Facultad de Ciencias, Universidad de la República

*Sección Genómica Funcional*

Instituto de Investigaciones Biológicas Clemente Estable

*Laboratorio de Bioinformática, Departamento de Genómica*

PEDECIBA Bioinformática

Tesis de Maestría en Bioinformática

# ANÁLISIS FUNCIONALES DE PERFILES DE CO-EXPRESIÓN GÉNICA EN TRYPANOSOMA CRUZI

**Autor**: Lic. Lucas Inchausti

**Orientador**: Dr. Pablo Smircich

**Co-orientador**: Dr. Ing. Álvaro Martín

Montevideo, Uruguay

Julio 2023

*A mis dos madres, a mis dos padres.*

# Agradecimientos

En primer lugar, agradecer a Pablo, quien despertó en mí el interés por la bioinformática mientras cursaba biología molecular en la carrera de grado, quien ha sido para mí un mentor a nivel profesional y humano, y por la confianza que deposita en mí día a día hace ya 5 años, motivándome a seguir transitando el arduo camino de la investigación. También agradecer a Álvaro, con quien nos hemos nutrido mutuamente de saberes de biología e informática, que ha sido una parte fundamental en este trabajo brindando perspectivas distintas a las que acostumbramos desde las ciencias biológicas, y ha realizado una lectura minuciosa de este trabajo cuyas correcciones mejoraron sin lugar a duda la calidad gramatical, estructura y adecuación al público objetivo de este manuscrito.

Agradecer con mucho cariño a cada uno de los y las integrantes de la Sección Genómica Funcional de Facultad de Ciencias, el Laboratorio de Bioinformática y el Departamento de Genómica del IIBCE, los nuevos, los viejos y los más viejos, con quienes hemos compartido hermosas vivencias en los últimos años, a nivel académico y personal, y han sido un apoyo importante para el desarrollo de este trabajo.

Al tribunal, mi más profundo agradecimiento por tomarse la labor de evaluar este trabajo en tiempo récord, con cuyos comentarios y correcciones recién culminaré en darle forma a esta historia que comenzó hace dos años y medio.

Por supuesto agradecer a mi familia: mis madres, mis padres, hermanos, primos y primas, tíos y tías, sobrino y sobrinas, por su apoyo, amor, contención, interés, aunque aún no haya logrado explicarles a qué me dedico. A mi otra familia, mis amigos de toda la vida, mis amigos y amigas de Facultad, particularmente a Sofi, gran amiga y compañera con quien transitamos juntos esta carrera. Su presencia constante y apoyo emocional fueron fundamentales para superar los desafíos que surgieron en este recorrido. Agradecer también a Fede, a Cami y a Martu por su invaluable apoyo y aliento, y por abrirme sus puertas para escribir grandes partes de este trabajo.

Por último, agradecer a la Comisión Académica de Posgrado por el reconocimiento y el apoyo a través de dos becas de posgrado, y a PEDECIBA por la oportunidad de desarrollar esta carrera y por los apoyos financieros recibidos para llevar a cabo este trabajo.

# Índice

# Glosario

ADN – ácido desoxiribonucléico

ARN – ácido ribonucléico

ARNm – ARN mensajero

ARNt – ARN de transferencia

CAP 5’ – caperuza 5’

CDS – *Coding sequences*

DAG – *Directed Acyclic Graph*

ENA – *European Nucleotide Archive*

GO – *Gene Ontology*

HMM – *Hidden Markov Model*

Hubgene – gen central

Module Eigengene – representación resumida de la expresión de un grupo de genes

NCBI – *National Center for Biotechnology Information*

PCA – *Principal Component Analysis*

RBP – *RNA binding protein*

TOM – *Topological Overlap Matrix*

t-SNE – *t-distributed stochastic neighbor embedding*

UTR – *untranslated region*

WGCNA – *Weighted Gene Co-expression Network Analysis*

## Resumen

*Trypanosoma cruzi* es un parásito protozoario causante de la tripanosomiasis americana también denominada enfermedad de Chagas, una enfermedad tropical desatendida que afecta a millones de personas y que prolifera en entornos empobrecidos. *T. cruzi* se caracteriza por un ciclo de vida complejo que involucran distintas etapas de diferenciación, tanto en su hospedero triatomino que funciona como vector como en su hospedero mamífero, cada una con características particulares**.** Los genes de *T. cruzi* se expresan de forma policistrónica, siendo los mecanismos post-transcripcionales los principales mecanismos de regulación de la expresión génica.

Los análisis de co-expresión génica son una valiosa herramienta para estudiar cambios en el nivel de expresión de grupos de genes que interactúan funcionalmente entre sí. En este estudio se buscó construir una red de co-expresión génica utilizando datos transcriptómicos de 12 puntos del ciclo de vida de *T. cruzi* abarcando todos los estadios de este organismo, con el fin de caracterizar y analizar funcionalmente grupos de genes co-expresados. A su vez, se buscó identificar posibles mecanismos que expliquen esta regulación conjunta, tales como la presencia de motivos en las secuencias 3'UTRs y el uso diferencial de codones de los genes co-expresados.

Se identificaron 13 módulos de genes co-expresados mediante la construcción de una *weighted gene co-expression network*. De ellos, se obtuvieron 10 módulos que estaban enriquecidos funcionalmente en roles asociados a metabolismo, patogénesis, regulación de la replicación del ADN, regulación del citoesqueleto y movimiento celular, y más. Por otro lado, se logró correlacionar ciertos módulos enriquecidos funcionalmente con determinados estadios del parásito y las características biológicas que los definen.

Por otro lado, se identificó un uso diferencial de codones entre los módulos de genes co-expresados, y la presencia de motivos a nivel de secuencia en las regiones 3'UTR que podrían explicar esta expresión conjunta de genes.

Por último, se identificaron los genes más conectados de cada módulo (*hubgenes)*, que cumplen roles clave en los procesos subyacentes a los módulos. Dado que muchos de estos *hubgenes* estaban anotados como proteínas hipotéticas se realizó su inferencia funcional mediante alineamiento estructural y comparación de perfiles HMM-HMM.

Se espera que los resultados obtenidos en este estudio realicen un aporte significativo a la comprensión de la biología molecular de este parasito de alta relevancia en nuestra región y provea información y priorice el estudio de genes clave en el desarrollo de este parásito, cuyas proteínas codificantes podrían ser posibles blancos moleculares de agentes tripanocidas.

**Palabras clave**: tripanosomiasis americana, enfermedad de Chagas, *Trypanosoma cruzi*, regulación génica, co-expresión génica

# 1 Introducción

## 1.1 Kinetoplástidos

Los kinetoplástidos son un grupo de protistas unicelulares flagelados pertenecientes al filo Euglenozoa. Se caracterizan por poseer una única gran mitocondria conocida como “kinetoplasto”, que le da nombre a este grupo. Aunque los distintos organismos del grupo presentan muchas similitudes, como la presencia de un único flagelo que se origina cerca del kinetoplasto, una organización genómica y estructuras celulares similares, y cambios morfológicos drásticos durante su ciclo de vida, estos organismos producen diversas enfermedades en humanos y son transmitidos por distintos vectores (Stuart et al., 2008).

Mediante estudios de ARNr 18S, la clase Kinetoplastea ha sido dividida en dos subclases: Prokinetoplastina y Metakinetoplastina, presentando este último cuatro órdenes: Eubodonida, Parabodonida, Neobodonida y Trypanosomatida, siendo este último el más estudiado (D’avila-Levy et al., 2015; Moreira et al., 2004) (**figura 1**).

Particularmente, todos los miembros del orden Trypanosomatida pertenecen a una única familia Trypanosomatidae. Dentro de esta familia se encuentran los denominados “TriTryps”, agentes causantes de la enfermedad del sueño africana (*Trypanosoma brucei*), la enfermedad de Chagas (*Trypanosoma cruzi*) y diversas formas de leishmaniasis (*Leishmania spp.*). Los TriTryps se caracterizan por ser organismos monoflagelados y diexénicos, con ciclos de vida y estrategias de supervivencia variados, y que sufren diversos cambios morfológicos al ser transmitidos a hospederos vertebrados mediante un vector invertebrado, principalmente insectos (Lukeš et al., 2018).

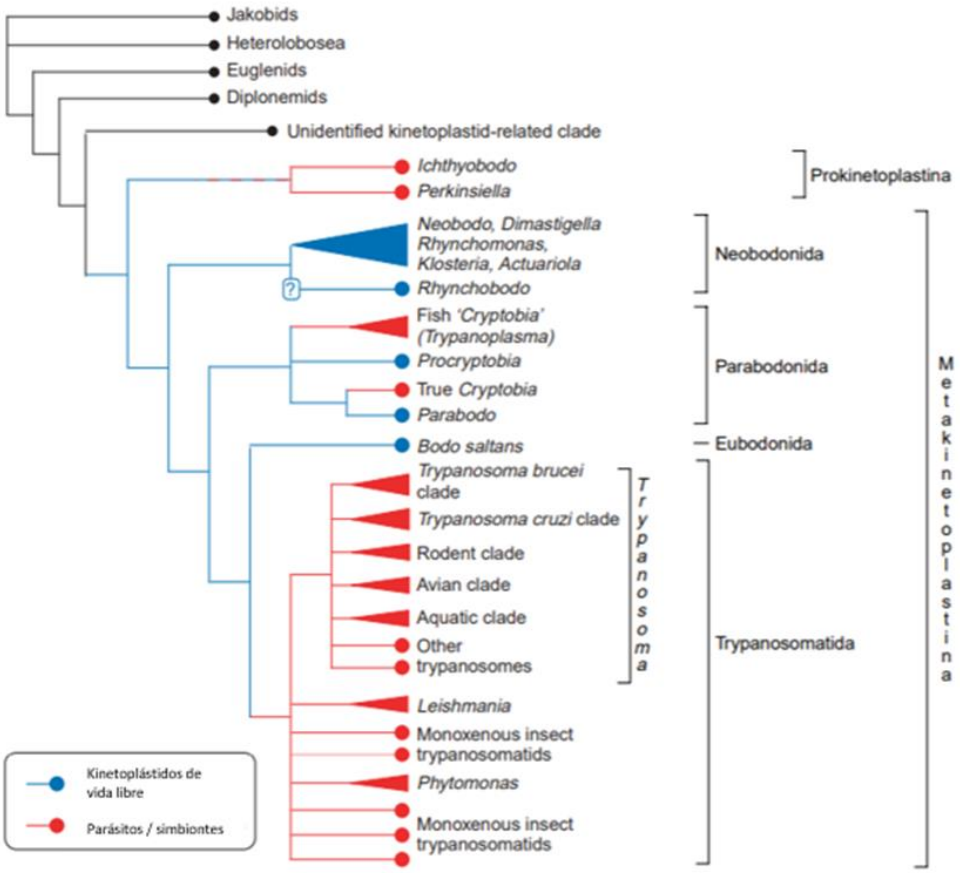



**Figura 1.** Árbol filogenético de Kinetoplástidos. Extraído y modificado de (Simpson et al., 2006)

## 1.2 *Trypanosoma cruzi*

### 1.2.1 Características clínicas y epidemiológicas de la enfermedad de Chagas

La tripanosomiasis americana, o enfermedad de Chagas, es una enfermedad causada por el parásito *Trypanosoma cruzi* (*T. cruzi*), que es transmitida principalmente por contacto con las heces o la orina infectadas de triatominos que se alimentan de sangre, particularmente de las especies *Triatoma infestans, Rhodnius prolixus* y *Triatoma dimidiata*. En general, pican al mamífero para alimentarse de su sangre y defecan/orinan cerca de la picadura, donde los parásitos ingresan al organismo cuando el individuo se frota instintivamente y empuja las heces o la orina hacia la picadura, los ojos, la boca o alguna lesión cutánea abierta. A su vez, existe la transmisión congénita de *T. cruzi*, que ocurre aproximadamente en el 5% de los niños nacidos de madres con infección crónica en áreas endémicas, con variaciones según la región (Carlier et al., 2015).

Inicialmente, la enfermedad de Chagas estaba confinada a las zonas rurales de América Latina. Debido a la mayor movilidad de la población en los últimos decenios, la mayoría de las personas infectadas ha pasado a vivir en entornos urbanos. A su vez, a través de las olas migratorias de las últimas décadas, la enfermedad se ha ido detectando cada vez más en Estados Unidos y Canadá, en muchos países europeos y en algunos países africanos, principalmente del Mediterráneo Oriental y del Pacífico Occidental (Rassi et al., 2010).

La enfermedad de Chagas presenta dos fases distintivas: la fase inicial o aguda, que dura entre 4 y 8 semanas después de la infección, y la fase crónica. Durante la fase aguda circula una gran cantidad de parásitos en el torrente sanguíneo, pero no suelen haber síntomas claros ni específicos de la enfermedad. En menos del 50% de las personas infectadas por un triatomino, un signo inicial característico puede ser una lesión cutánea o la hinchazón amoratada de un párpado. Además, esas personas pueden presentar fiebre, dolor de cabeza, agrandamiento de ganglios linfáticos, palidez, dolores musculares, dificultad para respirar, hinchazón y dolor abdominal o torácico (Rassi et al., 2010).

Por otro lado, la fase crónica de la enfermedad dura toda la vida y puede causar la muerte del individuo infectado. En esta fase, los parásitos permanecen ocultos principalmente en células musculares cardíacas y del aparato digestivo. Con el paso de los años, la infección puede causar arritmias o insuficiencia cardíaca progresiva como consecuencia de la destrucción del músculo cardíaco (“OMS | Enfermedades Tropicales Desatendidas: Preguntas Más Frecuentes,” 2010).

Actualmente, la enfermedad de Chagas puede tratarse con dos fármacos antiparasitarios: benznidazol y nifurtimox. Ambos medicamentos son sumamente eficaces si son administrados durante la fase aguda de la infección, incluso en los casos de transmisión congénita (Rassi et al., 2010). En este último caso, teniendo en cuenta que el tratamiento etiológico del recién nacido es siempre eficaz si se realiza antes de un año, el diagnóstico de infección en mujeres embarazadas y sus recién nacidos es de suma importancia. Las evidencias están demostrando que la transmisión congénita podría prevenirse mediante el tratamiento de mujeres infectadas antes de que queden embarazadas (Carlier et al., 2015). Por otro lado, la eficacia de estos fármacos disminuye a medida que transcurre el tiempo de la infección, y las reacciones adversas son más frecuentes en edades avanzadas. Estos medicamentos son sumamente tóxicos, con severos efectos secundarios; en el caso de nifurtimox, la pérdida de peso, alteraciones

psicológicas, excitación, somnolencia, vómitos, diarrea, mientras que en el caso del benznidazol, manifestaciones a nivel cutáneo tales como hipersensibilidad y dermatitis con erupciones y edemas, fiebre y dolores musculares (Castro et al., 2006). Ambos medicamentos mostraron también tener efectos mutagénicos y tumorigénicos (Castro et al., 2006; Teixeira et al., 1994), por lo que es de suma importancia continuar con el estudio de nuevos blancos moleculares y agentes antichagásicos.

### 1.2.2 Generalidades, características estructurales y de ciclo de vida

#### 1.2.2.1 Generalidades

El taxón *T. cruzi* presenta una amplia variabilidad genética, compuesto por numerosas cepas con distintas características en su perfil antigénico, virulencia, tasa de crecimiento, patogenicidad, tropismo y sensibilidad a fármacos antichagásicos (Buscaglia & Di Noia, 2003). Esta diversidad genética inicialmente permitió agrupar a las poblaciones de *T. cruzi* en dos categorías: TcI y TcII; TcI está asociado con el ciclo de transmisión en ambientes selváticos e infección de marsupiales, mientras que TcII se divide en cinco grupos relacionados, denominados TcIIa a TcIIe, y está asociado con el ciclo de transmisión intradomiciliario y la infección en mamíferos placentarios. En la actualidad, hay un consenso internacional que reconoce la existencia de estos principales seis linajes genéticos distribuidos en Unidades de Tipificación Discretas (DTUs) (Zingales et al., 2009, 2012).

#### 1.2.2.2 *Características estructurales*

Los tripanosomátidos presentan una considerable distancia filogenética respecto al resto de eucariotas, siendo uno de los géneros más ancestrales que se han estudiado. Debido a esto y a su complejo estilo de vida que se desarrolla en ambientes diversos, estos parásitos han desarrollado a lo largo de su evolución características adaptativas excepcionales (D. F. Smith & Parsons, 1996).

Una de las características más destacables de estos organismos es la presencia de una única y muy desarrollada mitocondria, que abarca gran parte del volumen celular. El ADN mitocondrial representa hasta un 30% del ADN celular total, y conforma una estructura particular denominada kinetoplasto, cuya localización varía dependiendo de la etapa del ciclo de vida del parásito y está físicamente ligada al cuerpo basal, en la base del flagelo (De Souza, 1984).

El núcleo, a diferencia de la mitocondria, no presenta características morfológicas distintivas con respecto al resto de las células eucariotas típicas. Sin embargo, *T. cruzi* sí presenta características particulares en sus procesos nucleares, tales como en la replicación, transcripción y reparación del ADN, considerados ancestrales con respecto al resto de eucariotas. Durante los estadios replicativos (ver sección 1.2.2.3), *T. cruzi* se reproduce mediante fisión binaria, y presenta un núcleo esférico y un evidente nucleolo central. Epimastigotas y amastigotas presentan gran diferencia a nivel de tamaño del núcleo, mientras que su forma es similar. Se desconoce aún con exactitud a qué se debe esta diferencia. Durante el estadio no replicativo tripomastigota (ver sección 1.2.2.3) existe una considerable disminución de la actividad transcripcional y una dramática reducción del tamaño celular. En este estadio el núcleo presenta una forma alargada, carente de nucleolo y con alto contenido de heterocromatina (Schenkman & Pascoalino, 2011).

En cuanto a la superficie celular, está conformada por una bicapa lipídica y otros componentes del lado extracelular que conforman la glicocálix. Las moléculas que se encuentran en la superficie celular incluyen las familias de glicoproteínas de mucinas, transialidasas y Tc85, entre otras. Estas características de su superficie celular son las que permiten que el parásito interaccione con las células de sus hospederos, en conjunto con la membrana de la vacuola parasitófora al momento de la infección (de Souza, 2009).

Un organelo característico de los tripanosomátidos es el glicosoma, donde tienen lugar varias etapas de la vía de la glucólisis. La compartimentalización de la glucólisis en estos organelos parece ser esencial para la regulación de dicho proceso, que le permite al parásito sobrellevar períodos de anaerobiosis que tienen lugar en determinados estadios del ciclo de vida, a través de la obtención de ATP mediante otras vías metabólicas (Michels et al., 2006). A su vez, el contenido enzimático de los glicosomas puede variar rápidamente durante los fenómenos de diferenciación celular.

Por último, otro organelo particular presente en todos los tripanosomátidos es el acidocalcisoma. Este organelo está involucrado en funciones tales como el almacenamiento de calcio, potasio, hierro, zinc, magnesio, la homeostasis del pH celular y la osmorregulación junto a la vacuola contráctil (Docampo & Moreno, 2011; Miranda et al., 2000; Rohloff et al., 2004).

#### 1.2.2.3 *Ciclo de vida de Trypanosoma cruzi*

*T. cruzi* se caracteriza por tener un ciclo de vida complejo, que involucra distintas etapas de diferenciación celular, cada una con características particulares de capacidad infectiva y replicativa, tanto en su hospedero triatomino que funciona como vector, como en sus hospederos vertebrados. Los estadios amastigota y tripomastigota se desarrollan en el hospedero mamífero, mientras que los estadios epimastigota y tripomastigota metacíclico lo hacen en el vector triatomino.

El ciclo de vida inicia con la ingesta de tripomastigotas sanguíneos presentes en la sangre del hospedero mamífero infectado por parte del vector triatomino. Generalmente, la mayor parte de los tripomastigotas mueren en el estómago, mientras que la fracción minoritaria que logra sobrevivir se diferencia a la forma epimastigota, proliferativa y no infectiva, en el tracto digestivo medio. En principio miden entre 10 a 20 µm pero crecen a medida que migran a través del intestino, donde se replican intensamente mediante fisión binaria y continúan migrando hasta las regiones más posteriores del intestino y la ampolla rectal del insecto, donde se da una nueva diferenciación a tripomastigotas metacíclicos, no proliferativos e infectivos. Cuando el insecto vuelve a alimentarse de la sangre de un nuevo mamífero, defeca cerca de la herida y libera junto a las heces tripomastigotas metacíclicos, que penetran al torrente sanguíneo a través de la herida. Una vez allí, los tripomastigotas infectan las células hospederas tales como macrófagos, fibras cardíacas y músculo liso, y son internalizados por sus vacuolas endocíticas conocidas como vacuolas parasitóforas. A partir de la acidificación del medio y ruptura de las vacuolas, los tripomastigotas son liberados al citoplasma celular y se diferencian en amastigotas, proliferativos y no infectivos. Dentro de la célula, los amastigotas sufren varias replicaciones mediante fisión binaria, diferenciándose una vez más en la forma de tripomastigotas sanguíneos, produciendo la lisis celular y la liberación de los parásitos con la capacidad de infectar nuevas células o alcanzar el torrente sanguíneo del hospedero, donde podrá ser ingerido nuevamente por el vector triatomino, completando así su ciclo de vida (**figura 2**).

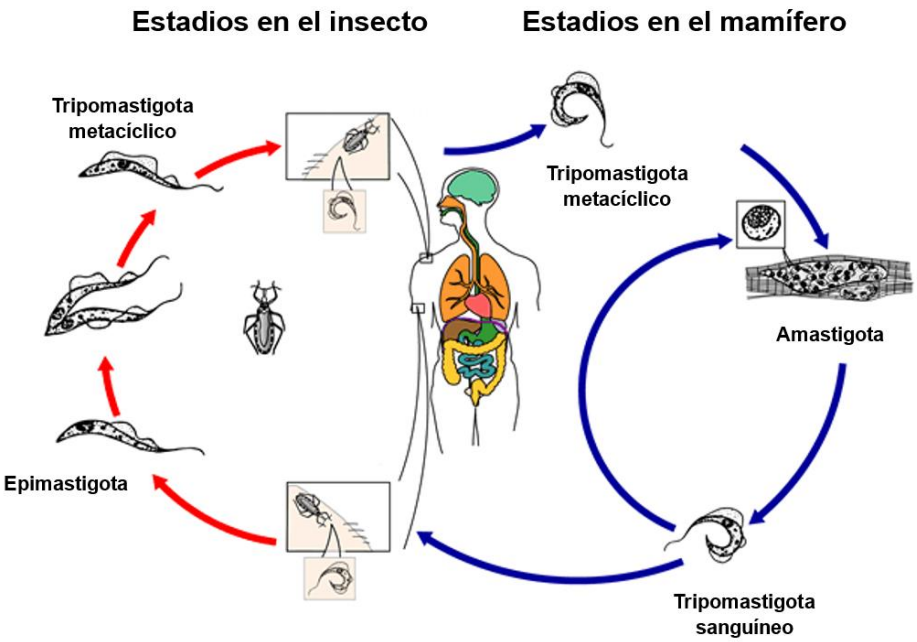



**Figura 2.** Ciclo de vida de *Trypanosoma cruzi*. Los epimastigotas se diferencian a tripomastigotas metacíclicos en el tracto digestivo del insecto vector hasta llegar a la ampolla rectal y son transmitidos al hospedero mamífero a través de sus heces. Una vez dentro del mamífero, invaden sus células y son rápidamente dirigidos hacia la vacuola parasitófora. Dentro de las vacuolas comienza su transformación hacia la forma amastigota. Esta vacuola se rompe y los parásitos pasan a localizarse en el citoplasma de las células. Los amastigotas intracelulares comienzan a replicarse intensamente hasta que, finalmente, se diferencian hacia tripomastigotas, quienes rompen la célula liberándose al torrente sanguíneo del hospedero, donde pueden infectar nuevas células o ser consumidos por un nuevo insecto que se alimente de la sangre del hospedero, diseminando la infección y completando así el ciclo de vida. Extraído y modificado del sitio web del Centers for Disease Control and Prevention (EEUU, www.cdc.gob)

### 1.2.3 Organización genómica

#### 1.2.3.1 *Genoma de Trypanosoma cruzi*

En el año 2005, un consorcio internacional publica los genomas de los TriTryps, e inmediatamente es publicado un *special issue* en la revista *Science* (Ash & Jasny, 2005) donde se exploran las secuencias de estos organismos, estudios que han permitido desde entonces caracterizar más en profundidad todas las particularidades biológicas que presentan. La cepa elegida para la secuenciación y ensamblado del genoma de *T. cruzi* (El-Sayed, Myler, Bartholomeu, et al., 2005) fue la CL Brener (TcVI), híbrida entre TcII y TcIII. Aparte de las dificultades provenientes de la cepa elegida para el secuenciado, el ensamblado del genoma (de aproximadamente 55 Mb para el genoma haploide) presentó otras dificultades, principalmente asociadas a la gran cantidad de secuencias repetidas.

El genoma haploide contiene unos 12.000 genes codificantes de proteínas (El-Sayed, Myler, Blandin, et al., 2005), de las cuales actualmente un 40% no tiene función conocida. Este genoma de referencia contiene unos 1994 genes de ARN no codificante y 3590 pseudogenes (El-Sayed, Myler, Blandin, et al., 2005). Más de la mitad del genoma está formado por secuencias repetidas, tales como retrotransposones, repetidos en tándem y subteloméricos y genes de familias multigénicas. Las principales familias multigénicas de este organismo corresponden a proteínas de tipo transialidasas, mucinas, metaloproteasas, DGF-1, proteínas RHS y las proteínas de superficie asociadas a mucinas (MASP). Algunas de las familias multigénicas codificantes para antígenos de superficie son compartidas en tripanosomátidos mientras que otras son exclusivas de alguno de ellos. A diferencia de *T.* brucei, *T. cruzi* no posee mecanismos de variación antigénica, sino que expresa varias proteínas de varias familias, que funciona como una estrategia fundamental para la evasión del sistema inmune del hospedero y demás procesos asociados con la infección (El-Sayed, Myler, Bartholomeu, et al., 2005; El-Sayed, Myler, Blandin, et al., 2005).

La secuenciación de los genomas de TriTryps demostró la existencia de una organización particular de los genes, que se encuentran agrupados en grandes regiones con igual orientación a las que se denominó *directional gene clusters* (DGC), y que son característicos de estos organismos, cuyos genes carecen de intrones, a excepción de muy pocos genes, como el gen de la poliA polimerasa y el gen que codifica para el ARNt de tirosina (Mair et al., 2000). Estos agrupamientos de genes recuerdan a los clásicos operones de los organismos procariotas; sin embargo, los genes incluidos en DGCs no presentan asociación funcional en tripanosomátidos (Palenchar & Bellofatto, 2006). A su vez, este tipo de organización de los genes determina la existencia de sitios denominados *strand switch regions* donde se invierte el sentido de la transcripción, y que juegan roles clave en el inicio de esta. La ausencia de promotores canónicos para la ARN polimerasa II y la escasa presencia de genes codificantes para factores de transcripción en el genoma (Palenchar & Bellofatto, 2006), sumado a este tipo de organización genómica, sugieren que la regulación de la expresión génica se da principalmente a nivel post-transcripcional (Kramer, 2012).

### 1.2.4 Regulación de la expresión génica

En la mayoría de los eucariotas, la primera etapa de regulación de la expresión génica está dada a nivel de la transcripción, mediante mecanismos tales como la alteración del estado de compactación de la cromatina, metilación del ADN, expresión de factores de transcripción, presencia de elementos potenciadores o represores, regulación de la formación del complejo de iniciación, etc.

En el caso de *T. cruzi*, así como de los tripanosomátidos en general, la regulación de la expresión génica a nivel transcripcional es limitada, por lo que la expresión de sus genes está regulada principalmente a nivel post-transcripcional. Esta afirmación está apoyada por el hecho de que genes que son transcriptos en la misma unidad policistrónica pueden presentar diferentes niveles de ARNm en estado estacionario, evidenciando la existencia de mecanismos de regulación que operan luego de la transcripción. Estudios transcriptómicos mediante microarreglos en las diferentes etapas del ciclo de vida de *T. cruzi* encontraron diferencias en los niveles de estado estacionario de los transcriptos (Minning et al., 2009).

La expresión de los genes codificantes de *T. cruzi* ocurre de forma bidireccional entre DGCs y de forma policistrónica, generando transcriptos primarios que contienen secuencias codificantes para varias proteínas, sin intrones, todas en una misma molécula de ARN. Estos ARNm primarios son procesados co-transcripcionalmente mediante dos mecanismos moleculares para generar transcriptos monocistrónicos maduros: *trans-splicing* y poliadenilación (Araújo & Teixeira, 2011a). Por un lado, el *trans-splicing* consiste en la adición de un miniexón de 39 pares de bases, que incorpora una estructura de caperuza a las regiones 5’ de los diferentes genes incluidos en un mismo ARNm primario. Este miniexón proviene de un ARN SL de aproximadamente 120 pares de bases que se encuentra repetido en tándem en el genoma permitiendo de esta forma suministrar el ARN SL en grandes cantidades que son requeridas por la célula, y que es posteriormente procesado. La adición del miniexón se produce en una secuencia consenso formada por dinucléotidos AG corriente arriba del codón de iniciación del gen (Daniels et al., 2010). Con respecto al resto de eucariotas superiores, la estructura del CAP en el extremo 5’ presenta un mayor número de modificaciones, que consiste en una 7-metilguanosina además de grupos 2’O-metilo en los cuatros primeros nucleótidos (Bangs et al., 1992). En cuanto a la poliadenilación del extremo 3’, se sabe que dicho proceso está acoplado al *trans-splicing*, pero no está descripta una secuencia consenso que actúe como señal (**figura 3**).

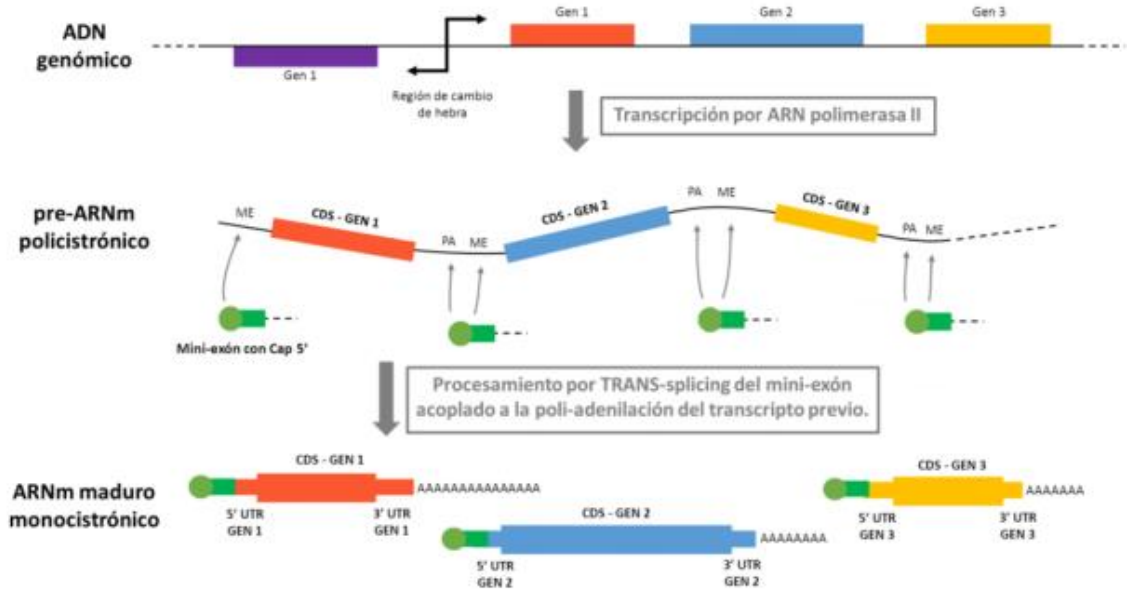



**Figura 3.** Expresión génica en tripanosomátidos. Se esquematizan las etapas de transcripción y procesamiento de los ARNm primarios. Tomado de la tesis de maestría de Santiago Chávez, 2016.

Las moléculas de ARNm maduro son uno de los principales blancos para la regulación génica, que se da mediante diferentes mecanismos. Uno de ellos es la regulación de los niveles de estado estacionario de los ARNm: la estabilización o degradación modula su vida media en los diferentes estadios del ciclo de vida del parásito o en las diferentes condiciones en las que vive, y está principalmente determinada por secuencias presentes en las regiones no traducidas 5' y 3' (UTRs) del ARNm y proteínas que interactúan con ellas, principalmente a nivel del 3' UTR (Coughlin et al., 2000; Di Noia et al., 2000; Vanhamme & Pays, 1995).

En células eucariotas existe lo que se denominan regulones post-transcripcionales: grupos de ARNm que codifican para proteínas relacionadas funcionalmente y los factores de acción en *trans* que modulan de forma coordinada su expresión. Esta agrupación de ARNm está conformada por proteínas de unión a ARN (RBPs) específicas de secuencia que se unen a elementos reguladores en *cis* en las regiones UTR de los ARNm y determinan su destino. La existencia de los regulones permite a las células controlar la síntesis proteica de genes que están dispersos a lo largo del genoma pero que están relacionados funcionalmente. Estos elementos a menudo se caracterizan por poseer motivos conservados a nivel de secuencia y/o estructuras secundarias de ARN, y es probable que muchos de estos elementos permanezcan sin descubrir (Keene, 2007). La

evidencia actual sugiere que los regulones gobiernan la expresión génica en los tripanosomátidos (Noé et al., 2008a; Ouellette & Papadopoulou, 2009; Queiroz et al., 2009). Por ejemplo, el estudio de (De Gaudenzi et al., 2013) observó la existencia de motivos de ARN conservados enriquecidos en las regiones UTR de 38 de 53 grupos de ARNm metabólicamente relacionados. Estos motivos tienen una estructura de horquilla y están ubicados muy cerca de los marcos de lectura abiertos (ORFs); 15 de estos 38 grupos contienen motivos únicos en los que la mayoría de los elementos presentes en el 3' UTR son específicos del grupo. A su vez, 13 motivos mostraron una fuerte correlación con grupos de genes co-expresados en el desarrollo de *T. cruzi*.

Otro mecanismo altamente conservado en eucariotas es la degradación de los ARNm mediante la acción de exonucleasas que actúan removiendo el CAP 5’ y la cola poliA. Se ha descrito que la maquinaria de la vía de degradación de la CAP, junto con ARNm y otras proteínas de unión al ARN, se concentran formando gránulos en el citoplasma (*P-bodies*), que han sido reportados en *T. cruzi* (Barbieri Holetz et al., 2007). Este tipo de estructuras funcionarían como reservorios de ARNm que permiten modular su degradación o devolverlos para ser traducidos, en función de las condiciones a las que esté expuesta la célula.

La traducción parece ser otro punto clave de regulación de la expresión génica en tripanosomátidos, sin embargo, aún no están del todo dilucidados los mecanismos asociados. En organismos que presentan una fuerte regulación traduccional, los niveles de transcriptos, cuantificados mediante microarreglos o *RNA-seq*, no reflejan de forma adecuada la cantidad de proteína presente en la célula. Estudios transcriptómicos y traductómicos realizados por nuestro grupo mediante *RNA-seq* y *ribosome-profiling* indican que existe una mayor correlación entre el traductoma y el proteoma de *T. cruzi*, con respecto al transcriptoma (Smircich et al., 2015), lo que sugiere la presencia de mecanismos de regulación actuando a nivel traduccional. A su vez, este estudio demuestra la presencia de grandes diferencias en la eficiencia traduccional de diferentes transcriptos presentes en el mismo estadio del ciclo de vida del parásito, y de iguales transcriptos presentes en diferentes estadios del ciclo de vida (específicamente epimastigota y tripomastigota metacíclico).

La regulación de la traducción estaría dada principalmente a nivel de la formación del complejo de iniciación de la traducción y la posterior elongación. En eucariotas, el CAP 5’ del ARNm se une a un complejo de exportación nuclear y, una vez fuera del núcleo, el factor

de inicio de la traducción eIF4E se une al CAP 5’ quien forma parte del complejo eIF4F. El extremo 5’ UTR del ARNm es escaneado por el ribosoma hasta el primer codón AUG, donde comienza la síntesis del polipéptido. Interacciones entre eIF4F y la cola poliA resulta en la circularización del ARNm, acercando los extremos 3’ y 5’ UTR, que pueden presentar unidas proteínas represoras de la síntesis proteica (Clayton & Shapira, 2007). En *T. cruzi* se han reportado factores de inicio de la traducción homólogos a los mencionados (Zinoviev & Shapira, 2012).

Hace varias décadas se ha reconocido que existen diferencias en el uso de codones tanto entre genes como entre especies. Estas diferencias se refieren a la frecuencia con la que se presentan los codones sinónimos. Durante mucho tiempo se creyó que las sustituciones sinónimas de codones no tenían ningún impacto en la expresión de los genes, debido a su aparente naturaleza silenciosa. Sin embargo, la tasa traduccional puede estar influida por el uso diferencial de codones sinónimos de cada gen: genes de alta expresión estarían optimizados para mejorar la eficiencia y fidelidad traduccional (Hershberg & Petrov, 2008). Este fenómeno se conoce en tripanosomátidos desde hace tiempo, con estudios previos a la publicación de los genomas completos y por lo tanto realizados en un set reducido de genes (Alvarez et al., 1994). Estudios realizados en los últimos años han evidenciado que incluso una única sustitución sinónima puede tener consecuencias significativas en los niveles de expresión génica, el plegamiento de proteínas y su función celular (Angov, 2011; Jeacock et al., 2018; Plotkin & Kudla, 2011). A su vez, se ha observado que el uso de codones influye en el procesamiento co-traduccional del péptido en formación. El uso de codones "raros" puede ralentizar la traducción y facilitar dicho procesamiento. Además, afecta la estabilidad y degradación de los ARNm, entre otros aspectos. Se ha reportado que en tripanosomátidos, como *Trypanosoma brucei*, el uso de codones y la disponibilidad de ARNt correspondientes desempeñan un papel importante en el control de la expresión génica (Horn, 2008a). En un estudio realizado por el grupo de Horn (Horn, 2008b), se demostró experimentalmente en *T. brucei* que el uso de codones tiene influencia en la regulación de la abundancia relativa de proteínas. Además, lograron predecir la abundancia relativa de ARNm y proteínas basándose exclusivamente en el uso de codones presentes en las secuencias codificantes (Horn, 2008b; Jeacock et al., 2018).

## 1.3 Estudio de redes

### 1.3.1 Teoría de grafos

La teoría de grafos se ha desarrollado dentro del campo de las matemáticas para el estudio y análisis de las relaciones entre entidades. Un grafo se compone por un conjunto de nodos o vértices, y un conjunto de aristas que conectan vértices entre sí. La teoría de grafos es aplicable a una amplia variedad de disciplinas, desde la matemática, la estadística, la física, la informática, la ingeniería y la sociología hasta la biología y la medicina. Los grafos se utilizan como modelos para una amplia variedad de sistemas complejos, desde redes sociales hasta redes de transporte, circuitos eléctricos y redes biológicas (Koutrouli et al., 2020).

Existen diferentes tipos de grafos que se utilizan para modelar diferentes tipos de sistemas. Uno de los tipos más comunes es el grafo no dirigido (*undirected*), en el que las aristas no tienen una dirección específica, es decir, no establece ningún orden entre los nodos que conectan. Por otro lado, los grafos dirigidos (*directed*) tienen aristas con una dirección específica, que indica el sentido en que fluye la relación entre dos nodos (Koutrouli et al., 2020; Pavlopoulos et al., 2011).

Otro tipo de grafo es el grafo ponderado (*weighted*), en el que cada arista tiene un peso o valor asignado que indica la fuerza o importancia de la relación entre dos nodos. Por otra parte, en los grafos no ponderados (*unweighted)* la importancia de la relación entre dos nodos radica únicamente en si existe o no una arista que los une (Pavlopoulos et al., 2011) (**figura 4**).

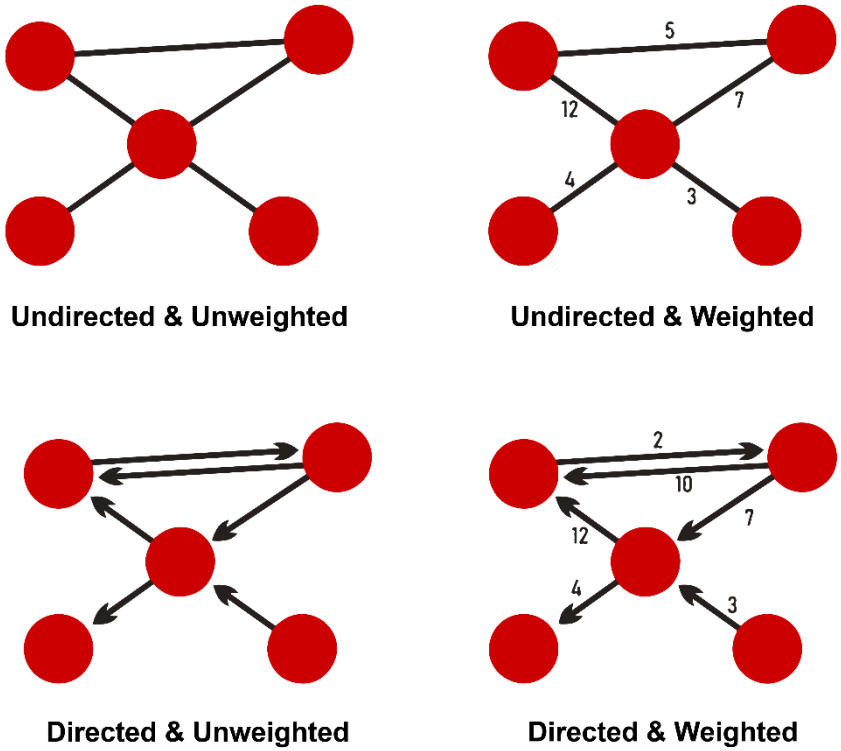



**Figura 4.** Esquema de tipos de grafos. Extraído y modificado de https://medium.com/tebs-lab/types-of-graphs-7f3891303ea8

En cuanto a las propiedades de los grafos, se pueden distinguir dos tipos, dependiendo del enfoque. Por un lado, las propiedades locales de un grafo se refieren a las características de los nodos y las aristas individuales en el grafo. Una de las propiedades más importantes de los nodos es cuan conectados están con el resto; en el caso de los grafos *unweighted,* se define el grado como número de aristas que están conectadas a él, mientras que en el caso de los grafos *weighted* definimos conectividad como la sumatoria de los pesos de todas las aristas que están conectados a él. En determinadas aplicaciones, los nodos con alta conectividad se consideran de gran relevancia porque tienen una gran cantidad de conexiones y pueden ser críticos para la conectividad global de la red.

Otra propiedad local importante de los grafos es la centralidad, que mide la importancia de un nodo con respecto a cierto criterio. Por ejemplo, la centralidad de intermediación mide la frecuencia con la que un nodo se encuentra en los caminos más cortos entre otros nodos, o la centralidad de cercanía que mide la distancia promedio entre un nodo y todos los demás nodos.

Por otro lado, las propiedades globales de los grafos se refieren a las características del grafo en su conjunto. Una de ellas es la conectividad, que básicamente es una medida

que refleja la facilidad con la que se puede llegar de un nodo a otro en la red. Los grafos pueden ser conexos o desconexos, dependiendo de si existe un camino entre todos los pares de nodos o no.

Otra propiedad global importante de los grafos es la transitividad, que mide la probabilidad de que al seleccionar uniformemente al azar tres nodos $x$, $y$, $z$ tales que $x$ y $z$ están ambos conectados a $y$, ocurra que también $x$ y $z$ están conectados entre sí. Un grafo con alta transitividad es más propenso a formar comunidades o módulos, donde un grupo de nodos está altamente conectado entre sí y menos conectado con los nodos fuera del grupo (Koutrouli et al., 2020; Pavlopoulos et al., 2011).

Otro aspecto que destacar de los grafos es su topología, que se refiere al estudio de la estructura y las propiedades que emergen de su análisis. Por un lado, los grafos *assortative* donde los nodos tienden a conectarse preferentemente con otros nodos que tienen características similares. Es decir, los nodos con alto grado de conectividad suelen estar conectados entre sí, mientras que los nodos con bajo grado de conectividad también se conectan entre sí. Esta estructura puede ser común en muchos sistemas biológicos y sociales, donde los nodos con características similares tienden a formar comunidades o agrupaciones. Por otro lado, en los grafos *disassortative*, los nodos tienden a conectarse preferentemente con otros nodos que tienen características diferentes. Es decir, los nodos con alto grado de conectividad tienden a estar conectados con nodos de bajo grado, y viceversa. Por último, las redes *scale-free* son aquellas en las que la distribución de grados de los nodos sigue una ley de potencia, lo que significa que la frecuencia de nodos con grado $k$ es aproximadamente proporcional a $k^{-\alpha}$, para cierta constante positiva $\alpha$ o, equivalentemente, el logaritmo de la cantidad de nodos con grado *k* es aproximadamente lineal en el logaritmo de *k*. En términos cualitativos, esto se traduce en que hay unos pocos nodos altamente conectados (llamados "*hubs*") y la mayoría de los nodos tienen un grado bajo de conectividad (Barabási & Albert, 1999; Newman, 2002).

### 1.3.2 Redes biológicas

Una red biológica es un término utilizado para describir un sistema complejo compuesto por elementos biológicos interconectados que interactúan entre sí desempeñando un papel fundamental en el funcionamiento y la regulación de los sistemas biológicos. Esta se puede representar como un grafo donde los nodos representan las

entidades biológicas, tales como genes, proteínas, o metabolitos, y las aristas representan las interacciones entre ellas. Estas interacciones pueden ser físicas, como la unión de una proteína a otra, o funcionales, como la regulación de la expresión génica. La representación de una red biológica como un grafo permite analizar la estructura de la red y descubrir patrones que podrían no ser evidentes de otra manera, es decir, obtener información que no se podría obtener analizando los componentes por separado, dando lugar a lo que se denominan propiedades emergentes (Barabási & Oltvai, 2004).

En la biología podemos identificar diferentes tipos de redes; a nivel de organismos, las redes de interacciones bióticas como las redes tróficas se pueden entender como un conjunto de organismos que están conectados dependiendo de qué organismo se alimenta de qué otro, representándose como un grafo de tipo dirigido; a nivel celular o de tejidos, se puede estudiar cómo son las interacciones que se dan entre las neuronas, por ejemplo las interacciones neuronales que se dan en determinadas regiones del cerebro a la hora de realizar cierta tarea o frente a determinados estímulos; y a nivel molecular se pueden identificar redes como las metabólicas, las de señalización, las de interacciones entre proteínas, o las redes de co-expresión génica.

Un aspecto fundamental del estudio de las redes biológicas son las potencialidades que presenta:

- Encontrar componentes particularmente importantes dentro de la red.
- Identificar comunidades donde sus componentes tienen mayores relaciones entre ellos que con el resto de los elementos de la red.
- Identificar nuevos genes asociados a enfermedades en el caso de redes de interacción génica.
- Formular nuevas hipótesis sobre la función de proteínas en el caso de redes de interacción de proteínas.

La comprensión biológica a partir de las redes se obtiene a través del análisis de la topología de la red. La topología de la red se refiere a la disposición de nodos y aristas dentro de una red (Albert, 2005), cuyas propiedades se utilizan para obtener información biológica relevante sobre los nodos y las aristas. Los nodos con un alto grado de conexiones y un alto coeficiente de agrupamiento se convierten en centros y es probable que estén asociados con genes esenciales en la red (Albert, 2005). Basándose en el grado de los nodos en las redes biológicas, la mayoría de las redes se consideran redes *scale-free*

(Albert, 2005; Barabási & Oltvai, 2004). Como se mencionó en el apartado anterior, las redes *scale-free* se caracterizan por el hecho de que la mayoría de los nodos están conectados a unos pocos vecinos, mientras que un pequeño número de nodos, los centros, están conectados a un alto número de vecinos.

Las redes biológicas también forman módulos que están formados por elementos altamente conectados en la red, que forman subredes que corresponden a características biológicas específicas. Los módulos en una red biológica actúan como puntos de partida para estudios adicionales y, por lo tanto, reducen la complejidad global de la red.

### 1.3.3 Redes de co-expresión génica

Los avances tecnológicos de los últimos años han permitido el desarrollo de instrumentos capaces de secuenciar a gran profundidad las moléculas de ARNm presentes en las células, particularmente las técnicas de *RNA-seq*, permitiendo estimar con eficiencia y precisión los niveles de expresión y estado estacionario de miles de genes de forma simultánea (Z. Wang et al., 2009). A su vez, estos avances y la disminución de los costos asociados generaron una gran producción y acumulación de datos producidos por la comunidad científica.

Por otro lado, esta gran cantidad de datos generados y disponibles en las bases de datos públicas permitieron el surgimiento de nuevas aproximaciones para su análisis, como lo son la generación y el estudio de redes de co-expresión génica. Brevemente, una red de co-expresión génica es un grafo que muestra la relación existente entre los genes en función de su co-expresión, es decir, la medida en que se expresan de forma conjunta en, por ejemplo, diferentes condiciones experimentales, distintos momentos del ciclo de vida celular, diferentes condiciones ambientales en que se encuentra la célula, etc. Estas redes se generan a partir de datos de expresión génica y proporcionan una visión global de las interacciones entre los genes, donde los nodos en la red representan genes y las aristas representan la fuerza de la correlación de la expresión entre ellos.

Este tipo de análisis ha permitido a los científicos identificar grupos de genes que se expresan de forma conjunta y, por lo tanto, que pueden estar involucrados en procesos biológicos comunes (Kharchenko et al., 2005). Por ejemplo, se han utilizado redes de co-expresión génica para identificar grupos de genes implicados en la enfermedad de

Alzheimer, el cáncer, la diabetes y otras enfermedades (Liang et al., 2018; Oldham et al., 2012; Riquelme Medina & Lubovac-Pilav, 2016; Tian et al., 2018). Además, estas redes se han utilizado para identificar genes centrales que son esenciales para la función celular y pueden ser potenciales blancos moleculares para el desarrollo de fármacos.

La construcción de las redes de co-expresión génica se basa en establecer la asociación de la expresión de genes en un conjunto de muestras, que se puede medir a través de diferentes métodos estadísticos como la correlación de Pearson, la correlación de Spearman, u organizando las muestras en forma vectorial en términos de distancias como distancia euclídea, coseno, etc. A partir de esta medida de asociación, se pueden construir matrices de adyacencia que representan la fuerza de la co-expresión entre los genes. Estas matrices se pueden transformar en grafos ponderados no dirigidos, donde los nodos representan los genes y las aristas representan la fuerza de la co-expresión entre ellos.

En última instancia las redes de co-expresión génica permiten identificar grupos/módulos/*clusters* de genes con perfiles de expresión similares y con funciones biológicas comunes, y genes que cumplen funciones clave para la regulación de los procesos biológicos específicos subyacentes al grupo (**figura 5**). Entre los análisis más utilizados para el estudio de este tipo de redes se encuentran:

- Análisis de modularidad: la modularidad es una medida que cuantifica la división de la red en módulos de genes altamente interconectados. La identificación de módulos de co-expresión puede ayudar a identificar genes que participan en procesos biológicos específicos.
- Análisis de enriquecimiento funcional: el análisis de enriquecimiento funcional permite identificar las funciones biológicas y procesos celulares que están sobrerrepresentados en un módulo de genes co-expresados.
- Identificación de *hubgenes*: los genes que actúan como nodos centrales en la red (*hubgenes*), es decir, que tienen un alto grado de conectividad con otros genes, pueden ser candidatos para la regulación de procesos biológicos de importancia.
- Anotación funcional de genes: al estar los módulos asociados con procesos biológicos específicos, es posible inferir funciones para genes con anotación desconocida en función de los genes conocidos que se agrupan con ellos en la red de co-expresión, basándose en el principio conocido como *guilt by association*.

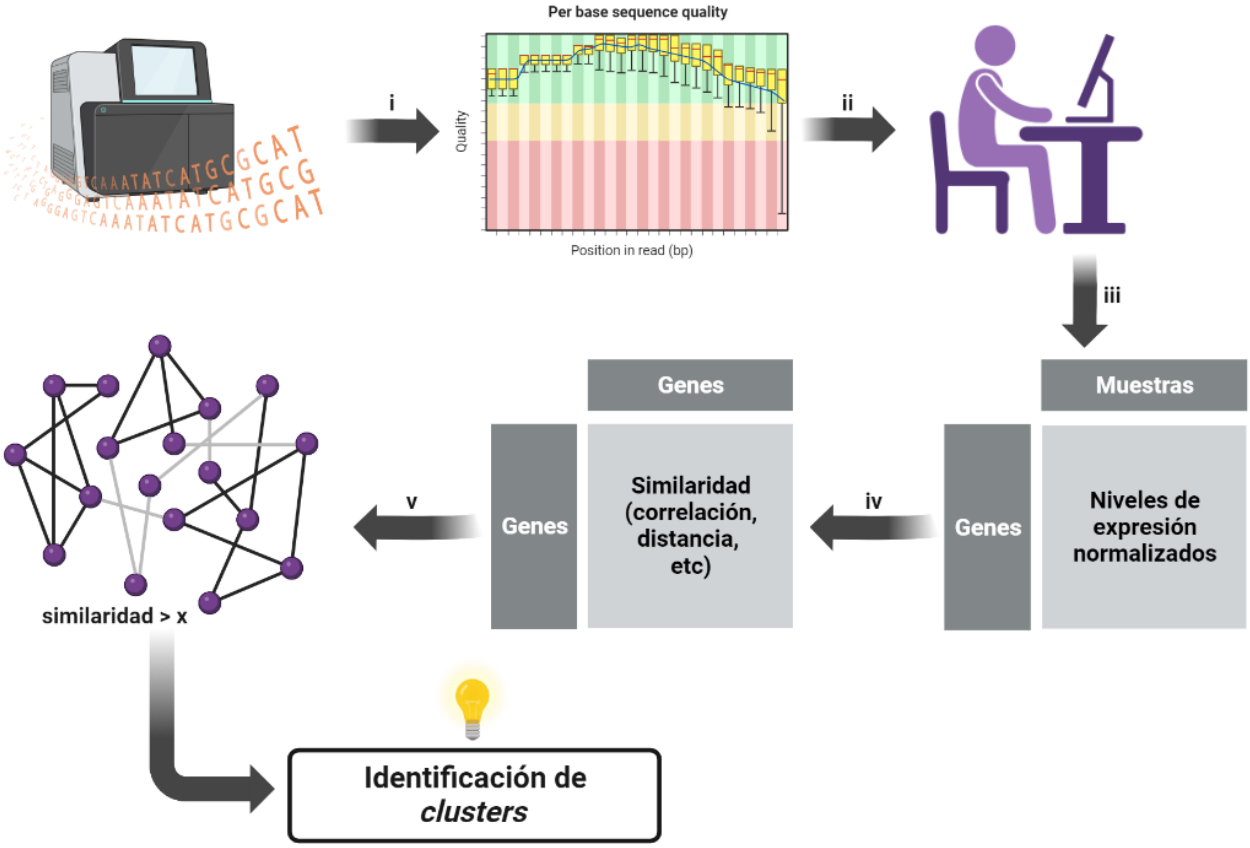



**Figura 5.** Esquema de la construcción de una red de co-expresión génica. Se parte de datos de expresión génica, se realiza un análisis de calidad de estos datos, se realiza la estimación de los perfiles de expresión, se construyen matrices de similaridad y se establece un valor de *cutoff* para la identificación de grupos de genes co-expresados.

### 1.3.4 *Gene Ontology Resource*

El Gene Ontology (GO) es un recurso de anotación y una herramienta bioinformática que se utiliza para describir y clasificar las funciones biológicas, los procesos celulares y las localizaciones subcelulares de los genes y sus productos (Ashburner et al., 2000; Carbon et al., 2021). La ontología del GO está organizada como un grafo acíclico dirigido (DAG) en el que los términos se organizan en diferentes niveles jerárquicos donde los términos más generales se encuentran en la parte superior de la jerarquía y los términos más específicos se encuentran en la parte inferior, y las relaciones entre los términos representan la relación entre los conceptos biológicos que describen.

Los términos del GO se dividen en tres categorías principales: “función molecular” (MF), “proceso biológico” (BP) y “componente celular” (CC). La categoría MF describe las

funciones moleculares de los genes y proteínas, como “catalizador” y “transportador”, la categoría de BP describe los procesos biológicos en los que están involucrados los genes y proteínas, como “metabolismo”, “transporte” y “señalización”, mientras que la categoría de CC describe las partes celulares en las que se encuentran los genes y proteínas, como “núcleo”, “mitocondria” y “membrana plasmática”.

Los términos GO están conectados entre sí por diversas relaciones, incluyendo:

- “*is_a*”: los términos que representan subtipos o subordinados de otro término. Por ejemplo, “fosforilación” es un subtipo del término “modificación de proteínas”.
- “*part_of*”: los términos que representan partes de otro término. Por ejemplo, “cilios” es un término que es una parte de “membrana celular”.
- “*regulates*”: los términos que representan la regulación de otro término. Por ejemplo, “proliferación celular” es un término que puede regular “apoptosis”.
- “*positively_regulates*”: los términos que representan la regulación positiva de otro término. Por ejemplo, “proliferación celular” es un término que puede tener un efecto positivo en “ciclo celular”.
- “*negatively_regulates*”: los términos que representan la regulación negativa de otro término. Por ejemplo, “inhibición de la apoptosis” es un término que puede tener un efecto negativo en “apoptosis”.
- “*has_part*”: los términos que representan las partes de otro término. Por ejemplo, “ribosoma” es un término que puede tener “proteínas ribosomales” como parte.
- “*has_function*”: los términos que representan las funciones de otro término. Por ejemplo, “receptor de membrana” es un término que puede tener una “actividad de unión a ligandos” como función.

El DAG del GO permite la búsqueda y la comparación de genes en función de sus funciones biológicas, procesos y componentes celulares. La estructura jerárquica del DAG permite la navegación y la exploración de los términos relacionados y las relaciones entre ellos. Además, los términos pueden tener múltiples padres y pueden estar relacionados con diferentes aspectos de la biología celular, lo que permite la exploración de la complejidad de las funciones biológicas (**figura 6**).

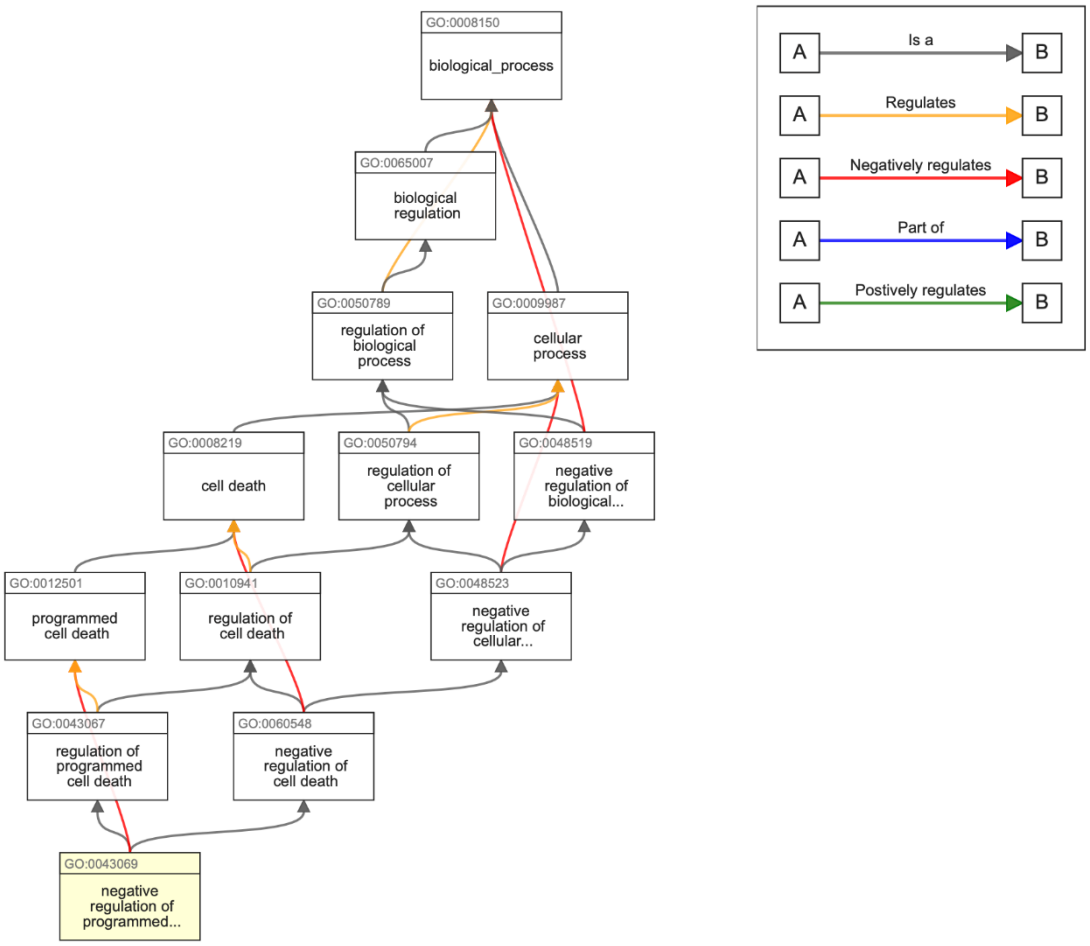



**Figura 6.** Ejemplo de la estructura del DAG del *Gene Ontology* para la ontología *biological process.* Tomado de https://advaitabio.com/faq-items/understanding-gene-ontology/

Por todo lo anterior, en este estudio se buscó contribuir a la comprensión de la regulación de la expresión génica de *T. cruzi* mediante la construcción de una red de co-expresión génica utilizando datos de *RNA-seq* de los cuatro estadios su ciclo de vida, y la caracterización y análisis funcional de los grupos de genes co-expresados obtenidos. A su vez, se buscó identificar posibles mecanismos que expliquen esta regulación conjunta.

# *2* Objetivos

## 2.1 Objetivo general

Profundizar en la comprensión de la regulación de la expresión génica de *Trypanosoma cruzi* mediante la caracterización y análisis funcional de grupos de genes co-expresados a lo largo de todo su ciclo de vida.

## 2.2 Objetivos específicos

1. Estimar los perfiles de expresión génica a partir de datos de *RNA-seq* evaluando el desempeño de diferentes estrategias.
2. Identificar grupos de genes co-expresados.
    - 2.a. Implementar diferentes métodos de agrupamiento no supervisado a partir de datos de expresión.
    - 2.b. Evaluar el desempeño de las metodologías implementadas mediante la evaluación de la calidad de los grupos generados.
3. Estudiar las características funcionales de los grupos de genes co-expresados.
    - 3.a. Realizar un estudio de enriquecimiento funcional mediante la evaluación de sobrerrepresentación de términos GO.
    - 3.b. Correlacionar funciones y momentos de expresión de los grupos con la biología parasitaria.
4. Analizar características comunes en los grupos de genes co-expresados.
    - 4.a. Buscar motivos en regiones reguladoras.
    - 4.b. Analizar el uso diferencial de codones en los distintos grupos.
5. Realizar inferencias funcionales para genes de interés de función desconocida.

# *3* Materiales y Métodos

## 3.1 Obtención y reporte de calidad de los datos

Los datos de *RNA-seq* de los diferentes estadios de *Trypanosoma cruzi* fueron obtenidos de diferentes trabajos publicados: epimastigotas en distintas fases de la curva de crecimiento *in vitro* (7, 14, 21 y 28 días) de (Smircich et al., 2023) disponibles en el *National Center for Biotecnology* (NCBI) *Sequence Read Archive* (SRA) en el proyecto PRJNA915394; tripomastigotas metacíclicos de (Cruz-Saavedra et al., n.d.) disponibles en el *European Nucleotide Archive* (ENA) en el proyecto PRJEB33521; amastigotas celulares a distintas horas post-infección (4, 6, 12, 24, 48 y 72 horas) y tripomastigotas celulares de (Y. Li et al., 2016) disponibles en el NCBI SRA en los proyectos PRJNA251582 y PRJNA251583 respectivamente. Se realizó el análisis de calidad de cada uno de los *sets* de datos utilizando la herramienta *FastQC* (Andrews, 2010) versión 0.11.9. El reporte generado por *FastQC* incluye diversas secciones, como la distribución de calidad de las bases en cada posición de las lecturas, la calidad promedio de las lecturas, la presencia de bases ambiguas, la presencia de adaptadores o secuencias no deseadas, la distribución de longitudes de las lecturas, entre otros aspectos.

## 3.2 Construcción de perfiles de expresión

A partir de los datos de lecturas de secuenciación mencionados anteriormente se estimaron los perfiles de expresión génica.

### 3.2.1 Evaluación de métodos de construcción de perfiles de expresión

La construcción de los perfiles de expresión fue optimizada evaluando el desempeño en la asignación de lecturas simuladas a los genes de *T. cruzi* de 3 de los programas más utilizados: *kallisto* (Bray et al., 2016)*, Bowtie2* (Langmead & Salzberg, 2012) y *STAR* (Dobin et al., 2013). Para la simulación de lecturas se procedió en primer lugar a generar una tabla primaria de conteos de lecturas asignadas a genes utilizando los archivos de lecturas *.fastq* (*paired-end*) de epimastigotas a 7 días de crecimiento (réplica 1) de (Smircich et al., 2023) mediante el *software kallisto*, con el fin de partir de una tabla de conteos construida a partir

de datos reales. Esta tabla contiene la cantidad de lecturas que fueron asignadas a cada uno de los genes por *kallisto*.

Obtenida esta tabla de conteos se plantearon 3 estrategias complementarias para simular lecturas (*paired-end*):

1- Se simularon las lecturas para cada uno de los genes respetando la cantidad de lecturas determinada por la tabla de conteos.
2- Se alteró de forma aleatoria la tabla de conteos original, estableciendo un nuevo valor de conteo en un rango de -15% a +15% del valor original. Se repitió esto 5 veces generando 5 nuevas tablas de conteos con las cuales se simularon las lecturas.
3- Se alteró toda la tabla, generando una nueva tabla de conteos para cada gen en un rango de [0, 8000] utilizando la función de R *sample()* con el argumento *replace= TRUE*, con el cual se samplea utilizando el algoritmo Mersenne twister (Matsumoto & Nishimura, 1998) de generación de números pseudoaleatorios con distribución uniforme, con la cual se simularon las lecturas.

En todos los casos se utilizó el paquete de R *Polyester* (Frazee et al., 2015) a partir del CDS (secuencias codificantes) del genoma de referencia de *T. cruzi CL-Brener Esmeraldo-like* (v.50). Esta simulación consiste en generar fragmentos de 100 nucleótidos a partir de los transcriptos incluídos en el CDS, incorporando una tasa de error del 0,5% que refleja los errores típicos de la secuenciación de ARN mediante la tecnología Illumina, bajo un modelo de error uniforme, donde cada nucleótido en una lectura tiene la misma probabilidad ($p_e = 0{,}005$) de ser secuenciado incorrectamente, y todos los posibles errores de secuenciación son igualmente probables. Por ejemplo, si hay un error en un nucleótido que se suponía que era una T, la base incorrecta tiene la misma probabilidad de ser una G, C, A o N.

Generados estos 7 nuevos sets de datos de lecturas pareadas (*.fastq paired-end*) se procedió a realizar las cuantificaciones de expresión génica con *kallisto*, *Bowtie2* y *STAR* (en los últimos dos casos realizando el conteo con el *software featureCounts* (Liao et al., 2014)), obteniendo 7 nuevas tablas con conteos de lecturas asignadas a los genes de *T. cruzi*.

Utilizando estas tablas se calculó un índice de correlación de Pearson entre cada una de las tablas de conteos obtenidas con las diferentes estrategias y la tabla de conteos

a partir de la cual se realizó cada simulación de lecturas; obteniendo, en última instancia, un índice de correlación de Pearson para cada combinación *software*-estrategia; en el caso de la estrategia de alteración al azar de algunos conteos, para obtener este único índice se promediaron los índices de las 5 réplicas.

Para visualizar las diferencias entre los conteos originales y los obtenidos por las diferentes estrategias, se seleccionaron los genes que tenían un conteo de lecturas original distinto de 0 y el s*oftware* le asignó más o menos lecturas y se calculó un porcentaje de diferencia de conteo de la forma $\frac{\text{conteo}-\text{conteo original}}{\text{conteo original}} \text{x } 100$. Por otro lado, para aquellos genes que tenían un conteo de lecturas original igual a 0 y el *software* le asignó más o menos lecturas la diferencia no se calculó en forma de porcentaje, sino en cantidad de lecturas incorrectamente asignadas de la forma $|conteo - conteo\ original|$. Las diferencias fueron visualizadas mediante histogramas en R.

### 3.2.2 Construcción de perfiles de expresión final

A la luz de los resultados del proceso de optimización descrito anteriormente, para la cuantificación final de las lecturas asignadas a los genes de *T. cruzi* se utilizó el *software kallisto* versión 0.46.1 (ver sección 4.2). Una vez obtenidos los conteos se utilizó la función *rlog* del paquete de R *DESeq2* (Love et al., 2018) que, brevemente, los transforma a una escala logarítmica donde atenúa la diferencia de las varianzas entre genes de altos y bajos conteos, y normaliza respecto al tamaño de las librerías de las diferentes muestras.

## 3.3 Estrategias para la construcción de redes de co-expresión génica

Se utilizaron los paquetes de R *WGCNA*, *CEMiTool* y *coseq* (Langfelder & Horvath, 2008; Rau & Maugis-Rabusseau, 2018; Russo et al., 2018) y el *software clust* (Abu-Jamous & Kelly, 2018) para la construcción de las diferentes redes de co-expresión génica.

Por un lado, *WGCNA* construye una *weighted gene co-expression network.* Brevemente, el algoritmo toma la matriz de *genes x muestras* con los valores de expresión génica y calcula una matriz de correlación de Pearson *genes x genes* a partir de la cual se realizan sucesivas transformaciones:

1- Construye una matriz de similaridad llevando los valores a correlación al intervalo [0-1] mediante la siguiente fórmula: $s_{ij} = |cor(x_i, x_j)|$

2- Toma la matriz de similaridad y eleva los valores a un parámetro β que permite acentuar la diferencia entre genes de alta y baja similaridad: $a_{ij} = s_{ij}^{\beta}$. Para determinar este parámetro, se evalúa qué tan bien se ajusta la red a una topología *scale-free*, ajustando un modelo lineal para el logaritmo de la cantidad de nodos con conectividad $k$ en función del logaritmo de $k$, obteniendo un índice $R^2$, denominado *scale-free topology index*, en función de la bondad de dicho ajusto. Por último, se escoge el menor parámetro β que obtenga un *scale-free topology index* > 0,9.

3- Construye una *Topological Overlapping Matrix* (TOM): $w_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}$ donde $l_{ij} = \sum_u a_{iu} a_{uj}$ es la conectividad de los nodos i y j con genes terceros u, y $k_i = \sum_u a_{iu}$ es la conectividad del nodo i. Esta matriz define una medida de similaridad entre dos genes que no depende únicamente de su correlación de expresión, sino también de la correlación que tienen en común con genes terceros.

Por último, utilizando la matriz TOM se construye la matriz $diss\mathrm{TOM} = 1 - \mathrm{TOM}$ y se realiza *clustering* jerárquico utilizando la función *hclust* de R con *average linkage* para crear un árbol o dendrograma. A continuación, se utiliza el algoritmo *Dynamic Tree Cut* de la función *cutreeDynamicTree* del paquete de R *dynamicTreeCut* (Langfelder et al., 2008) y se agrupan los genes en módulos o *clusters* de genes co-expresados. Para cada uno de ellos, se creó un archivo de texto (*.txt)* con los identificadores de sus genes.

Por otro lado, *CEMiTool* implementa y automatiza las funcionalidades de *WGCNA* para la construcción de una red de co-expresión génica, con ciertas modificaciones principalmente en el método de selección del parámetro β.

En el caso de *coseq*, este método toma los conteos de los perfiles de expresión génica, realiza una normalización TMM (*trimmed mean of M-values*) y una transformación de tipo logarítmica denominada *logCLR* que ajusta los datos a un *Gaussian mixture model,* que básicamente asume que los datos contienen una mezcla de subpoblaciones de medidas de expresión. Por último, aplica el algoritmo de *clustering K-means*, tomando un rango de K (cantidad de módulos) de 2 a 30 entre los cuales selecciona el K óptimo, determinado por el algoritmo evaluando la inercia y entropía de los módulos generados , obteniendo en última instancia K módulos de genes co-expresados.

Por último, *clust* comienza con el preprocesamiento de los datos, que incluye la normalización y filtrado de genes con baja expresión. Luego, se aplican múltiples iteraciones de agrupamiento *K-means* con diferentes valores de K para producir un conjunto de *clusters* semilla. Si el conjunto de datos de entrada incluye más de un conjunto de datos, se calculan los *clusters* de consenso utilizando el método de binarización de matrices de partición de consenso (Bi-CoPaM) (Abu-Jamous et al., 2013). Estos *clusters* semilla se analizan para aprender las distribuciones de la dispersión dentro del *cluster* en los conjuntos de datos seleccionados. Esta información se utiliza para eliminar los valores atípicos de los *clusters* y para identificar los genes que se ajustan a los *clusters* pero que se han perdido en los pasos anteriores.

## 3.4 Métodos de evaluación y comparación de redes de co-expresión génica

Para la selección del método de construcción de la red de co-expresión génica final para realizar los análisis propuestos en este trabajo se utilizaron dos métodos basados en la determinación de la consistencia funcional de los módulos obtenidos por cada método. Por un lado, *Biological Homogeneity* Index (BHI, (Datta & Datta, 2006)) y *Wang Index* (J. Z. Wang et al., 2007).

*BHI* es una medida utilizada para evaluar la capacidad de un algoritmo de *clustering* para producir grupos biológicamente significativos. Básicamente mide cuán biológicamente homogéneos son los grupos, lo que indica qué tan bien el algoritmo agrupa genes con características biológicas similares. Una puntuación alta de *BHI* (1) indica que el algoritmo de agrupamiento ha producido con éxito grupos biológicamente significativos. Para calcular el *BHI* se realiza una comparación entre las clases funcionales de los genes anotados dentro de cada grupo, utilizando como criterio los términos GO que tienen en común. Considerando dos genes $\mathrm{x}$ e $\mathrm{y}$ que pertenecen al mismo *cluster* $\mathrm{D}$, $\mathrm{C(x)}$ y $\mathrm{C(y)}$ son todos los términos GO del gen $\mathrm{x}$ y del gen *y*, respectivamente, y la función indicatriz $\mathrm{I(C(x) = C(y))}$ que toma valor 1 si al menos uno de los de los términos GO del gen $\mathrm{x}$ coincide con los términos GO del gen $\mathrm{y}$, se calcula $\mathrm{BHI} = \frac{1}{\mathrm{k}}\sum_{\mathrm{j=1}}^{\mathrm{k}} BHI_j$ siendo $BHI_j = \frac{1}{\mathrm{n_j(n_j-1)}}\sum_{\mathrm{x\neq y\in D_j}} \mathrm{I(C(x) = C(y))}$ donde $\mathrm{k}$ es el número de *clusters* y $\mathrm{n_j}$ el número de genes anotados en $\mathrm{D_j}$. Para este estudio se calculó $BHI_j$ con el fin de evaluar la consistencia funcional de cada módulo independientemente.

Por otro lado, el *Wang Index* es una medida de similitud funcional entre genes dentro de los *clusters* que se basa en la idea de que la similitud semántica entre dos términos del GO se puede calcular a partir de la similitud semántica de sus términos GO ancestrales comunes. El *Wang Index* se calcula de la siguiente forma:

Formalmente, el término GO A se puede representar como $\mathrm{DAG_A} = (\mathrm{A}, \mathrm{T_A}, \mathrm{E_A})$, donde $\mathrm{T_A}$ es el conjunto de términos GO de $\mathrm{DAG_A}$ incluyendo al término GO A y todos sus ancestros, y $\mathrm{E_A}$ es el conjunto de relaciones semánticas (aristas en el DAG) que conectan los términos GO de $\mathrm{T_A}$. Se define la contribución del término GO t a la semántica del término GO A como el S-valor del término GO t respecto al término GO A, $\mathrm{S_A(t)}$, definido como:

$$\begin{cases} \mathrm{S_A(A)} = 1 \\ \mathrm{S_A(t)} = \max\{\mathrm{w_e} * \mathrm{S_A(t')} | \mathrm{t'} \in \text{children of (t)}\} \text{ if t} \neq \mathrm{A} \end{cases}$$

donde $\mathrm{w_e}$ es el factor de contribución semántica para la arista $\mathrm{e} \in \mathrm{E_A}$ que une el término GO t con su hijo t', y donde $0 < \mathrm{w_e} < 1$, el valor de este factor está implementado en el algoritmo y fue calculado por sus desarrolladores. Luego de obtener los S-valores de todos los términos GO de $\mathrm{DAG_A}$, se puede calcular el valor semántico del término GO A, $\mathrm{SV(A)}$ , como $\mathrm{SV(A)} = \sum_{\mathrm{t} \in \mathrm{T_A}} \mathrm{S_A(t)}$.

Por otro lado, es posible medir la similaridad semántica de dos términos GO A y B. Considerando $\mathrm{DAG_A} = (\mathrm{A}, \mathrm{T_A}, \mathrm{E_A})$ y $\mathrm{DAG_B} = (\mathrm{B}, \mathrm{T_B}, \mathrm{E_B})$ para los términos GO A y B respectivamente, la similaridad entre ellos dos, $\mathrm{S_{GO}(A, B)}$ se calcula como: $\mathrm{S_{GO}(A, B)} = \frac{\sum_{\mathrm{t} \in \mathrm{T_A} \cap \mathrm{T_B}} (\mathrm{S_A(t)} + \mathrm{S_B(t)})}{\mathrm{SV(A)} + \mathrm{SV(B)}}$. La ventaja de esta fórmula es que determina la similaridad semántica entre dos términos GO teniendo en cuenta su localización en el DAG y las relaciones semánticas entre sus términos GO ancestros.

Por último, es posible medir la similaridad entre dos genes. Asumiendo que $\mathrm{GO_1} = \{\mathrm{go_{11}}, \mathrm{go_{12}}, \ldots, \mathrm{go_{1m}}\}$ y $\mathrm{GO_2} = \{\mathrm{go_{21}}, \mathrm{go_{22}}, \ldots, \mathrm{go_{2n}}\}$ son dos conjuntos de términos GO de los genes $\mathrm{G_1}$ y $\mathrm{G_2}$, respectivamente, en primer lugar, se mide la similaridad semántica entre un término GO y un *set* de términos GO como $\mathrm{Sim(go, GO)} = \max_{1 \leq \mathrm{i} \leq \mathrm{k}} (\mathrm{S_{GO}(go, go_i)})$. Teniendo en cuenta esto, se calcula la similaridad funcional entre dos genes según sus términos GO como $\mathrm{Sim(G1, G2)} = \frac{\sum_{1 \leq \mathrm{i} \leq \mathrm{m}} \mathrm{Sim(go_{1i}, GO_2)} + \sum_{1 \leq \mathrm{j} \leq \mathrm{n}} \mathrm{Sim(go_{2j}, GO_1)}}{\mathrm{m+n}}$. Finalmente, el *Wang Index* para un grupo de genes se define como el promedio de las similaridades entre todos sus genes.

Para evaluar la existencia de diferencias estadísticamente significativas de los índices de cada algoritmo se realizó test de Wilcoxon (Wilcoxon, 1945) tomando de a pares los algoritmos y estableciendo la significancia como *p-value* < 0,05.

## 3.5 Análisis de enriquecimiento funcional de módulos de genes co-expresados

Para el análisis de enriquecimiento funcional de términos GO de la ontología "*biological process*" se utilizó la herramienta de interfaz gráfica *Gene Ontology Enrichment* disponible en TriTrypDB (Aslett et al., 2009), utilizando como entrada los identificadores de los genes para cada módulo, y se conservaron aquellos módulos que tuvieran términos GO sobrerrepresentados con un *p-value* < 0,05 y un *p-value* ajustado por Benjamini-Hochberg < 0,1 (Benjamini & Hochberg, 1995).

## 3.6 Análisis de correlación módulo – estadio

Para el análisis de correlación de la expresión de los genes de cada módulo enriquecido funcionalmente y el estadio del ciclo de vida de *T. cruzi* se calculó en primer lugar el *Module Eigengene* (ME) de cada uno de ellos. El ME es un vector que resume la expresión de todos los genes del módulo, lo que permite una descripción más compacta y simplificada de la información de expresión de los genes. Brevemente, los valores de expresión de los genes en el módulo se normalizan para que tengan una media cero y una desviación estándar de uno. Luego, se realiza un análisis de componentes principales (PCA) para calcular los componentes principales de la matriz de expresión génica del módulo. El primer componente principal (PC1) se define como el vector propio que explica la mayor parte de la variabilidad en los datos, donde cada valor del vector resume la expresión de los genes del módulo en cada una de las muestras. Por último, se construyó una única matriz con todos los MEs.

Por otro lado, se construyó una matriz *muestras x estadios*, es decir, 38 x 4 (epimastigota, tripomastigota metacíclico, amastigota y tripomastigota celular), donde cada celda tomaba un valor de 1 o 0 dependiendo si la muestra correspondía o no a cada estadio.

Por último, se realizó un análisis de correlación puntobiserial entre estas dos matrices, que permite medir la relación entre una variable dicotómica y una variable continua, a partir del cual se construyó un *heatmap* con las correlaciones ME – estadio.

## 3.7 Obtención de regiones 3'UTR

Se descargó el genoma y la anotación de la cepa CLBrener de *T. cruzi* del TriTrypDB (v.50) y, junto con las lecturas de secuenciación descritas anteriormente, se utilizó el *software UTRme* (Radío et al., 2018) para la predicción de las regiones 3'UTR de cada gen, conservando para cada uno de ellos la región 3'UTR reportada más larga.

## 3.8 Búsqueda de motivos de secuencia y estructurales en regiones 3'UTR

Para identificar motivos de secuencia en las regiones 3'UTR de los genes de cada módulo se tomaron los archivos *.fasta* generados con *UTRme* como entrada para el *software XSTREME* (Grant & Bailey, 2021) de *MEME-Suite* (Bailey et al., 2009). Brevemente, este *software* permite realizar un análisis completo de motivos a partir de secuencias de ADN, ARN o proteínas en cualquier región de las mismas (que pueden ser de cualquier largo y cuyo largo entre las secuencias puede variar), identificando motivos previamente reportados o no sobrerrepresentados en ellas, y reportando a su vez proteínas de unión previamente descritas. Para ello, se ingresa el archivo de secuencias donde se quieren identificar los motivos y un archivo de secuencias de *background* contra el que se compara si el motivo está o no sobrerrepresentado. Por esta razón se creó para cada uno de los archivos *.fasta* de entrada un archivo *.fasta* del resto de las regiones 3'UTR de los genes que no pertenecen al módulo evaluado.

Obtenidos los motivos para cada módulo, se procedió a utilizar el *software FIMO* (Grant et al., 2011) también de *MEME-Suite*, con el fin de obtener las proporciones de esos motivos tanto en los módulos como en los *backgrounds*, conservando en última instancia aquellos motivos que tuvieran una representación de al menos el doble en el módulo que en el *background*, es decir $\frac{\text{proporción en el módulo}}{\text{proporción en el background}} \geq 2$.

Para la búsqueda de motivos estructurales se utilizó la función *cmsearch* del *software Infernal* (Nawrocki & Eddy, 2013). En este caso, se realizó la búsqueda de motivos estructurales ya caracterizados para *Trypanosoma cruzi* en (Noé et al., 2008b; Sabalette et al., 2019) y *Trypanosoma brucei* (Estévez, 2008), utilizando como entrada las secuencias de estos motivos, los archivos *.fasta* de las regiones 3'UTR de los módulos enriquecidos

funcionalmente y los archivos *.fasta* del resto de las regiones 3'UTR de los genes que no pertenecen al módulo evaluado.

## 3.9 Análisis de uso diferencial de codones

Para el análisis de uso diferencial de codones se calculó en primer lugar la frecuencia de codones sinónimos de cada uno de los genes de cada módulo, generando una tabla *genes x codones* para cada uno de ellos, a partir de la cual se calculó la frecuencia promedio de cada codón sinónimo para cada módulo, obteniendo así una tabla *módulo x codones*. A continuación, se utilizó el paquete de R *Rtsne* (https://github.com/jkrijthe/Rtsne) para realizar un análisis de reducción de dimensionalidad mediante *t-Distributed Stochastic Neighbor Embedding (t-SNE).*

Por otro lado, se tomó la información de expresión génica de cada *cluster* de módulos identificados en el *t-SNE*, se calculó el promedio de expresión agrupando las muestras por estadio del ciclo de vida de *T. cruzi* y se graficaron *boxplots* evaluando la existencia de diferencias estadísticamente significativas de la expresión génica en cada estadio para cada *cluster* mediante *test* de Fisher.

## 3.10 Inferencia funcional de genes de función desconocida con *DARK* y *FoldSeek*

En primer lugar, se identificaron los 5 genes más conectados (*hubgenes*) de cada módulo que presentaba sobrerrepresentación de términos GO utilizando la función *get_hubs()* del paquete de R *CEMiTool* con el cual fue construida la red. A continuación, se identificaron aquellos anotados como proteínas hipotéticas y se procedió a intentar determinar su función mediante dos metodologías: comparación de perfiles HMM-HMM y alineamiento estructural.

Para la primera metodología se utilizó el *software DARK* (disponible en https://github.com/sradiouy/DARK). Este *software* permite visualizar e interrogar las anotaciones de proteínas producidas por estrategias de comparación HMM-HMM. Los modelos de Markov ocultos (HMM) son similares a los perfiles de secuencia simples, pero además de las frecuencias de aminoácidos en cada posición del alineamiento múltiple, contienen información sobre la frecuencia de inserciones y eliminaciones. Estas

comparaciones son muy sensibles y como resultado obtuvo la anotación para más de 2500 proteínas con función desconocida en tripanosomátidos.

Para la segunda metodología, se buscó en el servidor web de *AlphaFold* (*AlphaFold Reveals the Structure of the Protein Universe*, n.d.; Jumper et al., 2021) la estructura tridimensional predicha para cada una de las proteínas hipotéticas codificantes por los *hubgenes* identificados previamente. A continuación, se descargó el archivo *.pdb* que contiene la información sobre la estructura tridimensional de cada proteína y se utilizaron como entrada en el servidor web de *FoldSeek* (Kempen et al., 2023) que permite realizar búsquedas de homología mediante alineamiento estructural en numerosas bases de datos utilizando dos algoritmos denominados *TM-align* y *3Di/AA*.

*TM-align* es un método de alineamiento estructural de proteínas que utiliza las coordenadas de los átomos de carbono alfa (Cα) para comparar dos estructuras proteicas. El algoritmo se basa en la superposición de los dos conjuntos de coordenadas Cα, minimizando la distancia cuadrática media entre los átomos superpuestos. El algoritmo primero calcula una matriz de distancia entre los átomos Cα de las dos proteínas. A continuación, se utiliza el algoritmo de programación dinámica de Needleman-Wunsch para encontrar la superposición óptima de los dos conjuntos de coordenadas Cα. Una vez que se ha encontrado la superposición óptima, se calcula el *TM-score*, que es una medida de la similitud estructural entre las dos proteínas. El *TM-score* se basa en la idea de que las proteínas con estructuras similares tendrán una mayor superposición de sus átomos Cα. El *TM-score* varía entre 0 y 1, siendo 1 la máxima similitud estructural. A su vez, este algoritmo también permite la utilización de diferentes funciones de puntuación, como la puntuación de la estructura secundaria o la distancia euclidiana entre las estructuras superpuestas.

Por otro lado, *3Di/AA* es otro método de alineamiento estructural de proteínas que utiliza una representación de la estructura tridimensional de las proteínas en términos de interacciones terciarias entre residuos adyacentes. El algoritmo combina la información de la secuencia de aminoácidos y la información de la estructura tridimensional para lograr una alta sensibilidad y especificidad en la detección de homólogos estructurales. Esta representación se llama *3Di* y se basa en la distancia entre los centros virtuales de los residuos adyacentes. El algoritmo comienza por discretizar las estructuras proteicas en secuencias de 3Di utilizando un alfabeto de 20 estados que representa las diferentes interacciones terciarias entre los residuos. A continuación, se realiza una búsqueda de secuencias similares en una base de datos de proteínas utilizando un prefiltrado basado en

la búsqueda de *k-mers* y la alineación sin *gaps*. Después de la búsqueda de secuencias similares, se realiza una alineación estructural de las proteínas utilizando el algoritmo de alineación local de Smith-Waterman, utilizando una matriz de puntuación que combina las puntuaciones de sustitución de aminoácidos y *3Di*. La puntuación de sustitución de *3Di* se calcula a partir de una matriz de sustitución de *3Di* que se entrena a partir de las frecuencias de sustitución observadas en las alineaciones estructurales. Finalmente, se calcula un puntaje de similitud estructural entre las proteínas alineadas utilizando una combinación geométrica de la puntuación TM (que mide la similitud global de la estructura) y la puntuación LDDT (que mide la calidad local de la alineación).

# *4* Resultados y Discusión

## 4.1 Adquisición y procesamiento de datos

Se utilizaron 38 muestras de experimentos de *RNA-seq* para llevar a cabo este estudio, abarcando los 4 principales estadios del ciclo de vida de *T.* cruzi: epimastigotas, tripomastigotas metacíclicos, amastigotas y tripomastigotas celulares, provenientes de 3 estudios independientes (**tabla 1**).

**Tabla 1.** Información sobre los experimentos de *RNA-seq* utilizados para este estudio, incluyendo su identificador en NCBI o ENA, la cepa, el estadio del ciclo de vida de *T. cruzi* y el estudio del que proviene.

| Muestra | Id NCBI/ENA | Cepa | Estadio | Estudio |
|---|---|---|---|---|
| Epi_7dias_R1 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_7dias_R2 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_7dias_R3 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_14dias_R1 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_14dias_R2 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_14dias_R3 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_21dias_R1 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_21dias_R2 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_21dias_R3 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_28dias_R1 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_28dias_R2 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Epi_28dias_R3 | PRJNA915394 | Dm28c | Epimastigota | Smircich et al., 2023 |
| Meta_1_R1 | ERR3501949 | MHOM/CO/04/MG | Tripomastigota metacíclico | Cruz-Saavedra et al., 2020 |
| Meta_1_R2 | ERR3501950 | MHOM/CO/04/MG | Tripomastigota metacíclico | Cruz-Saavedra et al., 2020 |
| Meta_1_R3 | ERR3501951 | MHOM/CO/04/MG | Tripomastigota metacíclico | Cruz-Saavedra et al., 2020 |
| Meta_1_R4 | ERR3501952 | MHOM/CO/04/MG | Tripomastigota metacíclico | Cruz-Saavedra et al., 2020 |
| Amas_4h_R1 | SRR1346027 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_4h_R2 | SRR1346028 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_4h_R3 | SRR1346029 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_6h_R1 | SRR1346031 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_6h_R2 | SRR1346032 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_6h_R3 | SRR1346033 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_12h_R1 | SRR1346035 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_12h_R2 | SRR1346036 | Y-strain | Amastigota | Li et al., 2016 |

| | | | | |
|---|---|---|---|---|
| Amas_24h_R1 | SRR1346038 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_24h_R2 | SRR1346039 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_24h_R3 | SRR1346040 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_48h_R1 | SRR1346044 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_48h_R2 | SRR1346045 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_48h_R3 | SRR1346046 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_48h_R4 | SRR1346047 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_72h_R1 | SRR1346050 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_72h_R2 | SRR1346051 | Y-strain | Amastigota | Li et al., 2016 |
| Amas_72h_R3 | SRR1346052 | Y-strain | Amastigota | Li et al., 2016 |
| Trypo_R1 | SRR1346053 | Y-strain | Tripomastigota celular | Li et al., 2016 |
| Trypo_R2 | SRR1346054 | Y-strain | Tripomastigota celular | Li et al., 2016 |
| Trypo_R3 | SRR1346055 | Y-strain | Tripomastigota celular | Li et al., 2016 |
| Trypo_R4 | SRR1346056 | Y-strain | Tripomastigota celular | Li et al., 2016 |

Para cada una de las muestras de realizó un análisis de calidad de las lecturas con el *software FastQC*, donde se observó que todas presentaban buena calidad de secuencias determinada por su *phred-score* mayor a 28 (**figura 7**) y no presentaban sobrerrepresentación de secuencias (que bien podrían ser de ARNr o adaptadores), por lo que se procedió con el estudio utilizando los 38 transcriptomas.

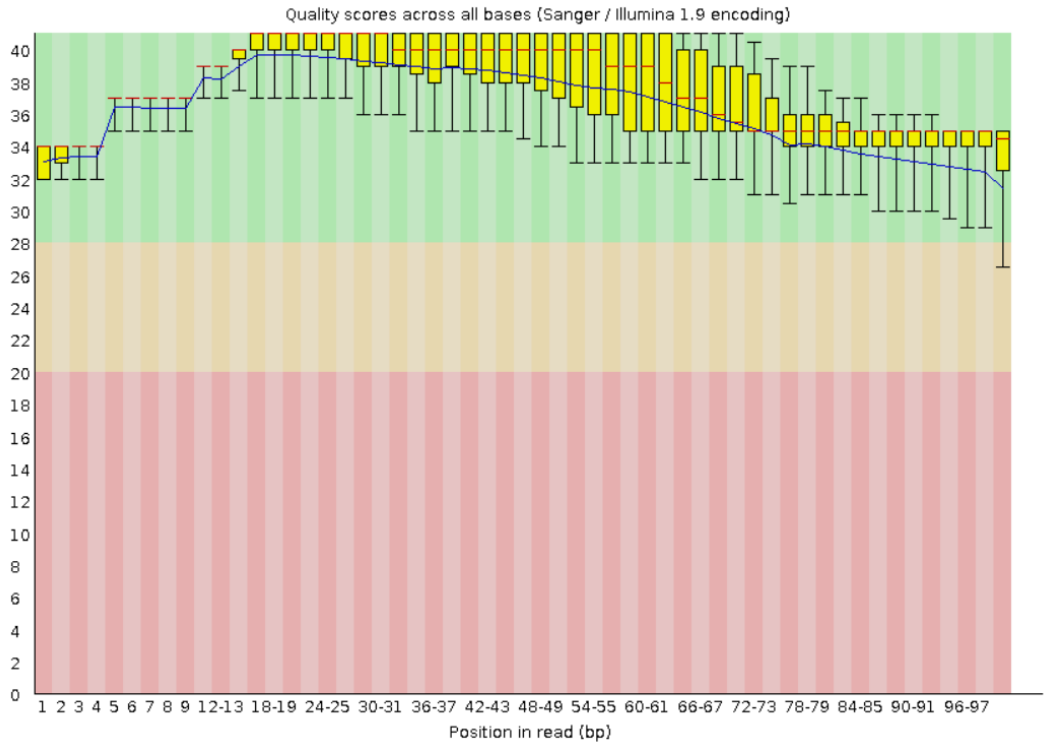



**Figura 7.** Ejemplo de reporte de calidad de bases de las lecturas obtenido mediante *FastQC*, correspondiente a la muestra Amas_6h_R1.

## 4.2 Evaluación y selección de estrategia para estimación de perfiles de expresión

Una de las desventajas de las metodologías de secuenciación de lecturas cortas, tal como las utilizadas en los datos de *RNA-seq* utilizados para este estudio, es la dificultad de asignar las lecturas correctamente a los genes durante el mapeo de estas al genoma de referencia (Deschamps-Francoeur et al., 2020), que se da cuando para una misma lectura existen varias regiones de alineamiento con la mejor puntuación en el genoma de referencia. Por consiguiente, también se ve afectada la posterior cuantificación de las lecturas asignadas a los genes para la estimación de los niveles de expresión génica. Este fenómeno es particularmente relevante en *T. cruzi*, dado que posee un genoma con muchas secuencias repetidas, pseudogenes y familias multigénicas muy expandidas. Por esta razón, se propuso en primer lugar optimizar y seleccionar la estrategia óptima para la estimación de los perfiles de expresión a partir de los datos de *RNA-seq*.

Para ello, se evaluó el desempeño de 3 de las herramientas más utilizadas: *kallisto*, *Bowtie2* y *STAR*. Estas 3 herramientas utilizan distintos métodos para asignar correctamente las lecturas con multimapeos; por un lado, *kallisto* utiliza el algoritmo *expectation maximization* (EM) (Dempster et al., 1977; B. Li et al., 2010) donde básicamente construye un transcriptoma en forma de grafo de de Bruijn a partir de las lecturas de secuenciación. Si las lecturas tienen secuencias idénticas, ya sea que se originen de diferentes transcriptos del mismo gen o de transcriptos de diferentes genes, estas se asignarán a la misma parte del grafo de Bruijn y por lo tanto formarán parte de la misma clase de equivalencia. Por último, el algoritmo muestrea las clases de equivalencia siguiendo una distribución multinomial y alimenta al algoritmo EM, lo que resulta en estimaciones de abundancia de transcriptos.

Por otro lado, por defecto *Bowtie2* simplemente asigna la lectura con multimapeos al azar a una de las regiones de mapeo, mientras que STAR ignora esa lectura y continúa el alineamiento con el resto de las lecturas.

Para evaluar el desempeño de cada software se diseñó un flujo de trabajo resumido en (**figura 8**) y detallado en Materiales y Métodos (3.2.1).

Brevemente, se simularon lecturas *paired-end* a partir del CDS de *T. cruzi* partiendo de una tabla de conteos ("tabla de conteos original") mediante 3 estrategias: simulando para cada gen las lecturas determinadas por la tabla de conteos original (**Estrategia 1 (E1)**), alterando de forma aleatoria las lecturas de algunos genes (5 réplicas) y simulando las lecturas (**Estrategia 2 (E2)**), y por último alterando toda la tabla de conteos original y simulando las lecturas de esta nueva tabla (**Estrategia 3 (E3)**). Las estrategias E2 y E3 se diseñaron para evaluar el desempeño de los programas a partir de diferentes tipos de datos iniciales, con el fin de disminuir posibles sesgos en la tabla obtenida para la estrategia E1.

Una vez simuladas las lecturas de cada una de las estrategias se procedió a realizar las estimaciones de los perfiles de expresión génica utilizando *kallisto*, *Bowtie2* y *STAR*. En el caso de *Bowtie2* (*Bwt2*) y *STAR*, se utilizó el *software featureCounts* (*fC*) para cuantificar las lecturas asignadas a los genes luego del mapeo, mientras que el algoritmo de *kallisto* asigna directamente conteos a los transcritos sin un paso de mapeo previo al genoma. De esta forma, se generaron 7 nuevas tablas de conteos para cada *software*: 1 para la estrategia E1, 5 para la estrategia E2 y 1 para la estrategia E3. Para el caso E2, se procedió a generar una única tabla promediando los resultados de las 5 réplicas, obteniendo entonces en última instancia 3 tablas por *software* (una por estrategia).

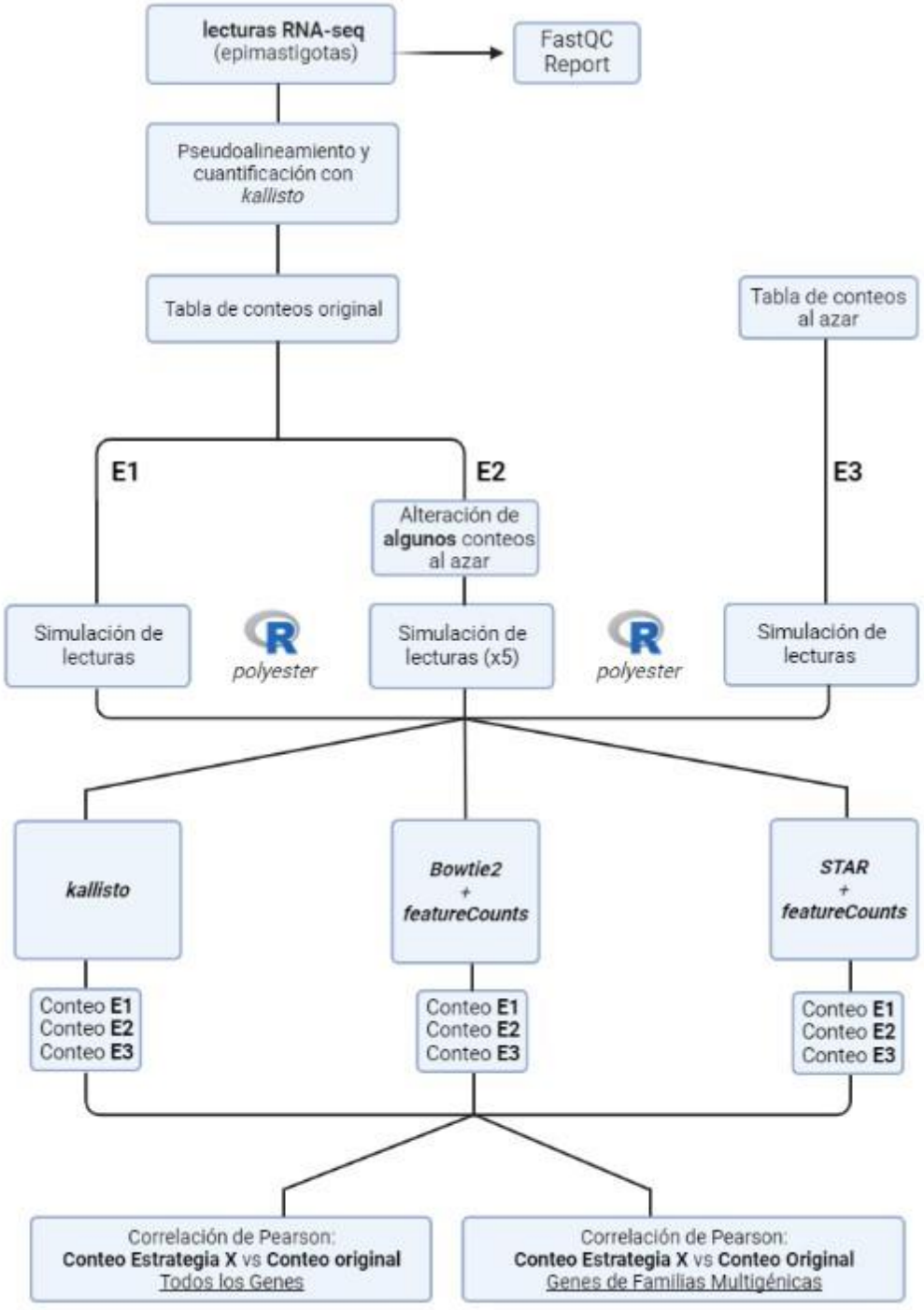



**Figura 8.** Flujo de trabajo para la optimización de la estimación de los perfiles de expresión génica y selección del *software* a utilizar entre *kallisto, Bowtie2* y *STAR*. Para el caso de la estrategia **E2** la tabla de conteos final se obtuvo a partir del cálculo del promedio de lecturas asignadas a los genes utilizando las 5 réplicas de lecturas simuladas.

Utilizando estas tablas se calculó un índice de correlación de Pearson entre cada una de ellas y la tabla de conteos desde la cual se partió para simular las lecturas, que permitió determinar qué tan bien se ajustaban los valores de conteo de cada estrategia a los valores originales. A su vez, dado que el fundamento de este análisis radicaba en la

dificultad de asignar las lecturas correctamente a los genes con muchos repetidos o genes pertenecientes a familias multigénicas, se procedió también a calcular un índice de correlación de Pearson en este subconjunto de genes (**tabla 2**).

**Tabla 2.** Correlaciones de Pearson entre las tablas de conteo de cada uno de los programa analizados para la cuantificación de la expresión génica y la tabla de conteos obtenida según la estrategia utilizada (**E1, E2** y **E3**).

| **Tabla Inicial** | ***software*** | **Todos los Genes** | | **Genes de Familias Multigénicas** | |
|---|---|---|---|---|---|
| | | **Coef. Pearson** | **p-valor** | **Coef. Pearson** | **p-valor** |
| **Estrategia 1** | *Kallisto* | 1.000 | < 2.2e-16 | 1.000 | < 2.2e-16 |
| | *Bwt2 + fC* | 0.980 | < 2.2e-16 | 0.997 | < 2.2e-16 |
| | *STAR + fC* | 0.892 | < 2.2e-16 | 0.991 | < 2.2e-16 |
| **Estrategia 2** | *Kallisto* | 1.000 | < 2.2e-16 | 1.000 | < 2.2e-16 |
| | *Bwt2 + fC* | 0.980 | < 2.2e-16 | 0.997 | < 2.2e-16 |
| | *STAR + fC* | 0.891 | < 2.2e-16 | 0.990 | < 2.2e-16 |
| **Estrategia 3** | *Kallisto* | 0.997 | < 2.2e-16 | 0.995 | < 2.2e-16 |
| | *Bwt2 + fC* | 0.987 | < 2.2e-16 | 0.978 | < 2.2e-16 |
| | *STAR + fC* | 0.922 | < 2.2e-16 | 0.889 | < 2.2e-16 |

Por un lado, se puede observar cómo todos los programas tienen un buen desempeño para cuantificar los perfiles de expresión génica en las 3 estrategias. Sin embargo, *kallisto* y *Bowtie2* parecen ser los que mejor realizan esta tarea, con coeficientes de Pearson > 0,95 en todas las estrategias, independientemente de si consideramos todos los genes o el subconjunto de genes de familias multigénicas. Por otro lado, resulta curioso que cuando se evalúa el desempeño en los genes de familias multigénicas todos los programas obtuvieron un coeficiente de Pearson mayor o igual al obtenido evaluando todos los genes. Estos resultados sugieren que el problema con que se enfrentan estos algoritmos a la hora de estimar la expresión de genes multicopia o repetidos está relativamente bien resuelto.

Dados estos resultados, se procedió a analizar en detalle los errores en la asignación de lecturas a los genes, identificando para cada *software* aquellos genes que tuvieron diferencias en esta asignación respecto a la tabla de conteo con la cual se simularon las lecturas (**figura 9**). Para esto se distinguieron dos tipos de comportamientos: aquellos genes que tenían un conteo de lecturas original distinto de 0 y el s*oftware* le asignó más o menos lecturas (**figura 9a**), y aquellos genes que tenían un conteo de lecturas original igual a 0 y el *software* le asignó lecturas (**figura 9b**). Se debió hacer esta distinción dado que

para el último caso no es posible calcular un porcentaje de cambio ya que implicaría dividir entre cero.

En el caso de los genes con el primer tipo de comportamiento, se calculó un porcentaje de diferencia de conteo entre *software* - Estrategia 1 vs tabla conteo original del tipo $\frac{\text{conteo}-\text{conteo original}}{\text{conteo original}} \text{x } 100$ (**figura 9a**), mientras que para los genes con el segundo tipo de comportamiento no se calculó en forma de porcentaje, sino en cantidad de lecturas incorrectamente asignadas según $|conteo - conteo\ original|$ (**figura 9b**).

Se puede observar varias cosas en este resultado; por un lado, *Bowtie2* parece ser el *software* que en menos cantidad de genes asigna incorrectamente las lecturas (1641), seguido por *kallisto* (2754) y luego *STAR* (5474)*,* donde este último asigna incorrectamente las lecturas a prácticamente la mitad de los genes. Sin embargo, se observa en los histogramas que por más que *Bowtie2* erre en menos cantidad de genes, lo hace por una mayor cantidad de lecturas, principalmente sobreestimando la expresión de muchos genes, mientras que *kallisto* tiende a tener errores mucho menores, concentrados alrededor del 0. Por otro lado, *STAR* parecería subestimar la expresión de una gran cantidad de genes, asignando a muchos de ellos más de un 50% de sus conteos originales, lo cual es razonable dado que descarta las lecturas de mapeo múltiple.

A su vez, cuando se evalúa la **figura 9b** se puede observar cómo en este caso *Bowite2* es el que peor desempeño tiene, asignando una gran cantidad de lecturas (hasta más de 9000) a genes cuyo conteo original era de 0, mientras que *kallisto* y *STAR* lo hacen en una proporción mucho menor.

Por otro lado, resultó de interés caracterizar aquellos genes en los cuales hubo dificultad al momento de asignar correctamente las lecturas independientemente del *software* y la estrategia utilizada. Se identificaron 130 genes, entre los cuales 46 estaban anotados como proteínas hipotéticas, 15 como pseudogenes, 12 como transialidasas y 14 como MASPs.

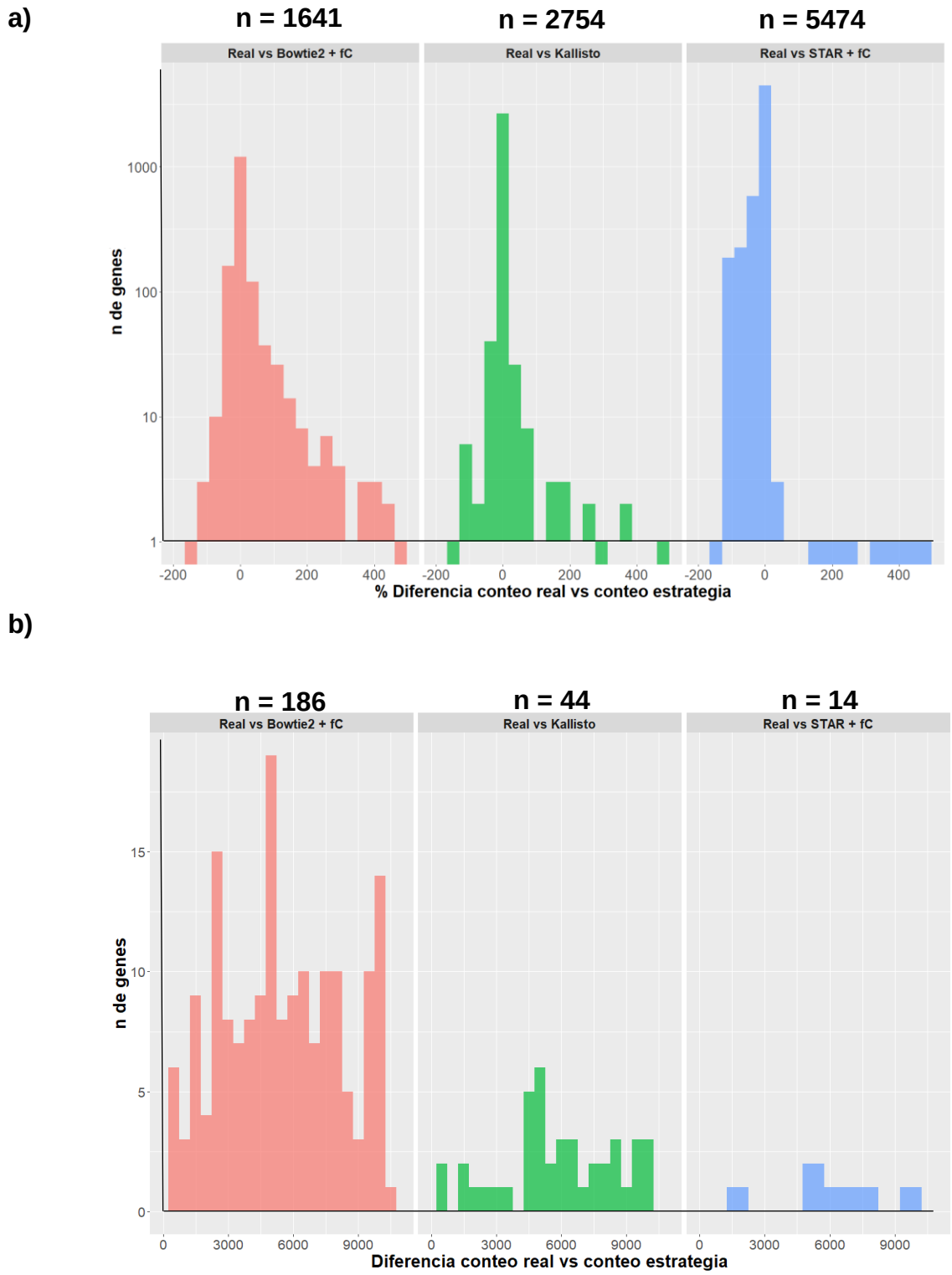



**Figura 9.** Histograma de distribución del porcentaje de diferencias de conteo para los genes con un conteo original distinto de 0 (**a**) y del valor absoluto de las diferencias de conteo para aquellos genes con conteos originales iguales a 0 (**b**).

Dados los resultados previamente descriptos es que seleccionamos el *software kallisto* para realizar las estimaciones de los perfiles de expresión génica de los 38 transcriptomas disponibles para este estudio, teniendo en cuenta que *kallisto* fue el *software*

con mejores correlaciones de Pearson y que globalmente estima mejor el número de lecturas proveniente de cada gen.

### 4.3 Estimación de perfiles de expresión génica

Una vez definido el método para la cuantificación de la expresión génica, *kallisto*, se procedió a utilizarlo para cada una de las 38 muestras. Este *software* devuelve una tabla con los conteos de lecturas asignadas a cada uno de los 10.338 genes de *T. cruzi*.

A partir de estas 38 tablas de conteos se generó una única tabla conteniendo los conteos de las lecturas asignadas a los 10.338 genes para las 38 muestras. Utilizando esta tabla se procedió a realizar una normalización utilizando la función *rlog()* del paquete de R *DESeq2* que transforma los conteos a una escala logarítmica donde atenúa la diferencia de las varianzas entre genes de altos y bajos conteos, y normaliza respecto al tamaño de las librerías de las diferentes muestras. Se decidió optar por este tipo de normalización ya que es sugerida para el posterior uso de técnicas de aprendizaje automático tales como *clustering*, una de las que fueron utilizadas para este estudio (Love et al., 2014).

Utilizando estos datos se determinó mediante *boxplots* la distribución de los conteos por muestra antes y después de la normalización (**figura 10**), donde se observó que luego de la normalización todas las muestras presentaron una media de expresión de los genes similar a grandes rasgos. Sin embargo, se observó que las muestras de tripomastigotas metacíclicos (Meta_1_R1, Meta_1_R2, Meta_1_R3, Meta_1_R4) presentaban una distribución de los conteos considerablemente más disperso respecto al resto de las muestras y una media ligeramente inferior.

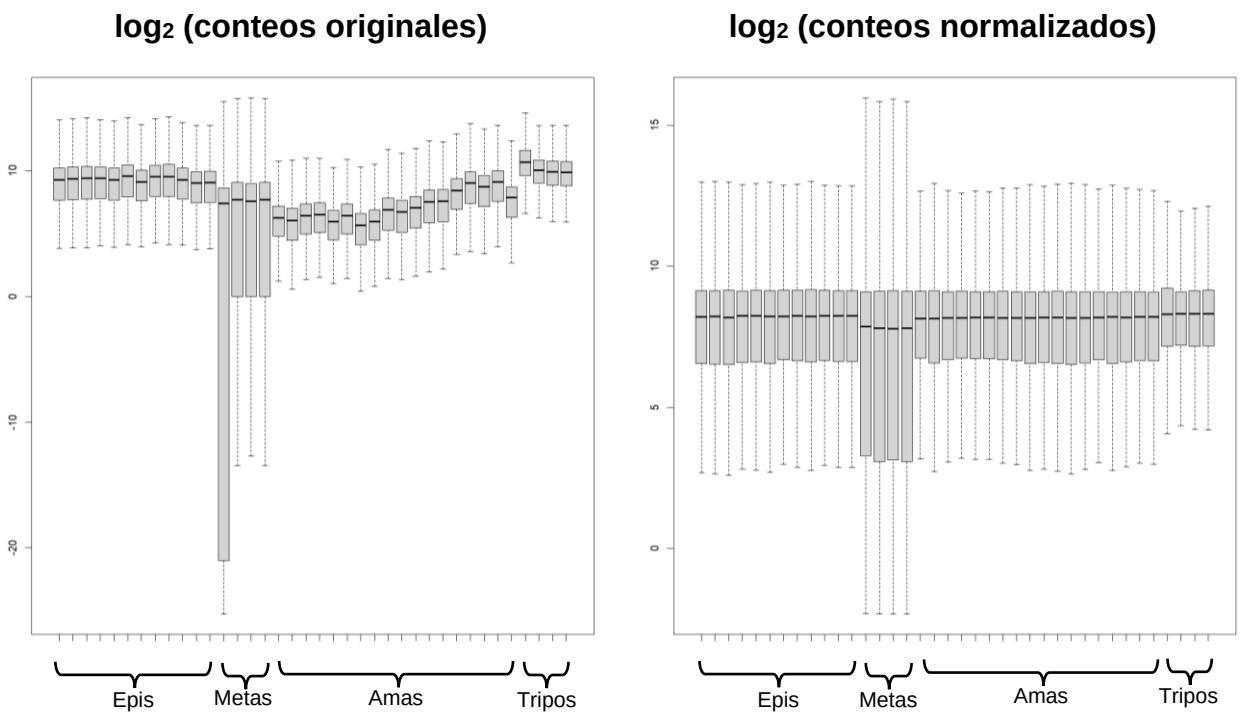



**Figura 10.** Distribuciones de conteos de lecturas de los genes obtenido utilizando *kallisto* para cada una de las 38 muestras antes (**izquierda**) y después (**derecha**) de la normalización mediante *rlog* del paquete de R *DESeq2*.

## 4.4 Selección de estrategia para la identificación de grupos de genes co-expresados

Una vez estimados y normalizados los niveles de expresión génica, se propuso la identificación de grupos de genes co-expresados durante el ciclo de vida de *T. cruzi*. Para ello se planteó la evaluación de diferentes programas disponibles de uso libre.

Actualmente existen numerosos programas para realizar esta tarea, cada uno con sus particularidades a nivel de las aproximaciones algorítmicas que utiliza. Por ejemplo, los que utilizan métodos de *clustering* basados en particiones como *K-means*, donde los genes se agrupan en un número fijo de *clusters*, asignando aleatoriamente los genes a los *clusters* y luego ajustando las asignaciones iterativamente para maximizar una medida de similitud dentro del *cluster* y maximizar una medida de disimilitud entre *clusters* (Hartigan & Wong, 1979); programa que utilizan este tipo de aproximaciones (con sus variaciones) son el paquete de R *coseq* (Godichon-Baggioni et al., 2019) y *clust* (Abu-Jamous & Kelly, 2018). Por otro lado, están aquellos que utilizan métodos de *clustering* basados en redes: estos métodos utilizan grafos para identificar *clusters* de genes co-expresados. Las redes se construyen a partir de los perfiles de expresión génica, donde los nodos representan genes

y las aristas representan correlaciones entre genes. Los *clusters* se identifican como módulos en la red, que son grupos densamente conectados de nodos. Ejemplos de programas de *clustering* basados en redes son los paquetes de R *WGCNA* (Langfelder & Horvath, 2008) y *CEMiTool* (Russo et al., 2018).

En este contexto, se procedió a la construcción de redes de co-expresión (para el caso de *WGCNA* y *CEMiTool*) y clusterización (para el caso de *coseq* y *clust*) de los genes utilizando los 38 transcriptomas disponibles para este estudio. A su vez, se replicó la construcción de una red de co-expresión génica generada para *Trypanosoma brucei* por (Mwangi et al., 2021), con el objetivo de validar nuestra estrategia comparándola con la utilizada en un artículo ya publicado**.** Los resultados de estos análisis de resumen en la **tabla 3**.

**Tabla 3.** Resumen de los algoritmos utilizados para la construcción de redes de co-expresión génica (*WGCNA* y *CEMiTool*) o clusterización directa (*coseq* y *clust*).

| Estrategia | # Genes | # Módulos | # Genes módulo más grande | # Genes módulo más chico | # Genes promedio por módulo |
|---|---|---|---|---|---|
| ***CEMiTool*** | 10338 | 14 | 2646 | 74 | 691 |
| ***WGCNA*** | 10338 | 12 | 3198 | 108 | 792 |
| ***WGCNA (T. brucei)*** | 7390 | 27 | 732 | 61 | 269 |
| ***coseq*** | 10338 | 15 | 6103 | 18 | 232 |
| ***clust*** | 4047 | 10 | 2279 | 82 | 202 |

Por un lado, se observó cómo *clust* es el único algoritmo que realiza un filtrado de genes para el análisis, descartando genes de baja expresión o con poca varianza, conservando un total de 4047 genes de los 10338 que presenta *T. cruzi*, obteniendo un total de 10 módulos de genes co-expresados. Esto no resulta particularmente de interés para este estudio dado que se planteó utilizar todos los genes independientemente de si son o no codificantes para proteínas funcionales, su varianza entre las diferentes muestras, etc. A su vez, más de la mitad de sus genes están clusterizados en un único módulo de 2279 genes.

Por otro lado, *coseq* no realiza un filtrado de genes, generó 15 módulos de genes co-expresados, aunque nuevamente más de la mitad de los genes están clusterizados en un único módulo de 6103 genes.

Por último, tanto *WGCNA* como *CEMiTool* no realizan un filtrado de genes, generando 12 y 13 módulos de genes co-expresados, respectivamente. En el caso de *WGCNA,* el módulo de mayor tamaño contenía 3198 genes, mientras que el más grande obtenido por *CEMiTool* contenía 2646 genes.

Dados estos resultados, se planteó implementar alguna estrategia que permitiera cuantificar la calidad de los módulos generados por cada uno de los algoritmos utilizados. Para ello se realizó una búsqueda bibliográfica y se observó la existencia de dos índices que evalúan la consistencia funcional de los módulos, evaluado en términos de los términos GO de los genes anotados de cada módulo, ellos son el *Biological Homogeneity Index* (BHI) (Datta & Datta, 2006) y *Wang Index* (B. Li et al., 2015). Por un lado, BHI toma las anotaciones funcionales (términos GO) de los genes de cada módulo y compara gen contra gen si tienen en común alguno de sus términos GO; basta con que esos dos genes compartan un único término GO para que se compute que tienen funciones compartidas. La desventaja de este algoritmo es que no tiene en cuenta la estructura jerárquica del DAG de las ontologías génicas ni la similaridad semántica de los términos GO. Por ejemplo, si se evalúa la similaridad utilizando los términos GO de la ontología MF de los genes *Adh4* and *Ldb3* da un BHI que está condicionado por los únicos dos términos GO que comparten de un total de 9 términos GO con los que están anotados. Esto sugiere que las funciones moleculares de *ADh4* y *Ldb3* no son similares dado que tienen un BHI bajo. Sin embargo, si se observan detalladamente los términos GO de estos dos genes, se encuentra que “metal ion binding” and “zinc ion binding” (términos asociados a *Ldb3*), y “protein binding” y “heme binding” (términos asociados a *ADh4*), son semánticamente similares dado que se encuentran muy cercanos en el DAG de la ontología. Lo mismo sucede con el término “electron carrier activity” de *Ldb3*, que es hijo del término “oxidoreductase activity” del gen *ADh4*. Por lo tanto, la semántica de estos términos debería ser similar. A su vez, los términos “alcohol dehydrogenase activity”, “alcohol dehydrogenase activity, zinc-dependent”, “alcohol dehydrogenase activity”, “metal ion-independent” y “electron-transferring-flavoprotein dehydrogenase activity” son hijos lejanos del término “oxidoreductase activity” en el DAG, por lo que también tendrían similaridades semánticas

con el término “oxidoreductase activity”. De todo lo anterior se desprende que las funciones moleculares de *ADh4* y *Ldb3* podrían ser similares (B. Li et al., 2015).

En este contexto es que surge el *Wang Index*, que incorpora además de la similaridad funcional de los términos GO, su similaridad semántica, evaluando las distancias de los términos en el DAG y asignándoles de esta forma un peso para el cálculo del índice de consistencia funcional del módulo.

Se calculó para cada uno de los módulos de cada uno de los 4 algoritmos utilizados para la clusterización de genes con patrones de expresión similares el BHI y *Wang Index,* con el fin de evaluar el desempeño de cada uno de los algoritmos, y se graficó la distribución de ambos índices para cada algoritmo mediante un *boxplot* (**figura 11**).

En el caso del índice BHI, se observa que *coseq* parecería ser el algoritmo con mejor calidad en términos de similaridad funcional de cada uno de sus módulos, seguido por *WGCNA*, *CEMiTool* y *clust*. De todas formas, se realizaron *tests* de Wilcoxon para evaluar la existencia de diferencias significativas entre los algoritmos (sin considerar el caso de *T. brucei*) y no se observaron diferencias entre ninguno de ellos

Cuando se evalúa en términos de similaridad funcional y semántica (*Wang Index*), *WGCNA* parecería ser el algoritmo con mejor desempeño, seguido por *CEMiTool*, *clust* y *coseq*. En este caso, cuando se realizaron los *tests* de Wilcoxon se encontraron diferencias estadísticamente significativas únicamente entre *CEMiTool* y *coseq* (*p-value = 0,007*) y entre *WGCNA* y *coseq* (*p-value* = 0,009).

En cuanto a los índices obtenidos para los módulos de *T. brucei*, utilizado a modo de control como una red de co-expresión génica de un organismo muy emparentado filogenéticamente y que ha sido publicada en una revista científica arbitrada, se observa que tanto para BHI como para *Wang* los valores se encuentran más dispersos que los obtenidos para *T. cruzi*, estando las medias en el mismo rango.

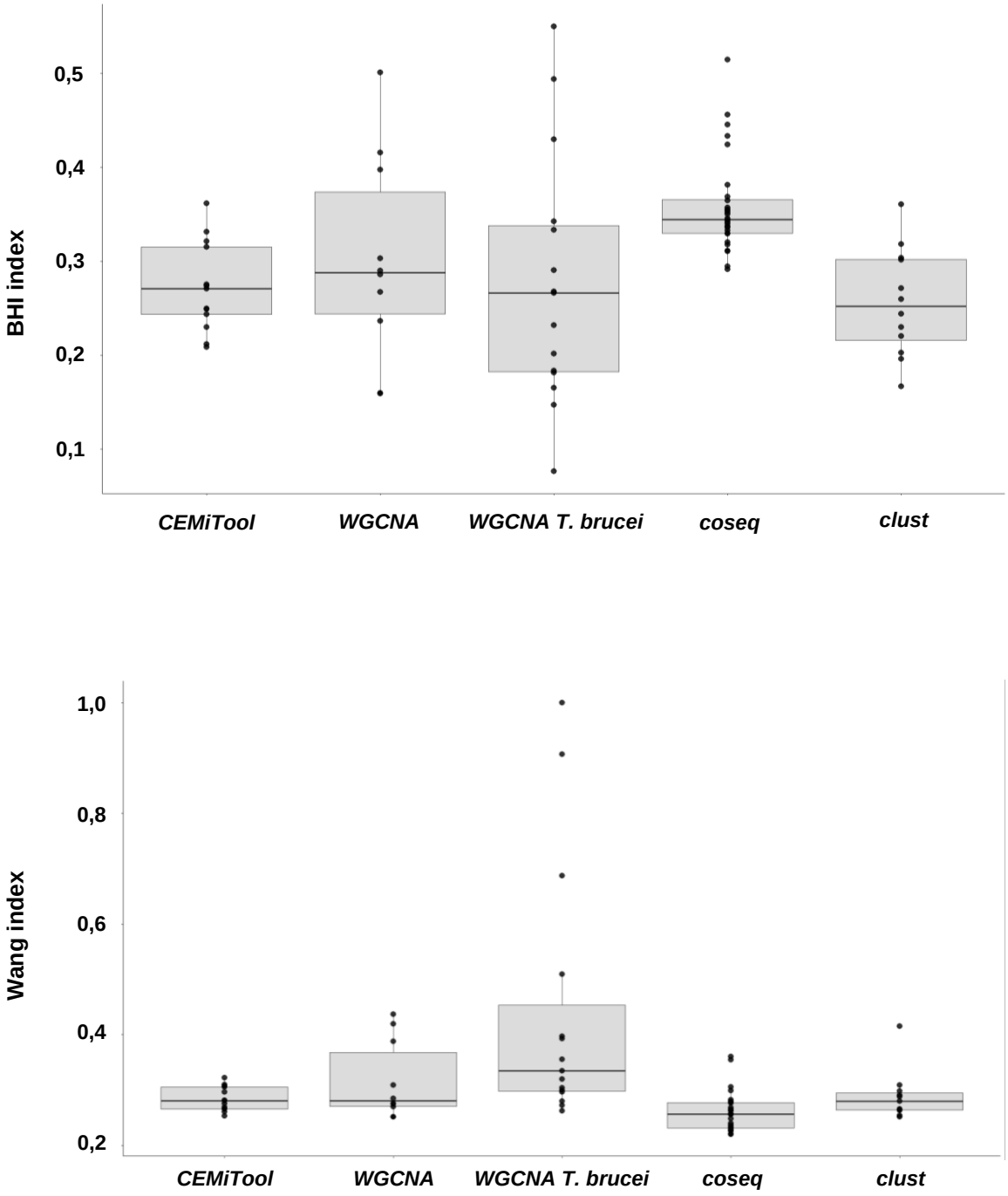



**Figura 11.** *Biological Homogeneity Index* **(arriba)** y *Wang Index* (**abajo**) para los diferentes programas utilizados para la identificación de módulos de genes co-expresados

Realizando búsquedas bibliográficas asociadas a este tema, existe un estudio realizado por (Cheng et al., 2020), donde realizan la comparación de algunos algoritmos utilizados para este tipo de análisis de co-expresión génica, y concluyen que, considerando la disponibilidad de documentación, la simplicidad de uso, el tiempo de ejecución y los recursos computacionales *CEMiTool* supera a otras herramientas, entre ellas *coseq* y

*WGCNA*. En particular, frente a *WGCNA* el estudio reporta una optimización del *software CEMiTool* al momento de la selección del parámetro β. Por otro lado, en nuestro caso no se encontraron diferencias estadísticamente significativas entre *WGCNA* y *CEMiTool* para los índices de BHI y Wang (**figura 11**).

Dado lo discutido anteriormente se decidió optar por la herramienta *CEMiTool* para la construcción de la red de co-expresión génica final que fue utilizada para el resto de este estudio.

## 4.5 Construcción de la red de co-expresión génica utilizando *CEMiTool*

Se utilizaron los 10.338 genes de *T. cruzi* y las 38 muestras de experimentos de *RNA-seq* detallados previamente para la construcción de una matriz de adyacencia de correlación de expresión de los genes entre las diferentes muestras para la posterior construcción de la red de co-expresión génica. Previo a su construcción se determinó el *soft-thresholding power* (β) que se utilizó para ajustar la red a una topología *scale-free* eliminando el ruido producido por las correlaciones débiles en la matriz de adyacencia. Se sabe que este tipo de topología es la que prevalece en las redes biológicas (Albert, 2005; Barabási & Oltvai, 2004) y se caracteriza por una distribución desigual de conexiones entre los nodos (genes) de la red: se esperan pocos genes (denominados *hubgenes*) que tienen muchas conexiones y muchos genes con pocas conexiones. Para ello se evaluaron diferentes β (1-20) (**figura 12**). Para cada uno se ajustó un modelo lineal entre el logaritmo de la conectividad, k, y el logaritmo de la frecuencia de nodos con conectividad k, p(k), obteniendo un *scale-free topology index,* $R^2$, para cada uno de los β. Por último, se escogió el menor β que obtuvo un $R^2 > 0,9$ con el fin de obtener una red lo más conectada posible (evaluada en términos de la conectividad media de los nodos de la red (*mean connectivity*)) pero manteniendo una topología *scale-free*. En la **figura 12b** se observa cómo disminuye la conectividad media de la red a medida que β aumenta. Para la construcción de la red de co-expresión génica se seleccionó un β = 7, siendo este el menor β que obtuvo un R2 > 0,9 (**figura 12a**). El ajuste lineal entre $\log_{10}(k)$ y $\log_{10}(p(k))$ se ilustra en la **figura 12c**.

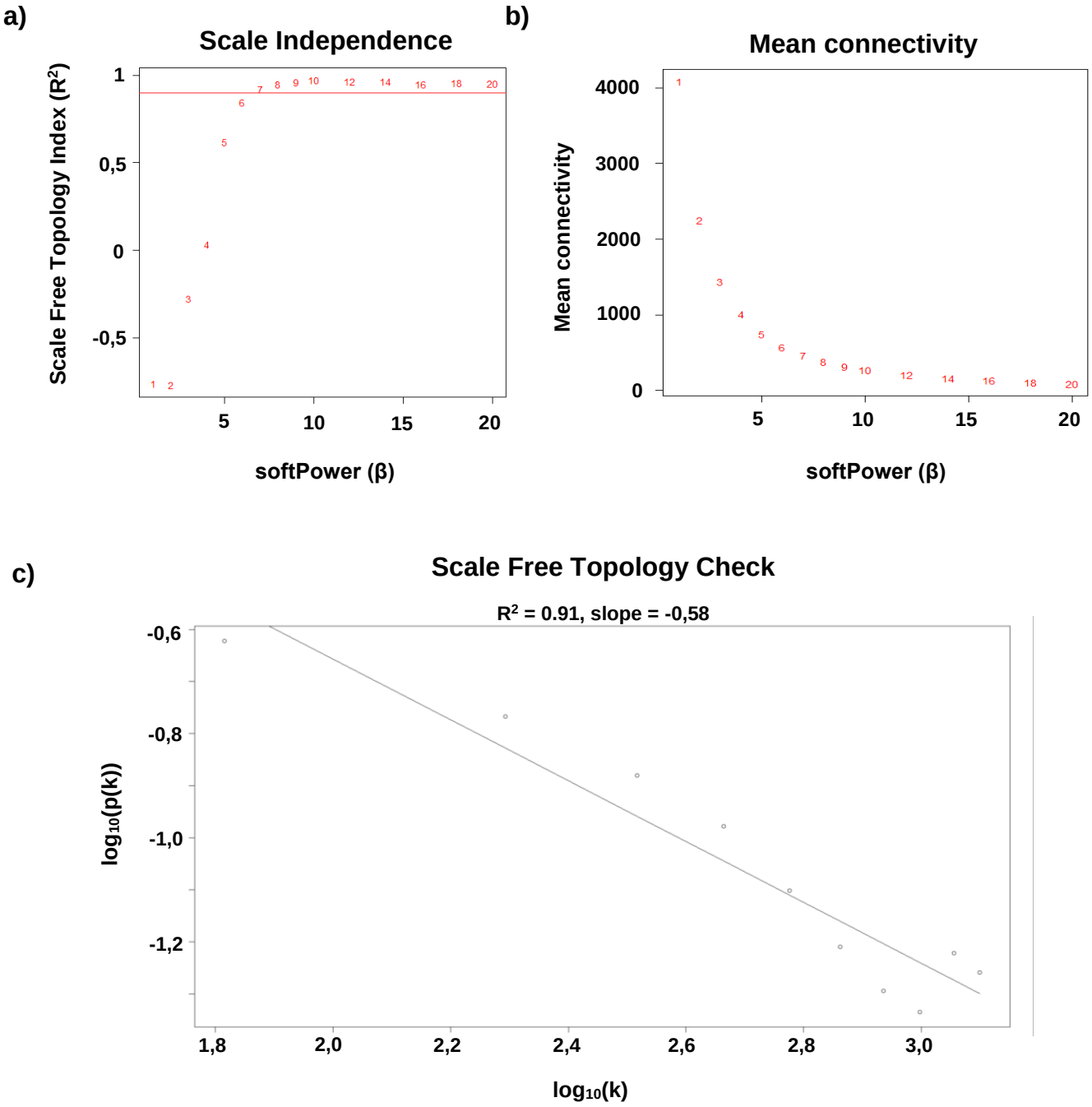



**Figura 12.** Selección del parámetro β para la construcción de la red de co-expresión génica con topología *scale-free*. Para ello se seleccionó el menor β donde el *scale-free topology index* ($R^2$) fuese > 0,9. **a)** índice de ajuste de la red a una topología *scale-free* ($R^2$) en función de los diferentes β evaluados (se marca en rojo $R^2$ = 0,9), **b)** conectividad media de la red en función de los diferentes β evaluados y **c)** ajuste lineal entre el logaritmo de la cantidad de nodos con conectividad k y el logaritmo de la conectividad k para la red obtenida con un β = 7.

Una vez construida la red, se identificaron 14 módulos de genes co-expresados mediante *clustering* jerárquico utilizando para ello el algoritmo *dynamic tree cut*. De los 14 módulos, uno de ellos contenía 10 genes que no pudieron ser asignados a ninguno de los

otros 13 módulos ya que no se detectó ningún tipo de asociación en sus niveles de co-expresión, por lo que se descartó de los siguientes análisis. De los 13 módulos conservados, el que más genes presentaba fue el módulo M1 con 2646 genes y el que menos genes presentaba fue el módulo M13 con 74 genes (**figura 13**).

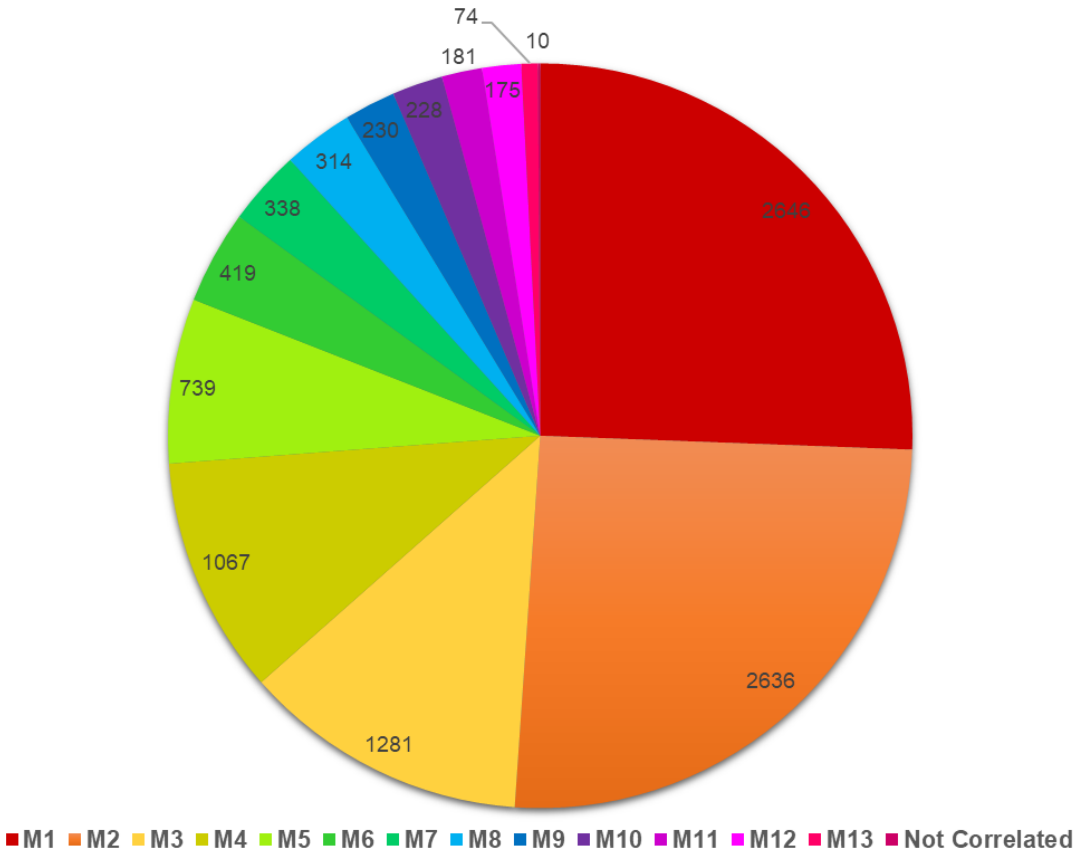



**Figura 13.** Número de genes identificados en cada uno de los 14 módulos obtenidos. El módulo "Not Correlated" contenía 10 genes no pudieron ser asignados a ninguno de los otros módulos por lo que fue excluido de los siguientes análisis.

## 4.6 Análisis funcional de la red de co-expresión génica seleccionada

Una vez construida la red de co-expresión génica se procedió a realizar un análisis de enriquecimiento funcional evaluando la sobrerrepresentación de términos GO. Se discutirán los términos asociados a la ontología BP en cada uno de los módulos. Respecto a las ontologías MF y CC, el análisis inicial (no mostrado) no arrojó resultados en principio interesantes. Esto último puede ser esperable dado que para determinados procesos biológicos las proteínas que actúan pueden cumplir funciones moleculares muy variadas y

encontrarse en diferentes compartimentos celulares. De los 13 módulos evaluados, se encontró que 10 estaban enriquecidos en procesos biológicos asociados al metabolismo, patogénesis, replicación del ADN, regulación del citoesqueleto y movimiento celular, entre otros (**tabla 4**).

**Tabla 4.** Módulos con términos GO sobrerrepresentados y sus términos GO sobrerrepresentados más significativos (*p-value < 0,05* y *p-value ajustado < 0,1*)

| Módulo | Términos GO sobrerrepresentados | p-valor ajustado Benjamini-Hochberg (<0.1) |
|---|---|---|
| **M1** | *metabolic process, biological process, organic substance metabolic process* | 2.10E-3, 2.10E-3, 2.26E-3 |
| **M4** | *obsolete pathogenesis* | 5.77E-14 |
| **M5** | *obsolete pathogenesis* | 2.14E-41 |
| **M7** | *oxoacid metabolic process, pyruvate metabolic process, ATP metabolic process, ribonucleotide metabolic process* | 2.86E-4, 2.86E-4, 1.95E-3, 2.32E-3 |
| **M8** | *chromosome organization, DNA packaging, DNA conformation change, nucleosome assembly, cellular component organization* | 6.30E-5, 8.88E-4, 2.65E-3, 1.64E-2 |
| **M9** | *regulation of supramolecular fiber organization, regulation of cytoskeleton organization, cilium or flagellum dependent cell motility* | 1.24E-2, 1.24E-2, 1.24E-2, 1.24E-2 |
| **M10** | *regulation of mRNA stability, regulation of mRNA catabolic process, regulation of mRNA metabolic process* | 7.07E-2, 7.07E-2, 7.07E-2 |
| **M11** | *cell-matrix adhesion, cell-substrate adhesion* | 1.92E-2, 1.92E-2 |
| **M12** | *movement of cell or subcellular component, microtubule-based process* | 2.83E-2, 2.83E-2 |
| **M13** | *obsolete pathogenesis* | 9.38E-02 |

Resulta interesante observar cómo los principales fenómenos que caracterizan los cambios biológicos que sufre *T. cruzi* durante los cambios de estadio y hospedero en su ciclo de vida están representados en este análisis, principalmente los asociados al metabolismo, la infectividad, replicación, cambios morfológicos y la regulación de los ARNm como mecanismo de regulación de la expresión génica.

Cabe destacar que en el caso del módulo M1, el módulo más grande de la red con 2646 genes, se halló una sobrerrepresentación de términos GO muy generales en la ontología. Dado el comportamiento que se observa al graficar los niveles de expresión de los genes para cada módulo (**figura suplementaria 1**), en el caso del módulo M1 se podría hipotetizar que este comportamiento en los perfiles de co-expresión se debe a que los genes que pertenecen a ese módulo entran dentro de los denominados genes *housekeeping* (Joshi et al., 2022), aunque no es posible confirmar esto sin realizar un estudio exhaustivo de los genes que integran este módulo.

## 4.7 Identificación de grupos de genes co-expresados de expresión estadio-específica

Con el fin de identificar grupos de genes que muestren patrones de coexpresión en estadios específicos del ciclo de vida del *T. cruzi*, se llevó a cabo un análisis de correlación entre cada uno de los estadios (epimastigota, tripomastigota metacíclico, amastigota y y tripomastigota celular) y los *Module Eigengenes* (ME) de cada módulo (**figura 14**). El ME de un módulo es utilizado como un *proxy* de la expresión génica global de un módulo, y se calcula como el primer vector propio (*eigenvector*) de la matriz de correlación de expresión de los genes dentro del módulo, que representa la dirección de la mayor variabilidad de esta matriz. Se utilizó la matriz de expresión génica para calcular los MEs de cada módulo enriquecido funcionalmente, se generó una matriz con las muestras y los estadios donde cada celda toma un valor de 1 o 0 dependiendo de si la muestra corresponde o no a cada estadio del ciclo de vida del parásito, respectivamente. Por último, se calculó un índice de correlación de puntobiserial (Gupta, 1960), que es un tipo de coeficiente de correlación utilizado para medir la relación entre una variable dicotómica (con dos posibles valores) y una variable continua. La correlación de puntobiserial se basa en la suposición de que la variable dicotómica representa un grupo o condición experimental, mientras que la variable continua representa una medida de interés, en este caso de expresión génica, y se calcula a partir de una fórmula que involucra la covarianza entre las dos variables y el desvío estándar de la variable continua.

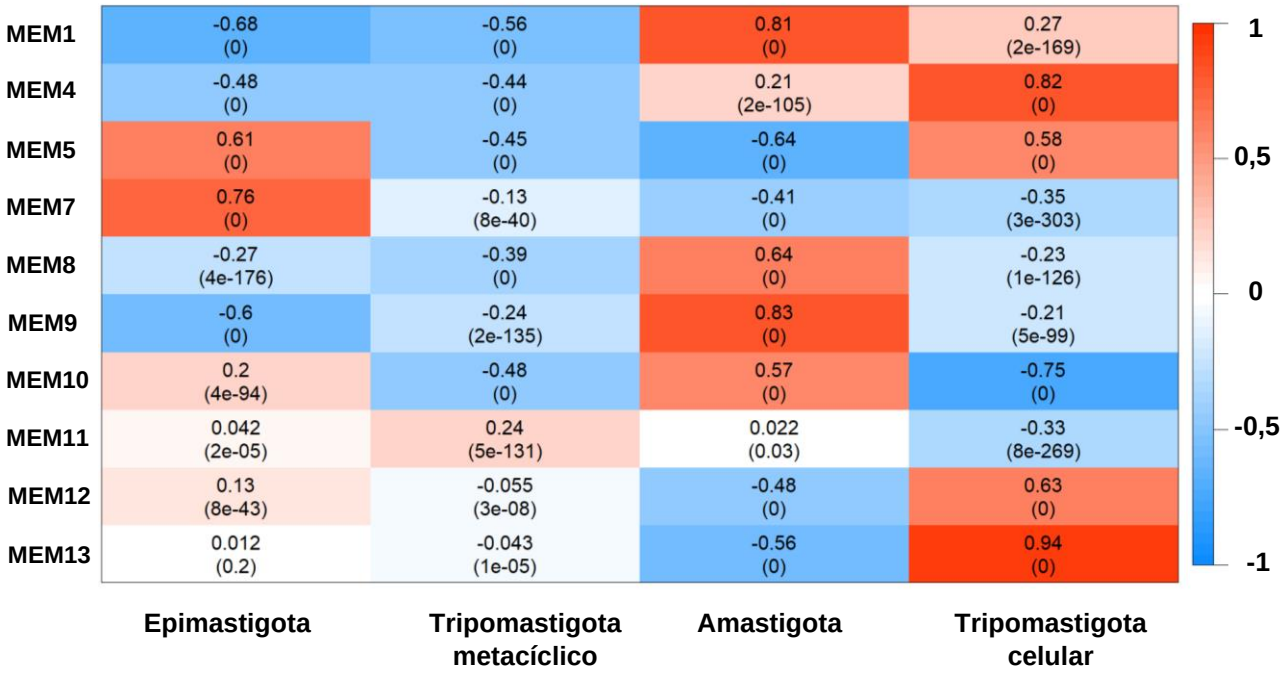



**Figura 14.** *Heatmap* de correlaciones de puntobiserial entre *Module Eigengenes* (ME) de módulos con sobrerrepresentación de términos GO y cada uno de los estadios de *T. cruzi*. Cada celda contiene el índice de correlación y su p-valor asociado debajo entre paréntesis.

Se puede observar la presencia de algunos módulos cuyos genes están sobre o sub-expresados en alguno de los estadios del parásito. En particular, en el caso de epimastigotas se observa una sobreexpresión de genes pertenecientes a módulos asociados funcionalmente a procesos metabólicos diversos (M7), lo que concuerda con las características de la biología parasitaria de este organismo alojado en el tracto digestivo de su hospedero triatomino. Llama la atención el término patogénesis (M5), aunque esto puede deberse a que el parásito comience con la expresión de genes a nivel de ARNm que lo preparen para la metaciclogénesis y adquirir sus características infectivas de forma completa. Es curioso que el módulo M8, asociado a procesos de replicación del ADN, esté subexpresado en este estadio, dado que se caracteriza por ser uno de los dos únicos estadios en donde el parásito se multiplica intensamente mediante fisión binaria.

Por otro lado, los estadios amastigota y tripomastigota celular son los más interesantes que surgen de este análisis: en el caso de amastigotas, se observa una sobreexpresión de genes pertenecientes a los módulos M1, M8, M9 y M10, asociados a procesos metabólicos, replicación del ADN, regulación del citoesqueleto y flagelo, y regulación de ARNm, respectivamente. Este estadio se caracteriza por ser un estadio

donde el parásito se encuentra dentro de las células de su hospedero mamífero, ser metabólica y replicativamente activo, y sufrir cambios morfológicos importantes, principalmente a nivel del “redondeo” del parásito y la pérdida del flagelo. En cuanto a los tripomastigotas celulares, estadio caracterizado por su alto nivel de infectividad dentro del hospedero mamífero, presenta una sobreexpresión de genes asociados a los módulos M4, M5, M12 y M13. M4, M5 y M13 son módulos enriquecidos en genes asociados a procesos vinculados a la patogénesis, que incluyen numerosas proteínas de superficie como transialidasas, mucinas y MASPs, mientras que M12 está enriquecido en genes asociados al movimiento celular, lo que nuevamente concuerda con la biología parasitaria, donde en la diferenciación amastigota-tripomastigota celular estos últimos desarrollan nuevamente el flagelo para movilizarse en el torrente sanguíneo y lograr así infectar nuevas células o ser ingerido por el vector triatomino, cerrando así el ciclo de vida del parásito.

En cuanto al estadio tripomastigota metacíclico, prácticamente todos los módulos parecen estar subexpresados en él, lo cual es razonable considerando lo observado en la **figura 10** que indica que los datos no pudieron ser adecuadamente normalizados con el resto. Aunque esto no impide su uso para la construcción de las redes (ya que la asignación de genes a módulos es robusta a los sesgos de *batch*), el hecho de que los datos provengan de experimentos diferentes, sí es una limitante para este último análisis.

### 4.8 Motivos compartidos en UTRs

En los últimos años, herramientas bioinformáticas permitieron la identificación de probablemente la mayoría de las proteínas de unión al ARN (RBPs) de tripanosomátidos y numerosos elementos de secuencia principalmente involucrados en el procesamiento del ARN (Benz et al., 2005; Duhagon et al., 2001; Pastro et al., 2013; M. Smith et al., 2008). A su vez, varios estudios demostraron la presencia de elementos ricos en U en las regiones 3'UTR de los ARNm de tripanosomátidos (Araújo & Teixeira, 2011b; Haile & Papadopoulou, 2007), y el papel funcional de las secuencias repetidas de CA en las regiones 3'UTR de *T. cruzi* como una señal para la modulación de la expresión génica a lo largo del ciclo de vida del *T. cruzi* (Pastro et al., 2013).

En este contexto, se propuso caracterizar posibles motivos a nivel de secuencia que regulen la expresión conjunta de los genes pertenecientes a cada módulo utilizando los programa *XSTREME* y *FIMO* de *MEME-Suite* (Bailey et al., 2009). Para ello, en primer

lugar, se determinaron las regiones 3'UTR de los genes de cada módulo utilizando el *software UTRme* (Radío et al., 2018).

Una vez definidas las 3'UTRs se procedió a la identificación de motivos *de-novo* para cada set de 3'UTRs de cada módulo utilizando *XSTREME* (**tabla suplementaria 2**). Este software reporta el número de sitios donde encuentra los motivos en los genes del módulo, así como un estadístico que permite inferir la sobrerepresentación con respecto al *background*, pero no el total de genes en los cuales se encuentra el motivo. Por lo tanto, se procedió a utilizar el *software FIMO* utilizando como entrada los motivos obtenidos por *XSTREME* con el fin de obtener las proporciones de esos motivos tanto en los módulos como en los *backgrounds*, pudiendo establecer de esa forma qué tan específicos eran los motivos a cada uno de los módulos (**tabla suplementaria 3**). A su vez, estos programas reportan proteínas de unión a esos motivos reportadas en otros estudios.

Los resultados muestran una serie de motivos significativos, muchos de los cuales han sido previamente reportados en otros organismos como funcionales e incluso con proteínas de unión descritas. Además, varios de ellos presentan una alta especificidad, siendo muy poco frecuentes fuera de los genes del módulo estudiado. Por lo tanto, podrían jugar un papel clave como moduladores de la regulación de los transcriptos. Sería interesante evaluar su funcionalidad en trabajos posteriores por metodologías de laboratorio húmedo.

Dado que, como mencionamos antes, existen en la literatura motivos funcionales descritos en *T. cruzi*, se buscó si estos estaban presentes y eran específicos de los módulos construidos en este trabajo. Por un lado, estudiamos la presencia de un motivo descrito por el grupo del Dr. De Gaudenzi (Sabalette et al., 2019) (5'-CAACUGCUCCACUCGCACACCCACCGACACGCUCAUGACGACGGCCCUGU-3') denominado *surface glycoprotein motif* (SGPm) que une la proteína TcUBP1, y que está altamente conservado en regiones 3'UTR de ARNm codificantes para proteínas de superficie como las transialidasas, entre otros ARNm (Noé et al., 2008c). Esta RBP puede formar un complejo con la proteína de unión a poli(A) en ciertos ARNm y regular el transcripto de una manera estadio específica (De Gaudenzi et al., 2011; Z. H. Li et al., 2012). En este contexto, se propuso la búsqueda de este motivo en la **tabla suplementaria 3**, encontrando su presencia en 2 módulos: M5 y M13. El resultado es coherente ya que estos módulos están enriquecidos en funciones asociadas a la patogénesis, y en particular en genes codificantes para proteínas de superficie, entre ellas las transialidasas.

Por otro lado, TcUBP1 también reconoce elementos estructurales conservados que también han sido descriptos en (Noé et al., 2008a), así como TcRBP3 (Noé et al., 2008a) y DRBD3/PTB1 de *T. brucei* (Estévez, 2008). En este sentido, se realizó la búsqueda de los motivos estructurales reconocidos por estas 3 proteínas utilizando el *software cmsearch* de *Infernal* en las regiones 3'UTR para cada uno de los módulos, aunque no se obtuvo una sobrerrepresentación significativa para ninguno de ellos (resultados no mostrados).

### 4.9 Estudio de uso diferencial de codones en los módulos

Sesenta y un tripletes de bases alternativos (codones) en el ADN y el ARNm codifican para veinte aminoácidos diferentes, de forma que varios aminoácidos están codificados por dos o hasta seis codones distintos o "sinónimos".

Se ha reconocido hace varias décadas las diferencias en el uso de codones entre genes, es decir, diferencias en la frecuencia de ocurrencia de codones sinónimos. En la última década se ha evidenciado a partir de estudios que incluso una única sustitución sinónima puede tener impactos significativos en los niveles de expresión génica, plegamiento de las proteínas codificantes y su función celular (Angov, 2011; Jeacock et al., 2018; Plotkin & Kudla, 2011).

El uso diferencial de codones influye en la tasa de traducción de los ARNm, el procesamiento co-traduccional del péptido naciente facilitado por el enlentecimiento de la traducción mediado por el uso codones "raros", la estabilidad y decaimiento de los ARNm, y más (Angov, 2011; Jeacock et al., 2018; Plotkin & Kudla, 2011).

A su vez, el sesgo en el uso de codones influye también la tasa de traducción de los ARNm, fenómeno que estaría ligado a los niveles de ARNt presentes en la célula; codones menos frecuentes están asociados a ARNt menos frecuentes en las poblaciones citoplasmáticas. La hipótesis establece que en una región que presente codones "raros", el ribosoma deberá esperar más tiempo a que llegue el ARNt cargado correcto, produciendo una "pausa ribosomal", o como mínimo un enlentecimiento de su movimiento (Lesnik et al., 2000). Este fenómeno podría estar asociado a la optimización de los tiempos de síntesis proteica con el fin de permitir el correcto plegamiento de la proteína sintetizada (Thanaraj & Argos, 1996), así como a la estabilización de los ARNm (Collart & Weiss, 2020; Presnyak et al., 2015; Radhakrishnan et al., 2016). En este caso, se ha reportado en levaduras la asociación entre el uso no óptimo de codones y el enriquecimiento de Dhh1, proteína

asociada al decaimiento de ARNm, y que ésta se une preferencialmente a ribosomas enlentecidos en el transcripto (Radhakrishnan et al., 2016).

Por estas razones, se planteó evaluar la existencia de un posible uso diferencial de codones sinónimos en los módulos enriquecidos funcionalmente. Para ello, se realizó un *t-Distributed Stochastic Neighbor Embedding (t-SNE)* utilizando información de las frecuencias promedio de los diferentes codones en los genes de cada uno de los módulos de la red (**figura 15**). Se puede observar que los módulos se agrupan en 3 *clusters* distintos, donde se interpreta que cada uno de ellos presenta un uso de codones distinto mientras que los módulos que los integran presentan un perfil de uso de codones similar entre sí.

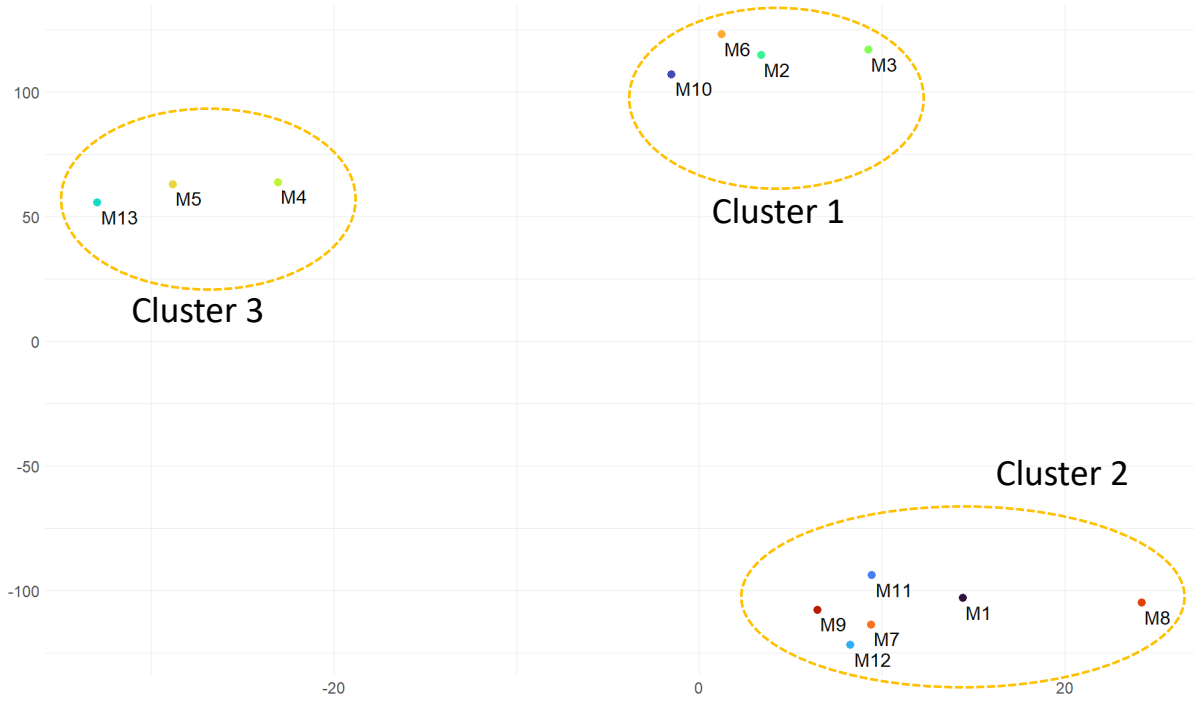



**Figura 15.** *t-SNE* del uso de codones de cada uno de los módulos de la red de co-expresión génica

Los *clusters* identificados se nombraron Cluster 1, Cluster 2 y Cluster 3. Por un lado, el Cluster 1 está integrado por los módulos M2, M3, M6 y M10. El Cluster 2 está integrado por los módulos M1, M7, M8, M9, M11 y M12, que están enriquecidos funcionalmente en procesos asociados al metabolismo, regulación de la replicación del ADN y a la regulación del citoesqueleto y movimiento celular. Por último, el Cluster 3 está integrado por los módulos M4, M5 y M13, todos ellos enriquecidos funcionalmente en términos asociados a la patogénesis. Este último caso resulta de particular interés, dado que los genes que integran estos módulos son en su mayoría proteínas de superficie, donde parecería que hay un uso de codones particular que podría aportar a su expresión característica en el ciclo de vida.

Dado este resultado, se propuso indagar si existía algún tipo de relación entre el uso de codones de cada *cluster* y la expresión de los genes que los integran. Para ello, se tomó la información de expresión génica de los genes de cada *cluster* de módulos identificados previamente en el *t-SNE*, se calculó el promedio de expresión agrupando las muestras por estadio del ciclo de vida de *T. cruzi* y se graficaron *boxplots* evaluando la expresión génica en cada estadio para cada *cluster* (**figura 16**).

Por un lado, se observan diferencias estadísticamente significativas evaluadas mediante *test* de Wilcoxon para la mayoría de los estadios de cada *cluster*. Sin embargo, resulta de particular interés el Cluster 3, integrado por módulos enriquecidos en genes de proteínas de superficie: se observa cómo hay un aumento de su expresión en el estadio tripomastigota celular, caracterizado por ser un estadio infectivo que expresa muchos de estos genes para lograr infectar las células.

Dado que el tamaño de los módulos que integran cada *cluster* es muy variable, es importante destacar la influencia que pueden llegar a tener los módulos grandes respecto a los pequeños en cuanto a que probablemente determinar en mayor proporción el uso de codones del *cluster*. Por esta razón, sería interesante estudiar el uso de codones de forma individual para cada módulo e intentar correlacionarlo con la expresión del módulo en cada estadio, con el fin de determinar si hay una asociación entre el uso de codones y la expresión génica módulo a módulo. A su vez, para esto sería interesante también determinar experimentalmente mediante metodologías de secuenciación masiva los niveles de ARNt en cada uno de los estadios, e incluir esos resultados con la asociación uso de codones – expresión, de forma de establecer fehacientemente el impacto del uso de codones en la expresión génica de *T. cruzi*. Actualmente existen diversos protocolos optimizados para esto, tales como YAMAT-seq (Shigematsu et al., 2017), TGIRT-seq (Xu et al., 2019) o más recientemente el desarrollo de un protocolo para secuenciación específica de ARNt y ARNs pequeños mediante la tecnología Oxford Nanopore (Lucas et al., 2023). Una vez obtenido el perfil de expresión de ARNt, se podría integrar con los resultados obtenidos y comparar el perfil de uso de codones entre los módulos de cada *cluster* y la expresión de los genes en cada estadio.

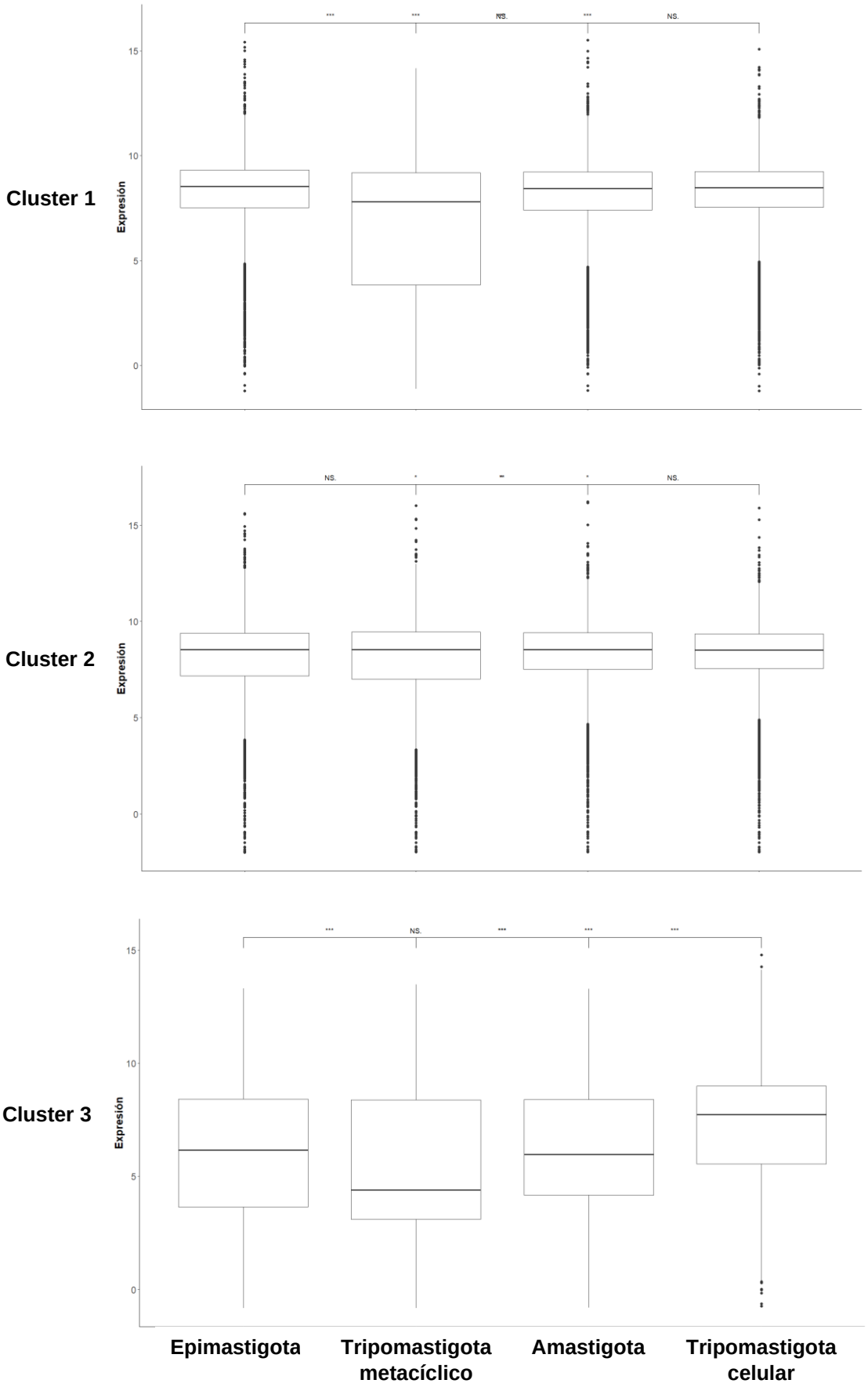



**Figura 16.** *Boxplots* de la distribución de los niveles de expresión génica para cada estadio de *T. cruzi* de los módulos pertenecientes a cada uno de los *clusters* identificados.

## 4.10 Identificación y análisis funcional de *hubgenes*

Los *hubgenes* son genes que tienen una alta conectividad con otros en las redes de co-expresión génica. Debido a esta alta conectividad, hay evidencia de que los *hubgenes* son importantes para la organización y el funcionamiento de la red biológica y pueden desempeñar un papel crucial en los procesos biológicos subyacentes a cada uno de los módulos (Barabási & Oltvai, 2004; Langfelder & Horvath, 2008).

En este contexto, se determinaron los 5 genes más conectados (Top 5 *hubgenes*) de cada módulo que presentaba términos GOs sobrerepresentados (10 módulos) y se procedió a obtener su anotación del TriTrypDB (**tabla 5**). De esta forma, se identificaron 50 *hubgenes* de los cuales 16 estaban anotados como proteínas hipotéticas. De esos 16, 3 estaban anotados como pseudogenes de proteínas hipotéticas conservadas, mientras que 9 estaban anotados como proteínas hipotéticas conservadas. Este resultado presenta particular relevancia ya que está evidenciando que se desconoce la función de varios genes *hub*, los cuales suelen jugar roles clave en los procesos biológicos subyacentes a los módulos. Nuestra aproximación bioinformática permitió su identificación, posibilitando nuevos estudios funcionales a nivel informático y a su vez priorizar su estudio a nivel experimental. Es particularmente interesante destacar que 4 de los 5 *hubgenes* del módulo M1 estaban anotados como pseudogenes conservados. Los pseudogenes son secuencias de ADN que pueden haber surgido a partir de eventos de duplicación o degradación de genes de copia única. En este sentido, estudios recientes han demostrado que el 94% de los pseudogenes de *T. cruzi* son transcripcionalmente activos (Abrahim et al., 2022), y que los pseudogenes podrían haber desarrollado nuevas funciones como reguladores de la expresión génica, servir como sustratos para la evolución de nuevos genes o incluso ser reactivados para desempeñar funciones biológicas específicas (Pink & Carter, 2013). Sin embargo, aún es necesario profundizar en el estudio de los pseudogenes en *T. cruzi* para determinar sus posibles funciones biológicas.

**Tabla 5.** Top 5 *hubgenes* de los 10 módulos con sobrerrepresentación de términos GO con su anotación funcional obtenida de TriTrypDB y los términos GO sobrerrepresentados del módulo.

| Módulo | Top | Top 5 hubgenes | Proteínas codificantes | GOs sobrerrepresentados |
|---|---|---|---|---|
| M1 | 1 | TcCLB.509205.100 | surface protease GP63, putative | metabolic process, biological process, organic substance metabolic process |
| | 2 | TcCLB.508207.10 | glycine dehydrogenase (pseudogene), putative | |
| | 3 | TcCLB.510849.30 | hypothetical protein, conserved (pseudogene) | |
| | 4 | TcCLB.508835.10 | hypothetical protein, conserved (pseudogene) | |
| | 5 | TcCLB.506113.80 | hypothetical protein, conserved (pseudogene) | |
| M4 | 1 | TcCLB.510553.30 | Mucin-associated surface protein (MASP), subgroup S101 | obsolete pathogenesis |
| | 2 | TcCLB.455171.9 | trans-sialidase, putative | |
| | 3 | TcCLB.507905.39 | Mucin-associated surface protein (MASP) | |
| | 4 | TcCLB.509081.20 | Mucin-associated surface protein (MASP), subgroup S122 | |
| | 5 | TcCLB.511553.30 | Mucin-associated surface protein (MASP), subgroup S122 | |
| M5 | 1 | TcCLB.510275.272 | hypothetical protein | obsolete pathogenesis |
| | 2 | TcCLB.511173.470 | trans-sialidase, Group V, putative | |
| | 3 | TcCLB.509233.10 | hypothetical protein | |
| | 4 | TcCLB.506599.420 | Mucin-associated surface protein (MASP), subgroup S117 | |
| | 5 | TcCLB.508541.90 | Mucin-associated surface protein (MASP), subgroup S078 | |
| M7 | 1 | TcCLB.508387.20 | methylthioadenosine phosphorylase, putative | oxoacid metabolic process, pyruvate metabolic process, ATP metabolic process, ribonucleotide metabolic process |
| | 2 | TcCLB.508771.50 | hypothetical protein, conserved | |
| | 3 | TcCLB.503487.70 | Inhibitor of apoptosis-promoting Bax1, putative | |
| | 4 | TcCLB.506367.30 | 1,2-Dihydroxy-3-keto-5-methylthiopentene dioxygenase, putative | |
| | 5 | TcCLB.509157.220 | hypothetical protein, conserved | |
| M8 | 1 | TcCLB.506563.10 | pumilio/PUF RNA binding protein 9, putative | chromosome organization, DNA packaging, DNA conformation change, nucleosome assembly, cellular component organization |
| | 2 | TcCLB.511417.70 | Histone-lysine N-methyltransferase, H3 lysine-76 specific | |
| | 3 | TcCLB.511871.30 | 2OG-Fe(II) oxygenase superfamily, putative | |
| | 4 | TcCLB.506795.44 | Bifunctional NAD(P)H-hydrate repair enzyme | |

| | | | | |
|---|---|---|---|---|
| | 5 | TcCLB.507129.30 | C-14 sterol reductase, putative | |
| M9 | 1 | TcCLB.508731.40 | hypothetical protein, conserved | regulation of supramolecular fiber organization, regulation of cytoskeleton organization, cilium or flagellum dependent cell motility |
| | 2 | TcCLB.510347.29 | hypothetical protein, conserved | |
| | 3 | TcCLB.508955.10 | syntaxin, putative | |
| | 4 | TcCLB.511215.119 | Paraflagellar rod protein 2 | |
| | 5 | TcCLB.506927.20 | hypothetical protein, conserved | |
| M10 | 1 | TcCLB.503811.45 | hypothetical protein, conserved | regulation of mRNA stability, regulation of mRNA catabolic process, regulation of mRNA metabolic process |
| | 2 | TcCLB.508169.90 | eukaryotic translation initiation factor 3 subunit l | |
| | 3 | TcCLB.506661.40 | zinc finger domain, LSD1 subclass, putative | |
| | 4 | TcCLB.509455.100 | exonuclease, putative | |
| | 5 | TcCLB.506285.40 | Mucin-associated surface protein (MASP) (pseudogene) | |
| M11 | 1 | TcCLB.506287.209 | DNA ligase, putative | cell-matrix adhesion, cell-substrate adhesion |
| | 2 | TcCLB.511623.20 | hypothetical protein, conserved | |
| | 3 | TcCLB.511729.60 | MORN repeat-containing protein 1 | |
| | 4 | TcCLB.506735.10 | mitochondrial processing peptidase alpha subunit, putative | |
| | 5 | TcCLB.510759.120 | rieske iron-sulfur protein, mitochondrial precursor, putative | |
| M12 | 1 | TcCLB.511127.90 | Domain of unknown function (DUF4586), putative | movement of cell or subcellular component, microtubule-based process |
| | 2 | TcCLB.503903.70 | hypothetical protein, conserved | |
| | 3 | TcCLB.453917.9 | hypothetical protein, conserved | |
| | 4 | TcCLB.510285.20 | PQQ-like domain/WD domain, G-beta repeat/Utp21 specific WD40 associated putative domain containing protein, putative | |
| | 5 | TcCLB.511693.70 | hypothetical protein, conserved | |
| M13 | 1 | TcCLB.506683.110 | trans-sialidase, Group VIII, putative | obsolete pathogenesis |
| | 2 | TcCLB.508165.170 | Mucin-associated surface protein (MASP), subgroup S030 | |
| | 3 | TcCLB.510025.30 | Mucin-associated surface protein (MASP) (pseudogene) | |
| | 4 | TcCLB.508165.400 | mucin TcMUCII, putative | |
| | 5 | TcCLB.510105.310 | Mucin-associated surface protein (MASP), subgroup S085 | |

Realizando un análisis más profundo de los *hubgenes* para los cuales sí hay anotación funcional definida, podemos destacar que, en la mayoría de los casos, existe una concordancia entre los términos GO sobrerrepresentados en los módulos y la función de

estos *hubgenes*, reforzando el hecho de que estos genes suelen tener roles relevantes en los módulos que integran.

En el caso de los módulos M4, M5 y M13, asociados con términos GO vinculados a la patogénesis, todos los *hubgenes* anotados son proteínas de superficie de tipo transialidasa, mucina o MASP, que son precisamente importantes factores de virulencia de este parásito.

Al analizar las funciones de los *hubgenes* del módulo M7, asociado a diversos procesos metabólicos, se observa que TcCLB.508387.20 es una enzima responsable al metabolismo de poliaminas, mientras que TcCLB.506367.30 es una enzima perteneciente a la familia de las oxidorreductasas que participa en el metabolismo de metionina.

El módulo M8 está relacionado a procesos biológicos vinculados a la replicación del ADN. En cuanto a sus *hubgenes* con anotación funcional se encuentra TcCLB.511417.70, una metiltransferasa de histonas, una de las principales enzimas de modificación epigenética de la cromatina que determinan su nivel de compactación. Por otro lado, TcCLB.511871.30 codifica para una proteína de la superfamilia de las 2OG-Fe(II) oxigenasas, que, entre sus diversas funciones, está reportada la de reparación, regulación y modificación del ADN en varios organismos (Jia et al., 2017). A su vez, TcCLB.506795.44, una "bifunctional NAD(P)H-hydrate repair enzyme" está reportada como una proteína de reparación y como parte de sistema de *proofreading* de nucleótidos de nicotinamida (Marbaix et al., 2011).

En el caso del módulo M9, asociado a regulación del citoesqueleto y movilidad del flagelo, sólo dos de sus top 5 *hubgenes* tienen función asignada: TcCLB.508955.10 y TcCLB.511215.119. El primero codifica para una sintaxina (Yu Hsuan Teng et al., 2001), vinculada a los procesos de exocitosis y endocitosis, mientras que el segundo codifica para una proteína Rod2 paraflagelar implicada en la movilidad del flagelo (Zhang et al., 2021).

En lo que respecta al módulo M10 vinculado a procesos de regulación de los ARNm, los *hubgenes* TcCLB.508169.90 y TcCLB.509455.100 codifican para la subunidad 1 del factor de iniciación de la traducción eIF3 y para una exonucleasa, enzima que cataliza la degradación de ARN desde el extremo 5' y 3', respectivamente.

En cuanto al módulo M1, es importante destacar que los términos GO sobrerrepresentados son términos que se encuentran muy altos en la estructura jerárquica

del DAG, por lo que resulta difícil asignar un vínculo estrecho entre los *hubgenes* anotados y los procesos biológicos asociados al módulo.

Por último, en cuanto a los módulos M11 y M12 y algunos de los *hubgenes* del resto, no se logró encontrar una asociación obvia entre sus funciones y los términos GO sobrerrepresentados en estos módulos.

## 4.11 Inferencia funcional para *hubgenes* de función desconocida

Los grandes avances tecnológicos para la secuenciación de genomas en los últimos años han facilitado enormemente la caracterización de los genes y sus productos proteicos a gran escala, en contraposición a los métodos tradicionales a nivel experimental donde se estudia la función de un gen/proteína a la vez. Sin embargo, una gran cantidad de genes permanecen sin ser caracterizados, a cuyos productos proteicos se les denominan “proteínas hipotéticas”. Básicamente una proteína hipotética se define como un marco de lectura que ha sido predicho como codificante por las herramientas computacionales de anotación automática (que evalúan precisamente la presencia de marcos de lectura, sitios de *splicing*, motivos de unión de factores de transcripción, contenido GC y más), pero que no existe evidencia de su función. A su vez, existen proteínas hipotéticas homólogas que están conservadas en diferentes linajes evolutivos que son denominadas “proteínas hipotéticas conservadas”.

Los métodos más utilizados en la actualidad para la anotación de genes y proteínas de función desconocida se basan en el alineamiento y búsqueda de similaridad a nivel de secuencia primaria, con el objetivo de encontrar secuencias homólogas de genes o proteínas ya anotadas en las bases de datos que permitan inferir posibles funciones moleculares y celulares. Sin embargo, en *T. cruzi* muchas proteínas no pueden ser anotadas mediante estas metodologías posiblemente resultado de la divergencia temprana a nivel evolutivo de la especie lo que conllevó a una gran divergencia a nivel de secuencia.

Actualmente, se han desarrollado numerosas herramientas bioinformáticas para la anotación funcional de genes que utilizan otras aproximaciones, tales como homología estructural (Kempen et al., 2023; Krissinel & Henrick, 2005), anotación basada en ontologías (Törönen et al., 2018), métodos de aprendizaje automático (Kulmanov et al., 2018), y más.

Dado que un alto porcentaje de los genes de *T. cruzi* no posee anotación funcional (aproximadamente el 40%), se propuso en esta sección la anotación de ciertos genes de interés, definidos a partir de su conectividad dentro de la red. Para ello, en primer lugar, se caracterizó por módulo la cantidad de genes sin anotación funcional (**tabla 6**), que van desde un 18% en el módulo M13 hasta un 51% en el módulo M6.

**Tabla 6** Resumen de cantidad de genes de función desconocida para cada uno de los módulos identificados en la red de co-expresión génica.

| Módulo | # de genes | # genes no anotados | % de genes no anotados |
|---|---|---|---|
| **M1** | 2646 | 967 | 37 |
| **M2** | 2636 | 1093 | 41 |
| **M3** | 1281 | 468 | 37 |
| **M4** | 1067 | 231 | 22 |
| **M5** | 739 | 169 | 23 |
| **M6** | 419 | 213 | 51 |
| **M7** | 338 | 141 | 42 |
| **M8** | 314 | 108 | 34 |
| **M9** | 230 | 84 | 37 |
| **M10** | 228 | 92 | 40 |
| **M11** | 181 | 70 | 39 |
| **M12** | 175 | 81 | 46 |
| **M13** | 74 | 13 | 18 |
| **Not.Correlated** | 10 | 3 | 30 |

A partir de la **tabla 5** se obtuvo una lista reducida de *hubgenes* de función desconocida, por lo que se procedió a intentar determinar su función *in-silico* mediante dos metodologías, por un lado, mediante comparación de perfiles HMM-HMM y por otro mediante la búsqueda de homología por alineamiento estructural, y comparar las funciones predichas por ambos algoritmos.

Para la primera metodología se utilizó el software *DARK*. Este software permite visualizar e interrogar las anotaciones de proteínas producidas por estrategias de comparación HMM-HMM. Los modelos de Markov ocultos (HMM) son similares a los perfiles de secuencia simples, pero además de las frecuencias de aminoácidos en cada posición del alineamiento múltiple, contienen información sobre la frecuencia de inserciones

y eliminaciones. Estas comparaciones son muy sensibles y como resultado obtuvo la anotación para más de 2500 proteínas con función desconocida en tripanosomátidos.

Por otro lado, para la búsqueda de homología mediante alineamiento estructural se obtuvo en primer lugar la estructura tridimensional de cada una de las proteínas codificadas por los *hubgenes*, disponibles en la base de datos de *AlphaFold*. Luego, se utilizó el servidor web *FoldSeek* para realizar búsquedas de homología mediante alineamiento estructural de las proteínas utilizando los dos algoritmos que dispone *FoldSeek*: *TM-align* y *3Di/AA*. *TM-align* es un método de alineamiento estructural que utiliza las coordenadas de los átomos de carbono alfa para comparar dos estructuras proteicas (ver Materiales y Métodos). El software calcula el *TM-score*, que varía entre 0 y 1 y es una medida de la similitud estructural entre las dos proteínas, siendo 1 la máxima similitud estructural. Por otro lado, *3Di/AA* utiliza una representación de la estructura tridimensional de las proteínas en términos de interacciones terciarias entre residuos adyacentes (ver sección 3.10 de Materiales y Métodos). El algoritmo calcula un puntaje de similitud estructural entre las proteínas alineadas.

Los resultados obtenidos se resumen en la **tabla 7**, y se detallan en la **tabla suplementaria 4**. Por un lado, se puede observar cómo de los 16 *hubgenes* no anotados, fue posible inferir la posible función de 10 de ellos mediante al menos una de las dos metodologías. A su vez, se observa una correlación entre la función inferida y los términos GO enriquecidos de los módulos a los que pertenecen estos *hubgenes*: TcCLB.510275.272 y TcCLB.509233.10, pertenecientes a un módulo enriquecido en el término *obsolete pathogenesis* fueron anotados por *DARK* como proteínas de superficie cruciales para las propiedades infectivas del parásito. Por otro lado, TcCLB.509157.220 fue anotado como una proteína transmembrana y una proteína exportadora de ácidos grasos, siendo este gen perteneciente a un módulo asociado a diversos procesos metabólicos. A su vez, del módulo M9 asociado a la regulación del citoesqueleto y movilidad del flagelo, TcCLB.508731.40 fue anotada tanto por *DARK* como por *FoldSeek* como una proteína mitocondrial, en este caso la relación entre la función predicha y el módulo no es tan evidente, aunque es un hecho que el flagelo utiliza energía generada por la mitocondria para su movimiento. Por otro lado, TcCLB.510347.29 fue anotada por *FoldSeek* como una "*HORMA domaing containing protein*"; trabajos recientes han descubierto otros roles a los ya reportados para esta proteína, como la regulación de la dinámica del centriolo (Muniyappa et al., 2014) y

TcCLB.506927.20 fue anotada por ambos algoritmos como una proteína de interacción con la actina.

Con respecto al módulo M11, asociado a procesos de adhesión celular, TcCLB.511623.20 fue anotado por *FoldSeek* como una "*Pleckstrin homology-like domain, family B, member 2*" (PHLDB2); está reportado que esta proteína actúa como un factor de anclaje al microtúbulo que une la proteína *CLASP*, involucrada en la interacción entre los extremos distales de los microtúbulos y la corteza celular (Chao et al., 2016).

Por último, en cuanto al módulo M12 asociado a procesos biológicos vinculados al movimiento celular, 3 de sus 4 *hubgenes* codificantes para proteínas hipotéticas fueron anotados funcionalmente por *DARK* o por ambos algoritmos: por un lado, TcCLB.511127.90 fue anotado por *DARK* como una proteína asociada al flagelo, mientras que TcCLB.503903.70 y TcCLB.511693.70 fueron anotadas por ambos algoritmos como una proteína *MORN1* y una proteína asociada a la actina y al flagelo, respectivamente. En trabajos recientes se ha reportado que *MORN1* actúa como un complejo multiproteico que podría tener un rol facilitador para la entrada de proteínas al bolsillo flagelar. En este estudio, se analizaron los efectos fenotípicos de la depleción de *MORN1* en *T. brucei*, que resultó en un rápido agrandamiento del bolsillo flagelar que no permitió la entrada al mismo de proteínas como *concanavalin A* y *bovine serum albumin* utilizadas para este estudio.

**Tabla 7.** Genes de función desconocida con su función predicha utilizando *DARK* y *FoldSeek***.** Se indica con un guion cuando la estrategia no dio resultados. Si indica si la función concuerda de manera evidente con los términos GOs sobrerepresentados para cada módulo.

| Módulo | *Hubgenes no anotados* | Anotación TriTrypDB | Anotación DARK | Anotación FoldSeek | ¿Coherente con GO? |
|---|---|---|---|---|---|
| **M1** | TcCLB.510849.30 | hypothetical protein, conserved (pseudogene) | - | - | - |
| | TcCLB.508835.10 | hypothetical protein, conserved (pseudogene) | - | - | - |
| | TcCLB.506113.80 | hypothetical protein, conserved (pseudogene) | - | - | - |
| **M5** | TcCLB.510275.272 | hypothetical protein | Trans-sialidase | - | Si |
| | TcCLB.509233.10 | hypothetical protein | Mucin-associated surface protein (MASP), putative | - | Si |
| **M7** | TcCLB.508771.50 | hypothetical protein, conserved | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | TcCLB.509157.220 | hypothetical protein, conserved | Putative transmembrane protein | Protein Fatty Acid Export 7 | Si |
| **M9** | TcCLB.508731.40 | hypothetical protein, conserved | Mitochondrial glycoprotein-like protein | Head domain of the mt-SSU assemblosome from *Trypanosoma brucei* | Si |
| | TcCLB.510347.29 | hypothetical protein, conserved | HP_Q4D059 | HORMA domain containing protein | Si |
| | TcCLB.506927.20 | hypothetical protein, conserved | Actin interacting protein 1 | Actin-interacting protein 1 | Si |
| **M10** | TcCLB.503811.45 | hypothetical protein, conserved | - | - | - |
| **M11** | TcCLB.511623.20 | hypothetical protein, conserved | - | Pleckstrin homology-like domain, family B, member 2 | Si |
| **M12** | TcCLB.511127.90 | Domain of unknown function (DUF4586), putative | Flagellar associated protein | - | Si |
| | TcCLB.503903.70 | hypothetical protein, conserved | Morn repeat protein | Crystal structure of Trypansoma brucei Morn 1 | Si |
| | TcCLB.453917.9 | hypothetical protein, conserved | - | - | - |
| | TcCLB.511693.70 | hypothetical protein, conserved | Actin interacting protein 1 | Cilia- and flagella-associated protein 52 | Si |

Por último, es importante destacar la correspondencia entre las anotaciones informáticas que fueron realizadas mediante dos estrategias que utilizan metodologías completamente distintas pero que obtuvieron resultados concordantes entre sí, a la vez que las funciones de estos genes correlacionan con los procesos biológicos subyacentes a cada módulo extraído de su sobrerrepresentación de términos GO.

Consideramos que los resultados obtenidos dan luz sobre genes que posiblemente tengan roles claves en la biología de *T. cruzi* y cuya función no ha sido explorada aún en detalle. Como mencionamos antes, abre puerta para su futura caracterización mediante método experimentales (localización subcelular, genética reversa, etc.).

# *5* Conclusiones

*Trypanosoma cruzi* es un parásito protozoario causante de la tripanosomiasis americana también denominada enfermedad de Chagas, una enfermedad tropical desatendida. *T. cruzi* se caracteriza por un ciclo de vida complejo que involucran distintas etapas de diferenciación cada una con características particulares. Sus genes se expresan de forma policistrónica, siendo los mecanismos post-transcripcionales los principales mecanismos de regulación de la expresión génica.

Los análisis de co-expresión génica son una valiosa herramienta para estudiar cambios en el nivel de expresión de grupos de genes que interactúan funcionalmente entre sí. En este estudio se buscó contribuir a la comprensión de la regulación de la expresión génica de este parásito mediante la construcción de una red de co-expresión génica utilizando datos de *RNA-seq* de los cuatro estadios su ciclo de vida, y la caracterización y análisis funcional de los grupos de genes co-expresados obtenidos. A su vez, se buscó identificar posibles mecanismos que expliquen esta regulación conjunta.

Se identificaron 13 módulos de genes co-expresados mediante la construcción de una *weighted gene co-expression network* utilizando el paquete de R *CEMiTool*. De ellos, se obtuvieron 10 módulos que estaban enriquecidos funcionalmente en roles asociados a diversos, entre ellos diversos procesos metabólicos, replicación del ADN, movimiento celular, regulación del citoqueleto, regulación de ARNm y patogénesis. Por otro lado, se logró establecer una asociación entre ciertos módulos enriquecidos funcionalmente y algunos de los estadios del parásito, en conjunto con las características biológicas que definen estos estadios.

A su vez, se observó la existencia de un perfil de uso de codones entre 3 *clusters* de módulos diferente entre sí, pero similar entre los módulos pertenecientes a cada *cluster*, y la presencia de motivos a nivel de secuencia en las regiones 3’UTR de los genes pertenecientes a módulos enriquecidos funcionalmente. Ambas observaciones apuntan a mecanismos específicos que aportan a la co-regulación de los genes de cada módulo.

Por otro lado, se contribuyó en la generación de un flujo de trabajo potencialmente útil para la priorización (mediante construcción de redes y selección de genes hub) y anotación funcional *in-silico* de genes, en particular para kinetoplástidos, utilizando dos algoritmos bioinformáticos con distintas aproximaciones: *DARK* y *FoldSeek*. En este sentido, se

realizaron inferencias funcionales de genes de función desconocida bajo el principio de "*guilt by association*" a partir de los grupos de genes co-expresados, reduciendo en primera instancia a un pequeño set de genes de gran relevancia para cada módulo de genes co-expresados. Los resultados de esta estrategia abren puertas para el estudio experimental de estos genes centrales en los módulos a partir de hipótesis funcionales específicas.

Se espera que los resultados obtenidos en este estudio realicen un aporte significativo a la comprensión de la biología molecular de este parásito de alta relevancia en nuestra región y provea información permitiendo priorizar el estudio de genes clave en el ciclo de vida de este parásito, cuyas proteínas codificantes podrían ser posibles blancos moleculares de agentes tripanocidas.

# *6* Perspectivas

El algoritmo *CEMiTool* utilizado para la construcción de la red de co-expresión génica permitió identificar 13 de módulos de genes co-expresados. Sin embargo, dos de ellos, M1 y M2, tenían una cantidad de genes relativamente grande respecto a los demás módulos. A su vez, se observó que M1 estaba sobrerrepresentado en términos GO bastante generales en el DAG de BP, mientras que el módulo M2 no presentó sobrerrepresentación significativa de ningún término GO. En lugar de tratar a los módulos grandes como entidades indivisibles, se podría examinar la aplicación de métodos de *clustering* adicionales para identificar subgrupos dentro de ellos. Seguramente esto permitiría explorar aún más la estructura de la red de co-expresión, potencialmente reduciendo la complejidad de los módulos grandes obteniendo una mayor cantidad de módulos de genes co-expresados donde es posible que haya un enriquecimiento de términos GO de procesos biológicos más específicos que los obtenidos en este estudio para esos módulos. Además, se podría incluir en la etapa de construcción de la red nuevos datos de diferentes escenarios biológicos cuya inclusión haga evidente la separación de la expresión génica de ciertos grupos que en las condiciones biológicas estudiadas en este trabajo se espera *a priori* sean co-expresados.

Por otro lado, dado que los métodos de selección del algoritmo para la construcción de la red de co-expresión estaban basados únicamente en la consistencia funcional de los módulos obtenidos por cada uno de ellos, sería interesante explorar otro tipo de métodos, por ejemplo basados en la topología de la red, como la modularidad, la conectividad o coeficientes de agrupamiento, para obtener una comprensión más profunda de la organización de la co-expresión génica en *T. cruzi*, como los propuestos por (B. Li et al., 2015).

En cuanto a los análisis de enriquecimiento funcional de los módulos, sería interesante incorporar al análisis dos otras estrategias: por un lado, un análisis de sobrerrepresentación de términos GO de la ontología MF, que podría complementar la información obtenida en el análisis de la ontología BP, y por otro lado sería interesante realizar un análisis de sobrerrepresentación de vías utilizando la base de datos KEGG para identificar vías biológicas enriquecidas en los módulos identificados. Esto ampliaría la información sobre las funciones y procesos biológicos relacionados con los genes co-expresados en cada módulo.

Recientemente el estudio de los ARN largos no codificantes (lncRNAs) ha tomado relevancia, particularmente en tripanosomátidos. Los lncRNAs son secuencias de RNA que tienen más de 200 nucleótidos y no se traducen en proteínas. Estos lncRNAs han sido identificados como participantes en mecanismos de regulación génica en diversos niveles, influyendo en diferentes procesos biológicos como la diferenciación y el ciclo celulares. Además, tienen la capacidad de interactuar con ADN, ARN y proteínas. Recientemente, se ha investigado el repertorio de lncRNAs en tripanosomátidos, y se ha descrito esta población en *T. brucei*. Sin embargo, hasta ahora, solo se ha informado de un lncRNA funcional que está implicado en los procesos de diferenciación de este organismo. En este contexto, actualmente nuestro grupo de trabajo se encuentra desarrollando una línea de investigación que consiste en la caracterización mediante aproximaciones bioinformáticas y de secuenciación masiva del repertorio de lncRNAs de *T. cruzi*. Recientemente, un análisis de redes de co-expresión génica en *Schistosoma mansoni* logró identificar lncRNAs como *hubgenes* de algunos de los módulos de genes co-expresados detectados que tendrían roles clave en los diferentes estadios de este organismo (Maciel et al., 2019). Siguiendo el enfoque de este estudio, sería interesante, una vez caracterizados los lncRNAs de *T. cruzi*, incorporarlos a nuestro estudio de redes de co-expresión génica. Esto podría revelar la posible participación de los lncRNAs en los procesos biológicos subyacentes a los módulos identificados y brindar información adicional sobre nuevos mecanismos de regulación génica en el parásito.

Por otro lado, además de buscar motivos en las regiones 3’UTR, se podría extender la búsqueda de motivos a las regiones 5'UTR. Existen informes de que los 5'UTRs pueden contener elementos regulatorios importantes, por lo que explorar esta región podría revelar nuevos conocimientos sobre la regulación de la expresión génica en *T. cruzi*.

A su vez, sería interesante analizar si las proteínas de unión a los motivos observados en las regiones 3'UTRs e identificadas en otros organismos tienen homólogos en *T. cruzi*. Esto abriría la posibilidad de explorar su perfil de expresión, su asociación con los módulos de co-expresión y su potencial papel como nodos clave en la red de regulación génica.

Por otra parte, para los genes que comparten motivos en cada módulo nos planteamos realizar un análisis de enriquecimiento de términos GO. Esto permitiría determinar si estos genes tienen funciones biológicas comunes o están involucrados en procesos celulares específicos. Además, se podría investigar si las proteínas de unión a los motivos

identificados regulan de manera particular algunos genes dentro de cada módulo, lo que brindaría información adicional sobre la regulación génica en *T. cruzi*.

Respecto al análisis de uso diferencial de codones entre los genes pertenecientes a los diferentes módulos, sería beneficioso realizar un análisis más exhaustivo determinando qué codones son preferidos en los módulos y si existe una asociación entre los codones sinónimos preferidos, la expresión estadio-específica y otros factores relevantes, como los niveles de ARNt en los diferentes estadios. Esto brindaría información sobre los patrones de uso de codones y su relación con la expresión génica en diferentes contextos. En este sentido, una de las líneas de investigación del grupo de laboratorio involucra la cuantificación de los niveles de expresión de ARNt de los estadios epimastigota y tripomastigota metacíclico mediante un protocolo específico de secuenciación masiva (Xu et al., 2019), aunque no se descarta la cuantificación en otros estadios del parásito.

En cuanto a la anotación funcional de genes, sería de gran interés poder validar experimentalmente la función de los *hubgenes* anotados en este estudio mediante técnicas de biología molecular como estudios de genética reversa, localización subcelular, etc.

A su vez, sería oportuno analizar la conectividad de los *hubgenes* en la red de co-expresión y determinar con qué otros genes se encuentran conectados. Utilizando esta información, se podrían priorizar genes para su anotación y evaluación experimental de su función. Se podría considerar realizar estudios de expresión diferencial en diferentes condiciones o etapas de *T. cruzi* para identificar genes cuya función podría ser evaluada de manera más precisa.

Por último, más allá de la implementación de un método de priorización para la anotación de genes seleccionando los *hubgenes* más conectados de cada módulo, actualmente existen métodos que realizan una anotación funcional global mediante la asignación de términos GO utilizando redes de co-expresión génica. Actualmente nuestro grupo de laboratorio se encuentra en colaboración con los desarrolladores del algoritmo *exp2GO* (Di Persia et al., 2022) y nos encontramos en vías de optimizar los parámetros que utiliza este algoritmo para anotar funcionalmente la mayor cantidad de genes posibles. Brevemente, este algoritmo toma en cuenta la estructura de la ontología mediante el uso de las distancias semánticas entre genes anotados, y además la evidencia empírica de sus co-expresiones, infiriendo las distancias semánticas faltantes con el resto de los genes mediante la factorización de matrices no negativas. Esta técnica permite descomponer una

matriz de datos X (no negativos) como el producto de dos matrices (A y H). La matriz A es como un diccionario de coeficientes para la matriz H y permite reconstruir la matriz X, si es que esta última tiene datos faltantes. En particular, para nuestro problema a partir de la matriz similaridad de expresión TOM se buscará una matriz A (usando NNMF) que se podrá utilizar para reconstruir la parte desconocida de la matriz de distancias semánticas $D_{GO}$. Una vez obtenida la matriz de distancia de términos GO para los genes desconocidos, se obtendrá el mapeo de ésta a etiquetas de función que caracterizan al gen correspondiente. De este modo, cada gen desconocido puede obtener etiquetas de anotación GO a partir de las anotaciones de los genes más cercanos en el espacio semántico reconstruido (**figura suplementaria 2**).

En resumen, la implementación de estas perspectivas proporcionaría una visión más completa del estudio de los perfiles de co-expresión génica de *Trypanosoma cruzi* y contribuiría a una mejor comprensión de su biología y patogenia.

# *7* Referencias bibliográficas


Abrahim, M., Machado, E., Alvarez-Valín, F., de Miranda, A. B., & Catanho, M. (2022). Uncovering Pseudogenes and Intergenic Protein-coding Sequences in TriTryps' Genomes. *Genome Biology and Evolution*, *14*(10). https://doi.org/10.1093/GBE/EVAC142

Abu-Jamous, B., Fa, R., Roberts, D. J., & Nandi, A. K. (2013). Paradigm of Tunable Clustering Using Binarization of Consensus Partition Matrices (Bi-CoPaM) for Gene Discovery. *PLOS ONE*, *8*(2), e56432. https://doi.org/10.1371/JOURNAL.PONE.0056432

Abu-Jamous, B., & Kelly, S. (2018). Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biology*, *19*(1), 1–11. https://doi.org/10.1186/S13059-018-1536-8/FIGURES/5

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, *118*(21), 4947–4957. https://doi.org/10.1242/JCS.02714

*AlphaFold reveals the structure of the protein universe*. (n.d.). Retrieved September 7, 2022, from https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe

Alvarez, F., Robeilo, C., & Vignalp, M. (1994). Evolution of codon usage and base contents in kinetoplastid protozoans. *Molecular Biology and Evolution*, *11*(5), 790–802. https://doi.org/10.1093/oxfordjournals.molbev.a040159

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*.

Angov, E. (2011). Codon usage: Nature's roadmap to expression and folding of proteins. In *Biotechnology Journal* (Vol. 6, Issue 6, pp. 650–659). Wiley-Blackwell. https://doi.org/10.1002/biot.201000332

Araújo, P. R., & Teixeira, S. M. (2011a). Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in trypanosoma cruzi - A review. In *Memorias do Instituto Oswaldo Cruz* (Vol. 106, Issue 3, pp. 257–266). Fundacao Oswaldo Cruz. https://doi.org/10.1590/S0074-02762011000300002

Araújo, P. R., & Teixeira, S. M. (2011b). Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in Trypanosoma cruzi: a review. *Memórias Do Instituto Oswaldo Cruz*, *106*(3), 257–266. https://doi.org/10.1590/S0074-02762011000300002

Ash, C., & Jasny, B. R. (2005). Trypanosomatid Genomes. *Science*, *309*, 399–436. https://science.sciencemag.org/content/309/5733/399

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25. https://doi.org/10.1038/75556

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., Depledge, D. P., Fischer, S., Gajria, B., Gao, X., Gardner, M. J., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Hertz-Fowler, C., Houston, R., Innamorato, F., Iodice, J., … Wang, H. (2009). TriTrypDB: A functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research*, *38*(SUPPL.1), D457. https://doi.org/10.1093/nar/gkp851

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, *37*(SUPPL. 2), W202–W208. https://doi.org/10.1093/nar/gkp335

Bangs, J., Crain, P., Hashizume, T., McCloskey, J., & Boothroyd, J. (1992). Mass spectrometry of mRNA cap 4 from trypanosomatids reveals two novel nucleosides. *Undefined*.

Barabási, A. L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/SCIENCE.286.5439.509

Barabási, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics 2004 5:2*, *5*(2), 101–113. https://doi.org/10.1038/nrg1272

Barbieri Holetz, F., Correa, A., Rodrigues, A., Vila, A. ´, Nakamura, V., Krieger, M. A., & Goldenberg, S. (2007). *Evidence of P-body-like structures in Trypanosoma cruzi*. https://doi.org/10.1016/j.bbrc.2007.03.104

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X

Benz, C., Nilsson, D., Andersson, B., Clayton, C., & Guilbride, D. L. (2005). Messenger RNA processing sites in Trypanosoma brucei. *Molecular and Biochemical Parasitology*, *143*(2), 125–134. https://doi.org/10.1016/J.MOLBIOPARA.2005.05.008

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology 2016 34:5*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Buscaglia, C. A., & Di Noia, J. M. (2003). Trypanosoma cruzi clonal diversity and the epidemiology of Chagas' disease. *Microbes and Infection*, *5*(5), 419–427. https://doi.org/10.1016/S1286-4579(03)00050-9

Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., Mushayahama, T., … Elser, J. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, *49*(D1), D325–D334. https://doi.org/10.1093/NAR/GKAA1113

Carlier, Y., Sosa-Estani, S., Luquetti, A. O., & Buekens, P. (2015). Congenital Chagas disease: an update. *Memorias Do Instituto Oswaldo Cruz*, *110*(3), 363–368. https://doi.org/10.1590/0074-02760140405

Castro, J. A., De Mecca, M. M., & Bartel, L. C. (2006). Toxic side effects of drugs used to treat Chagas' disease (American trypanosomiasis). *Human and Experimental Toxicology*, *25*(8), 471–479. https://doi.org/10.1191/0960327106het653oa

Chao, T., Zhou, X., Cao, B., Liao, P., Liu, H., Chen, Y., Park, H. W., Zeng, S. X., & Lu, H. (2016). Pleckstrin homology domain-containing protein PHLDB3 supports cancer growth via a negative feedback loop involving p53. *Nature Communications 2016 7:1*, *7*(1), 1–12. https://doi.org/10.1038/ncomms13755

Cheng, C. W., Beech, D. J., & Wheatcroft, S. B. (2020). Advantages of CEMiTool for gene co-expression analysis of RNA-seq data. *Computers in Biology and Medicine*, *125*. https://doi.org/10.1016/J.COMPBIOMED.2020.103975

Clayton, C., & Shapira, M. (2007). Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. In *Molecular and Biochemical Parasitology* (Vol. 156, Issue 2, pp. 93–101). https://doi.org/10.1016/j.molbiopara.2007.07.007

Collart, M. A., & Weiss, B. (2020). Ribosome pausing, a dangerous necessity for co-translational events. In *Nucleic acids research* (Vol. 48, Issue 3, pp. 1043–1055). NLM (Medline). https://doi.org/10.1093/nar/gkz763

Coughlin, B. C., Teixeira, S. M. R., Kirchhoff, L. V., & Donelson, J. E. (2000). Amastin mRNA abundance in Trypanosoma cruzi is controlled by a 3'- untranslated region position-dependent cis-element and an untranslated region-binding protein. *Journal of Biological Chemistry*, *275*(16), 12051–12060. https://doi.org/10.1074/jbc.275.16.12051

Cruz-Saavedra, L., Muñoz, M., Patiño, L. H., Vallejo, G. A., Guhl, F., & David Ramírez, J. (n.d.). *Slight temperature changes cause rapid transcriptomic responses in Trypanosoma cruzi metacyclic trypomastigotes*. https://doi.org/10.1186/s13071-020-04125-y

Daniels, J.-P., Gull, K., & Wickstead, B. (2010). Cell Biology of the Trypanosome Genome. *MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS*, *74*(4), 1092–2172. https://doi.org/10.1128/MMBR.00024-10

Datta, S., & Datta, S. (2006). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, *7*(1), 1–9. https://doi.org/10.1186/1471-2105-7-397/FIGURES/6

D'avila-Levy, C. M., Boucinha, C., Kostygov, A., Santos, H. L. C., Morelli, K. A., Grybchuk-Ieremenko, A., Duval, L., Votýpka, J., Yurchenko, V., Grellier, P., & Lukeš, J. (2015). Exploring the environmental diversity of kinetoplastid flagellates in the high-throughput DNA sequencing era. *Memórias Do Instituto Oswaldo Cruz*, *110*(8), 956–965. https://doi.org/10.1590/0074-02760150253

De Gaudenzi, J. G., Carmona, S. J., Ağuero, F., & Frasch, A. C. (2013). Genome-wide analysis of 3'-untranslated regions supports the existence of post-transcriptional

regulons controlling gene expression in trypanosomes. *PeerJ*, *1*(1). https://doi.org/10.7717/PEERJ.118

De Gaudenzi, J. G., Noé, G., Campo, V. A., Frasch, A. C., & Cassola, A. (2011). Gene expression regulation in trypanosomatids. *Essays in Biochemistry*, *51*(1), 31–46. https://doi.org/10.1042/BSE0510031/78270

De Souza, W. (1984). Cell Biology of Trypanosoma cruzi. *International Review of Cytology*, *86*(C), 197–283. https://doi.org/10.1016/S0074-7696(08)60180-1

de Souza, W. (2009). Structural organization of Trypanosoma cruzi. *Memorias Do Instituto Oswaldo Cruz*, *104*(SUPPL. 1), 89–100. https://doi.org/10.1590/s0074-02762009000900014

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/J.2517-6161.1977.TB01600.X

Deschamps-Francoeur, G., Simoneau, J., & Scott, M. S. (2020). Handling multi-mapped reads in RNA-seq. *Computational and Structural Biotechnology Journal*, *18*, 1569–1576. https://doi.org/10.1016/J.CSBJ.2020.06.014

Di Noia, J. M., D'Orso, I., Sánchez, D. O., & Frasch, A. C. C. (2000). AU-rich elements in the 3'-untranslated region of a new mucin-type gene family of Trypanosoma cruzi confers mRNA instability and modulates translation efficiency. *Journal of Biological Chemistry*, *275*(14), 10218–10227. https://doi.org/10.1074/jbc.275.14.10218

Di Persia, L., Lopez, T., Arce, A., Milone, D. H., & Stegmayer, G. (2022). exp2GO: improving prediction of functions in the Gene Ontology with expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *01*, 1–1. https://doi.org/10.1109/TCBB.2022.3167245

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15. https://doi.org/10.1093/BIOINFORMATICS/BTS635

Docampo, R., & Moreno, S. N. J. (2011). Acidocalcisomes. In *Cell Calcium* (Vol. 50, Issue 2, pp. 113–119). Elsevier Ltd. https://doi.org/10.1016/j.ceca.2011.05.012

Duhagon, M. A., Dallagiovanna, B., & Garat, B. (2001). Unusual Features of Poly[dT-dG]·[dC-dA] Stretches in CDS-Flanking Regions of Trypanosoma cruzi Genome. *Biochemical and Biophysical Research Communications*, *287*(1), 98–103. https://doi.org/10.1006/BBRC.2001.5545

El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C., Nilsson, D., Aggarwal, G., Tran, A. N., Ghedin, E., Worthey, E. A., Delcher, A. L., Blandin, G., Westenberger, S. J., Caler, E., Cerqueira, G. C., Branche, C., Haas, B., Anupama, A., Arner, E., Åslund, L., Attipoe, P., … Andersson, B. (2005). The genome sequence of Trypanosoma cruzi, etiologic agent of chagas disease. *Science*, *309*(5733), 409–415. https://doi.org/10.1126/science.1112631

El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E. A., Hertz-Fowler, C., Ghedin, E., Peacock, C., Bartholomeu, D. C., Haas, B. J., Tran, A. N., Wortman, J. R., Alsmark, U. C. M., Angiuoli, S., Anupama, A., … Hall, N. (2005). Comparative genomics of trypanosomatid parasitic protozoa. *Science*, *309*(5733), 404–409. https://doi.org/10.1126/science.1112181

Estévez, A. M. (2008). The RNA-binding protein Tb DRBD3 regulates the stability of a specific subset of mRNAs in trypanosomes. *Nucleic Acids Research*, *36*(14), 4573–4586. https://doi.org/10.1093/NAR/GKN406

Frazee, A. C., Jaffe, A. E., Langmead, B., & Leek, J. T. (2015). Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, *31*(17), 2778. https://doi.org/10.1093/BIOINFORMATICS/BTV272

Godichon-Baggioni, A., Maugis-Rabusseau, C., & Rau, A. (2019). Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. *Journal of Applied Statistics*, *46*(1), 47–65. https://doi.org/10.1080/02664763.2018.1454894

Grant, C. E., & Bailey, T. L. (2021). XSTREME: Comprehensive motif analysis of biological sequence datasets. *BioRxiv*, 2021.09.02.458722. https://doi.org/10.1101/2021.09.02.458722

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018. https://doi.org/10.1093/BIOINFORMATICS/BTR064

Gupta, S. Das. (1960). Point biserial correlation coefficient and its generalization. *Psychometrika*, *25*(4), 393–408. https://doi.org/10.1007/BF02289756/METRICS

Haile, S., & Papadopoulou, B. (2007). Developmental regulation of gene expression in trypanosomatid parasitic protozoa. *Current Opinion in Microbiology*, *10*(6), 569–577. https://doi.org/10.1016/J.MIB.2007.10.001

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. In *Source: Journal of the Royal Statistical Society. Series C (Applied Statistics)* (Vol. 28, Issue 1).

Hershberg, R., & Petrov, D. (2008). Selection on Codon Bias Comparative population genomics of latitudinal variation in Drosophila simulans and Drosophila melanogaster View project Adaptive dynamics of cuticular hydrocarbons in Drosophila View project. *Article in Annual Review of Genetics*. https://doi.org/10.1146/annurev.genet.42.110807.091442

Horn, D. (2008a). Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. *BMC Genomics*, *9*(1), 1–11. https://doi.org/10.1186/1471-2164-9-2

Horn, D. (2008b). Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. *BMC Genomics*, *9*, 1–11. https://doi.org/10.1186/1471-2164-9-2

Jeacock, L., Faria, J., & Horn, D. (2018). Codon usage bias controls mRNA and protein abundance in trypanosomatids. *ELife*, *7*, 1–20. https://doi.org/10.7554/eLife.32496

Jia, B., Jia, X., Kim, K. H., & Jeon, C. O. (2017). Integrative view of 2-oxoglutarate/Fe(II)-dependent oxygenase diversity and functions in bacteria. *Biochimica et Biophysica Acta - General Subjects*, *1861*(2), 323–334. https://doi.org/10.1016/j.bbagen.2016.12.001

Joshi, C. J., Ke, W., Drangowska-Way, A., O'Rourke, E. J., & Lewis, N. E. (2022). What are housekeeping genes? *PLoS Computational Biology*, *18*(7). https://doi.org/10.1371/JOURNAL.PCBI.1010295

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Keene, J. D. (2007). RNA regulons: coordination of post-transcriptional events. *Nature Reviews. Genetics*, *8*(7), 533–543. https://doi.org/10.1038/NRG2111

Kempen, M. van, Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *BioRxiv*, 2022.02.07.479398. https://doi.org/10.1101/2022.02.07.479398

Kharchenko, P., Church, G. M., & Vitkup, D. (2005). Expression dynamics of a cellular metabolic network. *Molecular Systems Biology*, *1*. https://doi.org/10.1038/MSB4100023

Koutrouli, M., Karatzas, E., Paez-Espino, D., & Pavlopoulos, G. A. (2020). A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology*, *8*, 34. https://doi.org/10.3389/FBIOE.2020.00034/BIBTEX

Kramer, S. (2012). Developmental regulation of gene expression in the absence of transcriptional control: The case of kinetoplastids. In *Molecular and Biochemical Parasitology* (Vol. 181, Issue 2, pp. 61–72). Elsevier. https://doi.org/10.1016/j.molbiopara.2011.10.002

Krissinel, E., & Henrick, K. (2005). Multiple alignment of protein structures in three dimensions. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3695 LNBI*, 67–78. https://doi.org/10.1007/11560500_7/COVER

Kulmanov, M., Khan, M. A., & Hoehndorf, R. (2018). DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, *34*(4), 660–668. https://doi.org/10.1093/BIOINFORMATICS/BTX624

Langfelder, P., & Horvath, S. (2008). *WGCNA: an R package for weighted correlation network analysis*. https://doi.org/10.1186/1471-2105-9-559

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, *24*(5), 719–720. https://doi.org/10.1093/BIOINFORMATICS/BTM563

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lesnik, T., Solomovici, J., Deana, A., Ehrlich, R., & Reiss, C. (2000). Ribosome traffic in E. coli and regulation of gene expression. *Journal of Theoretical Biology*, *202*(2), 175–185. https://doi.org/10.1006/jtbi.1999.1047

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493–500. https://doi.org/10.1093/BIOINFORMATICS/BTP692

Li, B., Zhang, Y., Yu, Y., Wang, P., Wang, Y., Wang, Z., & Wang, Y. (2015). Quantitative assessment of gene expression network module-validation methods. *Scientific Reports 2015 5:1*, *5*(1), 1–14. https://doi.org/10.1038/srep15258

Li, Y., Shah-Simpson, S., Okrah, K., Belew, A. T., Choi, J., Caradonna, K. L., Padmanabhan, P., Ndegwa, D. M., Temanni, M. R., Corrada Bravo, H., El-Sayed, N. M., & Burleigh, B. A. (2016). Transcriptome Remodeling in Trypanosoma cruzi and Human Cells during Intracellular Infection. *PLoS Pathogens*, *12*(4). https://doi.org/10.1371/journal.ppat.1005511

Li, Z. H., De Gaudenzi, J. G., Alvarez, V. E., Mendiondo, N., Wang, H., Kissinger, J. C., Frasch, A. C., & Docampo, R. (2012). A 43-nucleotide U-rich element in 3′-untranslated region of large number of Trypanosoma cruzi transcripts is important for mRNA Abundance in intracellular amastigotes. *Journal of Biological Chemistry*, *287*(23), 19058–19069. https://doi.org/10.1074/jbc.M111.338699

Liang, J. W., Fang, Z. Y., Huang, Y., Liuyang, Z. Y., Zhang, X. L., Wang, J. L., Wei, H., Wang, J. Z., Wang, X. C., Zeng, J., & Liu, R. (2018). Application of Weighted Gene Co-Expression Network Analysis to Explore the Key Genes in Alzheimer's Disease. *Journal of Alzheimer's Disease : JAD*, *65*(4), 1353–1364. https://doi.org/10.3233/JAD-180400

Liao, Y., Smyth, G. K., & Shi, W. (2014). *Sequence analysis featureCounts: an efficient general purpose program for assigning sequence reads to genomic features*. *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Love, M. I., Anders, S., & Huber, W. (2018). *Analyzing RNA-seq data with DESeq2*. https://doi.org/10.1186/s13059-014-0550-8

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9

Lucas, M. C., Pryszcz, L. P., Medina, R., Milenkovic, I., Camacho, N., Marchand, V., Motorin, Y., Ribas de Pouplana, L., & Novoa, E. M. (2023). Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. *Nature Biotechnology 2023*, 1–15. https://doi.org/10.1038/s41587-023-01743-6

Lukeš, J., Butenko, A., Hashimi, H., Maslov, D. A., Votýpka, J., & Yurchenko, V. (2018). Trypanosomatids Are Much More than Just Trypanosomes: Clues from the Expanded Family Tree. In *Trends in Parasitology* (Vol. 34, Issue 6, pp. 466–480). Elsevier Ltd. https://doi.org/10.1016/j.pt.2018.03.002

Maciel, L. F., Morales-Vicente, D. A., Silveira, G. O., Ribeiro, R. O., Olberg, G. G. O., Pires, D. S., Amaral, M. S., & Verjovski-Almeida, S. (2019). Weighted Gene Co-Expression Analyses Point to Long Non-Coding RNA Hub Genes at Different Schistosoma mansoni Life-Cycle Stages. *Frontiers in Genetics*, *10*, 823. https://doi.org/10.3389/FGENE.2019.00823/BIBTEX

Mair, G., Shi, H., Li, H., Djikeng, A., Aviles, H. O., Bishop, J. R., Falcone, F. H., Gavrilescu, C., Montgomery, J. L., Santori, M. I., Stern, L. S., Wang, Z., Ullu, E., & Tschudi, C. (2000). A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. *RNA*, *6*(2), 163–169. https://doi.org/10.1017/S135583820099229X

Marbaix, A. Y., Noël, G., Detroux, A. M., Vertommen, D., Van Schaftingen, E., & Linster, C. L. (2011). Extremely conserved ATP- or ADP-dependent enzymatic system for nicotinamide nucleotide. *Journal of Biological Chemistry*, *286*(48), 41246–41252. https://doi.org/10.1074/jbc.C111.310847

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *8*(1), 3–30. https://doi.org/10.1145/272991.272995

Michels, P. A. M., Bringaud, F., Herman, M., & Hannaert, V. (2006). Metabolic functions of glycosomes in trypanosomatids. In *Biochimica et Biophysica Acta - Molecular Cell Research* (Vol. 1763, Issue 12, pp. 1463–1477). Elsevier. https://doi.org/10.1016/j.bbamcr.2006.08.019

Minning, T. A., Weatherly, D. B., Atwood, J., Orlando, R., & Tarleton, R. L. (2009). The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi. *BMC Genomics*, *10*. https://doi.org/10.1186/1471-2164-10-370

Miranda, K., Benchimol, M., Docampo, R., & De Souza, W. (2000). The fine structure of acidocalcisomes in Trypanosoma cruzi. *Parasitology Research*, *86*(5), 373–384. https://doi.org/10.1007/s004360050682

Moreira, D., López-García, P., & Vickerman, K. (2004). An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: Proposal for a new classification of the class Kinetoplastea. *International Journal of Systematic and Evolutionary Microbiology*, *54*(5), 1861–1875. https://doi.org/10.1099/IJS.0.63081-0/CITE/REFWORKS

Muniyappa, K., Kshirsagar, R., & Ghodke, I. (2014). The HORMA domain: an evolutionarily conserved domain discovered in chromatin-associated proteins, has unanticipated diverse functions. *Gene*, *545*(2), 194–197. https://doi.org/10.1016/J.GENE.2014.05.020

Mwangi, K. W., Macharia, R. W., & Bargul, J. L. (2021). Gene co-expression network analysis of Trypanosoma brucei in tsetse fly vector. *Parasites and Vectors*, *14*(1), 1–11. https://doi.org/10.1186/S13071-021-04597-6/FIGURES/4

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935. https://doi.org/10.1093/BIOINFORMATICS/BTT509

Newman, M. E. J. (2002). Assortative Mixing in Networks. *Physical Review Letters*, *89*(20), 208701. https://doi.org/10.1103/PHYSREVLETT.89.208701/FIGURES/1/MEDIUM

Noé, G., De Gaudenzi, J. G., & Frasch, A. C. (2008a). Functionally related transcripts have common RNA motifs for specific RNA-binding proteins in trypanosomes. *BMC Molecular Biology*, *9*(1), 1–19. https://doi.org/10.1186/1471-2199-9-107/TABLES/6

Noé, G., De Gaudenzi, J. G., & Frasch, A. C. (2008b). Functionally related transcripts have common RNA motifs for specific RNA-binding proteins in trypanosomes. *BMC Molecular Biology*, *9*(1), 1–19. https://doi.org/10.1186/1471-2199-9-107/TABLES/6

Noé, G., De Gaudenzi, J. G., & Frasch, A. C. (2008c). Functionally related transcripts have common RNA motifs for specific RNA-binding proteins in trypanosomes. *BMC Molecular Biology*, *9*(1), 1–19. https://doi.org/10.1186/1471-2199-9-107/TABLES/6

Oldham, M. C., Langfelder, P., & Horvath, S. (2012). Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Systems Biology*, *6*. https://doi.org/10.1186/1752-0509-6-63

OMS | Enfermedades tropicales desatendidas: preguntas más frecuentes. (2010). *WHO*.

Ouellette, M., & Papadopoulou, B. (2009). Coordinated gene expression by post-transcriptional regulons in African trypanosomes. *Journal of Biology*, *8*(11), 1–4. https://doi.org/10.1186/JBIOL203/FIGURES/1

Palenchar, J. B., & Bellofatto, V. (2006). Gene transcription in trypanosomes. *Molecular & Biochemical Parasitology*, *146*, 135–141. https://doi.org/10.1016/j.molbiopara.2005.12.008

Pastro, L., Smircich, P., Pérez-Díaz, L., Duhagon, M. A., & Garat, B. (2013). Implication of CA repeated tracts on post-transcriptional regulation in Trypanosoma cruzi. *Experimental Parasitology*, *134*(4), 511–518. https://doi.org/10.1016/J.EXPPARA.2013.04.004

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., & Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Mining*, *4*(1), 1–27. https://doi.org/10.1186/1756-0381-4-10/FIGURES/11

Pink, R. C., & Carter, D. R. F. (2013). Pseudogenes as regulators of biological function. *Essays in Biochemistry*, *54*(1), 103–112. https://doi.org/10.1042/BSE0540103

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: The causes and consequences of codon bias. In *Nature Reviews Genetics* (Vol. 12, Issue 1, pp. 32–42). Nature Publishing Group. https://doi.org/10.1038/nrg2899

Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R., & Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell*, *160*(6), 1111–1124. https://doi.org/10.1016/j.cell.2015.02.029

Queiroz, R., Benz, C., Fellenberg, K., Hoheisel, J. D., & Clayton, C. (2009). Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons. *BMC Genomics*, *10*(1), 495. https://doi.org/10.1186/1471-2164-10-495/FIGURES/7

Radhakrishnan, A., Chen, Y. H., Martin, S., Alhusaini, N., Green, R., & Coller, J. (2016). The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell*, *167*(1), 122-132.e9. https://doi.org/10.1016/j.cell.2016.08.053

Radío, S., Fort, R. S., Garat, B., Sotelo-Silveira, J., & Smircich, P. (2018). UTRme: A Scoring-Based Tool to Annotate Untranslated Regions in Trypanosomatid Genomes. *Frontiers in Genetics*, *9*, 671. https://doi.org/10.3389/fgene.2018.00671

Rassi, A., Rassi, A., & Marin-Neto, J. A. (2010). Chagas disease. In *The Lancet* (Vol. 375, Issue 9723, pp. 1388–1402). https://doi.org/10.1016/S0140-6736(10)60061-X

Rau, A., & Maugis-Rabusseau, C. (2018). Transformation and model choice for RNA-seq co-expression analysis. *Briefings in Bioinformatics*, *19*(3), 425–436. https://doi.org/10.1093/bib/bbw128

Riquelme Medina, I., & Lubovac-Pilav, Z. (2016). Gene Co-Expression Network Analysis for Identifying Modules and Functionally Enriched Pathways in Type 1 Diabetes. *PLOS ONE*, *11*(6), e0156006. https://doi.org/10.1371/JOURNAL.PONE.0156006

Rohloff, P., Montalvetti, A., & Docampo, R. (2004). Acidocalcisomes and the contractile vacuole complex are involved in osmoregulation in Trypanosoma cruzi. *Journal of Biological Chemistry*, *279*(50), 52270–52281. https://doi.org/10.1074/jbc.M410372200

Russo, P. S. T., Ferreira, G. R., Cardozo, L. E., Bürger, M. C., Arias-Carrasco, R., Maruyama, S. R., Hirata, T. D. C., Lima, D. S., Passos, F. M., Fukutani, K. F., Lever, M., Silva, J. S., Maracaja-Coutinho, V., & Nakaya, H. I. (2018). CEMiTool: A Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/S12859-018-2053-1

Sabalette, K. B., Romaniuk, M. A., Noé, G., Cassola, A., Campo, V. A., & De Gaudenzi, J. G. (2019). The RNA-binding protein TcUBP1 up-regulates an RNA regulon for a cell surface–associated Trypanosoma cruzi glycoprotein and promotes parasite infectivity. *Journal of Biological Chemistry*, *294*(26), 10349–10364. https://doi.org/10.1074/JBC.RA118.007123/ATTACHMENT/4EEF03FE-5132-46F6-8E14-9B2453059C2F/MMC1.ZIP

Schenkman, S., & Pascoalino, B. (2011). Nuclear Structure of Trypanosoma cruzi Toxoplasma epigenetics View project New Medicine for Trypanosomatidic Infections _ FP7 Research&Innovation project View project. *Article in Advances in Parasitology*. https://doi.org/10.1016/B978-0-12-385863-4.00012-5

Shigematsu, M., Honda, S., Loher, P., Telonis, A. G., Rigoutsos, I., & Kirino, Y. (2017). YAMAT-seq: An efficient method for high-throughput sequencing of mature transfer RNAs. *Nucleic Acids Research*, *45*(9), e70. https://doi.org/10.1093/nar/gkx005

Smircich, P., Eastman, G., Bispo, S., Duhagon, M. A., Guerra-Slompo, E. P., Garat, B., Goldenberg, S., Munroe, D. J., Dallagiovanna, B., Holetz, F., & Sotelo-Silveira, J. R. (2015). Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in Trypanosoma cruzi. *BMC Genomics*, *16*(1), 1–14. https://doi.org/10.1186/s12864-015-1563-8

Smircich, P., Pérez-Díaz, L., Hernández, F., Duhagon, M. A., & Garat, B. (2023). Transcriptomic analysis of the adaptation to prolonged starvation of the insect-dwelling Trypanosoma cruzi epimastigotes. *Frontiers in Cellular and Infection Microbiology*, *13*, 374. https://doi.org/10.3389/FCIMB.2023.1138456

Smith, D. F., & Parsons, Marilyn. (1996). *Molecular biology of parasitic protozoa*. IRL Press at Oxford University Press. https://agris.fao.org/agris-search/search.do?recordID=US201300300462

Smith, M., Blanchette, M., & Papadopoulou, B. (2008). Improving the prediction of mRNA extremities in the parasitic protozoan Leishmania. *BMC Bioinformatics*, *9*(1), 1–12. https://doi.org/10.1186/1471-2105-9-158/TABLES/3

Stuart, K., Brun, R., Croft, S., Fairlamb, A., Gürtler, R. E., McKerrow, J., Reed, S., & Tarleton, R. (2008). Kinetoplastids: Related protozoan pathogens, different diseases. *Journal of Clinical Investigation*, *118*(4), 1301–1310. https://doi.org/10.1172/JCI33945

Teixeira, A. R. L., Calixto, M. A., & Teixeira, M. L. (1994). Chagas' disease: carcinogenic activity of the antitrypanosomal nitroarenes in mice. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, *305*(2), 189–196. https://doi.org/10.1016/0027-5107(94)90239-9

Thanaraj, T. A., & Argos, P. (1996). Ribosome-mediated translational pause and protein domain organization. *Protein Science*, *5*(8), 1594–1612. https://doi.org/10.1002/pro.5560050814

Tian, H., Guan, D., & Li, J. (2018). Identifying osteosarcoma metastasis associated genes by weighted gene co-expression network analysis (WGCNA). *Medicine*, *97*(24). https://doi.org/10.1097/MD.0000000000010781

Törönen, P., Medlar, A., & Holm, L. (2018). PANNZER2: a rapid functional annotation web server. *Nucleic Acids Research*, *46*(W1), W84–W88. https://doi.org/10.1093/NAR/GKY350

Vanhamme, L., & Pays, E. (1995). Control of gene expression in trypanosomes. *Microbiological Reviews*, *59*(2), 223. https://doi.org/10.1128/MR.59.2.223-240.1995

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics (Oxford, England)*, *23*(10), 1274–1281. https://doi.org/10.1093/BIOINFORMATICS/BTM087

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57. https://doi.org/10.1038/NRG2484

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80. https://doi.org/10.2307/3001968

Xu, H., Yao, J., Wu, D. C., & Lambowitz, A. M. (2019). Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Scientific Reports 2019 9:1*, *9*(1), 1–17. https://doi.org/10.1038/s41598-019-44457-z

Yu Hsuan Teng, elicia, Wang, Y., & Luen Tang, B. (2001). The syntaxins. *Genome Biology*, *2*(11). http://genomebiology.com/2001/2/11/reviews/3012.1http://genomebiology.com/2001/2/11/reviews/3012

Zhang, J., Wang, H., Imhof, S., Zhou, X., Liao, S., Atanasov, I., Hui, W. H., Hill, K. L., & Zhou, Z. H. (2021). Structure of the trypanosome paraflagellar rod and insights into non-planar motility of eukaryotic cells. *Cell Discovery*, *7*(1). https://doi.org/10.1038/s41421-021-00281-2

Zingales, B., Andrade, S. G., Briones, M. R. S., Campbell, D. A., Chiari, E., Fernandes, O., Guhl, F., Lages-Silva, E., Macedo, A. M., Machado, C. R., Miles, M. A., Romanha, A. J., Sturm, N. R., Tibayrenc, M., & Schijman, A. G. (2009). A new consensus for Trypanosoma cruzi intraspecific nomenclature: second revision meeting recommends TcI to TcVI. *Memorias Do Instituto Oswaldo Cruz*, *104*(7), 1051–1054. https://doi.org/10.1590/S0074-02762009000700021

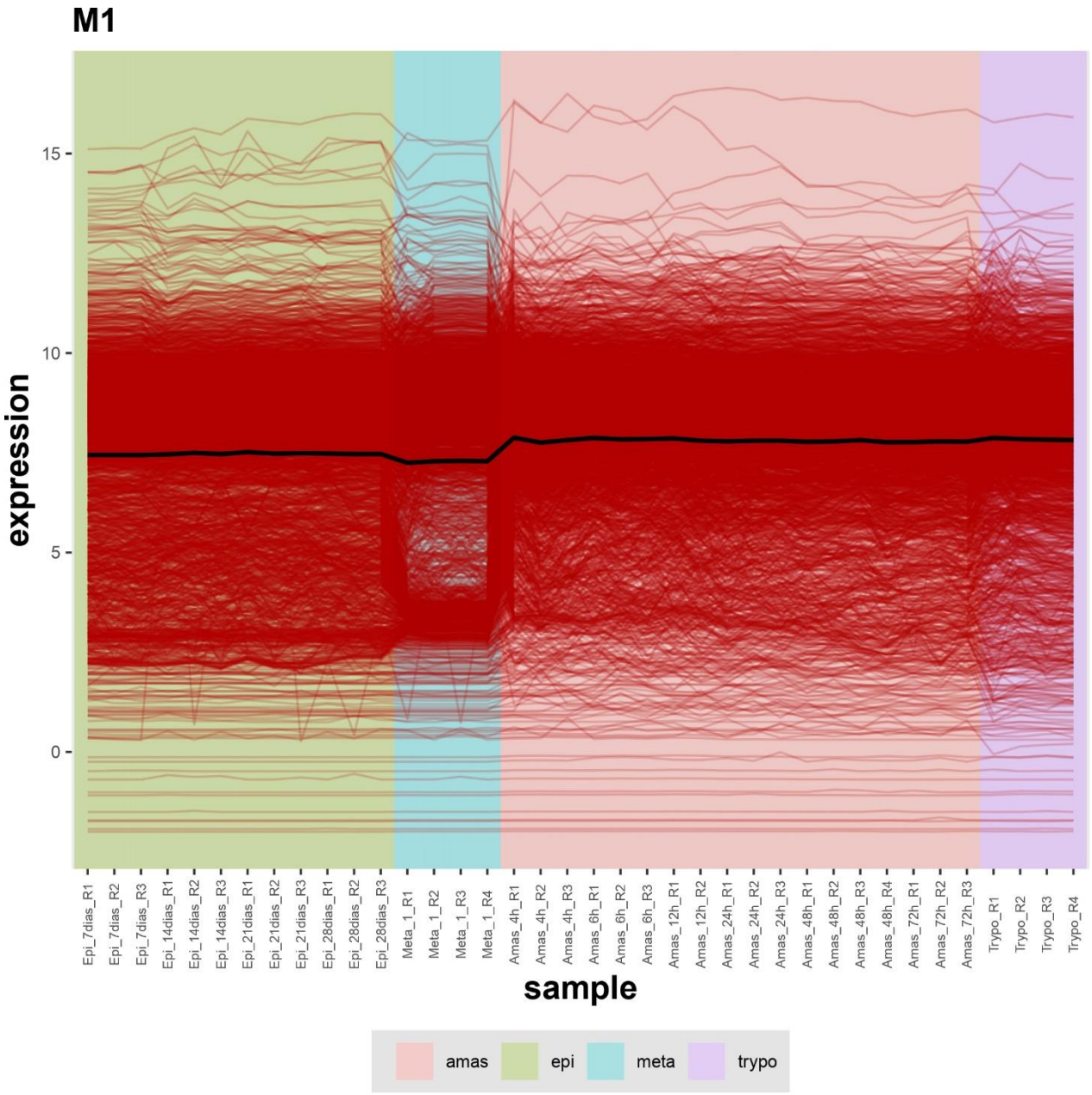
Zingales, B., Miles, M. A., Campbell, D. A., Tibayrenc, M., Macedo, A. M., Teixeira, M. M. G., Schijman, A. G., Llewellyn, M. S., Lages-Silva, E., Machado, C. R., Andrade, S. G., & Sturm, N. R. (2012). The revised Trypanosoma cruzi subspecific nomenclature: rationale, epidemiological relevance and research applications. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *12*(2), 240–253. https://doi.org/10.1016/J.MEEGID.2011.12.009

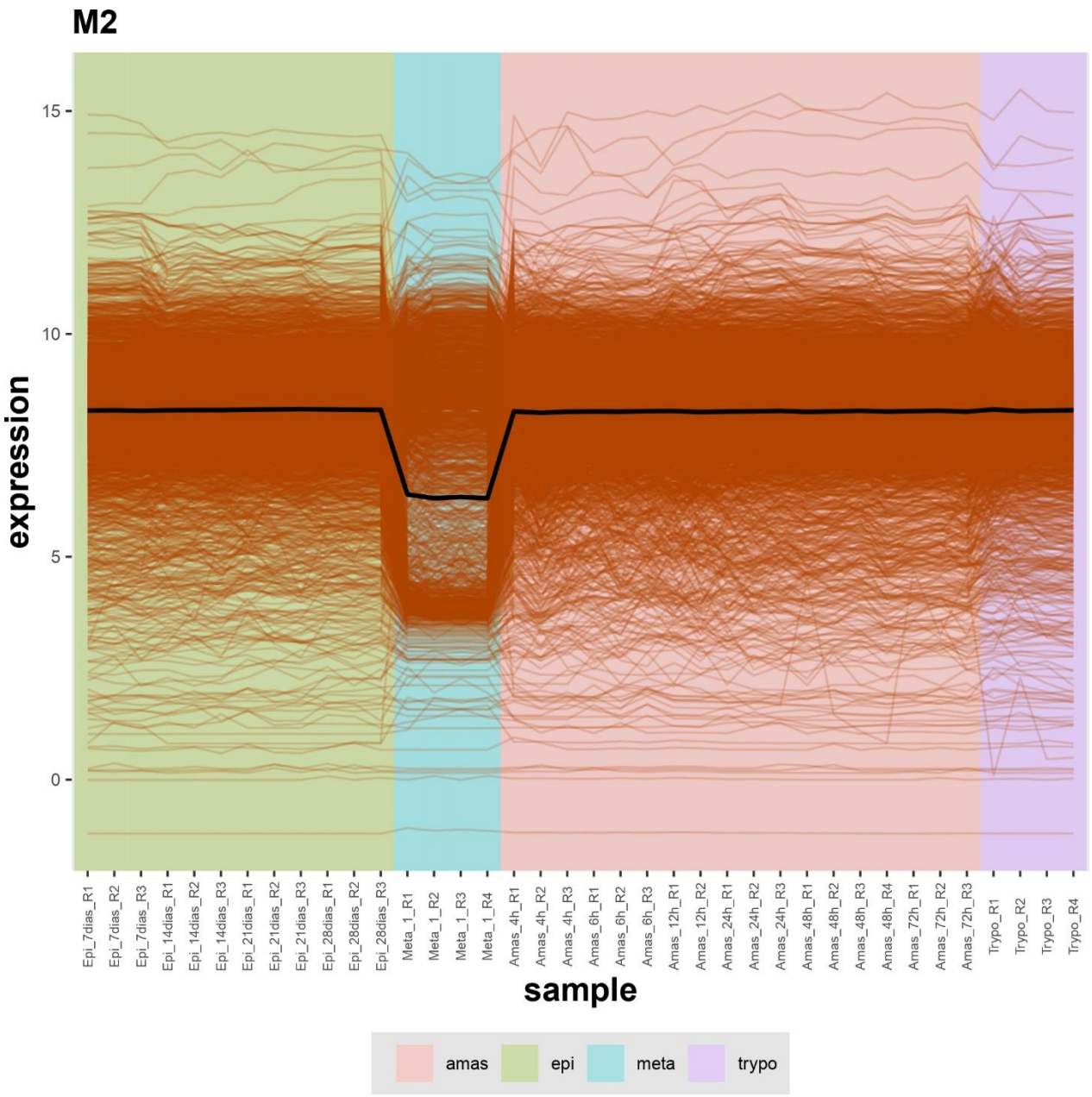
Zinoviev, A., & Shapira, M. (2012). Evolutionary conservation and diversification of the translation initiation apparatus in trypanosomatids. In *Comparative and Functional Genomics* (Vol. 2012). https://doi.org/10.1155/2012/813718

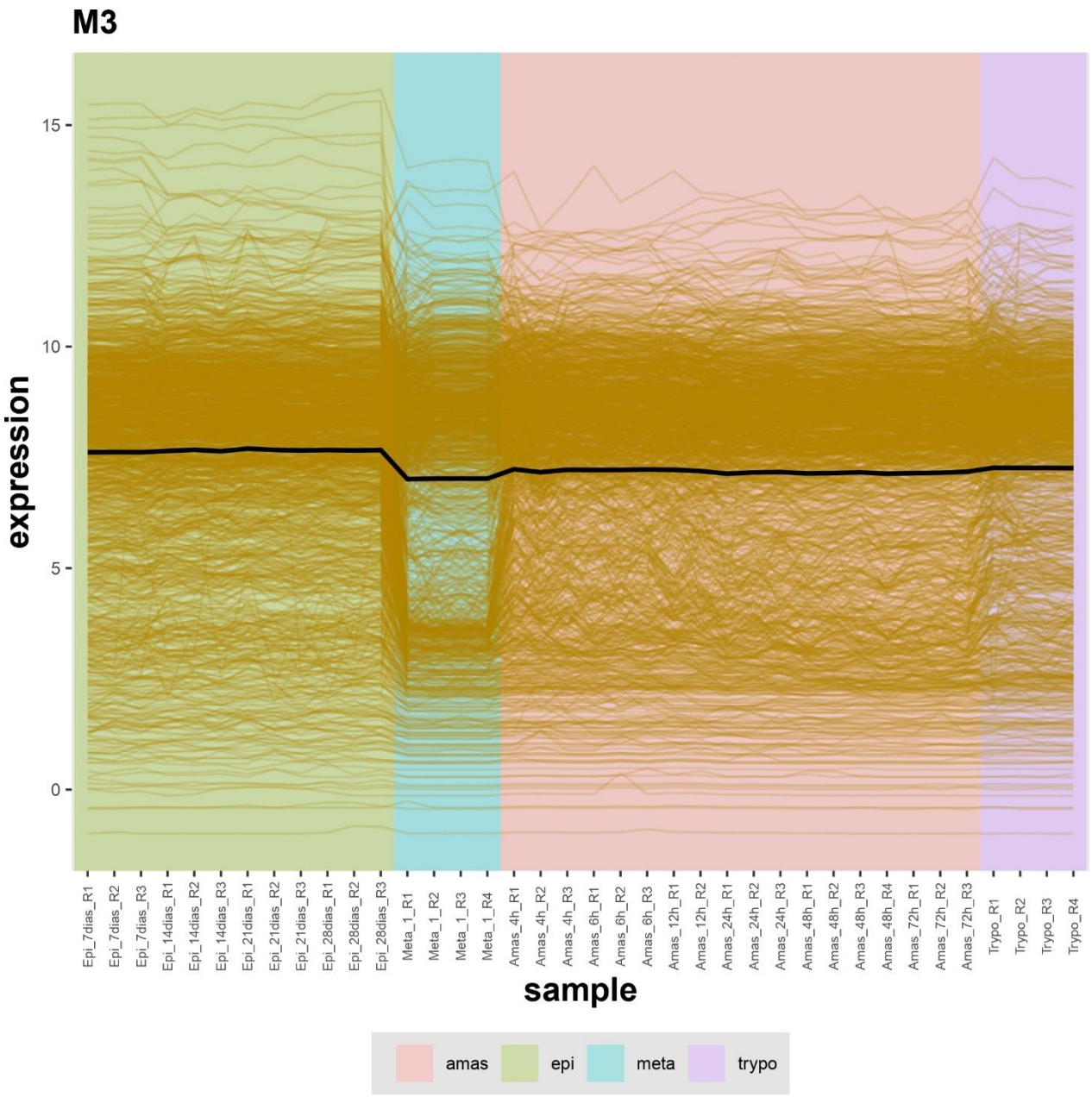
# *8* Anexo

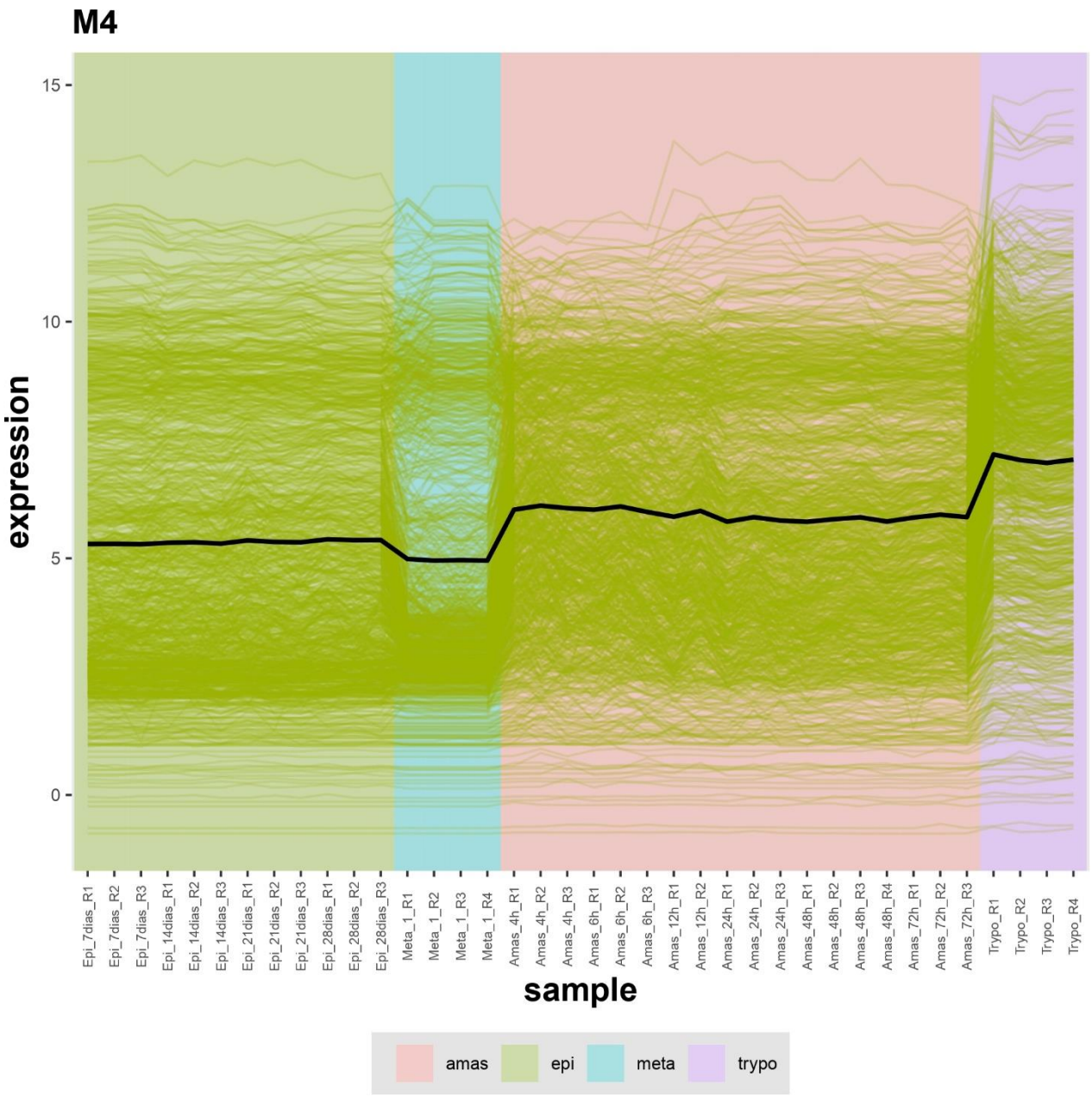
**Figura Suplementaria 1.**

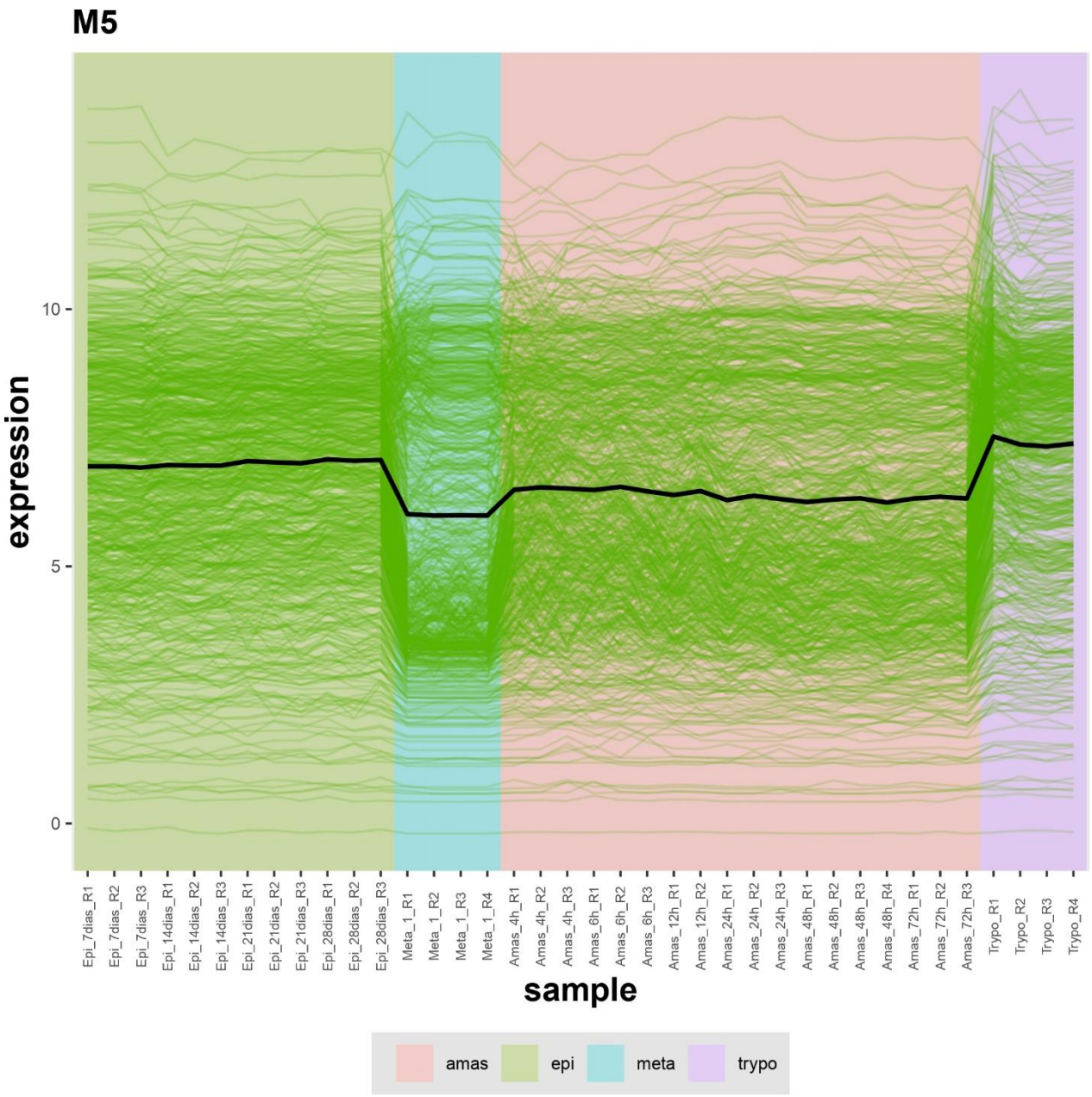
*Time-plot analysis* de cada uno de los 14 módulos identificados a partir de la red de co-expresión génica generada con CEMiTool.

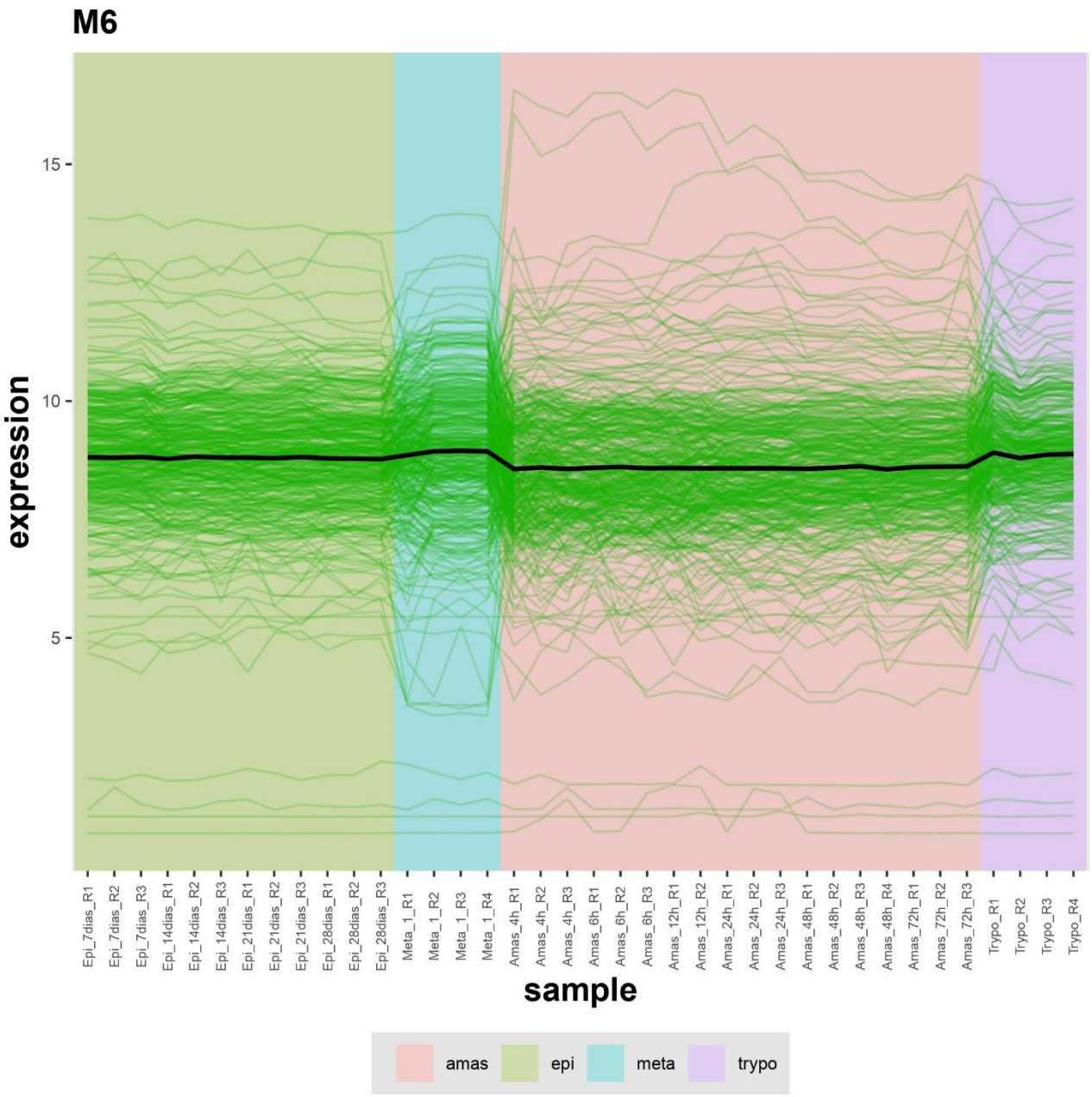


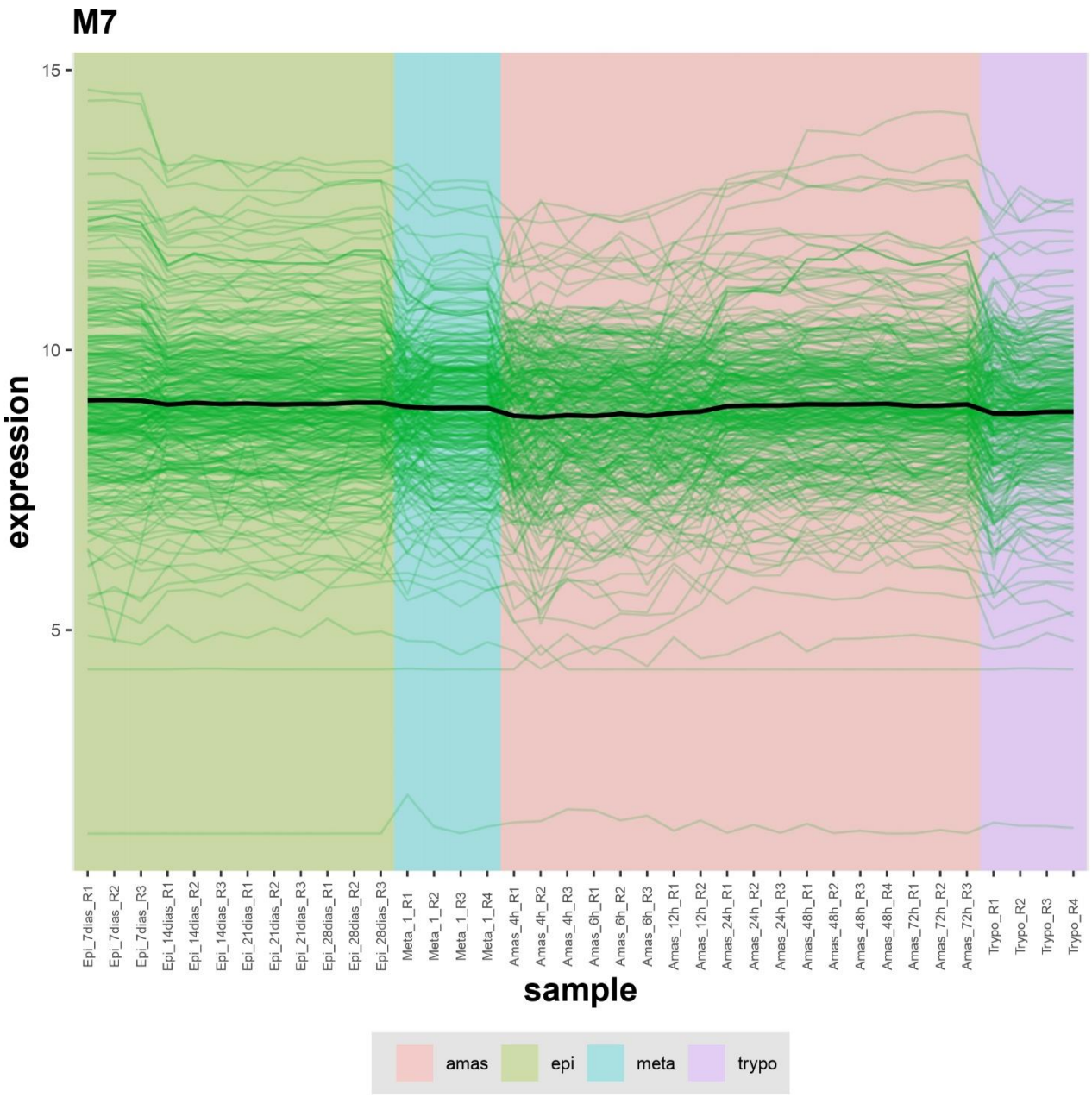

M2
expression
15
10
5
0
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo

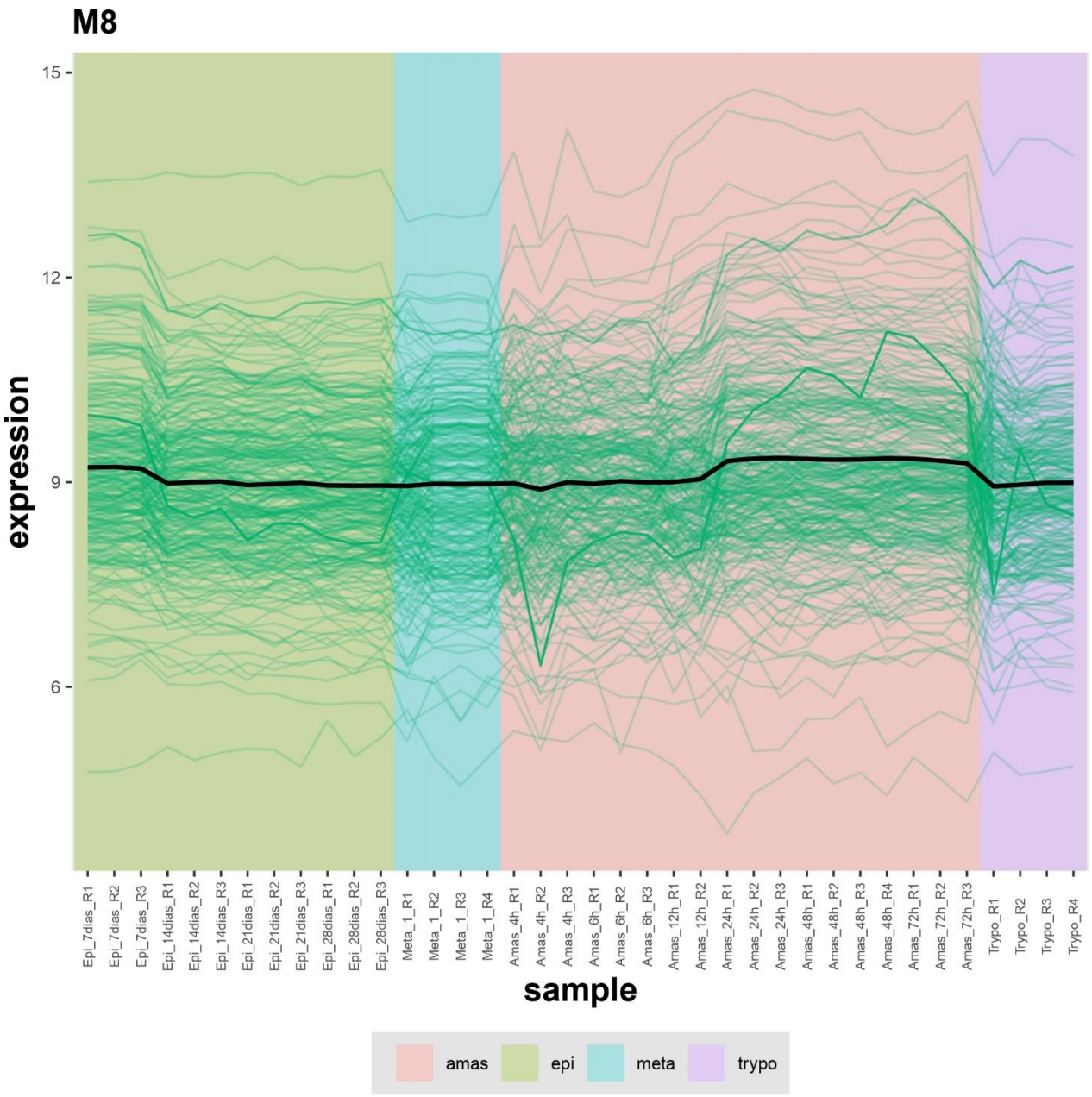

M3
expression
15
10
5
0
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo

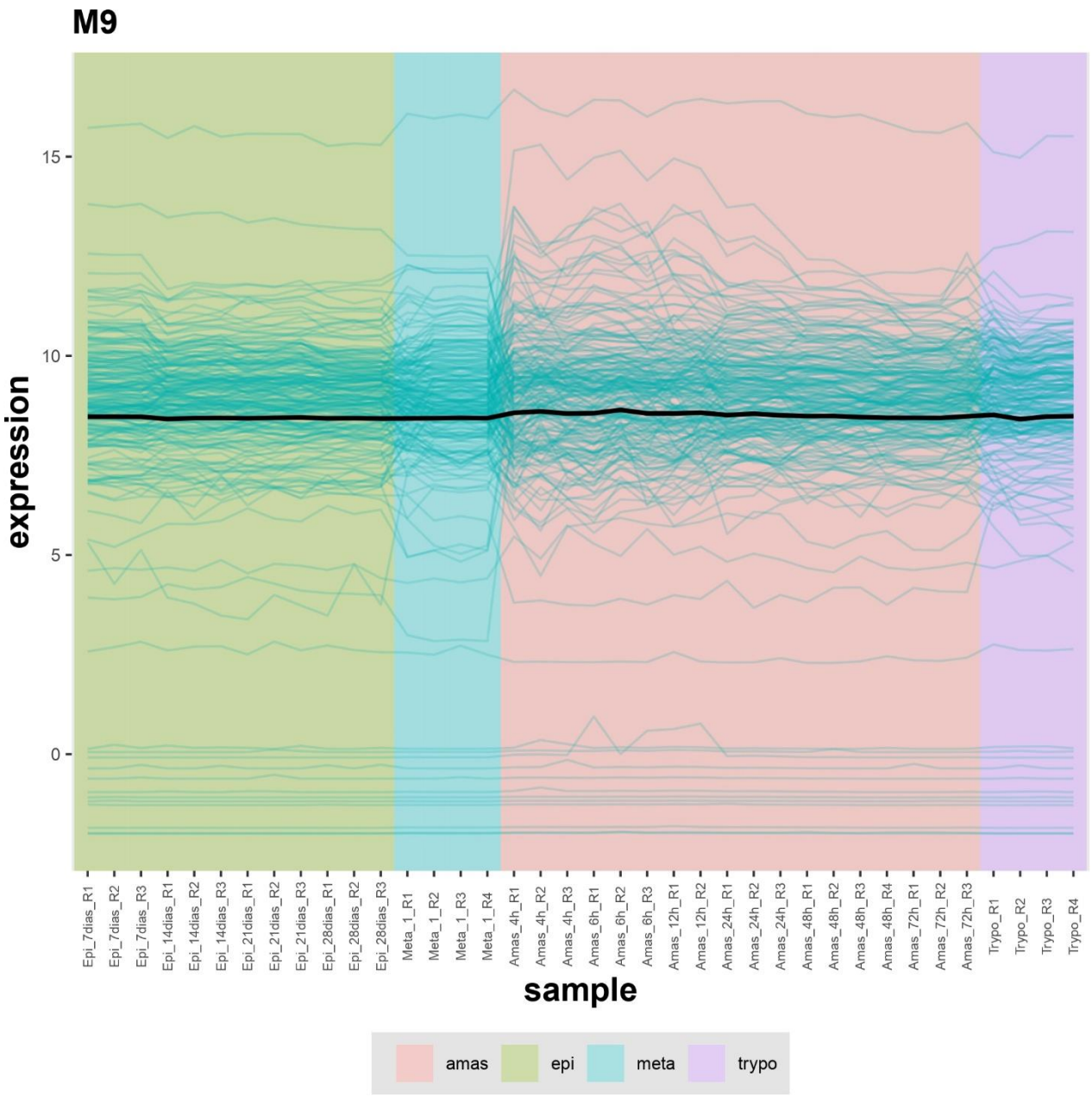

M4
expression
15
10
5
0
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo

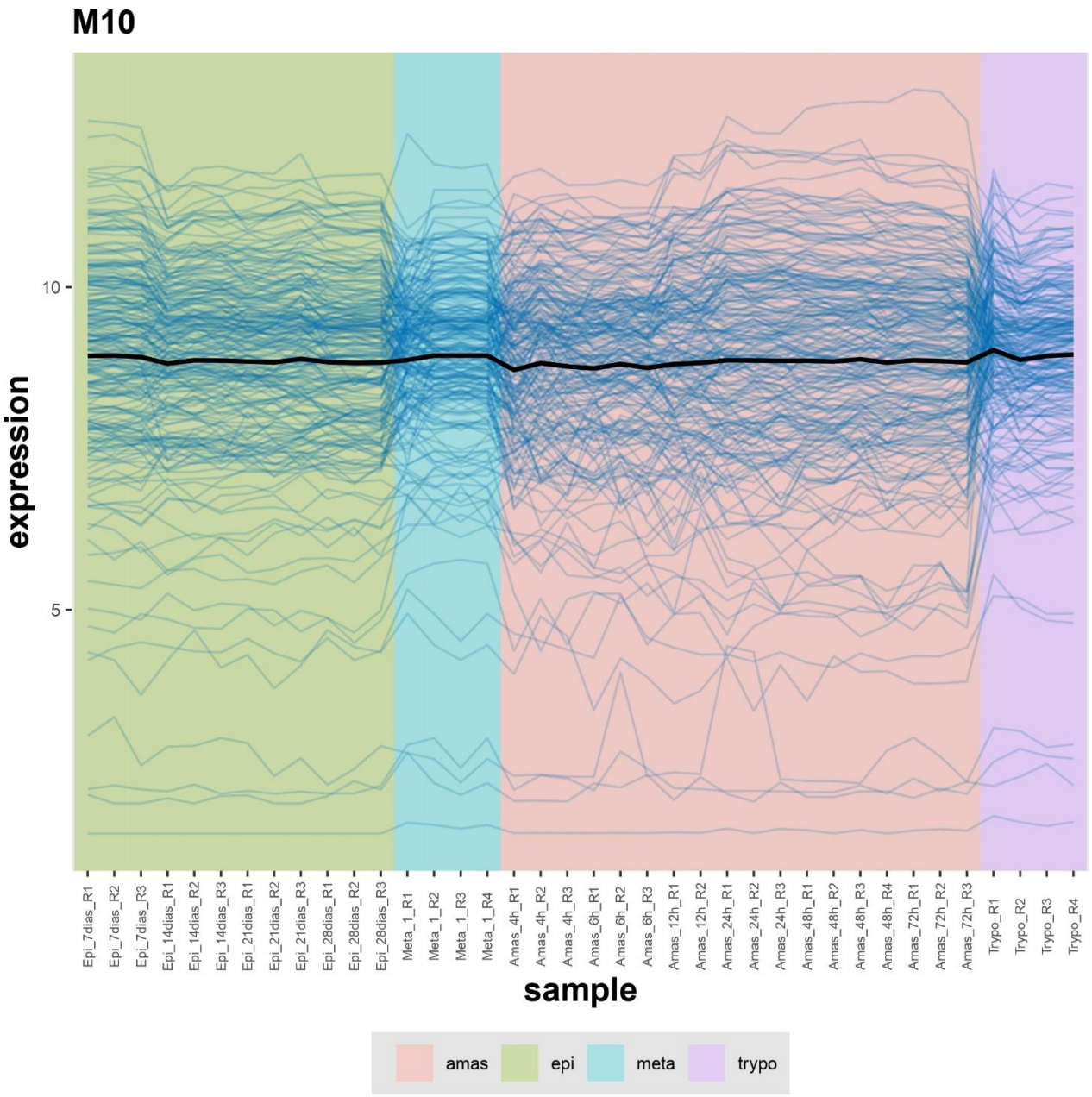

M5
expression
10
5
0
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo

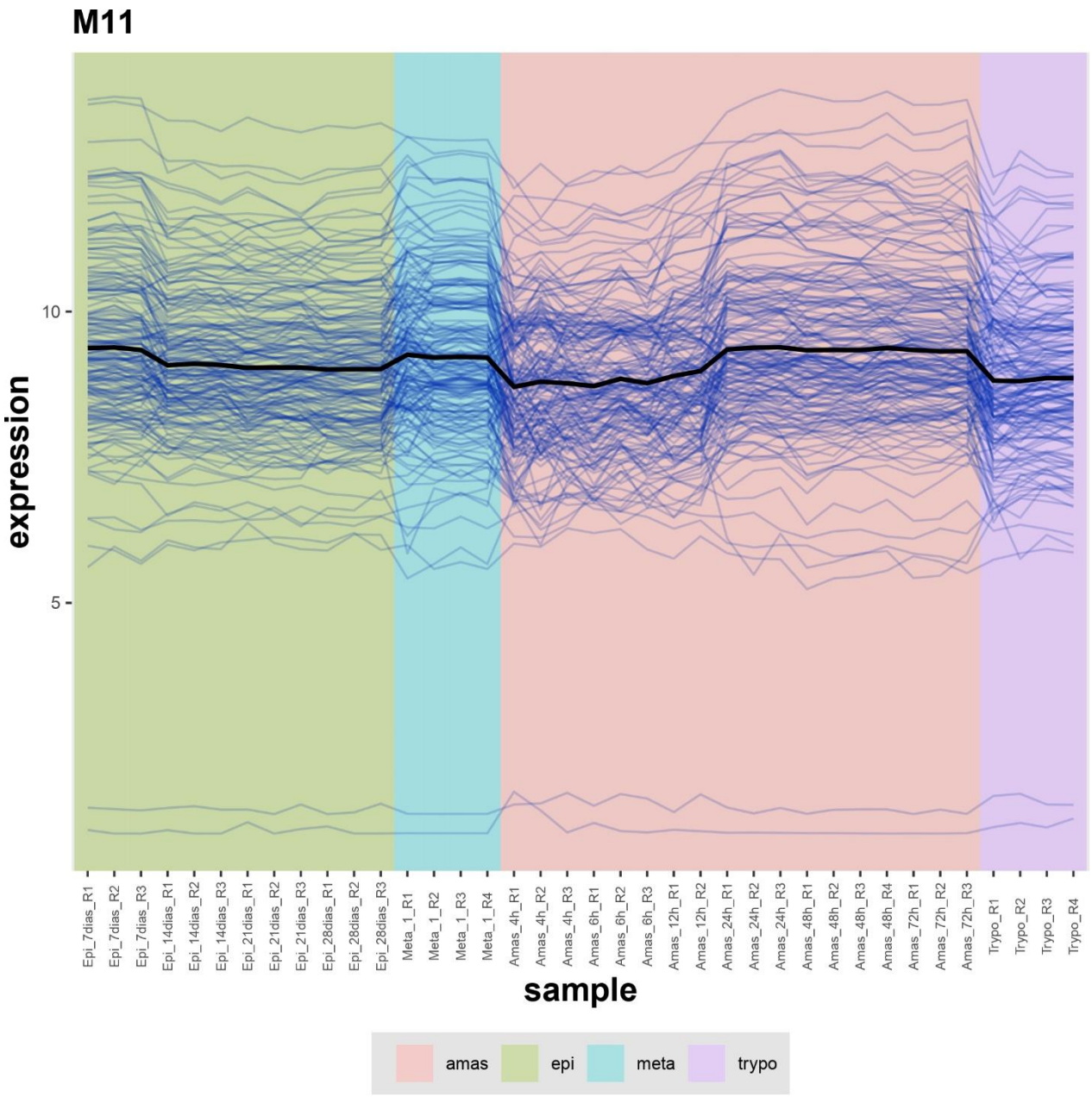

M6
expression
15
10
5
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo

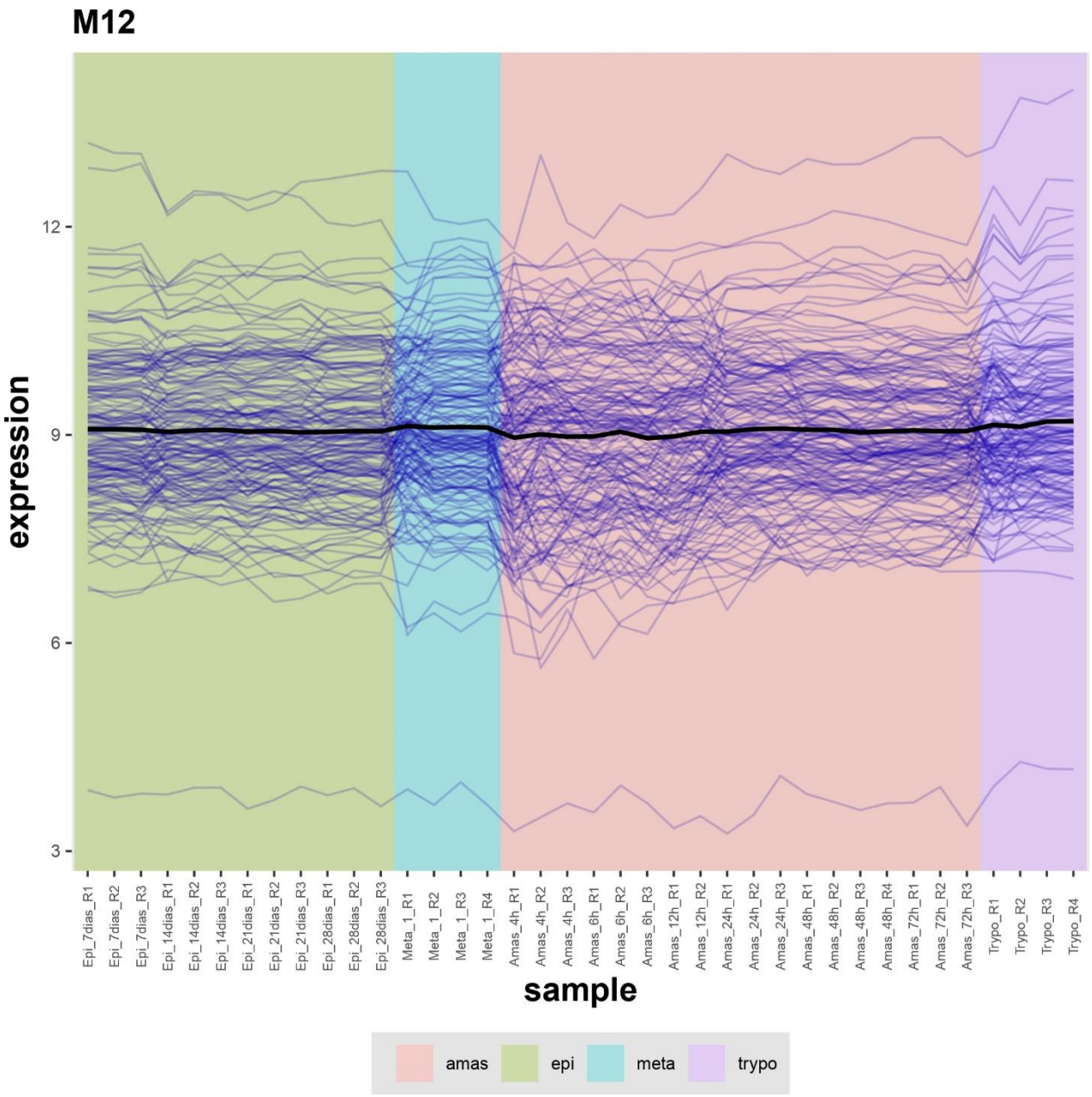

M7
15
10
5
expression
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo

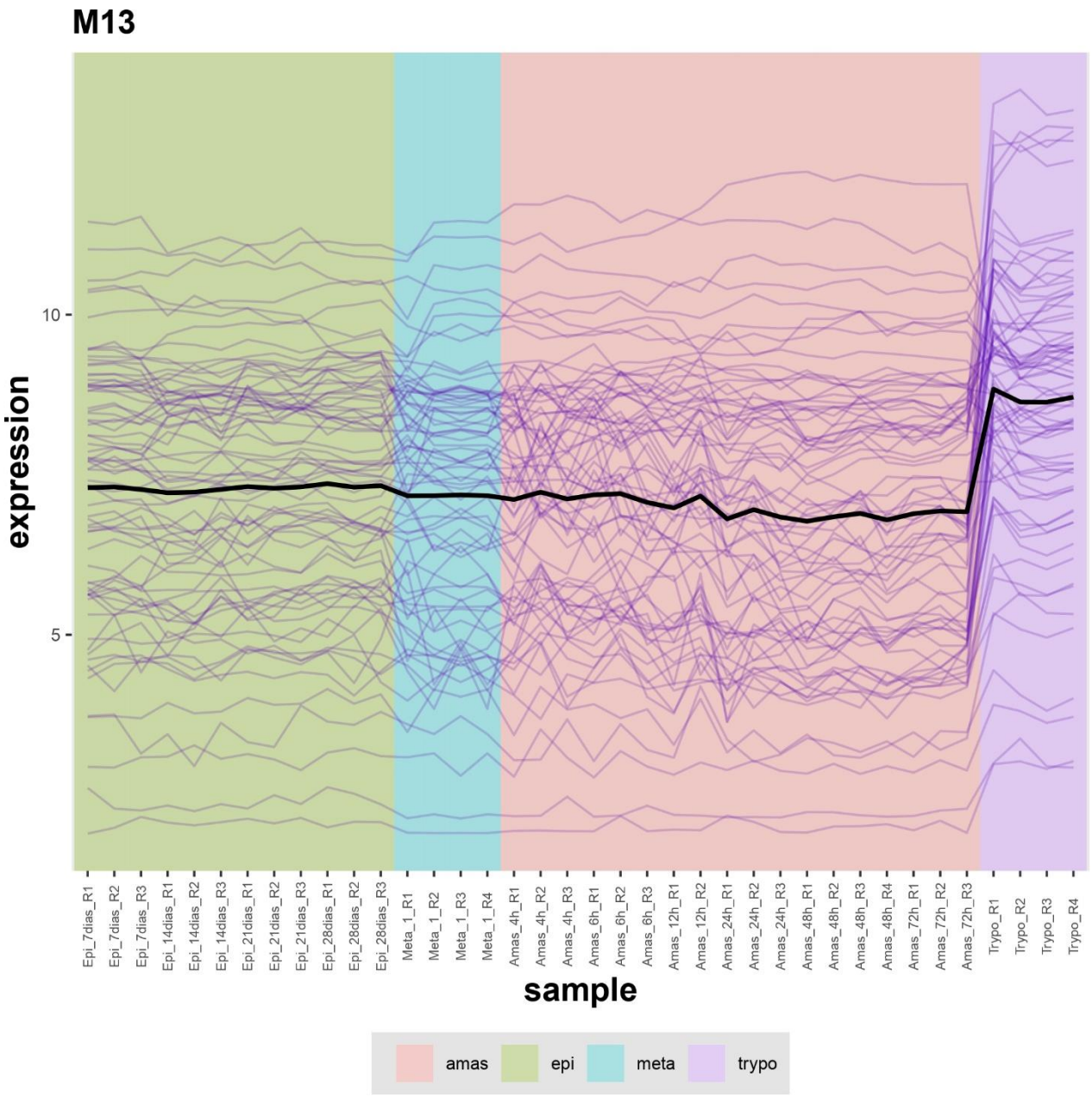

M8
expression
15
12
9
6
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo


M9
expression
15
10
5
0
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo


M10
expression
10
5
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo


M11
expression
10
5
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo


M12
expression
12
9
6
3
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo


M13
expression
10
5
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo

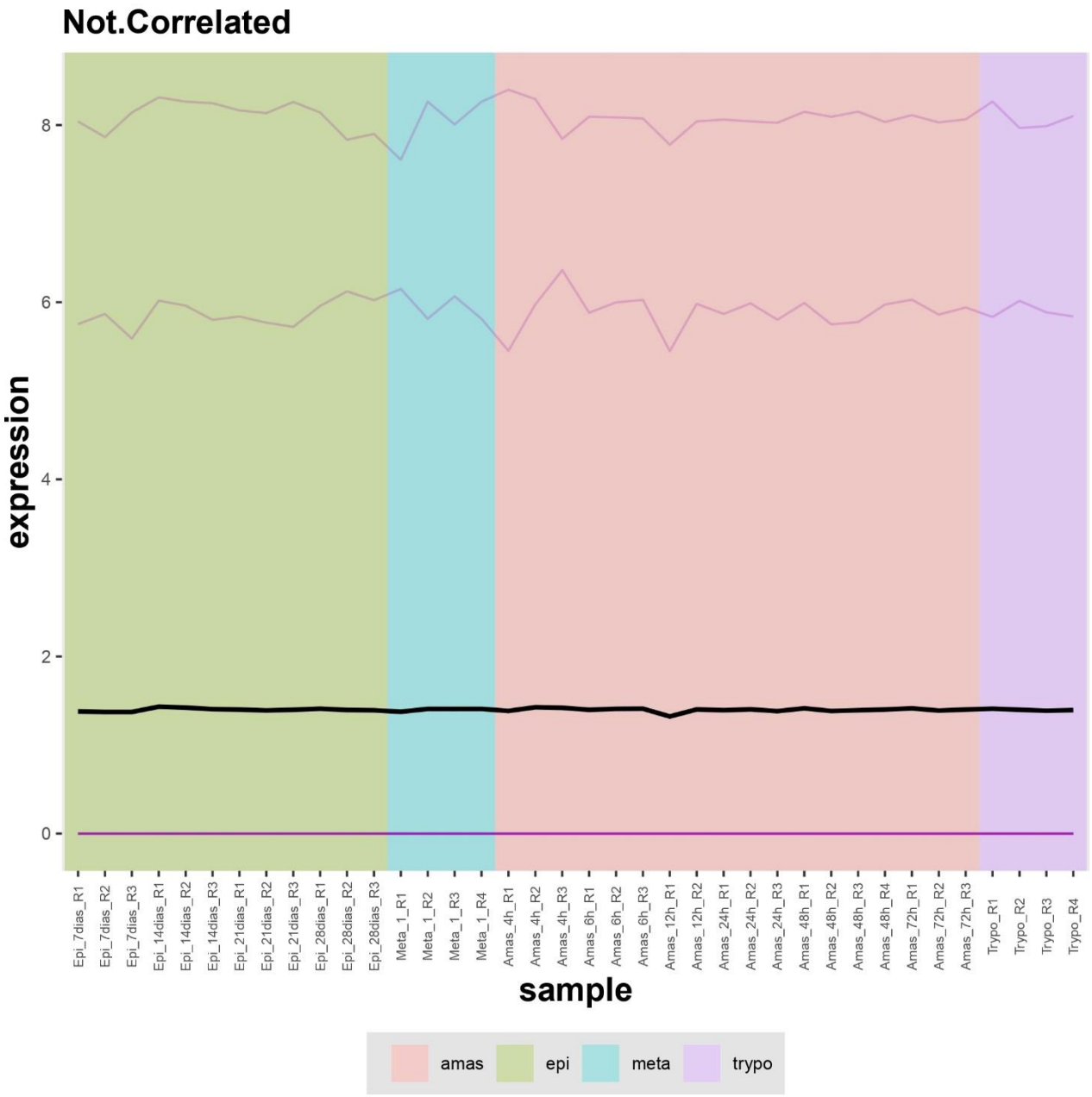

Not.Correlated
expression
8
6
4
2
0
Epi_7dias_R1
Epi_7dias_R2
Epi_7dias_R3
Epi_14dias_R1
Epi_14dias_R2
Epi_14dias_R3
Epi_21dias_R1
Epi_21dias_R2
Epi_21dias_R3
Epi_28dias_R1
Epi_28dias_R2
Epi_28dias_R3
Meta_1_R1
Meta_1_R2
Meta_1_R3
Meta_1_R4
Amas_4h_R1
Amas_4h_R2
Amas_4h_R3
Amas_6h_R1
Amas_6h_R2
Amas_6h_R3
Amas_12h_R1
Amas_12h_R2
Amas_24h_R1
Amas_24h_R2
Amas_24h_R3
Amas_48h_R1
Amas_48h_R2
Amas_48h_R3
Amas_48h_R4
Amas_72h_R1
Amas_72h_R2
Amas_72h_R3
Trypo_R1
Trypo_R2
Trypo_R3
Trypo_R4
sample
amas
epi
meta
trypo

**Figura Suplementaria 2.**

*Pipeline* del *software* exp2GO para la inferencia de términos GO. A) Datos de expresión génica. B) Matriz de distancia de expresión entre genes. C) Información de términos GO de genes anotados. D) Matriz de distancia semántica entre genes anotados. E) y F) exp2GO completa D usando B y D. G) Matriz de distancia semántica reconstruída y asignación de términos GO.

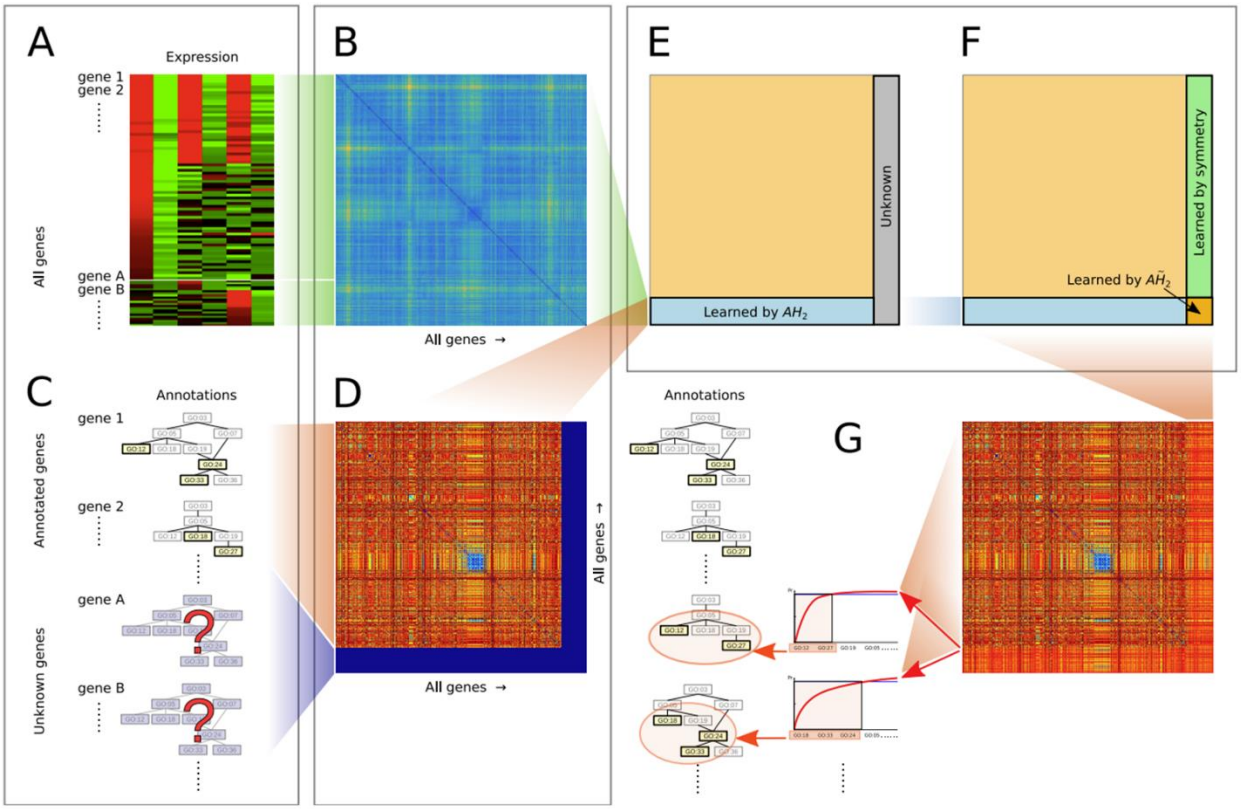


**Tabla Suplementaria 1.**

IUPAC *nucleotide code*

| IUPAC nucleotide code | Base |
|---|---|
| A | Adenine |
| C | Cytosine |
| G | Guanine |
| T (or U) | Thymine (or Uracil) |
| R | A or G |
| Y | C or T |
| S | G or C |
| W | A or T |
| K | G or T |
| M | A or C |
| B | C or G or T |
| D | A or G or T |
| H | A or C or T |
| V | A or C or G |
| N | any base |
| . or - | gap |

**Tabla Suplementaria 2.**

Motivos de secuencia primaria identificados en las regiones 3'UTR de los genes de cada uno de los módulos con sobrerrepresentación de términos GO utilizando *XSTREME*.

| Módulo | Motivo | # Genes | SEA p-value | e-value |
|---|---|---|---|---|
| **M1** | DUURUURUURUURUU | 924 | 0.0383 | 1.40E-22 |
| **M4** | CCGCUGCCSUSCUGC | 430 | 2.82E-131 | 1.40E-180 |
| | CCASSAMKCAMMCAC | 362 | 5.45E-117 | 6.90E-126 |
| | GCUGCCGUSCUGYS | 349 | 1.45E-115 | 2.29E-11 |
| | AGWGUGUGUGCGCUG | 362 | 4.91E-113 | 2.80E-104 |
| | CRCASACACACUCRU | 433 | 7.89E-108 | 2.60E-176 |
| | ACACACUCRUM | 376 | 5.59E-107 | 1.65E-13 |
| | SUGUGKYCCSWUGCA | 272 | 3.56E-105 | 1.40E-100 |
| | UGCAUGCACACCSRU | 303 | 7.63E-102 | 3.40E-109 |
| | UGCAUGCACACYS | 289 | 6.12E-101 | 8.27E-11 |
| | UGUUUUGCUUYAYKK | 335 | 3.34E-99 | 9.87E-10 |
| | UKUGUKUUUUGCUUU | 649 | 8.94E-97 | 1.10E-114 |
| | GUGUGYGCKGGYYGU | 291 | 7.78E-94 | 1.21E-11 |
| | GCCMGCAC | 307 | 3.93E-85 | 0.000135 |
| | AGCACACA | 334 | 4.67E-85 | 4.51E-07 |
| | UUUGCACSACACGCA | 232 | 5.70E-83 | 9.20E-93 |
| | UGCACCGCS | 288 | 1.66E-82 | 1.10E-08 |

| | | | |
|---|---|---|---|
| UUGCAUGGACUCWCG | 262 | 7.74E-82 | 1.14E-13 |
| CSACACGCAC | 281 | 7.89E-77 | 7.27E-07 |
| GAGCCGUGCACCGCC | 276 | 2.08E-76 | 4.30E-115 |
| ACCUCUCUCUCYCUC | 269 | 7.16E-76 | 2.50E-92 |
| CCCGUUGGUGCUCUC | 242 | 9.75E-76 | 1.60E-98 |
| CUGGCUGCAUGGGCG | 272 | 1.09E-75 | 1.80E-92 |
| CACGUUGAACGCAUC | 143 | 1.55E-75 | 0.000717 |
| UAAYUGUUUY | 269 | 1.82E-75 | 9.58E-10 |
| GCAUGGGCGGAGCA | 223 | 3.65E-73 | 2.88E-10 |
| CAUAAUGUGCCC | 221 | 1.16E-72 | 1.17E-12 |
| GAGCAAUGA | 249 | 2.07E-71 | 3.95E-08 |
| UUUYYAUUGCAUUG | 215 | 3.90E-70 | 7.08E-09 |
| CGUCACGUUGAACGC | 116 | 2.12E-69 | 1.40E-70 |
| ACGCUUCGGC | 259 | 1.43E-67 | 9.26E-08 |
| GUCGUCCCACGCUUC | 180 | 3.32E-67 | 2.50E-55 |
| GUUGGUGCUCUCUUU | 211 | 3.04E-66 | 2.90E-10 |
| AGUGGGCGACCU | 215 | 6.71E-66 | 6.99E-09 |
| UGUGGUCCGAUGUAW | 127 | 2.68E-63 | 0.00028 |
| CGGGAGCAAUG | 281 | 5.58E-62 | 4.70E-43 |
| CCCUCACCUCAY | 198 | 6.48E-62 | 5.52E-11 |
| UGCAUUGAGUGGGCG | 150 | 4.03E-60 | 2.80E-61 |
| CUCUCGCCCUCACCU | 252 | 3.39E-58 | 9.30E-100 |
| UGCAGGCCC | 237 | 8.14E-57 | 6.89E-07 |
| AGCACCGUAAC | 175 | 6.62E-56 | 6.20E-41 |
| GCUCCGCGUUGYUBC | 109 | 8.27E-56 | 0.0224 |
| AUAAUGUGCCCAUUG | 173 | 3.90E-55 | 1.50E-83 |
| CCRCGGGGAYWUGUG | 124 | 7.64E-53 | 0.00298 |
| GUGUGUGUCCC | 251 | 5.91E-51 | 4.04E-08 |
| AUUGUAUAUA | 169 | 2.16E-50 | 4.13E-07 |
| CUCUCCCUGUG | 235 | 1.27E-49 | 5.55E-09 |
| UGUGUGYGUGY | 416 | 3.24E-49 | 1.80E-34 |
| GAGAGAGCC | 260 | 1.18E-48 | 1.44E-06 |
| GUGAGGGAGAG | 276 | 1.57E-47 | 3.40E-31 |
| CACUCACUCU | 190 | 2.04E-44 | 4.42E-05 |
| AAGUAAGAGUAAUUA | 34 | 2.98E-42 | 3.30E-05 |
| GAGUGAGGGA | 173 | 2.43E-40 | 3.31E-08 |
| CYUUCACUCACUCUC | 160 | 1.39E-37 | 3.10E-64 |
| AAAAUAAUGAUUG | 84 | 8.20E-37 | 0.00208 |
| CGGCCCUG | 102 | 2.62E-32 | 0.0403 |
| ACGCAUGAGGA | 139 | 2.95E-32 | 0.017 |
| GAGGGGCCGCGACGA | 117 | 1.49E-31 | 6.80E-41 |
| NUGCAUGV | 303 | 3.56E-31 | 1.19E-28 |
| CAGACACCCGCAGC | 117 | 8.91E-31 | 0.0323 |
| CUGUGGGGCG | 138 | 2.20E-28 | 0.00255 |
| CACAGCCACACC | 110 | 1.24E-26 | 0.0215 |

| | | | |
|---|---|---|---|
| CCUGUGGGGCGGAGC | 100 | 3.50E-26 | 3.60E-28 |
| BGCUGGCC | 204 | 3.53E-26 | 1.18E-23 |
| CGGCCAGGGCGGGCA | 87 | 1.49E-25 | 2.20E-17 |
| DWGCACA | 246 | 9.09E-25 | 3.03E-22 |
| UGUGUUUCCCGCCCA | 86 | 1.22E-24 | 1.00E-25 |
| CACSCACACAA | 123 | 1.69E-24 | 4.00E-14 |
| BUGCHGCUSCUUSUK | 302 | 1.88E-24 | 1.10E-05 |
| CCGCACUCACRCGCC | 58 | 5.39E-24 | 7.50E-07 |
| WAGCACAN | 162 | 2.08E-23 | 6.96E-21 |
| UUUUWAUUKUUUYUU | 191 | 4.21E-23 | 1.00E-08 |
| GCAAUCACUAUGGAC | 62 | 7.76E-23 | 1.60E-18 |
| ACGGCGCGCUG | 89 | 1.45E-22 | 1.20E-09 |
| CCACGA | 258 | 2.50E-22 | 0.00702 |
| CAUWGUD | 172 | 2.53E-22 | 8.45E-20 |
| BGCUGGCC | 179 | 3.15E-22 | 1.05E-19 |
| GCUGGMC | 175 | 1.14E-21 | 3.80E-19 |
| UUUCUUUCCCU | 90 | 1.30E-21 | 0.00604 |
| GUGUGUG | 319 | 2.22E-21 | 7.43E-19 |
| DWGCACA | 277 | 5.06E-21 | 1.69E-18 |
| ACGCGACUCCCGUCA | 55 | 5.86E-21 | 2.30E-08 |
| UUGCACA | 255 | 6.06E-21 | 2.02E-18 |
| AACACGC | 391 | 2.33E-20 | 7.77E-18 |
| AAUAARARAAAAWAA | 262 | 3.15E-20 | 1.70E-12 |
| UGCAUGM | 387 | 3.91E-20 | 1.31E-17 |
| UGUUUYUGUUUKKUU | 784 | 4.41E-20 | 8.30E-102 |
| UGCGACGAGG | 103 | 1.22E-19 | 0.0465 |
| WGCAUGM | 393 | 1.34E-19 | 4.47E-17 |
| GACGACGGCCCUGUG | 166 | 1.38E-19 | 7.00E-33 |
| CAGUCCCCACACACA | 38 | 3.77E-19 | 2.00E-08 |
| CYWGCACA | 277 | 4.91E-19 | 1.64E-16 |
| CSWCGSCWGCYKCAS | 57 | 5.52E-19 | 1.50E-07 |
| CCUGUGAGGGGUGCC | 85 | 5.82E-19 | 4.30E-22 |
| GUGUGUG | 316 | 4.45E-18 | 1.49E-15 |
| ACACACA | 380 | 6.59E-18 | 2.20E-15 |
| CCCUGCCGCCGUGKU | 111 | 9.14E-18 | 3.50E-23 |
| SSGCGCS | 400 | 2.05E-17 | 6.84E-15 |
| UUGCGGCCAGG | 101 | 2.59E-17 | 0.0199 |
| ACGAUGUGCGGACAC | 55 | 6.69E-16 | 1.10E-07 |
| WGCAUGM | 115 | 9.95E-16 | 3.32E-13 |
| UKUGUGU | 280 | 5.88E-15 | 1.96E-12 |
| UGUGUGU | 271 | 1.02E-14 | 3.42E-12 |
| UGUGUGU | 273 | 8.01E-14 | 2.67E-11 |
| ACAMMCACARCMACA | 215 | 1.62E-13 | 1.40E-44 |
| WGCAUGA | 472 | 8.29E-13 | 2.77E-10 |
| AGGGGAGCAGCAACA | 58 | 1.50E-12 | 2.10E-05 |

| | | | | |
|---|---|---|---|---|
| | UGUGUGU | 279 | 3.00E-12 | 1.00E-09 |
| | SCAGCGC | 110 | 3.40E-12 | 1.13E-09 |
| | HUUUCCCU | 88 | 4.18E-12 | 1.40E-09 |
| | UGUGUGU | 319 | 6.09E-12 | 2.03E-09 |
| | GCGCGSG | 200 | 1.45E-11 | 4.86E-09 |
| | ACACAWA | 240 | 2.19E-11 | 7.30E-09 |
| | UGCGCGC | 177 | 8.47E-11 | 2.83E-08 |
| | UUGCACR | 253 | 1.14E-10 | 3.80E-08 |
| | ACGACGA | 150 | 2.30E-10 | 7.69E-08 |
| | GACGACGR | 142 | 3.20E-10 | 1.07E-07 |
| | UGUAYAK | 315 | 1.86E-09 | 6.20E-07 |
| | HGAACGM | 169 | 5.58E-09 | 1.86E-06 |
| | GGCCACACGGCAGCC | 23 | 7.99E-09 | 0.0011 |
| | GACGACM | 156 | 1.06E-07 | 3.55E-05 |
| | UGUUSGU | 265 | 1.43E-07 | 4.79E-05 |
| | UGUUUUK | 274 | 3.02E-07 | 0.000101 |
| | UGUGUGUK | 408 | 7.55E-07 | 0.000252 |
| | UUUUUYUUUWUUDYU | 545 | 8.65E-07 | 2.10E-58 |
| | UUACAUR | 256 | 1.21E-06 | 0.000404 |
| | GWGUGUGD | 316 | 1.59E-06 | 0.000531 |
| | NCGCGCGG | 192 | 2.82E-06 | 0.000941 |
| | GCGCGGG | 230 | 2.98E-06 | 0.000994 |
| | UUUUUUK | 218 | 3.85E-06 | 0.00129 |
| | ACACACA | 343 | 4.33E-06 | 0.00145 |
| | AGAYASAK | 486 | 9.38E-06 | 0.00313 |
| | UGUACAK | 403 | 1.37E-05 | 0.00459 |
| | ACUAAWC | 15 | 1.72E-05 | 0.00573 |
| | GCGCGCG | 237 | 3.29E-05 | 0.011 |
| | WGUGUGA | 159 | 4.24E-05 | 0.0142 |
| | ACACACA | 363 | 5.11E-05 | 0.0171 |
| | UAGUWRG | 11 | 6.26E-05 | 0.0209 |
| | UAGUWRG | 11 | 6.26E-05 | 0.0209 |
| | UGUACAK | 404 | 6.52E-05 | 0.0218 |
| | NGCUUGC | 489 | 6.75E-05 | 0.0226 |
| | KUAAUUS | 522 | 8.07E-05 | 0.027 |
| **M5** | GGCGGAGCACAS | 208 | 2.76E-66 | 8.30E-05 |
| | CUCGCACACCCAC | 210 | 6.07E-66 | 0.00111 |
| | SWKGGGCGGAGCACM | 216 | 1.26E-65 | 1.80E-151 |
| | CCRCYCGCACACMCA | 269 | 3.19E-64 | 6.20E-222 |
| | CKGGYMGWCMCMCGC | 208 | 1.65E-62 | 1.40E-122 |
| | GSWGAGRGAGCCGUG | 205 | 3.77E-62 | 6.70E-133 |
| | CUGCCGCCSUG | 252 | 7.24E-62 | 0.000146 |
| | CGACACGCWCM | 206 | 8.18E-61 | 0.00151 |
| | MRCGMMRCACACACA | 213 | 1.89E-60 | 3.30E-174 |
| | CKCSUSCUKCACKCA | 226 | 6.65E-60 | 2.80E-115 |

| | | | |
|---|---|---|---|
| CYGUGUGUKUCCCSY | 251 | 2.14E-59 | 5.80E-133 |
| AYGUGUGKGYGUGCG | 479 | 2.38E-58 | 7.90E-102 |
| CACRCYCAUGMMGRC | 216 | 1.43E-57 | 4.00E-144 |
| MCACRCMCYGCYGCC | 217 | 1.17E-55 | 9.80E-163 |
| MKUUSKUKCYCUCUU | 217 | 6.65E-55 | 7.50E-93 |
| CWCUCUYCCUGUGWG | 217 | 2.43E-54 | 3.90E-154 |
| UUUUGCUUUWSGGG | 199 | 3.49E-53 | 0.00509 |
| GUGUGGGCGUS | 201 | 1.67E-51 | 0.00356 |
| CCGCACUCACVCGCS | 120 | 5.42E-50 | 2.10E-39 |
| ACUCUCSC | 213 | 8.64E-50 | 0.00173 |
| CCCUGUGCGCC | 116 | 4.69E-49 | 2.10E-51 |
| GGGUGCCGUGUGUUU | 127 | 5.25E-49 | 0.00621 |
| CGACGGCCCUGUGC | 119 | 1.20E-48 | 0.00153 |
| GCCCAACUGCUC | 151 | 3.00E-48 | 0.00496 |
| GACACCCGCAGCGC | 116 | 3.16E-48 | 0.00253 |
| AUGAGGAGGGGCC | 119 | 2.87E-47 | 0.00253 |
| ACACAGCCACA | 135 | 1.70E-45 | 0.000177 |
| UUGCGGCCAGGGC | 119 | 7.46E-45 | 0.000883 |
| CGUSCUUCWCUCAC | 171 | 8.55E-45 | 0.00152 |
| CUGCGGGCCGCAY | 102 | 3.43E-44 | 0.0256 |
| CUGCCUGGCUG | 212 | 2.67E-43 | 4.90E-46 |
| CAAUCACUAUGGACG | 113 | 3.68E-43 | 5.50E-68 |
| GCCUCGCCUGCUGC | 119 | 4.62E-43 | 0.00253 |
| AAUCACUAUGGACGU | 114 | 8.26E-43 | 0.00403 |
| MCCCAUGMA | 158 | 2.05E-42 | 0.00392 |
| UUUUAAUUGUUUYUU | 207 | 6.78E-42 | 2.10E-50 |
| UUCCCUCUUGUUUU | 115 | 9.37E-41 | 0.00403 |
| ACGAACGGCGC | 125 | 3.74E-40 | 0.00402 |
| CGAGGGCC | 110 | 4.26E-40 | 0.0063 |
| CRCCRMCAMACACAC | 148 | 2.53E-39 | 1.40E-09 |
| ACUCUUCCUGUG | 118 | 2.98E-39 | 0.0114 |
| GAGRGAGCCGUG | 199 | 4.92E-39 | 0.00508 |
| CCCYUGCAUGG | 164 | 6.20E-38 | 0.00769 |
| GACGAGGGCCUGUGG | 121 | 1.93E-36 | 0.0135 |
| GCCGCGACGAACGGC | 112 | 2.56E-36 | 5.70E-87 |
| UCUUUUGCA | 182 | 1.40E-35 | 0.00456 |
| UUUUUGCUUUUGGGG | 78 | 1.80E-35 | 1.10E-39 |
| CUGGCUGC | 190 | 3.09E-35 | 0.0287 |
| GCCCGUUGGUGCUC | 129 | 1.65E-34 | 0.000157 |
| GCGCUGGUCGUCCCA | 137 | 2.76E-33 | 0.0136 |
| ACCCCAYGAAR | 157 | 1.51E-32 | 0.00108 |
| ACUCUCGCCCUCACC | 194 | 7.51E-32 | 9.00E-128 |
| GACGACGR | 123 | 9.89E-30 | 3.25E-27 |
| CACACACAUAAU | 118 | 7.52E-29 | 0.000978 |
| AAUUGUUUCUU | 165 | 2.23E-28 | 0.00142 |

| | | | |
|---|---|---|---|
| GUAACUGUUUC | 69 | 8.93E-28 | 3.10E-07 |
| AAUGUGCCCAUUGUA | 109 | 9.29E-28 | 8.40E-85 |
| GUGCCCAUUGU | 131 | 9.35E-27 | 0.00284 |
| CCUCACCUC | 116 | 1.12E-26 | 0.000281 |
| UGAGUGGGCGACCUC | 121 | 1.23E-26 | 5.30E-92 |
| CCAACUGC | 105 | 2.51E-26 | 0.022 |
| CUCCCUGUGUGUG | 108 | 8.77E-25 | 0.0256 |
| SCAGCGC | 125 | 3.35E-24 | 1.10E-21 |
| HUUUCCCU | 97 | 4.66E-24 | 1.53E-21 |
| UUUUUYUUUUYUWYU | 521 | 3.11E-23 | 2.70E-100 |
| ACCCCAUGAAAUAUA | 71 | 7.17E-23 | 2.20E-25 |
| UGCUUUACGGGAGCA | 95 | 8.48E-23 | 1.90E-66 |
| ACACGAGUGACUGUG | 56 | 2.28E-22 | 2.10E-27 |
| AACAGUAAUGGACGA | 54 | 5.98E-22 | 1.60E-14 |
| CGCUCACUCUCCCGA | 56 | 1.50E-20 | 3.00E-37 |
| ACGACGA | 151 | 3.23E-20 | 1.06E-17 |
| CGAUGUGCGGACACA | 49 | 1.04E-19 | 6.40E-33 |
| AGAGGGCUGAGGGGA | 57 | 1.88E-18 | 2.30E-24 |
| HUCYCYCWUCCCUYY | 137 | 1.95E-18 | 0.00011 |
| GACGACM | 149 | 2.35E-18 | 7.74E-16 |
| ACGAGUGACU | 70 | 3.24E-18 | 0.0209 |
| CUUCUCUCAGUCCCC | 54 | 3.99E-18 | 3.00E-28 |
| UUUUGUUUUUDUUGY | 510 | 2.25E-17 | 1.90E-127 |
| GCAGCAACACGCAGA | 72 | 1.16E-16 | 1.50E-38 |
| GACGCGACUCCCGUC | 55 | 5.98E-16 | 1.00E-30 |
| GCGCGCG | 119 | 3.23E-15 | 1.06E-12 |
| SSGCGCS | 120 | 5.34E-14 | 1.76E-11 |
| GUCAGACGGUUGAAA | 54 | 6.28E-14 | 3.50E-19 |
| UGCGCGC | 140 | 2.03E-13 | 6.69E-11 |
| CAUWGUD | 121 | 6.28E-13 | 2.07E-10 |
| GGCCACACGGCAGCC | 32 | 1.86E-12 | 1.40E-22 |
| DWGCACA | 169 | 2.94E-12 | 9.66E-10 |
| ACACACA | 284 | 2.94E-12 | 9.66E-10 |
| CCCMCACCACCKCC | 38 | 9.34E-12 | 3.20E-09 |
| WAGCACAN | 108 | 3.28E-11 | 1.08E-08 |
| BGCUGGCC | 157 | 3.99E-11 | 1.31E-08 |
| YUGCYGCUGCUKSKG | 115 | 5.18E-11 | 3.70E-10 |
| GCGCGSG | 153 | 9.03E-11 | 2.97E-08 |
| NCGCGCGG | 179 | 1.03E-10 | 3.40E-08 |
| AHGGACA | 143 | 2.11E-10 | 6.93E-08 |
| GCGCGGG | 217 | 3.95E-10 | 1.30E-07 |
| GCUGGMC | 173 | 5.99E-10 | 1.97E-07 |
| BGCUGGCC | 266 | 9.95E-10 | 3.27E-07 |
| AACACGC | 77 | 1.54E-09 | 5.08E-07 |
| DWGCACA | 192 | 2.36E-09 | 7.77E-07 |

| | | | | |
|---|---|---|---|---|
| | GAUGGAGU | 31 | 2.68E-09 | 8.80E-07 |
| | MCMYKCGGUGYCGGC | 84 | 2.68E-09 | 5.50E-18 |
| | UUGCACA | 168 | 1.47E-08 | 4.84E-06 |
| | BGGCGUG | 220 | 1.65E-08 | 5.44E-06 |
| | WGCAUGA | 266 | 2.62E-08 | 8.62E-06 |
| | UUUUUUUUUUUBUUU | 308 | 9.63E-08 | 1.10E-92 |
| | UCGCGCG | 200 | 2.90E-07 | 9.55E-05 |
| | AAYGACRA | 147 | 4.49E-07 | 0.000148 |
| | ACACACA | 265 | 4.82E-07 | 0.000159 |
| | ACACACA | 296 | 5.09E-07 | 0.000167 |
| | UGUUUUK | 174 | 5.66E-07 | 0.000186 |
| | UUUUUUK | 189 | 5.89E-07 | 0.000194 |
| | UUGCACR | 181 | 2.78E-06 | 0.000916 |
| | UGUGUGU | 21 | 1.94E-05 | 0.00638 |
| | GUGUGUG | 215 | 2.74E-05 | 0.00901 |
| | UGCAUGM | 214 | 4.99E-05 | 0.0164 |
| | WGCAUGM | 352 | 8.52E-05 | 0.028 |
| | GACGGCAGCACCGCG | 20 | 0.00117 | 2.70E-05 |
| | UUGUUGUUGUUGUUG | 74 | 2.27E-14 | 0.00969 |
| | WUUDUUDUUDUUDUU | 285 | 8.16E-13 | 2.80E-271 |
| | AAAAAAAAAAAAAAA | 159 | 1.58E-11 | 0 |
| | YUUUUUUUUUUUUUU | 181 | 3.19E-11 | 6.80E-275 |
| | AUAUAUAUAUAUAUA | 109 | 2.26E-10 | 4.40E-120 |
| | UAUAUGUA | 170 | 4.95E-10 | 0.033 |
| | AUWWAUWUAUUUAUU | 111 | 4.07E-09 | 0.014 |
| | UUUUUUY | 258 | 8.02E-08 | 2.08E-05 |
| | UUUUUUU | 248 | 4.36E-07 | 0.000113 |
| | CUUUUUU | 223 | 5.25E-07 | 0.000136 |
| | RWUCAAG | 233 | 7.52E-07 | 0.000195 |
| | AASAAHAANAANAAV | 163 | 1.24E-06 | 3.50E-54 |
| | UGUUGUG | 186 | 1.29E-06 | 0.000335 |
| **M7** | UGUKUGUGUGUKUGU | 162 | 1.36E-06 | 9.80E-122 |
| | UAUAUUA | 111 | 1.88E-06 | 0.00049 |
| | AUAUUWA | 180 | 1.96E-06 | 0.000511 |
| | UUUUUUC | 255 | 2.29E-06 | 0.000594 |
| | UUUUUUY | 255 | 2.52E-06 | 0.000656 |
| | GAAAACM | 251 | 3.60E-06 | 0.000936 |
| | UUUUUUU | 257 | 1.16E-05 | 0.00301 |
| | WUUUUUU | 249 | 1.34E-05 | 0.00348 |
| | AGAAAAR | 260 | 2.31E-05 | 0.00601 |
| | AUUUUUU | 233 | 2.52E-05 | 0.00656 |
| | UAURUAR | 122 | 3.08E-05 | 0.008 |
| | UUUUUUU | 252 | 4.09E-05 | 0.0106 |
| | AAUUUUG | 189 | 4.29E-05 | 0.0111 |
| | CUKUUGU | 257 | 4.49E-05 | 0.0117 |

| | | | | |
|---|---|---|---|---|
| | AUUUUUK | 234 | 4.96E-05 | 0.0129 |
| | UUUDGUU | 246 | 7.24E-05 | 0.0188 |
| | UUWUUUU | 229 | 7.38E-05 | 0.0192 |
| | CUUUUUU | 215 | 7.46E-05 | 0.0194 |
| | UUUUUUU | 260 | 8.19E-05 | 0.0213 |
| | RRAGRARRARAGRAR | 217 | 8.31E-05 | 3.80E-85 |
| | NUUUUUU | 204 | 9.86E-05 | 0.0256 |
| | NCAUUUU | 255 | 0.000103 | 0.0269 |
| | CWUUUUU | 248 | 0.00012 | 0.0312 |
| | AAAAAAG | 245 | 0.000124 | 0.0323 |
| | AGAAGAN | 249 | 0.000136 | 0.0355 |
| | GWGUAGW | 130 | 0.00019 | 0.0495 |
| | ACAMACAMACAMACA | 87 | 0.00264 | 1.50E-92 |
| **M8** | AAAGARRGAAAARAA | 118 | 8.61E-15 | 2.20E-27 |
| | AGRRAGRGARRRARA | 170 | 4.34E-11 | 1.70E-98 |
| | UGUGUGUGUGUYUSU | 49 | 4.50E-11 | 0.0477 |
| | AAAARAAAAAAAAAA | 181 | 6.73E-11 | 5.40013750904482e-321 |
| | UAGAUAGA | 32 | 1.87E-09 | 0.0334 |
| | AUAUAUAUAUAUAUA | 97 | 1.17E-07 | 8.30E-141 |
| | AMAAMAAMAAMAAAA | 145 | 5.26E-07 | 4.90E-120 |
| | AUAUWWA | 150 | 1.12E-06 | 0.000293 |
| | WAAWWAWAWWAAWAA | 85 | 4.90E-06 | 1.90E-42 |
| | ACAYACAYACACACA | 58 | 7.41E-06 | 7.20E-60 |
| | AGAGARR | 85 | 1.97E-05 | 0.00513 |
| | AGAAGAN | 169 | 3.46E-05 | 0.00904 |
| | AUAUAUM | 72 | 5.24E-05 | 0.0137 |
| | AUAUUWA | 174 | 0.000136 | 0.0356 |
| | ARAAAAA | 181 | 0.000141 | 0.0368 |
| | UUWUUUUUWUUWUUU | 221 | 0.000745 | 4.70E-195 |
| | CUYUYUYUCYYUCYY | 175 | 0.0329 | 3.40E-52 |
| **M9** | RARAAAAAAAAADAA | 106 | 1.13E-11 | 3.00E-104 |
| | AAAAAAAAAAAAAAA | 117 | 1.59E-10 | 4.60E-293 |
| | AUAAAAR | 63 | 2.16E-06 | 0.000554 |
| | AGAAAAR | 150 | 3.93E-06 | 0.00101 |
| | UAAAAGG | 138 | 6.81E-06 | 0.00174 |
| | AAAMAAA | 124 | 9.54E-06 | 0.00244 |
| | ARAAAAA | 127 | 1.08E-05 | 0.00277 |
| | ARAAAAA | 127 | 1.08E-05 | 0.00277 |
| | AGAGAMA | 135 | 5.13E-05 | 0.0131 |
| | AAAAAAA | 108 | 7.13E-05 | 0.0183 |
| | GAAAAHV | 131 | 7.15E-05 | 0.0183 |
| | RAUAAAM | 119 | 0.00014 | 0.0358 |
| | MCAGAUGV | 157 | 0.000163 | 0.0417 |
| | GUGUGUGUGUGUGUG | 62 | 0.000338 | 1.90E-121 |
| | UUKUUKUUKUUKUUU | 155 | 0.00382 | 7.70E-207 |

| | | | | |
|---|---|---|---|---|
| | AUAUAUAUAUAUAUA | 59 | 0.00729 | 5.80E-90 |
| | AGRGRGAGRRAGRRA | 86 | 0.0105 | 3.50E-64 |
| **M10** | GAGRGARRGAGRGAG | 92 | 1.18E-06 | 3.50E-79 |
| | RAADGAANRRAARDA | 136 | 2.25E-05 | 2.60E-38 |
| | UUAAGUU | 144 | 9.49E-05 | 0.0253 |
| | RARRARGARRARGAA | 85 | 0.000562 | 3.80E-11 |
| | AHAAYAAHAAHAAHA | 74 | 0.0013 | 3.20E-36 |
| | CRCMCRCACACACAC | 32 | 0.00273 | 3.00E-64 |
| | UAUAUAUWUAUAUAU | 53 | 0.00614 | 3.50E-06 |
| | UGUGUGUGUGUGUGU | 86 | 0.0081 | 8.30E-142 |
| | UAUAUAUAUAUAUAU | 76 | 0.0313 | 3.40E-108 |
| | CUYUYUYUCUYUCUY | 126 | 0.0388 | 5.00E-81 |
| **M11** | AAAMAAAAAAAAAAA | 98 | 2.61E-08 | 4.20E-260 |
| | AARGRAVVRVAASAA | 95 | 6.77E-06 | 4.80E-19 |
| | ARAAAAA | 107 | 7.48E-06 | 0.00192 |
| | ARAAAAA | 103 | 3.01E-05 | 0.00773 |
| | ARAAAAA | 103 | 3.13E-05 | 0.00805 |
| | GGUAGGG | 123 | 4.07E-05 | 0.0105 |
| | AABAAWAAWAAAAAW | 48 | 5.55E-05 | 3.40E-30 |
| | UUUKUUKUUKUUDUU | 146 | 5.64E-05 | 1.80E-196 |
| | AGAAAAR | 104 | 0.00017 | 0.0438 |
| | AAAAAAA | 93 | 0.000176 | 0.0451 |
| | GURGUKU | 140 | 0.000194 | 0.0498 |
| | AAARAAARARAGAAA | 131 | 0.00067 | 8.40E-128 |
| | UUYUUUUUUYYYUUU | 149 | 0.00331 | 7.00E-81 |
| **M12** | AAAAAMAAAAAMAAA | 102 | 2.49E-10 | 8.70E-273 |
| | UAUWUAUAUAUWUAU | 131 | 4.51E-08 | 2.20E-116 |
| | AAGGARWGRRA | 80 | 7.11E-08 | 2.30E-12 |
| | CGUUUAA | 32 | 1.94E-07 | 0.0036 |
| | AAAMAAMAAMAAAAA | 55 | 3.51E-05 | 7.00E-48 |
| | UUYYUUUYUUUUYUU | 109 | 5.11E-05 | 6.10E-213 |
| | AUAUWWA | 103 | 8.36E-05 | 0.0217 |
| | AUAUUWA | 133 | 9.02E-05 | 0.0234 |
| | AAAAGGAAAAA | 98 | 0.000129 | 0.0284 |
| | GAAAACM | 110 | 0.000161 | 0.0418 |
| | AUAAAAR | 56 | 0.000186 | 0.0481 |
| | GUGUGUGUGUG | 67 | 0.00557 | 5.40E-42 |
| | CACACACACACACAC | 22 | 0.00814 | 2.60E-45 |
| | GAARAAAAAAAARAA | 88 | 0.025 | 1.30E-133 |
| **M13** | GUCCCACGSUU | 29 | 1.25E-13 | 0.00271 |
| | UGCACACCCAUKC | 29 | 1.14E-12 | 0.0088 |
| | CGCUGCCGUCCU | 32 | 1.51E-12 | 0.0013 |
| | GGUGCUCUSUU | 28 | 6.67E-12 | 0.00271 |
| | YCCCACGAAGCACAC | 30 | 7.98E-12 | 0.0173 |
| | CGCSKYGCCUGCUGC | 28 | 1.14E-11 | 1.90E-37 |

| | | | |
|---|---|---|---|
| GGUUGAGWGCACCCA | 22 | 1.58E-11 | 1.10E-27 |
| UCUCGCCCUCACYUC | 27 | 1.89E-11 | 0.00557 |
| UUWUUAUUGCAUU | 25 | 5.18E-11 | 0.0497 |
| UGUGUGUUUUGCUU | 27 | 1.02E-10 | 0.0047 |
| CUCUCUCUCCCUGUG | 36 | 1.07E-10 | 9.10E-102 |
| CCGCACACACAC | 30 | 1.96E-10 | 0.0194 |
| GGCUGCAUGG | 30 | 2.10E-10 | 0.00177 |
| GCCCAUUGUAUAUAU | 25 | 2.58E-10 | 2.30E-94 |
| CACGCACCGCUGCCG | 31 | 2.89E-10 | 6.20E-123 |
| UGUGUGCGCUGGUS | 31 | 3.02E-10 | 0.00179 |
| UGUGUGUUUUGCUUU | 37 | 4.03E-10 | 2.20E-72 |
| CCCGUUGGUGCUCUC | 27 | 5.47E-10 | 4.70E-103 |
| CACGAAGCACACACA | 34 | 8.53E-10 | 2.10E-117 |
| CAUUGAGUGGGCGAC | 24 | 8.97E-10 | 2.00E-82 |
| CRCCCGCACACACAC | 42 | 9.44E-10 | 3.40E-143 |
| AUGUSCCCAU | 30 | 1.08E-09 | 0.00064 |
| GGGAGCAAUGAGA | 23 | 1.32E-09 | 0.0378 |
| CAUGCACACCCAUGC | 29 | 1.39E-09 | 1.70E-111 |
| GAGUGGGCGACCUCU | 28 | 1.96E-09 | 0.00472 |
| GCUGGKCGUCCCACG | 40 | 2.65E-09 | 1.30E-121 |
| ACUCUCGCCCUCACC | 27 | 2.68E-09 | 2.80E-108 |
| USUSUGUGUGUGUCY | 52 | 2.81E-09 | 2.70E-106 |
| AKGCCCGUU | 25 | 3.02E-09 | 0.000767 |
| CAUGGGCGGAGCACC | 31 | 3.38E-09 | 4.30E-115 |
| GCGGAGCACC | 31 | 3.85E-09 | 0.0116 |
| UGCGUGUGUGCUAUC | 9 | 7.10E-09 | 2.50E-15 |
| UGUGUGUGUCCCCU | 31 | 9.03E-09 | 0.000649 |
| UCGUCYUUCACUCAC | 17 | 1.81E-08 | 1.30E-34 |
| UCCUGCCUGGC | 28 | 1.81E-08 | 6.30E-50 |
| CCUUGCAU | 30 | 2.45E-08 | 5.20E-10 |
| UUGCACGACACG | 23 | 1.08E-07 | 0.00467 |
| ACGGGAGCAAUGAGA | 23 | 2.91E-07 | 6.70E-83 |
| CGCUGUGUGUUCCCG | 8 | 3.63E-07 | 1.30E-10 |
| GAGAGAGCCGUGCAC | 34 | 5.59E-07 | 6.90E-104 |
| UCUCCCUGUG | 30 | 6.24E-07 | 0.0141 |
| CCACMCMCYGCCVCC | 15 | 1.17E-06 | 1.90E-08 |
| CUGUGAGGGGUGMYG | 15 | 5.60E-06 | 4.00E-23 |
| CAUWGUD | 21 | 6.10E-06 | 0.00188 |
| UUCCGCGGGCUGCGU | 9 | 1.14E-05 | 1.40E-14 |
| UUYUUGUUUUGUUUU | 63 | 1.29E-05 | 7.50E-108 |
| UWCURYUCUUCCYUC | 12 | 1.33E-05 | 5.10E-14 |
| GUGUGUSUGUGUGYG | 20 | 2.27E-05 | 7.20E-27 |
| USCCGCCCARCUGCU | 22 | 3.23E-05 | 9.50E-34 |
| BGCUGGCC | 18 | 5.52E-05 | 0.017 |
| WAGCACAN | 18 | 6.58E-05 | 0.0203 |

| | | | | |
|---|---|---|---|---|
| GAGGGGCCGCGACGA | 13 | 6.99E-05 | 2.30E-31 |
| GCUGGMC | 19 | 8.07E-05 | 0.0249 |
| BGCUGGCC | 18 | 0.00012 | 0.0369 |
| DWGCACA | 25 | 0.000139 | 0.043 |
| DWGCACA | 27 | 0.000145 | 0.0446 |
| BUUUUUKUUUUKKUU | 58 | 0.000407 | 9.80E-51 |
| CACGCAGAUGACCCC | 7 | 0.000916 | 5.90E-09 |
| CGCACKCACRCRCCG | 12 | 0.00142 | 8.10E-11 |
| UUGCGGYCAGGGCGG | 9 | 0.00349 | 7.60E-12 |
| MCAMACRCACA | 33 | 0.0126 | 1.30E-22 |
| ACGCGACUCCCGUCA | 5 | 0.0129 | 0.009 |
| CUCAUGACGACGGCC | 9 | 0.0158 | 1.10E-19 |

**Tabla Suplementaria 3.**

Motivos a nivel de secuencia primaria identificados en las regiones 3'UTR de los genes de cada uno de los módulos con sobrerrepresentación de términos GO utilizando *FIMO*.

| Módulo | Motivo | RBP | Proporción en módulo | Proporción en *background* | Relación módulo/*background* |
|---|---|---|---|---|---|
| **M1** | GUUCCACGAUGACUG | - | 0.028 | 0.001 | 22.63 |
| | ACGGCUGUGCAGC | - | 0.028 | 0.001 | 25.86 |
| | RGCAUCCACYC | RNCMPT00170 (RBM6) | 0.023 | 0.000 | Inf |
| | UUGAGGUGGAG | RNCMPT00199 (PF10_0068) | 0.021 | 0.000 | Inf |
| | AACGGAUUAUCGA | - | 0.019 | 0.001 | 20.28 |
| **M4** | UGUUUUGCUUYAYKK | RNCMPT00229 (Pp_0229) | 0.599 | 0.134 | 4.47 |
| | CCGCUGCCSUSCUGC | RNCMPT00111 (Vts1p) | 0.584 | 0.068 | 8.55 |
| | CCASSAMKCAMMCAC | RNCMPT00178 (HNRPLL) | 0.564 | 0.065 | 8.72 |
| | AGWGUGUGUGCGCUG | RNCMPT00283 (Rbm38) | 0.529 | 0.059 | 8.91 |
| | ACCUCUCUCUCYCUC | RNCMPT00220 (Tb_0220) | 0.496 | 0.208 | 2.39 |
| | BUGCHGCUSCUUSUK | - | 0.465 | 0.116 | 4.02 |
| | UGCACCGCS | RNCMPT00249 (Pr_0249) | 0.447 | 0.039 | 11.35 |
| | UUUGCACSACACGCA | RNCMPT00249 (Pr_0249) | 0.405 | 0.057 | 7.15 |
| | SUGUGKYCCSWUGCA | RNCMPT00051 (RBM38) | 0.387 | 0.037 | 10.33 |
| | CCCGUUGGUGCUCUC | - | 0.381 | 0.038 | 10.14 |
| | CUCUCCCUGUG | RNCMPT00215 (PCBP3) | 0.376 | 0.044 | 8.45 |
| | UGCAGGCCC | RNCMPT00082 (Vts1p) | 0.357 | 0.000 | Inf |
| | UUGCAUGGACUCWCG | RNCMPT00123 (A2BP1) | 0.340 | 0.033 | 10.22 |
| | UAAYUGUUUY | - | 0.328 | 0.000 | Inf |

| | | | | | |
|---|---|---|---|---|---|
| | GAGAGAGCC | RNCMPT00090 (SRSF10) | 0.327 | 0.000 | Inf |
| | ACGCUUCGGC | - | 0.319 | 0.040 | 8.08 |
| | GCAUGGGCGGAGCA | RNCMPT00273 (EIF-2ALPHA) | 0.314 | 0.038 | 8.34 |
| | CAUAAUGUGCCC | - | 0.283 | 0.028 | 10.05 |
| | CCCUCACCUCAY | RNCMPT00186 (PCBP1) | 0.270 | 0.032 | 8.34 |
| | AGCACCGUAAC | - | 0.246 | 0.026 | 9.38 |
| | CCCUGCCGCCGUGKU | - | 0.245 | 0.051 | 4.82 |
| | GAGCAAUGA | - | 0.234 | 0.000 | Inf |
| | UGCAUUGAGUGGGCG | - | 0.233 | 0.027 | 8.70 |
| | CSWCGSCWGCYKCAS | - | 0.212 | 0.042 | 5.03 |
| | AUUGUAUAUA | RNCMPT00046 (PUM) | 0.206 | 0.000 | Inf |
| | CACGUUGAACGCAUC | RNCMPT00254 (Lm_0254) | 0.192 | 0.008 | 23.69 |
| | CGACGAACGGCGC | RNCMPT00061 (RSF1) | 0.187 | 0.028 | 6.75 |
| | CUGUGGGGCG | - | 0.177 | 0.021 | 8.38 |
| | GAGGGGCCGCGACGA | RNCMPT00061 (RSF1) | 0.177 | 0.032 | 5.50 |
| | ACGCAUGAGGA | RNCMPT00180 (ASD-1) | 0.171 | 0.028 | 6.04 |
| | GCUCCGCGUUGYUBC | - | 0.171 | 0.009 | 18.74 |
| | CCRCGGGGAYWUGUG | RNCMPT00150 (ESRP2) | 0.167 | 0.011 | 15.66 |
| | UGUGUUUCCCGCCCA | - | 0.143 | 0.025 | 5.82 |
| | GUGACGUGUGGG | RNCMPT00179 (SUP-12) | 0.134 | 0.024 | 5.66 |
| | CGGCCAGGGCGGGCA | RNCMPT00160 (HNRNPH2) | 0.133 | 0.025 | 5.38 |
| | CGGCCCUG | - | 0.129 | 0.000 | Inf |
| | UGCGACGAGG | RNCMPT00061 (RSF1) | 0.128 | 0.022 | 5.69 |
| | GCAAUCACUAUGGAC | - | 0.125 | 0.023 | 5.46 |
| | AGGGGAGCAGCAACA | RNCMPT00205 (RO3G_00049) | 0.124 | 0.024 | 5.12 |
| | AAAAUAAUGAUUG | - | 0.113 | 0.005 | 22.85 |
| | ACGCGACUCCCGUCA | - | 0.096 | 0.015 | 6.37 |
| | ACGAUGUGCGGACAC | - | 0.083 | 0.014 | 5.96 |
| | CYGUGUGUKUCCCSY | RNCMPT00011 (PAPI) | 0.561 | 0.110 | 5.11 |
| | **CUCGCACACCCAC** | RNCMPT00178 (HNRPLL) | 0.523 | 0.234 | 2.24 |
| | CUGCCGCCSUG | - | 0.492 | 0.082 | 5.98 |
| | CWCUCUYCCUGUGWG | RNCMPT00215 (PCBP3) | 0.490 | 0.168 | 2.91 |
| **M5** | MCACRCMCYGCYGCC | RNCMPT00178 (HNRPLL) | 0.478 | 0.078 | 6.16 |
| | CUUCUCUCAGUCCCC | - | 0.440 | 0.102 | 4.33 |
| | CKCSUSCUKCACKCA | RNCMPT00203 (PFI1695c) | 0.432 | 0.062 | 7.02 |
| | CACRCYCAUGMMGRC | RNCMPT00259 (Tv_0259) | 0.407 | 0.063 | 6.48 |
| | GCAGCAACACGCAGA | RNCMPT00225 (Tp_0225) | 0.402 | 0.176 | 2.28 |

| | | | | | |
|---|---|---|---|---|---|
| | GGCGGAGCACAS | RNCMPT00133 (CPO) | 0.380 | 0.064 | 5.92 |
| | CKGGYMGWCMCMCGC | RNCMPT00225 (Tp_0225) | 0.372 | 0.054 | 6.93 |
| | GSWGAGRGAGCCGUG | - | 0.370 | 0.068 | 5.43 |
| | UUUUGCUUUWSGGG | RNCMPT00136 (HuR) | 0.365 | 0.051 | 7.21 |
| | CUGCCUGGCUG | - | 0.360 | 0.062 | 5.86 |
| | CUGCGGGCCGCAY | RNCMPT00111 (Vts1p) | 0.317 | 0.027 | 11.87 |
| | CCCYUGCAUGG | RNCMPT00123 (A2BP1) | 0.284 | 0.040 | 7.10 |
| | GCCCAACUGCUC | - | 0.279 | 0.028 | 9.84 |
| | YUGCYGCUGCUKSKG | - | 0.277 | 0.073 | 3.82 |
| | ACCCCAYGAAR | RNCMPT00007 (CG2950) | 0.272 | 0.046 | 5.93 |
| | CAAUCACUAUGGACG | - | 0.259 | 0.036 | 7.10 |
| | UGAGUGGGCGACCUC | - | 0.257 | 0.044 | 5.82 |
| | GACGAGGGCCUGUGG | RNCMPT00073 (SRSF7) | 0.236 | 0.033 | 7.11 |
| | GCCUCGCCUGCUGC | - | 0.226 | 0.024 | 9.24 |
| | GCCCGUUGGUGCUC | - | 0.226 | 0.038 | 5.89 |
| | CGACGGCCCUGUGC | RNCMPT00061 (RSF1) | 0.218 | 0.025 | 8.68 |
| | AUGAGGAGGGGCC | RNCMPT00134 (B52) | 0.214 | 0.026 | 8.21 |
| | UUGCGGCCAGGGC | RNCMPT00144 (CG7903) | 0.211 | 0.021 | 9.91 |
| | ACGAACGGCGC | RNCMPT00248 (Rbm4.3) | 0.208 | 0.020 | 10.25 |
| | CCCUGUGCGCC | - | 0.206 | 0.020 | 10.10 |
| | CCUCACCUC | RNCMPT00186 (PCBP1) | 0.204 | 0.036 | 5.74 |
| | AAUGUGCCCAUUGUA | RNCMPT00216 (Tb_0216) | 0.204 | 0.037 | 5.56 |
| | GCAAUGAGAGUG | - | 0.183 | 0.037 | 4.87 |
| | MCMYKCGGUGYCGGC | - | 0.131 | 0.023 | 5.69 |
| | CGACUCCCGUC | - | 0.130 | 0.012 | 11.12 |
| | GCGUGGGACAMUU | RNCMPT00052 (RBM4) | 0.121 | 0.006 | 21.27 |
| | CGAUGUGCGGACACA | - | 0.120 | 0.014 | 8.43 |
| | AACAGUAAUGGACGA | | 0.113 | 0.015 | 7.75 |
| | AGAGGGCUGAGGGGA | RNCMPT00205 (RO3G_00049) | 0.113 | 0.020 | 5.64 |
| | GUCAGACGGUUGAAA | RNCMPT00002 (ANKHD1) | 0.103 | 0.014 | 7.46 |
| | GGCCACACGGCAGCC | - | 0.088 | 0.009 | 9.79 |
| **M7** | AAAGGGAGAU | - | 0.029 | 0.000 | Inf |
| **M9** | AAGAUGGAUUUG | - | 0.043 | 0.000 | Inf |
| | AUUGAUGUAD | - | 0.027 | 0.000 | Inf |
| **M10** | AUUUUCUCUKUAWU | RNCMPT00268 (PTBP1) | 0.052 | 0.000 | Inf |
| **M11** | GGAAGAGGGGAAGGK | RNCMPT00205 (RO3G_00049) | 0.192 | 0.022 | 8.77 |
| | GAAAGGGGAAGA | RNCMPT00205 (RO3G_00049) | 0.162 | 0.002 | 103.60 |

| | | | | | |
|---|---|---|---|---|---|
| | AAACGGGAAAGG | RNCMPT00205 (RO3G_00049) | 0.102 | 0.000 | Inf |
| | UGUGUGUUUUGCUU | RNCMPT00270 (ARET) | 0.700 | 0.006 | 120.39 |
| | GUGUGUSUGUGUGYG | RNCMPT00270 (ARET) | 0.657 | 0.005 | 120.39 |
| | CUCUCUCUCCCUGUG | RNCMPT00215 (PCBP3) | 0.600 | 0.005 | 120.39 |
| | CCGCACACACAC | RNCMPT00069 (SM) | 0.586 | 0.005 | 120.39 |
| | UUWUUAUUGCAUU | RNCMPT00200 (PF13_0315) | 0.571 | 0.005 | 120.39 |
| | UCUCGCCCUCACYUC | RNCMPT00203 (PFI1695c) | 0.514 | 0.004 | 120.39 |
| | CACGCACCGCUGCCG | RNCMPT00111 (Vts1p) | 0.514 | 0.004 | 120.39 |
| | USCCGCCCARCUGCU | - | 0.514 | 0.004 | 120.39 |
| | GAGAGAGCCGUGCA | - | 0.500 | 0.004 | 120.39 |
| | CGCUGCCGUCCU | RNCMPT00111 (Vts1p) | 0.500 | 0.004 | 120.39 |
| | UGCACACCCAUKC | RNCMPT00178 (HNRPLL) | 0.471 | 0.004 | 120.39 |
| | YCCCACGAAGCACAC | RNCMPT00007 (CG2950) | 0.471 | 0.004 | 120.39 |
| | UAACUGUUUCU | - | 0.457 | 0.004 | 120.39 |
| | CAUGGGCGGAGCACC | RNCMPT00238 (NCU02404) | 0.457 | 0.004 | 120.39 |
| | UCGUCYUUCACUCAC | RNCMPT00251 (Tb_0251) | 0.457 | 0.004 | 120.39 |
| | AUGUSCCCAU | - | 0.443 | 0.004 | 120.39 |
| **M13** | GGCUGCAUGG | RNCMPT00123 (A2BP1) | 0.443 | 0.004 | 120.39 |
| | GUCCCACGSUU | - | 0.443 | 0.004 | 120.39 |
| | GGUGCUCUSUU | - | 0.429 | 0.004 | 120.39 |
| | MAAAAAAAAAAAAAA | RNCMPT00043 (PABPC4) | 0.429 | 0.004 | 120.39 |
| | CGCSKYGCCUGCUGC | - | 0.400 | 0.003 | 120.39 |
| | UCCUGCCUGGC | RNCMPT00123 (A2BP1) | 0.400 | 0.003 | 120.39 |
| | CGGCUUUUUSUUU | RNCMPT00025 (HNRNPC) | 0.371 | 0.003 | 120.39 |
| | UUGCACGACACG | RNCMPT00249 (Pr_0249) | 0.371 | 0.003 | 120.39 |
| | GCCCAUUGUAUAUAU | RNCMPT00216 (Tb_0216) | 0.371 | 0.003 | 120.39 |
| | AKGCCCGUU | - | 0.357 | 0.003 | 120.39 |
| | CAUUGAGUGGGCGAC | - | 0.357 | 0.003 | 120.39 |
| | GGGAGCAAUGAGA | - | 0.343 | 0.003 | 120.39 |
| | GGUUGAGWGCACCCA | RNCMPT00133 (CPO) | 0.314 | 0.003 | 120.39 |
| | UWCURYUCUUCCYUC | - | 0.257 | 0.002 | 120.39 |
| | CUGUGAGGGGUGMYG | RNCMPT00187 (BRUNOL6) | 0.229 | 0.002 | 120.39 |
| | GAGGGGCCGCGACGA | RNCMPT00061 (RSF1) | 0.186 | 0.002 | 120.39 |
| | UUCCGCGGGCUGCGU | RNCMPT00113 (RBM4) | 0.143 | 0.001 | 120.39 |
| | **CUCAUGACGACGGCC** | RNCMPT00073 (SRSF7) | 0.129 | 0.001 | 120.39 |

| | | | | | |
|---|---|---|---|---|---|
| | UUGCGGYCAGGGCGG | RNCMPT00144 (CG7903) | 0.129 | 0.001 | 120.39 |
| | ACGCGACUCCCGUCA | - | 0.100 | 0.001 | 120.39 |

## Tabla Suplementaria 4.

Tabla detallada con información de la anotación funcional mediante *DARK* y *FoldSeek* de *hubgenes* identificados.

| Módulo | *hubgenes no anotados* | Anotación TriTrypDB | Anotación DARK | Anotación AlphaFold + FoldSeek | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Algoritmo | Descripcion | Taxonomia | Probabilidad | % id. sec. | e-value/TM-score | DB |
| M1 | TcCLB.510849.30 | hypothetical protein, conserved (pseudogene) | - | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | - | - | - | - | - | - |
| | TcCLB.508835.10 | hypothetical protein, conserved (pseudogene) | - | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | - | - | - | - | - | - |
| | TcCLB.506113.80 | hypothetical protein, conserved (pseudogene) | - | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | - | - | - | - | - | - |
| M5 | TcCLB.510275.272 | hypothetical protein | UNICLUST30(3) - Trans-sialidase [13\|2/8\|V5B9P9: 99.4\|1.3e-14\|137-294] | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | - | - | - | - | - | - |
| | TcCLB.509233.10 | hypothetical protein | UNICLUST30(1) - Mucin-associated surface protein (MASP), putative [54\|4/7\|K2MUL5: 100.0\|3.8e-32\|310-480] | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | - | - | - | - | - | - |
| M7 | TcCLB.508771.50 | hypothetical protein, conserved | - | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | - | - | - | - | - | - |
| | TcCLB.509157.220 | hypothetical protein, conserved | UNICLUST30(1) - Putative transmembrane protein [31\|1/60\|G0U922: 100.0\|9.8e-52\|1-105] | 3Di/AA | Protein FATTY ACID EXPORT 7 | *Arabidopsis thaliana* | 1 | 36,6 | 3,31e-3 | AFDB SWISSPROT |
| | | | | TM-align | Protein FATTY ACID EXPORT 7 | *Arabidopsis thaliana* | 0,97 | 29,8 | 0,752 | AFDB SWISSPROT |
| M9 | TcCLB.508731.40 | hypothetical protein, conserved | UNICLUST30(1) - Mitochondrial glycoprotein-like protein [15\|1/8\|A0A1G4HYY9: 96.9\|7.9e-05\|1-368] / POSIBLE P22 O P32 | 3Di/AA | Head domain of the mt-SSU assemblosome from Trypanosoma brucei | *Trypansoma brucei* | 1 | 45,5 | 3.33E-36 | PDB100 |
| | | | | TM-align | Head domain of the mt-SSU assemblosome from Trypanosoma brucei | *Trypansoma brucei* | 0,44 | 43,3 | 0,715 | PDB100 |
| | TcCLB.510347.29 | hypothetical protein, conserved | PDB70(1) - HP_Q4D059 [31\|1/1\|2MNI_A: 100.0\|1.6e-51\|7-91] | 3Di/AA | Chemical Shift Assignments and structure of Q4D059, a hypothetical protein from Trypanosoma cruzi | *Trypanosoma cruzi* | 1 | 100 | 3.43E-13 | PDB100 |
| | | | | TM-align | HORMA domain containing proTEIN | *B. bacterium* | 0,35 | 4 | 0,42 | AFDB50 |
| | TcCLB.506927.20 | hypothetical protein, conserved | PFAM(3) - Actin interacting protein 1 [11\|4/26\|d1nr0a2: 99.7\|1.4e-23\|2-298] | 3Di/AA | Actin-interacting protein 1 | *Drosophila melanogaster* | 1 | 10.7 | 2.07E-15 | AFDB SWISSPROT |
| | | | | TM-align | Angio-associated migratory cell protein | *Bos taurus* | 0.12 | 7.2 | 0.49 | AFDB SWISSPROT |
| M10 | TcCLB.503811.45 | hypothetical protein, conserved | - | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | - | - | - | - | - | - |
| M11 | TcCLB.511623.20 | hypothetical protein, conserved | - | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | Pleckstrin homology-like domain, family B, member 2 | *Rattus norvegicus* | 0,10 | 1,3 | 0,215 | AFDB-PROTEOME |
| M12 | TcCLB.511127.90 | Domain of unknown function (DUF4586), putative | UNICLUST30(3) - Flagellar associated protein [13\|2/8\|A8J2Y0: 99.9\|3.6e-26\|7-345] | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | - | - | - | - | - | - |
| | TcCLB.503903.70 | hypothetical protein, conserved | UNICLUST30(1) - Morn repeat protein [60\|24/166\|A0A078ATY6: 100.0\|2.1e-32\|98-376] | 3Di/AA | Crystal structura of trypansoma brucei morn 1 | Trypanosoma brucei | 1 | 23,1 | 1,04e-15 | PDB100 |
| | | | | TM-align | Crystal structure of Trypanosoma brucei Morn1 | Trypanosoma brucei | 0,54 | 16,3 | 0,665 | PDB100 |
| | TcCLB.453917.9 | hypothetical protein, conserved | - | 3Di/AA | - | - | - | - | - | - |
| | | | | TM-align | - | - | - | - | - | - |
| | TcCLB.511693.70 | hypothetical protein, conserved | PFAM(2) - Actin interacting protein 1 [31\|4/35\|d1nr0a1: 100.0\|2.9e-36\|33-311] | 3Di/AA | Cilia- and flagella-associated protein 52 | *Chlamydomonas reinhardtii* | 1 | 47.4 | 1.69E-74 | AFDB SWISSPROT |
| | | | | TM-align | Cilia- and flagella-associated protein 52 | *Chlamydomonas reinhardtii* | 1 | 46.4 | 0.963 | AFDB SWISSPROT |