt correspondientes a las bases de datos Protein Families, PANTHER, Pfam e InterPro, tanto para la proteína de kinetoplástido (“K”) como para los OrgRef (“OR”). Los resultados de la comparación se clasificaron según el grado de coincidencia en cinco categorías: TODO, K_in_OR, OR_in_K, PARCIAL y CERO (M&M 6).

Dado que ofrecen un único nivel de clasificación (coincide o no coincide) en las bases de datos Protein Families y PANTHER únicamente resultan aplicables las categorías TODO y CERO. En Protein Families, el Conjunto Bruto muestra un 90 % de coincidencias en TODO, lo que indica una clasificación familiar correcta en la gran mayoría de los casos; tras aplicar el filtrado por TM-score, esta cifra asciende al 95 %, reflejando una mejora clara en la identificación. En paralelo, la proporción de asignaciones incorrectas (CERO) disminuye en igual medida. De forma análoga, en PANTHER la precisión en TODO crece del 77 % al 88 % tras el filtrado, con una reducción correspondiente de la categoría CERO (**Figura 5A y 5B**). Estos resultados demuestran que las proteínas con asignación de familia previa mantienen un elevado grado de concordancia con la clasificación SRBH, alcanzando niveles comparables a los observados en organismos filogenéticamente más cercanos (clado Opisthokonta) [73,74]. Los resultados de Protein Families y PANTHER muestran que nuestra transferencia de anotaciones es consistente a nivel de familia de proteínas con algunas excepciones que no exploramos en profundidad. Tras el filtrado, la precisión aumenta, reflejándose en un mayor número de coincidencias exactas (categoría TODO) y una reducción de las anotaciones no coincidentes (categoría CERO) (**Figura 5A y 5B**).

Al centrarnos en el nivel de dominios, la base Pfam requiere considerar la categoría PARCIAL (cuando los SRBH comparten algunos, pero no todos, los dominios) además tenemos dos categorías que en la figura aparecen sumadas a la categoría TODO: K_in_OR y OR_in_K (M&M 6). K_in_OR se refiere a los casos en que todos los dominios detectados en el kinetoplástido están contenidos en el organismo de referencia, mientras que OR_in_K designa el escenario inverso. Si sumamos TODO, K_in_OR y OR_in_K como coincidencias aceptables, el Conjunto Bruto alcanza un 86 %, cifra que sube al 93 % en el Conjunto Final. En detalle, en el Conjunto Bruto K_in_OR representa aproximadamente un ~14 %, frente a un ~4 % de OR_in_K; tras el filtrado, K_in_OR aumenta a ~16 % y OR_in_K desciende a ~2 %. Hipotetizamos que esta particularidad, la cual implica que los genes de kinetoplástidos a nivel de dominios identificado por HMM son subconjunto de los organismos de referencia, puede ser un reflejo de la falta de sensibilidad de los HMM en este tipo de organismos tan divergentes con respecto a las referencias. Los mismos HMM que logran identificar dominios en los organismos de

referencia no logran lo mismos en la secuencia de proteína de lo kinetoplástidos y por lo tanto esto se reflejó como K_in_OR en lugar de TODO. Llama poderosamente la atención que el caso inverso OR_in_K se reduzca a valores marginales, lo cual se alinearía con lo antes mencionado. Esto demuestra que nuestra estrategia aprovecha los 45 genomas de referencia con alto nivel de curación para enriquecer genes que ya cuentan con alguna anotación previa: podemos agregar dominios que los HMM no logran detectar en los kinetoplástidos. De este modo, transferimos un volumen significativo de información fiable y profundizamos la anotación.

Finalmente, las inconsistencias definidas como PARCIAL y CERO en Pfam representan el 3 % y el 11 %, respectivamente, en el Conjunto Bruto, y disminuyen tras aplicar el filtrado. Esta reducción adicional subraya el beneficio de incorporar un umbral de TM-score para mejorar tanto la sensibilidad como la especificidad de la transferencia de anotaciones a nivel de dominios.

Por último, recurrimos a InterPro, que condensa la información de trece bases de datos independientes integradas y curadas manualmente. En esta plataforma, las firmas de proteínas que describen la misma familia, dominio o sitio se fusionan en una única entrada, y los curadores añaden datos biológicos complementarios. Al igual que en Pfam, tras aplicar el filtrado por TM-score observamos un aumento notable de las coincidencias aceptables (TODO + K_in_OR + OR_in_K): pasamos de un 79 % en el Conjunto Bruto a un 91 % en el Conjunto Final. No obstante, InterPro evidencia de forma más marcada el fenómeno de inclusión parcial: en el Conjunto Bruto, K_in_OR representa un 39 % frente a un 6 % de OR_in_K; tras el filtrado, K_in_OR asciende al 46 % y OR_in_K desciende al 4 %. Este desequilibrio refuerza la inquietud de que los perfiles HMM, aunque eficaces para organismos bien representados en las bases de datos, carecen de sensibilidad suficiente para capturar con igual precisión los dominios en kinetoplástidos.

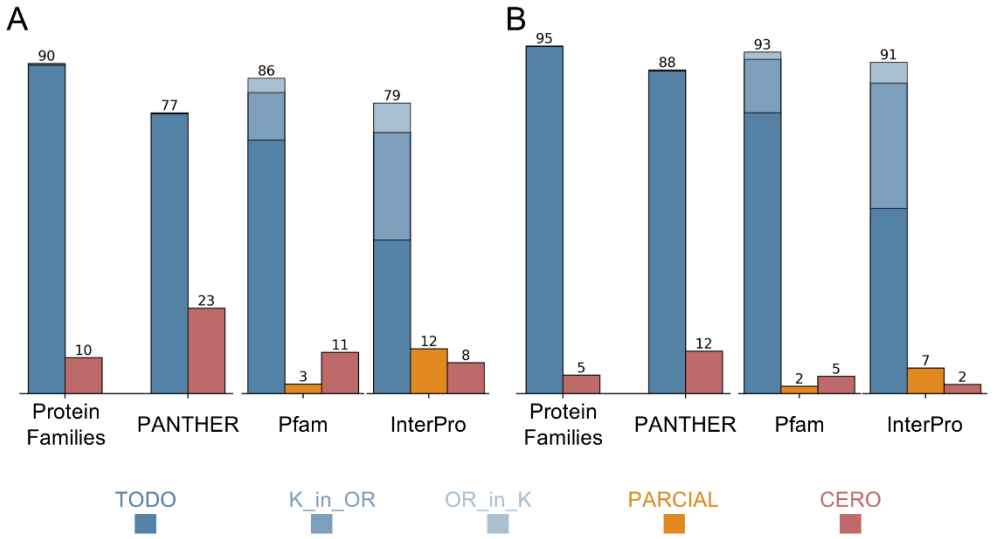

Figura 5. Evaluación de las anotaciones SRBH antes y después del filtrado. Comparación de los resultados de anotaciones SRBH para proteínas query (kinetoplastidos) y de target (OrgRef) usando cuatro bases de datos: Protein Families, PANTHER, InterPro y Pfam. El panel izquierdo (A) muestra los resultados de anotación antes de aplicar el proceso de filtrado, mientras que el panel derecho (B) muestra los resultados tras el filtrado (Conjunto Final). Cada barra representa el porcentaje de anotaciones clasificadas como TODO (azul), K_in_OR (azul claro), OR_in_K (azul más claro), PARCIAL (naranja) y CERO (rojo). Los números en la parte superior de cada barra indican el porcentaje total de la columna, redondeado.

Validación basada en datos experimentales

TrypTag es una base de datos experimental de alcance genómico que proporciona un mapa detallado de la localización subcelular de proteínas en *T. brucei*. Este recurso fue desarrollado mediante un proyecto de cuatro años que empleó técnicas de etiquetado fluorescente (N y/o C-terminal) y microscopía de alta resolución para determinar la ubicación intracelular de las proteínas codificadas en el genoma de *T. brucei*. Cada proteína etiquetada fue anotada manualmente respecto a su localización subcelular utilizando una ontología estandarizada [62]. Esta base de datos constituye una referencia inigualable para nuestro abordaje siendo que nuclea resultados experimentales con un abordaje genómico de alta precisión. Nos propusimos entonces evaluar nuestros resultados de SRBH con lo evidenciado en TrypTag. Para ello, evaluamos la precisión de la transferencia de información (tentativa) sobre localización subcelular desde homólogos estructurales en los OrgRef hacia proteínas de kinetoplástidos, utilizando como referencia la localización reportada por TrypTag. En otras palabras, para cada gen de kinetoplástido vamos a tener un gen estructuralmente homologo y vamos a utilizar la información de localización subcelular de este último para comparar con TrypTag. A

modo de ejemplo, si el representante del clúster de kinetoplástido es Tb927.8.4480 y tiene como homologo estructural a Ssu72L4 de humano, compararemos la localización reportada en UniProt para Ssu72L4 con la localización de Tb927.8.4480 en TrypTag. De esta manera emulamos la transferencia de anotación desde los genes de OrgRef hacia los genes de kinetoplástidos dándonos una idea de precisión. Como control, evaluamos la consistencia de las anotaciones actuales de localización subcelular de proteínas de kinetoplástidos con respecto a los datos de TrypTag. Volviendo al ejemplo anterior utilizamos la anotación subcelular disponible en UniProt para Tb927.8.4480 y la comparamos con la anotación de Tb927.8.4480 en TrypTag. Esto nos permite evaluar nuestro método de anotación con respecto a los métodos actuales por los cuales se predijo dicha localización subcelular.

Para aumentar la confiabilidad de nuestro análisis, eliminamos entradas de baja señal de fluorescencia y seleccionamos localizaciones celulares ubicuas en todos los eucariotas (M&M 7), utilizamos la anotación de nuestro Conjunto Final para comparar. Evaluamos la similitud semántica de los términos GO utilizando GOGO [75]. La **Figura 7** muestra la distribución de los valores de similitud semántica de los términos GO CC para cada comparación. Los gráficos de violín en gris representan la distribución de los puntajes de GOGO al comparar la anotación de genes de kinetoplástidos (anotación actual de UniProt) con las anotaciones asignadas por TrypTag, que utilizamos como referencia/control. Los gráficos de violín en rojo representan los resultados de la comparación entre los datos de anotación OrgRef derivados de los SRBH y las anotaciones de TrypTag, lo que aproxima el nivel de confianza que podemos obtener con nuestra estrategia de anotación para proteínas hipotéticas.

Utilizando la prueba de Mann-Whitney-U para comparar las medias, no observamos diferencias significativas para la mayoría de las categorías, lo que indica que la anotación actual es igualmente coherente con TrypTag que nuestra anotación tentativa basada en SRBH. Sin embargo, observamos una tendencia en las categorías de citoplasma, envoltura nuclear y aparato de Golgi, lo que sugiere que la estrategia SRBH asigna localizaciones subcelulares más alineadas con los resultados experimentales de TrypTag. En conclusión, nuestra anotación es al menos tan precisa como los métodos actualmente

aceptados cuando se compara con TrypTag, y aportando nueva información para proteínas que hasta ahora no habían sido caracterizadas.

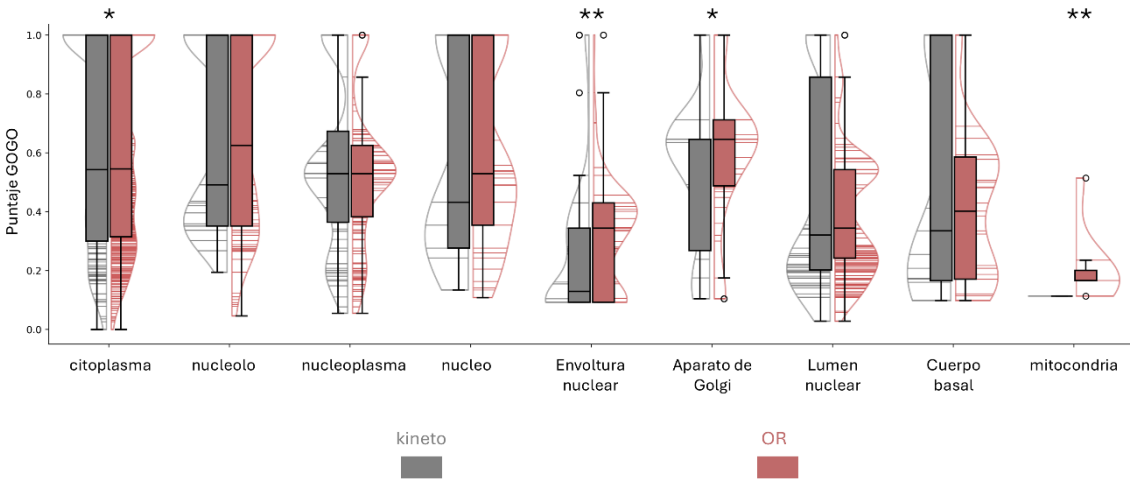

Figura 7. Comparación de las distribuciones de los puntajes de similitud semántica de GO CC para las anotaciones de kinetoplástidos y de organismos de referencia. Distribución de las puntuaciones de similitud semántica de términos GO (puntaje GOGO) para dos categorías: kinetoplástidos (gris) y organismo de referencia (rojo). Cada gráfico de violín muestra el rango y la distribución de puntuaciones para componentes celulares ubicuos, incluidos citoplasma, nucleolo, nucleoplasma, núcleo, envoltura nuclear, aparato de Golgi, lumen nuclear, cuerpo basal y mitocondria. El diagrama de caja dentro de cada violín representa el rango intercuartílico (IQR) e indicando la mediana. * p-valor < 0,05; ** p-valor < 0,01.

Casos de estudio

Tras validar nuestro enfoque, buscamos ejemplos biológicamente relevantes dentro de nuestros resultados, centrándonos en los Dark Clusters y los clústeres anotados carentes de anotaciones precisas. Con este objetivo, corrimos el software BUSCO utilizando la base de datos para eucariotas, para todas las secuencias de proteína disponibles en TriTrypDB y los OrgRef. Esta base contiene 255 "grupos BUSCO", correspondientes a genes ortólogos de copia única comunes a la mayoría de los eucariotas. Estos genes se caracterizan por estar presentes en prácticamente todos los eucariotas lo que indirectamente indica su relevancia biológica y probablemente un rol funcional basal. Nuestro objetivo en este paso es comparar los resultados entre los SRBH resultantes de nuestro abordaje. La idea detrás es identificar aquellos grupos busco en OrgRef que no sean encontrados en los genes de TriTrypDB y tengan SRBH con algún Dark Cluster. En otras palabras, genes eucariotas de copia única presentes en la gran mayoría de la diversidad eucariota que no estén anotados en TriTrypDB. Usar BUSCO nos permite evaluar nuestra anotación evitando la selección de casos favorables alias *cherry-picking*. Es una forma indirecta de seleccionar genes para su identificación y luego contrastarlo con información en artículos científicos.

Con el fin de refinar nuestra búsqueda e identificar auténticos Dark Clusters entre los clasificados como tales a partir de TriTrypDB, añadimos la descripción del producto disponible en UniProt. Posteriormente, nos centramos en aquellos donde ningún miembro del grupo posea anotación precisa en ninguna de las dos bases de datos. Dado que no hay una forma sencilla de automatizar este paso y para aumentar la sensibilidad la clasificación fue hecho "manualmente" leyendo los campos y seleccionando aquellos clústeres que estuvieran clasificados como hipotéticos o que tengan anotación que no permita una identificación precisa.

BUSCO identificó 68 grupos reportados como "Missing" en todo TriTrypDB, lo que indica que ninguna proteína de kinetoplástidos fue identificada por los perfiles HMM de esos grupos. En otras palabras, esas proteínas faltan, están muy fragmentadas o son demasiado divergentes para generar un hit significativo [64]. Nuestro enfoque de comparación estructural halló homólogos putativos para 48 de estos 68 grupos BUSCO

ausentes. Cabe señalar que 39 de esos genes ya estaban anotados en TriTrypDB o UniProt, probablemente mediante otros métodos de anotación como evidencia experimental, otras bases BUSCO o herramientas automáticas. Los 9 grupos BUSCO restantes fueron identificados usando nuestra estrategia SRBH y serán descriptos en detalle a continuación. Los IDs de gen para todas las proteínas de estos clústeres se incluyen en las **Tablas S3 y S4**. Estos genes abarcan categorías funcionales centrales distintas, entre ellas transcripción, reparación de ADN, traducción y regulación del ciclo celular.

Nuestro abordaje es automático y por lo tanto estos 9 grupos BUSCO identificados son suficiente información para confirmar que esta estrategia al menos para estos casos es más sensible y precisa que los pipelines de anotación de TriTrypDB y UniProt. De todos modos, hicimos una búsqueda manual, uno a uno y supervisada por un investigador utilizando herramientas como BLASTP e InterProScan para confirmar incluso que esta estrategia es especialmente relevante para algunos casos. En este sentido, para confirmar que estos genes no son identificados por estrategias convencionales realizamos un análisis detallado caracterizando su arquitectura de dominios con InterProScan, buscando hits BLASTP contra la base de datos de secuencias de proteínas no redundante de NCBI (excluyendo Discoba) y comparando estos hallazgos con nuestra transferencia de anotaciones vía SRBH. A continuación, se describen los resultados obtenidos y su relevancia. Los ejemplos están ordenados según cuan relevante fue nuestra estrategia para su identificación. Los primeros tres relacionados a los genes Ssu72, Cbp3 y Tfb4 son evidencia de la potencia de nuestro abordaje ya que ningún otro método automático e incluso métodos manuales con supervisión profesional fueron capaces de asignarles identidad. El resto de los genes, se caracterizan por no ser identificados por métodos automáticos más específicamente los pipelines de TriTrypDB e UniProt/InterPro, pero la observación detallada y supervisada puede dar indicios de su identidad. En estos casos nuestro abordaje es un aporte considerable al establecimiento de homología.

Fosfatasa Ssu72 del dominio C-terminal de la subunidad A de la ARN polimerasa II

El gen Ssu72 codifica una fosfatasa altamente conservada cuya función molecular principal es eliminar grupos fosfato de residuos de serina específicos (principalmente Ser5 y Ser7) dentro de las repeticiones heptapéptidas (YS_2PTS_5PS) del dominio C-terminal (CTD) de la subunidad A de la ARN polimerasa II [76,77]. Esta desfosforilación constituye un mecanismo central de regulación que modula el ciclo de transcripción al controlar la transición entre las fases de iniciación, elongación y terminación [77]. Funcionalmente, al desfosforilar Ser5P (y, en menor medida, Ser7P), Ssu72 desempeña un papel fundamental en la terminación de la transcripción asegurando que la ARN polimerasa II se desenganche correctamente del ADN y en la coordinación de los eventos de procesamiento del extremo 3' del ARN mediante su pertenencia al complejo de corte y poliadenilación (CPF) [78]. Esta actividad fosfatasa también facilita una elongación transcripcional adecuada al evitar la acumulación excesiva de ARN polimerasa II fosforilada en las regiones proximales al promotor, manteniendo así un equilibrio entre el reclutamiento de ARN polimerasa II, la pausa y la elongación [77,78]. En conjunto, la función molecular de Ssu72 se centra en la eliminación precisa de grupos fosfato del CTD de la ARN polimerasa II, un proceso integral para la regulación de la dinámica del ciclo de transcripción, el procesamiento del ARN y la estructura de la cromatina. A través de sus interacciones con diversos factores de transcripción y procesamiento, Ssu72 coordina una red de eventos que garantizan una expresión génica robusta y confiable, sosteniendo tanto la homeostasis transcripcional como la integridad genómica.

Hasta este momento se desconocía la presencia de Ssu72 en kinetoplástidos y de hecho experimentos de purificación por afinidad en tándem del CPF en *T. brucei* propusieron un complejo parcialmente convencional que se caracterizaba por la ausencia de fosfatasa de CTD entre los componentes y por dos subunidades específicas de trypanosomas (Tb927.11.13860 y Tb927.8.4480) [45,79]. Estas dos subunidades son las únicas para las que no se logró asignar homología mediante análisis comparativo y búsquedas basadas en secuencia [79]. En nuestros resultados identificamos un Dark Cluster representado por TcCLB.509569.160 (270 aminoácidos de largo) como homólogo estructural de Ssu72 (**Figura 8**).

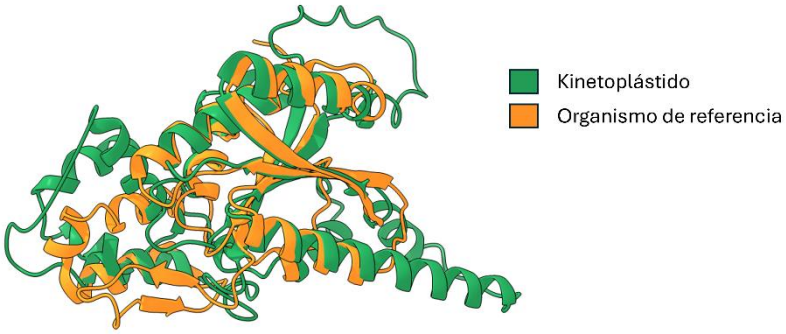

Figura 8. Alineamiento estructural de la fosfatasa del dominio C-terminal de la subunidad A de la ARN polimerasa II (Ssu72) de *Drosophila melanogaster* con homólogos en kinetoplástidos identificados mediante nuestro enfoque. La proteína del organismo de referencia se muestra en naranja, y la proteína de kinetoplástidos se muestra en verde. Códigos de UniProt: Q9VWE4_DROME y Q4DKS2_TRYCC (TcCLB.509569.160).

Nuestro análisis revela que el gen Tb927.8.4480 de *T. brucei* forma parte del clúster TcCLB.509569.160 (**Tabla S3**) y consecuentemente es homólogo estructural de Ssu72, lo que coincide con los resultados de purificación por afinidad en tándem del CPF y localización subcelular de TrypTag (**Figura 9**), descartando su clasificación inicial como un gen especifico de trypanosoma [79]. Los resultados experimentales indican que Tb927.8.4480 ejerce un papel represor en la poliadenilación, sugiriendo su implicación en pasos interconectados del procesamiento de ARNm [79]. Es pertinente indicar que el clúster está compuesto únicamente por organismos del género *Trypanosoma* incluyendo africanos y americanos. Resultaría interesante profundizar en los demás organismos como *B. saltans* o *Leishamania*, donde salvo que hayan sufrido pérdidas secundarias, deberían de poseer Ssu72.

El dominio característico de InterPro para identificar Ssu72 es IPR006811, que abarca casi toda la longitud de la proteína, se detecta en todos los organismos de referencia y es el único dominio identificado; no obstante, este HMM no logra identificar a ninguno de los miembros del clúster TcCLB.509569.160/Tb927.8.4480 y no se detecta ningún otro dominio de InterPro (**Tabla S6**). De manera consistente, no se encontraron hits BLASTP para los miembros del clúster en la base de datos no redundante de NCBI. Este

es un ejemplo de un gen donde la homología de secuencia es tan remota que ni la búsqueda de expertos con métodos manuales inspeccionados por investigadores capacitados logran identificar la proteína. Incluso cuando hay evidencia experimental como la precipitación del complejo CPF.

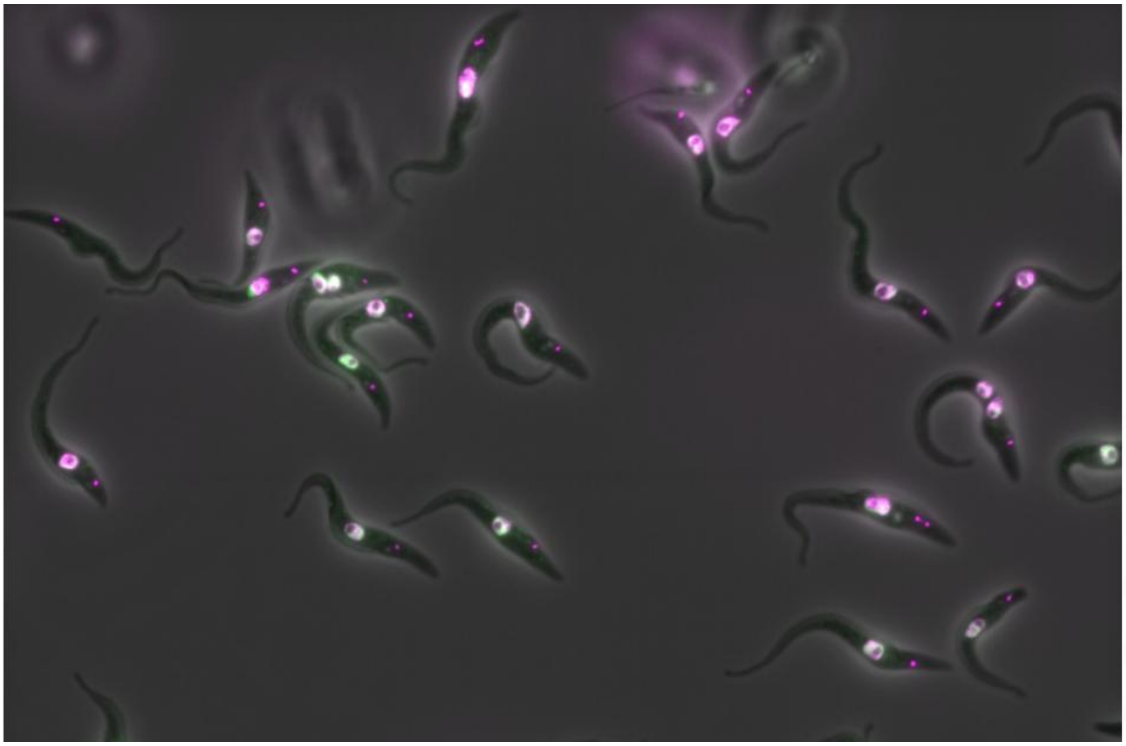

Figura 9. Localización subcelular del homologo estructural de la proteína Ssu72 en los datos de TrypTag de *T. brucei*. Imágenes de microscopía de fluorescencia de proteínas de *T. brucei* (marcadas con mNeonGreen), mostrando su localización subcelular determinada por TrypTag. Verde (mNeonGreen): proteína de interés, Cian (Hoechst): ADN. Para Tb927.8.4480 la localización reportada es nucleoplasma.

En cuanto a Tb927.11.13860, el otro gen de *T. brucei* identificado como especifico de kinetoplástido, su clúster tuvo como SRBH a CSRP1 (proteína rica en cisteína y glicina 1). El conocimiento sobre CSRP1 sigue siendo limitado; sin embargo, la identificación de homólogos sugiere que no es una proteína específica de kinetoplástidos. Siendo que este gen no es parte de los grupos BUSCO no profundizaremos en su estudio.

Chaperona de ubiquinol-citocromo c

El gen Chaperona de ubiquinol-citocromo c (Cbp3) en levaduras codifica una proteína mitocondrial esencial para el ensamblaje del complejo III de la cadena respiratoria, también conocido como ubiquinol–citocromo c reductasa o complejo bc1 [80]. Cbp3 forma un complejo funcional con otra proteína llamada Cbp6, con la cual interactúa estrechamente durante las primeras etapas del ensamblaje del complejo III. El complejo Cbp3-Cbp6 cumple una doble función en la biogénesis de la única subunidad del complejo III codificada en la mitocondria, el citocromo b (Cytb). Se ha demostrado que tiene dos funciones importantes, es necesario tanto para la síntesis eficiente de esta proteína como para su protección frente a la proteólisis inmediatamente después de ser sintetizada [81]. Este complejo Cbp3–Cbp6 se asocia al ribosoma mitocondrial, específicamente en la salida del túnel de polipéptidos, permitiendo el reconocimiento inmediato y la estabilización del Cytb recién sintetizado [81]. Además de su función como chaperona, Cbp3 participa en la regulación de la traducción del ARNm del Cytb (COB), lo que sugiere un mecanismo de retroalimentación que coordina la síntesis de Cytb con su incorporación al complejo respiratorio [82].

Este gen ha sido especialmente caracterizado en levadura y en nuestros resultados encontramos SRBH contra la estructura CBP3_YEAST disponible en UniProt. Otros organismos de referencia con los cuales obtuvimos SRBH son: *Glycine max, Candida albicans, Plasmodium faciparum, Oryza sativa*, etc donde todas sugieren la misma anotación.

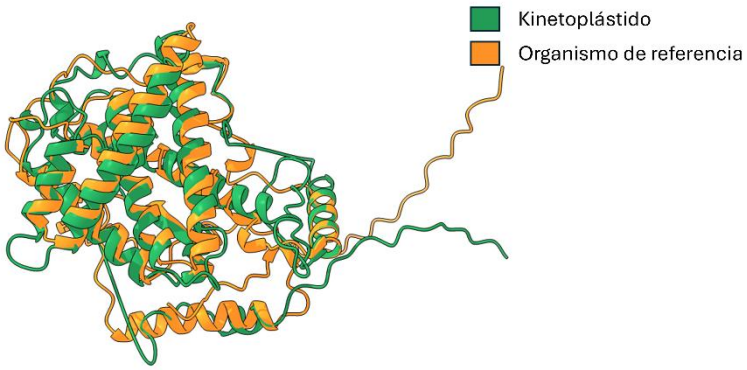

Figura 10. Alineamiento estructural de la Chaperona de ubiquinol-citocromo c (CBP3) de *Saccharomyces cerevisiae* con homólogos en kinetoplástidos identificados mediante nuestro enfoque. La proteína del organismo de referencia se muestra en naranja, y la proteína de kinetoplástidos se muestra en verde. Códigos de UniProt: CBP3_YEAST y A4I4I7_LEIIN (LINF_290018900).

El gen de kinetoplástido representante del clúster es LINF_290018900 representado por la estructura A4IAI7 (UniProt) con un largo de 296 aminoácidos. El clúster incluye genes representativos de toda la diversidad de kinetoplástidos como trypanosomas africanos y americanos, *Leishmania* e incluso al kinetoplástido de vida libre *B. saltans*. Demostrando que este gen es considerablemente conservado a lo largo de la evolución de los kinetoplástidos. Para todos los miembros del clúster representado por LINF_290018900 (**Figura 10**), no se reconocieron dominios mediante InterProScan **(Tabla S6**). Esta base de datos incluye los dominios distintivos de CBP3 "Chaperona ubiquinol-citocromo c, CBP3" (IPR007129) y "Chaperona ubiquinol-citocromo c" (IPR021150), que no tuvieron la capacidad de detectar ningún miembro del clúster y que sí están presentes en todos sus principales SRBH (**Tabla S6**). Tampoco se obtuvieron alineamientos significativos mediante BLASTP en la base de datos no redundante del NCBI. Al mismo tiempo no hay reportes en la literatura científica para los genes de kinetoplástidos del clúster. Sin embargo, la localización sugerida por TrypTag para el gen Tb927.3.3890 de *T. brucei* perteneciente al clúster es mitocondrial, consistente con lo esperado por nuestra identificación de homología (**Figura 11**).

Por todo lo antes mencionado resulta de particular interés ahondar mediante métodos experimentales esta proteína, por ejemplo, para identificar si forma un complejo con Cbp6 y aspectos vinculados a la función. Sin embargo, hasta el momento Cbp6 no ha sido identificada en estos organismos, incluyendo nuestro abordaje de SRBH. Este es un

caso particularmente complejo siendo que es una proteína pequeña entre los 100 y 150 aminoácidos con una estructura simple conformado por pequeñas alfa hélices.

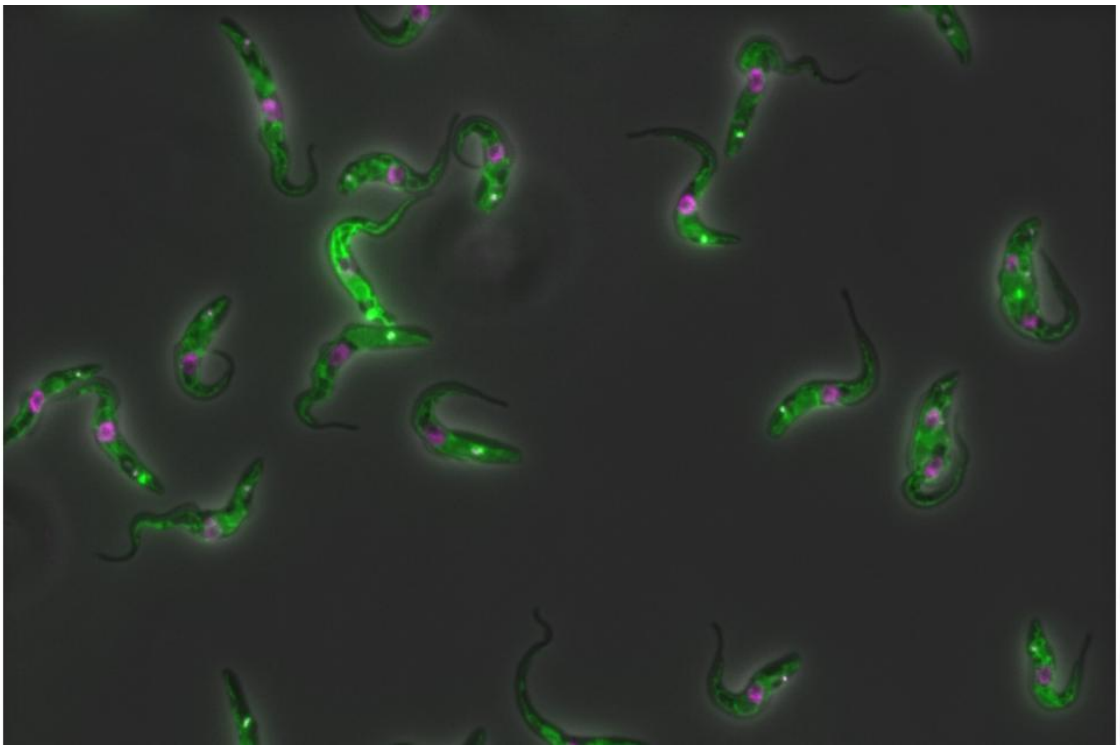

Figura 11. Localización subcelular del homologo estructural de la proteína CBP3 en los datos de TrypTag de *T. brucei*. Imágenes de microscopía de fluorescencia de proteínas de *T. brucei* (marcadas con mNeonGreen), mostrando su localización subcelular determinada por TrypTag. Verde (mNeonGreen): proteína de interés, Cian (Hoechst): ADN. Para Tb927.3.3890 la localización reportada es mitocondria.

Subunidad Tfb4 del factor de transcripción basal TFIIH

La subunidad Tfb4 (también conocida como p34 en levaduras) del factor de transcripción basal TFIIH, es un componente estructural conservado e indispensable del complejo TFIIH, necesario tanto para la iniciación de la transcripción por ARN polimerasa II como para la reparación por escisión de nucleótidos (NER) [83]. Tfb4 no realiza actividad enzimática propia, sino que actúa como andamiaje, asegurando el ensamblaje correcto y la cohesión del núcleo de TFIIH, cuya porción principal incluye las helicasas ATP-dependientes XPB y XPD [84,85]. Estas helicasas despliegan la doble hélice del ADN para abrir el promotor en transcripción y rodear las lesiones en NER, funciones que dependen de la integridad estructural facilitada por Tfb4 [85]. Estudios cristalográficos y bioquímicos han identificado dominios de interacción, un motivo von Willebrand A-like y una zona de dedo de zinc que median los contactos con otras subunidades, como p44, moduladora de la actividad helicasa de XPD [84]. Las mutaciones o pérdidas de Tfb4 provocan defectos graves en transcripción y reparación, lo que subraya su carácter esencial en la expresión génica y el mantenimiento del genoma [83].

Dos clústeres de kinetoplástidos mostraron SRBH con Tfb4 de distintos OrgRef (**Figura 12**). Un clúster, representado por Lsey_0010_0360, es un Dark Cluster, mientras que el otro, representado por TcCLB.508707.149, contiene varios miembros anotados como “subunidad del factor de transcripción basal TFIIH” en TriTrypDB. Este caso se diferencia de los dos anteriores, donde siempre teníamos un único clúster con SRBH. En esta situación, existe tal divergencia dentro de kinetoplástidos, que los parámetros de 50% identidad en 80% cobertura es demasiado exigente y separa homólogos de distintas especies en dos clústeres. Incluso siendo que estos genes son de copia única en eucariotas (no sabemos estrictamente cual es la situación en kinetoplástidos) podríamos estar ante ortólogos. De hecho, el primer clúster mencionado engloba a los géneros *Leishmania, Leptomonas, Crithidia* y *Porcisia*, mientras que el segundo clúster engloba del género *Trypanosoma* las especies *T. cruzi* y *T. rangeli*. Dado que utilizamos una estrategia de SRBH, si se utilizara una única especie de referencia o un proteoma agrupado único, solo lograríamos anotar uno de los clústeres. Mediante la estrategia de utilizar 45 organismos de referencia de forma independiente, logramos evitar que en algunos casos el SRBH identifique un clúster y en otros otro clúster. Esto no es

necesariamente en todos los casos cierto, y se debería de explorar en mayor profundidad.

La anotación del clúster TcCLB.508707.149 se basa en estudios experimentales en *T. brucei*, donde mediante purificación por afinidad en tándem seguida de espectrometría de masas y análisis supervisado de secuencias se identificó a Tb927.11.16070 como Tfb4 [86]. Cabe destacar que los autores informaron que las búsquedas BLASTP no detectaron homología fuera de los tripanosomátidos.

El gen de *T. brucei* está ausente en nuestros clústeres identificados, pero ambos clústeres y Tb927.11.16070 pertenecen al mismo grupo de ortólogos definido por OrthoMCL (OG6_158123) base de datos que utiliza una estrategia más permisiva para establecer sus agrupamientos. De este agrupamiento es que resulta la extensión de la anotación en el clúster que incluye a los trypanosomas, pero no en el otro. El Tfb4 del organismo de referencia se caracteriza por los dominios InterPro "subunidad Tfb4/GTF2H3 de TFIIH" (IPR004600) y "superfamilia de dominio similar al factor von Willebrand A" (IPR036465) (**Tabla S6**). Sin embargo, en las secuencias de kinetoplástidos de los clústeres identificados por similitud estructural no se detectaron dominios InterPro (**Tabla S6**), ni se obtuvieron hits significativos con BLASTP.

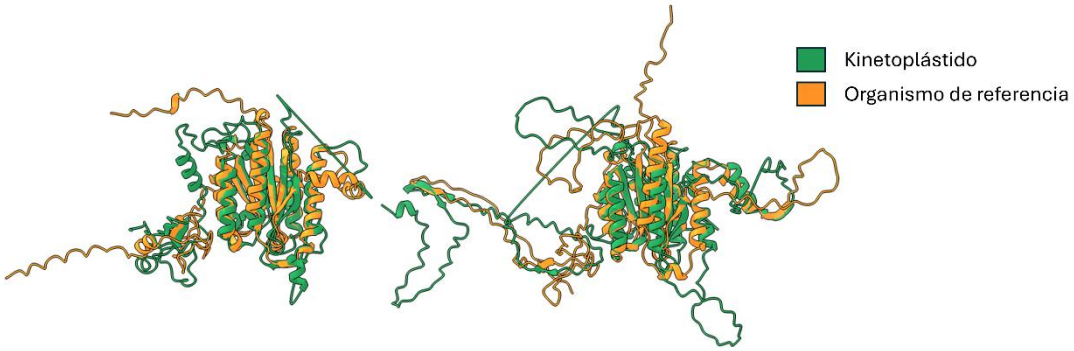

Figura 12. Alineamiento estructural de la Subunidad TFB4 del factor de transcripción basal TFIIH de *Saccharomyces cerevisiae* (izq) y *Ajellomyces capsulatus* (der) con homólogos en kinetoplástidos identificados mediante nuestro enfoque. La proteína del organismo de referencia se muestra en naranja, y la proteína de kinetoplástidos se muestra en verde. Izquierda, códigos de UniProt: Q12004_YEAST y Q4E262_TRYCC (TcCLB.508707.149). Derecha, códigos de UniProt: C0NNU2_AJECG y A0A0N1IME3_LEPSE (Lsey_0010_0360).

Este ejemplo, además de incluir dos clústeres, tiene la particularidad de que no encontramos información de resultados experimentales para ningún miembro de los clústeres. La evidencia experimental deriva de *T. brucei* y de su vínculo mediante la base de datos OrthoMCL. Lo cual implica una validación más indirecta que los ejemplos antes mencionados pero consistente. Al mismo tiempo, los resultados de TrypTag sugieren la misma identidad con localización subcelular en nucleoplasma, citoplasma (leve) y núcleo (puntos) (consulta en TrypTag: Tb927.11.16070).

Subunidad 2 del factor de iniciación de la transcripción TFIID

TFIID es un factor de transcripción esencial que reconoce el promotor central en células eucariotas y actúa como andamiaje para el ensamblaje de otros factores de transcripción, formando así el complejo de preiniciación (PIC, por sus siglas en inglés) necesario para iniciar la transcripción [87,88]. Tiene una estructura trilobulada y está compuesto por la proteína de unión a la caja TATA (TBP, por sus siglas en inglés) y trece factores asociados a TBP (TAF1 a TAF13, por sus siglas en inglés), numerados según su peso molecular en orden descendente [89]. Seis de estos TAFs están presentes por duplicado, lo que confiere al complejo una simetría bilateral. La mayoría de los TAFs dimerizan mediante el dominio de plegamiento de histona (HFD, por sus siglas en inglés), una estructura conservada de tres hélices α que facilita interacciones heterodiméricas específicas [89].

Dentro de TFIID, la subunidad TAF2 desempeña un papel central en la identificación del promotor [87,88]. Su dominio similar a una aminopeptidasa posee bucles ricos en arginina y lisina que se unen a la secuencia iniciadora (Inr, por sus siglas en inglés) del promotor [90,91]. Junto con TAF1 y TAF7, forma un subcomplejo de unión al ADN promotor, reforzando la especificidad de TFIID. Al interactuar directamente con el Inr, TAF2 estabiliza la unión de TFIID al ADN y asegura la selección precisa del sitio de inicio de la transcripción [90,91].

En nuestros resultados identificamos homólogos estructurales de TAF2, el segundo componente de mayor peso molecular del complejo TFIID (**Figura 13**). El representante del clúster de kinetoplástidos, LINF_120016700 con un largo de 1.372 aminoácidos, está actualmente anotado como “proteína similar a aminopeptidasa sensible a puromicina”, una descripción que no permite la identificación de TAF2. Este es un ejemplo de proteínas que consideramos PAnot pero que la información no es suficientemente precisa como la que aportamos en este trabajo.

El clúster está compuesto por los géneros *Crithidia, Endotrypanum, Leishmania, Leptomonas* y *Porcisia*. Nuestro enfoque SRBH sugiere que LINF_120016700 (E9AGI7_LEIIN) comparte similitud estructural con TAF2 de varios organismos de referencia como: *Rattus norvegicus* y *C. albicans*, entre otros (**Tabla S5**). Sin embargo, la

anotación de dominios basada únicamente en modelos HMM solo puede ubicar con confianza estas proteínas dentro de la misma familia o superfamilia que TAF2, dado que ambas contienen los dominios “Peptidasa M4/M1, superfamilia CTD” (IPR027268) y “Superfamilia de dominio N-terminal tipo aminopeptidasa N” (IPR042097). Cabe destacar que el dominio específico de TAF2 en InterPro (IPR037813) no reconoce a ninguna de las secuencias del clúster LINF_120016700/E9AGI7_LEIIN (**Tabla S6**). Esto sugiere que la identidad precisa de la proteína no puede ser inferida mediante el uso de los perfiles características de InterPro.

Análisis complementarios de BLASTP contra la base de datos de NCBI arrojaron e-valores significativos (~1e–60) para secuencias anotadas como “aminopeptidasa sensible a puromicina”. No obstante, estos hits se limitaron a la región N-terminal de ~500 aa (~40 % de cobertura), correspondiente al dominio tipo aminopeptidasa N, con solo ~30 % de identidad de secuencia. Por otro lado, no encontramos en la bibliografía información sobre los genes del clúster. Esto pone de manifiesto las limitaciones de los métodos basados en secuencia para identificar homólogos de proteínas altamente divergentes, que nuestro enfoque de homología estructural asocia con TAF2.

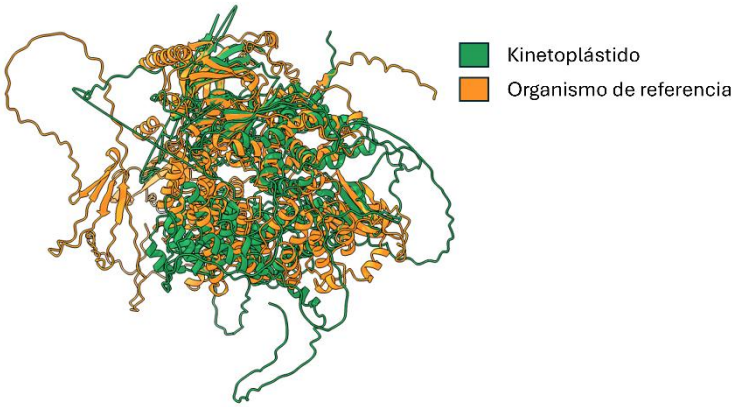

Figura 13. Alineamiento estructural de la subunidad 2 del factor de iniciación de la transcripción TFIID de *Candida albicans* con homólogo en kinetoplástidos identificados mediante nuestro enfoque. La proteína del organismo de referencia se muestra en naranja, y la proteína de kinetoplástidos se muestra en verde. Códigos de UniProt: A0A1D8PQF6_CANAL y E9AGI7_LEIIN (LINF_120016700).

Subunidad 3 del Factor de Estimulación de Corte (CstF-77 o CSTF3)

En las células de mamíferos, la formación de ARNm maduros depende, entre otras cosas, del procesamiento preciso del extremo 3’ de los transcritos precursores. El corte y la poliadenilación constituyen un paso crucial en esta maduración, ya que garantizan que los transcritos sean correctamente escindidos y dotados de una cola poli(A) que determina la estabilidad del ARNm, su exportación y su eficiencia traduccional [92]. En parte determina la diversidad del transcriptoma generada mediante la utilización de sitios alternativos de inicio de transcripción, empalmes de exones variables y el procesamiento alternativo del extremo 3’ del ARN. Este último constituye un mecanismo de regulación clave tanto en el desarrollo normal como en diversas patologías [93]. El procesamiento del extremo 3’ de la mayoría de los ARNm comprende primero el corte del ARN naciente y, a continuación, la adición de una cola homopolimérica de 150–250 residuos de adenosina, en un proceso conocido como corte y poliadenilación [94]. El Factor de estimulación de corte (CstF) es un complejo proteico esencial de la gran maquinaria de procesamiento del extremo 3’ del ARN pre-mensajero, imprescindible para una poliadenilación eficiente y la escisión del extremo 3’ en ARN pre-mensajeros [94,95]. El complejo CstF está formado por tres proteínas: CstF-50, CstF-64 y CstF-77 [95]. Entre ellas, CstF-77 alias CstF3 desempeña un papel estructural fundamental, al actuar como andamio que enlaza CstF-64 y CstF-50 con otro complejo proteico denominado “Factor de especificidad de corte y poliadenilación” (CPSF), garantizando así el correcto ensamblaje del complejo [92]. Además, CstF-77 es la única subunidad que posee una señal de localización nuclear (NLS), imprescindible para el transporte de todo el complejo al núcleo [96].

En nuestros resultados encontramos dos Dark Cluster representados por LINF_320046700 y TcCLB.504005.50 como homólogos estructurales de CstF3, con largos de 994 y 622 aminoácidos de largo respectivamente (**Figura 14**). Al igual que sucede con el caso de Tfb4, nuestros clústeres basados en identidad de secuencia separan los resultados dentro de los kinetoplástidos en dos grupos. El primero de los clústeres contiene los géneros *Crithidia, Endotrypanum, Leishmania, Leptomonas* y *Porcisia* y el otro el género *Trypanosoma*. Ninguno de los clústeres incluye a *B. saltans* lo que tal vez indique que el CstF3 de este organismo es suficientemente divergente como para no

agruparse con ninguno de los anteriores. Siendo que es un gen de vital importancia seria de especial interés buscar CstF3 en *B. saltans* ya sea con las secuencias identificadas en este trabajo o por estructura.

Una diferencia sustancial es que la mayoría de los SRBH fueron con el clúster de LINF_320046700 el cual obtuvo diez SRBH de Foldseek y los cinco mejores aciertos contra CstF3 de distintas especies. En el caso del clúster TcCLB.504005.50 solamente obtuvo un SRBH el cual pasó todos los filtros establecidos en el pipeline también contra CstF3 en este caso de *Zea mays*. En este caso el único acierto contra el clúster de *Trypanosoma* nos permitió identificar este gen en estos organismos. Igualmente, esto sugiere que en otros casos posiblemente uno de los dos clústeres se lleve todos los SRBH y, por lo tanto, perdamos capacidad de anotación. Como se mencionó anteriormente, se podría incorporar para una versión futura un paso previo de agrupamiento estructural dentro de los kinetoplástidos para evitar este tipo de separación de homólogos en subclusteres.

Para identificar a CstF3 la base de datos InterPro incluye un HMM nombrado “proteína de procesamiento del extremo 3’ de ARNm similar a Rna14” (IPR045243). Ambos clústeres de kinetoplástidos son reconocidos por IPR045243; sin embargo, el algoritmo automático de anotación de UniProt no asigna la descripción correspondiente. Además, como referencia, todos los mejores aciertos en organismos modelo contienen el dominio “Supresor de forked” (IPR008847), que no se detecta en nuestros clústeres (**Tabla S6**). En estos casos, nuestro enfoque aporta evidencia adicional sobre la identidad de las proteínas, aunque no es tan determinante para inferir homología como en los ejemplos anteriores. Los resultados de BLASTP arrojaron pocos hits, con baja diversidad taxonómica, aproximadamente 20 % de cobertura (coincidente con el dominio “Supresor de forked”), ~25 % de identidad de secuencia y e-valores superiores a 1e-6. La mayoría de los hits carecían de anotaciones, salvo uno en *Tieghemostelium lacteum* (KYQ88973.1), etiquetado como “subunidad 3 del factor de estimulación de corte”.

Dentro del clúster que engloba al género *Trypanosoma* tenemos el gen Tb927.11.12050 de *T. brucei* pero su localización subcelular no logro ser identificada en TrypTag.

Figura 14. Alineamiento estructural de la subunidad 3 del factor de estimulación de corte de *Homo sapiens* (izq) y *Zea mays* (der) con homólogos en kinetoplástidos identificados mediante nuestro enfoque. La proteína del organismo de referencia se muestra en naranja, y la proteína de kinetoplástidos se muestra en verde. Izquierda, códigos de UniProt: CSTF3_HUMAN y A4I8L3_LEIIN (LINF_320046700). Derecha, códigos de UniProt: A0A1D6M355_MAIZE y Q382X6_TRYB2 (Tb927.11.12050).

Proteína 4 del tráfico de Golgi al retículo endoplasmático (GET4)

El complejo de reconocimiento de dominio transmembrana (TRC, por sus siglas en inglés) es la vía principal encargada de guiar las proteínas ancladas por su cola C-terminal (TA, de *tail-anchored*) desde el citosol hasta sus membranas diana, que incluyen el retículo endoplasmático, el aparato de Golgi y las mitocondrias [97]. Debido a que las proteínas de membrana deben atravesar un entorno citosólico acuoso, corren el riesgo de agregarse, plegarse incorrectamente o localizarse en compartimentos equivocados. Para evitar estos problemas, intervienen chaperonas y co-chaperonas especializadas que protegen los segmentos transmembrana expuestos, seleccionan de forma precisa a sus sustratos y los guían activamente hasta la membrana apropiada [97]. En el caso de la vía TRC, la chaperona SGTA se une primero a la cola C-terminal hidrofóbica liberada del ribosoma y la transfiere al complejo heterotetramérico GET4/GET5, que actúa como un "*pre-targeting complex*". A continuación, GET4 entrega la proteína TA a la ATPasa GET3, la cual facilita su inserción en la membrana del retículo endoplasmático. GET4 cumple un papel clave en la etapa inicial de la vía TRC: su región N-terminal se acopla tanto con GET3 como con el resto del complejo GET4/GET5 (levadura) o GET4/GET5/BAG6 (mamíferos), lo que garantiza una transferencia precisa de la proteína TA hacia la ATPasa [98]. Cuando GET4 sufre mutaciones patogénicas, la integridad de este complejo se ve comprometida, reduciendo drásticamente la capacidad de dirigir proteínas al aparato de Golgi y provocando defectos en el tráfico retrogrado entre el retículo endoplasmático y el Golgi [99].

Al momento, GET4 ha sido descripta como conservada en eucariotas dado que fue identificada en plantas, hongos, mamíferos y ronda los 300 aminoácidos [100], pero no ha sido reportada en kinetoplástidos. En nuestros resultados dos clústeres obtuvieron SRBH con genes de OrgRefs correspondientes al grupo BUSCO "Proteína 4 del tráfico de Golgi al retículo endoplasmático" (**Figura 15**). El primero, representado por TcCLB.510187.140 (313 aminoácidos), un clúster compuesto únicamente por secuencias de *T. cruzi.* Este tuvo entre sus mejores hits GET4 correspondiente a *Arabidopsis thaliana* y *Dictyostelium discoideum* (entre otros) con valores TM-scores alrededor de 0,7. Los resultados de BLASTP no arrojó hits contra NCBI y no se detectaron dominios con InterProScan (**Tabla S6**).

El segundo clúster está representado por LtaPh_3406000 (389 aminoácidos) e integrado, al igual que en casos anteriores, por los géneros *Crithidia, Endotrypanum, Leishmania, Leptomonas y Porcisia*. En esta oportunidad tampoco se observaron hits en BLASTP, pero sí presenta un emparejamiento significativo con el HMM de "proteína de anclaje de NSF" (IPR000744), dominio identificado en los GET4 de RefOrgs (**Tabla S6**). De manera coherente, muchos miembros de este clúster están anotados como "proteína gamma-soluble de anclaje de NSF (SNAP-gamma)". En nuestros resultados, tres de los cuatro mejores aciertos por SRBH son identificados como GET4 (IPR007317), mientras que el restante figura como "proteína gamma-soluble de anclaje de NSF". Huelga decir que ambos dominios/HMM (el de "proteína de anclaje de NSF" y el de GET4) pertenecen a la superfamilia de dominio helicoidal tipo tetratricopeptídico (IPR011990). Este es el único caso entre nuestros "estudios de caso" en el que se presentan ambigüedades de anotación, lo que indica que se justifica un análisis más detallado.

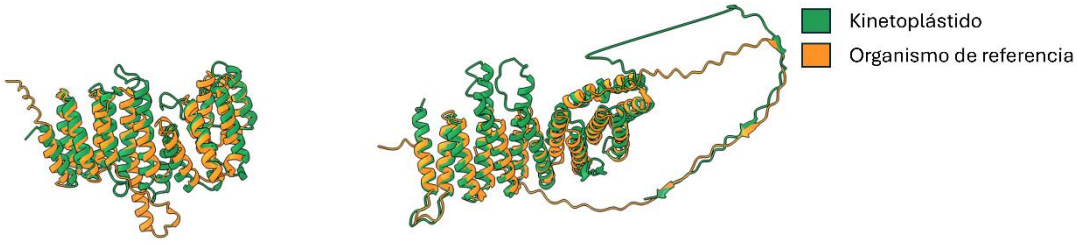

Figura 15. Alineamiento estructural de la Proteína 4 del tráfico de Golgi al retículo endoplasmático de *Arabidopsis thaliana* (izq) y *Schistosoma mansoni* (der) con homólogos en kinetoplástidos identificados mediante nuestro enfoque. La proteína del organismo de referencia se muestra en naranja, y la proteína de kinetoplástidos se muestra en verde. Izquierda, códigos de UniProt: GET4_ARATH y A0A2V2WCY5_TRYCR (TcCLB.510187.140). Derecha, códigos de UniProt: A0A3Q0KFW4_SCHMA y A0A640KRB8_LEITA (LtaPh_3406000).

E3 Ubiquitina-Proteína Transferasa MAEA (MAEA)

La proteína "E3 Ubiquitina-Proteína Transferasa MAEA" (MAEA) ronda los 400 aminoácidos de largo y actúa como subunidad esencial del complejo CTLH (*C-terminal to LisH*). Este complejo es fundamental para mantener la homeostasis proteica y regular la diferenciación y proliferación celular [101]. MAEA colabora con la proteína RMND5A para transferir ubiquitina desde la enzima E2 UBE2H hacia sustratos específicos, promoviendo la formación de cadenas de poliubiquitina que marcan estas proteínas para su degradación por el proteasoma [102]. Uno de sus blancos bien caracterizados es el factor de transcripción HBP1; su ubiquitinación por el complejo CTLH conduce a la degradación de HBP1, lo que alivia la represión de genes del ciclo celular y favorece un estado proliferativo [103]. Más allá de su papel en el control transcripcional, MAEA es crucial para la eritropoyesis. En la hematopoyesis mamífera, MAEA contribuye a la enucleación de los eritroblastos: el proceso por el cual los precursores inmaduros de glóbulos rojos expulsan su núcleo para convertirse en eritrocitos funcionales y anucleados. A través del complejo CTLH, MAEA coordina los eventos de reorganización del citoesqueleto y las interacciones célula-célula necesarios para este paso final de maduración [101]. En levaduras, el homólogo de MAEA es Gid9, una subunidad del complejo Gid (*glucose-induced degradation deficient*) en *Saccharomyces cerevisiae*. Al igual que su contraparte en mamíferos, Gid9 forma parte de un ensamblaje ligasa E3 que dirige proteínas reguladoras clave hacia la ubiquitinación y degradación [104].

En nuestros resultados el clúster representado por TcCLB.504253.20 (615 aminoácidos), obtuvo en sus cinco mejores aciertos proteínas anotadas como MAEA en las especies *R. norvegicus, Mus musculus, Homo sapiens, Schistosoma mansoni* y *S. cerevisiae* (**Figura 16**). Lo que resulta en fuerte evidencia de la identidad de este clúster. Sin embargo, en este clúster prácticamente todas las secuencias están anotadas en TriTrypDB como hipotéticas y en UniProt como "CTLH domain-containing protein". Para la identificación de MAEA deberían detectarse dominios clave como CTLH_C (IPR006595), CTLH/CRA (IPR024964), CRA_dom (IPR013144) y ZF_RING_GID (IPR044063); sin embargo, estos dominios están ausentes en todos los miembros del clúster TcCLB.504253.20. No obstante, el análisis con InterProScan reveló que el HMM informativo y específico para MAEA (IPR045098) que abarca toda la secuencia logra identificar a los genes de

kinetoplástidos, respaldando la misma identidad sugerida por nuestro enfoque (**Tabla S6**). Al mismo tiempo la localización del gen correspondiente a *T. brucei* del clúster se localiza en citoplasma y endocítica (**Figura 17**), concordando con lo esperado para MAEA. No se obtuvieron hits en BLASTP contra NCBI.

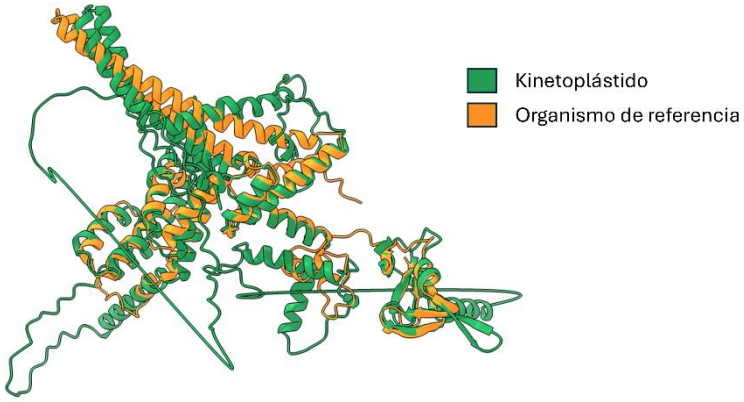

Figura 16. Alineamiento estructural de E3 Ubiquitina–Proteína Transferasa MAEA de *Homo sapiens* con homólogo en kinetoplástidos identificados mediante nuestro enfoque. La proteína del organismo de referencia se muestra en naranja, y la proteína de kinetoplástidos se muestra en verde. Códigos de UniProt: MAEA_HUMAN y Q4D4T7_TRYCC (TcCLB.504253.20).

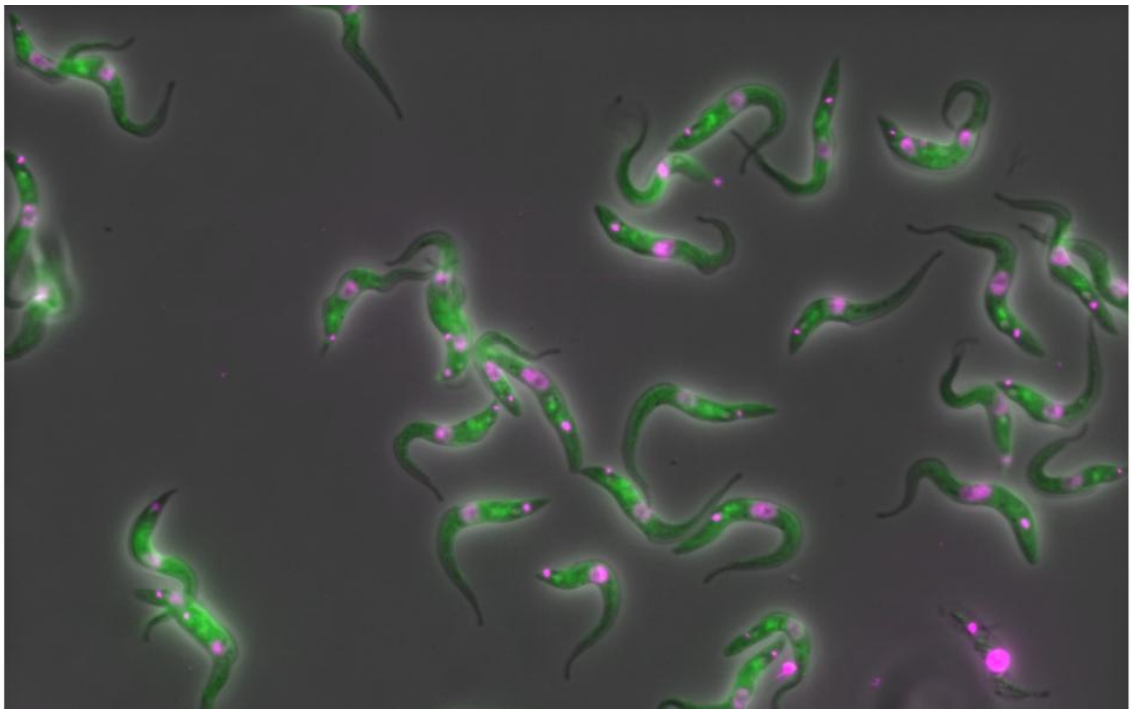

Figura 17. Localización subcelular del homologo estructural de la proteína MAEA en los datos de TrypTag de *T. brucei*. Imágenes de microscopía de fluorescencia de proteínas de *T. brucei* (marcadas con mNeonGreen), mostrando su localización subcelular determinada por TrypTag. Verde (mNeonGreen): proteína de interés, Cian (Hoechst): ADN. Para Tb927.4.2890 la localización reportada es citoplasma y endocítica (baja señal).

Subunidad no catalítica de la metiltransferasa de tRNA (guanina-N(7)-) (WDR4)

WDR4, también conocido como la subunidad no catalítica de la metiltransferasa de ARNt (guanina-N(7)-), es una proteína que ronda los 400 aminoácidos y desempeña un papel fundamental como andamiaje en los complejos de metiltransferasas de ARN. En el heterodímero METTL1–WDR4, WDR4 dirige la especificidad de sustrato y estabiliza el ensamblaje del complejo, asegurando que METTL1 pueda transferir un grupo metilo desde S-adenosilmetionina a la posición N7 de la guanina. Concretamente, WDR4 posiciona los ARNt para que la modificación m7G se incorpore en el nucleótido 46 del bucle variable de los ARNt que contienen el motivo 5'-RAGGU-3'. El m7G46 resultante interactúa con el par de bases C13-G22 en el bucle D, reforzando la estructura terciaria del ARNt y protegiéndolo de la degradación [105,106]. Más allá de su función canónica en la metilación de ARNt, WDR4 se ha implicado en eventos adicionales de metilación de ARN en ARNm y microARNs, y ejerce funciones no enzimáticas en la estabilidad genómica [107].

Obtuvimos dos clústeres con SRBH contra proteínas de OrgRef anotadas como WDR4 (**Figura 18**). El primero de los clústeres es un Dark Cluster (con la excepción de una secuencia con información parcial) y obtuvo SRBH con las especies *M. musculus, Danio rerio, Schizosaccharomyces pombe, O. sativa* y *S. cerevisiae*. El mismo está representado por la estructura del gen TcCLB.507711.120 (491 aminoácidos) y engloba varios genomas de *T. cruzi* y uno de *T. rangeli*. El segundo clúster, representado por TcIL3000_10_10140 de 462 aminoácidos de largo, obtuvo solo un SRBH únicamente contra *C. albicans* y está compuesto por *T. brucei, T. congolense, T. equiperdum* y *T. evansi* por lo que nuestro paso de agrupamiento dividió secuencias dentro del género *Trypanosoma*. Demostrando que incluso para genes de copia única y tan relevantes en kinetoplástidos se puede encontrar altos niveles de diversidad. Al mismo tiempo, no obtuvimos clústeres con el resto de los organismos con SRBH contra WDR4. Pero todo indica que deberían de estar presentes en los genomas, por lo que se podría intentar anotar utilizando las secuencias ahora descriptas o utilizar métodos estructurales alternativos.

Cuando analizamos las secuencias mediante InterProScan encontramos que el HMM específico de WDR4 (IPR028884) reconoce miembros de esos clústeres (**Tabla S6**), y los resultados de BLASTP corroboran esta anotación, aunque los parámetros estadísticos se sitúan en la zona gris para considerar un hallazgo BLASTP significativo para determinar homología. La localización subcelular en *T. brucei* es principalmente lumen nuclear (**Figura 19**). Por lo tanto, se pude concluir que a pesar de los algoritmos automáticos no logran anotar estas secuencias como WDR4, existe evidencia para su identificación mediante estos métodos. En este caso nuestros resultados estructurales constituyen una prueba más de que efectivamente estos genes son WDR4.

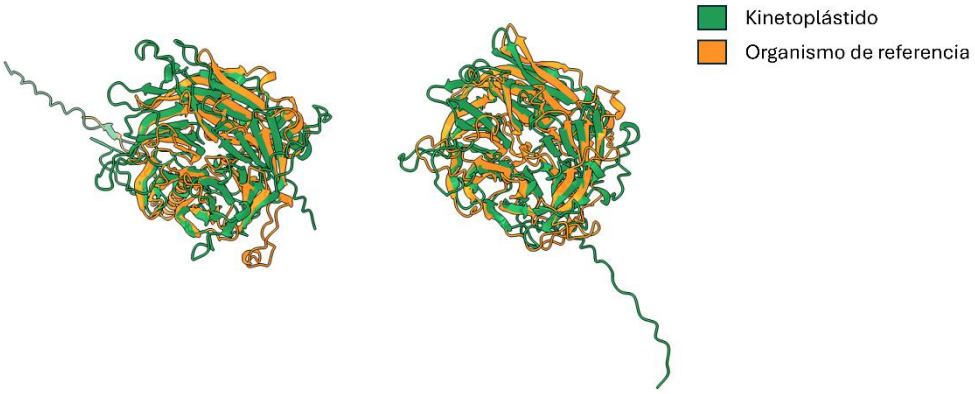

Figura 18. Alineamiento estructural de la subunidad no catalítica de la metiltransferasa de tRNA (guanina-N(7)-) de *Mus musculus* (izq) y *Candida albicans* (der) con homólogos en kinetoplástidos identificados mediante nuestro enfoque. La proteína del organismo de referencia se muestra en naranja, y la proteína de kinetoplástidos se muestra en verde. Izquierda, códigos de UniProt: WDR4_MOUSE y Q4DZR6_TRYCC (TcCLB.507711.120). Derecha, códigos de UniProt: TRM82_CANAL y G0UXW8_TRYCI (TcIL3000_10_10140).

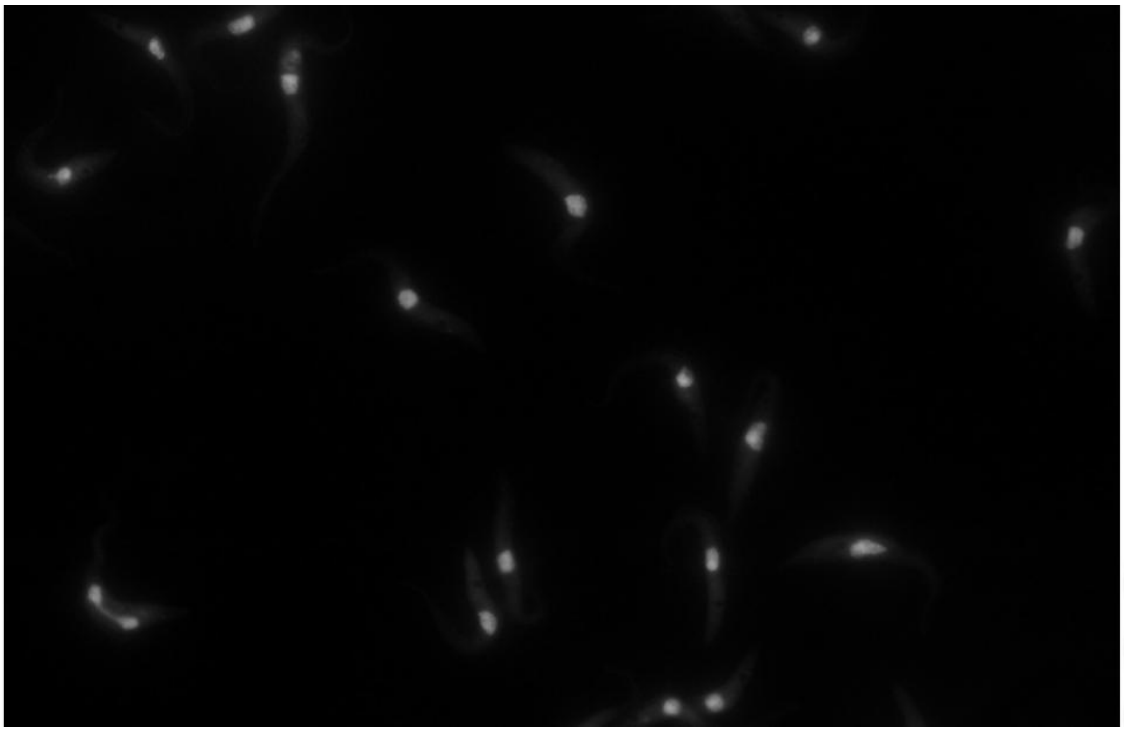

Figura 19. Localización subcelular del homologo estructural de la proteína WDR4 en los datos de TrypTag de *T. brucei*. Imágenes de microscopía de fluorescencia de proteínas de *T. brucei* (marcadas con mNeonGreen), mostrando su localización subcelular determinada por TrypTag. Blanco (mNeonGreen): proteína de interés. Para Tb927.10.11210 la localización reportada es lumen nuclear, citoplasma (baja señal) y citoplasma flagelar (baja señal).

Co-chaperona Hsc20/HscB/Jac1

La proteína co-chaperona HscB (bacteria), también conocida como HSC20 en mamíferos, desempeña un papel esencial en la biogénesis y transferencia de cúmulos hierro-azufre (Fe–S), fundamentales para la maduración de proteínas Fe–S en los compartimentos mitocondrial y citoplasmático. Aunque la mayor parte de estos cúmulos se sintetiza en las mitocondrias, HscB/HSC20 también participa activamente en la vía citosólica, donde contribuye a incorporar los Fe–S en numerosas proteínas involucradas en procesos metabólicos y reguladores [108].

En las mitocondrias, el ensamblaje de cúmulos Fe–S comienza con la síntesis de cúmulos nacientes sobre la proteína "*iron–sulfur cluster scaffold protein*" (ISCU). Para ello, una enzima desulfurasa de cisteína dona azufre al andamio, y posteriormente el hierro se incorpora para formar el cúmulo preliminar. HscB/HSC20 actúa inmediatamente después de esta etapa inicial: se une preferentemente a la forma ISCU que ya contiene el cúmulo y facilita su transferencia a través de su interacción con el chaperón Hsp70 específico de la mitocondria, conocido como HSPA9 en mamíferos [109]. En su función de co-chaperona tipo J, HscB utiliza su dominio J N-terminal para estimular la actividad ATPasa de HSPA9. La hidrólisis de ATP por parte de HSPA9 induce cambios conformacionales que permiten liberar el cúmulo Fe–S de ISCU y entregarlo a las proteínas receptoras que lo necesitan dentro de la mitocondria [108].

Sin embargo, la mitocondria no es el único lugar donde se ensamblan y transfieren estos cúmulos. En el citosol existe una vía paralela denominada Máquina de Ensamblaje de Hierro-Azufre Citosólica (CIA, por sus siglas en inglés) y una fracción significativa se localiza en el citoplasma, desempeñando allí una función análoga para la formación y transferencia de cúmulos Fe-S a proteínas citosólicas y nucleares [110]. En este compartimento, HscB/HSC20 interactúa directamente con la misma proteína andamio ISCU y con componentes específicos de la vía CIA, como CIAO1. Esta interacción molecular sirve de puente entre las vías mitocondrial y citosólica de biogénesis de Fe-S [110].

En este trabajo identificamos homólogos de la co-chaperona Hsc20/HscB/Jac1 separados en tres clústeres de kinetoplástidos con TM-scores entre 0,6 y 0,7 (todas las

secuencias involucradas rondan los 230-250 aminoácidos de largo) (**Figura 20**). Los representantes de los clústeres fueron LtaPh_3329200, LmjF.25.1690 y Tb927.3.1760. En TriTrypDB, al menos un miembro de estos clústeres está anotados a nivel de superfamilia como "proteína con dominio J" o equivalente. El primero de los clústeres, se aleja del largo antes mencionado donde su representante alcanza los 276 aminoácidos de largo y engloba los géneros *Crithidia, Endotrypanum, Leishmania, Leptomonas y Porcisia*. Sin embargo, obtuvo cuatro de sus cinco mejores SRBH anotados como Hsc20 o HscB. Interesantemente, el segundo también engloba estos géneros (e incluye *Blechomonas*) y comparte algunos de los genomas. Cuando observamos sus SRBH dos de sus mejores cinco hits están anotados como JAC1 el homólogo de Hsc20/HscB de hongos. Esto es inusual dado que según BUSCO en cada organismo solo debe haber una copia de este gen. A modo de ejemplo, en el genoma "*Leishmania major Friedlin 2021*" encontramos en el primer clúster el gen LMJFC_330040500 y en el segundo LMJFC_250027500. En el genoma de "*Crithidia fasciculata strain Cf-Cl*" encontramos CFAC1_210037700 y CFAC1_230016200, respectivamente. Esto resulta interesante, en parte porque uno de los genes es más largo y además, porque es el único ejemplo de nuestro trabajo donde ocurre que encontramos dos copias de un gen que en teoría es copia única en la mayoría de los eucariotas. En este sentido, aún queda espacio para investigar sobre estos genes *in silico* y sería interesante/pertinente abordarlo experimentalmente. Por ultimo y espejando lo antes mencionado y discutido, el último de los clústeres engloba el género *Trypanosoma*.

En organismos de referencia, Hsc20/HscB/Jac1 incluye dos HMM distintos: HscB_oligo_C (IPR009073) y HscB (IPR004640). Es notable que en los miembros del clúster LtaPh_3329200 no se detectaron estos HMM característicos, mientras que en los otros dos clústeres se reconoció con éxito el HscB (IPR004640) (**Tabla S6**). Demostrando que al menos dos de los clústeres logran ser reconocidos por algunos HMM característicos. Por otro lado, los resultados de BLASTP arrojaron coincidencias etiquetadas como "proteína con dominio J" para LtaPh_3329200, "Jac1" y "co-chaperona HscB para ensamblaje de proteína Fe–S" para los otros dos clústeres, con *e-values* aproximados entre 1e-5 y 1e-6, coberturas del 20 % al 80 % e identidades de secuencia en torno al 30–40 %. Aunque estos hallazgos respaldan la identidad Hsc20/HscB/Jac1 para

LmjF.25.1690 y Tb927.3.1760 y la clasificación a nivel de superfamilia para LtaPh_3329200, la calidad general de los aciertos es moderada. En este caso particular la localización subcelular reportada es citoplasmática, aunque dispuesta en lugares discretos que parecen puntos (**Figura 21**).

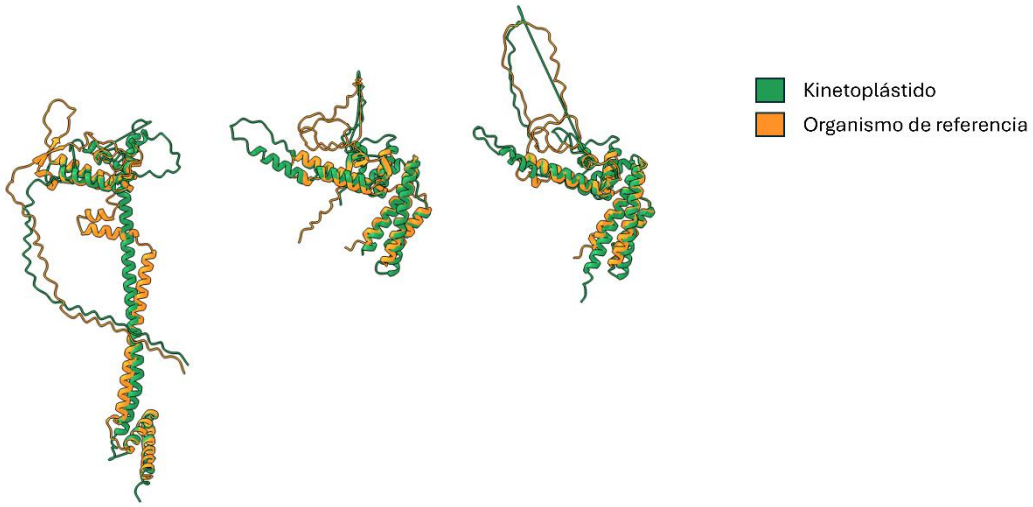

Figura 20. Alineamiento estructural de co-chaperona Hsc20/HscB/Jac1 de *Saccharomyces cerevisiae* (izq), *Mus musculus* (centro) y *Homo sapiens* (der) con homólogos en kinetoplástidos identificados mediante nuestro enfoque. La proteína del organismo de referencia se muestra en naranja, y la proteína de kinetoplástidos se muestra en verde. Izquierda, códigos de UniProt: JID1_YEAST y A0A640KWU3_LEITA (LtaPh_3329200). Centro, códigos de UniProt: HSC20_MOUSE y A0A088RUU2_9TRYP (LPAL13_000047100). Derecha, códigos de UniProt: HSC20_HUMAN y Q57ZD5_TRYB2 (Tb927.3.1760).

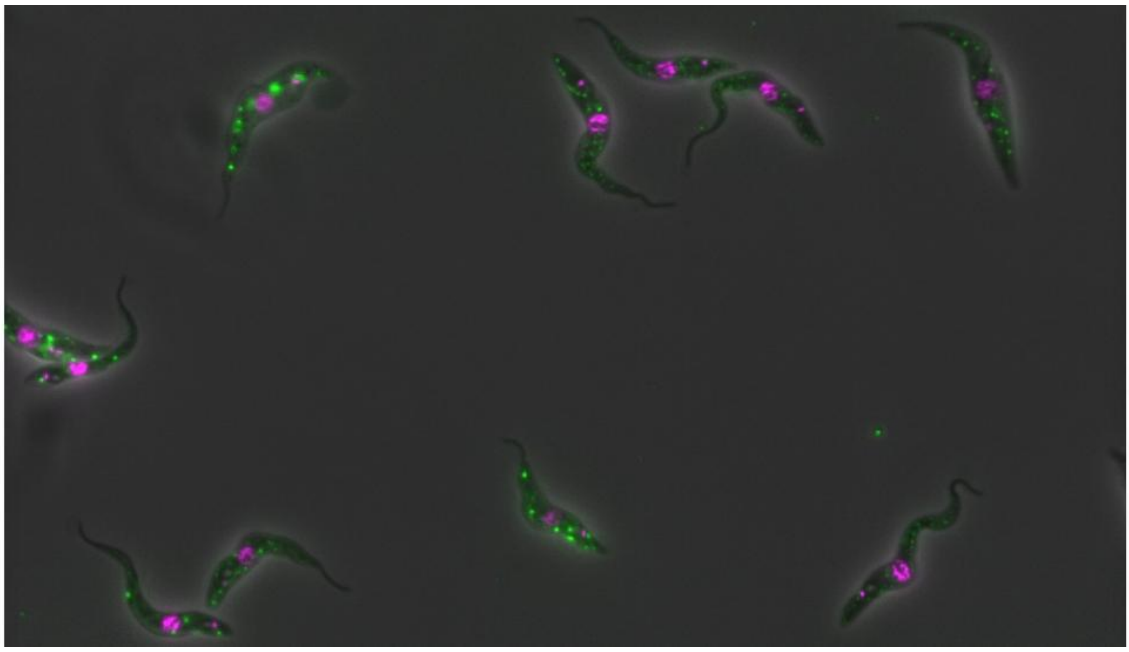

Figura 21. Localización subcelular del homologo estructural de la co-chaperona Hsc20/HscB/Jac1 en los datos de TrypTag de *T. brucei*. Imágenes de microscopía de fluorescencia de proteínas de *T. brucei* (marcadas con mNeonGreen), mostrando su localización subcelular determinada por TrypTag. Verde (mNeonGreen): proteína de interés, Cian (Hoechst): ADN. Para Tb927.3.1760 la localización reportada es citoplasma (en puntos).

Conclusiones

Este trabajo partió de la hipótesis de que la comparación estructural de proteínas permite evidenciar relaciones de homología remota entre genes cuya similitud a nivel de secuencia ha sido erosionada. Hasta hace poco tiempo, esta estrategia era poco factible a gran escala, dado que la comparación estructural dependía de modelos tridimensionales obtenidos experimentalmente, una tarea compleja y limitada. Esta situación cambió drásticamente en 2021 con la publicación de AlphaFold [54], una herramienta que permite predecir con alta precisión la estructura de una proteína a partir de su secuencia de aminoácidos. Estrictamente AlphaFold parte de un MSA para obtener información de covarianza entre aminoácidos y por ende predecir proximidad o interacción. Con esto busco destacar que AlphaFold continúa dependiendo, en parte, de la disponibilidad de secuencias homólogas para construir un MSA informativo. En este sentido, los kinetoplástidos presentan una situación particularmente favorable: si bien son muy divergentes respecto a organismos modelo, existe una gran cantidad de genomas secuenciados dentro del linaje, lo que permite generar alineamientos informativos usando solo otras secuencias de kinetoplástidos. Esto habilita la predicción de estructuras confiables incluso cuando no existe homología detectable con organismos externos. A partir de estas estructuras, buscamos realizar comparaciones tridimensionales con proteínas de referencia y revelar relaciones de homología que han escapado a los métodos basados en secuencia. Trabajos publicados al momento que han utilizado métodos de comparación estructural a nivel de genomas para establecer homología fueron realizados utilizando organismos de referencia cercanos al proteoma que se quería anotar, por ejemplo, en Porífera versus otros Opisthokonta [74] y Fungi versus *H. sapiens* [73,111]. En estos trabajos se comparó la capacidad de métodos basados en secuencia versus métodos basados en estructura y se evidenció poco incremento en la capacidad de anotación, lo cual en parte podría explicarse por el uso de especies cercanas donde los métodos de estructura tal vez no aporten sensibilidad o especificidad superiores a los de secuencia. Un abordaje similar pero basado en dominios logro identificar 400 dominios previamente no identificados en el genoma de

T. brucei [112], demostrando que la comparación estructural es más sensible que los HMM de Pfam en este modelo.

En este contexto, desarrollamos y evaluamos ASC, un pipeline diseñado para automatizar y escalar el análisis de homología estructural. ASC está implementado en Snakemake, lo que permite una ejecución reproducible y escalable. Además, por cómo está diseñado el flujo de trabajo no se precisa GPU ni gran poder de cómputo haciendo accesible la herramienta. ASC es amigable con el usuario el cual necesita habilidades minimas de programación o manejo de terminal para utilizarlo. El diseño del pipeline contempla dos insumos obligatorios: un archivo FASTA con las secuencias proteicas de cada especie y un archivo TSV que asocia cada encabezado a su acceso correspondiente en UniProt el cual se puede obtener para muchos organismos de la base de datos VEuPathDB. Las modificaciones más relevantes pueden ser orquestadas desde un `config file` facilitando el uso por el usuario. Nuestro algoritmo comparte similitudes con otros enfoques publicados, pero también presenta diferencias sustanciales. Dado que existen pocas herramientas dedicadas al modelado y comparación estructural a gran escala, la mayoría de las estrategias actuales se basan en AlphaFold y Foldseek, que se han consolidado como el estándar en el área. Sin embargo, nuestro abordaje incorpora un paso adicional que consideramos especialmente relevante: la comparación directa y reciproca entre estructuras tridimensionales mediante archivos PDB, utilizando herramientas clásicas como FATCAT y TM-align para una validación final rigurosa. De este modo, nuestra estrategia combina la velocidad y sensibilidad de los métodos modernos con la robustez de técnicas consolidadas, aplicando umbrales claros de calidad estructural, como un TM-score > 0.5. Esta integración representa una de las principales diferencias y fortalezas frente a otros métodos existentes.

Luego de su desarrollo, ASC fue aplicado de forma exhaustiva a los proteomas disponibles en TriTrypDB. Estos organismos fueron seleccionados por su relevancia para nuestras líneas de investigación, pero también por tratarse de un grupo

filogenéticamente divergente y con abundancia de genomas secuenciados. Los proteomas de estos organismos fueron evaluados en conjunto formando una base de datos conjunta contra “organismos de referencia” (OrgRef) también referidos en AFDB como “organismos modelo”. Los resultados los evaluamos utilizando dos estrategias complementarias una basada en identificación de dominios por HMM (InterPro) y la otra utilizando resultados experimentales de localización subcelular en la base de datos TrypTag. La primera de las estrategias mostro que los SRBH son principalmente idénticos en términos de dominios presentes. Al mismo tiempo se observó que una proporción significativa de los resultados mostraba una falta de sensibilidad en la detección de dominios en kinetoplástidos caracterizados en la categoría K_in_M, consistente con lo reportado [112]. Esta evidencia, puede explicar poqué algoritmos automáticos de anotación no logran anotar gran parte de los genomas de kinetoplástidos, a modo de ejemplo, el pipeline de UniProt basado en UniRules utiliza esta identificación de dominios de InterPro para la anotación funcional de genes.

El segundo método de evaluación consto de la comparación de localización subcelular disponible de resultados experimentales en TrypTag con la anotación que nosotros le daríamos a los genes basados en nuestra estrategia. Este resultado, demostró que nuestra capacidad de anotación es igual de buena y en algunas localizaciones subcelulares mejor que los métodos actuales.

En sumatoria, ambos métodos de evaluación sugieren que ASC es una buena estrategia a la hora de buscar homología estructural en kinetoplástidos. Al mismo tiempo, nos resulta imposible saber si estos homólogos van a cumplir la misma función por muchas razones, la más evidente es la divergencia a nivel de secuencia, pero también todas las particularidades de la biología celular y molecular de estos parásitos. Este tipo de preguntas solo pueden ser abordadas experimentalmente y escapan a los objetivos de esta tesis.

Es importante destacar que, dado que esta es una de las primeras veces que se aplica esta estrategia en estos organismos, optamos por un enfoque exigente. El objetivo fue obtener resultados robustos e informativos, evitando casos complejos o ruidosos. Esto nos permitió evaluar si estas estrategias son realmente útiles para estos organismos y,

en caso afirmativo, plantear su extensión o la exploración de casos que quedaron fuera del criterio de rigurosidad actual.

En primer lugar, el método RBH presenta diversas limitaciones, especialmente en familias multigénicas, inparalogos, entre otros [113]. Sin embargo, la ventaja principal es que cuando se produce un acierto, este resulta altamente confiable y se suele asumir como un método heurístico para identificar ortólogos. Para aumentar esa confianza, incorporamos un paso adicional de alineamiento estructural con FATCAT y TM-align. A estos resultados les aplicamos un filtro exigente: un TM-score ≥ 0.5 en ambos sentidos (cadena 1 vs. cadena 2 y viceversa). Esto garantiza que los aciertos tienen longitudes similares y un alineamiento extenso. Nuevamente, la estrategia demuestra una alta confiabilidad, aunque implica que quedan excluidos algunos genes, en particular aquellos con errores de predicción, especialmente genes pequeños, donde dichos errores representan una parte significativa de la longitud total de la proteína. Además, las regiones sin estructura, que suelen mostrar mayor variación en longitud e identidad, pueden comprometer este filtro en proteínas homólogas, provocando que algunas sean descartadas.

Para evaluar de manera objetiva la sensibilidad del método y evitar seleccionar ejemplos de éxito sesgados, incorporamos un análisis sistemático mediante BUSCO. BUSCO es una herramienta utilizada para identificar genes ortólogos de copia única, en este trabajo vamos a usar una base de datos que engloba toda la diversidad eucariota. Es pertinente indicar que estos genes por sus características antes mencionadas suelen estar involucrados en procesos celulares extremadamente relevantes para la biología del organismo. Dentro de los grupos de BUSCO muchos de estos genes ya estaban anotados en los genomas de kinetoplástidos. Pero encontramos un subconjunto conformado por 9 grupos BUSCO de los cuales no había ningún tipo de anotación para los genes de kinetoplástidos y mediante ASC pudimos detectar homólogos estructurales. En algunos casos, estos hallazgos fueron validados a través de evidencias experimentales, bibliografía o métodos *in silico* complementarios. No encontramos evidencia que contradijera estas asignaciones, lo que sugiere que ASC es capaz de rescatar funciones genuinas ocultas por niveles extremos de divergencia de secuencia.

En este proceso también se generó una gran cantidad de anotaciones complementarias para genes que ya contaban con algún grado de anotación funcional. Si bien no abordamos en detalle estos casos en este trabajo, es importante destacar que ASC no solo permite detectar nuevos genes, sino también enriquecer las anotaciones existentes. La evaluación semántica de los "grados" de anotación funcional es compleja y queda como un aspecto interesante a desarrollar en futuras versiones del análisis de datos. Como parte de este trabajo, se generó una base de datos web pública denominada KASC donde se puede consultar cada gen anotado, sus estructuras, alineamientos y homologías. Esta iniciativa apunta a favorecer la transparencia, el acceso y la reutilización de los resultados por parte de la comunidad científica.

En resumen, este trabajo demuestra que la comparación estructural de proteínas, apoyada en predicciones confiables y alineamientos tridimensionales validados, permite rescatar relaciones evolutivas profundas entre genes altamente divergentes. La herramienta ASC es una plataforma autoportante, flexible y fácil de usar que puede aplicarse a genomas modelo y no modelo y servir como complemento para mejorar la cobertura funcional de proteomas poco caracterizados. Al mismo tiempo, nuestros resultados resaltan la necesidad de seguir integrando evidencia experimental, modelos de lenguaje y bibliografía curada para continuar validando e interpretando las funciones de los genes identificados. La anotación por homología estructural, lejos de ser una alternativa, se presenta como una vía con proyección para avanzar en la caracterización de la biodiversidad genética.

Bibliografía

1. Hug LA. The ever-changing tree of life. Nat Microbiol 2024 98. 2024;9: 1906–1908. doi:10.1038/S41564-024-01768-W

2. Jewari CA, Baldauf SL. An excavate root for the eukaryote tree of life. Sci Adv. 2023;9. doi:10.1126/SCIADV.ADE4973

3. Williamson K, Eme L, Baños H, McCarthy CGP, Susko E, Kamikawa R, et al. A robustly rooted tree of eukaryotes reveals their excavate ancestry. Nature. 2025. doi:10.1038/s41586-025-08709-5

4. Kostygov AY, Albanaz ATS, Butenko A, Gerasimov ES, Lukeš J, Yurchenko V. Phylogenetic framework to explore trait evolution in Trypanosomatidae. Trends Parasitol. 2024;40: 96–99. doi:10.1016/J.PT.2023.11.009

5. Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, et al. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J Eukaryot Microbiol. 2019;66: 4–119. doi:10.1111/JEU.12691

6. Keeling PJ, Burki F. Progress towards the Tree of Eukaryotes. Curr Biol. 2019;29: R808–R817. doi:10.1016/J.CUB.2019.07.031

7. Kostygov AY, Karnkowska A, Votýpka J, Tashyreva D, Maciszewski K, Yurchenko V, et al. Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. Open Biol. 2021;11: 200407. doi:10.1098/RSOB.200407

8. Jackson AP. Genome evolution in trypanosomatid parasites. Parasitology. Cambridge University Press; 2015. pp. S40–S56. doi:10.1017/S0031182014000894

9. Berriman M, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, et al. The Genome of the African Trypanosome Trypanosoma brucei. 2012;416. doi:10.1126/science.1112642

10. El-sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Westenberger SJ, et al. The Genome Sequence of Trypanosoma cruzi , Etiologic Agent of Chagas

Disease. 2005;4975: 409–415.

11. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The Genome of the Kinetoplastid Parasite, Leishmania major. Science. 2005;309: 436. doi:10.1126/SCIENCE.1112680

12. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, et al. Comparative genomics of trypanosomatid parasitic protozoa. Science (80-). 2005;309: 404–409. doi:10.1126/science.1112181

13. Choi J, El-Sayed NM. Functional genomics of trypanosomatids. Parasite Immunol. 2012;34: 72–79. doi:10.1111/J.1365-3024.2011.01347.X

14. Kostygov AY, Skýpalová K, Kraeva N, Kalita E, McLeod C, Yurchenko V, et al. Comprehensive analysis of the Kinetoplastea intron landscape reveals a novel intron-containing gene and the first exclusively trans-splicing eukaryote. BMC Biol. 2024;22: 281. doi:10.1186/s12915-024-02080-z

15. Borst P, Sabatini R. Base J: Discovery, biosynthesis, and possible functions. Annual Review of Microbiology. Annual Reviews; 2008. pp. 235–251. doi:10.1146/annurev.micro.62.081307.162750

16. Campbell DA, Thomas S, Sturm NR. Transcription in kinetoplastid protozoa: why be normal? Microbes Infect. 2003;5: 1231–1240. doi:10.1016/J.MICINF.2003.09.005

17. Benne R, Van Den Burg J, Brakenhoff JPJ, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell. 1986;46: 819–826. doi:10.1016/0092-8674(86)90063-2

18. Aphasizhev R, Aphasizheva I, Nelson RE, Gao G, Simpson AM, Kang X, et al. Isolation of a U-insertion/deletion editing complex from Leishmania tarentolae mitochondria. EMBO J. 2003;22: 913–924. doi:10.1093/emboj/cdg083

19. Hannaert V, Bringaud F, Opperdoes FR, Michels PAM. Evolution of energy metabolism and its compartmentation in Kinetoplastida. Kinetoplastid Biol Dis

2003 21. 2003;2: 1–30. doi:10.1186/1475-9292-2-11

20. de Souza W, Attias M, Rodrigues JCF. Particularities of mitochondrial structure in parasitic protists (Apicomplexa and Kinetoplastida). Int J Biochem Cell Biol. 2009;41: 2069–2080. doi:10.1016/J.BIOCEL.2009.04.007

21. Butenko A, Hammond M, Field MC, Ginger ML, Yurchenko V, Lukeš J. Reductionist Pathways for Parasitism in Euglenozoans? Expanded Datasets Provide New Insights. Trends Parasitol. 2021;37: 100–116. doi:10.1016/J.PT.2020.10.001

22. Souto-Padron T, De Souza W, Heuser JE. Quick-freeze, deep-etch rotary replication of Trypanosoma cruzi and Herpetomonas megaseliae. J Cell Sci. 1984;VOL. 69: 167–178. doi:10.1242/JCS.69.1.167,

23. Hoffmann A, Jakob M, Ochsenreiter T. A novel component of the mitochondrial genome segregation machinery in trypanosomes. Microb Cell. 2016;3: 352–354. doi:10.15698/MIC2016.08.519,

24. Berna L, Greif G, Pita S, Faral-Tello P, Dıaz-Viraque F, De Cassia Moreira De Souza R, et al. Maxicircle architecture and evolutionary insights into trypanosoma cruzi complex. PLoS Negl Trop Dis. 2021;15. doi:10.1371/journal.pntd.0009719

25. Jackson AP. Genome evolution in trypanosomatid parasites. Parasitology. 2015;142: S40–S56. doi:10.1017/S0031182014000894

26. Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, et al. Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. Curr Biol. 2016;26: 161–172. doi:10.1016/j.cub.2015.11.055

27. Dávila López M, Marŕne Guerra JJ, Samuelsson T. Analysis of Gene Order Conservation in Eukaryotes Identifies Transcriptionally and Functionally Linked Genes. PLoS One. 2010;5: e10654. doi:10.1371/JOURNAL.PONE.0010654

28. Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, et al. Gene synteny and evolution of genome architecture in trypanosomatids. Mol Biochem Parasitol. 2004;134: 183–191. doi:10.1016/J.MOLBIOPARA.2003.11.012

29. Clayton C. Regulation of gene expression in trypanosomatids: Living with polycistronic transcription. Open Biol. 2019;9. doi:10.1098/rsob.190072

30. Díaz-Viraqué F, Chiribao ML, Libisch MG, Robello C. Genome-wide chromatin interaction map for Trypanosoma cruzi. Nat Microbiol. 2023;8: 2103–2114. doi:10.1038/s41564-023-01483-y

31. Respuela P, Ferella M, Rada-Iglesias A, Åslund L. Histone Acetylation and Methylation at Sites Initiating Divergent Polycistronic Transcription in Trypanosoma cruzi. J Biol Chem. 2008;283: 15884. doi:10.1074/JBC.M802081200

32. Smircich P, El-Sayed NM, Garat B. Intrinsic DNA curvature in trypanosomes. BMC Res Notes. 2017;10: 1–7. doi:10.1186/s13104-017-2908-y

33. Menezes AP, Murillo AM, de Castro CG, Bellini NK, Tosi LRO, Thiemann OH, et al. Navigating the boundaries between metabolism and epigenetics in trypanosomes. Trends Parasitol. 2023;39: 682–695. doi:10.1016/J.PT.2023.05.010

34. Martínez-Calvillo S, Nguyen D, Stuart K, Myler PJ. Transcription Initiation and Termination on Leishmania major Chromosome 3. Eukaryot Cell. 2004;3: 506. doi:10.1128/EC.3.2.506-517.2004

35. Gilinger G, Bellofatto V. Trypanosome spliced leader RNA genes contain the first identified RNA polymerase II gene promoter in these organisms. Nucleic Acids Res. 2001;29: 1556. doi:10.1093/NAR/29.7.1556

36. Sather S, Agabian N. A 5’ spliced leader is added in trans to both α- and β-tubulin transcripts in Trypanosoma brucei. Proc Natl Acad Sci U S A. 1985;82: 5695–5699. doi:10.1073/PNAS.82.17.5695,

37. Agabian N. Trans splicing of nuclear pre-mRNAs. Cell. 1990;61: 1157–1160. doi:10.1016/0092-8674(90)90674-4

38. Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, et al. A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. RNA. 2000;6: 163. doi:10.1017/S135583820099229X

39. Perry KL, Watkins KP, Agabian N. Trypanosome mRNAs have unusual “cap 4” structures acquired by addition of a spliced leader. Proc Natl Acad Sci. 1987;84: 8190–8194. doi:10.1073/PNAS.84.23.8190

40. Bangs JD, Crain PF, Hashizume T, McCloskey JA, Boothroyd JC. Mass spectrometry of mRNA cap 4 from trypanosomatids reveals two novel nucleosides. J Biol Chem. 1992;267: 9805–9815. doi:10.1016/s0021-9258(19)50165-x

41. LeBowitz JH, Smith HQ, Rusche L, Beverley SM. Coupling of poly(A) site selection and trans-splicing in Leishmania. Genes Dev. 1993;7: 996–1007. doi:10.1101/GAD.7.6.996,

42. Campos PC, Bartholomeu DC, DaRocha WD, Cerqueira GC, Teixeira SMR. Sequences involved in mRNA processing in Trypanosoma cruzi. Int J Parasitol. 2008;38: 1383–1389. doi:10.1016/j.ijpara.2008.07.001

43. Hummel HS, Gillespie RD, Swindle J. Mutational analysis of 3’ splice site selection during trans-splicing. J Biol Chem. 2000;275: 35522–35531. doi:10.1074/jbc.M002424200

44. Clayton C, Michaeli S. 3’ processing in protists. Wiley Interdiscip Rev RNA. 2011;2: 247–255. doi:10.1002/wrna.49

45. Kramer S. Nuclear mRNA maturation and mRNA export control: From trypanosomes to opisthokonts. Parasitology. Cambridge University Press; 2021. pp. 1196–1218. doi:10.1017/S0031182021000068

46. De Gaudenzi JG, Noé G, Campo VA, Frasch AC, Cassola A. Gene expression regulation in trypanosomatids. Essays Biochem. 2011;51: 31–46. doi:10.1042/BSE0510031

47. Rennie ML, Oliver MR. Emerging frontiers in protein structure prediction following the AlphaFold revolution. Journal of the Royal Society Interface. Royal Society Publishing; 2025. doi:10.1098/rsif.2024.0886

48. Ongus JR. Biotechnological Advances in Methods for Functional Analysis of Genes. J Nat Sci Res www.iiste.org ISSN. Online; 2017. Available: www.iiste.org

49. Pearson WR. Protein Function Prediction: Problems and Pitfalls. Curr Protoc Bioinforma. 2015;51: 4.12.1-4.12.8. doi:10.1002/0471250953.bi0412s51

50. Rost B. Twilight zone of protein sequence alignments. Protein Eng Des Sel. 1999;12: 85–94. doi:10.1093/PROTEIN/12.2.85

51. Shehu A, Barbará D, Molloy K. A survey of computational methods for protein function prediction. Big Data Analytics in Genomics. Springer International Publishing; 2016. pp. 225–298. doi:10.1007/978-3-319-41279-5_7

52. Chen J, Guo M, Wang X, Liu B. A comprehensive review and comparison of different computational methods for protein remote homology detection. Brief Bioinform. 2018;19: 231–244. doi:10.1093/BIB/BBW108

53. Gane A, Bileschi ML, Dohan D, Speretta E, Héliou A, Meng-Papaxanthos L, et al. ProtNLM: Model-based Natural Language Protein Annotation. 2022 Oct.

54. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nat 2021 5967873. 2021;596: 583–589. doi:10.1038/s41586-021-03819-2

55. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42: 243–246. doi:10.1038/s41587-023-01773-0

56. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35: 1026–1028. doi:10.1038/nbt.3988

57. Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. FATCAT 2.0: Towards a better understanding of the structural diversity of proteins. Nucleic Acids Res. 2020;48: W60–W64. doi:10.1093/NAR/GKAA443

58. Köster J, Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, et al. Sustainable data analysis with Snakemake. F1000Research. 2021;10: 33. doi:10.12688/f1000research.29032.2

59. Shanmugasundram A, Starns D, Böhme U, Amos B, Wilkinson PA, Harb OS, et al.

TriTrypDB: An integrated functional genomics resource for kinetoplastida. PLoS Negl Trop Dis. 2023;17: e0011058. doi:10.1371/JOURNAL.PNTD.0011058

60. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33: 2302–2309. doi:10.1093/NAR/GKI524

61. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25: 1605–1612. doi:10.1002/JCC.20084

62. Billington K, Halliday C, Madden R, Dyer P, Barker AR, Moreira-Leite FF, et al. Genome-wide subcellular protein map for the flagellate parasite Trypanosoma brucei. Nat Microbiol. 2023;8: 533–547. doi:10.1038/s41564-022-01295-6

63. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

64. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38: 4647–4654. doi:10.1093/MOLBEV/MSAB199

65. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47: D309–D314. doi:10.1093/NAR/GKY1085

66. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30: 1236–1240. doi:10.1093/BIOINFORMATICS/BTU031

67. Hernández-Salmerón JE, Moreno-Hagelsieb G. Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. BMC Genomics. 2020;21: 1–9. doi:10.1186/s12864-020-07132-6

68. Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, et al.

Clustering-predicted structures at the scale of the known protein universe. Nature. 2023. doi:10.1038/s41586-023-06510-w

69. Xia Y, Zhao K, Liu D, Zhou X, Zhang G. Multi-domain and complex protein structure prediction using inter-domain interactions from deep learning. Commun Biol. 2023;6. doi:10.1038/s42003-023-05610-7

70. Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, et al. Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. Curr Protoc Bioinforma. 2011;35: 6.12.1-6.12.19. doi:10.1002/0471250953.BI0612S35

71. Salzberg SL. Next-generation genome annotation: We still struggle to get it right. Genome Biology. BioMed Central Ltd.; 2019. doi:10.1186/s13059-019-1715-2

72. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010;26: 889. doi:10.1093/BIOINFORMATICS/BTQ066

73. Monzon V, Paysan-Lafosse T, Wood V, Bateman A. Reciprocal best structure hits: Using AlphaFold models to discover distant homologues. Gromiha M, editor. Bioinforma Adv. 2022;2. doi:10.1093/bioadv/vbac072

74. Ruperti F, Papadopoulos N, Musser JM, Mirdita M, Steinegger M, Arendt D. Cross-phyla protein annotation by structural prediction and alignment. Genome Biol. 2023;24: 113. doi:10.1186/s13059-023-02942-9

75. Zhao C, Wang Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. Sci Rep. 2018;8. doi:10.1038/s41598-018-33219-y

76. Meinhart A, Silberzahn T, Cramer P. The mRNA transcription/processing factor Ssu72 is a potential tyrosine phosphatase. J Biol Chem. 2003;278: 15917–15921. doi:10.1074/jbc.M301643200

77. Krishnamurthy S, He X, Reyes-Reyes M, Moore C, Hampsey M. Ssu72 is an RNA polymerase II CTD phosphatase. Mol Cell. 2004;14: 387–394. doi:10.1016/S1097-2765(04)00235-7

78. Dichtl B, Blank D, Ohnacker M, Friedlein A, Roeder D, Langen H, et al. A role for SSU72 in balancing RNA polymerase II transcription elongation and termination. Mol Cell. 2002;10: 1139–1150. doi:10.1016/S1097-2765(02)00707-4

79. Koch H, Raabe M, Urlaub H, Bindereif A, Preußer C. The polyadenylation complex of Trypanosoma brucei: Characterization of the functional poly(A) polymerase. RNA Biol. 2016;13: 221. doi:10.1080/15476286.2015.1130208

80. Wu M, Tzagoloff A. Identification and Characterization of a New Gene (CBP3) Required for the Expression of Yeast Coenzyme QH2-Cytochrome c Reductase. J Biol Chem. 1989;264: 11122–11130. doi:10.1016/S0021-9258(18)60438-7

81. Gruschke S, Kehrein K, Römpler K, Gröne K, Israel L, Imhof A, et al. Cbp3–Cbp6 interacts with the yeast mitochondrial ribosomal tunnel exit and promotes cytochrome b synthesis and assembly. J Cell Biol. 2011;193: 1101–1114. doi:10.1083/JCB.201103132

82. Dieckmann CL, Pape LK, Tzagoloff A. Identification and cloning of a yeast nuclear gene (CBP1) involved in expression of mitochondrial cytochrome b. Proc Natl Acad Sci. 1982;79: 1805–1809. doi:10.1073/PNAS.79.6.1805

83. Feaver WJ, Huang W, Friedberg EC. The TFB4 subunit of yeast TFIIH is required for both nucleotide excision repair and RNA polymerase II transcription. J Biol Chem. 1999;274: 29564–29567. doi:10.1074/jbc.274.41.29564

84. Schmitt DR, Kuper J, Elias A, Kisker C. The Structure of the TFIIH p34 Subunit Reveals a Von Willebrand Factor A Like Fold. PLoS One. 2014;9: e102389. doi:10.1371/JOURNAL.PONE.0102389

85. Warfield L, Luo J, Ranish J, Hahn S. Function of Conserved Topological Regions within the Saccharomyces cerevisiae Basal Transcription Factor TFIIH. Mol Cell Biol. 2016;36: 2464–2475. doi:10.1128/MCB.00182-16

86. Lee JH, Jung HS, Günzl A. Transcriptionally active TFIIH of the early-diverged eukaryote Trypanosoma brucei harbors two novel core subunits but not a cyclin-activating kinase complex. Nucleic Acids Res. 2009;37: 3811–3820. doi:10.1093/NAR/GKP236

87. Louder RK, He Y, Ramón López-Blanco J, Fang J, Chacón P, Nogales E. Structure of promoter-bound TFIID and insight into human PIC assembly. Nature. 2016 [cited 26 May 2025]. doi:10.1038/nature17394

88. Anandapadamanaban M, Andresen C, Helander S, Ohyama Y, Siponen MI, Lundström P, et al. High-resolution structure of TBP with TAF1 reveals anchoring patterns in transcriptional regulation. Nat Struct Mol Biol. 2013;20: 1008–1014. doi:10.1038/NSMB.2611

89. Nogales E, Louder RK, He Y. Structural Insights into the Eukaryotic Transcription Initiation Machinery. Annual Review of Biophysics. Annual Reviews Inc.; 2017. pp. 59–83. doi:10.1146/annurev-biophys-070816-033751

90. Hansen SK, Tjian R. TAFs and TFIIA mediate differential utilization of the tandem Adh promoters. Cell. 1995;82: 565–575. doi:10.1016/0092-8674(95)90029-2

91. Papai G, Tripathi MK, Ruhlmann C, Werten S, Crucifix C, Weil PA, et al. Mapping the Initiator Binding Taf2 Subunit in the Structure of Hydrated Yeast TFIID. Structure. 2009;17: 363–373. doi:10.1016/j.str.2009.01.006

92. Grozdanov PN, Masoumzadeh E, Latham MP, MacDonald CC. The structural basis of CstF-77 modulation of cleavage and polyadenylation through stimulation of CstF-64 activity. Nucleic Acids Res. 2018;46: 12022–12039. doi:10.1093/NAR/GKY862

93. Mayr C, Bartel DP. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. Cell. 2009;138: 673–684. doi:10.1016/j.cell.2009.06.016

94. Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol 2016 181. 2016;18: 18–30. doi:10.1038/nrm.2016.116

95. Takagaki Y, Manley JL, MacDonald CC, Wilusz J, Shenk T. A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. Genes Dev. 1990;4: 2112–2120. doi:10.1101/GAD.4.12A.2112,

96. Benoit B, Juge F, Iral F, Audibert A, Simonelig M. Chimeric human CstF-

77/Drosophila suppressor of forked proteins rescue suppressor of forked mutant lethality and mRNA 3' end processing in Drosophila. Proc Natl Acad Sci U S A. 2002;99: 10593–10598. doi:10.1073/pnas.162191899

97. Shan S ou. Guiding tail-anchored membrane proteins to the endoplasmic reticulum in a chaperone cascade. J Biol Chem. 2019;294: 16577–16586. doi:10.1074/jbc.REV119.006197

98. Wang F, Brown EC, Mak G, Zhuang J, Denic V. A chaperone cascade sorts proteins for posttranslational membrane insertion into the endoplasmic reticulum. Mol Cell. 2010;40: 159–171. doi:10.1016/j.molcel.2010.08.038

99. Tambe MA, Ng BG, Shimada S, Wolfe LA, Adams DR, Gahl WA, et al. Mutations in GET4 disrupt the transmembrane domain recognition complex pathway. J Inherit Metab Dis. 2020;43: 1037–1045. doi:10.1002/jimd.12249

100. Asseck LY, Mehlhorn DG, Monroy JR, Ricardi MM, Breuninger H, Wallmeroth N, et al. Endoplasmic reticulum membrane receptors of the GET pathway are conserved throughout eukaryotes. Proc Natl Acad Sci U S A. 2021;118: e2017636118. doi:10.1073/pnas.2017636118

101. Maitland MER, Onea G, Chiasson CA, Wang X, Ma J, Moor SE, et al. The mammalian CTLH complex is an E3 ubiquitin ligase that targets its subunit muskelin for degradation. Sci Rep. 2019;9: 1–14. doi:10.1038/s41598-019-46279-5

102. Komander D, Rape M. The ubiquitin code. Annu Rev Biochem. 2012;81: 203–229. doi:10.1146/annurev-biochem-060310-170328

103. Lampert F, Stafa D, Goga A, Soste MV, Gilberto S, Olieric N, et al. The multi-subunit GID/CTLH e3 ubiquitin ligase promotes cell proliferation and targets the transcription factor Hbp1 for degradation. Elife. 2018;7. doi:10.7554/ELIFE.35528

104. Santt O, Pfirrmann T, Braun B, Juretschke J, Kimmig P, Scheel H, et al. The yeast GID complex, a novel ubiquitin ligase (E3) involved in the regulation of carbohydrate metabolism. Mol Biol Cell. 2008;19: 3323–3333. doi:10.1091/mbc.E08-03-0328

105. Jin X, Guan Z, Hu N, He C, Yin P, Gong Z, et al. Structural insight into how WDR4 promotes the tRNA N7-methylguanosine methyltransferase activity of METTL1. Cell Discov. 2023;9: 65–65. doi:10.1038/S41421-023-00562-Y

106. Li J, Wang L, Hahn Q, Nowak RP, Viennet T, Orellana EA, et al. Structural basis of regulated m7G tRNA modification by METTL1–WDR4. Nature. 2023;613: 391–397. doi:10.1038/s41586-022-05566-4

107. Frye M, Harada BT, Behm M, He C. RNA modifications modulate gene expression during development HHS Public Access. Science (80-). 2018;361: 1346–1349. doi:10.1126/science.aau1646

108. Maio N, Rouault TA. Iron –sulfur cluster biogenesis in mammalian cells: New insights into the molecular mechanisms of cluster delivery. Biochim Biophys Acta - Mol Cell Res. 2015;1853: 1493–1512. doi:10.1016/J.BBAMCR.2014.09.009

109. Cai K, Frederick RO, Kim JH, Reinen NM, Tonelli M, Markley JL. Human mitochondrial chaperone (mtHSP70) and cysteine desulfurase (NFS1) bind preferentially to the disordered conformation, whereas co-chaperone (HSC20) binds to the structured conformation of the iron-sulfur cluster scaffold protein (ISCU). J Biol Chem. 2013;288: 28755–28770. doi:10.1074/jbc.M113.482042

110. Lill R, Mühlenhoff U. Iron-sulfur protein biogenesis in eukaryotes: Components and mechanisms. Annual Review of Cell and Developmental Biology. Annu Rev Cell Dev Biol; 2006. pp. 457–486. doi:10.1146/annurev.cellbio.22.010305.104538

111. Svedberg D, Winiger RR, Berg A, Sharma H, Tellgren-Roth C, Debrunner-Vossbrinck BA, et al. Functional annotation of a divergent genome using sequence and structure-based similarity. BMC Genomics. 2024;25: 1–18. doi:10.1186/S12864-023-09924-Y

112. Borujeni PM, Salavati R. Functional domain annotation by structural similarity. NAR Genomics Bioinforma. 2024;6: 1–11. doi:10.1093/nargab/lqae005

113. Ambrosino L, Chiusano ML. Transcriptologs: A transcriptome-based approach to predict orthology relationships. Bioinform Biol Insights. 2017;11. doi:10.1177/1177932217690136

Material suplementario