Tesis – Maestría en Bioinformática

Aprendizaje automático para la identificación de genes involucrados en la fosforilación oxidativa de *Caenorhabditis elegans* a partir de datos de bulk y single cell RNA-seq.

Lic. Sofía Zeballos-Gorón

30 de diciembre de 2024

Orientación:
Dr. Flavio Pazos

Co-orientación:
Dr. Gustavo Salinas

Laboratorio de Biología de Gusanos – Instituto Pasteur Montevideo
Departamento de Biología del Neurodesarrollo – Instituto de Investigaciones Biológicas Clemente Estable

Índice

Glosario

ATP	Adenosín trifosfato
ADP	Adenosin difosfato
AA	Aprendizaje automático
ADN	Ácido desoxirribonucleico
ADNmt	ADN mitocondrial
ADNn	ADN nuclear
ARN	Ácido ribonucleico
AUC-ROC	Área bajo la curva ROC
CI	Complejo I
CII	Complejo II
CIII	Complejo III
CIV	Complejo IV
CLR	*Centered log-ratio transformation*
CTE	Cadena de Transporte de Electrones
CV	Complejo V
DCPM	*Depth of Coverage Per Million mapped reads*
DEG	Genes diferencialmente expresados
EIM	Espacio intermembrana
F1	*Score F1*
FN	Falsos negativos
FP	Falsos positivos
GEO	*Gene Expression Omnibus*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
KNN	*K-Nearest Neighbors*
LOO	*Leave One Out*
MDS	Escalado Multidimensional
MME	Membrana Mitocondrial Externa
MMI	Membrana Mitocondrial Interna
NCBI	*National Center for Biotechnology Information*

O_2	Oxígeno molecular
Pi	Fosfato inorgánico
RE	Retículo Endoplasmático
RF	*Random Forest*
RMSD	*Root Mean Square Deviation*
RNA-seq	Secuenciación de ARN
ROC	*Receiver Operating Characteristic*
RPKM	*Reads Per Kilobase Million*
sc-RNAseq	Secuenciación de célula única de ARN
SRA	*Sequence Read Archive*
SVM	*Support Vector Machine*
TM score	*Template Modeling score*
TMM	*Trimmed Mean of M-values*
TPM	*Transcripts Per Million*
UQ	Ubiquinona
UQH2	Ubiquinol
VN	Verdaderos negativos
VP	Verdaderos positivos

1. Introducción

1.1. Fosforilación oxidativa.

La fosforilación oxidativa es un proceso esencial en la producción de adenosín trifosfato (ATP), la principal fuente de energía en las células de la mayoría de los organismos vivos. En células eucariotas, este proceso es llevado a cabo por cinco complejos proteicos que se encuentran en la membrana mitocondrial interna (MMI). Cuatro de estos complejos componen lo que se conoce como la cadena de transporte de electrones (CTE), mientras que el quinto es conocido como la ATP sintasa (Figura 1). En la fosforilación oxidativa los electrones obtenidos de la oxidación de moléculas orgánicas son utilizados para reducir el NAD+ y el FAD a NADH y $FADH_2$, y de estas moléculas los electrones son cedidos al oxígeno molecular (O_2), transportados por la CTE a su aceptor final. Durante el transporte de electrones en la CTE, estos transcurren por varios dadores y aceptores, pasando a estados energéticos más bajos. La energía que los electrones liberan en este proceso es utilizada para el bombeo de protones contra gradiente desde la matriz mitocondrial hacía el espacio intermembrana (EIM) de la mitocondria. Estos protones generan un gradiente electroquímico que es utilizado por la ATPasa para la fosforilación de adenosin difosfato (ADP).

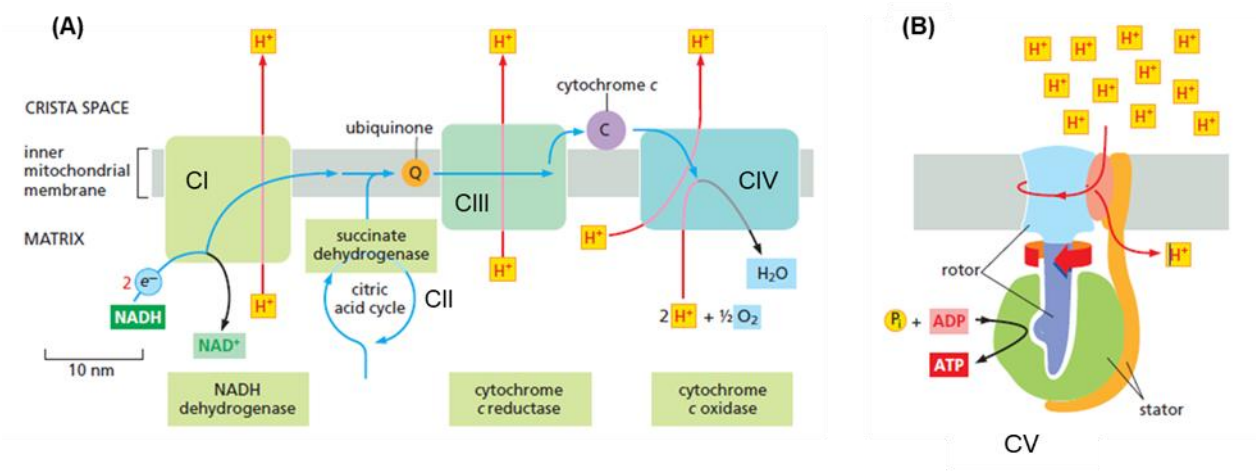


Figura 1: Diagrama del proceso de fosforilación oxidativa. (A) Transferencia de electrones en la cadena de transporte de electrones (CTE) mitocondrial, a través de los cuatro complejos (CI, CII, CIII y CIV). (B) Diagrama de la ATP sintasa (CV) mitocondrial. Esquemas adaptados del Alberts, 6ta ed.

El complejo I de la CTE, también conocido como NADH deshidrogenasa, lleva a cabo la reacción de oxidación del NADH, principalmente producido por el ciclo de Krebs. Los dos electrones que se liberan de la reacción van a la ubiquinona (UQ, conocida como coenzima Q), molécula hidrofóbica que se encuentra en la MMI, que se reduce a ubiquinol (UQH_2). Es el complejo más grande de los cinco. En mamíferos está compuesto por 14 proteínas del *core* (seis de ellas codificadas en el genoma mitocondrial - ADNmt), en donde se da el proceso catalítico, y 31 proteínas accesorias[1–3]. Su forma en "L" deriva de tres módulos estructural y funcionalmente diferentes: el módulo N, expuesto a la matiz mitocondrial y responsable de la oxidación de NADH; el módulo Q, que lleva a cabo el transporte de electrones a través del complejo haciendo uso de varios centros redox; y el módulo P, dispuesto en la MMI y encargado del bombeo de protones a través de la membrana hacia el EIM[4,5]. El módulo P está formado por cuatro motivos estructurales

similares a un antiporter, los cuales son capaces de transportar un protón cada uno[6]. Siendo así, de la oxidación de una molécula de NADH dos electrones son cedidos a la UQ y cuatro protones son transportados al EIM.

Por su lado, el complejo II, también conocido como la enzima succinato deshidrogenasa, es el único complejo de la cadena que no bombea protones hacia el EIM. Juega un rol crucial en el ciclo de Krebs catalizando la reacción de oxidación de succinato a fumarato, acoplado a la reducción de UQ a UQH_2, vinculando directamente el ciclo del ácido cítrico a la CTE. Este complejo está formado por cuatro subunidades, en mamíferos nombradas SDHA, SDHB, SDHC y SDHD, en donde las subunidades SDHA y SDHB se encargan del proceso de oxido-reducción, mientras que las subunidades SDHC y SDHD son el anclaje a la membrana[4,7]. Este es el único complejo para el cual no existen subunidades codificadas en el ADNmt.

Más allá de los dos primeros complejos de la CTE, existen otras enzimas que son capaces de reducir UQ a UQH_2 y contribuir al transporte de electrones a través de la MMI. Estas enzimas participan de otras vías metabólicas, como lo son la β-oxidación de ácidos grasos (ETFDH, deshidrogenasa de la flavoproteína de transferencia de electrones), el ciclo de la metionina, moléculas de un carbono y biosíntesis de lípidos (CHDH, colina deshidrogenasa), la síntesis de pirimidinas (DHODH, dehidroorotato deshidrogenasa), el transporte glicerol-fosfato de la mitocondria (G3PDH, glicerol 3-fosfato deshidrogenasa), la oxidación de prolina (prolina deshidrogenasa, PRODH) y la metabolización de H_2S (SQOR, sulfuro quinona óxidorreductasa). Adicionalmente, la enzima SUOX (sulfito oxidasa), partícipe de la metabolización de H_2S, y GFER, oxidasa de flavina para el ensamblaje de enlaces de disulfuro, también ceden electrones a la CTE, pero a nivel del citocromo C.

El complejo III, citocromo C reductasa o complejo citocromo bc1, es un homodímero donde cada monómero del complejo está conformado por tres proteínas core, siendo la proteína CYB (humana) codificada por el ADNmt. Adicionalmente tiene una cantidad variable de proteínas accesorias dependiendo de la especie[4]. Este complejo enzimático cataliza la reacción de oxidación del UQH_2 a UQ acoplada a la reducción del citocromo C, una pequeña proteína que es capaz de transportar un electrón entre el complejo III y complejo IV a través del EIM. Una vez que ingresa una molécula de UQH_2, dos protones son transportados al EIM mientras que uno de los electrones es trasportado a una molécula de citocromo C y el otro ingresa al denominado ciclo Q (o ciclo de la ubiquinona). Este electrón que proviene de una molécula de UQH_2 es cedido a la UQ transformándose en semiubiquinona con un electrón desapareado. Al ingresar una segunda molécula de UQH_2, el segundo electrón reacciona con la semiubiquinona para dar lugar nuevamente a UQH_2. En consecuencia, de la reacción de oxido-reducción, se transportan cuatro protones al EIM, dos por cada electrón cedido por el UQH_2 al citocromo C[8].

El complejo IV, también conocido como la enzima citocromo C oxidasa, es el último complejo de la CTE. Cataliza la reacción de oxidación de cuatro moléculas de citocromo C acoplado a la reducción de una molécula de O_2 a H_2O, transportando cuatro protones hacia el EIM. Estudios de cristalografía han encontrado que este complejo se encuentra en forma de homodímeros, en donde cada monómero de complejo posee tres proteínas *core* de membrana codificadas por el ADNmt y diez proteínas codificadas por el ADN núclear[4].

Finalmente, el complejo V, o la enzima ATP sintasa, cataliza la síntesis de ATP a parir de ADP y fosfato inorgánico (Pi) haciendo uso del gradiente de protones generado por la CTE. Esta enzima

está compuesta por dos subunidades, F0 que se encuentra en la MMI y funciona como un rotor por el que pasan los protones del EIM, y la subunidad F1 que se encuentra orientada en la matriz mitocondrial, en la cual se da la reacción de fosforilación gracias a la fuerza protón motriz generada por la subunidad F0[6]. Este complejo puede actuar en dirección contraria, expulsando los protones de la matriz hacia el EIM con el uso de ATP[4]. En la MMI se encuentra como dímeros, con dos proteínas mitocondriales en cada monómero. La estructura dimérica de la ATP sintasa es crucial para crear y mantener la forma curva de las membranas de las crestas en las mitocondrias [4,6].

Durante muchos años la organización de los complejos mitocondriales estuvo en discusión en términos de dos modelos muy distintos: el modelo sólido y el modelo fluido. En 1947 Keilin y Hartree[9] propusieron que el estado más probable del sistema de fosforilación oxidativa era aquel en el que los complejos estuvieran físicamente anclados a la MMI de forma tal que facilitara su acceso y mejorara su capacidad catalítica. En 1963 Chance[10] introduce el concepto del "oxisoma", definido como una unidad funcional para la transferencia de electrones y la fosforilación oxidativa. En conjunto, la teoría de Keilin, Chance y sus respectivos colegas se conoció como el modelo sólido.

En 1986 Hackenbrock *et al.*[11] propusieron el modelo de colisión, o modelo fluido, en el que establece que los componentes de la cadena de transporte de electrones se encuentran libres en la MMI y el citocromo C difunde libremente en el EIM, lo que implica que este proceso se pueda dar a distancia en la mitocondria. Este era el modelo globalmente aceptado hasta que en el año 2000 Schägger y Pfeiffer[12] utilizando electroforesis en gel de poliacrilamida nativa azul (BN-PAGE), técnica que permite conservar interacciones moleculares entre las proteínas de los complejos en un gel, encontraron bandas correspondientes a agrupaciones de complejos, más allá de los complejos individuales, llamados respirasomas. En el 2008 Acín-Pérez *et al.*[13] encontraron que el ensamblado del respirasoma se daba luego de que los complejos se encontraran ensamblados individualmente, que existía un pool de UQ y citocromo C asociado a estos supercomplejos, y que era posible transferir electrones del NADH al O_2, por lo que los respirasomas eran entidades funcionales. Con esta evidencia propusieron el modelo de plasticidad, en donde establecieron que los complejos no solo se encontraban libres en la MMI, sino que se agrupaban en diferentes estequiometrias para formar supercomplejos funcionales. A partir del trabajo reciente de Zheng *et. al*[14], en el cual se utiliza microscopia crio-electrónica *in situ* para estudiar la mitocondria de cerdos, se identificó que la organización de supercomplejos se daba mayormente de la forma: $CI_1CIII_2CIV_1$, $CI_1CIII_2CIV_2$, $CI_2CIII_2CIV_2$ y $CI_2CIII_4CIV_2$, en donde la estequiometria de los complejos está indicada por el subíndice.

Mutaciones en los complejos que forman parte del proceso de fosforilación oxidativa en humanos se han asociado diversas enfermedades, entre ellas el síndrome de Leigh, la encefalopatía mitocondrial, la encefalomiopatía, miopatía, la acidosis láctica, la atrofia óptica hereditaria de Leber (LHON) y el síndrome MELAS[15]. Por otro lado, se ha encontrado que trastornos como Alzheimer[16,17], Parkinson[18,19], autismo[20,21] y bipolaridad[22] están relacionados con disrupciones o el funcionamiento anormal de estos complejos.

El proceso de fosforilación oxidativa ha sido objeto de estudio durante más de 60 años, lo que ha permitido grandes avances en la comprensión de su funcionamiento y la identificación de sus componentes en diversos organismos. Es evidente que su correcto funcionamiento es crucial para la supervivencia y la salud de los organismos que lo utilizan. Sin embargo, aún existen

aspectos desconocidos sobre su funcionamiento que despiertan el interés para seguir trabajando en este tema.

Como se mencionó anteriormente, los complejos I, III, IV y V contienen proteínas codificadas en el genoma mitocondrial. No es claro cómo se coordina la expresión de genes de ambos genomas para el correcto ensamblado en la MMI[5]. Tampoco se entiende completamente la regulación del número de copias del cromosoma mitocondrial en tejidos con diferentes requerimientos energéticos, lo que parece ser una forma de regular la cantidad de complejos expresados[5]. Además, los detalles sobre cómo se inserta el ADN codificado mitocondrialmente en la MMI siguen siendo desconocidos[5].

En cuanto al complejo I, se sabe poco sobre el papel de las subunidades accesorias[6]. Respecto al complejo II, no se conoce el camino exacto por el cual se transfieren electrones a la UQ[4]. Del complejo III, se desconoce la función de sus proteínas accesorias[5], el proceso de dimerización (aunque se sabe que la proteína BRAWNIN está relacionada[4]), y la bifurcación de electrones en el sitio Q0[6]. El complejo IV presenta duplicaciones génicas en ciertas subunidades, algunas de las cuales se expresan específicamente en ciertos tejidos[5,6], pero se necesita más investigación al respecto. En el complejo V, las subunidades alfa de la cabeza F1 poseen sitios de unión a nucleótidos cuya función aún no se ha determinado[6]. Además, los pasos completos de su ensamblaje aún no han sido dilucidados[4].

En relación a los supercomplejos, no se comprende completamente el proceso de formación de agregaciones de complejos[6,23–25]. Tampoco se conoce el papel de la mitocondria y supercomplejos durante el desarrollo[25], ni cómo las crestas mitocondriales sufren cambios conformacionales para la formación de los mismos[24]. Finalmente, durante la falta de alimento, se utiliza la vía de degradación de lípidos en lugar de la glucólisis, lo cual genera una diferente composición de agregados de complejos, y no se entiende cómo se regula este proceso[26].

Aunque la fosforilación oxidativa ha sido ampliamente estudiada en mamíferos, nuestro conocimiento sobre este proceso en otros linajes es limitado. En general, investigaciones en otros organismos han adoptado un enfoque basado en la identificación de homólogos de secuencia de genes previamente caracterizados en mamíferos. Si bien esta estrategia ha permitido inferir ciertas funciones conservadas, puede pasar por alto genes con roles clave que no presentan homología. Esto resalta la necesidad de explorar enfoques alternativos, como el análisis funcional y de datos genómicos y transcriptómicos, para comprender mejor las particularidades y adaptaciones de este proceso en diferentes linajes evolutivos.

1.2. Caenorhabditis elegans.

Caenorhabditis elegans es un nematodo de vida libre perteneciente a la familia Rhabditidae, usado desde hace más de 50 años como organismo modelo. La primera vez que *C. elegans* fue descrito como tal fue en una publicación de Emile Mapuas en el año 1900, llamado en su momento *Rhabditis elegans*. En este trabajo se describen varias especies de nematodos, particularmente su modo de reproducción[27]. Casi 50 años más tarde, Victor M. Nigon, en una serie de trabajos científicos, describe el ciclo de vida de este nematodo y las bases genéticas de la determinación de su sexo[27]. En el año 1973 Sidney Brenner publica su trabajo titulado *The genetics of Caenorhabditis elegans* donde describe métodos para su aislamiento, complementación y mapeo de mutantes, finalmente proponiendo que se utilice como organismo

modelo para el estudio del desarrollo y de la neurobiología[28]. Desde ese momento se ha utilizado en varios campos de la biología, tradicionalmente usado para estudiar la diferenciación celular, biología del desarrollo y neurobiología, siendo también un formidable modelo para numerosas preguntas biológicas de diversas áreas como envejecimiento, respuesta al estrés, metabolismo, interacciones patógeno-hospedero y hasta en estudios evolutivos[27].

Estos organismos se alimentan de bacterias, por lo que en la naturaleza se encuentran en sitios donde haya vegetación en descomposición[29]. Son animales hermafroditas auto fecundativos, aunque en la naturaleza se pueden encontrar machos producidos por la no disyunción del cromosoma X durante la meiosis en la línea germinal. Estos animales eclosionan del huevo con 558 células[30], y el ciclo de vida se desarrolla en 4 etapas larvarias hasta llegar al estadío adulto, en donde alcanzan una longitud máxima de 1 mm (Figura 2).

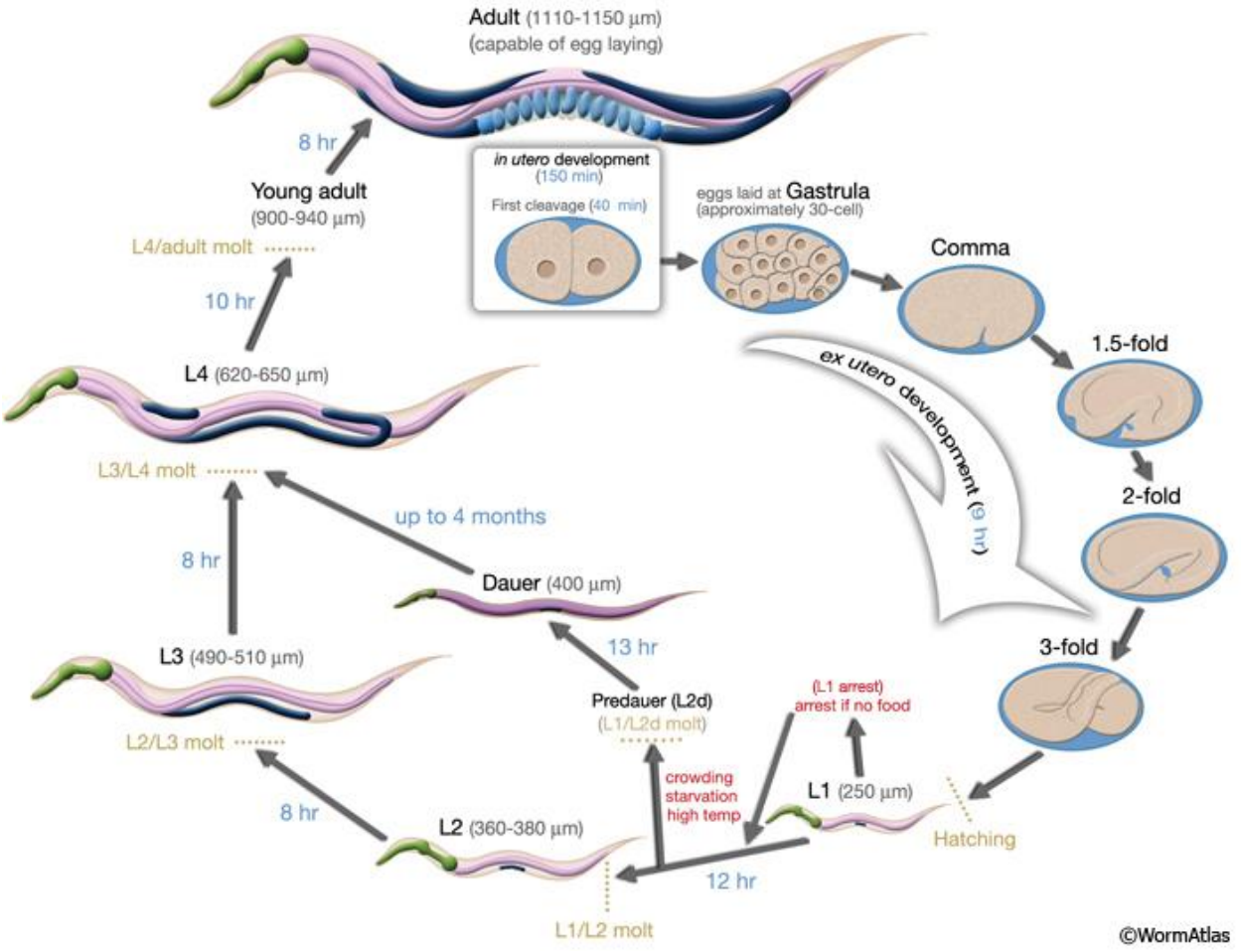


Figura 2: Esquema del ciclo de vida de C. elegans. Tomado de WormAtlas.

Las principales ventajas del uso de este organismo son su corto ciclo de vida, su capacidad de generar una gran progenie, la facilidad de cultivarlo y su bajo costo. En 1998 el genoma de *C. elegans* fue publicado como el primer organismo multicelular secuenciado[31,32]. En el 2000 se determinó que 83% del proteoma de *C. elegans* era compartido con el de humanos[33], lo que facilitó la anotación del genoma humano, publicado en el 2001[34]. Vías de señalización celular como la dependiente de insulina, mecanismos regulatorios de la inmunidad innata, la mayoría de

los sistemas de neurotransmisión y sistemas involucrados en la homeostasis de proteínas son altamente conservados entre gusanos y mamíferos[35]. Estas cualidades hacen que *C. elegans* sea un gran modelo para el estudio de enfermedades humanas. En particular, este organismo es utilizado para el estudio de enfermedades neurodegenerativas, muchas de ellas generadas por mutaciones que afectan procesos mitocondriales[36], como el Alzheimer[37,38], Parkinson[39–41], síndrome de Leigh[42] y esclerosis lateral amiotrófica[43]; trastornos como el autismo[38]; y también es utilizado como modelo para el estudio de la interacción entre la microbiota y organismos hospederos[44–46].

La facilidad en la manipulación de este organismo ha sido aprovechada para estudiar su expresión tanto a nivel de organismo completo (*RNA-seq*), como de célula única (*scRNA-seq*), con más de 30.000 muestras depositadas en *Gene Expression Omnibus* (*GEO*). Esta enorme cantidad de datos ha facilitado uso de algoritmos de aprendizaje automático para identificar patrones de expresión y explorar su relación con procesos biológicos. Existen trabajos que han aprovechado estos datos transcriptómicos para entrenar modelos predictivos, mostrando el potencial de esta aproximación en el estudio de la biología de *C. elegans*[47–53].

1.2.1. *C. elegans* como modelo de enfermedades mitocondriales.

Las enfermedades mitocondriales generan un grupo heterogéneo de trastornos, que pueden ser producto de mutaciones del ADNmt o del ADN nuclear (ADNn), y afectan la capacidad de las mitocondrias para producir energía. Es así como tejidos con altos requerimientos energéticos son los que se ven más afectados, como el cerebro, músculos, corazón e hígado[54,55]. Más allá de las características previamente descritas que hacen que *C. elegans* sea un buen organismo modelo, cuenta con ciertas ventajas adicionales para el estudio de estas enfermedades. Los gusanos hermafroditas adultos cuentan con 959 células somáticas, de las cuales 302 son neuronas, conociéndose el conectoma de cada una de ellas[56]. Esto hace que sea un modelo particularmente interesante para el estudio de enfermedades mitocondriales que afectan el sistema nervioso. Por otro lado, la mayoría de las proteínas mitocondriales encontradas en *C. elegans* tienen ortólogos humanos[57] y procesos claves como el ciclo de Krebs, la CTE, los perfiles de consumo de oxígeno y la formación de supercomplejos se comparten[58]. Adicionalmente, en un trabajo del 2003 se reporta que 72 de las 91 proteínas que componen la CTE de humanos están presentes en *C. elegans*[59]. Estas cualidades, sumado a la fácil manipulación genética, hace que *C. elegans* sea un buen modelo para comprender las bases genéticas de estas enfermedades e identificar posibles tratamientos[54].

1.2.2. La CTE como blanco de antihelmínticos.

Desde hace varias décadas que se utiliza *C. elegans* como modelo para el estudio de helmintos parásitos y para el desarrollo de antihelmínticos. La Organización Mundial de la Salud estima que las enfermedades causadas por helmintos afectan a casi un cuarto de la población mundial, en donde el sistema digestivo se ve principalmente afectado, causando importante morbilidad y llegando a causar la muerte en infecciones extremas. Los *screenings* genéticos y su capacidad para expresar genes heterólogos han sido fundamentales para identificar y caracterizar los blancos moleculares de varias clases de drogas antihelmínticas[60,61]. Piperazina y benzimidazoles fueron los primeros compuestos caracterizados por tener actividad antihemíntica a mediados del siglo XX y desde entonces se han creado derivados modificados que alteran su especificidad y eficacia[61].

Un enfoque que se ha dado en los últimos años se basa en dirigir el desarrollo de antihelmínticos a la CTE. En nematodos, existe una cadena alternativa en la cual los complejos I y II son capaces de utilizar rodoquinona como transportador de electrones[62], y este mecanismo que los diferencian de mamíferos es un blanco prometedor para el desarrollo de drogas[63]. En particular, recientemente se ha identificado la acción de benzimidazoles como inhibidores del complejo I de *C. elegans* únicamente en condiciones de anaerobiosis, indicando que su especificidad es hacia el complejo I modificado[64]. Adicionalmente se han encontrado otras drogas que inhiben la acción de los complejos I[64–66], II [65,67–69] y III[70]. En suma, el estudio de la CTE y la fosforilación oxidativa en *C. elegans* puede dar lugar al descubrimiento de nuevos blancos antihelmínticos que ayuden a combatir las enfermedades provocadas por estos organismos.

1.3. Aprendizaje automático.

La inteligencia artificial es un campo de la ciencia de la computación cuyo propósito es la creación de sistemas capaces de imitar la inteligencia humana para realizar tareas. El aprendizaje automático (AA) es una subdisciplina de la inteligencia artificial que crea y evalúa algoritmos que facilitan el reconocimiento de patrones, clasificación y predicción, basados en modelos derivados de datos[71]. De forma sencilla, en el AA se utilizan datos en donde a las i observaciones se les mide un conjunto de j características, lo que se conoce como vector de variables. Estos datos se ordenan en matrices $\mathrm{X} = (x_{ij})$, en donde cada entrada x_{ij} corresponde al valor de la variable j en la observación i (Figura 3).

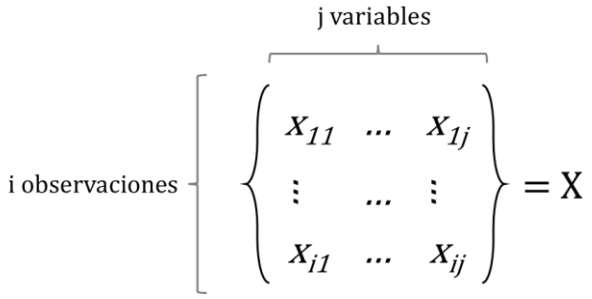


Figura 3: matrices de datos utilizados en AA.

El aprendizaje automático se puede dividir en: aprendizaje supervisado, no supervisado y semi-supervisado. En esta tesis se utilizaron los dos primeros.

1.3.1. Aprendizaje Automático Supervisado.

Los métodos de AA supervisado se utilizan cuando se tienen matrices de datos X que poseen un vector de etiquetas $\mathrm{y_i}$ de dimensión $\mathrm{i \times 1}$. Este vector de etiquetas (o clases) toma tantos valores como etiquetas posibles. Si las observaciones fueran animales domésticos y las variables características fisionómicas, el vector de etiquetas indicaría qué animal es (siendo 1 = perro, 2 = gato, etc). El objetivo de estos métodos es entrenar un modelo que tome como entrada una matriz con datos etiquetados y pueda "distinguir" entre las diferentes clases, basándose en las características asociadas a cada clase. De esta forma, el modelo entrenado podrá predecir la clase de una nueva observación sin etiqueta asociada. A este tipo de algoritmos se los denomina clasificadores. Es posible también predecir variables continuas, para lo cual se utilizan algoritmos de regresión.

Una pregunta que surge de utilizar algoritmos de aprendizaje supervisado es qué tasa de error se comete al predecir la etiqueta de una nueva observación. Para esto, una forma de evaluar el error de un predictor es reservar una parte de los datos etiquetados para entrenar el modelo (muestra de entrenamiento) y otra para evaluarlo (muestra de evaluación). La proporción utilizada para cada grupo de observaciones es variable y depende de cada caso de estudio, pero en general se utiliza entre tres cuartos y cuatro quintos para entrenar y el resto para evaluar. Con la muestra de evaluación es posible crear una matriz de confusión que permita calcular algunas métricas. En la Tabla 1 se muestra un ejemplo de matriz de confusión, resultante de clasificar con dos etiquetas, 1 y 2, en donde la clase de interés que se desea predecir es la clase 1, siendo las observaciones clasificadas en esta clase como las positivas y las de la clase 2 como las negativas.

Tabla 1: Ejemplo matriz de confusión

		Etiquetas predichas	
		Clase 1	**Clase 2**
Etiquetas reales	**Clase 1**	Verdaderos positivos (VP)	Falsos negativos (FN)
	Clase 2	Falsos positivos (FP)	Verdaderos negativos (VN)

El error de predicción de un clasificador es definido como el promedio de observaciones mal clasificadas, calculado como:

$$error = \frac{(FP + FN)}{Total\ de\ observaciones}$$

Por otro lado, la precisión de un modelo es calculada como:

$$precisión = 1 - error$$

Otras métricas para evaluar el desempeño de un modelo son el score F1 y el área bajo la curva ROC (AUC-ROC). Ambas métricas toman valores entre 0 y 1, siendo 1 el valor que corresponde a un modelo que no se equivoca. El score F1 es una medida que combina la precisión y el *recall* en un solo valor, calculado como la media armónica entre ambos:

$$score\ F1 = 2\frac{precisión\ x\ recall}{precisión + recall}$$

En donde el *recall* se calcula como:

$$recall = \frac{VP}{VP + FN}$$

Por otro lado, el área bajo la curva ROC (*Receiver Operating Characteristic*) que mide la capacidad del modelo para discriminar entre clases positivas y negativas, es la representación gráfica de la tasa entre verdaderos positivos (sensibilidad) y falsos positivos (1 - especificidad) para diferentes umbrales de clasificación del modelo.

Los algoritmos de aprendizaje supervisado poseen parámetros e hiperparámetros. Los primeros son aquellos que se “aprenden” durante el entrenamiento y los segundos son predefinidos por el operador. Estos hiperparámetros determinan el comportamiento del algoritmo y pueden influir significativamente en el rendimiento final del modelo. En toda instancia de aprendizaje supervisado es importante pasar por una etapa de optimización de hiperparámetros, en donde se encuentren aquellos valores que se ajustan mejor al conjunto de datos de entrenamiento. Para esto, se utiliza una pequeña parte de la muestra de entrenamiento denominada “muestra de validación”. Durante este proceso, se prueban diferentes combinaciones de hiperparámetros; para cada conjunto de valores, el modelo se entrena con los datos de entrenamiento y luego se evalúan las métricas de desempeño utilizando la muestra de validación. De este modo, se seleccionan los hiperparámetros que producen los mejores resultados en función de las métricas definidas.

Al dividir al azar las observaciones etiquetadas en muestras de entrenamiento y validación, se pueden generar muestras sesgadas, no representativas de la variabilidad original del conjunto de datos. Esto puede llevar a que el modelo se entrene y valide en subconjuntos que no reflejan la distribución original. Además, en matrices con pocas observaciones, se dispone de un número limitado de datos para la validación, lo que puede hacer que el cálculo de las métricas dependa de unos pocos ejemplos y se pierda capacidad de generalización. Por otra parte, los resultados del entrenamiento pueden variar significativamente según la partición realizada, lo que significa que, al cambiar la partición, las métricas también pueden cambiar.

Una forma de mitigar los problemas que pueden surgir de la división entre muestra de entrenamiento y de validación es aplicar la técnica de validación cruzada. La idea detrás de ésta es realizar diferentes particiones y así entrenar y validar con diferentes conjuntos. Con este abordaje se reduce el riesgo de que las muestras sean sesgadas o no representativas de la variabilidad original. Esto permite evaluar el modelo de manera más robusta, ya que el desempeño se calcula promediando los resultados obtenidos en cada partición, lo que disminuye la dependencia de los resultados en una única división de datos. Además, al utilizar diferentes subconjuntos para entrenamiento y validación, se mejora la capacidad de generalización del modelo, ya que se expone a una mayor diversidad de observaciones durante el proceso.

Existen varios métodos de validación cruzada, siendo leave-one-out (LOO) y k-fold los más utilizados. La técnica de LOO consiste en entrenar con n-1 observaciones y etiquetar la observación restante. Este proceso se repite con las n observaciones etiquetadas, lo que puede ser computacionalmente costoso si se tienen muchos datos. Por su parte, k-fold consiste en dividir las observaciones etiquetadas en k partes y entrenar con k-1 partes. Este proceso se repite k veces dejando afuera cada una de las k partes para evaluar. En esta técnica se entrenan tantos modelos como k divisiones se haga, por lo que es más rápido y se recomienda en casos que se tengan muchos datos.

La optimización de hiperparámetros y el uso de técnicas como la validación cruzada son esenciales para obtener un modelo robusto y confiable. En la Figura 4 se muestra el esquema de cómo se utilizan estas etapas en el entrenamiento de un modelo de aprendizaje automático supervisado.

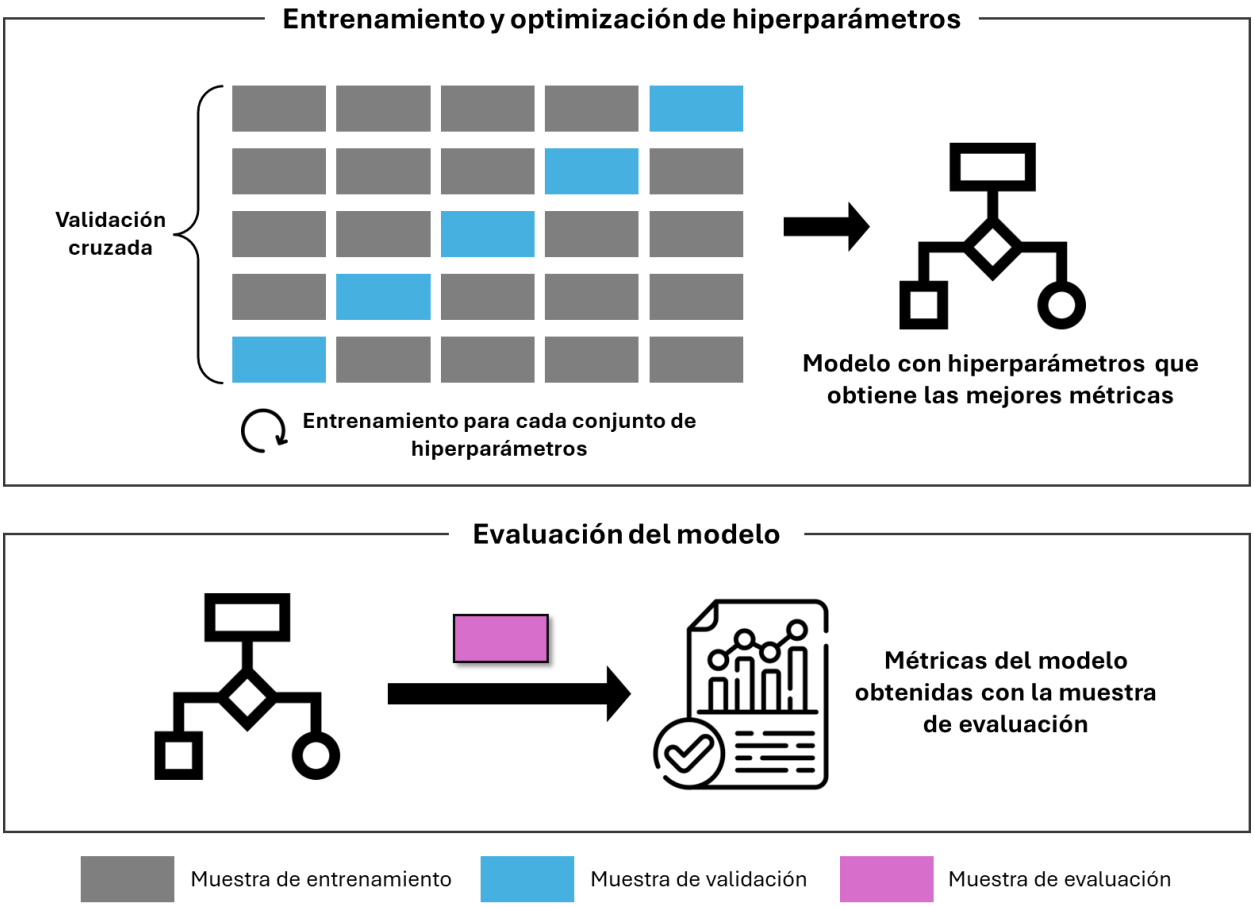


Figura 4: Esquema de entrenamiento de un modelo de aprendizaje automático supervisado. Del conjunto de datos etiquetados se aparta la muestra de evaluación para ser utilizada al final del proceso (violeta). Con el resto de los datos etiquetados se realiza la optimización de hiperparámetros a través de la técnica de validación cruzada, particionando sucesivamente los datos en muestra de entrenamiento (gris) y muestra de validación (azul). Se determina qué parámetros e hiperparámetros dan lugar a las mejores métricas utilizando la muestra validación. Una vez entrenado el modelo, se evalúa su capacidad de predecir correctamente etiquetas con los datos no utilizados para el entrenamiento (muestra de evaluación).

1.3.1.1. Algoritmos de Aprendizaje Automático Supervisado.

Existen muchos algoritmos de aprendizaje supervisado. En el marco de este trabajo se utilizaron tres algoritmos clásicos: *k-Nearest Neighbors (KNN), Support Vector Machine (SVM)* y *Random Forest (RF)*. KNN fue introducido por Evelin Fix y J. L. Hodges en 1951[72] en un trabajo que discute las propiedades de consistencia de métodos de discriminación no paramétricos. Ellos proponen un algoritmo sencillo de clasificación en donde una observación sin etiqueta se clasifica de acuerdo con la etiqueta de la mayoría de sus k vecinos más cercanos. La cercanía de los vecinos se puede establecer según la medida de distancia utilizada, que puede ser Euclidea, de Minkowski o Manhattan, entre otras.

El algoritmo de SVM fue presentado por Corinna Cortes y Vladimir Vapnik en 1995 para el problema de clasificación de dos clases[73]. Conceptualmente, el algoritmo mapea los datos en un espacio de dimensión mayor que el espacio al que pertenecen originalmente aplicando una función (*kernel*), y luego busca un hiperplano que separe ambas clases y maximice el margen

entre ambas, siendo el margen la distancia mínima que existe entre los datos de cada clase y el hiperplano.

Finalmente, RF fue propuesto por Tin Kam Ho en 1995 como una implementación de los árboles de decisión[74]. RF es un algoritmo de ensamble, lo que significa que combina las predicciones de múltiples modelos para generar un único resultado. En este caso, el algoritmo construye un conjunto de árboles de decisión. Durante la construcción de estos árboles, se emplea una técnica conocida como *bagging*, que consiste en dividir la muestra de entrenamiento de manera que cada árbol se construya con un subconjunto diferente de datos. Además, en cada nodo del árbol, RF utiliza un subconjunto aleatorio de las variables predictivas disponibles, lo que contribuye a incrementar la diversidad entre los árboles y a reducir el riesgo de sobreajuste. Al momento de clasificar, RF determina la clase de un dato sin etiquetar en función de la predicción de la mayoría de los árboles.

Los algoritmos utilizados en este trabajo tienen características, fortalezas y limitaciones diferentes. Mientras que KNN se basa en la proximidad de las observaciones, SVM maximiza márgenes entre clases y RF combina múltiples árboles de decisión. Estas diferencias pueden llevar a que cada modelo realice predicciones acertadas en diferentes escenarios, pero también a que presenten sesgos individuales. Una forma de aprovechar las fortalezas de todos estos algoritmos y reducir sus limitaciones es mediante el uso de técnicas de *ensemble*, como lo hace RF. Dentro de estas, *majority label voting*[75] es una de las más simples y efectivas para problemas de clasificación. En este enfoque, cada modelo contribuye con una predicción para una instancia, y la clase final se asigna según el voto mayoritario. Este voto puede ser unánime (todos los modelos coinciden), de mayoría simple (más de la mitad de los modelos coinciden) o de pluralidad (la clase con más votos, incluso si no es la mayoría absoluta). Al combinar las etiquetas predichas de esta manera, *majority label voting* reduce la probabilidad de errores individuales y mejora la precisión global del sistema.

1.3.2. Aprendizaje Automático No Supervisado.

Los algoritmos de AA no supervisado se aplican en datos que no tienen etiquetas asociadas y no necesariamente se conoce la cantidad de etiquetas que existen. Estos algoritmos realizan tareas de agrupamiento -o clusterización-, de aprendizaje por asociación de reglas y reducción de dimensionalidad[76]. En esta tesis nos centraremos en los algoritmos de clusterización.

El objetivo de los algoritmos de clusterización es agrupar los datos de manera que las observaciones dentro de un mismo grupo sean más similares entre sí que con las del resto de los grupos. Retomando el ejemplo de los animales domésticos, al aplicar un algoritmo de clusterización el objetivo sería identificar cuántos grupos de animales existen según sus características fisionómicas y por qué animales están formados. Para lograr esto, existen diversas estrategias, entre las cuales se encuentra la clusterización jerárquica. Esta se divide en dos métodos principales: aglomerativos y divisivos. En los métodos aglomerativos, se comienza con un número de *clusters* igual al número de observaciones, y en cada iteración se agrupan los *clusters* que son menos disímiles. Por otro lado, en los métodos divisivos, se parte de un único *cluster* que se va dividiendo sucesivamente hasta alcanzar el número de *clusters* k deseado. Existen otros algoritmos populares de clusterización que no fueron utilizados en esta tesis, como k-means.

1.3.3. Aprendizaje Automático en el campo de la biología.

En las últimas décadas ha habido un desarrollo acelerado de las tecnologías de secuenciación y en la obtención de datos ómicos (entre éstos, datos genómicos, transcriptómicos, proteómicos y metabolómicos), estructurales, imágenes y estudios clínicos, lo que ha impulsado a científicos y médicos a explorar el campo del AA para procesar, integrar y analizar esta cantidad masiva de datos. Es así como año a año crece la cantidad de publicaciones asociadas al uso de AA en datos biológicos y médicos. En efecto, el AA es ampliamente utilizado en estudios oncológicos[77–80], neurológicos[81,82], epidemiológicos[83,84] y otros estudios clínicos[85–88], en el desarrollo de nuevas drogas[89–91], en biología de sistemas[92,93], biología celular[94] y genética[95], entre otros.

Los avances en la secuenciación de ácidos nucleicos representan uno de los grandes hitos científicos de este siglo. En particular, la secuenciación de ARN de muestras biológicas, incluyendo la secuenciación de células única, ha impulsado significativamente la caracterización funcional de genes, el modelado de redes biológicas, el estudio del *splicing* y variantes, así como el descubrimiento de marcadores moleculares en organismos multicelulares. Estas técnicas generan grandes volúmenes de datos, especialmente en *scRNA-seq*, lo que ha llevado a la implementación de modelos de AA para su análisis.

Estos datos son especialmente valiosos para la predicción de la función génica, basándose en la premisa de que los genes con funciones similares tienden a expresarse con patrones similares. A partir de este concepto, hace algunos años surgió el uso de redes de co-expresión[96]. Estas redes se construyen a partir de datos de expresión, generando grafos en los que cada nodo representa un gen y aquellos genes con expresión correlacionada son unidos por una arista. Mediante técnicas de *clustering* aplicadas a estos grafos, es posible identificar grupos de genes con patrones de expresión correlacionados, lo que ha permitido generar hipótesis sobre la función de una gran cantidad de genes[97–100]. Este abordaje ha sido aplicado en *C. elegans* para la predicción de función génica en varios trabajos[101–104].

Sin embargo, estas aproximaciones presentan limitaciones. A menudo, se identifican muchos genes con expresión similar, lo que redunda en una baja especificidad. Además, la asignación de genes a un mismo clúster depende de las condiciones experimentales del *RNA-seq* y del método de *clustering* utilizado, lo que puede dificultar la interpretación de los resultados y hacer que las relaciones entre los genes dentro de un mismo cluster no siempre sean claras.

Por otro lado, el aprendizaje supervisado ha ganado popularidad en los últimos años en el análisis de datos masivos en el campo de la biología. Estos modelos han sido de gran interés para la clasificación de enfermedades[105,106], en el descubrimiento de biomarcadores[107–109], en el procesamiento de imágenes[110], metagenómica[111–113] y predicción de función de genes[114–116], entre otros. En estudios de *RNA-seq* han sido ampliamente utilizados, ya que estos algoritmos permiten identificar patrones complejos en la expresión que no se evidencian con el uso de métodos estadísticos clásicos[117–119]. En *C. elegans*, algunos trabajos han utilizado este abordaje para el estudio del envejecimiento[120], para predecir genes esenciales[121], ARNs no codificantes[122] y cambios en la accesibilidad de la cromatina durante el desarrollo del organismo[123].

2. Objetivos.

No hay estudios exhaustivos sobre la composición de la CTE y la ATP sintasa de *C. elegans*. Uno de los estudios que realiza el mapeo (*blastp*) de los genes que codifican para las subunidades de los complejos de la fosforilación oxidativa utilizando las proteínas de mamífero como queries[59], no logra identificar a 19 de las 91 proteínas de los complejos I-V de mamíferos. Más aún, en este trabajo se reporta que para 13 genes que codifican para las proteínas de la fosforilación oxidativa de mamíferos, existen al menos dos copias en *C. elegans*. Este trabajo tuvo como objetivo identificar genes candidatos a estar involucrados en la fosforilación oxidativa de *C. elegans* con un abordaje que combina aprendizaje automático supervisado y no supervisado.

2.1. General

Generar una lista de genes candidatos a estar involucrados en el proceso de fosforilación oxidativa en *C. elegans* a partir de datos transcriptómicos mediante el uso de algoritmos de aprendizaje automático.

2.2. Específicos

1. Seleccionar genes previamente identificados como parte de la fosforilación oxidativa en *C. elegans,* a través de una búsqueda bibliográfica, para ser utilizados como ejemplos positivos de la muestra de entrenamiento.
2. Seleccionar genes para ser utilizados como ejemplos negativos de la muestra de entrenamiento aplicando aprendizaje no supervisado sobre una red de co-expresión definida a partir de datos de *scRNA-seq*.
3. Entrenar tres algoritmos de aprendizaje automático supervisado (SVM, KNN, RF) con la lista de entrenamiento diseñada y utilizando datos de *bulk-RNAseq*.
4. Obtener una lista de genes candidatos a estar involucrados en el proceso de fosforilación oxidativa en *C. elegans* a través de un ensemble por mayoría unánime de votos de todos los modelos utilizados.
5. Encontrar evidencia adicional que permita seleccionar los mejores genes candidatos a participar de la fosforilación oxidativa a través de un abordaje bioinformático.

3. Metodología.

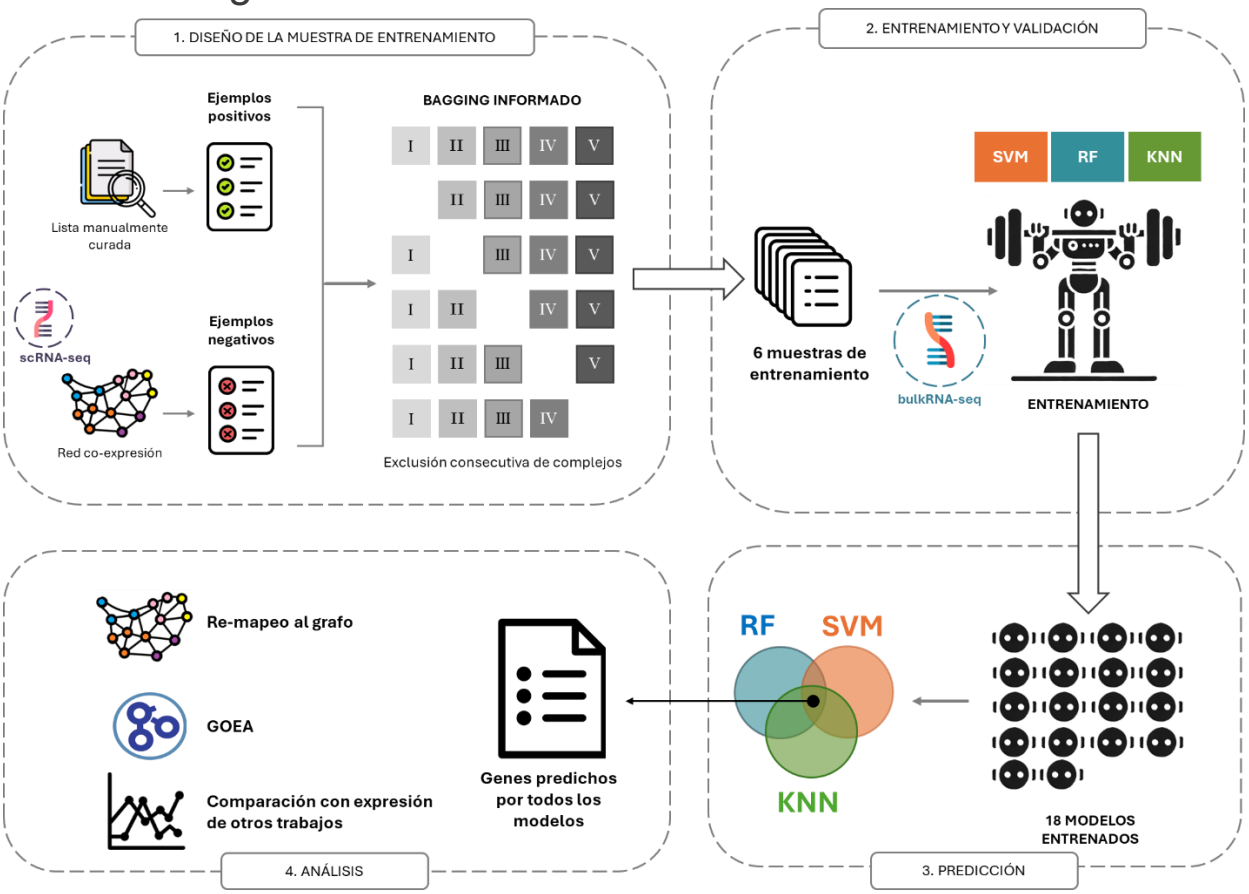


Figura 5: Pipeline del abordaje utilizado en esta tesis. En una primera etapa se seleccionaron los genes utilizados como ejemplos positivos en el entrenamiento a partir de una lista manualmente curada, y los genes utilizados como ejemplos negativos a partir de aplicar la política de hermanos en un grafo de co-expresión obtenido con datos de scRNAseq. Seis muestras de entrenamiento fueron diseñadas a partir de aplicar un bagging informado. En una segunda etapa se entrenaron modelos de aprendizaje supervisado con cada muestra de entrenamiento utilizando tres algoritmos diferentes (SVM, RF y KNN) y datos de bulk RNA-seq, resultando en 18 modelos. En tercer lugar, se utilizaron los modelos para predecir la probabilidad de pertenecer a la clase positiva del resto de los genes de C. elegans, aplicando la técnica de ensemble por mayoría unánime de votos para llegar a la lista consenso de genes candidatos. Finalmente, se buscó evidencia adicional que permitiera seleccionar los mejores candidatos a pertenecer al proceso de fosforilación oxidativa a través de un abordaje bioinformático.

3.1. Diseño de la muestra de entrenamiento.

Con el fin de identificar genes relacionados con el proceso de fosforilación oxidativa aun no descriptos, nos propusimos entrenar modelos de aprendizaje automático supervisado. La selección de los datos etiquetados que formarán la muestra de entrenamiento es crucial, ya que determina en buena medida qué es lo que el modelo aprenderá a reconocer. Para identificar genes relacionados con el proceso de fosforilación oxidativa, es necesario entonces entrenar modelos con dos clases: genes que ya se conoce que pertenecen a ese proceso (ejemplos positivos) y genes que, por algún motivo, se puede asumir que no están involucrados en el mismo (ejemplos negativos).

3.1.1. Selección de ejemplos positivos.

Utilizamos tres criterios para seleccionar ejemplos positivos. Una revisión bibliográfica exhaustiva nos permitió identificar un conjunto de genes con un rol bien establecido en el proceso de fosforilación oxidativa en *C. elegans*[124–126]. Por otro lado, se recurrió a la información depositada en *Wormbase*[127], *Gene ontology*[128] y *Uniprot*[129], bases de datos con información bastante actualizada de este organismo. Adicionalmente, incluimos algunos genes de *C. elegans* con alta homología con genes humanos para los que existe evidencia experimental de que están involucrados en este proceso[130].

3.1.2. Selección de ejemplos negativos.

La selección de ejemplos negativos para la predicción de función de genes es un problema abierto. Idealmente, se deben seleccionar genes cuya función no esté asociada al proceso de fosforilación oxidativa, pero en general no existen trabajos publicados que identifiquen genes sin una función determinada. Existen diferentes estrategias para seleccionar ejemplos negativos en situaciones en las cuales no se posee información curada por expertos[131,132]. En este trabajo decidimos utilizar como criterio los patrones de transcripción de los genes, en el entendido de que los genes involucrados en un mismo proceso biológico comparten patrones de transcripción espacial y/o temporal. Para ello obtuvimos un grafo de co-expresión con el paquete de Python pyWGCNA, el cual clusteriza jerárquicamente los genes cuyos patrones de transcripción estén correlacionados. Como se detalla en la sección 4.2.2, encontramos que los genes previamente identificados por pertenecer al proceso de fosforilación oxidativa se concentran en algunos cluster, por lo que decidimos aplicar la política de hermanos[133] (*siblings policy*) para seleccionar la muestra negativa.

La política de hermanos es una estrategia para diseñar muestras negativas en problemas donde las clases tienen una estructura jerárquica. El principio se basa en la idea de que los ejemplos negativos para una categoría cj deben ser elegidos entre los ejemplos que no son positivos para cj pero si lo son para las categorías hermanas a cj. Esta política se basa en dos intuiciones principales: si el clasificador asociado con el nodo padre de cj no ha generado falsos positivos, entonces el clasificador asociado con cj solo necesitará clasificar observaciones que pertenezcan a cj o a sus hermanos. Por otro lado, al tomar en consideración la estructura topológica de los datos, seleccionar a los hermanos de la categoría cj como ejemplos negativos permite al clasificador trazar una línea entre la clase de interés y ejemplos muy similares, pero que se sabe no pertenecen a la clase de interés.

Para identificar en qué *clusters* se concentran los genes previamente identificados por pertenecer al proceso de fosforilación oxidativa se calculó el enriquecimiento (*fold enrichment*) como:

$$fold\ enrichment\ = \frac{\begin{matrix}Cantidad\ de\ genes\ de\\ interés\ en\ el\ cluster\end{matrix}}{\begin{matrix}Cantidad\ de\ genes\ de\ interés\\ que\ debería\ haber\ en\ el\\ cluster\ por\ azar\end{matrix}}$$

Mientras que el cálculo de p-value se realizó según el test binomial de la siguiente forma:

$$p = \sum_{i=k(C)}^{K} \binom{K}{i} \big(p(C)\big)^{i} \big(1 - p(C)\big)^{K-i}$$

En donde:

- P = p-value.
- K = el número total de genes de interés.
- k(c) = el número de genes de interés que están presentes en el cluster.
- p(c) = la probabilidad de observar un gen del cluster en la lista de genes de interés

3.1.3. *Bagging* informado: exclusión secuencial de complejos.

El proceso de fosforilación oxidativa es llevado a cabo por los cuatro complejos de la CTE y la ATP sintasa. Los complejos I y II de la CTE son los responsables de ceder electrones a la ubiquinona, la cual se encarga de transportarlos al complejo III, que la oxida y cede los electrones al citocromo C, reduciéndolo. El citocromo C transporta estos electrones al complejo IV, el cual los transfiere al oxígeno que es reducido a agua. En este proceso, los complejos I, III y IV bombean protones al espacio intermembrana, los cuales son utilizados por la ATP sintasa para la síntesis de ATP con ayuda de la fuerza protón motriz. Siendo esta la función principal de estos complejos, también participan en otros procesos.

El complejo I es capaz de reducir NADH a NAD+. Estas moléculas no solo participan del transporte de electrones en la cadena, sino que también están involucradas en la homeostasis redox, en la respuesta al daño de ADN, en la expresión génica a través de la modificación de histonas y ADN, en el metabolismo de distintas vías como coenzima, en el ritmo circadiano y en el sistema inmune[134]. Por otro lado, la beta-oxidación de ácidos grasos produce $FADH_2$ y acetil-coA, que ingresa al ciclo de Krebs, para entre otras cosas producir NADH y succinato. El NADH es utilizado por el complejo I, mientras que el succinato es utilizado por el complejo II. El $FADH_2$ generado por la beta-oxidación es utilizado por la enzima ETFDH, la cual transfiere sus electrones a la UQ. Por su parte, los complejos III y IV junto al citocromo C participan de la respuesta antioxidante y de la apoptosis desencadenada por ROS[135]. Por otro lado, el complejo IV es el principal regulador de la CTE, ya que el ATP actúa como regulador negativo de este complejo[136]. En suma, el funcionamiento y la regulación del proceso de fosforilación oxidativa no dependen únicamente del requerimiento energético, sino que cada uno de los complejos está involucrado en numerosos procesos adicionales.

Es esperable entonces que los patrones de transcripción de los genes de un complejo en particular compartan características con los patrones de expresión de genes involucrados en otros procesos. Por lo tanto, al entrenar modelos de aprendizaje automático con esos patrones de transcripción, corremos el riesgo de que el modelo aprenda a reconocer genes cuya función no está directamente relacionada con la fosforilación oxidativa, sino con alguno de estos procesos vinculados.

Como estrategia frente a esta posible fuente de ruido, se entrenó cada algoritmo seis veces, variando la muestra de entrenamiento. El procedimiento está inspirado en la técnica de *bagging*, que consiste en entrenar múltiples modelos en subconjuntos aleatorios del conjunto de datos original (con reemplazo) y combinar sus predicciones. En este trabajo, en vez de muestrear aleatoriamente con reposición, decidimos excluir secuencialmente los genes asociados a cada uno de los complejos. En una de las muestras se incluyeron los genes de todos los complejos, mientras que en las otras cinco se excluyeron los genes correspondientes a cada uno de los complejos proteicos (Figura 5). Luego del entrenamiento y validación, el paso de agregación es similar al del *bagging* standard. Con este procedimiento buscamos disminuir la varianza y mejorar

la robustez del modelo identificando genes relacionados con la fosforilación oxidativa, independientemente de su participación en otros procesos asociados a algún complejo en particular.

3.2. Datasets.

En este trabajo consideramos dos *datasets*: uno para la selección de ejemplos negativos y otro para el entrenamiento los clasificadores. El primero se utilizó para realizar la red de co-expresión con el fin de evaluar si los ejemplos positivos se encontraban distribuidas en un mismo cluster, y así utilizar la política de hermanos para definir los ejemplos negativos. Por otro lado, el segundo *dataset* fue utilizado para entrenar los clasificadores y así predecir la función de genes asociados a este proceso.

3.2.1. Datos para el entrenamiento de los clasificadores.

Para identificar genes relacionados con la fosforilación oxidativa y que previamente no habían sido relacionados con este proceso, se entrenaron algoritmos de aprendizaje automático supervisado. A diferencia de los algoritmos de aprendizaje no supervisado, estos utilizan observaciones etiquetadas de diferentes clases, y aprenden los patrones específicos asociados a los ejemplos que pertenecen a cada clase. Una vez entrenado, el algoritmo es capaz de clasificar una nueva observación no etiquetada asignándole alguna de las clases con cierta probabilidad. Para ello se basa en variables predictivas (o *features*) asociadas a cada uno de los ejemplos. Las variables predictivas asociadas a los genes pueden ser de diferente tipo, ya sean datos cuantitativos continuos, como los datos ómicos; o discretos, como datos de interacción proteína-proteína, datos de variación genética o fenotipos asociados. También pueden ser datos categóricos, como la presencia o ausencia de ciertos dominios proteicos, anotaciones funcionales o datos de variación genética.

En este trabajo se utilizaron datos de transcripción, en el entendido de que los genes que participan de cierto proceso biológico tendrán un patrón de expresión característico en ciertas condiciones. Para esto, se hizo una búsqueda bibliográfica de trabajos con datos disponibles de transcripción en los que las condiciones estudiadas estuvieran relacionadas con la respiración o el consumo de oxígeno (Tabla 2).

El primer *dataset* seleccionado fue el de Boeck *et al*[137]. El objetivo de este trabajo fue generar un conjunto de datos de *RNA-seq* de *C. elegans* con alta resolución temporal durante la embriogénesis y las etapas post-embrionarias a lo largo de su ciclo de vida. Se sabe que el transporte de electrones durante la embriogénesis para la generación de energía a través de la ATP sintasa es esencial en el desarrollo embrionario de *C. elegans*, en donde mutaciones en los distintos complejos tienen diferentes grados de severidad[138], mientras que las vías de obtención de energía durante el desarrollo post-embrionario cambian en los diferentes estadíos larvarios[139]. Cabe esperar que los patrones de expresión de los genes seleccionados para la muestra positiva serán característicos, dependiendo del momento del desarrollo, por lo que este *dataset* muestra ser de interés para nuestro estudio.

Por otro lado, se consideró el trabajo presentado por Gómez-Orte *et al*[140]. En este estudio se realiza un análisis transcriptómico de adultos jóvenes de *C. elegans* alimentados con *Escherichia coli* y *Bacillus subtilis* incubados en tres temperaturas diferentes, con el fin de comparar la expresión génica en estas condiciones. Encontraron que en gusanos alimentados con *E. coli* a temperaturas más altas, se aumenta la expresión de genes relacionados con la defensa y se

disminuye la expresión de genes asociados con el metabolismo, mientras que al ser alimentados con *B. subtilis* se encuentra el efecto contrario. En particular, se encontró diferencias en la expresión de genes asociados al metabolismo de lípidos y ácidos carboxílicos, en donde ambas vías proporcionan sustratos y cofactores reducidos necesarios para la fosforilación oxidativa.

En tercer lugar, se incluyó el trabajo de Kaletsky *et al.*[47], el cual busca estudiar la expresión génica tejido-especifica de adultos de *C. elegans*. En particular, obtuvieron datos de expresión de hipodermis, intestino, neuronas y músculo, tejidos que no poseen la misma disponibilidad de oxígeno ni los mismos requerimientos energéticos, por lo que se espera que los patrones de expresión de la muestra positiva sean distintivos.

Por último, se incluyó el trabajo de Mirza *et al.*[141], que estudia la respuesta patogénica de larvas de *C. elegans* frente la presencia de *Pseudomonas aeruginosa* en diferentes tiempos. Esta bacteria es capaz de producir cianuro de hidrógeno, el cual inhibe la respiración mitocondrial y genera altos niveles de especies reactivas del oxígeno, por lo que la expresión de genes asociados al proceso de fosforilación oxidativa se puede ver afectada en exposiciones prolongadas.

Tabla 2: Datasets seleccionados para entrenar modelos de aprendizaje automático supervisado.

Datasets	Trabajo	Datos	Cantidad de features
***Dataset* 1**	Boeck *et al.*[137]	Expresión de embriones tomados cada 30 min desde la fecundación hasta la formación de la cutícula. Adicionalmente tiene datos de cada una de las 4 etapas larvarias, entrada, durante y salida de dauer, hermafroditas y machos L4, y adultos.	29
***Dataset* 2**	Gómez-Orte *et al.*[140]	Expresión de adultos alimentados con *E. coli* y *B. subtilis* a 15, 20 y 25°C.	6
***Dataset* 3**	Kaletsky *et al.*[47]	Expresión de hipodermis, intestino, neuronas y músculo de adultos.	4
***Dataset* 4**	Mirza *et al.*[141]	Expresión de larvas alimentadas con *P. areuginosa* patógena y no patógena durante 4 y 6 horas.	4

Los *datasets* 2, 3 y 4 se encontraban depositados en el *Sequence Read Archive* (SRA) del *National Center for Biotechnology Information* (NCBI) como archivos FASTQ sin procesar. Para estos casos fue necesario obtener las tablas de conteo normalizadas.

3.2.2. Datos para la selección de ejemplos negativos.

Como se dijo en la sección 3.2, para seleccionar los ejemplos negativos de la muestra de entrenamiento decidimos construir una red de co-expresión. Para ello los datos de expresión deben incluir condiciones, tejidos o células en las cuales se espera que los ejemplos positivos de la muestra de entrenamiento tengan un patrón característico. Siendo así, se seleccionó el trabajo de Packer *et. al*[127] para construir la red.

Este trabajo recoge datos de *scRNA-seq* en embriones de *C. elegans*. En el mismo se aislaron embriones a través de la disgregación de adultos jóvenes utilizando una solución de hipoclorito de sodio e hidróxido de sodio. La cáscara fue digerida con quitinasa y las células fueron

secuenciadas con la tecnología de 10X *Genomics*. Se utilizó el pipeline de CellRanger[143] para el análisis bioinformático y la clusterización de células. Dada la población heterogénea de células obtenidas debido a los diferentes estadíos del desarrollo embrionario en los que se encontraba la muestra, se obtuvieron dos grandes grupos: células terminalmente diferenciadas y células no terminalmente diferenciadas. Luego del análisis se obtuvieron 411 células terminales y 512 células no terminales, dando así un total de 923 células y 17321 genes.

La selección de este *dataset* se realizó bajo la hipótesis de que los genes de la muestra positiva no tendrán la misma expresión en todas las células, ya sea porque el requerimiento energético es diferente, o porque el acceso al oxígeno no es el mismo. Estos patrones de expresión distintivos darían lugar a *clusters* enriquecidos con genes de la muestra positiva. Posteriormente, al seleccionar como ejemplos negativos genes provenientes de *clusters* no enriquecidos con ejemplos positivos, se reduce la probabilidad de incluir falsos negativos como ejemplos negativos en el conjunto de entrenamiento. Esto se puede ver en la Figura 6, en donde se grafica la expresión de los genes en cada una de las células. Observando la figura, es claro que no todos los genes se expresan de la misma forma en todas las células y que algunas de ellas poseen alta expresión de todos los ejemplos positivos.

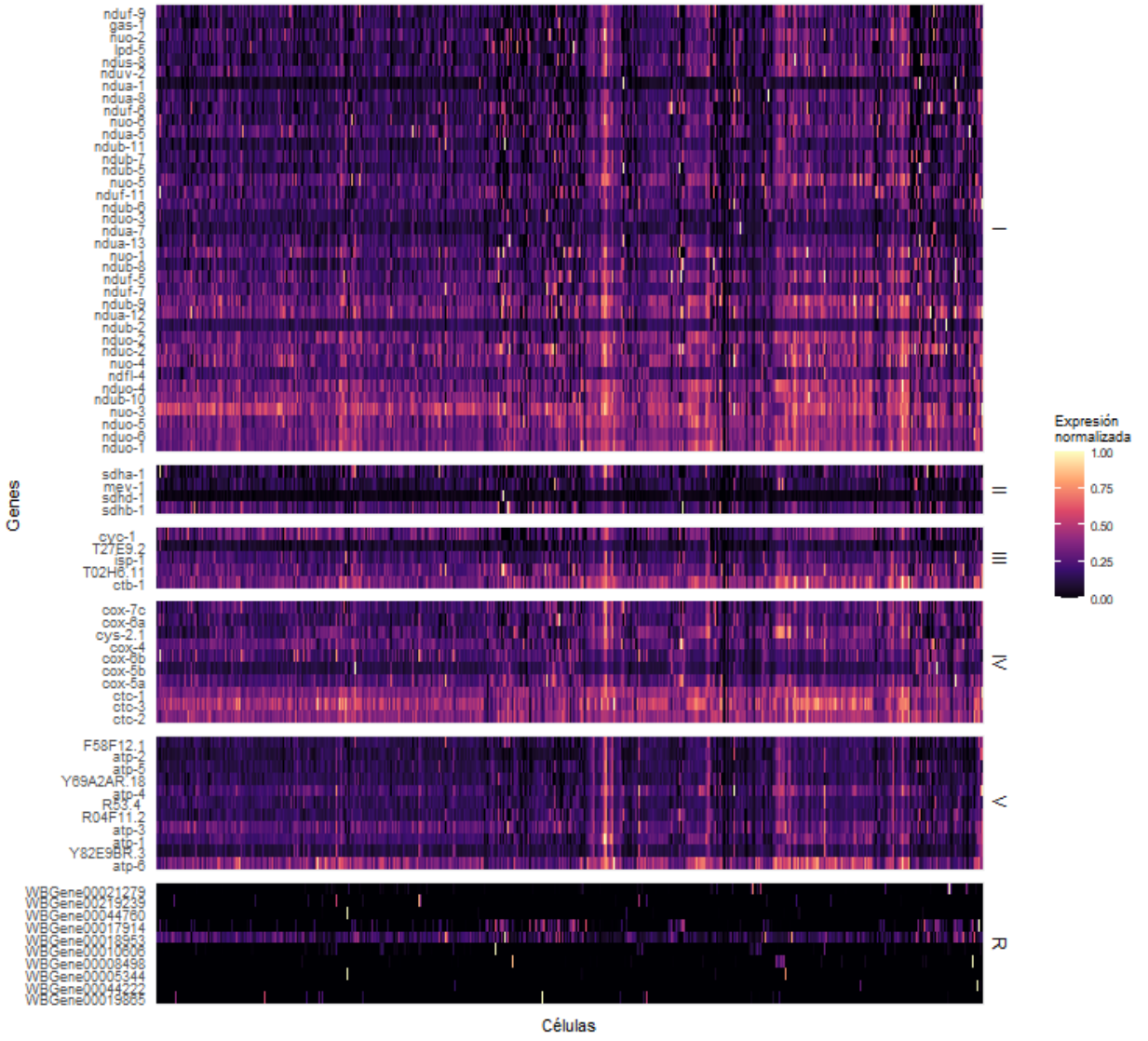


Figura 6: Expresión de los ejemplos positivos en los datos de Packer et al. La expresión se normalizó entre 0 y 1. Los bloques representan la expresión de cada complejo. El último bloque (R) representa la expresión de 10 genes tomados al azar.

3.3. Redes de co-expresión para la selección de ejemplos negativos.

Las redes de co-expresión fueron generadas con el paquete de Python pyWGCNA[144] utilizando los datos de expresión obtenidos por Packer *et. al*[142]. Este software permite construir y analizar redes de co-expresión génica ponderadas, basándose en cálculos de correlación que utilizan un umbral de potencia suave (*soft power thresholding*) y agrupando aquellos genes cuyos conteos de transcriptos estén correlacionados (Figura 7). El proceso de construcción de redes comenzó con el cálculo de una matriz de correlación entre los transcriptos de cada par de genes, a la cual se le aplicó un umbral de potencia suave que permite la detección de relaciones significativas y reduce el ruido. El umbral de potencia suave se utiliza para elevar cada correlación de forma que maximice las correlaciones altas y minimice las correlaciones bajas. Luego se construyó una matriz de superposición topológica, que cuantifica la superposición de vecinos compartidos entre

los genes, proporcionando una medida de interconexión y similitud entre genes más allá de los valores de correlación. Esta matriz de superposición topológica define una red, en la cual luego se buscó identificar módulos de genes co-expresados aplicado clustering jerárquico. Los *clusters* fueron identificados con la función cutreeHybrid() utilizando sus valores por default.

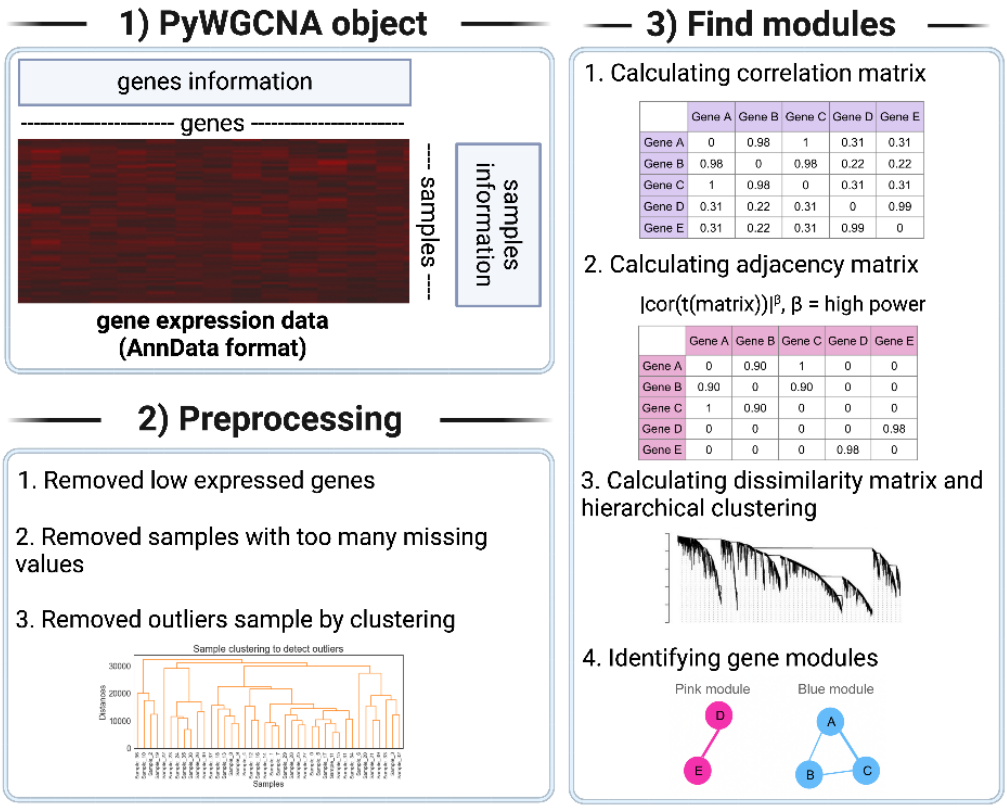


Figura 7: Esquema del pipline del paquete pyWGCNA, adaptado del repositorio oficial en github.

3.4. Conteos de bulk RNA-seq.

Para obtener las tablas de conteo, los archivos FASTQ fueron descargados de NCBI con el programa SRA toolkit v 3.0.10 (https://github.com/ncbi/sra-tools). Las bases y *reads* de baja calidad fueron eliminados (*trimming)* con el programa FastP[145] y su calidad fue evaluada con FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Para el análisis de los datos se utilizó el lenguaje de programación R, siguiendo los siguientes pasos.

a) Construcción del Índice del Genoma

Para alinear las secuencias leídas, fue necesario construir un índice del genoma de *C. elegans*, usando su versión WBcel235. Esto se realizó con la librería Rsubread[146], que permite indexar el genoma a partir de un archivo FASTA.

b) Mapeo de Lecturas

Una vez construido el índice del genoma, se procedió a mapear las lecturas de secuencias utilizando nuevamente Rsubread. Se listaron los archivos FASTQ de interés y se dividieron en

archivos de lecturas emparejadas. Los archivos resultantes del mapeo fueron almacenados en formato BAM.

c) Conteo de Lecturas Alineadas

Para contar las lecturas alineadas a segmentos específicos del genoma, como genes, se utilizó la función featureCounts de Rsubread. Esta función permite asignar lecturas a genes basándose en un archivo de anotación en formato GTF. Los resultados del conteo se guardaron para su posterior análisis.

d) Anotación de Genes

Con la librería biomaRt, se recuperaron datos de anotación de genes desde la base de datos Ensembl. Esta anotación incluyó identificadores de genes, descripciones, tipos de genes, y otros atributos relevantes.

e) Filtrado de Genes con Baja Expresión

Finalmente, se filtraron los genes de baja expresión, manteniéndose aquellos que presentaran por lo menos 1 CPM (counts per million) en las tres réplicas.

3.4.1. Normalización.

Una vez obtenida las tablas de conteo, fue necesario determinar el tipo de normalización a utilizar. Existen varios tipos de normalizaciones de conteos de genes, y en este trabajo consideramos dos: *Trimmed Mean of M-values* (TMM)[147], *Transcripts Per Million* (TPM)[148], y una variación de esta última (*centered log-ratio transformation*).

Por un lado, TMM es una estrategia empírica que fue propuesta originalmente para equilibrar los niveles de expresión génica entre muestras en análisis de expresión diferencial de datos de RNA-seq. Este método calcula un factor de normalización para cada gen tomando en cuenta el logaritmo de los *ratios* de expresión del gen entre muestras (M-values) y recortando un porcentaje de los genes con M-values altos (*trimming*). El método TMM ha demostrado ser robusto en simulaciones y comparaciones con otros métodos de normalización, mejorando la detección de genes diferencialmente expresados en datos de RNA-seq[149,150], y ha sido usado para aplicaciones de aprendizaje automático[151,152].

Por otro lado, TPM es una variación de *Reads Per Kilobase Million* (RPKM) que surgió ante la necesidad de comprar la abundancia relativa de genes en muestras diferentes[148]. Mientras que el RPKM normaliza por el número total de reads y puede verse afectado por la longitud de los transcritos y la distribución del tamaño del RNA, el TPM normaliza por el número total de transcritos muestreados y corrige de manera más efectiva las diferencias en la longitud de los genes y la composición de la librería[148]. TPM también ha sido utilizado para predicciones basadas en aprendizaje automático[153–155].

Finalmente, un trabajo de 2020[156] demostró que modelos de aprendizaje automático lineales se desempeñaban mejor aplicando una transformación de logaritmo centrada (*centered log-ratio transformation – CLR*) a los datos de expresión ya normalizados por TPM (TPMCLR).

En este trabajo se decidió normalizar los *datasets* con estos tres métodos, utilizando la función *calcNormFactors* de la librería de R *edgeR*[157] para normalizar por TMM. Para calcular el valor correspondiente en TPM se utilizó la siguiente ecuación:

$$\mathrm{TPM_i} = \frac{\mathrm{CPM_i} \times 10^6}{\sum_j \left(\mathrm{CPM_{ij}} \times \mathrm{norm\_factor_j}\right)}$$

En donde:

- $\mathrm{TPM_i}$ es el valor TPM de la muestra i.
- $\mathrm{CPM_i}$ es el valor CPM de la muestra i.
- $\mathrm{norm\_factor_j}$ el factor de normalización TMM para la muestra j.

3.5. Entrenamiento de modelos.

Los modelos de aprendizaje automático supervisado fueron entrenados en el lenguaje Python, utilizando las librerías Pandas, Numpy, Matplotlib y scikit-learn (sklearn). Se dividieron los datos etiquetados en una muestra de entrenamiento (X_train) y de evaluación (y_train). Los datos de X_train son los que efectivamente se utilizaron para entrenar y validar los modelos, mientras que los datos de y_train se utilizaron para evaluar la capacidad predictiva de los modelos entrenados. Esta división se realizó con la función train_test_split() de sklearn utilizando un 25% de los datos para evaluación.

Los algoritmos utilizados fueron RF (RandomForestClassifier()), SVM (scv()) y KNN (KNeighborsClassifier()). Para el entrenamiento de SVM y KNN los datos fueron previamente estandarizados con la función StandardScaler(). Los hiperparámetros fueron optimizados con la función GridSearchCV() de sklearn, optimizando sobre el score F1 (scoring='f1') y utilizando una validación cruzada de 5 fold. Los hiperparámetros y las grillas utilizadas se encuentran en la Tabla 3.

Tabla 3: Hiperparámetros optimizados con la función GridSearchCV() de sklearn.

Algoritmo	Hiperparámetro	Rango
RF	n_estimators	np.arange(1, 20, 1)
	max_depth	np.arange(1, 20, 1)
	min_samples_split	np.arange(1, 20, 1)
	min_samples_leaf	np.arange(1, 20, 1)
SVM	kernel	'linear', 'poly', 'rbf', 'sigmoid'
	C	np.arange(0.1, 20, 1)
	gamma	np.arange(0.01, 10, 0.5)
KNN	n_neighbors	np.arange(2, 30, 2)
	weights	'uniform', 'distance'
	algorithm	'auto', 'ball_tree', 'kd_tree', 'brute'
	leaf_size	np.arange(1, 100, 10)

Con la muestra de evaluación se calcularon los umbrales de clasificación que optimizan el score F1 y el área bajo la curva ROC, utilizando la función precision_recall_curve() de sklearn.

3.6. Caracterización de la lista consenso

Con el fin de encontrar evidencia adicional que vincule los genes de la lista consenso (los genes votados por todos los modelos) con el proceso de fosforilación oxidativa, se llevaron a cabo varios

análisis adicionales, buscando a su vez priorizar a los mejores candidatos para una eventual validación experimental.

3.6.1. Enriquecimiento funcional en términos de Gene Ontology (GOEA).

Los enriquecimientos funcionales en términos de Gene Ontology (GOEA) se realizaron con el paquete de Python GOATOOLS[158]. Las gráficas fueron diseñadas con el paquete matplotlib.

3.6.2. Expresión de genes de la lista consenso en otros trabajos.

Se realizó una búsqueda bibliográfica de trabajos en los que los ejemplos positivos se encontraran diferencialmente expresados. Se consideraron dos trabajos. El trabajo de Priebe *et al.*[159] estudia la prolongación de la vida de *C. elegans* mediante la alimentación con una dieta restrictiva en glucosa. Este trabajo proporciona en su material suplementario la lista de genes de *C. elegans*, diferencialmente expresados en animales alimentados con DOG en diferentes estadíos.

El trabajo de Hammarlund et. al.[160] incluye un atlas completo de expresión génica de todo el sistema nervioso de *C. elegans* a resolución de neuronas individuales. Este trabajo proporciona un archivo que contiene un objeto Seurat el cual puede ser manipulado en R para obtener datos del estudio. A partir de estos datos, identificamos los tipos celulares en los cuales se expresan los genes utilizados como ejemplos positivos, con el objetivo de verificar si los genes de la lista consenso también se expresan en esos tipos celulares. Para ello, seleccionamos los tipos celulares en los que al menos la mitad de los genes positivos se encontraran expresados. Sin embargo, dado que en algunos casos había muy pocas células dentro de un tipo celular que expresaran estos genes, decidimos conservar únicamente los tipos celulares en los cuales los genes positivos se expresaran en al menos el 20% de las células. Luego, dentro de estos tipos celulares seleccionados, filtramos los genes de la lista consenso para conservar aquellos que también se expresaban en al menos el 20% de las células del tipo celular. Estos cálculos se hicieron en el lenguaje R en RStudio.

3.6.3. Inferencia estructural y búsquedas con Foldseek.

Para la inferencia de la estructura de los genes que no tenían homología de secuencia con genes humanos, utilizamos AlphaFold[161], una herramienta desarrollada por DeepMind para predecir la estructura tridimensional de las proteínas a partir de su secuencia de aminoácidos. AlphaFold emplea redes neuronales profundas entrenadas con una amplia base de datos de estructuras proteicas conocidas. Estas redes aprenden a mapear secuencias de aminoácidos a estructuras tridimensionales mediante el uso de información evolutiva y física. La principal innovación de AlphaFold radica en su capacidad para predecir la posición de cada átomo en una proteína con una precisión comparable a la obtenida mediante técnicas experimentales como la cristalografía de rayos X y la resonancia magnética nuclear.

Se utilizó ColabFold[162], una implementación optimizada de AlphaFold que permite realizar inferencias estructurales utilizando Google Colab. Se fijó una relajación de 1 para inferir la estructura. La relajación de una estructura se da luego de su inferencia, realizando pequeñas perturbaciones en el descenso de gradiente en el campo de fuerza AMBAR, un paso que mejora la estabilidad y elimina interacciones no favorables.

Para evaluar la confiabilidad de las estructuras predichas se utilizó el índice pLDDT (predicted Local Distance Difference Test). El pLDDT es una métrica de confianza implementada en

AlphaFold que varía entre 0 y 100. Valores de pLDDT superiores a 70 indican alta confianza en la precisión de la estructura en esa región. Se cortaron los extremos de las proteínas que tenían pLDDT menores a 70 utilizando el programa PyMOL[163].

Para identificar posibles homologías estructurales con proteínas de humanos o levaduras, se empleó Foldseek, una herramienta de búsqueda de similitud estructural. Foldseek es capaz de comparar rápidamente estructuras proteicas y calcular varios *scores*, entre los cuales destacan el TM-score (*Template Modeling score*) y el RMSD (*Root Mean Square Deviation*). El TM-score proporciona una medida de la similitud global entre dos estructuras, con valores que van de 0 a 1, siendo 1 una coincidencia perfecta. El RMSD mide la distancia promedio entre los átomos correspondientes de dos estructuras superpuestas. Foldseek permitió buscar homología estructural entre las estructuras inferidas para genes de la lista consenso sin homólogos en humanos y las estructuras proteicas depositadas en bases de datos de proteínas de humanos y levaduras, evaluando si existían similitudes significativas que no se detectan a nivel de secuencia.

3.7. Códigos y scripts.

Todos los códigos y scripts se encuentran depositados en https://github.com/SofiaZeballos/Thesis_scripts.

4. Resultados

4.1. Selección de ejemplos positivos.

La selección de ejemplos positivos para la muestra de entrenamiento se hizo a partir de la integración de datos de varias fuentes. La secuencia de los genes de *C. elegans* y otros nematodos se encuentra depositada en la Wormbase, mientras que la función y localización de los productos génicos se encuentra en Gene Ontology. En *C. elegans*, diversos autores han descrito algunos de los genes que codifican componentes de los complejos de la cadena de transporte de electrones y la ATP sintasa[124–126]. Por otro lado, muchos de estos genes se han anotado por homología de secuencia con genes involucrados en este proceso en otros organismos. En la etapa inicial de este trabajo, unificamos la información dispersa mediante tres enfoques: búsqueda exhaustiva en la literatura científica, comparación mediante blast de genes caracterizados en humanos[130] con el genoma de *C. elegans*, y consulta en bases de datos como Wormbase, UniProt y Gene Ontology. Este proceso fue esencial, dado que, a pesar de la existencia de estudios previos, la información relevante no estaba centralizada en una única fuente.

En la Tabla 4 se resume la lista final de los 68 genes seleccionados como ejemplos positivos. Tras el relevamiento de genes involucrados en este proceso definimos tres clases de proteínas: las proteínas *core*, que conforman el núcleo estructural y funcional de los complejos; las proteínas accesorias (exclusivas del complejo I en *C. elegans*), que ayudan en la regulación, estabilidad y protección de los complejos de la cadena respiratoria y forman parte de estos sin estar en los sitios catalíticos; y las proteínas de ensamblaje, que son esenciales para el correcto ensamblaje de los componentes proteicos y de los cofactores dentro de los complejos, y que luego de participar del ensamblaje se separan de los mismos. Para la lista de ejemplos positivo sólo se incluyeron los genes *core* y accesorios. La lista total de genes se encuentra en la Tabla S1, en donde también se incluye enfermedades humanas asociadas a los ortólogos de humanos.

Tabla 4: Genes que participan en el proceso de fosforilación oxidativa de C. elegans y su respectivo ortólogo en humanos en los casos correspondientes. Los genes en gris de la columna "Genes en H. sapiens" corresponden a genes mitocondriales los cuales no tienen homología de secuencia con los genes de C. elegans, pero cumplen la misma función.

Complejo	Gen en C. elegans	Clase	Evidencia	Gen en *H. sapiens*
I	*gas-1*	Core	Bibliográfica[164]	*NDUS2*
I	*ndfl-4*	Core	Bibliográfica[165]	*ND4L*
I	*ndub-6*	Core	GO	*NDUFB6*
I	*nduf-7*	Core	Homología de secuencia de humano	*NDUFS7*
I	*nduo-1*	Core	Bibliográfica[165]	*ND1*
I	*nduo-2*	Core	Bibliográfica[165]	*ND2*
I	*nduo-3*	Core	Bibliográfica[165]	*ND3*
I	*nduo-4*	Core	Bibliográfica[165]	*ND4*
I	*nduo-5*	Core	Bibliográfica[165]	*ND5*
I	*nduo-6*	Core	Bibliográfica[165]	*ND5*
I	*ndus-8*	Core	Homología de secuencia de humano	*NDUFS8*
I	*nduv-2*	Core	Homología de secuencia de humano	*NDUFV2*
I	*nuo-1*	Core	Homología de secuencia de humano	*NDUFV1*
I	*nuo-2*	Core	Homología de secuencia de humano	*NDUS3*
I	*nuo-3*	Core	Homología de secuencia de humano	*NDUA6*

I	*nuo-5*	Core	Homología de secuencia de humano	*NDUFS1*
I	*nuo-6*	Core	Bibliográfica[166]	*NDUB4*
I	*lpd-5*	Accesoria	Homología de secuencia de humano	*NDUFS4*
I	*ndua-1*	Accesoria	Ortólogo y GO	*NDUFA1*
I	*ndua-12*	Accesoria	Homología de secuencia de humano	*NDUFA12*
I	*ndua-13*	Accesoria	Homología de secuencia de humano	*NDUFA13*
I	*ndua-5*	Accesoria	Homología de secuencia de humano	*NDUA5*
I	*ndua-7*	Accesoria	Homología de secuencia de humano	*NDUA7*
I	*ndua-8*	Accesoria	Homología de secuencia de humano	*NDUA8*
I	*ndub-10*	Accesoria	Homología de secuencia de humano	*NDUFB10*
I	*ndub-11*	Accesoria	Homología de secuencia de humano	*NDUFB11*
I	*ndub-2*	Accesoria	Homología de secuencia de humano	*NDUFB2*
I	*ndub-5*	Accesoria	Homología de secuencia de humano	*NDUFB5*
I	*ndub-7*	Accesoria	Homología de secuencia de humano	*NDUFB7*
I	*ndub-8*	Accesoria	Homología de secuencia de humano	*NDUFB8*
I	*ndub-9*	Accesoria	Homología de secuencia de humano	*NDUFB9*
I	*nduc-2*	Accesoria	Homología de secuencia de humano	*NDUFC2*
I	*nduf-11*	Accesoria	Bibliografica[167]	*NDUFA11*
I	*nduf-5*	Accesoria	Homología de secuencia de humano	*NDUFS5*
I	*nduf-6*	Accesoria	Homología de secuencia de humano	*NDUFS6*
I	*nduf-9*	Accesoria	Homología de secuencia de humano	*NDUFA9*
I	*nuo-4*	Accesoria	Homología de secuencia de humano	*NDUAA*
II	*mev-1*	Core	Bibliográfica[126]	*SDHC*
II	*sdha-1*	Core	Bibliográfica[126]	*SDHA*
II	*sdhb-1*	Core	Bibliográfica[126]	*SDHB*
II	*sdhd-1*	Core	Bibliográfica[126]	*SDHD*
III	*ctb-1*	Core	Bibliográfica[165]	*CYB*
III	*cyc-1*	Core	Homología de secuencia de humano	*CYC1*
III	*isp-1*	Core	Bibliográfica[168]	*UQCRFS1*
III	*T02H6.11*	Core	Homología de secuencia de humano	*UQCRB*
III	*T27E9.2*	Core	Homología de secuencia de humano	*UQCRH*
IV	*cox-4*	Core	Homología de secuencia de humano	*COX4I1*
IV	*cox-5a*	Core	Homología de secuencia de humano	*COX5A*
IV	*cox-5b*	Core	Homología de secuencia de humano	*COX5B*
IV	*cox-6a*	Core	Homología de secuencia de humano	*COX6A*
IV	*cox-6b*	Core	Homología de secuencia de humano	*COX6B2*
IV	*cox-7c*	Core	Homología de secuencia de humano	*COX7C*
IV	*ctc-1*	Core	Bibliográfica[165]	*COI*
IV	*ctc-2*	Core	Bibliográfica[165]	*COII*
IV	*ctc-3*	Core	Bibliográfica[165]	*COII*
IV	*cys-2.1*	Core	Homología de secuencia de humano	*CYCS*
V	*atp-1*	Core	Ortólogo	*ATP5F1A*
V	*atp-2*	Core	Homología de secuencia de humano	*ATP5F1B*
V	*atp-3*	Core	Homología de secuencia de humano	*ATP5PO*
V	*atp-4*	Core	Homología de secuencia de humano	*ATP5PF*
V	*atp-5*	Core	Homología de secuencia de humano	*ATP5PD*
V	*atp-6*	Core	Wormbase y GO	*ATP6*
V	*F58F12.1*	Core	Homología de secuencia de humano	*ATP5F1D*
V	*R04F11.2*	Core	Homología de secuencia de humano	*ATP5ME*
V	*R53.4*	Core	Homología de secuencia de humano	*ATP5MF*

V	*Y116A8C.27*	Core	Homología de secuencia de humano	*ATPAF2*
V	*Y69A2AR.18*	Core	Homología de secuencia de humano	*ATP5F1C*
V	*Y82E9BR.3*	Core	Homología de secuencia de humano	*ATP5MC1, ATP5MC2, ATP5MC3*

Al realizar la búsqueda de genes asociados al proceso de fosforilación oxidativa en *C. elegans*, se encontraron genes sin homología de secuencia con humanos ni levaduras, sin evidencia bibliográfica, pero con anotaciones asociadas a los complejos en Gene Ontology con códigos de evidencia IBA (*inferred from biological aspect of ancestor*) o IEA (*inferred from electronic annotation*). Estos genes no fueron incluidos en la muestra de entrenamiento y se reservaron para corroborar si eran recuperados en las predicciones. Por otro lado, se identificaron varios genes con duplicaciones génicas, para las cuales no se definió un criterio de selección entre las múltiples copias, excepto en algún caso particular, como en las duplicaciones génicas de *gas-1* (parálogo, *nduf-2.2*) y *sdha-1* (parálogo, *sdha-2*), ya que existe evidencia previa que *nduf-2.2* y *sdha-2* son de expresión exclusiva en línea germinal[169]. Los genes duplicados, entonces, fueron excluidos de la muestra de entrenamiento, salvo excepciones fundamentadas, y se reservaron para verificar si alguno de los parálogos era recuperado en las predicciones. Estos genes se encuentran listados en la Tabla 5.

Tabla 5: Lista de genes que han sido asociados con el proceso de fosforilación oxidativa, pero sobre lo cual no hay evidencia clara y que fueron excluidos de la muestra de entrenamiento.
**ND: no definido. **IBA: Inferred from Biological aspect of Ancestor. ***IEA: Inferred from Electronic Annotation*

Complejo	Gen en *C. elegans*	Clase	Evidencia	Motivo de la exclusión
I	*ndub-3*	ND	GO	Inconsistencias en su anotación. No está definido si es de ensamblaje o accesoria
I	*C06A5.3*	ND*	GO	Anotado en GO con código IBA**, sin sustento bibliográfico.
I	*ndab-1*	Accesoria	Ortólogo de humano	Duplicado con *ndab-2*.
I	*ndab-2*	Accesoria	Ortólogo de humano	Duplicado con *ndab-1*.
I	*nduv-3*	ND	GO	Anotado en GO con código IEA***, sin sustento bibliográfico.
III	*C14B9.10*	Accesoria	GO	Anotado en GO con código IEA, sin sustento bibliográfico.
III	*F45H10.2*	Core	Ortólogo de humano	Duplicado con *R07E4.3*.
III	*R07E4.3*	Core	Ortólogo de humano	Duplicado con *F45H10.2*.
III	*mppb-1*	Core	Ortólogo de humano	Duplicado con *ucr-1*.
III	*ucr-1*	Core	Ortólogo de humano	Duplicado con *mppb-1*.
III	*ucr-11*	ND	GO	Anotado en GO con código IEA, sin sustento bibliográfico.
III	*ucr-2.1*	Core	Ortólogo de humano	Triplicado con *ucr-2.2* y *ucr-2.3*.
III	*ucr-2.2*	Core	Ortólogo de humano	Triplicado con *ucr-2.1* y *ucr-2.3*.
III	*ucr-2.3*	Core	Ortólogo de humano	Triplicado con *ucr-2.1* y *ucr-2.2*.
IV	*B0035.18*	Core	Homología con cox-6c	Triplicado con *cox-6c* y *Y111B2A.2*.
IV	*cox-6c*	Core	Wormbase	Triplicado con *B0035.18* y *Y111B2A.2*.

IV	*Y111B2A.2*	Core	Homología con cox-6c	Triplicado con *cox-6c* y *B0035.18*.
V	*asb-1*	Core	Ortólogo de humano	Duplicado con *asb-2*.
V	*asb-2*	Core	Ortólogo de humano	Duplicado con *asb-1*.
V	*asg-1*	ND	Ortólogo de humano	Duplicado con *asg-2*.
V	*asg-2*	ND	Ortólogo de humano	Duplicado con *asg-1*.
V	*hpo-18*	Core	Ortólogo de humano	Triplicado con *R05D3.6* y *ZC262.5*.
V	*R05D3.6*	Core	Ortólogo de humano	Triplicado con *hpo-18* y *ZC262.5*.
V	*ZC262.5*	Core	Ortólogo de humano	Triplicado con *hpo-18* y *R05D3.6*.

En comparación a la descripción más detallada que se tenía sobre los genes de la fosforilación oxidativa en *C. elegans* en base a homología de secuencia (Tsang y Lemire[59]), identificamos 11 genes adicionales, dos por homología (*atp-4* y *ucr-2.3*), seis por referencias bibliográficas (*ndua-1, ndub-11, nduf-11* y *nuo-6*)[54,124,167] y tres por anotación en Gene Ontology (*C06A5.3, nduv-3* y *ucr-11*). Adicionalmente, agregamos los genes *cys-2.1* y *cys-2.2*. En cambio, el trabajo de Tsang y Lemire incluye los genes *ndua-2* y *mrpl-4* (duplicados) los cuales participan del ensamblado; los genes *mai-1* y *mai-2,* los cuales regulan la actividad del complejo V; el gen *F48E8.3,* el cual lo catalogan como un duplicado de *sdha-1,* el cual descartamos por la poca identidad con ellos; y el gen *F16B4.6*, el cual lo presentan como un duplicado de *ndab-1* y *ndab-2* pero no posee homología de secuencia con ninguno de los dos.

4.2. Selección de ejemplos negativos.

4.2.1. Exploración de los datos.

Dado el método de obtención de datos de sc-RNAseq de *Packer et al.*, la cantidad de transcriptos que se puede recuperar de cada gen es notablemente baja. Para muchos genes se observa un valor de expresión de 0 en varias células, que se puede deber a sesgos inherentes a la metodología utilizada o a la nula expresión en esas células. En este contexto, se evaluó la distribución de valores nulos por gen. Los resultados se presentan en la Figura 8.

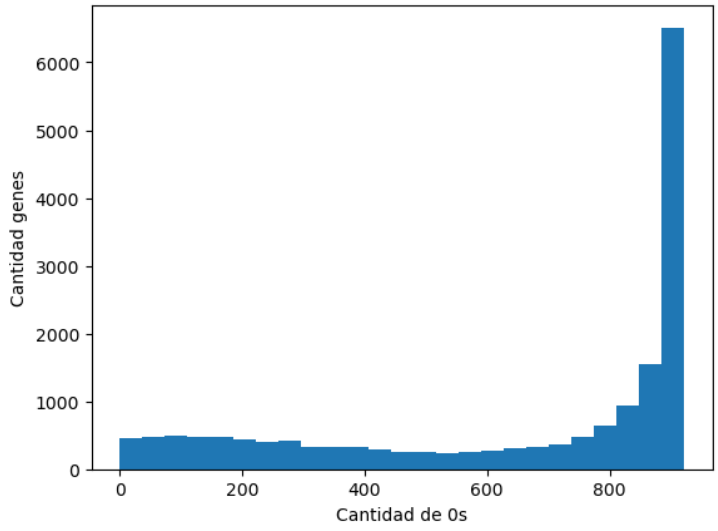


Figura 8: Cantidad de genes con valores nulos de expresión según Packer et. al, que considera 923 células y 17321 genes.

En la Figura 8 se observa que una proporción significativa de los genes presentan valores nulos en más de 800 células. Dada esta característica de los datos, se optó por construir tres redes de co-expresión utilizando tres conjuntos diferentes: todos los genes, el conjunto de genes con menos de 900 ceros (se expresan en más de 23 células) y el conjunto con menos de 800 ceros (se expresan en más de 123 células).

El paquete de Python utilizado para construir las tres redes de co-expresión fue pyWGCNA, y los resultados se encuentran en el anexo (Figura S1, Figura S2 y Figura S3). Con el fin de determinar cuál de estas redes se ajustaba mejor a los objetivos del estudio se llevó a cabo un *clustering* jerárquico de cada una y se identificó en cada caso el *cluster* con la mayor cantidad de genes usados como ejemplos positivos. Posteriormente, se calculó el enriquecimiento en ejemplos positivos de cada uno de estos tres *clusters* (Tabla 6).

Tabla 6: Resultados de construcción de redes de co-expresión de tres subconjuntos de los datos obtenidos por Packer et. al. *En cada caso, el cluster de interés es aquel que contiene la mayor cantidad de genes que fueron usados como ejemplos positivos.*

Genes considerados	Cantidad de genes	Cantidad de *clusters*	Total de genes en el cluster de interés	Cantidad de genes de la clase positiva en el *cluster* de interés	Fold Enrichment	p-value
Todos los genes	17320	57	718	48	16,7	1,94E-242
Genes con menos de 900 0s	11903	31	528	40	13,3	1,73E-173
Genes con menos de 800 0s	8102	16	329	38	13,8	8,77E-95

En Tabla 6 se observa que al considerar todos los genes se forma el cluster más enriquecido en ejemplos positivos. Por este motivo se decidió continuar con el conjunto total de genes.

4.2.2. Redes de co-expresión.

La Figura 9 muestra el dendrograma resultante del clustering jerárquico de la red de co-expresión. En el gráfico, las barras representan la cantidad de genes en cada *cluster*, y el gradiente de color indica el enriquecimiento (fold enrichment) de cada *cluster* en genes usados como ejemplos positivos. Se destaca la presencia de un *cluster* particularmente enriquecido en estos genes, con un fold enrichment de 16,7 y un p-valor de $1{,}94e10^{-242}$ (de ahora en adelante llamado “cluster A”), mientras que el fold enrichment de los demás clusters es menor a 2.

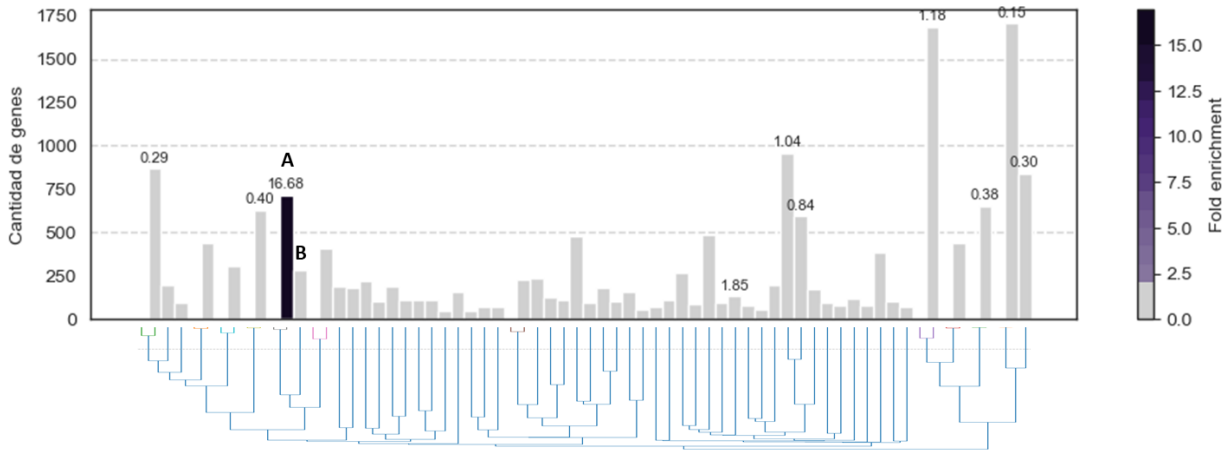


Figura 9: Dendrograma correspondiente al clustering jerárquico de la red de co-expresión generada con el software pyWGCNA. La altura de las barras indica la cantidad de genes en cada cluster. El gradiente de color indica el enriquecimiento en ejemplos positivos de cada cluster, con gris para fold enrichments menores a 2. En aquellos cluster con enriquecimiento diferente a 0, se indica su valor sobre las barras. Los p-valores de los enriquecimientos fueron menores a $1x10^{-20}$ en todos los casos.

Una vez generada la red, y habiendo encontrado que los ejemplos positivos se concentran en unos pocos *clusters* (particularmente en el *cluster* A) decidimos aplicar el criterio de la *siblings policy* para seleccionar ejemplos negativos[133]. Como se observa en la Figura 9, el *cluster* hermano al cluster A (de 282 genes, llamado de aquí en adelante "*cluster* B"), es aquel que comparte el mismo nodo parental y que por definición está compuesto por genes cuyos perfiles de expresión son similares en un nivel jerárquico anterior, pero que los diferencia del *cluster* A.

Con el fin caracterizar ambos *clusters*, se hizo un análisis de enriquecimiento funcional en términos de Gene Ontology. En la Figura 10 se puede ver que el cluster A está enriquecido, además del proceso de la fosforilación oxidativa, en procesos como traducción, síntesis de ATP asociada la fuerza de movimiento, y la organización e importación de proteínas a la mitocondria, todos términos relacionados con el proceso de fosforilación oxidativa. Por otro lado, el cluster B está enriquecido en genes asociados a la localización y diferenciación del músculo, a la formación de colágeno y a funciones asociadas a la unión a actina y zinc. Estos términos dan indicio de que a este *cluster* lo conforman genes responsables de la formación de músculo, tejido que requiere una gran cantidad de energía, el cual se encuentra enriquecido en mitocondrias.

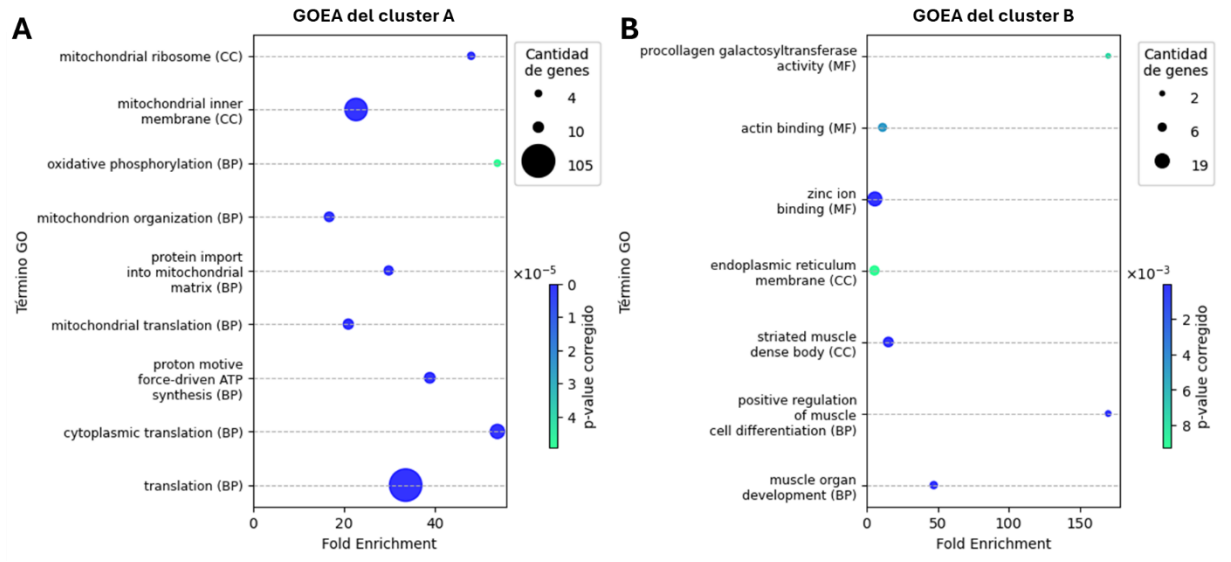


Figura 10: Análisis de enriquecimiento funcional en términos de Gene Ontology (GOEA). (A) GOEA del cluster enriquecido en genes de la muestra positiva y (B) GOEA del cluster contiguo.

Los genes *hub* de un cluster se definen como los genes más conectados dentro de ese *cluster* y son representativos del perfil de expresión de ese conjunto de genes[170]. Se calculó la conectividad (Figura S4) e identificaron los primeros 20 genes *hub* de cada *cluster*, ilustrados en la Figura 11. Para el *cluster* A, la mayoría de los genes *hub* codifican para proteínas ribosomales (Tabla S2). En el *cluster* B, los genes *hub* participan principalmente del ensamblado de la miofibrilla y del proceso catabólico de la timina (Tabla S3).

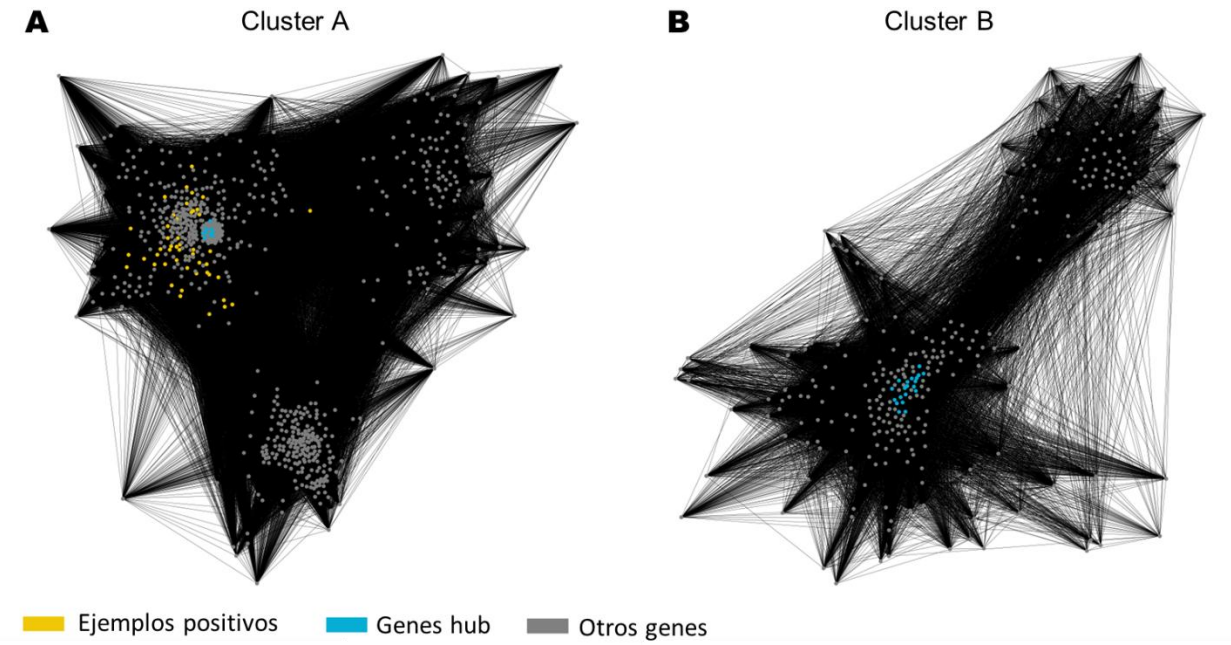


Figura 11: (A) Subgrafo del cluster de interés. Los genes de la muestra positiva están coloreados de amarillo, mientras que los genes hub se indican en azul. (B) Subgrafo del cluster contiguo. Los genes hub se indican en azul.

4.3. Aprendizaje automático supervisado

4.3.1. Conteos de bulk RNA-seq y normalización.

Para poder entrenar modelos de aprendizaje automático con los datos seleccionados, fue necesario primero normalizar la expresión de cada gen. El *dataset* 1 está disponible en valores de profundidad de cobertura por millón de reads mapeados (*depth of coverage per million mapped reads - dcpm*). Los otros *datasets* se encuentran depositados en NCBI como reads sin procesar. Frente a esto, fue necesario filtrar y mapear los reads en el genoma de referencia WBcel235 (GCF_000002985.6). Se obtuvieron las tablas de conteo para los *datasets* 2-4 que se encuentran resumidos en la

Tabla 7.

Todos los *datasets* fueron obtenidos utilizando la tecnología de Illumina, con diferentes estrategias en la construcción de la librería. Es relevante destacar que el *dataset* 3 estaba compuesto por entre 5 y 7 réplicas por cada tejido, con un promedio de 20,7 giga bases por réplica, teniendo un total 560.68 giga bases. Dado el tamaño del *dataset*, se decidió tomar 3 réplicas de cada tejido que tuvieran la misma cantidad de información, y de ellas se tomaron 25M de reads al azar de cada réplica.

Tabla 7: Información de datasets seleccionados. Se informa el tipo y tamaño de las librerías, el tamaño del dataset, la cantidad de reads y genes luego de mapear al genoma de referencia WBcel235 (GCF_000002985.6).

Dataset	BioProject	Tipo de librería	Tamaño	Cantidad de reads luego de filtrar (x10^6)	% reads mapeados	Cantidad de genes
Dataset 1	-	-	-	-	-	19928
Dataset 2	PRJNA394726	Paired-end	7.08 Gb	236.3	85%	18425
Dataset 3	PRJNA400796	Single-end	9.4 Gb	284.6	90%	18781
Dataset 4	PRJNA878786	Single-end	6.61 Gb	168.5	96%	23232
Cantidad de genes comunes a todos los *datasets*						**15714**

Luego de generar las tablas de conteo se realizó un análisis de Escalado Multidimensional (MDS) con el fin de evaluar si las réplicas eran consistentes en todos los conjuntos de datos[171,172]. Al graficar los resultados del MDS e observó que las réplicas para cada *dataset* se encontraban separadas espacialmente, lo que sugiere una alta consistencia y reproducibilidad entre réplicas de cada condición experimental (Figura 12). Este patrón de separación espacial en el MDS confirma la robustez de los datos normalizados y resalta las diferencias biológicas entre las condiciones experimentales evaluadas.

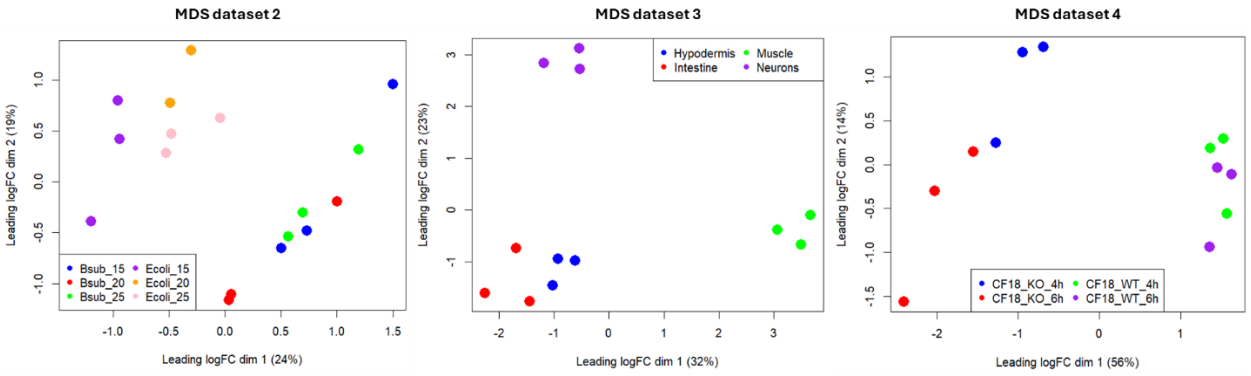


Figura 12: Multidimensional scaling (MDS) de cada réplica de los tres datasets.

Una vez obtenidas las tablas de conteo, se evaluaron tres métodos de normalización: TMM, TPM y TPMCLR. Para la evaluación se comparó el desempeño de modelos predictivos entrenados con distintas combinaciones de *datasets* y distintas normalizaciones. Los resultados se muestran en la

Tabla 8.

Tabla 8: Evaluación del desempeño de los 24 modelos combinando diferentes datasets y normalizaciones. El algoritmo de aprendizaje utilizado fue random forest. Las métricas utilizadas para evaluar el desempeño fueron: precisión, score F1 y área bajo la curva ROC (AUCROC). Los resultados fueron obtenidos con una muestra de evaluación (25% de los datos etiquetados disponibles).

Datasets	**Normalización**	**Umbral que maximiza F1**	**Precisión**	**F1**	**AUCROC**
***Dataset* 1, 2, 3 y 4**	TPM	0.59	0.97	0.97	0.98
	TPMCLR	0.33	0.97	0.97	0.99
	TMM	0.78	0.97	0.97	0.96
***Dataset* 1, 2 y 3**	TPM	0.75	0.94	0.95	0.96
	TPMCLR	0.5	0.88	0.89	0.92
	TMM	0.75	0.94	0.95	0.94
***Dataset* 1, 2 y 4**	TPM	0.78	0.94	0.95	0.98
	TPMCLR	1	0.85	0.87	0.9
	TMM	0.75	0.94	0.95	0.94
***Dataset* 1 y 2**	TPM	0.22	0.94	0.95	0.99
	TPMCLR	0.37	0.88	0.9	0.96
	TMM	0.54	0.94	0.95	0.97
***Dataset* 1, 3 y 4**	TPM	0.33	0.97	0.97	0.96
	TPMCLR	1	0.82	0.83	0.89
	TMM	0.58	0.97	0.97	0.96
***Dataset* 1 y 3**	TPM	0.65	0.94	0.95	0.98
	TPMCLR	0.52	0.88	0.89	0.98
	TMM	0.5	0.94	0.95	0.96
***Dataset* 1 y 4**	TPM	0.29	0.94	0.95	0.98
	TPMCLR	0.4	0.97	0.97	1
	TMM	0.44	0.97	0.97	0.96
***Dataset* 1**	TPM	0.21	0.91	0.93	0.98
	TPMCLR	0.12	0.91	0.93	0.98
	TMM	0.36	0.91	0.92	0.94

La

Tabla 8 muestra que el desempeño de los modelos es satisfactorio, con valores superiores a 0.8 en todos los casos. Se observa que la aplicación de CLR en lugar de la normalización de TPM solo da lugar a métricas más altas en una de las ocho combinaciones de *datasets*, mientras que en el resto de los casos no hay una mejora significativa e incluso se registra un peor desempeño en cinco de ellos. Por otra parte, no se evidencian diferencias claras entre la normalización por TPM o TMM. Finalmente, en relación con las distintas combinaciones de *datasets*, fue la combinación de todos ellos la que dio lugar a las mejores métricas, usando cualquiera de los tres métodos de normalización. En consecuencia, se optó por usar los 4 *datasets*. Dado que existe evidencia que su uso es mejor al integrar muestras de RNAseq provenientes de diferentes fuentes se optó por usar TMM como método de normailzación[150].

4.4. Entrenamiento y evaluación de los modelos.

Se entrenaron 18 modelos de aprendizaje automático para predecir nuevos genes asociados al proceso de fosforilación oxidativa. Se utilizó la misma muestra de entrenamiento y evaluación que la sección anterior, quitando los genes de cada uno de los complejos según correspondiera. Las métricas mediante las que se evaluó el desempeño de los modelos se muestran en la Tabla 9.

Tabla 9: Resultado de la evaluación de los 18 modelos, resultado de combinar 6 muestras de entrenamiento diferentes y 3 algoritmos de aprendizaje. Las métricas utilizadas para la evaluación fueron: precisión, score F1 y área bajo la curva ROC (AUCROC). Los resultados fueron obtenidos con una muestra de evaluación que incluía el 25% de los ejemplos positivos y negativos disponibles.

Algoritmo	Muestra de entrenamiento	Umbral	Cantidad de genes predichos	Precisión	F1	AUCROC
Random Forest	Todos los complejos	0.78	554	0.97	0.97	0.96
	Sin complejo I	0.33	945	0.9	0.86	0.95
	Sin complejo II	0.83	510	0.97	0.97	0.95
	Sin complejo III	0.75	581	0.94	0.95	0.99
	Sin complejo IV	0.33	2228	0.94	0.95	0.97
	Sin complejo V	0.67	595	0.94	0.94	0.93
Support Vector Machines	Todos los complejos	0.53	1174	0.97	0.97	0.93
	Sin complejo I	0.13	1214	0.95	0.92	0.91
	Sin complejo II	0.45	1169	0.97	0.97	0.92
	Sin complejo III	0.48	1176	0.97	0.97	0.94
	Sin complejo IV	0.46	1778	0.97	0.97	0.93
	Sin complejo V	0.38	1348	0.97	0.97	0.82
K-nearest neighbors	Todos los complejos	1	874	0.91	0.92	0.96
	Sin complejo I	0.62	549	0.9	0.83	0.81
	Sin complejo II	1	1208	0.97	0.97	0.97
	Sin complejo III	0.49	1838	0.97	0.97	0.97
	Sin complejo IV	1	807	0.91	0.92	0.96
	Sin complejo V	0.25	3147	0.84	0.86	0.9

Para los 18 modelos se obtuvieron métricas superiores a 0.8. Es interesante observar que los modelos entrenados sin incluir el complejo I presentan las métricas más bajas en los tres algoritmos evaluados. Esto puede explicarse por el tamaño de la muestra de entrenamiento, ya que los genes del complejo I representan casi el 50% de los ejemplos positivos disponibles.

El modelo entrenado con KNN y sin el complejo V obtuvo la mayor cantidad de predicciones (3147), mientras que el modelo de RF entrenado sin el complejo IV es el modelo con menos predicciones (510). En total, 4003 genes diferentes fueron predichos por al menos un modelo. La distribución de genes según la cantidad de modelos que los predicen se encuentra en Figura 13. Se puede ver que casi 1400 genes son predichos por un único modelo, mientras que 103 son los genes predichos por todos los modelos.

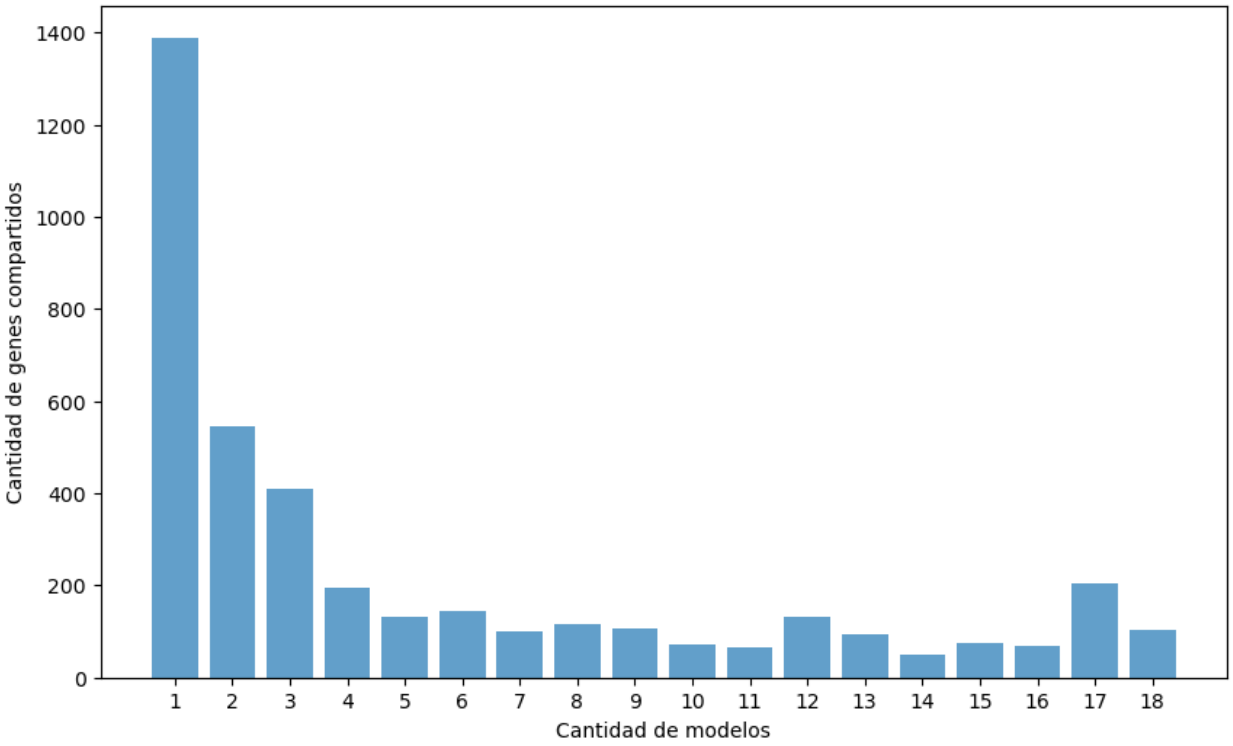


Figura 13: Distribución de genes según la cantidad de modelos que los predicen.

Una vez entrenados los modelos y clasificado el resto del genoma, se identificaron los genes clasificados como positivos por todos los modelos, independientemente del algoritmo utilizado y de la muestra de entrenamiento. La Figura 14 muestra un gráfico *upset* en el que cada conjunto es la intersección de los genes clasificados como positivos por los 6 modelos, para cada uno de los 3 algoritmos. Del total de 4003 genes clasificados como positivos por al menos un modelo, 103 lo fueron por todos los modelos. Al conjunto de estos genes se le denominó “lista de genes consenso” (Tabla 10).

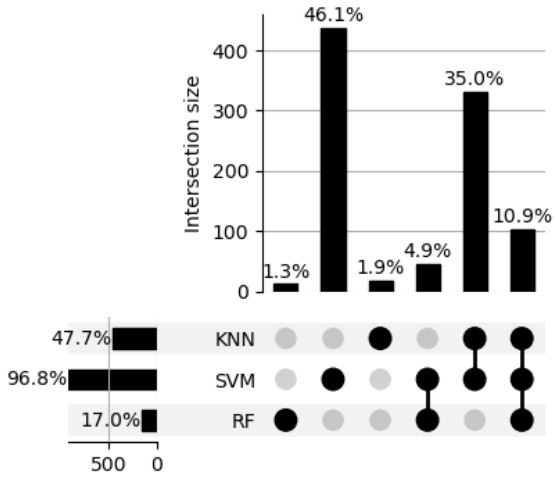


Figura 14: Intersección de genes predichos por los 18 modelos de aprendizaje automático. Cada clase representa la intersección de los genes predichos por cada algoritmo.

Tabla 10: Lista de genes consenso. Formada por los genes clasificados como positivos por los 18 modelos, ordenada de forma descendente por el promedio de la probabilidad de pertenecer a esa clase. La probabilidad de cada uno de los modelos está detallada en la Tabla S4. El (*) indica los genes que fueron excluidos de los ejemplos positivos por estar duplicados.

Gen	Nombre	Promedio de probabilidad	Gen	Nombre	Promedio de probabilidad
WBGene00004453	*rpl-39*	1.00	**WBGene00022114**	*Y71F9AL.9*	0.95
WBGene00004498	*rps-29*	1.00	**WBGene00013238**	*trap-4*	0.95
WBGene00004444	*rpl-30*	1.00	**WBGene00013766**	*prmt-1*	0.95
WBGene00004497	*rps-28*	1.00	**WBGene00009880**	*F49C12.11*	0.95
WBGene00006434	*prdx-2*	1.00	**WBGene00000209**	*asg-1**	0.94
WBGene00004888	*smo-1*	1.00	**WBGene00010627**	*nol-56*	0.94
WBGene00001427	*fkb-2*	1.00	**WBGene00021420**	*trap-3*	0.94
WBGene00017075	*nap-1*	1.00	**WBGene00022122**	*trap-1*	0.94
WBGene00017925	*F29B9.11*	1.00	**WBGene00003962**	*pdi-1*	0.94
WBGene00001946	*his-72*	1.00	**WBGene00000210**	*asg-2**	0.94
WBGene00006984	*zig-7*	1.00	**WBGene00010478**	*dkc-1*	0.94
WBGene00011156	*rbm-3.2*	1.00	**WBGene00002191**	*kin-3*	0.94
WBGene00006917	*vha-8*	1.00	**WBGene00016655**	*acbp-1*	0.93
WBGene00000263	*F23H11.5*	1.00	**WBGene00000534**	*cpi-2*	0.93
WBGene00012179	*rpl-37.1*	1.00	**WBGene00001501**	*ftn-2*	0.93
WBGene00000183	*arf-5*	1.00	**WBGene00004046**	*plp-1*	0.93
WBGene00013463	*kdp-1*	1.00	**WBGene00007708**	*nola-3*	0.93
WBGene00015755	*C14B9.10*	1.00	**WBGene00006913**	*vha-4*	0.93
WBGene00017982	*hpo-18**	0.99	**WBGene00012344**	*ola-1*	0.93
WBGene00010896	*snu-13*	0.99	**WBGene00003372**	*mlc-4*	0.92
WBGene00000881	*cyn-5*	0.99	**WBGene00009882**	*vha-17*	0.92
WBGene00000063	*act-1*	0.99	**WBGene00015514**	*srlf-1*	0.92

WBGene00020894	*T28D9.1*	0.99	**WBGene00021427**	*sec-61.B*	0.92
WBGene00011128	*adk-1*	0.98	**WBGene00004920**	*snr-7*	0.92
WBGene00010809	*lias-1*	0.98	**WBGene00021088**	*W08E12.7*	0.92
WBGene00009739	*F45H10.2**	0.98	**WBGene00020507**	*vha-15*	0.92
WBGene00009050	*mstr-1*	0.98	**WBGene00001423**	*fib-1*	0.92
WBGene00006918	*vha-9*	0.98	**WBGene00012602**	*Y38E10A.24*	0.92
WBGene00006919	*vha-10*	0.98	**WBGene00021248**	*Y22D7AL.10*	0.92
WBGene00001005	*dlc-1*	0.98	**WBGene00022599**	*daf-41*	0.91
WBGene00004302	*ran-1*	0.98	**WBGene00004914**	*snr-1*	0.91
WBGene00012097	*abcf-2*	0.98	**WBGene00020604**	*T20B12.7*	0.91
WBGene00000182	*arf-1*	0.98	**WBGene00021960**	*tmem-258*	0.91
WBGene00020868	*eif-1*	0.98	**WBGene00019680**	*K12H4.5*	0.91
WBGene00018963	*ucr-1**	0.97	**WBGene00001393**	*fat-1*	0.91
WBGene00003053	*lmp-1*	0.97	**WBGene00007630**	*har-1*	0.91
WBGene00001840	*hel-1*	0.97	**WBGene00000896**	*dad-1*	0.91
WBGene00077526	*C25A1.16**	0.97	**WBGene00003065**	*lpd-9*	0.90
WBGene00007684	*ndub-3*	0.97	**WBGene00018151**	*ndab-2**	0.90
WBGene00019007	*ucr-11*	0.97	**WBGene00020915**	*nol-58*	0.90
WBGene00010174	*F56H9.2*	0.97	**WBGene00010317**	*idh-1*	0.90
WBGene00004266	*rab-1*	0.97	**WBGene00012768**	*eef-1B.2*	0.89
WBGene00020297	*nog-1*	0.97	**WBGene00003123**	*mag-1*	0.89
WBGene00006910	*vha-1*	0.97	**WBGene00003214**	*mel-32*	0.89
WBGene00017926	*cox-6C**	0.96	**WBGene00004766**	*sel-9*	0.87
WBGene00014454	*MTCE.7*	0.96	**WBGene00011510**	*pdha-1*	0.86
WBGene00004918	*snr-5*	0.96	**WBGene00009915**	*F52A8.1*	0.84
WBGene00016746	*C48B6.10*	0.96	**WBGene00000379**	*cct-4*	0.83
WBGene00017984	*gmpr-1*	0.96	**WBGene00015778**	*got-2.2*	0.82
WBGene00019537	*K08D12.3*	0.96	**WBGene00009664**	*idha-1*	0.82
WBGene00020216	*trap-2*	0.96	**WBGene00000378**	*cct-2*	0.81
WBGene00015248	*mai-2*	0.95			

Estos 103 genes, predichos por todos los modelos al considerar en cada uno de ellos el umbral de clasificación que maximiza el F1, conforman la lista consenso. Los valores alcanzados con las distintas métricas de evaluación son muy buenos, indicando que los modelos aprendieron a reconocer patrones de expresión distintivos, que les permitieron identificar a genes ya conocidos por estar involucrados en este proceso y que fueron usados para evaluar su desempeño. Con el fin de evaluar la calidad de las predicciones en general y de encontrar evidencia adicional que vincule los genes de la lista consenso con el proceso de fosforilación oxidativa, se llevaron a cabo varios análisis adicionales, buscando a su vez priorizar a los mejores candidatos para una eventual validación experimental.

4.5. Caracterización de la lista consenso.

4.5.1. Predicción de genes excluidos.

En el relevamiento de genes a ser utilizados como ejemplos positivos, se excluyeron los genes de la Tabla 5 y los genes que participan del ensamblado de los complejos que llevan a cabo la fosforilación oxidativa. Luego de obtenidas las predicciones de los modelos entrenados se verificó la cantidad de veces que cada uno de estos genes habían sido predicho por alguno de los modelos. Esta información se resume en la Tabla 11 y en la Tabla 12.

Tabla 11: Lista de genes involucrados en el ensamblaje de los complejos excluidos del entrenamiento y la cantidad de modelos que los clasifica como involucrados en la fosforilación oxidativa.

Complejo	Gen	Cantidad de modelos que lo predicen
I	*acdh-12*	0/18
I	*B0035.15*	0/18
I	*B0334.5*	1/18
I	*K09E4.3*	0/18
I	*M04B2.4*	0/18
I	*ndua-2*	8/18
I	*nuaf-1*	0/18
I	*nuaf-3*	0/18
I	*nubp-1*	0/18
I	*Y116A8C.30*	0/18
I	*Y38F2AR.3*	0/18
I	*ZK1128.1*	0/18
II	*Y57A10A.29*	0/18
III	*bcs-1*	0/18
III	*ddl-1*	0/18
IV	*coa-1*	0/18
IV	*coa-3*	0/18
IV	*coa-4*	0/18
IV	*coa-5*	0/18
IV	*coa-6*	0/18
IV	*coa-7*	0/18
IV	*cox-10*	0/18
IV	*cox-11*	0/18
IV	*cox-14*	4/18
IV	*cox-15*	0/18
IV	*cox-16*	0/18
IV	*cox-17*	0/18
IV	*cox-18*	0/18
IV	*cox-19*	1/18
IV	*sco-1*	0/18

IV	*stf-1*	0/18
IV	*T20D3.6*	0/18
IV	*Y53F4B.14*	0/18

En la Tabla 11 se puede ver que de las 32 proteínas que forman parte del ensamblaje de los complejos, 28 no son clasificadas como positivas por ninguno de los modelos. De las otras 4, dos son clasificadas como positivas por solo un modelo, otra por 4 y el gen *ndua-2* por 8 modelos Esto puede interpretarse como resultado de que los patrones de expresión de las proteínas del ensamblaje y el de las proteínas *core* y accesorias no son similares, de forma que excluir las primeras de la muestra de entrenamiento fue una decisión acertada para predecir genes relacionados el proceso de fosforilación oxidativa. Este resultado refuerza la calidad de la lista consenso.

Por otro lado, en los genes excluidos por poseer duplicaciones génicas (Tabla 12) se observa un patrón notable: a excepción de tres casos la mayoría de los modelos predijeron consistentemente solo a una de las dos o tres copias del gen duplicado. Las duplicaciones en las que esto no sucedió son las de los genes *asg-1 y asg-2,* ambos predichos por todos los modelos, los genes *ndab-1 y ndab-2,* el primero predicho por 17 modelos y el segundo por los 18 y los genes *B0035.18, cox-6c y Y111B2A2* predichos por 11, 18 y ningún modelo respectivamente.

Este resultado indica que, en la mayoría de los casos, sólo una de las dos copias posee un perfil de expresión característico de los genes involucrados en el proceso de fosforilación oxidativa. Esto podría deberse a varios factores, como diferencias en la regulación transcripcional, la presencia de isoformas específicas, o variaciones en la expresión de distintos tipos de tejidos (como pasa con *nduf-2.2* y *sdha-2*) o condiciones experimentales. Cabe consignar que estos últimos dos genes, inicialmente excluidos por estudios previos que mostraban un patrón de expresión muy singular, únicamente en línea germinal, tampoco fueron recuperados por ningún modelo.

Este resultado es un claro ejemplo de que en ocasiones no es suficiente la homología de secuencia o estructural para comprender la función de un gen, y abre la puerta a estudiar la expresión temporal y espacial (o ante condiciones específicas) de estos genes duplicados, pobremente anotados, en mayor profundidad con el fin comprender si estas duplicaciones son parte de alguna(s) adaptación(es) bioquímica(s) particular(es), u otra optimización biológica y la historia evolutiva de estas duplicaciones.

Finalmente, los genes cuya asociación con el proceso de fosforilación oxidativa tenía sustento bibliográfico tan débil que se optó por excluirlos de la muestra de entrenamiento (*C06A5.3, nduv-3, C14B9.10 y ucr-11*) fueron clasificados como positivos por la mayoría de los modelos, resultado que refuerza esa asociación inicial.

Tabla 12: Lista de genes excluidos de la lista de entrenamiento y cantidad de modelos que los predicen. Los genes de esta lista no fueron incluidos por diversos motivos: bien porque estaban duplicados o porque no contaban con sustento bibliográfico suficiente para ser agregados a la muestra de entrenamiento

Complejo	Gen en *C. elegans*	Motivo de exclusión	Cantidad de modelos que lo predicen
I	***C06A5.3***	**Anotado en GO con código IBA, sin sustento bibliográfico.**	**17/18**
I	***ndab-1***	**Duplicado con ndab-2.**	**17/18**
I	***ndab-2***	**Duplicado con ndab-1.**	**18/18**
I	*nduf-2.2*	Solo se expresa en línea germinal. Duplicado con *nduf-2.1*.	2/18
I	***nduv-3***	**Anotado en GO con código IEA, sin sustento bibliográfico.**	**18/18**
I	***ndub-3***	**Inconsistencias en la anotación.**	**18/18**
II	*sdha-2*	Solo se expresa en línea germinal. Duplicado con *sdha-1*	1/18
III	***C14B9.10***	**Anotado en GO con código IEA, sin sustento bibliográfico.**	**18/18**
III	***F45H10.2***	**Duplicado con R07E4.3.**	**18/18**
III	*R07E4.3*	Duplicado con F45H10.2	06/18
III	*mppb-1*	Duplicado con ucr-1	0/18
III	***ucr-1***	**Duplicado con mppb-1**	**18/18**
III	***ucr-11***	**Anotado en GO con código IEA, sin sustento bibliográfico.**	**18/18**
III	***ucr-2.1***	**Triplicado con ucr-2.2 y ucr-2.3.**	**17/18**
III	*ucr-2.2*	Triplicado con ucr-2.1 y ucr-2.3.	6/18
III	*ucr-2.3*	Triplicado con ucr-2.1 y ucr-2.2.	1/18
IV	***B0035.18***	**Triplicado con cox-6c y Y111B2A.2.**	**11/18**
IV	***cox-6c***	**Triplicado con B0035.18y Y111B2A.2.**	**18/18**
IV	*Y111B2A.2*	Triplicado con cox-6c y B0035.18.	0/18
V	*asb-1*	Duplicado con asb-2.	7/18
V	***asb-2***	**Duplicado con asb-1.**	**17/18**
V	***asg-1***	**Duplicado con asg-2.**	**18/18**
V	***asg-2***	**Duplicado con asg-1.**	**18/18**
V	***hpo-18***	**Triplicado con R05D3.6 y ZC262.5.**	**18/18**
V	*R05D3.6*	Triplicado con hpo-18 y ZC262.5.	0/18
V	*ZC262.5*	Triplicado con hpo-18 y R05D3.6.	0/18

4.5.2. Análisis de enriquecimiento.

Con el fin de evaluar el procedimiento de considerar la lista consenso, se comparó el enriquecimiento funcional de esa lista con el de las listas obtenidas por cada algoritmo de clasificación por separado (conjuntos que se muestran en la Figura 14). Los enriquecimientos obtenidos se muestran en la Figura 15, en la que se puede ver que la lista consenso mantiene el mismo enriquecimiento en términos de la ontología “componente celular” asociados a la

mitocondria y a membranas, pero muestra un enriquecimiento mayor en términos asociados a procesos como la acidificación del lumen lisosomal (que se discutirá más adelante), la síntesis de ATP impulsada por la fuerza motriz de protones, y el transporte de protones transmembrana.

Además de resultar en una lista más enriquecida en términos GO de interés, una gran ventaja de esta estrategia es que da lugar a una lista más pequeña. La lista consenso tiene solo 103 genes, lo cual permite realizar un análisis meticuloso de los mismos y que puede derivar en la implementación de ensayos *in vivo* a futuro. En las próximas secciones, se presenta un análisis más detallado de estos genes, buscando aportar evidencia adicional que respalde su carácter de genes candidatos a estar involucrados en el proceso de fosforilación oxidativa o estar íntimamente relacionados a este proceso.

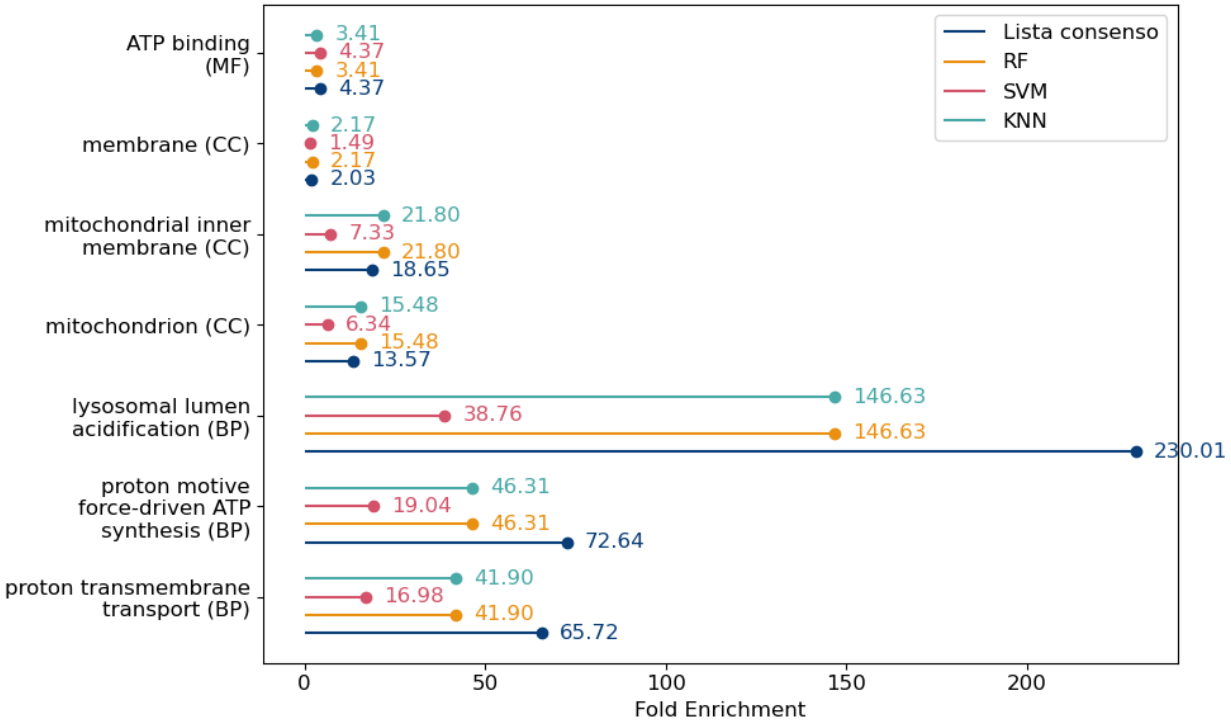


Figura 15: Análisis de enriquecimiento en término GO de la lista de genes consenso y de los genes predichos utilizando cada uno de los tres algoritmos.

4.5.3. Mapeo de los genes de lista consenso en la red de co-expresión.

Como primera forma de caracterizar la lista de genes clasificados como positivos por los 18 modelos, se los ubicó en la red de co-expresión construida en la sección 4.2.2. De los 103 genes de la lista consenso, solamente el gen *MTCE.7*, correspondiente al ARN ribosomal 12s de la mitocondria, no está presente en la red, lo cual se puede deber a que no fue detectado por la tecnología de secuenciación o a que no pasó los controles de calidad. Luego se comparó el enriquecimiento en ejemplos positivos y en genes de la lista consenso en los *clusters* que resultan de aplicar *clustering* jerárquico a la red de co-expresión (Figura 16).

Como era de esperar, el *cluster* más enriquecido en ejemplos positivos está también muy enriquecido en genes de la lista consenso. Sin embargo, como muestran la Figura 16, los genes de la lista consenso aparecen distribuidos en más *clusters* que los ejemplos positivos y 6 de estos

clusters muestran un enriquecimiento significativo y mayor a 2 (los *clusters* A, C, D, E, F y G). Esto se puede interpretar como evidencia de que para clasificar como positivo a un gen dado, estos modelos no se limitaron a eventuales correlaciones entre patrones de transcripción, sino que lograron aprender patrones compartidos más complejos, lo que resulta en la predicción de genes que se encuentran en otros *clusters* de la red. Es especialmente interesante notar que el *cluster* C posee un enriquecimiento de 6.47 en genes de la lista consenso, ligeramente mayor al enriquecimiento del *cluster* A (5.56), en el que se encuentran la mayoría de los ejemplos positivos.

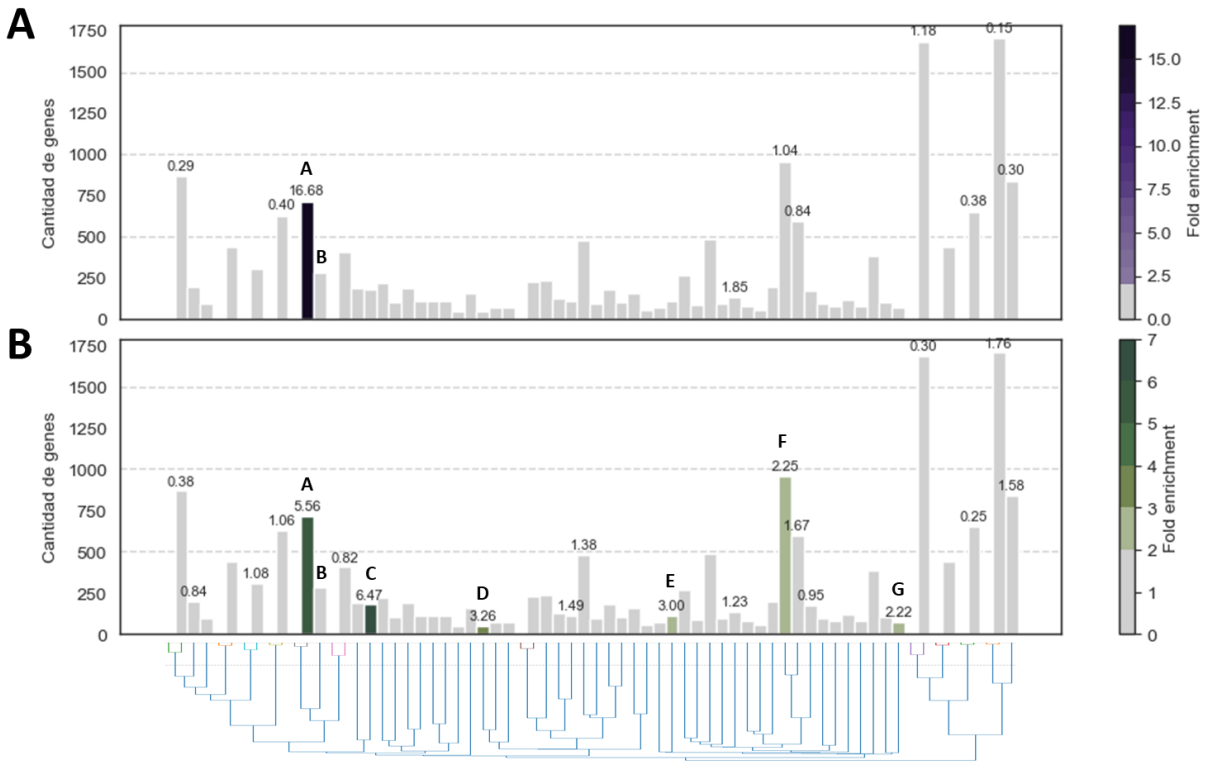


Figura 16: Distribución de los ejemplos positivos (A) y de los genes de la lista consenso (B) en la red de co-expresión. La altura de las barras indica la cantidad total de genes en cada cluster. El gradiente de color corresponde al enriquecimiento en ejemplos positivos (A) o en genes de la lista consenso (B) para cada cluster. En gris se indica que el fold enrichment es menor a 2. En aquellos clusters en los que el enriquecimiento fue diferente a 0 se indica su valor sobre las barras. Todos los p-values son menores a 0.05.

En la Tabla 13 se muestra la cantidad total de genes, la cantidad de genes de la muestra de entrenamiento y la cantidad de genes de la lista consenso para varios *clusters*. Se puede ver que a pesar de que los *clusters* C, D, E, F y G poseen un enriquecimiento en genes de la lista consenso mayor a 2, la cantidad de esos genes en cada uno de esos *clusters* es en realidad baja, sobre todo para los *clusters* D, E y G. Se realizó un enriquecimiento en términos de Gene Ontology de cada uno de estos *clusters* para explorar la composición de genes de estos. Se detectó un enriquecimiento en por lo menos un término GO en todos los *clusters*, excepto en el *cluster* D. Los resultados se muestran en la Figura 17.

Tabla 13: Cantidad de genes de la muestra de entrenamiento y lista consenso en algunos clústeres del grafo.

Cluster	Total de genes	Ejemplos positivos	Enriquecimiento en ejemplos positivos	Genes de la lista consenso	Enriquecimiento de los genes de la lista consenso
A	718	48	16.68	24	5.56
C	180	0	0	7	6.47
D	51	0	0	1	3.26
E	111	0	0	2	3.00
F	960	4	1.04	13	2.25
G	75	0	0	1	2.22
Otros	14873	16		54	
Total	**16968**	**68**		**102**	

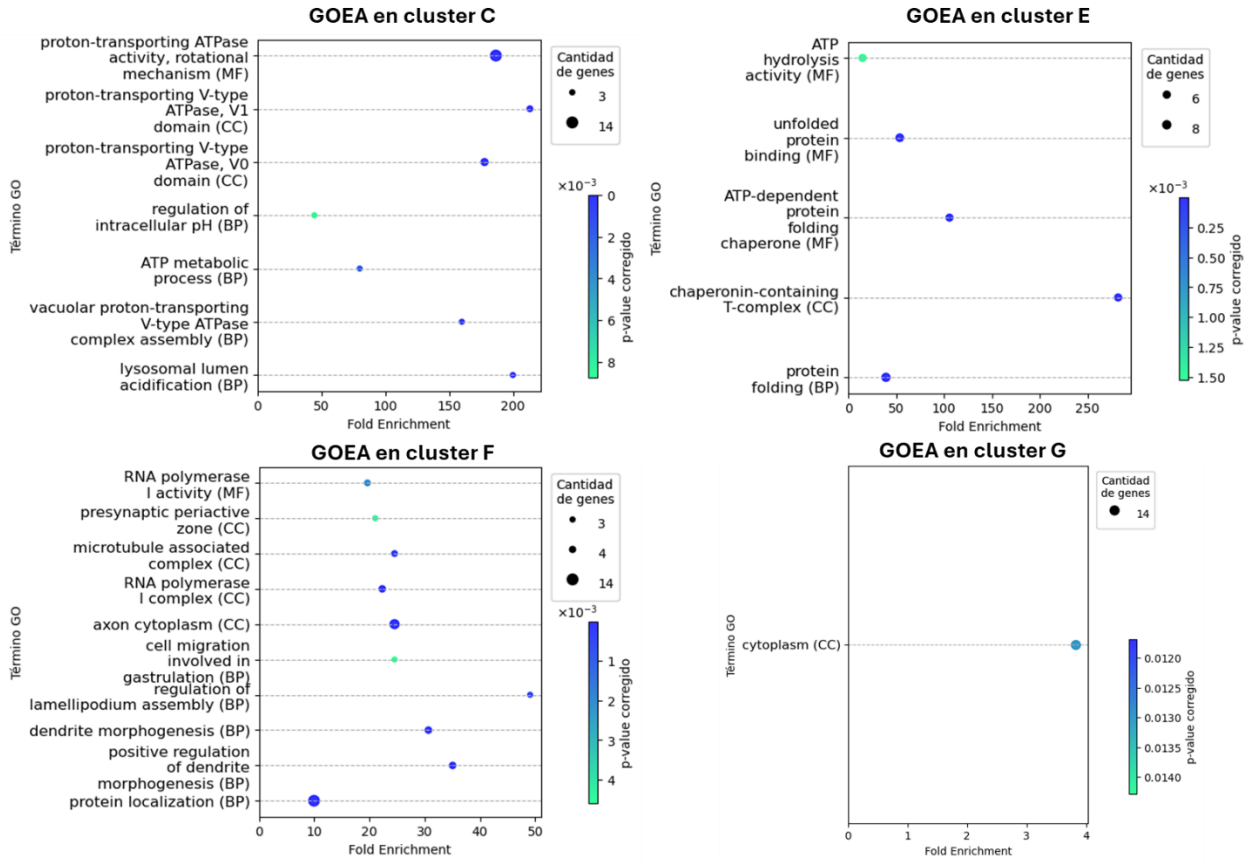


Figura 17: Gene Ontology Enrichment Analysis (GOEA) en 4 de los 5 clusters enriquecidos en genes de la lista consenso. El cluster D, de 51 genes, no presenta un enriquecimiento en genes de ningún término GO. En las gráficas se indica si los términos GO pertenecen a la ontología biological process (BP), molecular function (MF) o cellular component (CC).

En la Figura 17, se observa que el *cluster* C está enriquecido en genes asociados al proceso de acidificación de vacuolas. Se trata de los genes *vha,* que codifican proteínas que conforman ATPasas dependientes de protones. Los 21 genes que conforman estos complejos están presentes en este *cluster*, a excepción de los genes *vha-6* y *vha-7*. A su vez, la lista consenso

incluye a siete de estos 19 genes *vha,* y todos ellos se ubican en este clúster. Por otro lado, el *cluster* E muestra un enriquecimiento en genes asociados a términos GO vinculados a la unión de proteínas desplegadas y al plegamiento de proteínas mediado por chaperonas que dependen de ATP. Los dos genes de la lista consenso que se encuentran dentro de este grupo son cct-2 y cct-4, ambos con anotaciones que incluyen estos términos. Las actividades relacionadas con el uso de ATP de estos genes podrían explicar su presencia en la lista consenso.

El *cluster* F está compuesto por 960 genes y está enriquecido en 438 términos GO de las 3 ontologías (datos no mostrados). En la Figura 17 se muestran algunos de los términos con los que están anotados los genes de la lista consenso que aparecen en este *cluster*. Estos términos están relacionados con el desarrollo, los movimientos celulares vinculados a la gastrulación, y a la morfogénesis neuronal.

Finalmente, el *cluster* G está compuesto por 75 genes y se encuentra enriquecido únicamente en el término "citoplasma" de la ontología "componente celular" y no se considera particularmente informativo. El único gen de la lista consenso que pertenece a este *cluster* (*arf-1*) se encuentra anotado con este término.

El *cluster* A, enriquecido en ejemplos positivos, incluye a 24 de los 103 genes de la lista consenso. La localización de estos genes en la red de co-expresión se muestra en la Figura 18. En particular, los genes *W01D2.1, rps-28, rps-29, rpl-30 y rpl-3 (marcados en rojo),* que codifican para proteínas ribosomales, se encuentran próximos a los genes *hub* del *cluster*, que codifican para el mismo tipo de proteínas. Por otro lado, todas las duplicaciones génicas incluidas en la lista consenso se encuentran en este clúster.

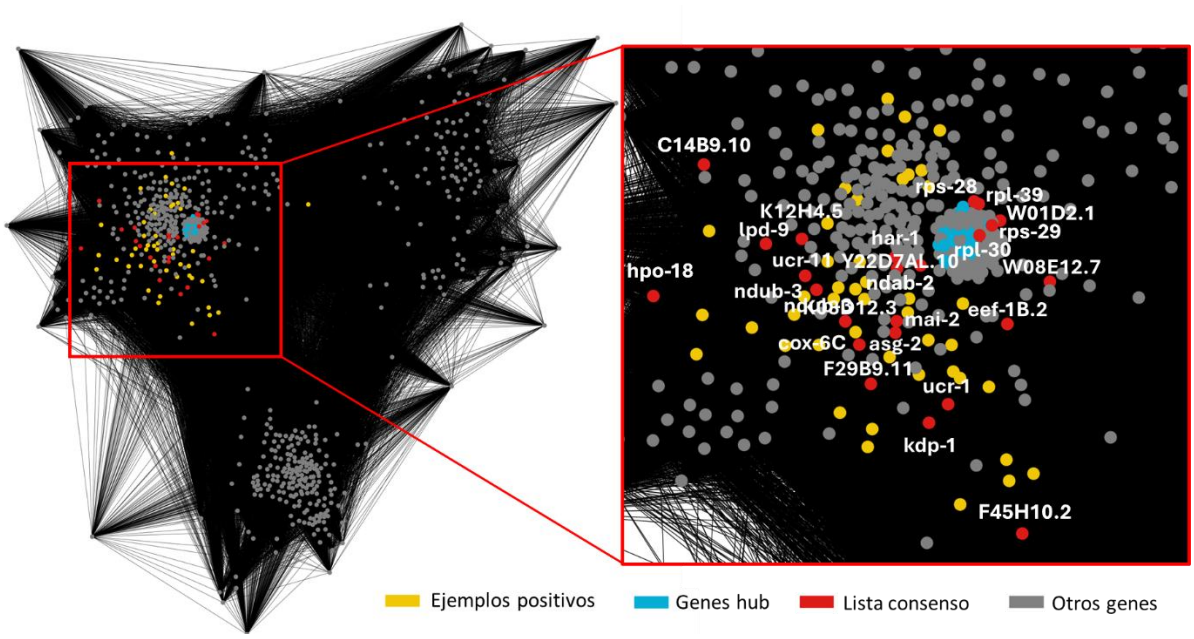


Figura 18: Subgrafo correspondiente al cluster A. En amarillo se indican los genes positivos de la muestra de entrenamiento, en azul los genes hub y en rojo los genes de la lista consenso. El resto de los genes están coloreados en gris.

4.5.4. Anotaciones GO y homólogos humanos de los genes de la lista consenso.

Se llevó a cabo un relevamiento de las anotaciones GO de los genes de la lista consenso, sin incluir las duplicaciones génicas (ver Tabla 12), por lo que se trabajó con un total de 93 genes. Para estos genes también se realizó una búsqueda de homólogos en *H sapiens* en Wormbase. La información recabada se resume en la Figura 19.

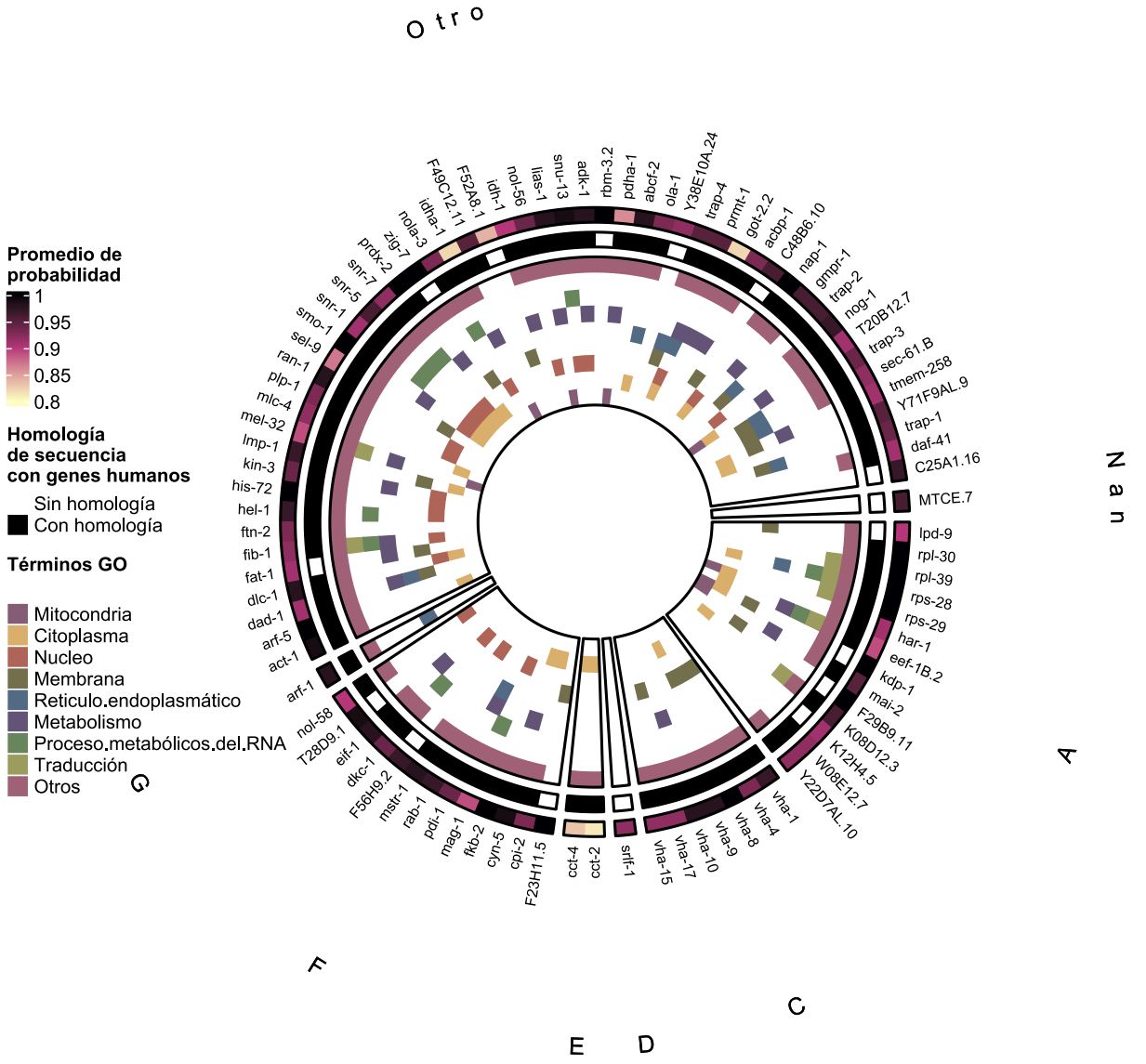


Figura 19: Anotaciones GO y ortólogos en humanos de los genes de la lista consenso. Los clusters de la red están indicados por fuera del heatmap. En cada fila se indica el promedio de la probabilidad de pertenecer a la clase de los ejemplos positivos, si el gen presenta o no ortólogos de secuencia con humanos y la presencia o ausencia de una anotación GO. Las anotaciones utilizadas fueron: Traducción = GO:0006412, GO:0006417, GO:0017148, GO:0045727; Procesamiento Metabólico de ARN = GO:0016070; Metabolismo = GO:0008152; Retículo Endoplasmático = GO:0005783; Membrana = GO:0016020; Núcleo = GO:0005634, Citosol = GO:0005737; Mitocondria = GO:0005739. Para todos los casos se tomó en

cuenta si el gen posee anotaciones en algún término descendiente de los términos enumerados. Para el caso de metabolismo, se eliminaron todos los términos descendientes de traducción y procesamiento metabólico de ARN.

En el *cluster* A se destaca la presencia del gen *mai-2*, ortólogo del gen humano *ATP5IF1*, que inhibe la actividad hidrolítica de la ATPsintasa cuando el gradiente electroquímico de protones es interrumpido en la mitocondria[173]. También se encuentra el gen *har-1*, homólogo del gen humano *CHCHD10*, que tiene funciones de mantenimiento de la integridad mitocondrial[174]. Teniendo en cuenta que la inclusión de un gen en la lista consenso depende exclusivamente de su patrón de transcripción, es interesante la presencia de estos dos genes, con ortólogos humanos que participan uno de la regulación de la ATPasa mitocondrial y el otro de la integridad de la mitocondria.

El resto de los genes de la lista consenso que se encuentra en el *cluster* A está anotado con diferentes términos GO, que incluyen regulación de la traducción, unión a ARN (*eef-1B2* y *K08D12.3*), plegamiento de proteínas (*Y22D7AL.10*), componentes estructurales de la célula (*kdp-1)* y almacenamiento de lípidos (*lpd-9*). Finalmente, los genes *W08E12.7, K12H4.5* y *F29B9.11* no tienen ninguna anotación GO, al tiempo que *K12H4.5* y *F29B9.11* no presentan homología con ningún gen humano.

Los genes *vha-1, vha-4, vha-8, vha-9, vha-10, vha-15* y *vha-17* de la lista consenso se encuentran en el *cluster* C. Los genes *vha* codifican para complejos de ATPasas dependientes de protones (V-ATPasas), proteínas esenciales para el mantenimiento del gradiente de pH a través de las membranas biológicas, y median la acidificación de organelos. Estos complejos están formados por los módulos V0 y V1, que estructuralmente son muy similares a los módulos F0 y F1 de la ATP sintasa. Durante el proceso de acidificación, las V-ATPasas bombean protones (H^+) desde el citosol al interior de los organelos en contra del gradiente en un proceso acoplado a la hidrólisis de ATP, disminuyendo así el pH y creando un ambiente ácido en el interior del organelo. Este proceso es esencial para mantener el pH ácido del lisosoma. Cabe destacar que de los 21 genes que conforman las V-ATPs, 18 fueron clasificados como positivos por al menos un modelo y siete lo fueron por los 18 modelos, formando parte de la lista consenso. Únicamente los genes *vha-1, vha-18* y *spe-5* no fueron clasificados como positivos por ningún modelo. Este resultado vincula el proceso de producción de ATP con un proceso dependiente de energía. Es plausible que la regulación de la expresión de ambos conjuntos de genes esté coordinada, dando lugar a patrones de expresión con características similares, lo cual explicaría la inclusión de genes de la V-ATPasa en la lista consenso.

El único gen de la lista consenso que pertenece al *cluster* D es *srlf-1*, el cual no tiene homología con ningún gen humano y carece de referencias bibliográficas relevantes. Según la descripción disponible en WormBase[175], en base a estudios de proteómica y de scRNA-seq, se sabe que este gen se expresa en células accesorias, células del arco anterior, hipodermis anterior, neuronas ciliadas y el sistema nervioso somático. Además, estudios de proteómica, microarrays y RNA-seq han determinado que está afectado por varios genes, como *daf-16, daf-2* y *skn-1*.

El *cluster* E contiene los genes *cct-2* y *cct-4*, ortólogos de los genes *CCT2* y *CCT4* de *Homo sapiens*, respectivamente. Estos genes humanos codifican para proteínas chaperonas que se encuentran en el citosol y forman parte del complejo de chaperonas que contienen TCP1 (CCT). Este complejo consiste en dos anillos apilados idénticos, formados por 8 proteínas cada uno. Péptidos desplegados entran a la cavidad central y se da el plegamiento dependiente de ATP. Si bien solo estos dos genes se encuentran en la lista consenso, los otros 6 genes fueron

clasificados como positivos por más de la mitad de los modelos, y 4 de ellos también se encuentran en el *cluster* C. Al igual que sucede con los genes *vha*, este es otro ejemplo de un conjunto de genes clasificados como positivos cuya función depende de un alto consumo de ATP.

Trece de los genes de la lista consenso se encuentran en el clúster F. Los genes *cpi-2* y *mag-1* están anotados con términos GO relacionados a la reproducción. El primero está asociado a la regulación del desarrollo de ovocitos y la vitelogénesis, mientras que el segundo está relacionado con la feminización de ovocitos en la línea germinal de hermafroditas, la regulación positiva de la puesta de huevos y la oogénesis. Por otro lado, genes como *cyn-5, fkb-2, pdi-1* y *rab-1* están asociados a procesos como el plegamiento de proteínas, la respuesta a proteínas mal plegadas y el transporte intracelular de proteínas. Además, los genes *mstr-1, nol-58, dkc-1* y *eif-1* están anotados con actividad de unión a ADN, a ARNsno, de modificación post-transcripcional de ARN e inicio de la traducción, respectivamente. Finalmente, los genes *F23H11.5, F56H9.2* y *T28D9.1* no tienen ortólogos en humanos ni anotaciones GO.

Por otro lado, el *cluster* G contiene un único gen de la lista consenso: *arf-1*. Este gen codifica para una proteína perteneciente a la familia de GTPasas Arf/Sar, asociada al tráfico intracelular de vesículas. La proteína Arf1, en su forma unida a GDP, se asocia con la membrana mitocondrial y participa activamente en la regulación de la homeostasis y la dinámica mitocondrial. Dado que las proteínas de la cadena de transporte de electrones y la ATP sintasa desempeñan su función en la membrana mitocondrial, la inclusión de este gen en la lista consenso podría deberse a la co-localización de su producto y no a que tengan funciones relacionadas, aunque esto último tampoco puede descartarse.

Finalmente, los genes distribuidos en el resto de los *clusters* están anotados con diversos términos GO. Hay un grupo numeroso de estos genes cuyos productos tienen actividad de unión a ADN, ARN y/o factores de transcripción, que incluye a *fib-1, hel-1, nap-1, nog-1, his-72, nol-56, nola-3, plp-1, snr-2, snr-5, snr-7, snu-13 y smo-1*. Por otro lado, genes como *act-1, arf-5, kin-3, mel-32 y ron-1* están vinculados principalmente a procesos asociados al desarrollo embrionario. Los genes *trap* del 1 al 4 se encuentran ubicados en el mismo clúster y las proteínas que codifican participan de la inserción de proteínas que se están sintetizando en la membrana del retículo endoplasmático[176]. Finalmente, los genes *C25A1.16, C48B6.10, F49C12.11, F52A8.1, rbm-3.2, Y38E10A.24, Y71F9AL.9 y zig-7* no tienen anotaciones GO.

4.5.5. Predicción de procesos metabólicos asociados a la fosforilación oxidativa.

La fosforilación oxidativa representa la etapa final de la producción de energía en el metabolismo de los organismos aerobios. En este proceso convergen los productos oxidados de diversas rutas catabólicas celulares, lo que resalta la importancia de identificar genes de la lista consenso relacionados con procesos catabólicos.

Entre los genes de la lista consenso se encuentra *got-2.2*, que codifica una transaminasa con actividad sobre L-aspartato:2-oxoglutarato, implicada en el catabolismo del aspartato. Además, se encuentran los genes *idh-1, idha-1* y *pdha-1*, todos ellos involucrados en el ciclo del ácido cítrico. Este ciclo, compuesto por una serie de reacciones que tienen lugar en la matriz mitocondrial, genera NADH y $FADH_2$, moléculas esenciales para la fosforilación oxidativa, ya que donan electrones a los complejos I y II de la cadena de transporte de electrones.

Asimismo, el succinato, uno de los intermediarios del ciclo, es utilizado directamente por el complejo II en la transferencia de electrones. Por esta razón, es esperable que los genes implicados en estos procesos compartan patrones de expresión similares. Para corroborar esta hipótesis, se evaluó la frecuencia con la que las enzimas involucradas en estos procesos, según la Kyoto Encyclopedia of Genes and Genomes (KEGG), fueron clasificadas como positivas por los 18 modelos predictivos. Esta información se resume en la Tabla 14.

Tabla 14: Genes del ciclo del ácido cítrico y cantidad de veces que fueron clasificados como positivos. Se excluyen los genes de la enzima succinato deshidrogenasa (complejo II) ya que la misma fue incluida en la muestra de entrenamiento.

Gen	Nombre	Cantidad de predicciones
acly-2	ATP Citrate Lyase	1/18
cts-1	Citrate Synthase	17/18
aco-1	Aconitase	15/18
aco-2	Aconitase	12/18
idh-1	Isocitrate Dehydrogenase (cytosol)	18/18
Idh-2	Isocitrate Dehydrogenase (mitochondrion)	0/18
idha-1	Isocitrate DeHydrogenase Alpha	18/18
idhb-1	Isocitrate DeHydrogenase Beta	5/18
idhg-1	Isocitrate DeHydrogenase Gamma	10/18
idhg-2	Isocitrate DeHydrogenase Gamma	0/18
ogdh-1	Oxoglutarate Dehydrogenase	17/18
dld-1	DihydroLipoamide Dehydrogenase	17/18
dlst-1	Dihydrolipoamide S-Succinyltransferase	17/18
sucl-1	Succinyl-CoA Ligase, alpha subunit	15/18
sucl-2	Succinyl-CoA Ligase, alpha subunit	1/18
sucg-1	Succinyl-CoA ligase, GTP-specific, beta subunit	2/18
suca-1	Succinyl-CoA ligase, ATP-specific, beta subunit	16/18
fum-1	Fumarase	17/18
mdh-1	Malate Dehydrogenase	17/18
mdh-2	Malate Dehydrogenase	16/18
pdha-1	Pyruvate DeHydrogenase Alpha subunit	18/18
pdhb-1	Pyruvate DeHydrogenase Beta	13/18
dlat-1	Dihydrolipoyllysine-residue Acetyltransferase	17/18
dlat-2	Dihydrolipoyllysine-residue Acetyltransferase	16/18

En la tabla se puede ver que la mayoría de los modelos clasifica como positivos a los genes que llevan a cabo cada una de las reacciones del ciclo. Una vez más, se trata de dos procesos estrechamente interconectados: la fosforilación oxidativa y el ciclo del ácido cítrico, por lo que es razonable que los patrones de expresión de los genes involucrados en cada uno de esos procesos sean similares. El hecho de que un modelo predictivo, entrenado con genes asociados a uno de estos procesos, identifique como candidatos a genes conocidos por participar en el otro proceso, valida el diseño del enfoque y respalda la solidez de los genes incluidos en la lista consenso como buenos candidatos. Sin embargo, los genes que codifican para proteínas partícipes de la beta-oxidación y la glicólisis no son predichas por estos modelos (Tabla S5 y Tabla S6).

4.5.6. Expresión de genes de la lista consenso en otros trabajos.

Con el fin de aportar evidencia adicional que refuerce la asociación de los genes de la lista consenso con la fosforilación oxidativa en *C. elegans*, se realizó una búsqueda bibliográfica de trabajos reportaran expresión diferencial (DEG) en diversas condiciones de genes ya vinculados este proceso. El trabajo de Priebe *et al.*[159] se centra en el estudio de cómo la restricción de glucosa mediante el tratamiento con 2-deoxi-D-glucosa (DOG) prolonga la vida de *C. elegans*. Uno de los principales hallazgos del trabajo es la identificación de 4891 genes diferencialmente expresados en gusanos de un día de edad tratados con DOG. En esta condición, observaron que la mayoría de los genes asociados a la fosforilación oxidativa presentaban una regulación positiva.

De los 68 genes seleccionados en este estudio como ejemplos positivos, 36 se encuentran entre los DEG al alta, lo que representa un fold change de 1.9. Además, cerca de la mitad de los genes de la lista consenso también están regulados positivamente en esta condición, con un fold change de 1.6. Los genes de la lista consenso que aparecen como diferencialmente expresados en el análisis de Priebe *et al.* se enumeran en la Tabla 15. Estos genes están distribuidos en todos los *clusters* definidos y, notablemente, algunos no tienen una función asignada, lo que los convierte en interesantes candidatos para estudios futuros.

Tabla 15: Genes de la lista consenso diferencialmente expresados en Priebe et. al.

WormBaseID	Nombre público	Nombre de secuencia	*Cluster*
WBGene00012768	eef-1B.2	Y41E3.10	A
WBGene00017925	F29B9.11	F29B9.11	A
WBGene00007630	har-1	C16C10.11	A
WBGene00019680	K12H4.5	K12H4.5	A
WBGene00013463	kdp-1	Y67H2A.5	A
WBGene00003065	lpd-9	T21C9.5	A
WBGene00015248	mai-2	B0546.1	A
WBGene00004444	rpl-30	Y106G6H.3	A
WBGene00004453	rpl-39	C26F1.9	A
WBGene00004497	rps-28	Y41D4B.5	A
WBGene00004498	rps-29	B0412.4	A
WBGene00021248	Y22D7AL.10	Y22D7AL.10	A
WBGene00006919	vha-10	F46F11.5	C
WBGene00009882	vha-17	F49C12.13	C
WBGene00006918	vha-9	ZK970.4	C
WBGene00015514	srlf-1	C06A8.3	D
WBGene00000534	cpi-2	R01B10.1	F
WBGene00000881	cyn-5	F31C3.1	F
WBGene00000263	F23H11.5	F23H11.5	F
WBGene00010174	F56H9.2	F56H9.2	F
WBGene00001427	fkb-2	Y18D10A.19	F
WBGene00020894	T28D9.1	T28D9.1	F
WBGene00014454	MTCE.7	MTCE.7	Nan
WBGene00016655	acbp-1	C44E4.6	Otro

WBGene00000063	act-1	T04C12.6	Otro
WBGene00016746	C48B6.10	C48B6.10	Otro
WBGene00000896	dad-1	F57B10.10	Otro
WBGene00001005	dlc-1	T26A5.9	Otro
WBGene00009880	F49C12.11	F49C12.11	Otro
WBGene00009915	F52A8.1	F52A8.1	Otro
WBGene00001423	fib-1	T01C3.7	Otro
WBGene00006434	prdx-2	F09E5.15	Otro
WBGene00011156	rbm-3.2	R09B3.3	Otro
WBGene00021427	sec-61.B	Y38F2AR.9	Otro
WBGene00004914	snr-1	Y116A8C.42	Otro
WBGene00004918	snr-5	ZK652.1	Otro
WBGene00010896	snu-13	M28.5	Otro
WBGene00021960	tmem-258	Y57E12AM.1	Otro
WBGene00021420	trap-3	Y38F2AR.2	Otro
WBGene00013238	trap-4	Y56A3A.21	Otro
WBGene00012602	Y38E10A.24	Y38E10A.24	Otro
WBGene00006984	zig-7	F54D7.4	Otro

Por otro lado, el *CeNGEN project* (Hammarlund *et al.*[160]), tiene como objetivo compilar un atlas completo de expresión génica de todo sistema nervioso de *C. elegans* a resolución de neuronas individuales. El proyecto incluye una aplicación web (https://cengen.shinyapps.io/CengenApp/) en la cual se puede explorar distintos datos de los genes de *C. elegans*, como su expresión en distintos tipos celulares, encontrar marcadores, genes diferencialmente expresados en cierto tipo celular, y realizar *heatmaps* de expresión celular. Con esta última aplicación, se generó la Figura 20, en donde se ve la expresión de los genes de la muestra positiva en las diferentes neuronas que tiene el proyecto. Observando la figura, es claro que no todos los genes de la muestra positiva se expresan en la misma cantidad ni en las mismas células, en donde células como ALN, AVK y RIS tienen una baja expresión de estos genes.

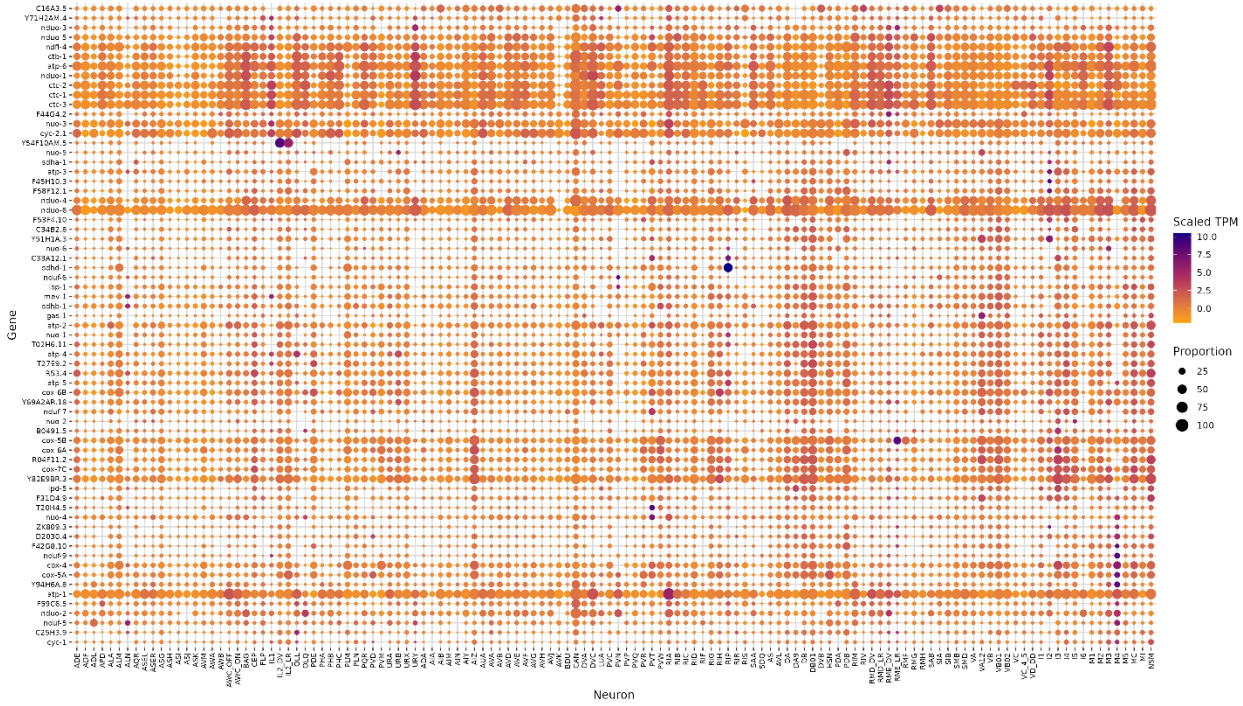


Figura 20: Heatmap de la expresión de los genes que conforman los ejemplos positivos en las células neuronales del CeNGEN Project.

A raíz de esta observación, se bajaron los datos curados de expresión y se buscó en qué tipos celulares más del 20% de las células expresan más de la mitad de los ejemplos positivos, y se encontró que las células AIZ, VB01, DB, DB01 y NSM cumplían esa condición. Posteriormente, se determinó qué genes de la lista consenso se expresan en más del 20% de las células de esos tipos celulares. Los genes que cumplen esta condición se listan en la

Tabla 16.

Tabla 16: Genes de la lista consenso que según el CeNGEN Project se expresan en las mismas neuronas que los ejemplos positivos.

WBID	Public_Name	Sequence_Name	*Cluster*
WBGene00017925	*F29B9.11*	*F29B9.11*	A
WBGene00015248	*mai-2*	*B0546.1*	A
WBGene00004453	*rpl-39*	*C26F1.9*	A
WBGene00004497	*rps-28*	*Y41D4B.5*	A
WBGene00004498	*rps-29*	*B0412.4*	A
WBGene00006910	*vha-1*	*R10E11.8*	C
WBGene00006919	*vha-10*	*F46F11.5*	C
WBGene00009882	*vha-17*	*F49C12.13*	C
WBGene00006917	*vha-8*	*C17H12.14*	C
WBGene00006918	*vha-9*	*ZK970.4*	C
WBGene00000534	*cpi-2*	*R01B10.1*	F

WBGene00020868	*eif-1*	*T27F7.3*	F
WBGene00000263	*F23H11.5*	*F23H11.5*	F
WBGene00004266	*rab-1*	*C39F7.4*	F
WBGene00009050	*mstr-1*	*F22D6.2*	F
WBGene00001427	*fkb-2*	*Y18D10A.19*	F
WBGene00020894	*T28D9.1*	*T28D9.1*	F
WBGene00000182	*arf-1*	*B0336.2*	G
WBGene00014454	*MTCE.7*	*MTCE.7*	Nan
WBGene00022599	*daf-41*	*ZC395.10*	Otro
WBGene00001946	*his-72*	*Y49E10.6*	Otro
WBGene00004046	*plp-1*	*F45E4.2*	Otro
WBGene00001005	*dlc-1*	*T26A5.9*	Otro

Este es otro abordaje simple que permite asociar genes de la lista consenso con genes de la muestra de entrenamiento a partir de información independiente a la utilizada en nuestro estudio, aportando así evidencia adicional que relaciona ambas listas.

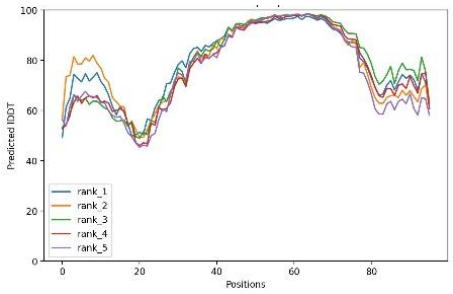
4.5.7. Homología estructural de los genes de la lista consenso sin ortólogos en humanos

De los 102 genes nucleares que conforman la lista consenso, 13 no tienen homología de secuencia con ningún gen de *H. sapiens*, y se poco se sabe de ellos. Para intentar caracterizar mejor a esos genes se decidió predecir su estructura utilizando Alphafold[161] y buscar homólogos estructurales con Foldseek[177].

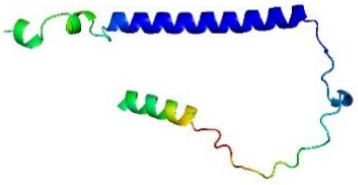
Las predicciones estructurales alcanzaron un buen nivel de confianza. De las 13 proteínas consideradas, 12 mostraron un IDDT superior a 60 en toda su estructura, exceptuando algunos aminoácidos de los extremos que fueron eliminados durante la búsqueda con Foldseek. La mayoría de las estructuras obtenidas consistieron en hélices alfa con baja probabilidad de similitud con otras proteínas conocidas (Figura S5).

Una excepción notable fue la proteína codificada por el gen *F23H11.5*, cuya estructura predicha mostró cierta similitud con la estructura de la proteína humana NDUFC1, que codifica para la subunidad C1 de la NADH deshidrogenasa (complejo I). Esta subunidad es accesoria y se cree que no está involucrada en la catálisis. Al seleccionar los ejemplos positivos para la muestra de entrenamiento, y tras búsqueda por homología de secuencia y revisión bibliográfica, esta proteína no había sido identificada en *C. elegans.* El IDDT predicho por posición (Figura 21A) fue mayor a 60 en la mayoría de las posiciones, excepto en los alrededores de la posición 20, que corresponde a un *loop* que conecta dos hélices alfa (Figura 21B). Al realizar una búsqueda de homología estructural con Foldseek, se encontró un match con la proteína NDUFC1 de humanos, con un TM-Score de 0.46 y un RMSD de 12.78 (Figura 21C). Si bien el alineamiento no es perfecto, un RMSD bastante alto nos motiva a especular que sería también parte del complejo I de *C. elegans* y estudiar más a fondo esta proteína de *C. elegans* sin función conocida.

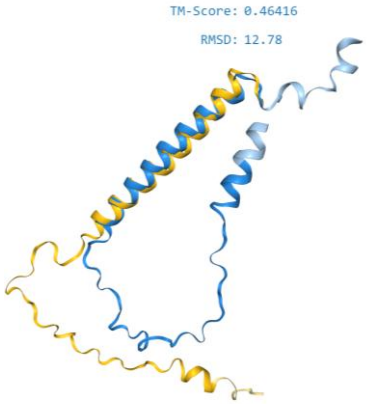
A. IDDT predicho por posición



B. Estructura predicha



C. Alineamiento de FoldSeek



F23H11.5
(WBGene00000263)

NADH dehydrogenase [ubiquinone] 1 subunit C1, mitocondrial (NDUFC1)
Homo sapiens

Figura 21: (A) IDDT predicho por posición de la proteína codificada por el gen F23H11.5. (B) Estructura predicha de la proteína codificada por el gen F23H11.5. (C) Alineamiento de la proteína codificada por el gen F23H11.5 con la proteína NDFUC1 humana.

5. Discusión.

El concepto de aprendizaje automático no es nuevo. En 1959, Arthur Samuel, empleado de IBM, introdujo la idea de diseñar computadoras capaces de realizar tareas sin ser programadas explícitamente para ello. Ya en la década de los 90, se exploraba la implementación de estos modelos para la inferencia de estructuras proteicas y su posible aplicación en ensayos clínicos. En el campo de la predicción funcional de genes, desde el comienzo del siglo se han utilizado modelos de aprendizaje supervisado para llevar a cabo estas tareas. En años recientes, el aprendizaje automático ha ganado una notable popularidad, impulsado por avances como *AlphaFold*. En este trabajo implementamos un *ensamble* de clasificadores binarios para identificar genes involucrados en la fosforilación oxidativa en *C. elegans.* Como variables predictivas utilizamos exclusivamente datos de transcripción y para entrenar los clasificadores proponemos un método que denominamos "*bagging* informado".

Abordando la fosforilación oxidativa mediante aprendizaje automático.

La predicción de función génica es un campo fundamental en la biología moderna que busca identificar las funciones de genes cuya función sigue siendo desconocida. A causa de los avances en la secuenciación genómica, la brecha entre la cantidad de genes secuenciados y aquellos con una función bien caracterizada continúa creciendo. Al dirigir los esfuerzos experimentales generando hipótesis contrastables, la asignación computacional de funciones a genes no solo facilita la comprensión de los procesos biológicos, sino que también permite aplicar ese conocimiento en campos como la medicina, la agricultura y la biotecnología. En las últimas décadas se han ensayado múltiples abordajes para la asignación computacional de funciones de genes. En este contexto, la iniciativa *Critical Assessment of Function Annotation* (CAFA)[178] ha sido un catalizador clave para el progreso en el campo, ya que proporciona una plataforma global donde probar, comparar y mejorar los métodos de anotación en un entorno estandarizado. Esta iniciativa ha sido fundamental para establecer métricas objetivas que evalúan la precisión, confiabilidad y robustez de los algoritmos predictivos.

C. elegans es un excelente modelo para aplicar y validar estas estrategias de predicción. Este organismo es fundamental en estudios biológicos debido a su simplicidad, su rápido ciclo de vida, facilidad de manipulación genética y cantidad de datos acumulados a través de décadas de investigación. Posee 19885 genes codificantes, de los cuales más del 75% no tiene una función experimentalmente establecida[179], lo que limita nuestra comprensión no solo de los procesos biológicos de *C. elegans* sino también de las similitudes y diferencias entre este y otros organismos. La predicción de función de genes mediante aprendizaje automático representa una oportunidad para reducir la brecha de conocimiento en la función de la mayoría de sus genes. En este contexto, se han utilizado algoritmos para analizar grandes cantidades de datos genómicos y fenotípicos de este organismo, revelando patrones que facilitan la asignación de funciones a genes no caracterizados[47,50,52,180–182]. Además, la robustez de *C. elegans* como modelo permite validar experimentalmente las predicciones generadas, lo que a su vez facilita seguir mejorando estos métodos[48,121,183,184]. Este tipo de estudios son particularmente valiosos, ya que los descubrimientos asociados se extienden más allá de este organismo, facilitando la anotación de genes en especies más complejas y menos estudiadas, incluyendo otros nematodos de importancia médica y agrícola[53].

La CTE ha emergido como un proceso clave para identificar genes con potencial como blancos terapéuticos en nematodos parásitos. *C. elegans*, como organismo modelo, ha sido fundamental para estudiar la CTE y su relevancia en helmintos de importancia médica, permitiendo descubrir que algunos de sus componentes pueden ser aprovechados como blancos antihelmínticos. Esto subraya la importancia de entender en detalle cómo funciona la fosforilación oxidativa, dado que su disfunción no solo está asociada con diversas enfermedades humanas, como trastornos neurodegenerativos y metabólicos, sino que también ofrece oportunidades terapéuticas en el control de parásitos.

En la literatura revisada no existen trabajos que busquen identificar genes involucrados en la fosforilación oxidativa entrenando modelos de aprendizaje supervisado. El trabajo más cercano es el de Han *et. al*[185], en el cual se utilizan modelos de regresión para identificar genes asociados al proceso de fosforilación oxidativa a partir de una lista de genes diferencialmente expresados en la enfermedad de moyamoya. Por otro lado, algunos estudios han empleado estos modelos para analizar datos transcriptómicos asociados a enfermedades o condiciones experimentales y, como resultado, han predicho genes relacionados con la CTE[186–189]. Estos estudios evidencian cómo los modelos de aprendizaje automático pueden ser herramientas efectivas para revelar genes clave implicados en procesos biológicos complejos. Sin embargo, la aplicación directa de estas herramientas al estudio de genes específicos del proceso de fosforilación oxidativa sigue siendo un área no explorada. El presente trabajo representa el primer esfuerzo en abordar este tema de manera específica, sentando las bases para futuras investigaciones que profundicen en el rol de los genes identificados como buenos candidatos a participar de este proceso y sus implicancias en la salud y la enfermedad.

Mejoramiento en la anotación génica de la fosforilación oxidativa en *C. elegans.*

Al inicio de este trabajo, nos encontramos con una carencia de bibliografía actualizada que detallara los genes implicados en el proceso de fosforilación oxidativa en *C. elegans*. Este vacío en el conocimiento se manifestaba en tres aspectos clave: por un lado, la existencia de duplicaciones génicas de ciertas proteínas para las cuales no se conocía el propósito funcional o evolutivo de dichas duplicaciones; anotaciones en Gene Ontology sin respaldo bibliográfico que dificultaban la validación de la información, y genes involucrados en la fosforilación oxidativa en mamíferos no identificados en *C. elegans* por homología de secuencia.

Nuestro laboratorio tenía experiencia previa con duplicaciones génicas que se expresaban únicamente en un tipo celular (*nduf-2.2* y *sdha-2*), lo que nos llevó a considerar que no era trivial incluir todas las duplicaciones en la lista de ejemplos positivos. Al excluirlas, observamos que, en la mayoría de los casos, uno de los genes duplicados era identificado por la mayoría de los modelos, mientras que el otro no. Sin embargo, hubo excepciones en las que ambas duplicaciones fueron predichas (como *ndab-1* y *ndab-2*, del complejo I y *asg-1* y *asg-2* del complejo V), y esto sugiere fuertemente que ambas duplicaciones génicas conformarían los complejos. Es importante señalar que se ha observado que algunos genes duplicados conforman los complejos de la CTE o la ATP sintasa en mamíferos, por lo que saber si tanto *ndab-1* como *ndab-2*, así como *asg-1* y *asg-2* conforman parte de los complejo I y V, respectivamente, requiere confirmación experimental.

Por otro lado, al diseñar la muestra de entrenamiento, se identificaron genes con anotaciones GO asociadas a este proceso, marcados con códigos IEA e IBA pero sin respaldo bibliográfico,

que fueron descartados del entrenamiento. Estos genes posteriormente fueron recuperados por la mayoría de los modelos (*C06A5.3*, *nduv-3* y *ucr-11*).

Finalmente, el hallazgo más relevante fue encontrar en la lista consenso el gen *F23H11.5,* anotado únicamente como un gen de membrana en la base de datos de GO. Este gen muestra homología de secuencia con genes predichos en otros nematodos. Al inferir la estructura de la proteína que codifica, se identificó una homología estructural con el gen *NDUFC1*, una subunidad del complejo I en *H. sapiens*.

A partir de este trabajo, reportamos la lista de genes que consideramos serían necesarios para el proceso constitutivo de fosforilación oxidativa en *C. elegans* en la Tabla 17. Los genes resaltados en negro corresponden a genes predichos por este trabajo. Once de estos genes fueron predichos por todos los modelos, lo que representa poco más del 10% de la lista consenso, mientras que 4 de ellos fueron predichos por 17 de los 18 modelos. Lógicamente, esta lista requiere de confirmación experimental. Asimismo, los resultados obtenidos en esta tesis indican que algunos genes de la cadena respiratoria podrían estar reservados para situaciones o tejidos específicos, mientras que otros se expresan de manera constitutiva. Profundizar en este hallazgo podría aportar información sobre las presiones selectivas o mecanismos evolutivos que han favorecido la conservación de duplicaciones de estos genes. Por otro lado, en esta tabla incluimos los genes *C06A5.3*, *nduv-3*, *ucr-11 y F23H11*.5, los cuales fueron predichos por todos nuestros modelos, pero hace falta validar experimentalmente. Este resultado sugiere que aún queda trabajo por hacer en la caracterización de este proceso en *C. elegans*.

Tabla 17: Lista de genes que participarían de forma constitutiva en el proceso de fosforilación oxidativa de C. elegans de acuerdo con los resultados de este trabajo. Los genes en negrita de la columna "Genes en C. elegans" corresponden a genes predichos por este trabajo. Los genes en gris de la columna "Genes en H. sapiens" corresponden a genes mitocondriales los cuales no tienen homología de secuencia con los genes de C. elegans, pero cumplen la misma función

Complejo	Gen en *C. elegans*	Clase	Gen en *H. sapiens*
I	*gas-1*	Core	*NDUS2*
I	*ndfl-4*	Core	*NDL4*
I	*ndub-6*	Core	*NDUFB6*
I	*nduf-7*	Core	*NDUFS7*
I	*nduo-1*	Core	*ND1*
I	*nduo-2*	Core	*ND2*
I	*nduo-3*	Core	*ND3*
I	*nduo-4*	Core	*ND4*
I	*nduo-5*	Core	*ND5*
I	*nduo-6*	Core	*ND6*
I	*ndus-8*	Core	*NDUFS8*
I	*nduv-2*	Core	*NDUFV2*
I	*nuo-1*	Core	*NDUFV1*
I	*nuo-2*	Core	*NDUS3*
I	*nuo-3*	Core	*NDUA6*
I	*nuo-5*	Core	*NDUFS1*
I	*nuo-6*	Core	*NDUB4*
I	***C06A5.3***	Core	-

I	***nduv-3***	Core	-
I	*lpd-5*	Accesoria	*NDUFS4*
I	*ndua-1*	Accesoria	*NDUFA1*
I	*ndua-12*	Accesoria	*NDUFA12*
I	*ndua-13*	Accesoria	*NDUFA13*
I	*ndua-5*	Accesoria	*NDUA5*
I	*ndua-7*	Accesoria	*NDUA7*
I	*ndua-8*	Accesoria	*NDUA8*
I	*ndub-10*	Accesoria	*NDUFB10*
I	*ndub-11*	Accesoria	*NDUFB11*
I	*ndub-2*	Accesoria	*NDUFB2*
I	*ndub-5*	Accesoria	*NDUFB5*
I	*ndub-7*	Accesoria	*NDUFB7*
I	*ndub-8*	Accesoria	*NDUFB8*
I	*ndub-9*	Accesoria	*NDUFB9*
I	*nduc-2*	Accesoria	*NDUFC2*
I	*nduf-11*	Accesoria	*NDUFA11*
I	*nduf-5*	Accesoria	*NDUFS5*
I	*nduf-6*	Accesoria	*NDUFS6*
I	*nduf-9*	Accesoria	*NDUFA9*
I	*nuo-4*	Accesoria	*NDUAA*
I	***ndab-1***	Accesoria	*NDUFAB1*
I	***ndab-2***	Accesoria	*NDUFAB1*
I	***ndub-3***	Accesoria	*NDUFB3*
II	*mev-1*	Core	*SDHC*
II	*sdha-1*	Core	*SDHA*
II	*sdhb-1*	Core	*SDHB*
II	*sdhd-1*	Core	*SDHD*
III	*ctb-1*	Core	*CYB*
III	*cyc-1*	Core	*CYC1*
III	*isp-1*	Core	*UQCRFS1*
III	*T02H6.11*	Core	*UQCRB*
III	*T27E9.2*	Core	*UQCRH*
III	***F45H10.2***	Core	*UQCRQ*
III	***ucr-1***	Core	*UQCRC1*
III	***ucr-11***	Core	-
III	***ucr-2.1***	Core	*UQCRC2*
III	***C14B9.10***	Accesoria	-
IV	*cox-4*	Core	*COX4I1*
IV	*cox-5a*	Core	*COX5A*
IV	*cox-5b*	Core	*COX5B*

IV	*cox-6a*	Core	*COX6A*
IV	*cox-6b*	Core	*COX6B2*
IV	*cox-7c*	Core	*COX7C*
IV	*ctc-1*	Core	*COI*
IV	*ctc-2*	Core	*COII*
IV	*ctc-3*	Core	*COIII*
IV	*cys-2.1*	Core	*CYCS*
IV	***cox-6c***	Core	-
V	*atp-1*	Core	*ATP5F1A*
V	*atp-2*	Core	*ATP5F1B*
V	*atp-3*	Core	*ATP5PO*
V	*atp-4*	Core	*ATP5PF*
V	*atp-5*	Core	*ATP5PD*
V	*atp-6*	Core	*ATP6*
V	*F58F12.1*	Core	*ATP5F1D*
V	*R04F11.2*	Core	*ATP5ME*
V	*R53.4*	Core	*ATP5MF*
V	*Y116A8C.27*	Core	*ATPAF2*
V	*Y69A2AR.18*	Core	*ATP5F1C*
V	*Y82E9BR.3*	Core	*ATP5MC1, ATP5MC2, ATP5MC3*
V	***asb-2***	Core	*ATP5PB*
V	***asg-1***	Core	*ATP5MG*
V	***asg-2***	Core	*ATP5MG*
V	***hpo-18***	Core	*ATP5F1E*
V	***asb-2***	Core	*ATP5PB*

En la Tabla 17 incorporamos los *genes C06A5.3, nduv-3, ucr-11, C14B9.10* y *cox-6c,* ninguno de los cuales presenta homología de secuencia con mamíferos. Esta característica los convierte en blancos potencialmente interesantes para el desarrollo de fármacos antihelmínticos, especialmente si estos genes están presentes en helmintos pero ausentes en sus hospederos mamíferos. Como perspectiva, se propone explorar más a fondo estos genes mediante análisis de homología estructural con proteínas de mamíferos, con el fin de evaluar su viabilidad como objetivos terapéuticos.

Respecto a las proteínas encargadas del ensamblado de los complejos, la mayoría no fueron recuperados por estos modelos. El gen que obtuvo más predicciones fue *ndua-2*, el cual fue predicho por 8 de 18 modelos. Este gen recientemente fue identificado no solo por llevar a cabo el ensamblado del complejo I, sino que se mantiene anclado al complejo[190], por lo que podríamos haberla considerado una proteína accesoria. Para próximos trabajos, se podría evaluar si las predicciones y las métricas cambian al agregar este gen a la lista de ejemplos positivos.

Las predicciones cantadas y no tan cantadas.

En este trabajo, se predijo que poco más de un centenar de genes están asociados a la fosforilación oxidativa según 18 modelos de aprendizaje supervisado. Dejando de lado los genes

que fueron incluidos en la Tabla 17, son 93 los genes de la lista consenso que inicialmente no se consideraban vinculados a este proceso. Al revisar estos genes, se observó que varios de ellos participan en procesos relacionados indirectamente con la fosforilación oxidativa, ya sea produciendo moléculas que este proceso utiliza o consumiendo ATP generado por él. El hecho de que los modelos hayan identificado genes involucrados en funciones asociadas a la fosforilación oxidativa sugiere que la metodología empleada es robusta y capaz de generar resultados coherentes.

Por un lado, la mayoría de las enzimas del ciclo de Krebs fueron predichas por casi todos los modelos en este trabajo. Se ha mencionado anteriormente cómo el ciclo de Krebs proporciona sustratos y co-factores que son utilizados por los complejos I y II, por lo que la predicción de estas enzimas no es sorprendente. Sin embargo, es interesante notar que ni los genes de la beta-oxidación ni los genes de la glucólisis fueron predichos. El grado de vinculación entre la CTE y el ciclo de Krebs es tal que en organismos que no se utiliza oxígeno como aceptor final de electrones, muchos veces no se tiene un ciclo de Krebs completo[191,192], o incluso es inexistente[193], en cambio sí presentan glucólisis y otras vías de degradación de lípidos. Este resultado sugiere que las predicciones están particularmente asociadas al proceso de fosforilación oxidativa y no simplemente al catabolismo general.

Por otro lado, están los genes *cct* y *trap*, cuyos productos están relacionados con el plegamiento y localización de proteínas, respectivamente. Ambos procesos son claves para que las proteínas sintetizadas puedan llevar a cabo su función en el lugar y la forma adecuada. La relación con las proteínas *cct* es más clara, ya que estas forman un complejo que se encarga del correcto plegamiento de las proteínas a consta de ATP. En particular, estas chaperonas se encargan del plegamiento de proteínas grandes, en donde se destacan las proteínas citoesqueléticas como actina y tubulina, entre otras[194]. Sin embargo, no explica por qué los modelos fueron capaces de predecir estas chaperonas y no otras, ya que no son las únicas que se expresan de forma constitutiva y hacen uso de ATP.

El vínculo con las proteínas *trap* no es tan directo. Este complejo se encuentra asociado al RE y se encarga de insertar proteínas que están siendo sintetizadas en la membrana de este. Sin embargo, las proteínas de los complejos de fosforilación oxidativa que son codificadas en el núcleo son sintetizadas en el citosol e insertadas en la MMI por el complejo de translocasas de la membrana interna (TIM)[5]. No es evidente por qué todas las proteínas *trap* son predichas por estos modelos más allá de la relación que existe entre la producción de ATP y la síntesis proteica, proceso que se da en condiciones de abundancia energética. De alguna forma la inserción de proteínas en la membrana del RE se vincula con la producción o la disponibilidad de energía, lo que permite se predigan estos genes al entrenar modelos con genes de la fosforilación oxidativa.

Por último, se identificaron también los genes *vha*, los cuales codifican para proteínas que llevan a cabo un proceso que requiere grandes cantidades de ATP. No obstante, hay una particularidad que no deja de llamar la atención: tanto el complejo V como las ATPasas son proteínas estructuralmente similares, y ambas comparten la función de transportar protones a través de una membrana, aunque con objetivos diferentes. Mientras que el complejo V está destinado a la síntesis de ATP, las ATPasas lisosomales se encargan de acidificar el interior del lisosoma a expensas de ATP. Siendo que un proceso hace uso del ATP producido por el otro, uno podría hipotetizar que la expresión de ambos grupos de genes están co-regulados, pero al observar el grafo de co-expresión vemos que se encuentran en *clusters* diferentes. Los modelos de aprendizaje supervisado fueron capaces de tomar patrones que no se limitan a la correlación de

la expresión de estos genes. Sería interesante indagar en este resultado para entender el vínculo entre ellos se limita a su interdependencia energética o si hay algo más.

Candidatos sin función conocida.

De los 103 genes de la lista consenso, de 15 de ellos no se conoce su función. A partir de este trabajo se pudo identificar que la proteína codificada por el gen *F23H11.5* posee cierta similitud estructural con la proteína NDUFC1 de humanos. De los restantes, algunos de ellos tienen anotaciones en GO respecto a su locación celular o función molecular (Tabla 18), pero no se conocen más detalles.

Los genes del *cluster* A son sin dudas los candidatos más interesantes, ya que en este *cluster* se encuentra la mayor cantidad de ejemplos positivos de la muestra de entrenamiento. Los genes *W08E12.7, C25A1.16 y Y71F9AL.9* no fueron recuperados por los estudios de expresión diferencial ni están las listas de CeNGEN, lo que sugiere que, en principio, no serían candidatos clave a estar involucrados en el proceso de fosforilación oxidativa. Por otro lado, las anotaciones de los genes *F56H9.2, rbm-3.2* y *zig-7* sugieren que podrían estar participando en procesos relacionados con la fosforilación oxidativa, pero no a la fosforilación en sí, aunque también podrían tratarse de falsos positivos. En cuanto a los genes restantes, únicamente dos poseen anotaciones en GO, ambas relacionadas con su localización en membranas, lo que resulta relevante, ya que los complejos de la fosforilación oxidativa también están asociados a esta estructura.

*Tabla 18: Anotación de genes de la lista consenso sin función definida. *IPI = Inferred from Physical Interaction*

Gen	*Cluster*	Anotación en GO	Nombre del término	Evidencia de GO
F29B9.11	A	GO:0016020	Membrana	IEA
K12H4.5	A	Sin anotaciones	-	-
W08E12.7	A	Sin anotaciones	-	-
srlf-1	D	Sin anotaciones	-	-
F56H9.2	F	GO:0005840	Ribosoma	IEA
T28D9.1	F	Sin anotaciones	-	-
C25A1.16	Otro	Sin anotaciones	-	-
C48B6.10	Otro	GO:0016020	Membrana	IEA
F49C12.11	Otro	Sin anotaciones	-	-
F52A8.1	Otro	GO:0016020	Membrana	IEA
rbm-3.2	Otro	GO:0003676	Unión a ácidos nucleicos	IEA
	Otro	GO:0003723	Unión a ARN	IEA
	Otro	GO:0005515	Unión a proteína	IPI*
Y38E10A.24	Otro	Sin anotaciones	-	-
Y71F9AL.9	Otro	Sin anotaciones	-	-
zig-7	Otro	GO:0005576	Región extracelular	IEA

Otro	GO:0007411	Axon guidance	IEA

El estudio de estos genes cuya función aún se desconoce podría llevar al descubrimiento de nuevos genes involucrados (o asociados) con la fosforilación oxidativa. Lo que hace que estos genes sean particularmente interesantes es que muchos de ellos no tienen homología con genes humanos, lo que abre la posibilidad de que puedan ser utilizados como blancos para el desarrollo de nuevos antihelmínticos. Explorar sus funciones no solo podría ampliar nuestro conocimiento sobre este proceso, sino también ofrecer oportunidades innovadoras para combatir infecciones por parásitos, aprovechando estas diferencias genéticas.

6. Conclusiones y perspectivas.

En este trabajo se diseñó un enfoque experimental novedoso para identificar genes involucrados en el proceso de fosforilación oxidativa en *C. elegans*. Hasta donde tenemos conocimiento, este es el primer estudio que utiliza una red de co-expresión como herramienta para seleccionar ejemplos negativos destinados al entrenamiento de modelos predictivos. Este enfoque permitió entrenar modelos con buenas métricas, mostrando su eficacia para abordar un problema complejo como la predicción de genes en procesos biológicos.

Mediante la implementación y ensamble de múltiples modelos entrenados con una estrategia de *bagging* informado, obtuvimos una lista consenso de genes reducida y enriquecida en términos GO asociados al proceso de fosforilación oxidativa. Esta lista no solo incluye genes previamente conocidos por su participación en este proceso, como los relacionados con el ciclo de Krebs, sino también genes involucrados en procesos que utilizan el ATP generado, lo que refuerza la robustez del abordaje propuesto para identificar genes funcionalmente vinculados al proceso de fosforilación oxidativa.

Un hallazgo destacado fue que no todas las duplicaciones génicas fueron predichas por todos los modelos. Este análisis reveló que, en la mayoría de los casos, solo una de las copias duplicadas o triplicadas era consistentemente predicha por los modelos, mientras que las otras no lo eran. No obstante, hubo excepciones en las que ambas duplicaciones fueron clasificadas como positivas (*ndab-1/ndab-2* y *asg-1/asg-2*), lo que sugiere que algunas duplicaciones podrían estar asociadas a roles específicos en ciertos tejidos o condiciones fisiológicas, mientras que otras parecen desempeñar funciones más generales. Este patrón plantea preguntas sobre los mecanismos evolutivos que han llevado a la conservación de estas duplicaciones y resalta la importancia de explorar su regulación y funcionalidad. Una estrategia sería medir por respirometría el consumo de oxígeno de gusanos mutantes en genes duplicados no esenciales para identificar si una de las copias provoca un mal funcionamiento de la respiración mitocondrial, o si por lo contrario respiran igual que las cepas silvestres.

Adicionalmente, la lista consenso incluye numerosos genes sin función conocida, muchos de los cuales carecen de homólogos en humanos. Entre ellos, el gen *F23H11.5* mostró una similitud estructural con la proteína humana NDUFC1, una subunidad del complejo I de la cadena de transporte de electrones, que habíamos identificado previamente como ausente en el genoma de *C. elegans*. Por otro lado, los demás genes sin función conocida no mostraron homologías relevantes con genes humanos, aunque varios están anotados en GO como genes de membrana. Esto plantea la posibilidad de que algunos de ellos puedan estar relacionados con la cadena de transporte de electrones o con la síntesis de ATP a través del complejo V. Una forma de abordar esta pregunta sería aislar los complejos a partir de mitocondrias de *C. elegans* y la separación de los complejos por electroforesis de geles nativos azules, los cuales permiten separar complejos proteicos según su tamaño, forma y carga[195], para luego realizar ensayos de proteómica y determinar las proteínas presentes.

Más allá de las pruebas experimentales, el próximo paso que tomaremos con estos resultados será re-entrenar los modelos utilizando la Tabla 17 como ejemplos positivos, incluyendo también el gen ndua-*2,* y explorar otras formas de selección de ejemplos negativos, con el fin de evaluar cambios en las métricas y predicciones.

En conjunto, estos resultados aportan evidencia sobre la efectividad del abordaje propuesto para la predicción de genes implicados en procesos complejos y resaltan la importancia de continuar caracterizando los genes identificados, tanto para comprender mejor el proceso de fosforilación oxidativa como para explorar su potencial como blancos específicos en el desarrollo de estrategias terapéuticas, particularmente en el contexto de organismos modelo como *C. elegans*.

7. Bibliografìa

1. Zhu J, Vinothkumar KR, Hirst J. Structure of mammalian respiratory complex I. *Nature*. 2016;536(7616):354-358. doi:10.1038/nature19095
2. Fiedorczuk K, Letts JA, Degliesposti G, Kaszuba K, Skehel M, Sazanov LA. Atomic structure of the entire mammalian mitochondrial complex I. *Nature*. 2016;538(7625):406-410. doi:10.1038/nature19794
3. Wu M, Gu J, Guo R, Huang Y, Yang M. Structure of Mammalian Respiratory Supercomplex I 1 III 2 IV 1. *Cell*. 2016;167(6):1598-1609.e10. doi:10.1016/j.cell.2016.11.012
4. Vercellino I, Sazanov LA. The assembly, regulation and function of the mitochondrial respiratory chain. *Nat Rev Mol Cell Biol*. 2022;23(2):141-161. doi:10.1038/s41580-021-00415-0
5. Cogliati S, Lorenzi I, Rigoni G, Caicci F, Soriano ME. Regulation of Mitochondrial Electron Transport Chain Assembly. *Journal of Molecular Biology*. 2018;430(24):4849-4873. doi:10.1016/j.jmb.2018.09.016
6. Sousa JS, D'Imprima E, Vonck J. Mitochondrial Respiratory Chain Complexes. In: Harris JR, Boekema EJ, eds. *Membrane Protein Complexes: Structure and Function*. Vol 87. Subcellular Biochemistry. Springer Singapore; 2018:167-227. doi:10.1007/978-981-10-7757-9_7
7. Rasheed MRHA, Tarjan G. Succinate Dehydrogenase Complex: An Updated Review. *Archives of Pathology & Laboratory Medicine*. 2018;142(12):1564-1570. doi:10.5858/arpa.2017-0285-RS
8. Xia D, Esser L, Tang WK, et al. Structural analysis of cytochrome bc1 complexes: Implications to the mechanism of function. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*. 2013;1827(11-12):1278-1294. doi:10.1016/j.bbabio.2012.11.008
9. Keilin D, Hartree EF. Activity of the cytochrome system in heart muscle preparations. *Biochemical Journal*. 1947;41(4):500-502. doi:10.1042/bj0410500
10. Chance B. Electron Transport in the Oxysome. *Science*. 1963;140(3565):370-380. doi:10.1126/science.140.3565.379-c.
11. Hackenbrock CR, Chazotte B, Gupte SS. The random collision model and a critical assessment of diffusion and collision in mitochondrial electron transport. *J Bioenerg Biomembr*. 1986;18(5):331-368. doi:10.1007/BF00743010
12. Schagger H. Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. *The EMBO Journal*. 2000;19(8):1777-1783. doi:10.1093/emboj/19.8.1777
13. Acín-Pérez R, Fernández-Silva P, Peleato ML, Pérez-Martos A, Enriquez JA. Respiratory Active Mitochondrial Supercomplexes. *Molecular Cell*. 2008;32(4):529-539. doi:10.1016/j.molcel.2008.10.021
14. Zheng W, Chai P, Zhu J, Zhang K. High-resolution in situ structures of mammalian respiratory supercomplexes. *Nature*. 2024;631(8019):232-239. doi:10.1038/s41586-024-07488-9
15. Fernandez-Vizarra E, Zeviani M. Mitochondrial disorders of the OXPHOS system. *FEBS Letters*. 2021;595(8):1062-1106. doi:10.1002/1873-3468.13995
16. Parker D, Parks J, Kleinschmidt-DeMasters BK. Electron transnort chain defects in Alzheimeh disease brain.
17. Yao PJ, Eren E, Goetzl EJ, Kapogiannis D. Mitochondrial Electron Transport Chain Protein Abnormalities Detected in Plasma Extracellular Vesicles in Alzheimer's Disease. *Biomedicines*. 2021;9(11):1587. doi:10.3390/biomedicines9111587
18. Parker WD, Boyson SJ, Parks JK. Abnormalities of the electron transport chain in idiopathic parkinson's disease. *Annals of Neurology*. 1989;26(6):719-723. doi:10.1002/ana.410260606

19. Li JL, Lin TY, Chen PL, et al. Mitochondrial Function and Parkinson's Disease: From the Perspective of the Electron Transport Chain. *Front Mol Neurosci*. 2021;14:797833. doi:10.3389/fnmol.2021.797833
20. Chauhan A, Gu F, Essa MM, et al. Brain region-specific deficit in mitochondrial electron transport chain complexes in children with autism. *Journal of Neurochemistry*. 2011;117(2):209-220. doi:10.1111/j.1471-4159.2011.07189.x
21. Ghanizadeh A, Berk M, Farrashbandi H, Alavi Shoushtari A, Villagonzalo KA. Targeting the mitochondrial electron transport chain in autism, a systematic review and synthesis of a novel therapeutic approach. *Mitochondrion*. 2013;13(5):515-519. doi:10.1016/j.mito.2012.10.001
22. Wang JF. Defects of Mitochondrial Electron Transport Chain in Bipolar Disorder: Implications for Mood-Stabilizing Treatment. *Can J Psychiatry*. 2007;52(12):753-762. doi:10.1177/070674370705201202
23. Moreno-Lastres D, Fontanesi F, García-Consuegra I, et al. Mitochondrial Complex I Plays an Essential Role in Human Respirasome Assembly. *Cell Metabolism*. 2012;15(3):324-335. doi:10.1016/j.cmet.2012.01.015
24. Baker N, Patel J, Khacho M. Linking mitochondrial dynamics, cristae remodeling and supercomplex formation: How mitochondrial structure can regulate bioenergetics. *Mitochondrion*. 2019;49:259-268. doi:10.1016/j.mito.2019.06.003
25. Cogliati S, Cabrera-Alarcón JL, Enriquez JA. Regulation and functional role of the electron transport chain supercomplexes. *Biochemical Society Transactions*. 2021;49(6):2655-2668. doi:10.1042/BST20210460
26. Enríquez JA. Supramolecular Organization of Respiratory Complexes. *Annu Rev Physiol*. 2016;78(1):533-561. doi:10.1146/annurev-physiol-021115-105031
27. Nigon VM, Felix MA. History of research on C. elegans and other free-living nematodes as model organisms. *WormBook*. Published online September 7, 2017:1-84. doi:10.1895/wormbook.1.181.1
28. Brenner S. THE GENETICS OF *CAENORHABDITIS ELEGANS*. *Genetics*. 1974;77(1):71-94. doi:10.1093/genetics/77.1.71
29. Frézal L, Félix MA. C. elegans outside the Petri dish. *eLife*. 2015;4:e05849. doi:10.7554/eLife.05849
30. Corsi AK, Wightman B, Chalfie M. A Transparent window into biology: A primer on Caenorhabditis elegans. *WormBook*. Published online June 18, 2015:1-31. doi:10.1895/wormbook.1.177.1
31. Stein LD. Internet access to the C. elegans genome. *Trends in Genetics*. 1999;15(10):425-427. doi:10.1016/S0168-9525(99)01805-3
32. Hodgkin J, Horvitz HR, Jasny BR, Kimble J. *C. elegans* : Sequence to Biology. *Science*. 1998;282(5396):2011-2011. doi:10.1126/science.282.5396.2011
33. Lai CH, Chou CY, Ch'ang LY, Liu CS, Lin W chang. Identification of Novel Human Genes Evolutionarily Conserved in *Caenorhabditis elegans* by Comparative Proteomics. *Genome Res*. 2000;10(5):703-713. doi:10.1101/gr.10.5.703
34. Venter JC, Adams MD, Myers EW, et al. The Sequence of the Human Genome. *Science*. 2001;291(5507):1304-1351. doi:10.1126/science.1058040
35. Giunti S, Andersen N, Rayes D, De Rosa MJ. Drug discovery: Insights from the invertebrate *Caenorhabditis elegans*. *Pharmacology Res & Perspec*. 2021;9(2):e00721. doi:10.1002/prp2.721
36. Alvarez J, Alvarez-Illera P, García-Casas P, Fonteriz RI, Montero M. The Role of Ca2+ Signaling in Aging and Neurodegeneration: Insights from Caenorhabditis elegans Models. *Cells*. 2020;9(1):204. doi:10.3390/cells9010204

37. Alvarez J, Alvarez-Illera P, Santo-Domingo J, Fonteriz RI, Montero M. Modeling Alzheimer's Disease in Caenorhabditis elegans. *Biomedicines*. 2022;10(2):288. doi:10.3390/biomedicines10020288
38. Calahorro F, Ruiz-Rubio M. Caenorhabditis elegans as an experimental tool for the study of complex neurological diseases: Parkinson's disease, Alzheimer's disease and autism spectrum disorder. *Invert Neurosci*. 2011;11(2):73-83. doi:10.1007/s10158-011-0126-1
39. Maulik M, Mitra S, Bult-Ito A, Taylor BE, Vayndorf EM. Behavioral Phenotyping and Pathological Indicators of Parkinson's Disease in C. elegans Models. *Front Genet*. 2017;8:77. doi:10.3389/fgene.2017.00077
40. Chege PM, McColl G. Caenorhabditis elegans: a model to investigate oxidative stress and metal dyshomeostasis in Parkinson's disease. *Front Aging Neurosci*. 2014;6. doi:10.3389/fnagi.2014.00089
41. Gaeta AL, Caldwell KA, Caldwell GA. Found in Translation: The Utility of C. elegans Alpha-Synuclein Models of Parkinson's Disease. *Brain Sciences*. 2019;9(4):73. doi:10.3390/brainsci9040073
42. Bakare AB, Lesnefsky EJ, Iyer S. Leigh Syndrome: A Tale of Two Genomes. *Front Physiol*. 2021;12:693734. doi:10.3389/fphys.2021.693734
43. Therrien M, Parker JA. Worming forward: amyotrophic lateral sclerosis toxicity mechanisms and genetic interactions in Caenorhabditis elegans. *Front Genet*. 2014;5. doi:10.3389/fgene.2014.00085
44. Anjum M, Laitila A, Ouwehand AC, Forssten SD. Current Perspectives on Gastrointestinal Models to Assess Probiotic-Pathogen Interactions. *Front Microbiol*. 2022;13:831455. doi:10.3389/fmicb.2022.831455
45. Kumar A, Baruah A, Tomioka M, Iino Y, Kalita MC, Khan M. Caenorhabditis elegans: a model to understand host–microbe interactions. *Cell Mol Life Sci*. 2020;77(7):1229-1249. doi:10.1007/s00018-019-03319-7
46. Poupet C, Chassard C, Nivoliez A, Bornes S. Caenorhabditis elegans, a Host to Investigate the Probiotic Properties of Beneficial Microorganisms. *Front Nutr*. 2020;7:135. doi:10.3389/fnut.2020.00135
47. Kaletsky R, Yao V, Williams A, et al. Transcriptome analysis of adult Caenorhabditis elegans cells reveals tissue-specific gene and isoform expression. Barsh GS, ed. *PLoS Genet*. 2018;14(8):e1007559. doi:10.1371/journal.pgen.1007559
48. Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG. Global Prediction of Tissue-Specific Gene Expression and Context-Dependent Gene Networks in Caenorhabditis elegans. Stormo GD, ed. *PLoS Comput Biol*. 2009;5(6):e1000417. doi:10.1371/journal.pcbi.1000417
49. Li YH, Dong MQ, Guo Z. Systematic analysis and prediction of longevity genes in Caenorhabditis elegans. *Mechanisms of Ageing and Development*. 2010;131(11-12):700-709. doi:10.1016/j.mad.2010.10.001
50. Baruch L, Itzkovitz S, Golan-Mashiach M, Shapiro E, Segal E. Using Expression Profiles of Caenorhabditis elegans Neurons To Identify Genes That Mediate Synaptic Connectivity. Sporns O, ed. *PLoS Comput Biol*. 2008;4(7):e1000120. doi:10.1371/journal.pcbi.1000120
51. Qi Y, Missiuro PE, Kapoor A, et al. Semi-supervised analysis of gene expression profiles for lineage-specific development in the *Caenorhabditis elegans* embryo. *Bioinformatics*. 2006;22(14):e417-e423. doi:10.1093/bioinformatics/btl256
52. Townes FW, Carr K, Miller JW. Identifying longevity associated genes by integrating gene expression and curated annotations. Lavrik I, ed. *PLoS Comput Biol*. 2020;16(11):e1008429. doi:10.1371/journal.pcbi.1008429

53. Ben Or G, Veksler-Lublinsky I. Comprehensive machine-learning-based analysis of microRNA–target interactions reveals variable transferability of interaction rules across species. *BMC Bioinformatics*. 2021;22(1):264. doi:10.1186/s12859-021-04164-x
54. Maglioni S, Ventura N. C. elegans as a model organism for human mitochondrial associated disorders. *Mitochondrion*. 2016;30:117-125. doi:10.1016/j.mito.2016.02.003
55. Haroon S, Li A, Weinert JL, et al. Multiple Molecular Mechanisms Rescue mtDNA Disease in C. elegans. *Cell Reports*. 2018;22(12):3115-3125. doi:10.1016/j.celrep.2018.02.099
56. Cook SJ, Jarrell TA, Brittin CA, et al. Whole-animal connectomes of both Caenorhabditis elegans sexes. *Nature*. 2019;571(7763):63-71. doi:10.1038/s41586-019-1352-7
57. Li J, Cai T, Wu P, et al. Proteomic analysis of mitochondria from *Caenorhabditis elegans*. *Proteomics*. 2009;9(19):4539-4553. doi:10.1002/pmic.200900101
58. Dancy BM. Mitochondrial bioenergetics and disease in Caenorhabditis elegans. *Front Biosci*. 2015;20(2):198-228. doi:10.2741/4305
59. Tsang WY, Lemire BD. The role of mitochondria in the life of the nematode, Caenorhabditis elegans. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. 2003;1638(2):91-105. doi:10.1016/S0925-4439(03)00079-6
60. Hahnel SR, Dilks CM, Heisler I, Andersen EC, Kulke D. Caenorhabditis elegans in anthelmintic research – Old model, new perspectives. *International Journal for Parasitology: Drugs and Drug Resistance*. 2020;14:237-248. doi:10.1016/j.ijpddr.2020.09.005
61. Southampton Neuroscience Group (SoNG), Centre for Biological Sciences, University of Southampton, Southampton SO17 1BJ, UK., Holden-Dye L, Walker RJ. Anthelmintic drugs and nematicides: studies in Caenorhabditis elegans. *WormBook*. Published online December 16, 2014:1-29. doi:10.1895/wormbook.1.143.2
62. Tielens AGM, Hellemond JJV. The electron transport chain in anaerobically functioning eukaryotes. *Biochimica et Biophysica Acta*.
63. Lautens MJ, Tan JH, Serrat X, Del Borrello S, Schertzberg MR, Fraser AG. Identification of enzymes that have helminth-specific active sites and are required for Rhodoquinone-dependent metabolism as targets for new anthelmintics. Jex AR, ed. *PLoS Negl Trop Dis*. 2021;15(11):e0009991. doi:10.1371/journal.pntd.0009991
64. Davie T, Serrat X, Snider J, et al. Identification of a novel family of benzimidazole species-selective Complex I inhibitors as potential anthelmintics.
65. Liu WC, Ren YX, Hao AY, et al. The activities of wortmannilactones against helminth electron transport chain enzymes, structure-activity relationships, and the effect on Trichinella spiralis infected mice. *J Antibiot*. 2018;71(8):731-740. doi:10.1038/s41429-018-0061-z
66. Ōmura S, Miyadera H, Ui H, et al. An anthelmintic compound, nafuredin, shows selective inhibition of complex I in helminth mitochondria. *Proc Natl Acad Sci USA*. 2001;98(1):60-62. doi:10.1073/pnas.98.1.60
67. Mathew MD, Mathew ND, Miller A, et al. Using C. elegans Forward and Reverse Genetics to Identify New Compounds with Anthelmintic Activity. Prichard RK, ed. *PLoS Negl Trop Dis*. 2016;10(10):e0005058. doi:10.1371/journal.pntd.0005058
68. Ishii I. Reprofiling a classical anthelmintic, pyrvinium pamoate, as an anti-cancer drug targeting mitochondrial respiration. *Frontiers in Oncology*.
69. Schleker ASS, Rist M, Matera C, et al. Mode of action of fluopyram in plant-parasitic nematodes. *Sci Rep*. 2022;12(1):11954. doi:10.1038/s41598-022-15782-7
70. Enkai S. Mitochondrial complex III in larval stage of Echinococcus multilocularis as a potential chemotherapeutic target and in vivo efficacy of atovaquone against primary hydatid cysts. *Parasitology International*. Published online 2020.

71. Tarca AL, Carey VJ, Chen X wen, Romero R, Drăghici S. Machine Learning and Its Applications to Biology. Lewitter F, ed. *PLoS Comput Biol*. 2007;3(6):e116. doi:10.1371/journal.pcbi.0030116
72. Fix E, Hodges JL. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique*. 1989;57(3):238. doi:10.2307/1403797
73. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018
74. Tin Kam Ho. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol 1. IEEE Comput. Soc. Press; 1995:278-282. doi:10.1109/ICDAR.1995.598994
75. Zhang C, Ma Y, eds. 1.2.3.1 Combining Class Labels. In: *Ensemble Machine Learning: Methods and Applications*. Springer New York; 2012:6-8. doi:10.1007/978-1-4419-9326-7
76. Noorbakhsh J, Chandok H, Karuturi RKM, George J. Machine Learning in Biology and Medicine. *Advances in Molecular Pathology*. 2019;2(1):143-152. doi:10.1016/j.yamp.2019.07.010
77. You Y, Lai X, Pan Y, et al. Artificial intelligence in cancer target identification and drug discovery. *Sig Transduct Target Ther*. 2022;7(1):156. doi:10.1038/s41392-022-00994-0
78. Papalia GF, Brigato P, Sisca L, et al. Artificial Intelligence in Detection, Management, and Prognosis of Bone Metastasis: A Systematic Review. *Cancers*. 2024;16(15):2700. doi:10.3390/cancers16152700
79. Shimizu H, Nakayama KI. Artificial intelligence in oncology. *Cancer Science*. 2020;111(5):1452-1460. doi:10.1111/cas.14377
80. Patil S, Moafa IH, Mosa Alfaifi M, et al. Reviewing the Role of Artificial Intelligence in Cancer. *Asian Pac J Cancer Biol*. 2020;5(4):189-199. doi:10.31557/apjcb.2020.5.4.189-199
81. Aneja S, Chang E, Omuro A. Applications of artificial intelligence in neuro-oncology. *Current Opinion in Neurology*. 2019;32(6):850-856. doi:10.1097/WCO.0000000000000761
82. Voigtlaender S, Pawelczyk J, Geiger M, et al. Artificial intelligence in neurology: opportunities, challenges, and policy implications. *J Neurol*. 2024;271(5):2258-2273. doi:10.1007/s00415-024-12220-8
83. Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. *Annu Rev Public Health*. 2020;41(1):21-36. doi:10.1146/annurev-publhealth-040119-094437
84. Rose S. Intersections of machine learning and epidemiological methods for health services research. *International Journal of Epidemiology*. 2021;49(6):1763-1770. doi:10.1093/ije/dyaa035
85. Pastore VP, Ciranni M, Bianco S, Fung JC, Murino V, Odone F. Efficient unsupervised learning of biological images with compressed deep features. *Image and Vision Computing*. 2023;137:104764. doi:10.1016/j.imavis.2023.104764
86. Dilsizian ME, Siegel EL. Machine Meets Biology: a Primer on Artificial Intelligence in Cardiology and Cardiac Imaging. *Curr Cardiol Rep*. 2018;20(12):139. doi:10.1007/s11886-018-1074-8
87. Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial Intelligence in Cardiology. *Journal of the American College of Cardiology*. 2018;71(23):2668-2679. doi:10.1016/j.jacc.2018.03.521
88. Belcher BT, Bower EH, Burford B, et al. Demystifying image-based machine learning: a practical guide to automated analysis of field imagery using modern machine learning tools. *Front Mar Sci*. 2023;10:1157370. doi:10.3389/fmars.2023.1157370
89. W. Cassidy J, Taylor B, eds. *Artificial Intelligence in Oncology Drug Discovery and Development*. IntechOpen; 2020. doi:10.5772/intechopen.88376
90. Patel V, Shah M. Artificial intelligence and machine learning in drug discovery and development. *Intelligent Medicine*. 2022;2(3):134-140. doi:10.1016/j.imed.2021.10.001

91. Borkotoky S, Joshi A, Kaushik V, Nath Jha A. Machine Learning and Artificial Intelligence in Therapeutics and Drug Development Life Cycle. In: Akhtar J, Badruddeen, Ahmad M, Irfan Khan M, eds. *Drug Development Life Cycle*. IntechOpen; 2022. doi:10.5772/intechopen.104753
92. Gilpin W, Huang Y, Forger DB. Learning dynamics from large biological data sets: Machine learning meets systems biology. *Current Opinion in Systems Biology*. 2020;22:1-7. doi:10.1016/j.coisb.2020.07.009
93. Procopio A, Cesarelli G, Donisi L, Merola A, Amato F, Cosentino C. Combined mechanistic modeling and machine-learning approaches in systems biology – A systematic literature review. *Computer Methods and Programs in Biomedicine*. 2023;240:107681. doi:10.1016/j.cmpb.2023.107681
94. Sommer C, Gerlich DW. Machine learning in cell biology – teaching computers to recognize phenotypes. *Journal of Cell Science*. Published online January 1, 2013:jcs.123604. doi:10.1242/jcs.123604
95. Fernandez A, Garcia S, Luengo J, Bernado-Mansilla E, Herrera F. Genetics-Based Machine Learning for Rule Induction: State of the Art, Taxonomy, and Comparative Study. *IEEE Trans Evol Computat*. 2010;14(6):913-941. doi:10.1109/TEVC.2009.2039140
96. D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*. 2000;16(8):707-726. doi:10.1093/bioinformatics/16.8.707
97. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W. Learning from Co-expression Networks: Possibilities and Challenges. *Front Plant Sci*. 2016;7. doi:10.3389/fpls.2016.00444
98. Sun Y, Yang H, Guo J, Du J, Han S, Yang X. Identification of HTRA1, DPT and MXRA5 as potential biomarkers associated with osteoarthritis progression and immune infiltration. *BMC Musculoskelet Disord*. 2024;25(1):647. doi:10.1186/s12891-024-07758-7
99. Zhang Y, Zhang Y, Yu Z, et al. Insights into ACO genes across Rosaceae: evolution, expression, and regulatory networks in fruit development. *Genes Genom*. Published online August 14, 2024. doi:10.1007/s13258-024-01551-5
100. Ma Y, Lai J, Wan Q, et al. Exploring the common mechanisms and biomarker ST8SIA4 of atherosclerosis and ankylosing spondylitis through bioinformatics analysis and machine learning. *Front Cardiovasc Med*. 2024;11:1421071. doi:10.3389/fcvm.2024.1421071
101. Godini R, Pocock R, Fallahi H. Caenorhabditis elegans hub genes that respond to amyloid beta are homologs of genes involved in human Alzheimer's disease. Padmanabhan J, ed. *PLoS ONE*. 2019;14(7):e0219486. doi:10.1371/journal.pone.0219486
102. Munns J, Newling K, James SR, Gilbert L, Davis SJ, Chawla S. The rhythmic transcriptional landscape in *Caenorhabditis elegans* : daily, circadian and novel 16-hour cycling gene expression revealed by RNA-sequencing. Published online July 7, 2024. doi:10.1101/2024.07.06.602329
103. Yu S, Zheng C, Zhou F, et al. Genomic identification and functional analysis of essential genes in Caenorhabditis elegans. *BMC Genomics*. 2018;19(1):871. doi:10.1186/s12864-018-5251-3
104. Liu W, Li L, He Y, et al. Functional Annotation of Caenorhabditis elegans Genes by Analysis of Gene Co-Expression Networks. *Biomolecules*. 2018;8(3):70. doi:10.3390/biom8030070
105. Chitra R, Seenivasagam DV. Heart Disease Prediction System Using Supervised Learning Classifier. *Bonfring International Journal of Software Engineering and Soft Computing*. 2013;3(1).
106. Chai H. A novel logistic regression model combining semi-supervised learning and active learning for disease classification. *SCIenTIFIC REPOrTS*. Published online 2018.

107. Xie Y. Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational Oncology*. Published online 2021.
108. Zhang X, Jonassen I, Goksøyr A. Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data.
109. Bazazeh D, Shubair RM, Malik WQ. Biomarker Discovery and Validation for Parkinson's Disease: A Machine Learning Approach.
110. Abdelsamea MM, Zidan U, Senousy Z, Gaber MM, Rakha E, Ilyas M. A survey on artificial intelligence in histopathology image analysis.
111. Soueidan H, Nikolski M. Machine learning for metagenomics: methods and tools. Published online March 8, 2016. Accessed August 21, 2024. http://arxiv.org/abs/1510.06621
112. Roy G, Prifti E, Belda E, Zucker JD. Deep learning methods in metagenomics: a review. Published online 2023.
113. Droit A. Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation. *Frontiers in Microbiology*. 2022;13.
114. Zhao XM, Wang Y, Chen L, Aihara K. Gene function prediction using labeled and unlabeled data. *BMC Bioinformatics*. Published online 2008.
115. Pazos Obregón F. Gene function prediction in five model eukaryotes exclusively based on gene relative location through machine learning. *Scientific Reports*. Published online 2022.
116. Liu R, Mancuso CA, Yannakopoulos A, Johnson KA, Krishnan A. Supervised learning is an accurate method for network-based gene classification.
117. Blanco JL. Differential Gene Expression Analysis of RNA-seq Data Using Machine Learning for Cancer Research.
118. Jabeen A, Ahmad N, Raza K. Machine Learning-Based State-of-the-Art Methods for the Classification of RNA-Seq Data. *Machine Learning*.
119. Bostanci E, Kocak E, Unal M, Guzel MS, Acici K, Asuroglu T. Machine Learning Analysis of RNA-seq Data for Diagnostic and Prognostic Prediction of Colon Cancer. Published online 2023.
120. Bairakdar MD, Tewari A, Truttmann MC. A meta-analysis of RNA-Seq studies to identify novel genes that regulate aging. *Experimental Gerontology*. 2023;173:112107. doi:10.1016/j.exger.2023.112107
121. Campos TL, Korhonen PK, Sternberg PW, Gasser RB, Young ND. Predicting gene essentiality in Caenorhabditis elegans by feature engineering and machine-learning. *Computational and Structural Biotechnology Journal*. 2020;18:1093-1102. doi:10.1016/j.csbj.2020.05.008
122. Lu ZJ, Yip KY, Wang G, et al. Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res*. 2011;21(2):276-285. doi:10.1101/gr.110189.110
123. Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. *Genome Res*. 2017;27(12):2096-2107. doi:10.1101/gr.226233.117
124. Falk MJ, Rosenjack JR, Polyak E, et al. Subcomplex Iλ Specifically Controls Integrated Mitochondrial Functions in Caenorhabditis elegans. Goldman G, ed. *PLoS ONE*. 2009;4(8):e6607. doi:10.1371/journal.pone.0006607
125. Knowlton WM, Hubert T, Wu Z, Chisholm AD, Jin Y. A Select Subset of Electron Transport Chain Genes Associated with Optic Atrophy Link Mitochondria to Axon Regeneration in Caenorhabditis elegans. *Front Neurosci*. 2017;11:263. doi:10.3389/fnins.2017.00263
126. Ichimiya H, Huet RG, Hartman P, Amino H, Kita K, Ishii N. Complex II inactivation is lethal in the nematode Caenorhabditis elegans. *Mitochondrion*. 2002;2(3):191-198. doi:10.1016/S1567-7249(02)00069-7

127. Sternberg PW, Van Auken K, Wang Q, et al. WormBase 2024: status and transitioning to Alliance infrastructure. Wood V, ed. *GENETICS*. 2024;227(1):iyae050. doi:10.1093/genetics/iyae050
128. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*. 2004;32(90001):258D - 261. doi:10.1093/nar/gkh036
129. The UniProt Consortium, Bateman A, Martin MJ, et al. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*. 2023;51(D1):D523-D531. doi:10.1093/nar/gkac1052
130. Signes A, Fernandez-Vizarra E. Assembly of mammalian oxidative phosphorylation complexes I–V and supercomplexes. Garone C, Minczuk M, eds. *Essays in Biochemistry*. 2018;62(3):255-270. doi:10.1042/EBC20170098
131. Pezoulas VC, Zaridis DI, Mylona E, et al. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*. 2024;23:2892-2910. doi:10.1016/j.csbj.2024.07.005
132. Yang Z, Ding M, Huang T, et al. Does Negative Sampling Matter? a Review With Insights Into its Theory and Applications. *IEEE Trans Pattern Anal Mach Intell*. 2024;46(8):5692-5711. doi:10.1109/TPAMI.2024.3371473
133. Fagni T, Sebastiani F. On the Selection of Negative Examples for Hierarchical Text Categorization.
134. Xie N, Zhang L, Gao W, et al. NAD+ metabolism: pathophysiologic mechanisms and therapeutic potential. *Sig Transduct Target Ther*. 2020;5(1):227. doi:10.1038/s41392-020-00311-7
135. Wang C qun, Li X, Wang M qiang, et al. Protective effects of ETC complex III and cytochrome ***c*** against hydrogen peroxide-induced apoptosis in yeast. *Free Radical Research*. 2014;48(4):435-444. doi:10.3109/10715762.2014.885116
136. Kadenbach B. Complex IV – The regulatory center of mitochondrial oxidative phosphorylation. *Mitochondrion*. 2021;58:296-302. doi:10.1016/j.mito.2020.10.004
137. Boeck ME, Huynh C, Gevirtzman L, et al. The time-resolved transcriptome of *C. elegans*. *Genome Res*. 2016;26(10):1441-1450. doi:10.1101/gr.202663.115
138. Green RA, Khaliullin RN, Zhao Z, et al. Automated profiling of gene function during embryonic development. *Cell*. 2024;187(12):3141-3160.e23. doi:10.1016/j.cell.2024.04.012
139. Wadsworth WG, Riddle DL. Developmental regulation of energy metabolism in Caenorhabditis elegans. *Developmental Biology*. 1989;132(1):167-173. doi:10.1016/0012-1606(89)90214-5
140. Gómez-Orte E, Cornes E, Zheleva A, et al. Effect of the diet type and temperature on the *C. elegans* transcriptome. *Oncotarget*. 2018;9(11):9556-9571. doi:10.18632/oncotarget.23563
141. Mirza Z, Walhout AJM, Ambros V. A bacterial pathogen induces developmental slowing by high reactive oxygen species and mitochondrial dysfunction in Caenorhabditis elegans. *Cell Reports*. 2023;42(10):113189. doi:10.1016/j.celrep.2023.113189
142. Packer JS, Zhu Q, Huynh C, et al. A lineage-resolved molecular atlas of C. elegans embryogenesis at single cell resolution. *Science*. Published online 2020:365(6459). doi:doi:10.1126/science.aax1971
143. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8(1):14049. doi:10.1038/ncomms14049
144. Rezaie N, Reese F, Mortazavi A. PyWGCNA: a Python package for weighted gene co-expression network analysis. Mathelier A, ed. *Bioinformatics*. 2023;39(7):btad415. doi:10.1093/bioinformatics/btad415

145. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *IMeta*. 2023;2:e107. doi:https://doi.org/10.1002/imt2.107
146. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*. 2019;47(8):e47-e47. doi:10.1093/nar/gkz114
147. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25. doi:10.1186/gb-2010-11-3-r25
148. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*. 2012;131(4):281-285. doi:10.1007/s12064-012-0162-3
149. Rosati D, Palmieri M, Brunelli G, et al. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and Structural Biotechnology Journal*. 2024;23:1154-1168. doi:10.1016/j.csbj.2024.02.018
150. Gupta R, Dewan I, Bharti R, Bhattacharya A. Differential Expression Analysis for RNA-Seq Data. *ISRN Bioinformatics*. 2012;2012:1-8. doi:10.5402/2012/817508
151. Goksuluk D, Zararsiz G, Korkmaz S, et al. MLSeq: Machine learning interface for RNA-sequencing data. *Computer Methods and Programs in Biomedicine*. 2019;175:223-231. doi:10.1016/j.cmpb.2019.04.007
152. Wang B, Sun F, Luan Y. Comparison of the effectiveness of different normalization methods for metagenomic cross-study phenotype prediction under heterogeneity. *Sci Rep*. 2024;14(1):7024. doi:10.1038/s41598-024-57670-2
153. Zaitsev A, Chelushkin M, Dyikanov D, et al. Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes. *Cancer Cell*. 2022;40(8):879-894.e16. doi:10.1016/j.ccell.2022.07.006
154. Sammut SJ, Crispin-Ortuzar M, Chin SF, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022;601(7894):623-629. doi:10.1038/s41586-021-04278-5
155. Heil BJ, Crawford J, Greene CS. The effect of non-linear signal in classification problems using gene expression. Liu J, ed. *PLoS Comput Biol*. 2023;19(3):e1010984. doi:10.1371/journal.pcbi.1010984
156. Smith AM, Walsh JR, Long J, et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*. 2020;21(1):119. doi:10.1186/s12859-020-3427-8
157. Chen Y, Chen L, Lun ATL, Baldoni PL, Smyth GK. edgeR 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. Published online January 24, 2024. doi:10.1101/2024.01.21.576131
158. Klopfenstein DV, Zhang L, Pedersen BS, et al. GOATOOLS: A Python library for Gene Ontology analyses. *Sci Rep*. 2018;8(1):10872. doi:10.1038/s41598-018-28948-z
159. Priebe S, Menzel U, Zarse K, et al. Extension of Life Span by Impaired Glucose Metabolism in Caenorhabditis elegans Is Accompanied by Structural Rearrangements of the Transcriptomic Network. Vera J, ed. *PLoS ONE*. 2013;8(10):e77776. doi:10.1371/journal.pone.0077776
160. Hammarlund M, Hobert O, Miller DM, Sestan N. The CeNGEN Project: The Complete Gene Expression Map of an Entire Nervous System. *Neuron*. 2018;99(3):430-433. doi:10.1016/j.neuron.2018.07.042
161. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2
162. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19(6):679-682. doi:10.1038/s41592-022-01488-1

163. The PyMOL Molecular Graphics System.
164. Kayser EB, Morgan PG, Sedensky MM. GAS-1. *Anesthesiology*. 1999;90(2):545-554. doi:10.1097/00000542-199902000-00031
165. Lemire B. Mitochondrial genetics. *WormBook*. Published online 2005. doi:doi/10.1895/wormbook.1.25.1,
166. Yang W, Hekimi S. Two modes of mitochondrial dysfunction lead independently to lifespan extension in *Caenorhabditis elegans*. *Aging Cell*. 2010;9(3):433-447. doi:10.1111/j.1474-9726.2010.00571.x
167. Knapp-Wilson A, Pereira GC, Buzzard E, et al. Maintenance of complex I and its supercomplexes by NDUF-11 is essential for mitochondrial structure, function and health. *Journal of Cell Science*. 2021;134(13):jcs258399. doi:10.1242/jcs.258399
168. Falk MJ, Zhang Z, Rosenjack JR, et al. Metabolic pathway profiling of mitochondrial respiratory chain mutants in C. elegans. *Molecular Genetics and Metabolism*. 2008;93(4):388-397. doi:10.1016/j.ymgme.2007.11.007
169. *Comunicación Personal de Cecilia Martinez*.
170. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559. doi:10.1186/1471-2105-9-559
171. Torgerson WS. Multidimensional scaling: I. Theory and method. *Psychometrika*. 1952;17(4):401-419. doi:10.1007/BF02288916
172. Tzeng J, Lu HHS, Li WH. Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*. 2008;9(1):179. doi:10.1186/1471-2105-9-179
173. Fernández-Cárdenas LP, Villanueva-Chimal E, Salinas LS, José-Nuñez C, Tuena De Gómez Puyou M, Navarro RE. Caenorhabditis elegans ATPase inhibitor factor 1 (IF1) MAI-2 preserves the mitochondrial membrane potential (Δψm) and is important to induce germ cell apoptosis. Santos J, ed. *PLoS ONE*. 2017;12(8):e0181984. doi:10.1371/journal.pone.0181984
174. Woo JAA, Liu T, Trotter C, et al. Loss of function CHCHD10 mutations in cytoplasmic TDP-43 accumulation and synaptic integrity. *Nat Commun*. 2017;8(1):15558. doi:10.1038/ncomms15558
175. Kishore R, Arnaboldi V, Van Slyke CE, et al. Automated generation of gene summaries at the Alliance of Genome Resources. *Database*. 2020;2020:baaa037. doi:10.1093/database/baaa037
176. Jaskolowski M, Jomaa A, Gamerdinger M, et al. Molecular basis of the TRAP complex function in ER protein biogenesis. *Nat Struct Mol Biol*. 2023;30(6):770-777. doi:10.1038/s41594-023-00990-0
177. Van Kempen M, Kim SS, Tumescheit C, et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol*. Published online May 8, 2023. doi:10.1038/s41587-023-01773-0
178. Radivojac P, Clark WT, Oron TR, et al. A large-scale evaluation of computational protein function prediction. *Nat Methods*. 2013;10(3):221-227. doi:10.1038/nmeth.2340
179. Xue B, Rhee SY. Status of genome function annotation in model organisms and crops. *Plant Direct*. 2023;7(7):e499. doi:10.1002/pld3.499
180. Ceron-Noriega A, Almeida MV, Levin M, Butter F. Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteome-wide evolutionary analysis. *Genome Res*. 2023;33(1):112-128. doi:10.1101/gr.277070.122
181. Vishnoi A, Sharma R. A machine learning based analysis to probe the relationship between odorant structure and olfactory behaviour in *C. elegans*. Published online July 26, 2021. doi:10.1101/2021.07.26.453815
182. Campos TL, Korhonen PK, Gasser RB, Young ND. An Evaluation of Machine Learning Approaches for the Prediction of Essential Genes in Eukaryotes Using Protein Sequence-

Derived Features. *Comput Struct Biotechnol J*. 2019;17:785-796. doi:10.1016/j.csbj.2019.05.008
183. Zhong W, Sternberg PW. Genome-Wide Prediction of *C. elegans* Genetic Interactions. *Science*. 2006;311(5766):1481-1484. doi:10.1126/science.1123287
184. Hakim A, Mor Y, Toker IA, et al. WorMachine: machine learning-based phenotypic analysis tool for worms. *BMC Biol*. 2018;16(1):8. doi:10.1186/s12915-017-0477-0
185. Han Z, Zhang J, Su Y, et al. Identification of oxidative phosphorylation-related genes in moyamoya disease by combining bulk RNA-sequencing analysis and machine learning. *Front Genet*. 2024;15:1417329. doi:10.3389/fgene.2024.1417329
186. Cheng J, Liu HP, Lin WY, Tsai FJ. Machine learning compensates fold-change method and highlights oxidative phosphorylation in the brain transcriptome of Alzheimer's disease. *Sci Rep*. 2021;11(1):13704. doi:10.1038/s41598-021-93085-z
187. Newell AJ, Jima D, Reading B, Patisaul HB. Machine learning reveals common transcriptomic signatures across rat brain and placenta following developmental organophosphate ester exposure. *Toxicological Sciences*. 2023;195(1):103-122. doi:10.1093/toxsci/kfad062
188. Wei Y, Ma L, Peng Q, Lu L. Establishing an oxidative stress mitochondria-related prognostic model in hepatocellular carcinoma based on multi-omics characteristics and machine learning computational framework. *Discov Onc*. 2024;15(1):287. doi:10.1007/s12672-024-01147-1
189. Finney CA, Delerue F, Gold WA, Brown DA, Shvetcov A. Artificial intelligence-driven meta-analysis of brain gene expression identifies novel gene candidates and a role for mitochondria in Alzheimer's Disease. Published online February 4, 2022. doi:10.1101/2022.02.02.22270347
190. Padavannil A, Ayala-Hernandez MG, Castellanos-Silva EA, Letts JA. The Mysterious Multitude: Structural Perspective on the Accessory Subunits of Respiratory Complex I. *Front Mol Biosci*. 2022;8:798353. doi:10.3389/fmolb.2021.798353
191. Han J, Collins LJ. Reconstruction of Sugar Metabolic Pathways of *Giardia lamblia*. *International Journal of Proteomics*. 2012;2012:1-9. doi:10.1155/2012/980829
192. Neumann-Schaal M, Jahn D, Schmidt-Hohagen K. Metabolism the Difficile Way: The Key to the Success of the Pathogen Clostridioides difficile. *Front Microbiol*. 2019;10:219. doi:10.3389/fmicb.2019.00219
193. Saavedra E, Encalada R, Vázquez C, Olivos-García A, Michels PAM, Moreno-Sánchez R. Control and regulation of the pyrophosphate-dependent glucose metabolism in Entamoeba histolytica. *Molecular and Biochemical Parasitology*. 2019;229:75-87. doi:10.1016/j.molbiopara.2019.02.002
194. Valpuesta JM, Martín-Benito J, Gómez-Puertas P, Carrascosa JL, Willison KR. Structure and function of a protein folding machine: the eukaryotic cytosolic chaperonin CCT. *FEBS Letters*. 2002;529(1):11-16. doi:10.1016/S0014-5793(02)03180-0
195. Wittig I, Braun HP, Schägger H. Blue native PAGE. *Nat Protoc*. 2006;1(1):418-428. doi:10.1038/nprot.2006.62

8. Anexo

8.1. Tablas.

Tabla S1: Lista completa de genes relevados por participar en el proceso de fosforilación oxidativa de C. elegans y H. sapiens. Se indican si los genes codifican para proteínas cores, accesorias o del ensamblaje.

Complejo	Gen en Celegans	Funcion	Gen en humanos	Enfermedad en humanos
I	-	-	DMAC1	-
I	-	-	NDUA3	-
I	-	-	NDUB1	-
I	-	-	NDUC1	-
I	-	-	NDUV3	-
I	-	-	NDUFA4L2	-
I	-	-	NU6M	Leber hereditary optic neuropathy (LHON), Leber hereditary optic neuropathy with dystonia (LDYT), Mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes syndrome (MELAS) and Leigh syndrome (LS)
I	-	-	NU2M	Leber hereditary optic neuropathy, Alzheimer disease mitochondrial (AD-MT) and leigh's syndrome
I	-	-	NU3M	Mitochondrial complex I deficiency, nuclear type 1 (MC1DN1) and Leigh syndrome
I	-	-	NDUF8	Mitochondrial complex I deficiency, nuclear type 34
I	-	-	NDUA4	Mitochondrial complex IV deficiency, nuclear type 21
I	acdh-12	Ensamblado	ACADVL	Alzheimer's disease and very long chain acyl-CoA dehydrogenase deficiency.
I	B0035.15	Ensamblado	NDUFAF4	Mitochondrial complex I deficiency, type 15
I	B0334.5	Ensamblado	NDUFAF6	Mitochondrial complex I deficiency, type 17 and Fanconi renotubular syndrome 5
I	gas-1	Core	NDUS2	Mitochondrial complex I deficiency, nuclear type 6
I	K09E4.3	Ensamblado	NDUFAF5	Mitochondrial complex I deficiency, type 16.
I	lpd-5	Accesoria	NDUFS4	Mitochondrial complex I deficiency, nuclear type 1 (MC1DN1) and Leigh syndrome
I	M04B2.4	Ensamblado	FOXRD1	Mitochondrial complex I deficiency, type 19.

I	ndfl-4	Core	-	
I	ndua-1	Accesoria	NDUFA1	Mitochondrial complex I deficiency, nuclear type 12
I	ndua-12	Accesoria	NDUFA12	Mitochondrial complex I deficiency, nuclear type 23
I	ndua-13	Accesoria	NDUFA13	Hurthle cell thyroid carcinoma (HCTC) and Mitochondrial complex I deficiency, nuclear type 28
I	ndua-2	Ensamblado	NDUFA2	Mitochondrial complex I deficiency, nuclear type 13
I	ndua-5	Accesoria	NDUA5	-
I	ndua-7	Accesoria	NDUA7	-
I	ndua-8	Accesoria	NDUA8	Mitochondrial complex I deficiency, nuclear type 37
I	ndub-10	Accesoria	NDUFB10	Mitochondrial complex I deficiency, nuclear type 35
I	ndub-11	Accesoria	NDUFB11	Linear skin defects with multiple congenital anomalies 3
I	ndub-2	Accesoria	NDUFB2	Mitochondrial complex I deficiency, type 30
I	ndub-5	Accesoria	NDUFB5	-
I	ndub-6	Core	NDUFB6	-
I	ndub-7	Accesoria	NDUFB7	Mitochondrial complex I deficiency, nuclear type 39
I	ndub-8	Accesoria	NDUFB8	Mitochondrial complex I deficiency, nuclear type 32 (MC1DN32)
I	ndub-9	Accesoria	NDUFB9	Mitochondrial complex I deficiency, type 24
I	nduc-2	Accesoria	NDUFC2	Mitochondrial complex I deficiency, type 36.
I	nduf-11	Accesoria	NDUFA11	Mitochondrial complex I deficiency, nuclear type 14
I	nduf-5	Accesoria	NDUFS5	-
I	nduf-6	Accesoria	NDUFS6	Mitochondrial complex I deficiency, nuclear type 9 (MC1DN9)
I	nduf-7	Core	NDUFS7	Mitochondrial complex I deficiency, nuclear type 3 (MC1DN3)
I	nduf-9	Accesoria	NDUFA9	Mitochondrial complex I deficiency, nuclear type 26 (MC1DN26)
I	nduo-1	Core	-	
I	nduo-2	Core	-	
I	nduo-3	Core	-	
I	nduo-4	Core	-	
I	nduo-5	Core	-	
I	nduo-6	Core	-	
I	ndus-8	Core	NDUFS8	Mitochondrial complex I deficiency, nuclear type 2 (MC1DN2)

I	nduv-2	Core	NDUFV2	Mitochondrial complex I deficiency 7 and Parkinson's disease
I	nuaf-1	Ensamblado	NDUFAF1	Mitochondrial complex I deficiency, type 11.
I	nuaf-3	Ensamblado	NDUFAF3	Mitochondrial complex I deficiency, type 18.
I	nubp-1	Ensamblado	NUBP1	-
I	nuo-1	Core	NDUFV1	Mitochondrial complex I deficiency, nuclear type 4 (MC1DN4)
I	nuo-2	Core	NDUS3	Mitochondrial complex I deficiency, nuclear type 8 (MC1DN8)
I	nuo-3	Core	NDUA6	Mitochondrial complex I deficiency, nuclear type 33
I	nuo-4	Accesoria	NDUAA	Mitochondrial complex I deficiency, nuclear type 22
I	nuo-5	Core	NDUFS1	Mitochondrial complex I deficiency, nuclear type 5
I	nuo-6	Core	NDUB4	-
I	Y116A8C.30	Ensamblado	NDUFAF2	Mitochondrial complex I deficiency, type 10.
I	Y38F2AR.3	Ensamblado	TIMMDC1	Mitochondrial complex I deficiency, nuclear type 31
I	ZK1128.1	Ensamblado	NDUFAF7	May be a cause of susceptibility to pathologic myopia
II	mev-1	Core	SDHC	Carney-Stratakis syndrome; gastrointestinal stromal tumor; and paraganglioma 3
II	sdha-1	Core	SDHA	Mitochondrial complex II deficiency, nuclear type 1 (MC2DN1), Leigh syndrome (LS, dilated cardiomyopathy 1GG, Neurodegeneration with ataxia and late-onset optic atrophy (NDAXOA) and Paragangliomas 5 (PGL5)
II	sdhb-1	Core	SDHB	Pheochromocytoma (PCC), Paragangliomas 4 (PGL4), Paraganglioma and gastric stromal sarcoma (PGGSS), Carney-Stratakis syndrome, and Mitochondrial complex II deficiency, nuclear type 4 (MC2DN4).
II	sdhd-1	Core	SDHD	Carney-Stratakis syndrome; mitochondrial complex II deficiency, nuclear type 3, Paraganglioma and gastric stromal sarcoma (PGGSS), Pheochromocytoma (PCC) and paraganglioma 1.
II	Y57A10A.29	Ensamblado	SDHAF2	Paragangliomas 2 (PGL2)
III	-	-	QCR10	-
III	-	-	QCR9	-
III	bcs-1	Ensamblado	BCS1L	Bjornstad syndrome; GRACILE syndrome; and mitochondrial complex III deficiency nuclear type 1.
III	ctb-1	Core	CYB	Cardiomyopathy, infantile histiocytoid (CMIH), Leber hereditary optic neuropathy (LHON)
III	cyc-1	Core	CYC1	Mitochondrial complex III deficiency nuclear type 6.

III	ddl-1	Ensamblado	TTC19	Mitochondrial complex III deficiency nuclear type 2.
III	isp-1	Core	UQCRFS1	Mitochondrial complex III deficiency, type 10.
III	T02H6.11	Core	UQCRB	Mitochondrial complex III deficiency nuclear type 3.
III	T27E9.2	Core	UQCRH	Mitochondrial complex III deficiency, nuclear type 11 (MC3DN11)
IV	coa-1	Ensamblado	COA1	-
IV	coa-3	Ensamblado	COA3	Mitochondrial complex IV deficiency, nuclear type 14 (MC4DN14), cytochrome-c oxidase deficiency disease.
IV	coa-4	Ensamblado	COA4	-
IV	coa-5	Ensamblado	COA5	Mitochondrial complex IV deficiency, nuclear type 9 (MC4DN9), fatal infantile cardioencephalomyopathy due to cytochrome c oxidase deficiency 3.
IV	coa-6	Ensamblado	PRORP	Combined oxidative phosphorylation deficiency 54 (COXPD54)
IV	coa-7	Ensamblado	COA7	Spinocerebellar ataxia, autosomal recessive, with axonal neuropathy 3 (SCAN3)
IV	cox-10	Ensamblado	COX10	Mitochondrial complex IV deficiency, nuclear type 3 (MC4DN3), cytochrome-c oxidase deficiency disease.
IV	cox-11	Ensamblado	COX11	Mitochondrial complex IV deficiency, nuclear type 23 (MC4DN23)
IV	cox-14	Ensamblado	COX14	Mitochondrial complex IV deficiency, nuclear type 10 (MC4DN10)
IV	cox-15	Ensamblado	COX15	Mitochondrial complex IV deficiency, nuclear type 6 (MC4DN6), Leigh disease; fatal infantile cardioencephalomyopathy due to cytochrome c oxidase deficiency 2; and hypertrophic cardiomyopathy.
IV	cox-16	Ensamblado	COX16	Mitochondrial complex IV deficiency, nuclear type 22 (MC4DN22)
IV	cox-17	Ensamblado	COX17	-
IV	cox-18	Ensamblado	COX18	-
IV	cox-19	Ensamblado	COX19	-
IV	cox-4	Core	COX4I1	Mitochondrial complex IV deficiency, nuclear type 16 (MC4DN16), cytochrome-c oxidase deficiency disease.
IV	cox-5a	Core	COX5A	Mitochondrial complex IV deficiency, nuclear type 20 (MC4DN20), cytochrome-c oxidase deficiency disease.
IV	cox-5b	Core	COX5B	-
IV	cox-6a	Core	COX6A	Mitochondrial complex IV deficiency, nuclear type 18 (MC4DN18), Charcot-Marie-Tooth disease recessive intermediate D and cytochrome-c oxidase deficiency disease.
IV	cox-6b	Core	COX6B2	Cytochrome-c oxidase deficiency disease.

IV	cox-7c	Core	COX7C	-
IV	ctc-1	Core	-	-
IV	ctc-2	Core	-	-
IV	ctc-3	Core	-	-
IV	cys-2.1	Core	CYCS	Huntington's disease; carcinoma (multiple); and thrombocytopenia 4.
IV	sco-1	Ensamblado	SCO1	Mitochondrial complex IV deficiency, nuclear type 4 (MC4DN4), cytochrome-c oxidase deficiency disease; fatal infantile cardioencephalomyopathy due to cytochrome c oxidase deficiency 1; and myopia.
IV	stf-1	Ensamblado	SURF1	Mitochondrial complex IV deficiency, nuclear type 1 (MC4DN1), Charcot-Marie-Tooth disease type 4K; Leigh disease; and cytochrome-c oxidase deficiency disease.
IV	T20D3.6	Ensamblado	HIGD2A	-
IV	Y53F4B.14	Ensamblado	PET100	Mitochondrial complex IV deficiency, nuclear type 12 (MC4DN12), cytochrome-c oxidase deficiency disease.
V	atp-1	Core	ATP5F1A	Combined oxidative phosphorylation deficiency 22; mitochondrial complex V (ATP synthase) deficiency nuclear type 4; and vascular dementia.
V	atp-2	Core	ATP5F1B	Hypermetabolism due to uncoupled mitochondrial oxidative phosphorylation 2 (HUMOP2)
V	atp-3	Core	ATP5PO	Mitochondrial complex V deficiency, nuclear type 7 (MC5DN7)
V	atp-4	Core	ATP5PF	Implicated in essential hypertension.
V	atp-5	Core	ATP5PD	-
V	atp-6	Core	-	-
V	F58F12.1	Core	ATP5F1D	Alzheimer's disease and mitochondrial complex V (ATP synthase) deficiency, type 5.
V	R04F11.2	Core	ATP5ME	-
V	R53.4	Core	ATP5MF	-
V	Y116A8C.27	Core	ATPAF2	Mitochondrial complex V (ATP synthase) deficiency, nuclear type 1.
V	Y69A2AR.18	Core	ATP5F1C	-
V	Y82E9BR.3	Core	ATP5MC1 ATP5MC2 ATP5MC3	Leber hereditary optic neuropathy, autosomal recessive (LHONAR), clear cell renal cell carcinoma; early-onset dystonia and/or spastic paraplegia; and urinary bladder cancer.

Tabla S2: Primeros 20 genes hub del cluster A, ordenados por conectividad.

Gen	Conectividad	Descripción Wormbase
rla-1	93.92684	Predicted to enable protein kinase activator activity and ribonucleoprotein complex binding activity. Predicted to be a structural constituent of ribosome. Predicted to be involved in cytoplasmic translation. Predicted to be located in cytosolic ribosome. Predicted to be part of cytosolic large ribosomal subunit. Is an ortholog of human RPLP1 (ribosomal protein lateral stalk subunit P1).
rps-6	93.39139	Predicted to be a structural constituent of ribosome. Involved in determination of adult lifespan. Predicted to be located in cytoplasm; nucleolus; and ribosome. Predicted to be part of small-subunit processome. Is an ortholog of human RPS6 (ribosomal protein S6).
rps-18	92.95331	Predicted to enable RNA binding activity. Predicted to be a structural constituent of ribosome. Predicted to be involved in translation. Predicted to be located in cytosol. Predicted to be part of small ribosomal subunit. Is an ortholog of human RPS18 (ribosomal protein S18).
rpl-17	92.67957	Predicted to be a structural constituent of ribosome. Predicted to be involved in translation. Predicted to be located in cytoplasmic side of rough endoplasmic reticulum membrane. Predicted to be part of cytosolic large ribosomal subunit. Human ortholog(s) of this gene implicated in Diamond-Blackfan anemia 16. Is an ortholog of human RPL27 (ribosomal protein L27).
rps-4	92.64201	Predicted to enable RNA binding activity. Predicted to be a structural constituent of ribosome. Predicted to be involved in translation. Predicted to be located in ribosome. Predicted to be part of cytosolic small ribosomal subunit. Is an ortholog of human RPS4X (ribosomal protein S4 X-linked); RPS4Y1 (ribosomal protein S4 Y-linked 1); and RPS4Y2 (ribosomal protein S4 Y-linked 2).
rpl-16	92.54629	Predicted to enable mRNA binding activity. Predicted to be a structural constituent of ribosome. Predicted to be involved in negative regulation of translation. Predicted to be located in ribosome. Predicted to be part of cytosolic large ribosomal subunit. Is an ortholog of human RPL13A
rps-22	92.50641	Predicted to be a structural constituent of ribosome. Involved in determination of adult lifespan. Predicted to be located in ribosome. Predicted to be part of cytosolic small ribosomal subunit. Human ortholog(s) of this gene implicated in Diamond-Blackfan anemia 20. Is an ortholog of human RPS15A (ribosomal protein S15a)
rps-1	92.15855	Predicted to be a structural constituent of ribosome. Predicted to be involved in translation. Predicted to be located in cytosol. Predicted to be part of cytosolic small ribosomal subunit. Is an ortholog of human RPS3A (ribosomal protein S3A).
rpl-3	91.97086	Predicted to enable RNA binding activity. Predicted to be a structural constituent of ribosome. Predicted to be involved in translation. Predicted to be located in cytoplasm and ribosome. Predicted to be part of cytosolic large ribosomal subunit. Human ortholog(s) of this gene implicated in dilated cardiomyopathy 2D. Is an ortholog of human RPL3 (ribosomal protein L3).
rack-1	91.59098	Predicted to enable protein kinase C binding activity and ribosome binding activity. Involved in several processes
rps-23	91.31157	Predicted to be a structural constituent of ribosome. Predicted to be involved in cytoplasmic translation. Predicted to be located in ribosome. Predicted to be part of cytosolic small ribosomal subunit and polysomal ribosome. Human ortholog(s) of this gene implicated in brachycephaly
rpl-21	91.25531	Predicted to be a structural constituent of ribosome. Predicted to be involved in translation. Predicted to be located in ribosome. Predicted to be part of cytosolic

		large ribosomal subunit. Human ortholog(s) of this gene implicated in hypotrichosis 12. Is an ortholog of human RPL21 (ribosomal protein L21).
rpl-26	90.77968	Predicted to enable RNA binding activity. Predicted to be a structural constituent of ribosome. Predicted to be involved in cytoplasmic translation and ribosomal large subunit biogenesis. Predicted to be located in ribosome. Predicted to be part of cytosolic large ribosomal subunit. Human ortholog(s) of this gene implicated in Diamond-Blackfan anemia 11. Is an ortholog of human RPL26 (ribosomal protein L26) and RPL26L1 (ribosomal protein L26 like 1).
rps-15	90.71421	Predicted to be a structural constituent of ribosome. Involved in determination of adult lifespan. Predicted to be located in ribosome. Predicted to be part of cytosolic small ribosomal subunit. Is an ortholog of human RPS15 (ribosomal protein S15).
rpl-6	90.65245	Predicted to enable RNA binding activity. Predicted to be a structural constituent of ribosome. Involved in determination of adult lifespan. Predicted to be located in cytosol; ribosome; and rough endoplasmic reticulum. Predicted to be part of cytosolic large ribosomal subunit. Is an ortholog of human RPL6 (ribosomal protein L6).
rpl-1	90.63793	Predicted to enable RNA binding activity. Predicted to be involved in maturation of LSU-rRNA. Predicted to be located in ribosome. Predicted to be part of cytosolic large ribosomal subunit. Is an ortholog of human RPL10A (ribosomal protein L10a).
rpl-2	90.43437	Predicted to enable RNA binding activity. Predicted to be a structural constituent of ribosome. Predicted to be involved in cytoplasmic translation. Predicted to be located in cytoplasm and ribosome. Predicted to be part of cytosolic large ribosomal subunit. Is an ortholog of human RPL8 (ribosomal protein L8).
rps-8	90.42096	Predicted to be a structural constituent of ribosome. Predicted to be involved in maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA
rpl-7	90.36611	Predicted to enable RNA binding activity. Predicted to be a structural constituent of ribosome. Predicted to be involved in maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA
rpl-17	90.26663	Predicted to be a structural constituent of ribosome. Predicted to be involved in cytoplasmic translation. Predicted to be located in ribosome. Predicted to be part of cytosolic large ribosomal subunit. Is an ortholog of human RPL17 (ribosomal protein L17).

Tabla S3: Primeros 20 genes hub del cluster B, ordenados por conectividad.

Gen	Conectividad	Descripción de Wormbase
Y71F9AR.2	31.98236	Predicted to enable GTP binding activity and GTPase activity. Is an ortholog of human RASL12 (RAS like family 12).
Y37D8A.2	31.83775	Predicted to enable phospholipase activity. Predicted to be involved in phospholipid catabolic process. Predicted to be located in extracellular region. Is an ortholog of human PLBD2 (phospholipase B domain containing 2).
osg-1	30.94161	Predicted to enable guanyl-nucleotide exchange factor activity. Involved in response to oxidative stress. Expressed in body wall musculature and head neurons. Is an ortholog of human ARHGEF17 (Rho guanine nucleotide exchange factor 17).
F46C5.6	29.95304	Predicted to enable protein phosphatase regulator activity. Predicted to be located in cytosol. Predicted to be part of protein serine/threonine phosphatase complex. Is an ortholog of human PPP4R4 (protein phosphatase 4 regulatory subunit 4).
F13G3.10	29.85067	Predicted to be involved in negative regulation of apoptotic process and protein stabilization. Is an ortholog of human HYPK (huntingtin interacting protein K).
D2092.4	29.42435	Located in myofibril.
alh-8	28.93436	Predicted to enable malonate-semialdehyde dehydrogenase (acetylating) activity and methylmalonate-semialdehyde dehydrogenase (acylating
F21C10.7	28.7259	Predicted to be involved in axon guidance and cell adhesion. Predicted to be located in plasma membrane.
ikb-1	28.30103	Predicted to enable NF-kappaB binding activity. Involved in defense response to Gram-negative bacterium. Located in cytoplasm. Human ortholog(s) of this gene implicated in common variable immunodeficiency 10. Is an ortholog of human NFKB2 (nuclear factor kappa B subunit 2).
R03E9.2	28.08444	Predicted to enable Glc3Man9GlcNAc2 oligosaccharide glucosidase activity. Predicted to be involved in protein N-linked glycosylation. Predicted to be located in endoplasmic reticulum membrane.
upb-1	27.73645	Enables beta-ureidopropionase activity. Involved in beta-alanine biosynthetic process via 3-ureidopropionate; thymine catabolic process; and uracil catabolic process. Located in striated muscle dense body. Expressed in body wall musculature. Is an ortholog of human UPB1 (beta-ureidopropionase 1).
T13H5.1	27.71321	Predicted to enable protein tyrosine phosphatase activity. Predicted to be involved in synaptic membrane adhesion. Human ortholog(s) of this gene implicated in gastric adenocarcinoma. Is an ortholog of human PTPRA (protein tyrosine phosphatase receptor type A) and PTPRE (protein tyrosine phosphatase receptor type E).
unc-45	27.48387	Enables identical protein binding activity; protein folding chaperone; and ubiquitin protein ligase binding activity. Involved in several processes
unc-98	27.40812	Enables cytoskeletal protein binding activity. Involved in several processes
unc-68	26.32678	Predicted to enable calcium ion binding activity and ryanodine-sensitive calcium-release channel activity. Involved in locomotion; positive regulation of programmed cell death; and protein localization to organelle. Located in I band and sarcoplasmic reticulum. Expressed in body wall musculature; intestine; neurons; and non-striated muscle. Used to study congenital myopathy 1A and malignant hyperthermia. Human ortholog(s) of this gene implicated in several diseases
pat-9	26.24714	Predicted to enable chromatin insulator sequence binding activity. Involved in regulation of striated muscle cell differentiation. Located in nucleus. Expressed in body wall musculature and gonad.

hsp-12.1	26.24698	Predicted to enable unfolded protein binding activity. Predicted to be involved in protein refolding and response to heat. Located in striated muscle dense body. Expressed in body wall musculature. Human ortholog(s) of this gene implicated in several diseases
tag-241	26.05431	Enriched in several structures
tmd-2	25.98732	Predicted to enable tropomyosin binding activity. Predicted to be involved in actin filament organization and myofibril assembly. Located in myofibril. Human ortholog(s) of this gene implicated in dilated cardiomyopathy 2G. Is an ortholog of human LMOD2 (leiomodin 2) and TMOD4 (tropomodulin 4).

Tabla S4: Probabilidad de pertenecer a la clase 1 de los 103 genes de la lista consenso.

	Random Forest						Support Vector Machines						K-Nearest Neighbors						
Gen	Todos	w/o CI	w/o CII	w/o CIII	w/o CIV	w/o CV	Todos	w/o CI	w/o CII	w/o CIII	w/o CIV	w/o CV	Todos	w/o CI	w/o CII	w/o CIII	w/o CIV	w/o CV	Promedio probabilidad
rpl-39	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
rps-29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
rpl-30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
rps-28	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
prdx-2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
smo-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
fkb-2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
nap-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F29B9.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
his-72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
zig-7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
rbm-3.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
vha-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F23H11.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
rpl-37.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
arf-5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
kdp-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
C14B9.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
hpo-18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
snu-13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
cyn-5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
act-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.99
T28D9.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.75	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99
adk-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.73	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
lias-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.78	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.98
F45H10.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.73	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.98

mstr-1	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.76	0.99	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.98
vha-9	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.83	0.97	0.98	0.99	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.98
vha-10	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
dlc-1	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
ran-1	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
abcf-2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.59	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.98
arf-1	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
eif-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.59	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.98
ucr-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.55	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
lmp-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.54	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97
hel-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.57	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.97
C25A1.16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.59	0.99	1.00	1.00	0.94	1.00	1.00	1.00	1.00	1.00	1.00	0.97
ndub-3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.56	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00	1.00	0.97
ucr-11	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
F56H9.2	1.00	0.67	0.83	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
rab-1	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.74	0.92	0.94	0.97	0.94	1.00	1.00	1.00	1.00	1.00	1.00	0.97
nog-1	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.75	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.97
vha-1	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.53	0.97	0.98	0.99	0.92	1.00	1.00	1.00	1.00	1.00	1.00	0.97
cox-6C	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.45	0.99	0.99	1.00	0.92	1.00	1.00	1.00	1.00	1.00	1.00	0.96
MTCE.7	1.00	0.67	1.00	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96
snr-5	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.66	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.96
C48B6.10	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.48	0.97	0.98	0.99	0.89	1.00	1.00	1.00	1.00	1.00	1.00	0.96
gmpr-1	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.58	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96
K08D12.3	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.58	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96
trap-2	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.72	1.00	1.00	1.00	1.00	1.00	0.81	1.00	1.00	1.00	1.00	0.96
mai-2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.40	0.99	0.99	0.99	0.96	1.00	0.83	1.00	1.00	1.00	1.00	0.95
Y71F9AL.9	1.00	0.67	1.00	1.00	1.00	1.00	0.98	0.59	0.97	0.98	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.95
trap-4	1.00	0.67	1.00	1.00	1.00	1.00	0.99	0.52	0.99	0.99	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.95
prmt-1	0.80	0.67	1.00	1.00	1.00	1.00	1.00	0.63	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95
F49C12.11	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.41	0.98	0.99	0.99	0.90	1.00	0.82	1.00	1.00	1.00	1.00	0.95

asg-1	1.00	0.33	1.00	1.00	1.00	1.00	1.00	0.65	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94
nol-56	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.45	0.99	0.99	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	0.94
trap-3	1.00	0.67	1.00	0.75	1.00	1.00	0.99	0.57	0.98	0.99	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.94
trap-1	1.00	0.67	1.00	0.75	1.00	1.00	0.99	0.54	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.94
pdi-1	1.00	0.33	1.00	1.00	1.00	1.00	1.00	0.57	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94
asg-2	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.40	0.98	0.98	0.99	0.86	1.00	0.68	1.00	1.00	1.00	1.00	0.94
dkc-1	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.41	0.99	1.00	1.00	0.99	1.00	0.83	1.00	1.00	1.00	1.00	0.94
kin-3	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.39	0.91	0.93	0.94	0.93	1.00	0.83	1.00	1.00	1.00	1.00	0.94
acbp-1	1.00	1.00	0.83	1.00	1.00	1.00	0.99	0.44	0.99	0.99	0.99	0.91	1.00	0.66	1.00	1.00	1.00	1.00	0.93
cpi-2	1.00	1.00	0.83	1.00	1.00	1.00	0.96	0.44	0.92	0.93	0.95	0.78	1.00	1.00	1.00	1.00	1.00	1.00	0.93
ftn-2	1.00	1.00	1.00	1.00	0.67	1.00	0.99	0.44	0.97	0.98	0.99	0.88	1.00	0.82	1.00	1.00	1.00	1.00	0.93
plp-1	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.38	0.82	0.85	0.87	0.92	1.00	1.00	1.00	1.00	1.00	1.00	0.93
nola-3	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.48	0.98	0.99	0.99	0.79	1.00	1.00	1.00	1.00	1.00	0.50	0.93
vha-4	1.00	0.67	1.00	1.00	1.00	1.00	0.97	0.47	0.95	0.96	0.98	0.89	1.00	0.82	1.00	1.00	1.00	1.00	0.93
ola-1	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.28	0.93	0.95	0.97	0.94	1.00	0.65	1.00	1.00	1.00	1.00	0.93
mlc-4	1.00	0.33	0.83	1.00	1.00	1.00	0.99	0.53	0.99	0.99	0.99	0.96	1.00	1.00	1.00	1.00	1.00	1.00	0.92
vha-17	1.00	0.67	1.00	1.00	1.00	1.00	0.99	0.49	0.97	0.98	0.99	0.85	1.00	0.65	1.00	1.00	1.00	1.00	0.92
srlf-1	1.00	0.67	1.00	0.75	1.00	1.00	1.00	0.61	1.00	1.00	1.00	0.92	1.00	0.62	1.00	1.00	1.00	1.00	0.92
sec-61.B	1.00	0.67	1.00	0.75	1.00	1.00	0.96	0.47	0.92	0.94	0.96	0.88	1.00	1.00	1.00	1.00	1.00	1.00	0.92
snr-7	1.00	0.67	1.00	1.00	1.00	1.00	0.99	0.37	0.98	0.99	0.99	0.92	1.00	0.64	1.00	1.00	1.00	1.00	0.92
W08E12.7	0.80	0.67	1.00	0.75	1.00	1.00	0.99	0.37	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92
vha-15	1.00	0.67	0.83	1.00	0.67	1.00	0.98	0.54	0.95	0.97	0.98	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.92
fib-1	1.00	0.67	1.00	1.00	0.67	1.00	1.00	0.38	0.99	0.99	0.99	1.00	1.00	0.83	1.00	1.00	1.00	1.00	0.92
Y38E10A.24	1.00	0.67	1.00	1.00	1.00	1.00	1.00	0.46	0.99	1.00	1.00	0.78	1.00	0.63	1.00	1.00	1.00	1.00	0.92
Y22D7AL.10	1.00	0.67	1.00	1.00	1.00	1.00	0.98	0.39	0.95	0.96	0.98	0.91	1.00	0.65	1.00	1.00	1.00	1.00	0.92
daf-41	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.30	0.88	0.91	0.94	0.94	1.00	0.80	1.00	1.00	1.00	0.75	0.91
snr-1	1.00	0.33	1.00	1.00	1.00	1.00	0.99	0.41	0.98	0.99	0.99	0.95	1.00	0.81	1.00	1.00	1.00	1.00	0.91
T20B12.7	1.00	0.67	1.00	1.00	1.00	1.00	0.99	0.32	0.97	0.98	0.98	0.93	1.00	0.62	1.00	1.00	1.00	1.00	0.91
tmem-258	1.00	0.67	1.00	1.00	1.00	1.00	0.95	0.38	0.91	0.92	0.92	0.83	1.00	0.81	1.00	1.00	1.00	1.00	0.91
K12H4.5	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.31	0.89	0.92	0.93	0.71	1.00	0.66	1.00	1.00	1.00	1.00	0.91

fat-1	0.80	0.67	0.83	1.00	0.67	1.00	0.99	0.49	0.97	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.91
har-1	1.00	0.67	1.00	1.00	1.00	1.00	0.96	0.31	0.93	0.94	0.96	0.88	1.00	0.70	1.00	1.00	1.00	1.00	0.91
dad-1	1.00	0.33	1.00	1.00	1.00	1.00	0.98	0.37	0.96	0.97	0.97	0.93	1.00	0.81	1.00	1.00	1.00	1.00	0.91
lpd-9	1.00	0.67	1.00	1.00	1.00	1.00	0.97	0.30	0.94	0.96	0.96	0.78	1.00	0.70	1.00	1.00	1.00	1.00	0.90
ndab-2	1.00	0.67	1.00	1.00	1.00	1.00	0.97	0.28	0.94	0.95	0.96	0.83	1.00	0.67	1.00	1.00	1.00	1.00	0.90
nol-58	1.00	0.33	1.00	1.00	1.00	1.00	0.99	0.30	0.97	0.98	0.98	0.99	1.00	0.65	1.00	1.00	1.00	1.00	0.90
idh-1	1.00	0.67	1.00	0.75	1.00	1.00	0.91	0.38	0.85	0.88	0.90	0.99	1.00	0.82	1.00	1.00	1.00	1.00	0.90
eef-1B.2	0.80	0.67	1.00	1.00	1.00	1.00	0.96	0.22	0.92	0.94	0.94	0.98	1.00	0.66	1.00	1.00	1.00	1.00	0.89
mag-1	1.00	0.33	0.83	0.75	1.00	1.00	1.00	0.42	0.99	0.99	0.99	0.95	1.00	0.82	1.00	1.00	1.00	1.00	0.89
mel-32	0.80	0.67	1.00	0.75	1.00	1.00	0.98	0.29	0.96	0.97	0.97	0.99	1.00	0.62	1.00	1.00	1.00	1.00	0.89
sel-9	1.00	0.33	1.00	1.00	1.00	1.00	0.93	0.25	0.88	0.90	0.89	0.81	1.00	0.63	1.00	1.00	1.00	1.00	0.87
pdha-1	0.80	0.67	0.83	0.75	0.67	1.00	0.97	0.26	0.93	0.95	0.95	0.94	1.00	0.68	1.00	1.00	1.00	1.00	0.86
F52A8.1	1.00	0.33	1.00	1.00	1.00	1.00	0.84	0.28	0.76	0.79	0.79	0.74	1.00	0.64	1.00	1.00	1.00	1.00	0.84
cct-4	0.80	0.67	1.00	0.75	0.67	1.00	0.84	0.20	0.76	0.80	0.85	0.95	1.00	0.65	1.00	1.00	1.00	1.00	0.83
got-2.2	0.80	0.67	0.83	0.75	0.33	1.00	0.93	0.25	0.87	0.89	0.90	0.84	1.00	0.63	1.00	1.00	1.00	1.00	0.82
idha-1	0.80	0.33	0.83	0.75	0.67	1.00	0.89	0.24	0.83	0.86	0.87	0.82	1.00	0.79	1.00	1.00	1.00	1.00	0.82
cct-2	0.80	0.33	1.00	0.75	1.00	1.00	0.76	0.16	0.67	0.71	0.75	0.96	1.00	0.64	1.00	1.00	1.00	1.00	0.81

Tabla S5: Genes de la beta-oxidación de ácidos grasos en C. elegans según KEGG module M00087 y cantidad de veces predichos por los 18 modelos.

KEGG Orthology	Gen	Nombre	Cantidad de veces predicho
KO:K00022	B0272.3	putative 3-hydroxyacyl-CoA dehydrogenase	7/18
KO:K07509	hadb-1	putative 3-ketoacyl-CoA thiolase	4/18
KO:K07515	ech-1.1	Enoyl-CoA hydratase	0/18
KO:K00232	acox-1.5	putative acyl-coenzyme A oxidase acox-1.5	0/18
KO:K09479	acdh-12	Very long-chain specific acyl-CoA dehydrogenase, mitochondrial	2/18
KO:K00022	ech-8	Enoyl-CoA Hydratase	0/18
KO:K00022	ech-9	Enoyl-CoA Hydratase	0/18
KO:K00232	acox-1.1	Acyl-coenzyme A oxidase	0/18
KO:K00232	acox-1.2	Acyl-coenzyme A oxidase	0/18
KO:K00232	acox-1.3	Acyl-coenzyme A oxidase	0/18
KO:K00232	acox-1.4	Acyl-coenzyme A oxidase	0/18
KO:K07508	acaa-2	3-ketoacyl-CoA thiolase, mitochondrial	8/18
KO:K00022	F54C8.1	putative 3-hydroxyacyl-CoA dehydrogenase	0/18
KO:K00232	acox-3	Acyl-coenzyme A oxidase	0/18
KO:K00232	acox-1.6	Acyl-coenzyme A oxidase	0/18
KO:K00249	acdh-8	medium-chain acyl-CoA dehydrogenase	0/18
KO:K00022	hacd-1	3-hydroxyacyl-CoA dehydrogenase	0/18
KO:K07511	ech-6	putative enoyl-CoA hydratase, mitochondrial	13/18
KO:K07515	ech-1.2	Enoyl-CoA hydratase	9/18
KO:K00249	acdh-10	putative medium-chain specific acyl-CoA dehydrogenase 10, mitochondrial	8/18
KO:K00249	acdh-7	Medium-chain specific acyl-CoA dehydrogenase, mitochondrial	11/18
KO:K07511	ech-7	Enoyl-CoA hydratase, mitochondrial	0/18

Tabla S6: Genes de la glucólisis en C. elegans según KEGG module M00001 y cantidad de veces predichos por los 18 modelos.

KEGG Ontology	Gen	Nombre	Cantidad de veces predicho
KO:K08074	C50D2.7	putative ADP-dependent glucokinase	0/18
KO:K00850	pfk-1.2	ATP-dependent 6-phosphofructokinase 2	0/18
KO:K01623	aldo-2	Fructose-bisphosphate aldolase 2	17/18
KO:K00844	hxk-1	Phosphotransferase	2/18
KO:K00873	pyk-1	Pyruvate kinase	9/18
KO:K00134	gpd-4	Glyceraldehyde-3-phosphate dehydrogenase 4	6/18
KO:K15633	ipgm-1	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	4/18
KO:K00134	gpd-3	Glyceraldehyde-3-phosphate dehydrogenase 3	15/18
KO:K00134	gpd-2	Glyceraldehyde-3-phosphate dehydrogenase 2	17/18
KO:K00927	pgk-1	putative phosphoglycerate kinase	12/18
KO:K01623	aldo-1	Fructose-bisphosphate aldolase 1	11/18
KO:K00134	gpd-1	Glyceraldehyde-3-phosphate dehydrogenase 1	9/18
KO:K01689	enol-1	Enolase	15/18
KO:K01803	tpi-1	Triosephosphate isomerase	16/18
KO:K00850	pfk-1.1	ATP-dependent 6-phosphofructokinase 1	1/18
KO:K01810	gpi-1	Glucose-6-phosphate isomerase	6/18
KO:K00873	pyk-2	Pyruvate kinase	0/18

8.2. Figuras.

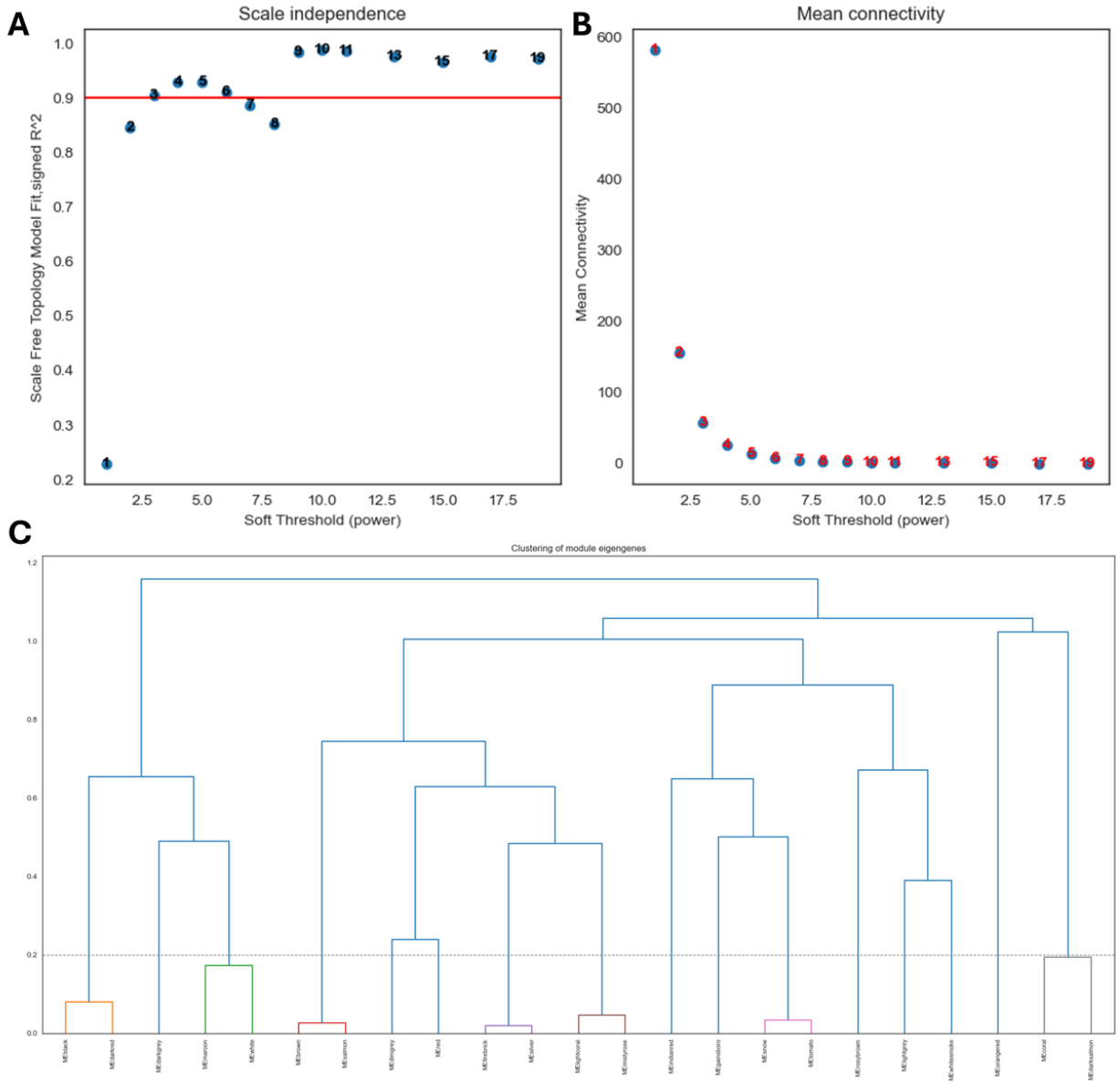


Figura S1: Resultado de la construcción de la red de co-expresión utilizando los genes que se expresan en por lo menos 123 células. (A) Análisis de la topología libre de escala (scale-free topology) de la red utilizando diferentes valores de potencia. El software elije el primer valor de potencia que genera un grafo con una topología libre de escala, obteniendo un R^2>0.9. (B) Conectividad media del grafo para las diferentes potencias evaluadas. (C) Dendograma obtenido del clustering jerárquico aplicado luego de seleccionar el valor de potencia para construir la red (3).

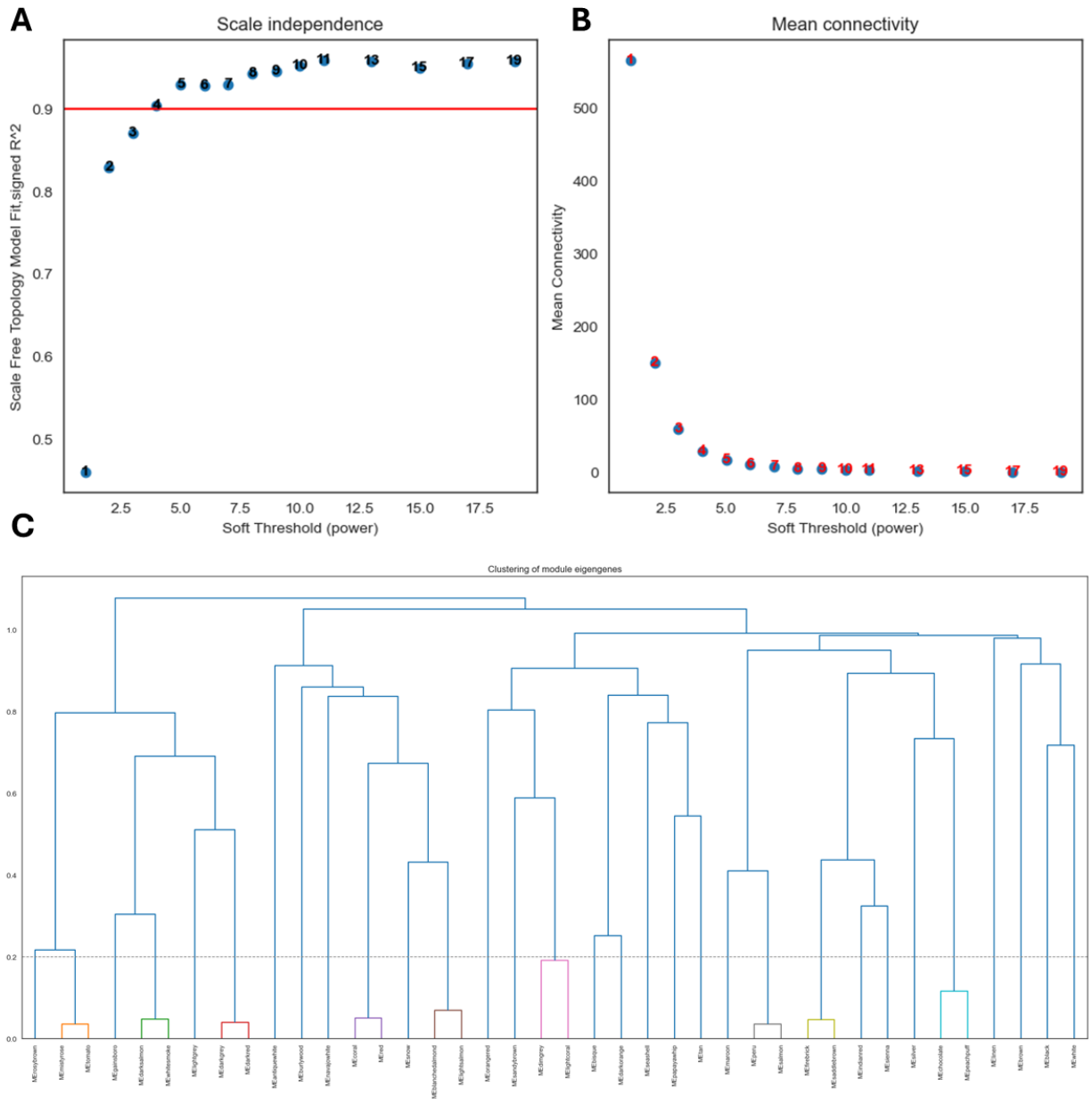


Figura S2: Resultado de la construcción de la red de co-expresión utilizando los genes que se expresan en por lo menos 23 células. (A) Análisis de la topología libre de escala (scale-free topology) de la red utilizando diferentes valores de potencia. El software elije el primer valor de potencia que genera un grafo con una topología libre de escala, obteniendo un R^2>0.9. (B) Conectividad media del grafo para las diferentes potencias evaluadas. (C) Dendograma obtenido del clustering jerárquico aplicado luego de seleccionar el valor de potencia para construir la red (4).

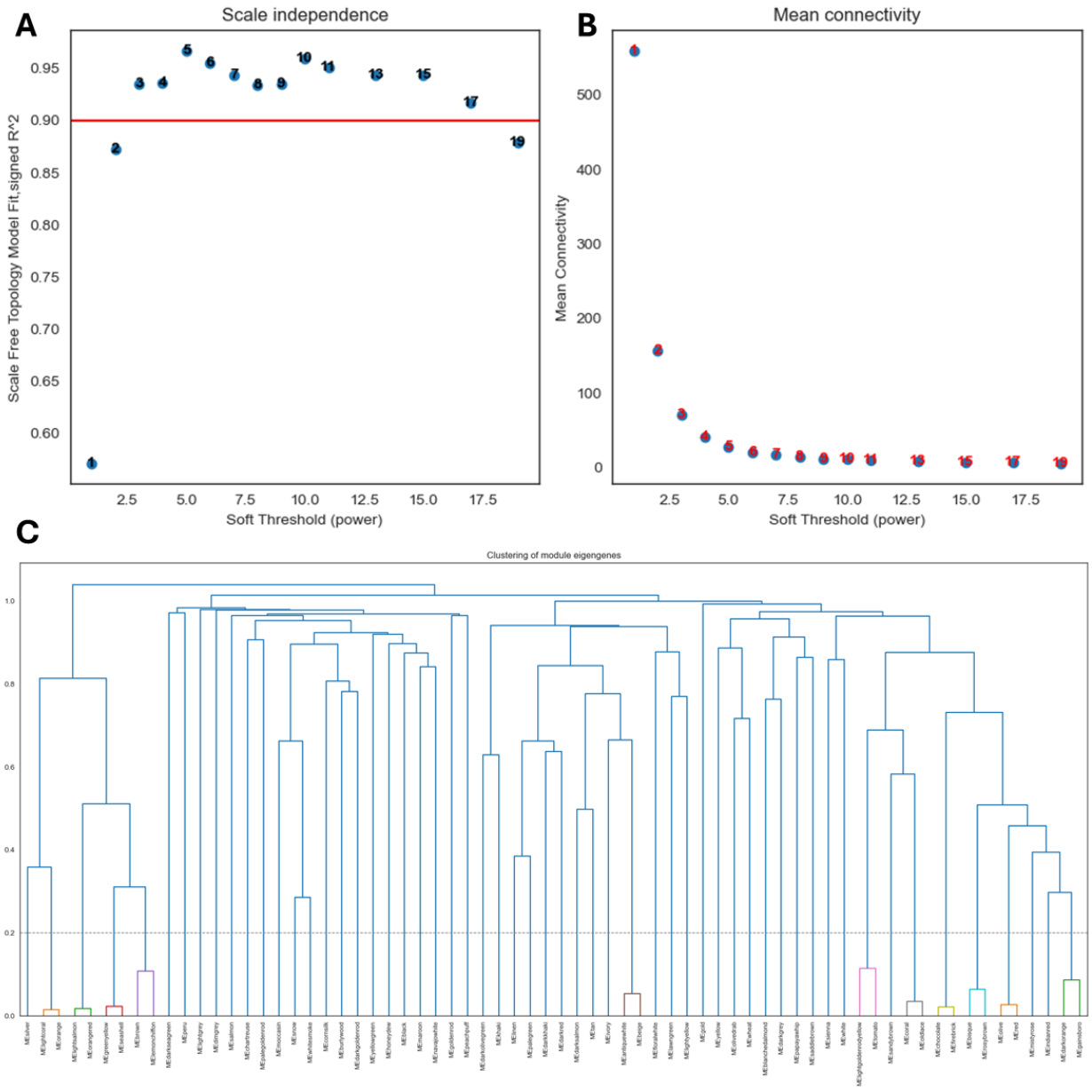


Figura S3: Resultado de la construcción de la red de co-expresión utilizando todos los genes. (A) Análisis de la topología libre de escala (scale-free topology) de la red utilizando diferentes valores de potencia. El software elije el primer valor de potencia que genera un grafo con una topología libre de escala, obteniendo un R^2>0.9. (B) Conectividad media del grafo para las diferentes potencias evaluadas. (C) Dendograma obtenido del clustering jerárquico aplicado luego de seleccionar el valor de potencia para construir la red (3).

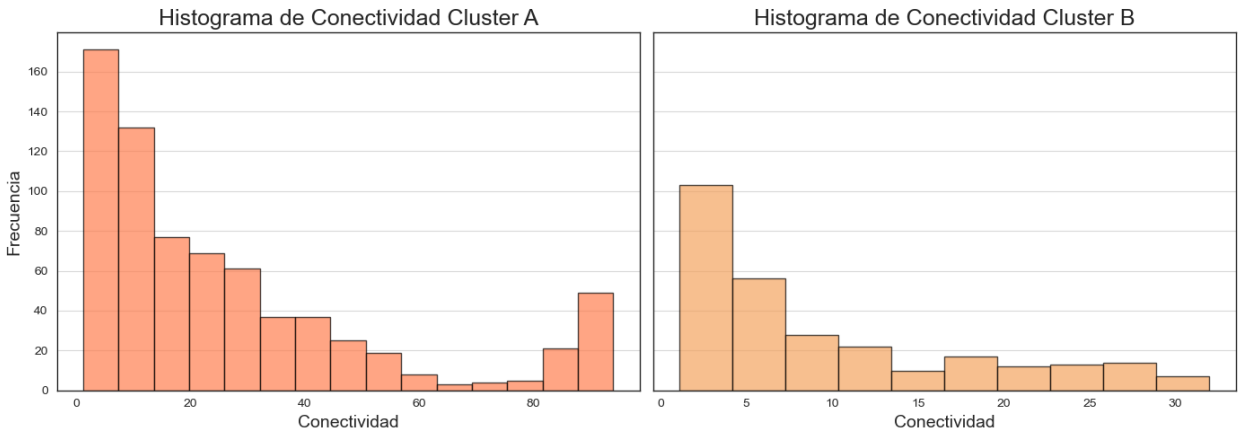
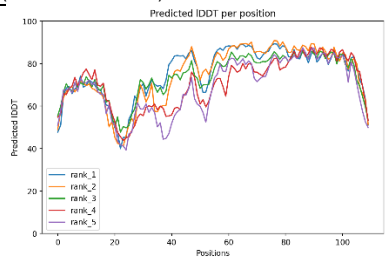
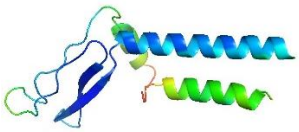
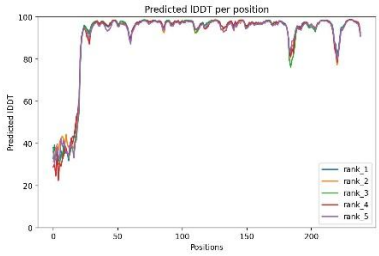
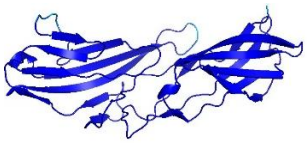
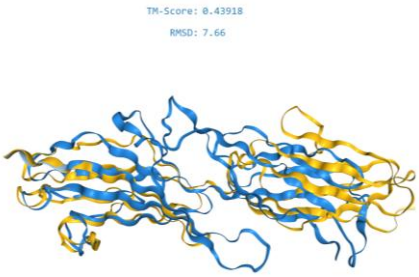
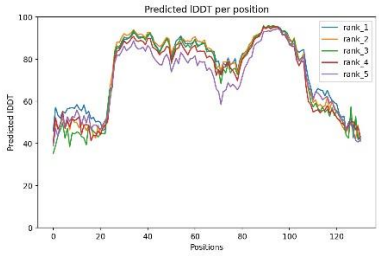
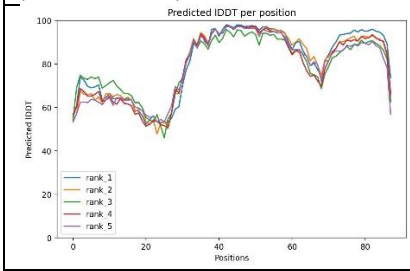
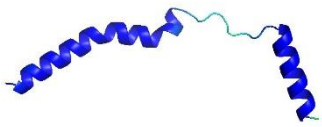
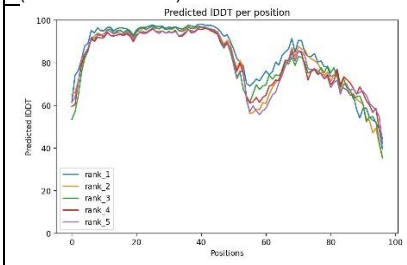
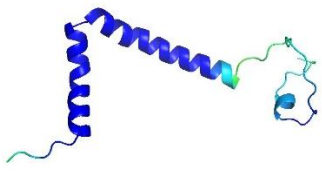
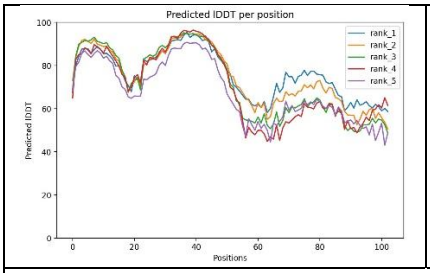
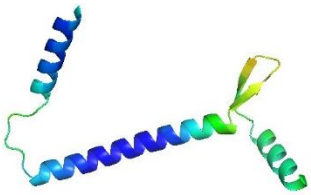
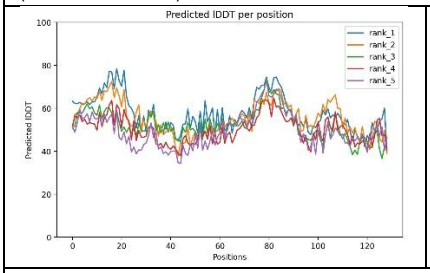
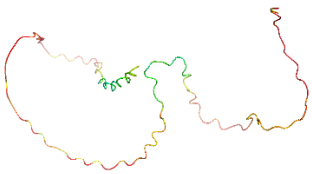
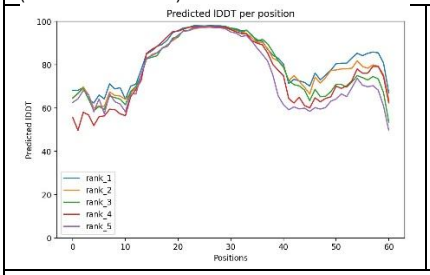
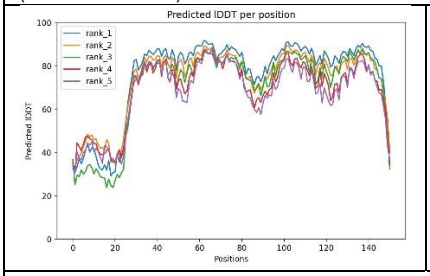
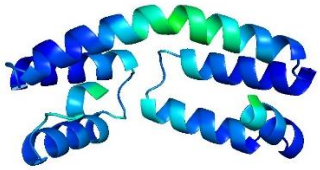
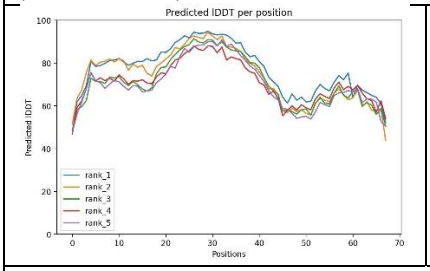


Figura S4: Histogramas de conectividad para el cluster A (izquierda) y B (derecha).

IDDT predicho por posición	Estructura predicha	Alineamiento de FoldSeek
F52A8.1 (WBGene00009915)		
		Sin hits significativos.
zig-7 (WBGene00006984)		CRYSTAL STRUCTURE OF NCAM2 IG1-2 Homo sapiens
		
lpd-9 (WBGene00003065_plddt)		
		Sin hits significativos.
F29B9.11 (WBGene00017925)		
		Sin hits significativos.
K12H4.5 (WBGene00019680)		
		Sin hits significativos.
C25A1.16 (WBGene00077526)		

		Sin hits significativos.
T28D9.1 (WBGene00020894)		
		No se evaluó por la baja calidad de la estructura.
C48B6.10 (WBGene00016746)		
		Sin hits significativos.
srlf-1 (WBGene00015514)		
		Sin hits significativos.
Y38E10A.24 (WBGene00012602)		
		Sin hits significativos.
rbm-3.2 (WBGene00011156)		CATH Superfamily 3.30.70.330 RRM (RNA recognition motif) domain Homo sapiens

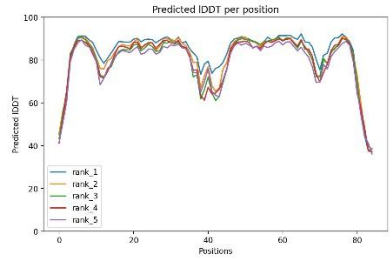
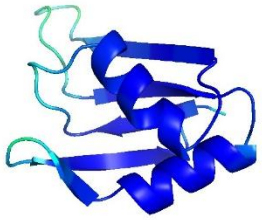
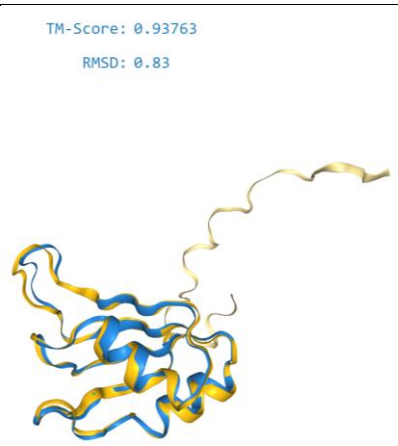
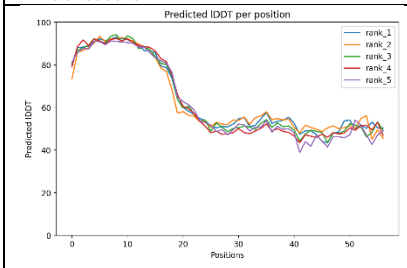
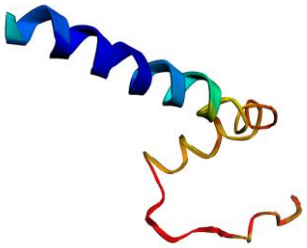
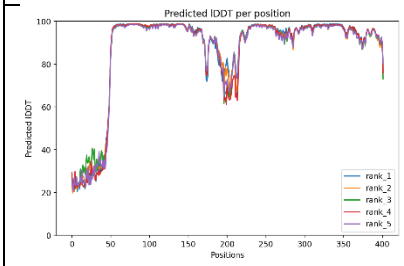
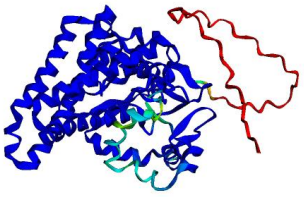
<table>
<tr>
<td> 
</td>
<td>  </td>
<td> 
</td>
</tr>
<tr>
<td colspan="3">F56H9.2
WBGene00010174</td>
</tr>
<tr>
<td> 
</td>
<td>  </td>
<td>Sin hits significativos.</td>
</tr>
<tr>
<td colspan="2">Fat-1
WBGene00001393</td>
<td>Fatty acid desaturase 3
Homo sapiens</td>
</tr>
<tr>
<td> 
</td>
<td>  </td>
<td> 
</td>
</tr>
</table>

Figura S5: Resultado de predicción estructural de los genes sin homología de secuencia con humanos de la lista consenso. En la primera columna se muestra el lDDT predicho por posición, seguido de la estructura predicha y de los matchs con foldseek en los casos que corresponde.