UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA QUÍMICA, BIOTECNOLOGÍA Y MATERIALES

META-ANÁLISIS TRANSCRIPTÓMICO DE HOSPEDEROS INFECTADOS CON WOLBACHIA PIPIENTIS, PARA SU APLICACIÓN EN CONTROL DE ENFERMEDADES ARBOVIRALES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN BIOTECNOLOGÍA

SEBASTIÁN ARIEL MEJÍAS OLEA

PROFESOR GUÍA:
DR. JOSÉ CRISTIAN SALGADO HERRERA

PROFESORA CO-GUÍA:
DRA. NATALIA JIMÉNEZ TAPIA

MIEMBROS DE LA COMISIÓN:
DR. CARLOS CONCA ROSENDE
DRA. ZIOMARA GERDTZEN HAKIM

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE: Ingeniero Civil en Biotecnología
POR: Sebastián Ariel Mejías Olea
FECHA: 2022
PROF. GUÍA: José Cristian Salgado Herrera

META-ANÁLISIS TRANSCRIPTÓMICO DE HOSPEDEROS INFECTADOS CON WOLBACHIA PIPIENTIS, PARA SU APLICACIÓN EN CONTROL DE ENFERMEDADES ARBOVIRALES

Por su capacidad para bloquear infecciones virales en mosquitos y diseminarse a través de sus poblaciones mediante manipulaciones reproductivas, la bacteria intracelular *Wolbachia pipientis* se está usando para el control de enfermedades arbovirales. A pesar de los alentadores resultados, preocupa que los fenotipos útiles de los sistemas *Wolbachia*-hospedero artificialmente establecidos (*transfectados*) se pierdan luego de un proceso de coevolución. El estudio de las infecciones nativas ayudaría a predecir rasgos de los sistemas transfectados en el futuro y así prever problemas que amenacen la sostenibilidad de las intervenciones. Distintos estudios RNA-Seq se han dedicado a develar los mecanismos subyacentes en dichas asociaciones, mediante la identificación y caracterización funcional de genes hospederos diferencialmente expresados (D.E.) por efecto de *Wolbachia*. Los resultados han sido muy diversos, despertando un interés por distinguir efectos comunes que, sin embargo, no se ha materializado en estudios dedicados. El objetivo de este trabajo fue identificar funciones comúnmente afectadas por infecciones *Wolbachia* nativas en hospederos Diptera. Para esto se obtuvieron 16 listas de genes D.E. por infección nativa en *Aedes fluviatilis*, *Drosophila melanogaster* y *Drosophila paulistorum*; se caracterizaron las listas mediante análisis de enriquecimiento de términos GO; y se identificaron términos enriquecidos en múltiples listas. Las funciones moleculares glicosil hidrolasa/transferasa, unión/hidrólisis de quitina, unión a iones de calcio, serina- y metalopeptidasa, y monooxigenasa (principalmente aquella ejercida por citocromos P450) fueron afectadas en los tres sistemas nativos. Evidencia de la afectación de las mismas funciones en sistemas transfectados fue hallada en bibliografía. Comparando el sentido de las regulaciones (sobre- o subregulación) se observaron indicios de adaptación de los efectos en actividad quitinasa y peptidasa, motivando su monitoreo en el tiempo. También, la consistencia entre funciones afectadas en sistemas nativos y transfectados derivó en hipótesis sobre aspectos universales de *Wolbachia* implicados. Por ejemplo, se postuló que la afectación de la quitina podría originarse en la necesidad de *Wolbachia* hacia su monómero constituyente, cuya síntesis podría no ser posible para la bacteria. Vinculando los efectos asociados a quitina con pérdidas de resistencia a la desecación en huevos de *A. aegypti*-wMel, se propuso que el estudio de esta interacción es prioritaria. Adicionalmente, se buscaron genes implicados en las alteraciones funcionales comunes, destacando *fmo-1* y *regucalcin*, que fueron D.E. en los tres sistemas nativos y en algunos sistemas transfectados, proveyendo blancos para la investigación futura. La obtención directa de algunas listas D.E. desde publicaciones (sin análisis RNA-Seq propio) y el uso de parámetros relajados en el análisis de enriquecimiento funcional se consideraron aspectos subóptimos pero necesarios de la metodología. En total, se concluyó el logro de los objetivos y se sostuvo la utilidad del trabajo en cuanto permitió proponer fundadamente direcciones de investigación que hasta ahora han sido desatendidas y podrían ser críticas para el éxito de las estrategias de control basadas en *Wolbachia*.

Agradecimientos

Entre tantas cosas que agradezco a mi padre y a mi madre, aquí les agradezco la educación que me brindaron.

A mis hermanos y hermana les agradezco comprender mi lejanía durante este tiempo de desarrollo de la memoria, con la promesa de que nos pondremos al día.

Agradezco a los amigos y amigas que hice durante mi paso por la universidad. Particularmente le agradezco a Yerko, Pancha, Leni, Josefa, Thais, Diego e Ignacia, por el privilegio de su amistad.

A las y los integrantes del equipo NEMBICA, profes Ziomara, Natalia, Cristian y Carlos, por el tiempo y atención que me concedieron, sé de buena fuente que no todas las comisiones son así de presentes.

En particular quiero agradecer a Natalia Jiménez, por su gran paciencia y sus sabios consejos metodológicos que, aunque no siempre logré implementar durante este trabajo, llevo muy presentes y sabré aplicar en el futuro.

Tabla de Contenido

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

1.1. Contexto

1.1.1. Enfermedades arbovirales como problemas de salud pública

Los arbovirus (*Arthropod Borne Virus*) son virus transmitidos por artrópodos, tales como mosquitos, pulgas y garrapatas. En este trabajo, el término se referirá específicamente a patógenos humanos transmitidos por mosquitos, incluyendo a los virus de ARN de la familia Flaviviridae como el virus del dengue (DENV), Zika (ZIKV), de la fiebre amarilla (YFV), de la encefalitis japonesa (JEV) y del Nilo occidental (WNV); y los de la familia Togaviridae como el virus del chikungunya (CHIKV) y el Sindbis (SINV) [1, 2].

Las enfermedades arbovirales representan un problema de salud pública a nivel mundial, la cantidad de personas en riesgo de contraerlas es del orden de los miles de millones y sigue creciendo debido a fenómenos como la urbanización no planificada, el cambio climático y la movilidad internacional [1, 3]. Por ejemplo, se estiman cientos de millones de infecciones anuales con el virus del dengue, cuyas consecuencias más graves incluyen hemorragias e insuficiencia orgánica, causando decenas de miles de muertes anuales [4]. El virus del Zika puede causar complicaciones graves durante el embarazo (incluyendo pérdida fetal), así como microcefalia y otros transtornos neurológicos congénitos que han sido reportados por 27 países de América desde 2015 [5, 6]. Igualmente grave es la fiebre amarilla, con un 15 % de los casos pasando a fase tóxica con fiebre alta y compromiso de la función renal, de los cuales la mitad termina en muerte. La fiebre amarilla es endémica en países de África y América, donde existe el riesgo de brotes de rápida expansión [7, 8].

El estado de las vacunas contra las enfermedades arbovirales no es óptimo [9]. Por ejemplo, a pesar de que muchos esfuerzos se han dirigido al desarrollo de vacunas contra el dengue, siguen existiendo dificultades críticas, derivadas de la necesidad de proteger contra los cuatro serotipos del virus (DENV 1-4) y de las dispares respuestas de individuos a la vacunación en función de su edad y seropositividad previa [10, 11]. En el caso del Zika, aunque hay vacunas en desarrollo, no hay claridad sobre el nivel de protección que éstas deben ofrecer para evitar los problemas congénitos derivados de la exposición al virus durante el embarazo, ni si las vacunas candidatas serán capaces de otorgar dicho nivel de protección [10].

La insuficiencia de las estrategias de inmunización es evidente para el caso de los virus que no tienen vacunas aprobadas o de aplicación general, pero incluso en los casos en que sí existen, la cobertura de la población en riesgo no está garantizada (por ejemplo, la cobertura de la vacunación contra la fiebre amarilla en los países americanos y africanos afectados no alcanza el 50 % [1]). El éxito de las campañas de vacunación puede verse mermado por factores como discontinuidad en la producción o disposición de las vacunas, dificultad de acceso a algunas regiones, reticencia de la población a la vacunación, interrupción de las actividades de inmunización por sobrecarga de los sistemas de salud, entre otros [12].

Un blanco alternativo en la lucha contra las enfermedades arbovirales son los vectores que las transmiten a los seres humanos, principalmente mosquitos del género *Aedes* [6, 9, 10]. Los enfoques tradicionales, que han consistido en el uso de insecticidas para la reducción de poblaciones de mosquitos, muestran importantes debilidades como pérdida de eficacia ante la selección de mosquitos resistentes, necesidad de ser aplicadas constantemente y riesgos ecológicos asociados con efectos inespecíficos [13]. Las estrategias de reemplazo de poblaciones apuntan a sustituir fenotipos nativos por otros con capacidad vectorial disminuida, habiéndose propuesto como una alternativa más sostenible que los métodos de reducción de poblaciones [9]. En el desarrollo de tales estrategias se ha encontrado un uso promisorio para la bacteria endosimbionte *Wolbachia pipientis*.

1.1.2. Control biológico de mosquitos mediante *Wolbachia pipientis*

Wolbachia pipientis (en adelante *Wolbachia*) es una bacteria gram-negativa, endosimbionte obligada y maternalmente heredada que pertenece al orden *Rickettsiales*, y cuyas numerosas cepas se estiman presentes de forma natural en más de la mitad de las especies de insectos del planeta [14, 15]. Una motivación para el estudio de *Wolbachia* es su capacidad para inducir una serie de alteraciones fenotípicas en sus hospederos, que pueden recuperarse en mosquitos artificialmente infectados (en adelante *transfectados*) y que actualmente están siendo explotadas en estrategias de control de vectores [14, 15, 16, 17, 18]. Específicamente, el uso de *Wolbachia* para el reemplazo de poblaciones implica la liberación de individuos transfectados con una cepa que disminuye su capacidad para transmitir virus. La diseminación de la cepa a través de la población objetivo y la reducción de su capacidad vectorial dependen críticamente de la manifestación de dos efectos: la *incompatibilidad citoplasmática* y el *bloqueo patogénico* [9, 18].

La incompatibilidad citoplasmática (IC) consiste en la muerte de embriones que surgen de la cruza de un macho infectado y una hembra no infectada o infectada con una cepa distinta [9, 19]. Tanto la IC como otras manipulaciones reproductivas favorecen la prevalencia de la bacteria en las poblaciones, mediante un aumento de la proporción de hembras infectadas, que son quienes las transmiten verticalmente [9, 19]. Por su parte, el bloqueo patogénico se da cuando un hospedero infectado con la bacteria resulta menos susceptible a determinadas infecciones secundarias, en comparación con los individuos sin *Wolbachia* [9, 14, 15]. En mosquitos, el bloqueo patogénico contra ciertos arbovirus puede traducirse en una densidad viral significativamente menor en la saliva, que es el medio que los transporta hacia el cuerpo humano durante una picadura [14].

Notablemente, se han podido realizar transfecciones estables de las cepas wMel y wAlbB (nativas de *Drosophila melanogaster* y *Aedes albopictus*, respectivamente) en el principal vector arboviral *Aedes aegypti*, una especie en que las infecciones *Wolbachia* nativas son muy escasas [20, 21]. Comprobándose la expresión de la IC y el bloqueo patogénico contra el DENV y otros virus en mosquitos *A. aegypti* transfectados, estos se han usado en campañas de reemplazo de poblaciones en América, Asia y Oceanía, algunas de las cuales ya han mostrado reducción significativa en la incidencia local de enfermedades arbovirales [14, 22, 23, 24, 25, 26, 27, 28]. A pesar de estos alentadores resultados, hay consenso en que queda mucho por comprender sobre las asociaciones *Wolbachia*-hospedero para poder optimizar las intervenciones y prever problemas que pudieran amenazar su sostenibilidad en el tiempo [15, 17, 23].

Una de las principales preocupaciones es que, tras un tiempo de co-evolución de los sistemas transfectados, los fenotipos útiles se pierdan o nuevos fenotipos indeseados emerjan (*p. ej.* potenciación de infecciones secundarias) [17, 28]. Críticamente, las asociaciones *Wolbachia*-hospedero no necesariamente transitan hacia el mutualismo, desafiando al paradigma clásico para los endosimbiontes verticalmente heredados [19]. Aunque se han descrito beneficios para el hospedero como el bloqueo patogénico o la suplementación de ciertos compuestos, la capacidad de *Wolbachia* para manipular la reproducción hospedera le significa cierta holgura para sostener efectos perjudiciales [19]. Así, predecir el futuro estado de las transfecciones es una tarea complicada. Se ha propuesto que el estudio de los sistemas *Wolbachia*-hospedero nativos puede ayudar en dicho propósito, en cuanto reflejarían posibles rasgos resultantes de los procesos de co-evolución [17].

Diversos estudios transcriptómicos se han dirigido a develar las bases mecanísticas de las infecciones *Wolbachia*, tanto nativas como artificiales. A grandes rasgos, dichos estudios comparan la composición de ARN de tejidos con y sin *Wolbachia*, detectando genes con niveles de transcripción sustancialmente distintos entre ambas condiciones (*i.e.*, genes *diferencialmente expresados*), cuya caracterización funcional permite inferir mecanismos de interacción [15, 17, 29, 30, 31]. Tales estudios han apuntado a una gran cantidad de genes y funciones; los resultados varían dependiendo de factores biológicos, como la combinación cepa/hospedero o el tejido analizado, y de factores técnicos, como las metodologías de análisis [31, 32]. En particular, la manera de inferir las funciones o procesos biológicos afectados por *Wolbachia* a partir de listas de genes diferencialmente expresados ha sido variable entre los estudios y a veces orientada a explorar hipótesis específicas [30, 31]. Ante tal incertidumbre se ha mostrado interés por identificar regularidades entre los transcriptomas de distintos sistemas *Wolbachia*-hospedero [17, 31, 32]. Hasta ahora, sin embargo, sólo Chung *et al.* (2020) se ha dedicado formalmente a este propósito, identificando patrones en la expresión génica de *Wolbachia* en distintos hospederos (principalmente nematodos), mediante un meta-análisis de conjuntos de datos transcriptómicos previamente publicados [32].

Recogiendo el interés por recapitular los efectos funcionales de *Wolbachia*, así como de caracterizar a las infecciones nativas en particular, en este trabajo se propuso identificar procesos biológicos o funciones moleculares comúnmente afectadas por infecciones *Wolbachia* nativas en hospederos Diptera. Para esto se obtuvieron listas de genes sobre- y subexpresados por efecto de infecciones nativas en distintos hospederos; se caracterizaron funcionalmente las listas mediante análisis de enriquecimiento de términos GO (*Gene Ontology*); y se selec-

cionaron los términos enriquecidos en listas de múltiples especies. La razón para reducir el análisis a hospederos del orden Diptera, es que este incluye a los mosquitos *Aedes* y al hospedero nativo de la cepa wMel, *D. melanogaster*; mientras que excluye a organismos infectados con *Wolbachia* pero con historias de vida radicalmente distintas, como insectos terrestres o nematodos.

Se espera que este trabajo aporte en la caracterización de los sistemas *Wolbachia*-Diptera nativos e informe sobre alteraciones funcionales candidatas a manifestarse en las transfecciones futuras. Se propone que la comparación entre tales efectos y los observados en las transfecciones actuales, permitiría prever posibles cambios adaptativos en el tiempo. Especialmente si los ámbitos funcionales en cuestión se han ligado a fenotipos de interés para el control de vectores, tales comparaciones podrían indicar blancos a monitorear y sugerir líneas de investigación tendientes a prevenir o reaccionar ante los cambios. También, se propone que la identificación de alteraciones funcionales comunes en sistemas nativos podría reflejar aspectos necesarios de la endosimbiosis (*p. ej.* requerimientos básicos para la vida intracelular de *Wolbachia*), aportando al conocimiento general de la bacteria.

Antes de formalizar los Objetivos y la Metodología, se presenta un Marco teórico con los conceptos básicos para comprender la lógica detrás de la aplicación de la transcriptómica para el estudio de los sistemas *Wolbachia*-hospedero. Específicamente, se presenta la técnica RNA-Seq que dio origen a las listas de genes diferencialmente expresados obtenidas, así como la forma en que de dichas listas se derivan funciones moleculares o procesos biológicos afectados por *Wolbachia*.

1.2. Marco teórico

1.2.1. RNA-Seq

En su acepción más general, la transcriptómica es el estudio de la identidad y abundancia de los transcritos de ARN expresados en una colección de células, incluyendo a los ARN mensajeros (ARNm) y a los ARN no codificantes (ARNnc) [33, 34]. Debido al rol de los ARNm como determinantes de la composición de las proteínas, y de los ARNnc en múltiples procesos celulares, la transcriptómica puede proveer información esencial sobre el estado de los sistemas biológicos [34, 35].

La secuenciación de ARN de alto rendimiento (RNA-Seq) permite cuantificar la expresión de transcritos en forma costo-efectiva [36, 37]. Críticamente, el RNA-Seq puede usarse para estudiar transcriptomas en amplitud genómica, sin la necesidad de conocer de antemano la identidad de los transcritos como en el caso de las técnicas basadas en hibridación de ácidos nucleicos (microarreglos) [36, 37]. En particular, el RNA-Seq ha sido la preferencia de los autores que han estudiado los sistemas *Wolbachia*-Diptera nativos [15, 17, 29, 30, 31].

En general, los protocolos RNA-Seq comprenden: (i) la extracción del ARN desde una muestra biológica, opcionalmente seguida del enriquecimiento de un tipo de ARN particular (usualmente ARNm), (ii) la síntesis de ADN complementario (ADNc) a partir del ARN extraído, (iii) la fragmentación del ADNc, la unión de adaptadores en los extremos de los fragmentos y, según la cantidad de material inicial, la amplificación de los fragmentos mediante

PCR, y (iv) la secuenciación de uno o ambos extremos de los fragmentos, originando lecturas *single-end* o *paired-end*, respectivamente [38]. En la Figura 1.1 se ilustran los pasos indicados.

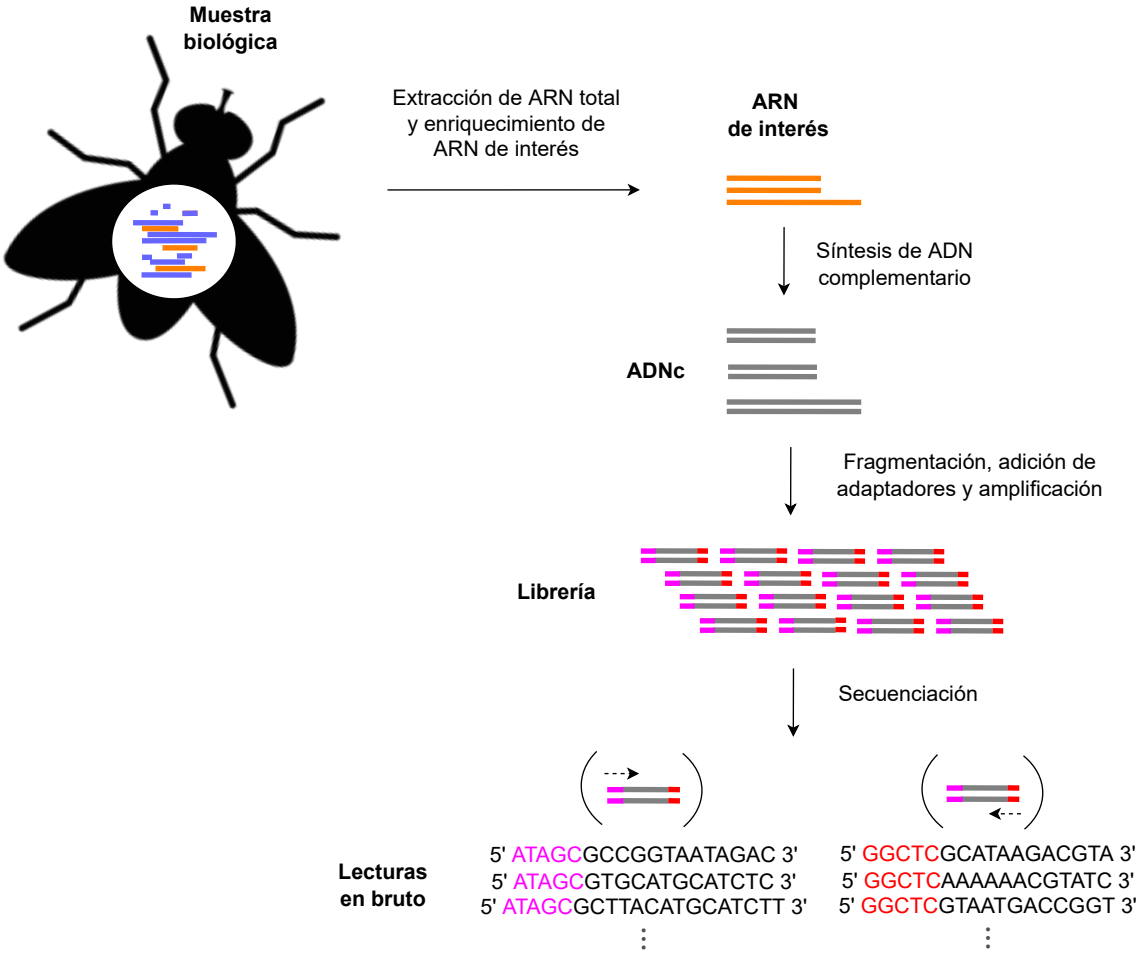

Figura 1.1: Etapas generales del RNA-Seq [38]. Se representa un caso en que se enriquece un tipo de ARN (naranja), se amplifica el ADNc y se generan lecturas de tipo *paired-end*. Elaboración propia.

1.2.2. Uso de RNA-Seq para el estudio de infecciones *Wolbachia*

Las técnicas RNA-Seq han sido frecuentemente utilizadas para la interrogación de los sistemas *Wolbachia*-hospedero [15, 17, 29, 30, 31]. Dicha aplicación involucra la identificación de genes o transcritos cuyos niveles de expresión promedio difieren significativamente entre tejidos en distintas condiciones relativas a la infección *Wolbachia*, denominados elementos diferencialmente expresados (en adelante, D.E.). En la Figura 1.2 se esquematiza la expresión diferencial de genes y transcritos entre dos condiciones.

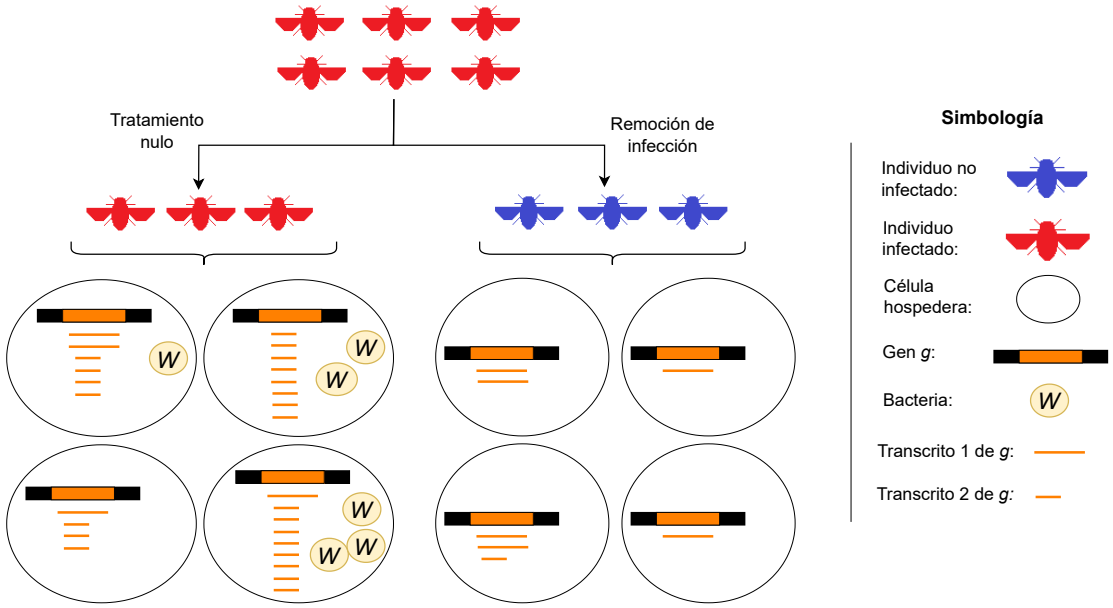

Figura 1.2: Representación de la expresión diferencial de genes y transcritos entre dos condiciones biológicas, originadas por un par de tratamientos sobre una población inicial de individuos infectados con *Wolbachia* (tratamiento nulo y remoción de la infección). Los transcritos totales expresados desde el gen *g* son, en promedio, más abundantes en células de individuos infectados: el gen *g* está *sobreexpresado* en dicha condición. Similarmente, el transcrito 2 se encuentra sobreexpresado en la condición infectada, mientras que el transcrito 1 no está afectado significativamente. Elaboración propia.

La caracterización de las listas de elementos D.E. permite identificar diferencias entre las condiciones biológicas a nivel funcional, cuya interpretación puede sugerir o desafiar hipótesis sobre los mecanismos de interacción *Wolbachia*-hospedero [15, 16, 17, 29, 30, 31, 39]. Por ejemplo, Lindsey *et al.* (2021) usó RNA-Seq para estudiar los efectos de infecciones wMel y virus Sindbis (SINV) sobre la expresión de genes hospederos en hembras *D. melanogaster* completas [15]. La comparación entre los conjuntos de genes con expresión sensible a *Wolbachia*, SINV y/o la interacción de ambos, sugirió que un posible ámbito funcional implicado en el bloqueo patogénico podría ser el metabolismo de nucleótidos [15]. Posterior evidencia a favor de esta hipótesis se obtuvo al hallar un efecto interactivo entre la infección *Wolbachia* y el silenciamiento de un gen involucrado en la síntesis de nucleótidos purínicos (*prat2*) sobre la densidad viral en individuos inyectados con SINV [15].

Por su parte, He *et al.* (2019) usó RNA-Seq para comparar transcriptomas de ovarios de *D. melanogaster* con y sin infección wMel. La expresión de genes D.E. con funciones reproductivas fue posteriormente estudiada en testículos de la misma especie con y sin wMel, identificándose un grupo de genes con cambios de expresión en sentido opuesto en ambos tejidos, interpretándose como evidencia a favor de un mecanismo particular para la incompatibilidad citoplasmática, denominado titración-restitución [30].

Si bien los detalles metodológicos de estas aplicaciones varían según el tipo de ARN estudiado, la complejidad de las muestras biológicas, los factores cuyo efecto sobre la expresión génica se estudia, y la disponibilidad de genomas de referencia con anotaciones de elementos genómicos, se pueden delinear etapas generales representadas en la Figura 1.3 [36, 40].

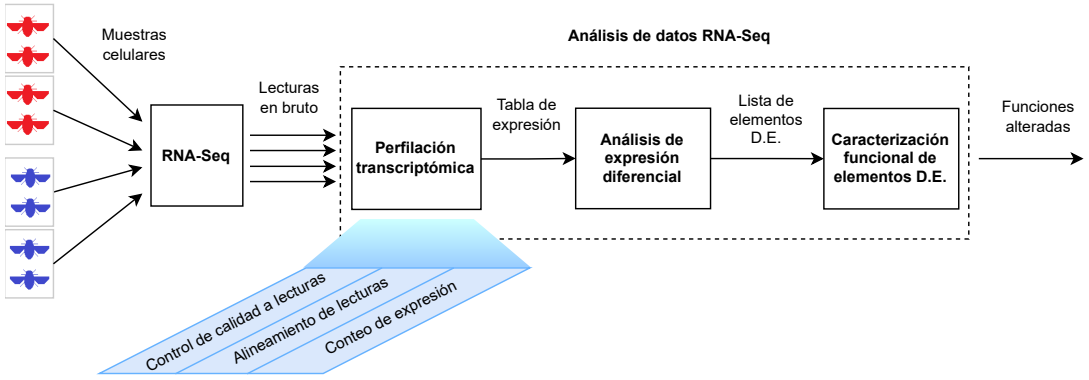

Figura 1.3: Etapas generales de la aplicación del RNA-Seq para el estudio de infecciones *Wolbachia*. Inicialmente, muestras celulares en distintas condiciones (típicamente mosquitos completos con y sin *Wolbachia*, representados en la imagen con colores distintos) son sometidas a RNA-Seq, generando archivos con lecturas en bruto para cada muestra. El análisis de datos RNA-Seq comienza con la perfilación transcriptómica de cada muestra, que incluye el control de calidad de las lecturas, el alineamiento de lecturas a un genoma o transcriptoma de referencia y el conteo de lecturas alineadas, agregadas a nivel de transcritos o genes. Los resultados se someten a un análisis estadístico para identificar elementos D.E. en virtud de los factores considerados (en este ejemplo, estado de infección *Wolbachia*). La caracterización de las listas de elementos D.E. permite inferir diferencias a nivel funcional, cuya interpretación puede ayudar a generar o poner a prueba hipótesis sobre los mecanismos de interacción *Wolbachia*-hospedero. Elaboración propia.

El resto del Marco teórico se dedica a precisar la etapa de Caracterización funcional de elementos D.E., que está en el centro de la metodología aquí seguida, pues todos los resultados fueron obtenidos realizando una caracterización funcional propia. Las etapas de Perfilación transcriptómica y Análisis de expresión diferencial se describen en el Anexo A, siendo útil para comprender y justificar la metodología aquí seguida para obtener listas D.E. mediante análisis RNA-Seq *de novo*, que sólo se hizo sobre algunos de los conjuntos de datos. En particular, la sección del Anexo A referida al Análisis de expresión diferencial contiene una definición matemática del concepto de expresión diferencial.

1.2.3. Caracterización funcional de elementos D.E.

1.2.3.1. Análisis de enriquecimiento de términos funcionales

Para extraer conocimiento biológico a partir de elementos D.E., estos se asocian a las funciones de sus productos génicos. El recurso *Gene Ontology* (GO) es el más utilizado para dicho fin [41]. GO se compone de dos elementos principales, a saber: la ontología y la anotación GO [41].

La ontología está estructurada en forma de grafo, donde cada nodo corresponde a un concepto o *término GO* y las aristas representan relaciones entre los términos [41]. Relaciones comúnmente usadas entre dos términos incluyen *es un* (en inglés *is a*), *es parte de* (en inglés *is part of*) y *regula* (en inglés *regulates*) [42]. Para facilitar el seguimiento de los términos GO específicos aludidos en este trabajo, se usarán los nombres originales de estos en inglés. Así, los siguientes son ejemplos de las relaciones previamente mencionadas: *glucose transport*

is a monosaccharide transport, *mitochondrial membrane is part of mitochondrial envelope*, y *latency-replication decision regulates release from viral latency* [42].

La anotación GO registra, para distintas especies, las asociaciones entre sus genes y los términos GO que describen a sus productos génicos, asignando a cada asociación un código que determina el tipo de evidencia que la respalda [41]. Así, dado un conjunto de genes U, un término GO arbitrario se puede asociar unívocamente con un conjunto $T \subseteq U$ de todos los genes en U que se encuentran anotados con dicho término. La asociación de un término funcional con dicho conjunto de genes, forma la base para un tipo de análisis que tiene como objetivo identificar funciones inusualmente representadas en listas derivadas de experimentos ómicos, denominado *Análisis de enriquecimiento de términos funcionales* [43]. Importantemente, este análisis constituye un procedimiento objetivo para asignar relevancia a las funciones invocadas por listas de elementos D.E., en oposición a la asignación realizada subjetivamente según el interés de investigadores por determinados ámbitos funcionales.

Formalmente, sea Q el conjunto de genes de interés según un experimento ómico (*p. ej.*, genes D.E.), U el conjunto de todos los genes monitoreados en el experimento (en adelante referido como *universo*), y $T \subseteq U$ el término cuyo enriquecimiento se quiere testear en Q. Determinar el enriquecimiento de T en Q implica rechazar la hipótesis nula (H_0) de que la pertenencia de un gen a T es independiente de su pertenencia a Q. Dicho de otra forma, si p_1 es la probabilidad de que un gen tomado al azar desde Q pertenezca a T y p_2 es la probabilidad de que un gen tomado al azar desde Q^c pertenezca a T, entonces $H_0 : p_1 = p_2$ [44]. La hipótesis H_0 se puede testear mediante una prueba estadística en la que los tamaños de U, Q y T (denotados n_U, n_Q y n_T, respectivamente) son valores fijos, mientras que el tamaño de la intersección $T \cap Q$ (denotado $n_{T,Q}$) se considera una realización de una variable aleatoria $N_{T,Q}$. Cuando sólo interesa identificar una sobrerrepresentación de T en Q y no una subrrepresentación, la región de rechazo de H_0 (el conjunto de los $n_{T,Q}$ que conducen al rechazo de H_0) se define de tal forma que una observación en dicha región entregue evidencia a favor de la hipótesis alternativa $H_a : p_1 > p_2$ [44]. Al igual que en las pruebas de expresión diferencial, la asignación de significancia de múltiples pruebas de enriquecimiento se basa en un cálculo inicial de *p-values* y un posterior ajuste para controlar el *False Discovery Rate* (una descripción del FDR y la motivación de su uso en testeo múltiple se puede hallar en el Anexo A.2) [44].

Para la caracterización funcional de listas de genes D.E., una decisión temprana es si el análisis de enriquecimiento se realizará sobre el conjunto completo de genes o separadamente para genes sobre- y subexpresados. En favor de la segunda opción se ha argumentado que, dado que los genes asociados con un mismo término funcional tienen cierta tendencia a ser corregulados, se puede perder poder estadístico al realizar las pruebas sobre el conjunto de todos los genes D.E. [45]. Una ventaja adicional aquí prevista para la separación de los genes D.E. en listas sobre- y subexpresadas es que facilitará la interpretación de los resultados, permitiendo detectar de forma gráfica si los términos enriquecidos representan funciones activadas o desactivadas.

Actualmente existen varias herramientas para realizar análisis de enriquecimiento de términos funcionales sobre listas de genes, entre ellas, el servidor público g:Profiler [46]. Entre las virtudes de g:Profiler se cuenta una actualización periódica para capturar cambios en

las bases de datos que utiliza (en particular, de *Gene Ontology*), una extensiva documentación y la provisión centralizada de funciones auxiliares para la manipulación de las listas de elementos genómicos, como mapeo de identificadores entre distintos espacios de nombres (*namespaces*) y búsqueda de ortólogos entre distintas especies [43, 46]. Al igual que varias otras herramientas para el análisis de enriquecimiento de términos funcionales, g:Profiler testea las hipótesis mediante una prueba hipergeométrica. En esta se asume que, si H_0 es cierta, entonces $N_{T,Q}$ tiene una distribución de probabilidad hipergeométrica, lo cual se denota como $N_{T,Q} \sim HG(n_U, n_Q, n_T)$ e implica que [44]:

$$Pr(N_{T,Q} = n_{T,Q}) = \frac{\binom{n_T}{n_{T,Q}}\binom{n_U - n_T}{n_Q - n_{T,Q}}}{\binom{n_U}{n_Q}}$$

Cuando existe un criterio con sentido biológico para ordenar la lista que se quiere someter a análisis de enriquecimiento (como log_2FC en el caso de listas de elementos D.E.), esta puede procesarse en g:Profiler como una lista ordenada [46]. En tal caso, la prueba hipergeométrica se realiza sobre cada posible prefijo de la lista (partiendo desde el primer elemento y agregando los posteriores uno a uno), y el menor *adjusted p-value* obtenido es reportado junto con el largo del prefijo para el cual se obtuvo [46].

La Figura 1.4 muestra la interfaz del módulo de g:Profiler que realiza el análisis de enriquecimiento (g:GOSt), y servirá como referencia para comentar las principales decisiones que deben tomarse para la ejecución de dicho análisis.

(El siguiente espacio se ha dejado en blanco intencionalmente)

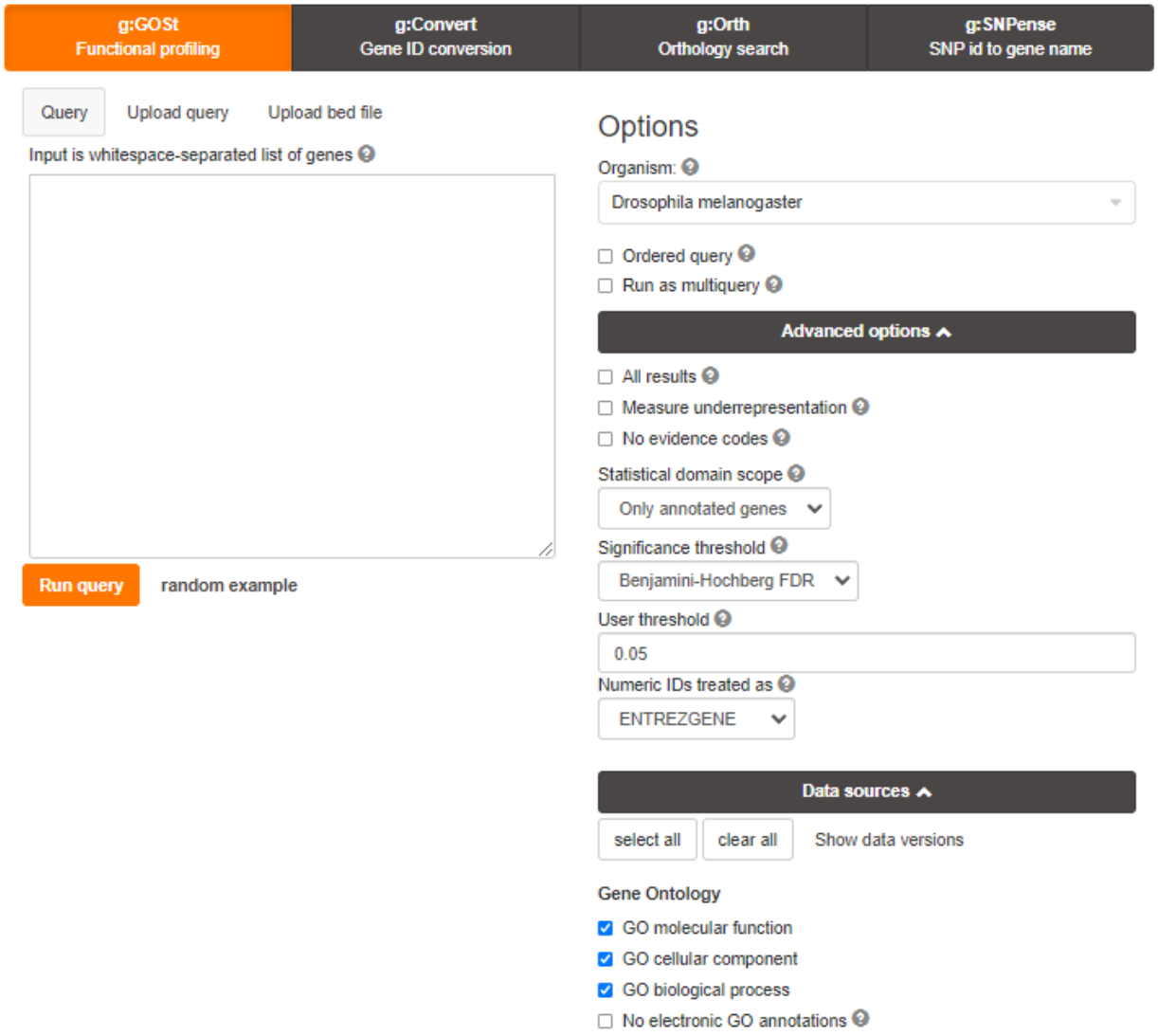

Figura 1.4: Parámetros a definir en el análisis de enriquecimiento funcional mediante el módulo g:GOSt de g:Profiler (https://biit.cs.ut.ee/gprofiler/gost).

En la interfaz presentada en la Figura 1.4, el campo *Ordered query* especifica si se prefiere la opción antes descrita para tratar la consulta como una lista ordenada. Los campos *Organism* y *Statistical domain scope* se utilizan para especificar el conjunto U. Usualmente, se considera apropiado definir U como el conjunto de todos los genes de un organismo de referencia que tienen al menos una anotación en la respectiva ontología, en cuyo caso *Statistical domain scope* se fija en *Only annotated genes* [47]. Otras opciones se utilizan cuando la lista de interés tiene muchos genes sin anotación (en cuyo caso se selecciona *All known genes*) o cuando el experimento que dio lugar a la lista de interés monitoreó un grupo de genes específicos (en cuyo caso se especifica dicho grupo, seleccionando la opción *Custom*) [47].

Críticamente, el campo *Significance threshold* indica el método a utilizar para la corrección de los *p-values* por testeo múltiple. Por defecto, g:Profiler utiliza un método de corrección propio denominado g:SCS, que toma en cuenta el hecho de que las pruebas de hipótesis no son independientes entre sí, como se asume por los otros dos métodos de corrección escogibles: Benjamini-Hochberg y Bonferroni [47]. Aún así, se ha hecho notar previamente que el método

g:SCS, tal como la corrección Bonferroni, puede resultar demasiado conservador, siendo el método Benjamini-Hochberg frecuentemente utilizado en estudios exploratorios [29, 43]. Otra forma de alterar el balance entre poder estadístico y control de falsos positivos es mediante el valor de corte del *adjusted p-value* para asignar significancia estadística a las pruebas, especificado mediante el campo *User threshold*, cuyo valor por defecto es 0.05. Cuando no se hallan términos enriquecidos usando Benjamini-Hochberg con un valor de corte de 0.05, otros valores más permisivos pueden ser utilizados, como 0.1 e incluso 0.25, a conciencia de arriesgar un FDR más alto [43].

Terminando con la inspección de la Figura 1.4, se observa que los campos del apartado Gene Ontology permiten especificar el tipo de términos GO a testear y si se quiere o no excluir las anotaciones a términos GO determinadas por inferencias basadas en análisis computacionales automáticos, asignadas con el código de evidencia IEA (acrónimo de *Inferred from Electronic Annotation*). Si bien la calidad de dichas anotaciones en comparación con otras manualmente curadas es un tema de debate, es común considerarlas en el análisis dada su gran abundancia y el hecho de que ignorarlas puede sesgar los resultados, favoreciendo a las funciones biológicas mejor estudiadas [43].

g:Profiler permite resumir y descargar los resultados del análisis de enriquecimiento funcional para una lista de genes, en un formato denominado GEM (acrónimo de *General Enrichment Map*). Los archivos GEM consisten en una tabla, donde cada fila se corresponde con un término funcional enriquecido e incluye el identificador del término funcional, su descripción y el *adjusted p-value* asociado con su respectiva prueba de hipótesis. También es posible descargar, en archivos de formato GMT (acrónimo de *Gene Matrix Transposed*), las anotaciones de genes de un organismo de referencia a términos funcionales o, retomando la notación previa, la definición de los conjuntos T cuyo enriquecimiento es testeado por g:Profiler en la lista Q [43]. Cómo se verá en la siguiente sección, los archivos GEM y GMT resultan útiles como insumo para la visualización de los resultados del análisis de enriquecimiento funcional.

1.2.3.2. Visualización de términos enriquecidos

Cytoscape es una herramienta con versión de escritorio para la visualización de redes en el contexto de la bioinformática, permitiendo la integración de distintos tipos de datos en estas [43, 48]. Dependiendo de los objetivos del usuario, distintos *plugins* pueden ser cargados en Cytoscape para proveer funcionalidades específicas. Para la visualización de términos funcionales enriquecidos en múltiples listas de genes, tres *plugins* de utilidad son *EnrichmentMap* [49], *clusterMaker2* [50] y *WordCloud* [51].

A partir de archivos GEM generados por g:Profiler para múltiples listas de genes y un archivo GMT, EnrichmentMap permite crear un *mapa de enriquecimiento*. Dicho mapa consiste en un grafo cuyos nodos corresponden a los términos funcionales incluidos en al menos un archivo GEM (es decir, enriquecido en al menos una de las listas de genes). Para visualizar el nivel de enriquecimiento de un término en las distintas listas, a estas se les asignan distintas secciones del respectivo nodo, tal como se ilustra en la Figura 1.5

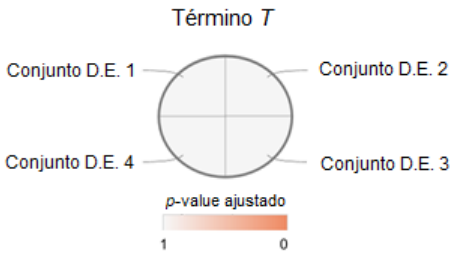

Figura 1.5: Estructura de un nodo de EnrichmentMap. En este ejemplo, el nivel enriquecimiento del término T en cada uno de cuatro conjuntos de genes (medido según el *adjusted p-value* de la respectiva prueba de hipótesis) puede ser visualizado mediante la coloración de su respectiva sección en el nodo.

Aristas de grosor variable entre los nodos representan similaridad entre los términos funcionales, según sus definiciones especificadas en el archivo GMT. La similaridad entre dos términos T_1 y T_2 se cuantifica mediante un coeficiente de similaridad, calculado como una suma ponderada de los valores *Jaccard Coefficient* (JC) y *Overlap Coefficient* (OC), definidos como [52]:

$$JC(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$
$$OC(T_1, T_2) = \frac{|T_1 \cap T_2|}{min(|T_1|, |T_2|)}$$

Por defecto, el coeficiente de similaridad se obtiene sumando los valores anteriores multiplicados por 0.5. El panel de opciones de EnrichmentMap permite fijar un valor de corte tal que las aristas entre términos con coeficiente de similaridad menor a dicho valor son omitidas.

La complejidad de los mapas de enriquecimiento en bruto generados por EnrichmentMap puede ser reducida notablemente agrupando los términos similares en *clusters* que, en el contexto de este trabajo, también serán llamados *módulos funcionales* [43]. Atendiendo a los pesos de las aristas (similaridades) entre nodos computados por EnrichmentMap, clusterMaker2 permite dividir el grafo en *clusters* de nodos mediante algoritmos como MCL (*Markov Cluster Algorithm*). En pocas palabras, los *clusters* definidos por MCL corresponden a regiones del grafo densas en aristas, dentro de las cuales el número de caminos de tamaño k (para k pequeños en $\mathbb{N}$) es relativamente alta [53]. Una vez definidos los módulos funcionales, WordCloud permite asignarles automáticamente títulos, identificando consensos entre los títulos de los términos individuales. La generación del mapa de enriquecimiento termina con la curación manual de los títulos de los módulos [43].

Como se mencionó al inicio de esta subsección, una de las virtudes de Cytoscape es que permite la integración de diversos tipos de datos a los grafos. En particular, para los mapas de enriquecimiento generados sobre listas de genes D.E., se pueden incorporar tablas que dan cuenta del patrón de expresión de los genes a nivel de muestras o condiciones [43]. Provisto que los identificadores de los genes en dichas tablas de expresión concuerden con los identificadores usados en la definición de los términos funcionales según el archivo GMT, la tabla de expresión puede ser filtrada en Cytoscape para mostrar exclusivamente la expresión de los genes asociados con nodos seleccionados [43]. Tales subtablas pueden ser descargadas

en formato de texto para posteriores análisis (en este trabajo las subtablas se usaron para la generación de *heatmaps* en R).

1.3. Motivación del trabajo

Como preámbulo a la formalización de los Objetivos, en esta breve sección se precisa la motivación del trabajo, en cuanto a la necesidad concreta que busca satisfacer.

Tal como se expuso en la Sección 1.1.2, aunque en la literatura se ha manifestado interés por identificar efectos de *Wolbachia* comunes a distintos hospederos Diptera, ningún estudio se ha dedicado específicamente a este propósito.

Obteniendo listas D.E. directamente desde bibliografía o mediante análisis *de novo* de datos RNA-Seq publicados, y procesándolas mediante análisis de enriquecimiento de términos GO y visualización en mapas de enriquecimiento, este trabajo pretende identificar efectos comunes de infecciones *Wolbachia* nativas sobre hospederos Diptera.

Como se argumentó en la Sección 1.1.2, conocer los efectos consistentes de las infecciones nativas podría ayudar a predecir cambios y continuidades en las interacciones actualmente observadas en transfecciones, sugiriendo medidas para conservar la efectividad de las estrategias de control de vectores arbovirales en el tiempo y contribuyendo al conocimiento general de la bacteria.

1.4. Objetivos

El objetivo general de este trabajo fue identificar procesos biológicos o funciones moleculares comúnmente afectadas por infecciones *Wolbachia* nativas en hospederos del orden Diptera.

Los objetivos específicos consistieron en:

- Obtener listas de genes sobre- y subexpresados en distintas especies Diptera por efecto de infecciones *Wolbachia* nativas.
- Identificar términos funcionales GO enriquecidos en cada lista de genes.
- Identificar términos GO enriquecidos en listas de múltiples especies.

Capítulo 2

Metodología

2.1. Obtención de listas de elementos D.E.

Se buscaron publicaciones en que se hubieran comparado transcriptomas de individuos Diptera con infecciones *Wolbachia* nativas removidas (total o parcialmente) y no removidas, mediante RNA-Seq. Tales búsquedas fueron realizadas en Google Scholar y en el repositorio de la Universidad de Chile mediante las palabras clave "wolbachia", "native", "transcriptomics" y "diptera" conectadas por un operador de conjunción.

Listas de genes D.E. fueron obtenidas bien a partir de tablas suplementarias en las publicaciones o bien mediante un análisis RNA-Seq propio sobre las lecturas en bruto almacenadas en repositorios públicos. En cualquier caso, la finalidad de esta etapa fue obtener listas de identificadores únicos de genes sobre- y subexpresados, ordenados según la magnitud del cambio de expresión, reconocibles por la herramienta escogida para el análisis de enriquecimiento (módulo g:GOSt de g:Profiler) y correspondientes a los organismos de referencia definidos para dicho análisis: *Drosophila melanogaster* para hospederos del género *Drosophila* y *Aedes aegypti* para hospederos del género *Aedes*.

2.1.1. Obtención desde tablas

En el caso de los análisis transcriptómicos publicados sobre el hospedero modelo *Drosophila melanogaster*, los genes D.E. se hallaron en las tablas mediante un identificador único, correspondiente al organismo de referencia definido para el análisis de enriquecimiento (también *D. melanogaster*) y directamente interpretable en g:Profiler. Así, bastó con distribuir los identificadores en dos archivos de texto dependiendo del sentido de la expresión diferencial. En un archivo se almacenaron los identificadores de elementos sobreexpresados (*adjusted p-value* < 0.05 y $log_2FC > 0$) y en otro los subexpresados (*adjusted p-value* < 0.05 y $log_2FC < 0$). Las listas fueron ordenadas decrecientemente según el valor absoluto de log_2FC.

En el caso de los análisis transcriptómicos publicados sobre hospederos no-modelo, se requirió una manipulación adicional de los identificadores en las tablas, que comprendió la búsqueda de ortólogos del respectivo organismo de referencia escogido para el análisis de enriquecimiento funcional. La razón es que la identificación de las secuencias expresadas en los hospederos no-modelo fue mediada por ensambles de transcriptoma *de novo*, y la posterior

caracterización de los transcritos ensamblados (basada en búsquedas BLAST) no siempre les asignó un identificador del organismo de referencia respectivo.

Los procesos de decisión para la recuperación de IDs mediada por ortología fueron definidos dependiendo de los tipos específicos de información contenida en las tablas. El contenido original de dichas tablas se expone en las Tablas B.1 y B.2 del Anexo B.1, mientras que los procesos de decisión definidos se detallan en las Figuras B.1 y B.2 del mismo Anexo. A pesar de sus particularidades, dichos métodos tuvieron en común el objetivo de recuperar tantas IDs del respectivo organismo de referencia como fuera posible, manteniendo criterios estándar para inferir ortología (esto es, usando los parámetros por defecto en las herramientas utilizadas: el buscador de proteínas por secuencia de STRING [54] y/o el módulo g:Orth de g:Profiler).

Los identificadores obtenidos tras estos pasos intermedios fueron separados en dos archivos, sobre- y subexpresados, según los mismos criterios previamente descritos. En los contados casos en que un mismo identificador se halló en una lista asociado con múltiples valores log_2FC, un promedio simple de estos valores fue asignado.

2.1.2. Obtención mediada por análisis RNA-Seq propio

La obtención mediada por análisis propio fue preferida en caso de que las tablas publicadas omitieran información necesaria según la metodología aquí escogida para el análisis de enriquecimiento de términos funcionales (descrita posteriormente). Los únicos conjuntos de datos que requirieron un análisis RNA-Seq propio provinieron de *Drosophila melanogaster*.

Los archivos con lecturas en formato FASTQ fueron adquiridos desde el *Sequence Read Archive* (SRA) del *National Center for Biotechnology Information* (NCBI) [55]. La calidad de los archivos fue evaluada mediante FastQC [56] y los reportes generados se reunieron en un único reporte MultiQC [57] para facilitar su inspección. Ante la presencia de adaptadores acusada por el módulo *Adapter content* de FastQC, los archivos fueron procesados mediante Trimmomatic [58], activando la opción ILLUMINACLIP para detectar y remover adaptadores usados en secuenciación Illumina. Principalmente para remover extremos 3' con calidad Q_{Phred} subóptima, la opción SLIDINGWINDOW fue activada con los parámetros por defecto (remoción de ventanas de 4 bases con promedio Q_{Phred} menor a 20). El resultado del tratamiento mediante Trimmomatic fue evaluado mediante la generación de un nuevo reporte MultiQC; las advertencias en los módulos fueron revisadas y su importancia determinada según lo descrito en el Anexo A para lecturas RNA-Seq. En general, anormalidades con respecto a los módulos *Per Base Sequence Content*, *Sequence Duplication Levels* y *Overrepresented Sequences* de FastQC fueron toleradas, mientras que anormalidades con respecto al módulo *Per Sequence GC Content* fueron aceptadas en la medida de que fueran comunes entre los archivos (no excepcionales) y en que, visualmente, las curvas presentadas por dicho módulo describieran una distribución aproximadamente normal.

Para el alineamiento de lecturas, el genoma de referencia dm6 [59] en formato FASTA y su anotación de elementos genómicos dmel_r6.32 [60] en formato GTF (*Gene Transfer Format*) fueron obtenidos desde la base de datos FlyBase [61]. Para cada archivo de lecturas (o par de archivos en el caso de lecturas *paired-end*), un alineamiento espaciado al genoma

de referencia fue realizado con la guía de los eventos de *splicing* reportados en dmel_r6.32, mediante la herramienta STAR [62] con parámetros por defecto. Los reportes de alineamiento entregados por STAR para cada archivo o par de archivos fueron reunidos mediante MultiQC, para revisar la distribución de las lecturas según la multiplicidad de sus alineamientos y detectar posibles anomalías. Provisto que la cantidad de *multireads* fuera relativamente baja, los mapas de alineamientos generados por STAR en formato BAM fueron procesados junto con el archivo dmel_r6.32 mediante la herramienta featureCounts [63], para realizar conteos brutos de lecturas a nivel de genes con parámetros por defecto (en particular, ignorando los *multireads*).

Las tablas de expresión resultantes fueron sometidas a análisis de expresión diferencial mediante DESeq2 [64] con parámetros por defecto, definiendo un único factor (tratamiento *Wolbachia*) y asignando las columnas a sus respectivos niveles (WT o GFR). El orden en que fueron definidos los niveles implicó que los tamaños de los efectos reportados correspondieran a los cambios de expresión en la condición WT con respecto a la GFR. La similaridad global de las muestras en términos de cuentas normalizadas fue evaluada cualitativamente desde los gráficos PCA entregados por DESeq2. En caso de haber muestras evidentemente aisladas según su representación en el PCA, estas fueron removidas de su respectiva tabla de expresión, y la tabla actualizada fue sometida nuevamente al análisis DESeq2. De la tabla con los resultados de las pruebas de hipótesis se extrajeron los identificadores de genes asociados con *adjusted p-value* < 0.05, distribuyéndose en dos archivos de texto según el sentido de su expresión diferencial.

2.2. Análisis de enriquecimiento de términos funcionales

Análisis de enriquecimiento de términos GO:BP, GO:MF y GO:CC se realizaron sobre los conjuntos de elementos sobre- y subexpresados mediante la API de g:Profiler (módulo g:GOSt) invocada desde Python 3.8.8. El código fuente escrito para este propósito y para toda acción ejecutada en Python aludida en el resto de la Metodología, se presenta y describe en el Anexo B.2.

En cada análisis se usó el respectivo organismo de referencia (*D. melanogaster* o *A. aegypti*), se trató a la consulta como una lista ordenada, no se impuso restricciones sobre los códigos de evidencia y se definió el conjunto de genes de fondo como aquellos anotados en el organismo de referencia. Dos opciones se probaron para el control del falsos positivos: ajuste g:SCS con *adjusted p-value* de corte 0.05 (en adelante referidos como parámetros *estrictos*), y ajuste Benjamini-Hochberg con *adjusted p-value* de corte 0.20 (en adelante referidos como parámetros *relajados*). Tamaños mínimos y máximos para los términos a testear fueron establecidos en 4 y 350, con el fin de reducir el número de pruebas de hipótesis y excluir términos muy generales (poco informativos), respectivamente.

La baja cantidad de términos enriquecidos bajo los parámetros estrictos no justificó una visualización en Cytoscape. Así, los términos relevantes fueron identificados directamente a partir de los resultados entregados por g:GOSt. Por el contrario, los términos enriquecidos bajo los parámetros relajados fueron más abundantes, justificando una posterior visualización.

En este caso, archivos GEM asociados con las listas D.E. se generaron mediante código Python para proveer a Cytoscape. Para visualizar términos candidatos a describir alteraciones comunes causadas por *Wolbachia*, en los archivos GEM sólo se incluyeron aquellos términos que se hallaran enriquecidos en 4 o más listas D.E.

2.3. Visualización de enriquecimiento en múltiples listas

Para cada clase de términos GO (BP, MF y CC) se construyeron mapas de enriquecimiento en Cytoscape, con la ayuda de los *plugins* EnrichmentMap, clusterMaker2 y WordCloud.

Inicialmente, los archivos GEM fueron cargados en EnrichmentMap junto con el archivo GMT de *D. melanogaster* asociado con el tipo de término GO correspondiente, generando un mapa en bruto. Posteriormente, los términos fueron agrupados en módulos usando clusterMaker2 con el algoritmo "MCL Cluster" y con base en los coeficientes de similaridad entre las categorías, donde la definición del coeficiente de similaridad fue la establecida por defecto. Notar que, dado que el archivo GMT provisto fue de *D. melanogaster*, la similaridad entre los términos (y por lo tanto, el agrupamiento en módulos) dependió de las anotaciones específicas para dicho organismo.

Nombres de consenso fueron generados automáticamente para cada grupo de términos mediante WordCloud, usando el algoritmo por defecto, y fueron posteriomente corregidas de forma manual. Finalmente, el valor de corte del coeficiente de similaridad para la visualización de aristas fue fijado en el mínimo valor que omitiera todas las aristas entre módulos.

2.4. Generación de *heatmaps* con niveles de expresión

Para comparar entre las listas la identidad y nivel de expresión diferencial de los genes que invocan a un módulo de términos múltiplemente enriquecidos, se generó un *heatmap* con los log_2FC de todos los genes D.E. que se asocian, en al menos una lista, al módulo en cuestión. Para obtener un punto de comparación entre los genes de las especies *Drosophila* y *Aedes*, los identificadores de *A. aegypti* fueron mapeados por ortología a *D. melanogaster*, mediante el módulo g:Orth de g:Profiler. Nuevamente, en los pocos casos en que se obtuvieron múltiples mapeos al mismo identificador, este fue asignado con un log_2FC promedio.

Una tabla con los $log2FC$ de los ortólogos de *D. melanogaster* a través de las listas D.E. (*archivo de expresión*) fue generado mediante Python, y posteriormente cargado a la sesión Cytoscape en que se construyeron los tres mapas de enriquecimiento. Seleccionando cada módulo funcional, una subtabla (*sub-archivo de expresión*) con los $log2FC$ de los identificadores que invocan a dicho módulo fue exportada. Finalmente, los sub-archivos de expresión fueron procesados con R 4.1.0 [65] para la generación de *heatmaps* asociados con cada módulo, usando el paquete *gplot*. El código R específico escrito para tal propósito se presenta y describe en el Anexo B.2.

Capítulo 3

Resultados y discusiones

3.1. Publicaciones seleccionadas

A continuación se presentan las publicaciones seleccionadas, describiendo aspectos relevantes para explicar el origen de las listas de elementos D.E. obtenidas. Además, se indica el tipo de caracterización funcional realizada originalmente sobre los elementos D.E. por los diversos autores, enfatizando las respectivas diferencias con la metodología propia.

3.1.1. Caragata *et al.* (2017)

En este estudio se comparó el transcriptoma global de *Aedes fluviatilis* con y sin su endosimbionte nativo wFlu, para explorar las interacciones biológicas entre mosquitos e infecciones *Wolbachia* nativas y, particularmente, investigar las bases de la potenciación de infecciones de *Plasmodium gallinaceum* en *A. fluviatilis* con wFlu [17].

El efecto de remover wFlu sobre la expresión génica de cuerpo completo fue estudiado en hembras *A. fluviatilis* de 6 días de edad, alimentadas con solución de sucrosa, mantenidas a 27 ± 1°C y con ciclos de luz/oscuridad de 12h/12h [17]. Cada condición biológica (WT y GFR) se representó por 3 muestras, cada una de ellas compuesta por 16 individuos. Los individuos GFR fueron tales que sus antecesores WT, varias generaciones atrás, fueron tratados con tetraciclina para eliminar la infección *Wolbachia* (el transcurso de varias generaciones permitió descartar razonablemente la influencia de antibiótico residual o microbiota intestinal alterada sobre los resultados). Las muestras derivaron en librerías de ADNc sin conservación de la información de hebra, que fueron secuenciadas en una plataforma Illumina MiSeq con lecturas en formato *paired-end* de 300 pares de bases [17].

Luego de la remoción de adaptadores y evaluación de calidad mediante FastQC, las lecturas se ensamblaron en un transcriptoma *de novo* mediante Trinity (parámetros por defecto), respondiendo a la ausencia de un genoma de referencia para *Aedes fluviatilis*. Posteriormente, las lecturas fueron alineadas a los *contigs* ensamblados mediante BowTie2 y contabilizadas mediante una rutina propia no especificada. Las cuentas se usaron para un análisis de expresión diferencial mediante baySeq, seleccionándose los *contigs* D.E. según el criterio *adjusted p-value* < 0.05 [17].

Nuevamente, dado el desconocimiento sobre el genoma de *Aedes fluviatilis*, los *contigs* ensamblados fueron asociados por homología de secuencias vía BLAST a identificadores de productos génicos de otras especies artrópodas, principalmente *Aedes aegypti* y *Culex quinquefasciatus*. Excluyendo asociaciones múltiples a un mismo identificador, los 159 *contigs* sobreexpresados rindieron 95 identificadores y los 98 *contigs* subexpresados rindieron 59 identificadores [17]. Términos GO fueron extraidos para caracterizar funcionalmente a los *contigs* identificados en la etapa anterior, y usados para organizar su discusión en ámbitos funcionales definidos de forma manual. Los identificadores no se sometieron a análisis de enriquecimiento de términos funcionales.

3.1.2. He *et al.* (2019)

En esta publicación se comparó el transcriptoma de ovarios de *Drosophila melanogaster* con y sin su infección nativa wMel para investigar cómo ocurre el rescate de la incompatibilidad citoplasmática por parte de hembras infectadas, y cómo la bacteria modifica los ovarios de su hospedero, beneficiando su propia supervivencia y propagación [30].

El efecto de remover wMel sobre la expresión génica de ovarios fue estudiado en hembras *D. melanogaster* vírgenes de 4 días de edad, alimentadas con medio Agar Harina de Maíz y melaza, mantenidas a 25°C y con ciclos luz/oscuridad de 12h/12h [30]. Cada condición (WT y GFR) se representó por 3 muestras, cada una de ellas compuesta por 80 pares de ovarios. Los individuos GFR fueron tales que sus antecesores WT, varias generaciones atrás, fueron tratados con tetraciclina para eliminar la infección *Wolbachia* (el transcurso de seis generaciones intermedias permitió descartar razonablemente la influencia de antibiótico residual o microbiota intestinal alterada sobre los resultados). El ARN mensajero de las muestras fue aislado del ARN total usando perlas con anclajes oligo-dT, convertido a ADNc sin conservar información de hebra y secuenciado en una plataforma Illumina HiSeq 4000, con lecturas en formato *paired-end* de 150 pares de bases [30].

El control de calidad de los archivos de lecturas consistió en remover adaptadores y lecturas ambiguas o de baja calidad. No se especificaron los procedimientos seguidos para el alineamiento ni para el conteo de lecturas. 116 genes sobreexpresados y 33 subexpresados se seleccionaron según un criterio $|log_2(fold\ change)| \geq 1$ y *adjusted p-value* < 0.05, aunque no se especificó el método usado para las pruebas estadísticas que dieron lugar a dichos indicadores [30]. Un análisis de enriquecimiento de términos GO fue realizado mediante el paquete de R GOSeq sobre la lista de genes D.E. identificados, sin separación entre sobre- y subexpresados y definiendo un *adjusted p-value* de corte de 0.05, no halládose ningún término significativamente enriquecido [30].

Dada la omisión de parte importante de la metodología y el hecho de que los genes D.E. fueron filtrados según la magnitud del cambio de expresión, se decidió rescatar las lecturas en bruto (almacenadas en el NCBI-SRA bajo el código SRP136211) y someterlas a un análisis RNA-Seq propio.

3.1.3. Baiao *et al.* (2019)

En este estudio se compararon los transcriptomas de cabezas y abdómenes de hembras y machos de tres semiespecies de *Drosophila paulistorum* (*Orinocan* (OR), *Amazonian* (AM) o *Centro American* (CA)), con infecciones *Wolbachia* nativas intactas o parcialmente eliminadas. Los objetivos de este trabajo fueron investigar los efectos de infecciones *Wolbachia* nativas en *D. paulistorum* y explorar la hipótesis de que la bacteria puede jugar un rol en el proceso de especiación de sus hospederos [31].

Se definieron bloques para cada combinación de semiespecie (OR, AM o CA), sexo (macho o hembra) y tejido (abdómenes o cabezas). El tratamiento cuyo efecto fue evaluado en cada bloque fue la remoción parcial de la infección *Wolbachia*, mediada por un tratamiento con rifampicina, una posterior recuperación de la microbiota intestinal, y el transcurso de una generación. Los individuos fueron alimentados con Formula 4-24®, mantenidos a 21-22°C y con ciclos luz/oscuridad de 12h/12h. Ningún esfuerzo se realizó para mantener a los individuos vírgenes. Cada combinación de semiespecie, sexo, tejido y estado de infección fue representada por tres réplicas, cada una compuesta por 20 cabezas o 10 abdómenes de individuos de 3 días, según correspondiera. Las muestras de ARN total fueron depletadas de ARN ribosomal, el ADNc se preparó preservando la información de hebra y la secuenciación se realizó en una plataforma Illumina HiSeq2500, con lecturas en formato *paired-end* de 125 pares de bases [31].

Luego de un control de calidad que incluyó la remoción de adaptadores, los archivos con lecturas fueron ensamblados mediante Trinity para formar transcriptomas *de novo* para cada semiespecie por separado, los cuales fueron evaluados según criterios como porcentaje de lecturas mapeables de vuelta, porcentaje de marcos de lectura abiertos completos e inclusión de secuencias conservadas en Arthropoda, Diptera e Insecta [31]. El alineamiento y conteo de lecturas fue realizado, respectivamente, mediante STAR y featureCounts (este último con la opción de contar *multireads* en cada alineamiento candidato y considerar la información de hebra). Los *contigs* fueron considerados diferencialmente expresados con base en un criterio *adjusted p-value* < 0.05, usando la prueba por defecto en DESeq2 [31]. La cantidad de *contigs* D.E. presentados por la subespecie OR fue sustancialmente mayor que para las semiespecies AM y CA, presumiblemente debido a la pérdida de señales transcriptómicas de células infectadas en dichas subespecies, que presentan baja densidad bacteriana [31]. Por este motivo, los autores centraron la discusión en los resultados de la semiespecie OR, proponiendo que posteriores estudios de las demás semiespecies fueran mediados por una selección de las células específicas infectadas [31]. En este trabajo se tomó una decisión similar, rescatando sólo la información asociada con la semiespecie OR.

Para identificar los *contigs* ensamblados, estos se sometieron a una búsqueda BLAST sobre una base de datos con genes de *Drosophila melanogaster*, *Drosophila willistoni*, *Wolbachia* y otras bacterias. Así, para las comparaciones de abdómenes de hembras OR, 325 y 164 *contigs* sobre- y subexpresados rindieron 203 y 143 identificadores de genes Drosophila, respectivamente. Para las comparaciones de cabezas de hembras OR, 36 y 202 *contigs* sobre- y subexpresados rindieron 23 y 188 identificadores de genes Drosophila, respectivamente. Para las comparaciones de abdómenes de machos OR, 223 y 324 *contigs* sobre- y subexpresados rindieron 108 y 282 identificadores de genes Drosophila, respectivamente. Para las comparaciones de cabezas de machos OR, 25 y 225 *contigs* sobre- y subexpresados rindieron 8 y 203

identificadores de genes Drosophila, respectivamente.

Términos funcionales GO asociados con los identificadores D.E. obtenidos fueron recuperados desde Flybase y utilizados para análisis de enriquecimiento funcional mediante TopGO con el criterio *adjusted p-value* < 0.05 según la prueba hipergeométrica y personalizando el universo para contener un gen por cada módulo Trinity Gene (esto es, un conjunto de *contigs* cuyo origen en un mismo gen es predicha por la herramienta Trinity) [31].

3.1.4. Lindsey *et al.* (2021)

En esta publicación se comparó el transcriptoma de *Drosophila melanogaster* con y sin infección wMel nativa y virus Sindbis (SINV), para estudiar las bases del bloqueo patogénico por parte de *Wolbachia* [15].

Como se mencionó en el Marco Teórico, las condiciones experimentales se definieron como combinaciones de los tres factores: tratamiento *Wolbachia* (con niveles WT y GFR), tipo de inyección (con niveles *SINV* y *suero*) y horas post-inyección (con niveles *8hpi*, *24hpi* y *48hpi*). Cada condición fue representada por cuatro réplicas conformadas por 5 hembras adultas vírgenes alimentadas con medio Agar Harina de Maíz, mantenidas a 25°C y con ciclos luz/oscuridad de 24h/24h [15]. Los individuos representantes de las condiciones sin *Wolbachia* fueron obtenidos mediante tratamiento con tetraciclina, posterior restauración de la microbiota intestinal y transcurso de varias generaciones. En la preparación de las librerías de ADNc se depletó el ARN ribosomal, se conservó la información de hebra y se usó una plataforma Illumina NextSeq con lecturas *single-end* de 75 pares de bases [15].

Luego de un control de calidad, las lecturas fueron alineadas mediante BowTie a transcritos extraídos desde el genoma anotado de *D. melanogaster* (*dmel6*), y cuantificadas mediante RSEM. La expresión diferencial fue testeada en edgeR a nivel de genes (esto es, agregando las lecturas de los transcritos derivados de un gen), usando el método de normalización TMM [15]. Para evaluar los efectos principales e interactivos de los niveles de los tres factores experimentales, las pruebas estadísticas se realizaron sobre los parámetros de un modelo lineal generalizado multivariable. En particular, 123 genes sobreexpresados y 114 subexpresados por efecto de *Wolbachia* se obtuvieron según el criterio *adjusted p-value* < 0.05 [15].

Análisis de enriquecimiento de términos GO fueron realizados para caracterizar subconjuntos de genes D.E. estrechamente ligados según una base de datos de interacciones de proteínas (STRING), sin embargo, las listas completas de genes sobre- y subexpresados por efecto de *Wolbachia* no fueron caracterizadas de esta forma (tampoco lo fue la lista de todos los genes D.E.) [15].

3.1.5. Detcharoen *et al.* (2021)

En este trabajo se compararon las alteraciones funcionales causadas por wMel en su hospedero nativo *Drosophila melanogaster* y en el hospedero transfectado *Drosophila nigrosparsa*, con base en la expresión de un grupo de genes ortólogos entre ambas especies [29]. Aquí se rescatarán exclusivamente los datos generados para *D. melanogaster*, dado que wMel no es una infección nativa en *D. nigrosparsa*. Más aún, considerando la pérdida de datos incurrida

al filtrar la lista de genes D.E. en *D. melanogaster* según ortología con *D. nigrosparsa*, se decidió realizar un análisis de expresión diferencial propio a partir de las lecturas en bruto. Por dicho motivo, la restante descripción metodológica explicará exclusivamente la producción de los datos brutos de *D. melanogaster*.

El factor cuyo efecto se estudió sobre hembras *D. melanogaster* completas de 5 días de edad fue el estado de infección *Wolbachia*, con niveles infectado y no infectado. La condición WT fue representada por 15 réplicas, cada una compuesta por 5 hembras infectadas completas, provenientes de tres líneas aisladas en laboratorio [29]. Por su parte, los individuos representantes de la condición GFR fueron obtenidos de una cuarta línea aislada en laboratorio, mediante un tratamiento con tetraciclina y el transcurso de 6 generaciones. La condición GFR fue representada por 5 réplicas de 5 hembras completas cada una. Todos los individuos fueron alimentados con Agar de Jugo de uva, malta y levadura, mantenidos a 19°C y con ciclos luz/oscuridad de 16h/8h [29].

El ARN ribosomal fue removido en la preparación de la librería de secuenciación y se generaron lecturas *single-end* de 75 pares de bases usando una plataforma Illumina NextSeq 500. Las lecturas en bruto fueron almacenadas en el repositorio SRA del NCBI, bajo el código de proyecto PRJNA602188 [29].

3.2. Listas de genes D.E. obtenidas

En la Tabla 3.1 se asigna un nombre a cada lista de elementos sobre- o subexpresados obtenida en este trabajo y se resumen las condiciones experimentales que les dieron origen.

Tabla 3.1: Resumen del origen experimental de las listas de elementos D.E. En cada caso, la expresión génica del tejido de hospederos infectados fue comparado frente a una eliminación (total o parcial) de la infección. *Nombre no oficial, asignado a la cepa nativa de *Drosophila paulistorum* (semiespecie OR) en el ámbito del presente informe.

Publicación	Cepa	Tejido hospedero	Listas de elementos D.E.
Caragata *et al.* (2017) [17]	wFlu	Hembras *A. fluviatilis* completas, 6 días de edad	Car_up Car_down
He *et al.* (2019) [30]	wMel	Ovarios de *D. melanogaster*, 4 días de edad	He_up He_down
Baiao *et al.* (2019) [31]	wPau*	Abdómenes de hembras *D. paulistorum* (OR), 3 días de edad	Bai_F_abd_up Bai_F_abd_down
		Abdómenes de machos *D. paulistorum* (OR), 3 días de edad	Bai_M_abd_up Bai_M_abd_down
		Cabezas de hembras *D. paulistorum* (OR), 3 días de edad	Bai_F_head_up Bai_F_head_down
		Cabezas de machos *D. paulistorum* (OR), 3 días de edad	Bai_M_head_up Bai_M_head_down
Lindsey *et al.* (2021) [15]	wMel	Hembras *D. melanogaster* completas, 5 días de edad	Lind_up Lind_down
Detcharoen *et al.* (2021) [29]	wMel	Hembras *D. melanogaster* completas, 5 días de edad	Det_up Det_down

Las listas He_up, He_down, Det_up y Det_down fueron obtenidas a partir de los datos transcriptómicos en bruto publicados por He *et al.* (2019) y Detcharoen *et al.* (2021), mediante un análisis RNA-Seq propio. Las listas Lind_up y Lind_down fueron obtenidas

directamente desde tablas suplementarias de Lindsey *et al.* (2021). El resto de las listas fueron obtenidas desde tablas suplementarias de Caragata *et al.* (2017) y Baiao *et al.* (2019), más una posterior búsqueda de ortólogos en los organismos de referencia definidos para el análisis de enriquecimiento funcional (*A. aegypti* y *D. melanogaster*, respectivamente). Las listas obtenidas completas se adjuntan en el Anexo C (Tablas C.1 a C.16). Los detalles de los procesos de obtención (incluyendo evaluaciones de calidad de lecturas, estadísticas de alineamiento y cantidad de identificadores perdidos en las búsquedas de ortología) se presentan y discuten en el Anexo D. En la Figura 3.1 se resume el tamaño de las listas obtenidas.

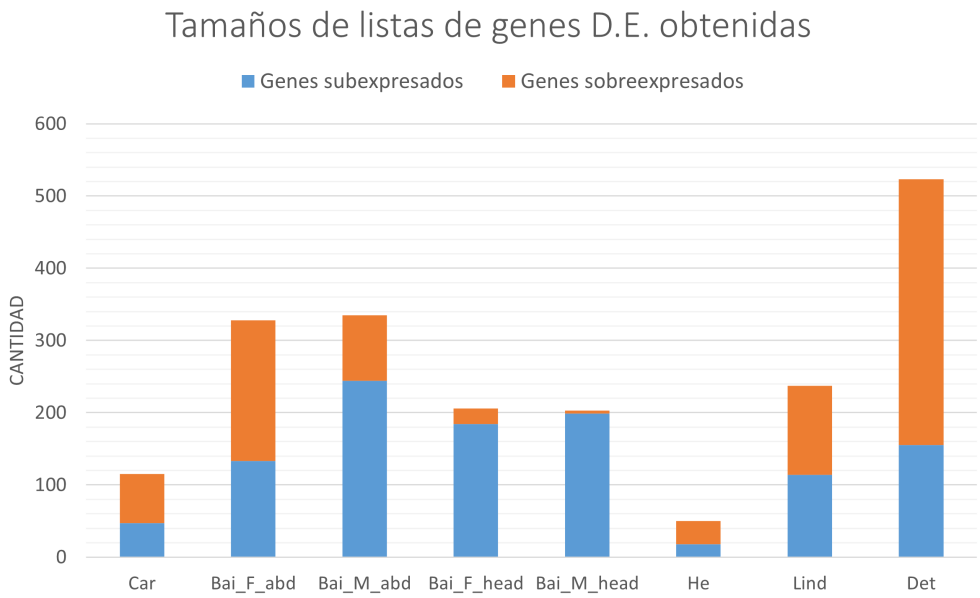

Figura 3.1: Cantidad de genes en cada lista D.E. obtenida. Cada ubicación en el eje horizontal se corresponde con una comparación WT vs GFR. La cantidad de genes sobre- y subexpresados en una comparación se representa por el tamaño (extensión a lo largo del eje vertical) de la respectiva barra roja y azul, respectivamente.

Las listas obtenidas desde Caragata *et al.* (2017) y He *et al.* (2019) llaman la atención por la baja cantidad de genes sobre- y subexpresados que acumulan. Como se discute en el Anexo D.1, hay razones para creer que la identificación de genes D.E. desde los datos de He *et al.* (2019) sufrió de una baja sensibilidad, en virtud de una combinación de alta variabilidad intra-condición, bajo nivel de replicación y baja cobertura por muestra; exacerbado por el uso de un algoritmo relativamente conservador para el análisis de expresión diferencial. Por su parte, según lo expuesto en el Anexo D.2, las listas provenientes de Caragata *et al.* (2017) resultaron sustancialmente reducidas en dos instancias de búsquedas de secuencias homólogas (una propia y una por parte de los autores).

Las listas provenientes de Baiao *et al.* (2019) son relativamente abundantes, aún cuando, tal como las de Caragata *et al.* (2017), fueron obtenidas en un organismo no modelo y una cantidad considerable de *contigs* ensamblados fueron descartados del posterior análisis por no recibir anotación. Esto se puede atribuir a que una cantidad particularmente alta de *contigs* D.E. fueron hallados inicialmente, posiblemente en virtud de la tejido-especificidad con que se realizaron las comparaciones, como fue discutido por los mismos autores [31]. También, como se discute en el Anexo D.3, el origen de los datos de Baiao *et al.* (2019) en *D. paulistorum*

−un organismo filogenéticamente cercano al modelo *D. melanogaster*− pudo significar una menor pérdida porcentual de elementos D.E. en las instancias de búsqueda ortóloga.

Las listas Det_up y Det_down contienen, en conjunto, la mayor cantidad de genes D.E. entre todas las comparaciones, lo cual es llamativo considerando que los datos RNA-Seq que les dieron origen corresponden a muestras de cuerpos completos. Como se expone en el Anexo D.4, la condición WT se representó por 15 muestras, 14 de las cuales fueron retenidas en el presente análisis de expresión diferencial y nueve de las cuales se agruparon estrechamente según un PCA, separándose claramente del grupo de muestras GFR. Una variabilidad intra-condición reducida y un importante nivel de replicación pudieron proveer una buena sensibilidad para la detección de cambios de expresión.

Para ilustrar tendencias generales de la expresión a nivel de genes, el *heatmap* de la Figura 3.2 muestra el patrón de expresión diferencial de genes *D. melanogaster* que fueron identificados como D.E. (o cuyos ortólogos en *A. fluviatilis* o *D. paulistorum* lo fueron).

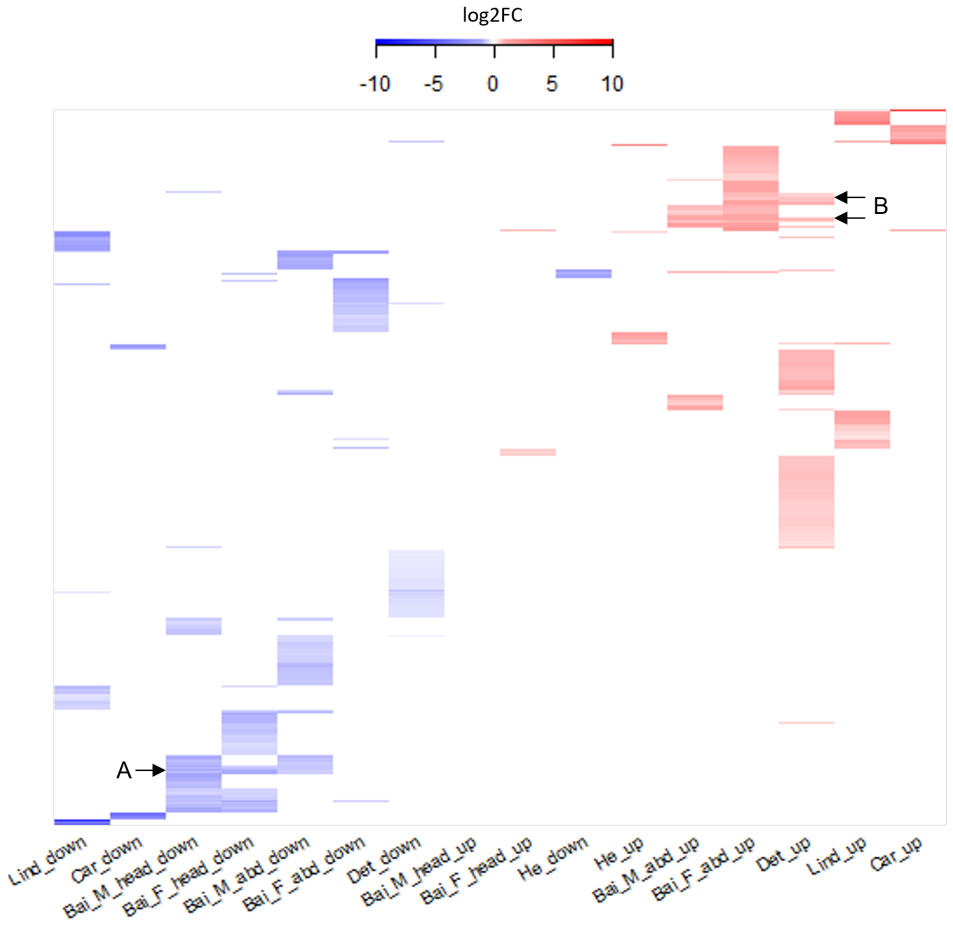

Figura 3.2: *Heatmap* con 1542 ortólogos *D. melanogaster* D.E. por infección *Wolbachia* nativa. Cada columna corresponde a una lista de genes D.E. (nombres en extremo inferior) y cada fila corresponde a un ortólogo de *D. melanogaster* (nombres omitidos). El color de la celda (i, j) representa el log_2FC del ortólogo i-ésimo en la lista j-ésima, según la paleta de colores en la cabeza del *heatmap*. Una celda en blanco indica que un ortólogo no se encuentra en una lista D.E. Las zonas A y B indican grupos de genes D.E. en múltiples listas.

La Figura 3.2 muestra 1542 genes *D. melanogaster* diferencialmente expresados (o con ortólogos D.E.) según al menos una lista, esto es, alrededor de un 10 % de todos los genes codificantes de proteínas en *D. melanogaster* [66], sugiriendo que *Wolbachia* puede influir en numerosos ámbitos de la biología de sus hospederos nativos.

La mayoría de los elementos diferencialmente expresados lo son exclusivamente en una lista. Excepciones a dicho patrón están dadas principalmente por genes que resultaron D.E. en más de una comparación realizada por Baiao *et al.* (2019), como los indicados en las zonas A y B de la Figura 3.2 (este último punto también muestra intersección con los genes D.E. provenientes de Detcharoen *et al.* (2021)). La falta de intersección entre listas sub- y sobreexpresadas provenientes de la misma comparación no es extraña, pues sólo podría darse en una situación bastante específica. Que un mismo gen apareciera sub- y sobreexpresado sólo pudo darse en aquellos casos en que la obtención de las listas fuese mediada por una búsqueda de ortólogos, esto es, en la obtención desde Caragata *et al.* (2017) o Baiao *et al.* (2019). Además, se requeriría que dos elementos bastante similares y con log_2FC de distinto signo fueran asociados por ortología al mismo gen del respectivo organismo de destino.

Más llamativa es la escasa intersección entre listas provenientes de comparaciones distintas, para lo cual se pueden considerar múltiples motivos de índole biológica y técnica. Un buen ejemplo del efecto de factores biológicos distintos del estado de infección *Wolbachia* sobre la identidad de los genes D.E., se extrae de la comparación que realizó Baiao *et al.* (2019) entre las listas obtenidas para las tres semiespecies de *D. paulistorum* y para distintas combinaciones de sexo y tejido. Por una parte, los autores encontraron que la mayoría de los genes diferencialmente expresados, lo fueron exclusivamente en una de las semiespecies [31]. Por otra parte, para la semiespecie OR, se halló que el 93 % de los genes subexpresados en abdómenes de machos no aparecieron diferencialmente expresados en ninguna otra condición [31]. Finalmente, un PCA realizado por los autores indicó que la semiespecie, el tejido y el sexo fueron más influyentes sobre los perfiles transcriptómicos que el estado de infección *Wolbachia* [31]. En cuanto los resultados anteriores fueron obtenidos mediante una misma metodología (lo cual supone cierto control con respecto a fuentes de variabilidad técnicas), estos sugieren que diferencias con respecto a factores biológicos como tejido, sexo y genética hospedera o bacteriana pueden provocar, por sí mismas, una gran discrepancia entre las identidades específicas de los genes D.E.

En particular, por provenir del mismo origen en términos de sexo, especie hospedera y cepa *Wolbachia*, resulta fundamental hallar sentido a la falta de intersección entre las listas provenientes de He *et al.* (2019), Lindsey *et al.* (2021) y Detcharoen *et al.* (2021), representada mediante los diagramas de Venn de la Figura 3.3.

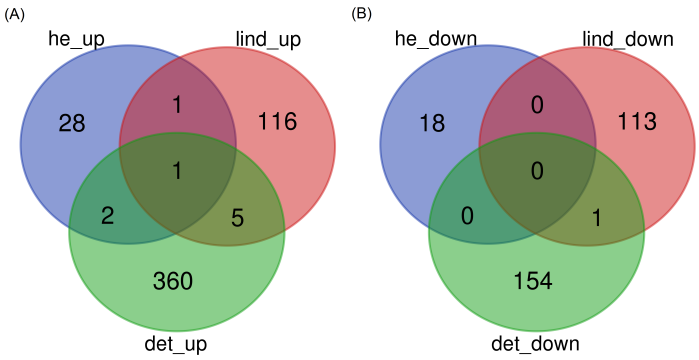

Figura 3.3: Diagramas de Venn representando la intersección de listas obtenidas desde He *et al.* (2019), Lindsey *et al.* (2021) y Detcharoen *et al.* (2021). (A) Intersección de listas sobreexpresadas. (B) Intersección de listas subexpresadas. Los números en cada área denotan la cardinalidad del subconjunto representado por esta

Evidentemente las listas son muy distintas estre sí. Cabe señalar que al reemplazar las listas He_up y He_down por las publicadas originalmente por He *et al.* (2019), el nivel de intersección con el resto de las listas sigue siendo mínimo, descartando que esta tendencia haya sido introducida por el cambio metodológico.

La diferencia entre las listas provenientes de He *et al.* (2019) y las de los otros dos estudios se puede entender como influida por uno de los factores biológicos críticos antes señalados, a saber, el tipo de tejido. Los orígenes de las listas obtenidas desde Lindsey *et al.* (2021) y Detcharoen *et al.* (2021), por su parte, son homogéneas con respecto a tejido, sexo, especie de hospedero y cepa *Wolbachia*: para explicar su inconsistencia cabe barajar otros factores. Por ejemplo, las descripciones de los orígenes experimentales de los datos RNA-Seq (Sección 3.1) dan cuenta de que estos se diferencian con respecto a alimentación, ciclos de luz/oscuridad (16h/8h vs 24h/24h), temperatura (25°C vs 19°C) y, posiblemente, estado de virginidad (sólo en Lindsey *et al.* (2021) se especifica que los individuos muestreados eran vírgenes) [15, 29]. Además, de forma excepcional, la lista de Lindsey *et al.* (2021) proviene de datos de individuos que recibieron inyecciones, bien fuera con suero o con SINV. Se constató en el mismo estudio que el transcriptoma de los individuos varió considerablemente entre tres tiempos posteriores a la inyección con suero, cambio que fue atribuido a una "recuperación" hacia la inyección [15]. Con este antecedente, y dado que en el experimento no se usaron controles sin inyección, no es posible descartar que el simple hecho de recibir una inyección modulara de alguna forma el efecto de *Wolbachia* a nivel de expresión de genes individuales. Finalmente, las descripciones de la Sección 3.1 muestran que los orígenes de las listas se diferencian con respecto a múltiples factores técnicos, como protocolo de enriquecimiento de ARN de interés, tipo de plataforma Illumina usada y elección de programas para el análisis RNA-Seq. Lo anterior podría terminar de explicar las inconsistencias entre las listas de genes D.E. provenientes de los mismos sistemas.

La inconsistencia de los efectos *Wolbachia* a nivel de genes D.E. ha sido previamente notada por autores que han intentado buscar regularidades entre distintos sistemas *Wolbachia*-hospedero. Por ejemplo, en Caragata *et al.* (2017) se compararon las listas de genes D.E. en *A. fluviatilis*-wFlu con las halladas previamente para transfecciones de wMel y wMelPop en *A. aegypti*; mientras que en Baiao *et al.* (2019) se buscaron genes consistentemente D.E. entre las distintas semiespecies de *D. paulistorum* [17, 31]. En ambos casos se halló muy poca

intersección a nivel de genes específicos pero, importantemente, se pudieron identificar regularidades a nivel de familias de genes D.E. o de términos funcionales enriquecidos [17, 31]. Otro ejemplo es el estudio proteómico de Geoghegan *et al.* (2017), donde se halló consistencia entre los ámbitos funcionales perturbados por *Wolbachia* en un línea celular y en el intestino de *A. aegypti*, pero en virtud de la expresión diferencial de distintas proteínas [67]. Así, la disparidad entre las listas de genes D.E. aquí obtenidas, no descartó *a priori* la posibilidad de hallar consistencias entre las funciones afectadas y, por el contrario, resaltó la utilidad de buscar efectos comunes a nivel funcional.

3.3. Enriquecimiento de términos funcionales bajo parámetros estrictos

La identificación de funciones comúnmente afectadas por infecciones *Wolbachia* nativas fue mediada por el análisis de enriquecimiento de términos GO sobre las listas D.E. En el Anexo E se comparan los resultados de los análisis de enriquecimiento propios con parámetros estrictos (ajuste g:SCS y *adjusted p-value* < 0.05) frente a los resultados análogos obtenidos por Baiao *et al.* (2019) y He *et al.* (2019). En general, se halló consenso entre los resultados propios y los publicados, identificándose razones plausibles para las discordancias, esto es: diferencias en la definición del universo de genes, mutabilidad de las anotaciones GO y distintos métodos de ajuste por testeo múltiple. Una comparación similar para los resultados del resto de los análisis no fue posible, dado que estos no tuvieron equivalentes en las publicaciones originales.

En línea con el objetivo de hallar regularidades a nivel del orden Diptera, se centran las discusiones en torno a aquellos resultados que reflejan alteraciones comunes entre las especies *Aedes* y *Drosophila*, descartando aquellas funciones que sólo aparecen alteradas en uno de los dos géneros. Bajo los parámetros estrictos, los únicos términos funcionales seleccionados fueron los asociados con la visión hospedera.

3.3.1. Visión

Tanto en hembras *A. fluviatilis* completas como en cabezas de hembras *D. paulistorum* se halló una afectación importante de funciones asociadas con la visión hospedera. Específicamente, en la lista Car_down se halló un enriquecimiento de los procesos biológicos *Detection of light stimulus*, *Response to light stimulus*, *Phototransduction*, *Visual perception* y *Protein-chromophore linkage*, así como de las funciones moleculares *Photoreceptor activity* y *G protein-coupled receptor activity*. Por otra parte, consistentemente con el análisis de enriquecimiento original de Baiao *et al.* (2019), en la lista Bai_F_head_up se halló un enriquecimiento de los procesos biológicos *Deactivation of rhodopsin mediated signaling*, *Rhabdomere development*, *Visual perception* y *Photoreceptor cell maintenance*. Las alteraciones fueron notadas y comentadas brevemente por los autores de las respectivas publicaciones originales, Caragata *et al.* (2017) y Baiao *et al.* (2019), quienes las catalogaron como efectos novedosos y consistentes con la conocida presencia de *Wolbachia* en las cabezas de *A. fluviatilis* y *D. paulistorum* [17, 31].

No se halló en la literatura evidencia directa de la afectación de funciones visuales por parte de *Wolbachia* en otros hospederos, aunque esto no descarta su posibilidad ya que no se

encontraron estudios dedicados a interrogar estos efectos, y bien podrían haber sido pasados por alto en análisis ómicos realizados sobre cuerpos completos [15, 29, 68], tejidos distintos a cabezas [30, 67] o líneas celulares [16, 67, 69, 70, 71, 72]. No sería sorprendente la manifestación de efectos visuales en otros hospederos cuyas cabezas son habitadas por *Wolbachia*, como *Culex pipiens*, *Drosophila simulans*, *Glossina austeni* y *D. melanogaster* (incluyendo lóbulo óptico y retina de este último) [73]. En *D. melanogaster* y *D. simulans* se han observado efectos de infecciones nativas sobre el olfato, sentando un precedente en estas especies para la afectación de funciones sensoriales [74]. La proximidad de *Wolbachia* a los tejidos involucrados en la recepción y procesamiento de señales lumínicas habilitaría una variedad de interacciones locales que podrían originar efectos sobre la visión hospedera, sin la necesidad de un mecanismo sistémico.

Naturalmente, ante la falta mencionada de estudios sobre los efectos de *Wolbachia* en la visión, tampoco está claro si dichos efectos se manifiestan en alguna medida en los sistemas transfectados. Un indicio sugestivo, sin embargo, proviene de una evaluación de los efectos de *Wolbachia* sobre la competitividad de larvas *A. aegypti* transfectadas, que mostró que estas demoraron un tiempo significativamente mayor en alejarse de zonas luminosas que sus contrapartes sin infección [75]. Tal como en individuos adultos, la percepción visual de las larvas *A. aegypti* depende de la expresión de rodopsinas, esto es, proteínas pertenecientes a la familia de receptores asociados a proteínas G, que inician cascadas de señalización inducidas por fotones [76, 77]. Interesantemente, cinco ortólogos a rodopsinas (*gprop1*, *gprop2*, *gprop4*, *gprop8* y *gprop9*) forman parte de los genes subexpresados que provocaron los enriquecimientos de términos GO para *Aedes fluviatilis* hallados en este trabajo. Un efecto análogo en larvas *Aedes aegypti* podría explicar su respuesta atenuada a los estímulos lumínicos.

La posibilidad de que existan efectos visuales de *Wolbachia* sobre hospederos transfectados −actuales o futuros− amerita ser seriamente evaluada, dado que podrían tener impactos mayores para las estrategias de control. Por ejemplo, la lentitud de la migración de larvas *A. aegypti* transfectadas ante estímulos lumínicos podría significar un mayor riesgo de depredación [75], lo cual podría traducirse en una mayor tasa de mortalidad para individuos transfectados y, por lo tanto, una resistencia al reemplazo de poblaciones.

También se ha determinado que el reconocimiento de objetivos humanos a corta distancia por parte de *A. aegypti* depende de su visión, habiéndose descrito que la mutación conjunta de las rodopsinas *gprop1* y *gprop2* impide dicho mecanismo [78, 79]. De lo anterior se sigue que, si una subexpresión de *gprop1* y *gprop2* como la detectada en *A. fluviatilis*-wFlu se manifestara también en *A. aegypti* transfectado, este último podría tener menor contacto efectivo con las personas, lo cual podría tener al menos dos consecuencias sobre los riesgos de contraer enfermedades arbovirales. Por una parte, el riesgo de transmisión de arbovirus desde mosquitos transfectados (que, a pesar del bloqueo patogénico, no es nulo [15]) podría disminuir en virtud del contacto reducido. Por otra parte, la menor capacidad de los mosquitos transfectados para obtener sangre −alimento necesario para la ovogénesis [80]− podría situarlos en desventaja competitiva con respecto a mosquitos normales, tendiendo a la purga de la bacteria de las poblaciones. El beneficio neto que ambos efectos en conflicto tendrían sobre el control de las enfermedades arbovirales podría depender de factores adicionales, como la efectividad de la incompatibilidad citoplasmática para contrarrestar los costos de aptitud, o la intensidad del bloqueo patogénico, que dictaría el riesgo de trasmisión de

determinados virus por parte de los individuos transfectados. En cualquier caso, queda claro que las consecuencias podrían ser significativas.

Finalmente, la afectación visual podría tener consecuencias sobre el comportamiento reproductivo de mosquitos transfectados. Tal como en *Drosophila*, se entiende que el comportamiento de *A. aegypti* en torno al apareamiento es influido por pistas visuales: una alteración en la forma de recibirlas y procesarlas podría significar la modificación de las preferencias de apareamiento [31, 81, 82]. Nuevamente, esto podría tener consecuencias de peso para las estrategias de control. En efecto, con base en simulaciones de dinámica de poblaciones, se ha predicho que la manifestación de preferencias de apareamiento en función del estado de infección *Wolbachia* podría, por lo bajo, imponer requerimientos sobre el tamaño o la frecuencia requerida en las liberaciones de mosquitos [82, 83].

3.4. Enriquecimiento de términos funcionales bajo parámetros relajados

El análisis de enriquecimiento de términos GO bajo los parámetros iniciales fue poco informativo para varias de las listas, con pocos o ningún término enriquecido en Car_up, Lind_up, Lind_down, He_up y He_down. Así, se decidió hacer un segundo análisis con parámetros menos estrictos, a saber, método de ajuste Benjamini-Hochberg y *adjusted p-value* de corte de 0.20. Nuevamente, los términos fueron seleccionados de acuerdo a su enriquecimiento en múltiples especies. El patrón de expresión de los términos funcionales a través de las listas fue representado mediante nodos, cuya estructura se precisa en la Figura 3.4.

Los términos funcionales similares fueron agrupados en módulos funcionales, de la forma especificada en la Metodología. El grosor de las aristas entre términos se correlaciona con la similaridad de estos. Cada módulo funcional se encuentra asociado con un *heatmap* adjunto en el Anexo F, que muestra los genes específicos que produjeron los enriquecimientos. En las discusiones se dirá que una función T está sobre- o subregulada por *Wolbachia* en un tejido, si es que hay un enriquecimiento de T en la lista de genes sobre- o subexpresados en dicho tejido, respectivamente. También, se dirá que un gen tuvo una sobre- o subexpresión importante si se cumple que $log_2FC > 1$ o $log_2FC < -1$, respectivamente. Finalmente, al referirse a resultados bibliográficos adicionales, la expresión diferencial detectada a nivel proteómico será distinguida de la detectada a nivel transcriptómico, denotando la primera como "abundancia diferencial" (sub- o sobreabundancia, a falta de mejores términos).

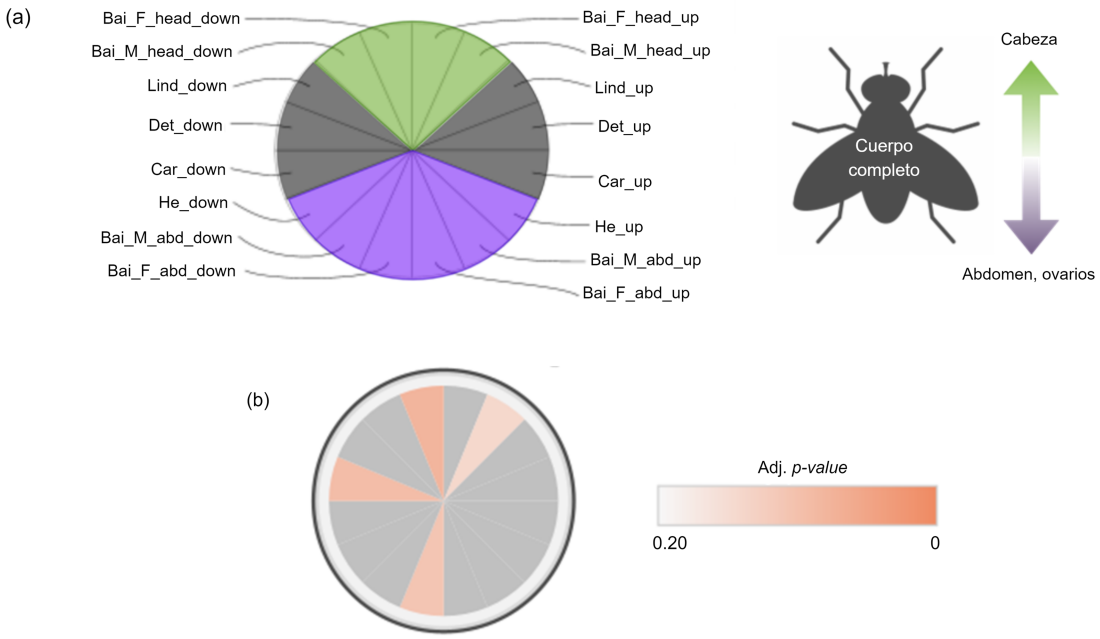

Figura 3.4: Estructura de un nodo EnrichmentMap, específica para este estudio. El nodo expresa el nivel de enriquecimiento de un término GO en cada lista D.E. (a) Ubicaciones asignadas a las listas. A la derecha se propone una mnemotecnia: las posiciones altas de los nodos (verde) corresponden a comparaciones de cabezas, las posiciones bajas (azul) a abdómenes/ovarios, y las posiciones medias (negro) a cuerpos completos. Las listas sub- y sobre-expresadas ocupan la mitad izquierda y derecha de los nodos, respectivamente. (b) Relación entre color y *adj. p-value* del enriquecimiento en una lista. El ejemplo representa el enriquecimiento moderado de un término en Bai_M_head_up, Bai_F_head_down, Det_down y Bai_F_abd_down.

3.4.1. Glicosiltransferasas/hidrolasas y unión/hidrólisis de quitina

En las Figuras 3.5 y 3.6 se presentan los módulos funcionales titulados *Actividad glicosil hidrolasa e hidrólisis/unión de quitina* y *Actividad glicosil transferasa*, respectivamente.

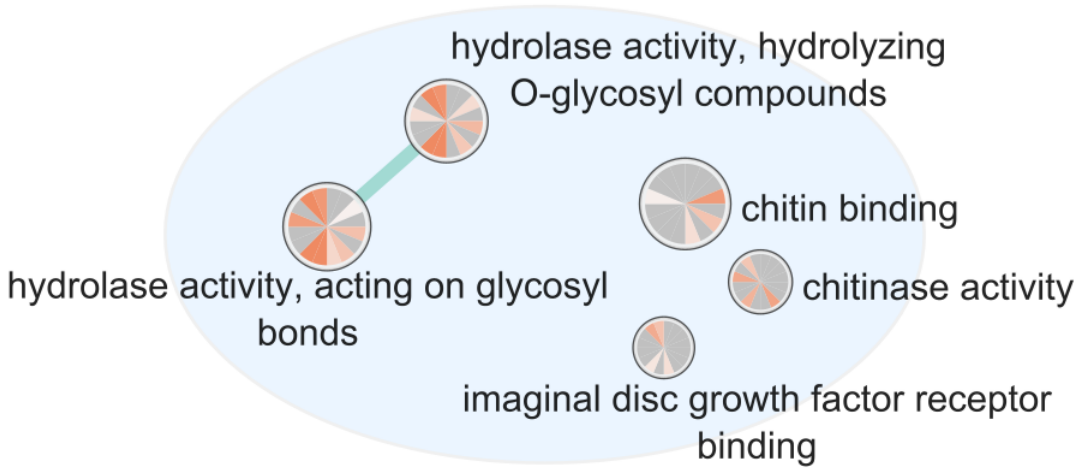

Figura 3.5: Módulo *Actividad glicosil hidrolasa e hidrólisis/unión de quitina.*

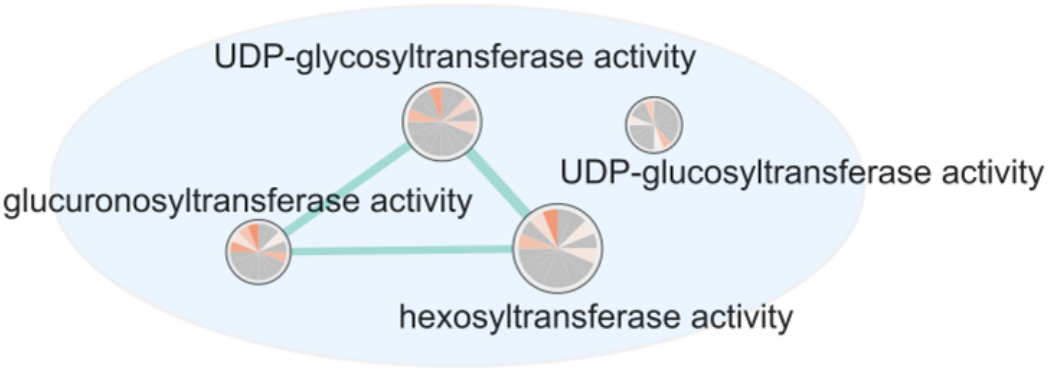

Figura 3.6: Módulo *Actividad glicosil transferasa.*

La hidrólisis de enlaces glicosídicos (principalmente O-glicosídicos) es comúnmente afectada por infecciones nativas, hallándose reguladas en las tres especies Diptera estudiadas; en cabezas, abdómenes y cuerpo completo; e incluyendo a machos.

Los enlaces glicosídicos se establecen entre un sacárido y otra molécula, que puede o no ser un sacárido, hallándose en una enorme variedad de compuestos como azúcares simples y complejos, glicoproteínas, glicolípidos, proteoglucanos, nucleótido-azúcares y otros [84]. Igualmente variados son los genes con función glicosil hidrolítica según el *heatmap* anexo en la Figura F.1, que incluyen lisozimas (*lysd*, *lyse*, *lysp*), amilasas (*amyrel*), quitinasas (*cht4*, *cht8*), maltasas (*mal-a4*, *mal-b2*), galactosidasas (*gal*), manosidasas lisosomales (*lmani*, *lmanii*, *lmaniii*, *lmaniv*), enzimas involucradas en la ramificación del glucógeno (*agbe*), y otros. La poca consistencia entre los genes D.E. a través de las listas, más la gran diversidad de procesos biológicos a los que pueden asociarse, hacen difícil proponer un relato común para las alteraciones de la hidrólisis de enlaces glicosídicos. Un aspecto a destacar, sin embargo, es que el enriquecimiento de la actividad glicosil hidrolasa en una lista según la Figura 3.5 frecuentemente viene acompañada de un enriquecimiento de la actividad glicosil transferasa en la misma lista según la Figura 3.6. Donde las glicosil hidrolasas se encargan de romper enlaces glicosídicos, las glicosil transferasas se encargan de generarlos, catalizando la transferencia de un grupo glicosil preactivado por factores como uridina difosfato (UDP), a aceptores como azúcares, lípidos, proteínas y ácidos nucleicos [85]. El fenómeno de corregulación descrito, que se manifiesta en muestras de las tres especies estudiadas, podría reflejar manipulaciones concertadas de *Wolbachia* o bien una regulación hospedera tendiente a compensar efectos de *Wolbachia* sobre uno de los dos ámbitos funcionales.

Una referencia más específica en la Figura 3.5 es hacia las funciones de unión e hidrólisis de quitina (en inglés *chitin*), un polisacárido conformado por residuos de N-acetilglucosamina unidos por enlaces β-1,4 O-glicosídicos que, en mosquitos, es un componente central del exoesqueleto, de revestimientos cuticulares internos como el de la tráquea o los ductos salivales, de la matriz peritrófica que cubre el epitelio intestinal y de la cutícula serosa que cubre a los huevos [86, 87, 88]. En principio, la Figura 3.5 sugiere una regulación de actividades asociadas a quitina exclusivamente en las especies *Drosophila*. Cabe destacar, sin embargo, que aunque no fue suficiente para presentar un enriquecimiento en dicho ámbito funcional, en *A. fluviatilis*-wFlu se halló una sobreexpresión importante de los ortólogos AAEL003066 (unión a quitina, hidrólisis de enlaces O-glicosídicos) y AAEL004931 (β-N-acetilhexosaminidasa), así como una subexpresión importante de AAEL009528 (unión y metabolismo de quitina).

En un estudio transcriptómico adicional sobre testículos larvales *D. melanogaster*-wMel se detectó sobreexpresión de tres genes con actividad de unión a quitina (*cg3348*, *cg5210* y *cpr49a*) y subexpresión de otro gen con la misma actividad (*cg13806*) [89], respaldando la afectación de funciones asociadas con la quitina en machos *Drosophila* y sugiriendo que esta podría manifestarse desde etapas tempranas del desarrollo.

Alteraciones en funciones asociadas con la quitina también son observadas en sistemas transfectados. Por ejemplo, en hembras completas *A. aegypti*-wMelPop se observó una sobreexpresión importante de tres quitinasas (AAEL003066, AAEL002959 y AAEL002969) y de una quitina sintasa (AAEL005618) [68]. Un estudio proteómico en intestinos de *A. aegypti*-wMel también mostró una sobreabundancia de AAEL003066, así como de una quitinasa adicional (AAEL013262) [67]. En hembras *Drosophila nigrosparsa* transfectadas con wMel se halló la sobreexpresión de una quitinasa (*cht2*) y la expresión diferencial (principalmente subexpresión) de varios genes anotados como constituyentes estructurales de la cutícula larval basada en quitina y con actividad predicha de unión a quitina (*cpr62bb*, *cpr62bc*, *cpr92f*, *cpr100a*, *twdlv*, *twdly* y *pcp*) [29]. Cabe notar que en los sistemas transfectados sólo se observaron aumentos de la actividad quitinasa hospedera, a diferencia de lo ocurrido para sistemas nativos, lo cual evoca nuevamente cierta tendencia adaptativa. Ilustrativamente, donde hembras *D. melanogaster*-wMel mostraron una subregulación de la actividad quitinasa según la Figura 3.5, las hembras *D. nigrosparsa* y *A. aegypti* transfectadas con la misma cepa mostraron sobreexpresión de 1 y 2 quitinasas, respectivamente.

Sólo recientemente se ha reparado en la importancia que la interacción entre *Wolbachia* y la quitina hospedera podría tener para las estrategias de control de vectores, proponiéndose que podría estar implicada en el bloqueo patogénico [88] y en la potenciación de infecciones secundarias por *Plasmodium* [17]. Acá se propone una razón adicional para estudiar este fenómeno.

Actualmente, el principal efecto perjudicial de las transfecciones wMel y wAlbB sobre *A. aegypti* es la disminución de la supervivencia de los huevos quiescentes, la cual se ha asociado al secuestro generalizado de nutrientes como aminoácidos y lípidos por parte de la bacteria [90]. Una componente importante de la supervivencia de huevos quiescentes en mosquitos es la resistencia a la desecación [90, 91], la cual se correlaciona fuertemente con el contenido de quitina en la cutícula serosa de los huevos [92]. Se puede especular que la interacción entre *Wolbachia* y la quitina, que parece caracterizarse por un aumento de la actividad quitinasa en hospederos transfectados, induce alguna deficiencia en la cutícula serosa, disminuyendo la resistencia a la desecación en huevos de *A. aegypti* transfectado. Si el anterior fuera el caso, un entendimiento mecanístico de la interacción podría revelar algún blanco para manipulación genética dirigida a eliminar el efecto perjudicial sobre los huevos (en particular, estudiar el fenómeno en sistemas nativos podría ilustrar la forma en que una infección puede persistir sin causar tal efecto).

El que *Wolbachia* parezca afectar actividades asociadas a quitina en todo tipo de hospedero, sugiere que puede haber una característica universal de la bacteria implicada. Resulta interesante seguir esta premisa en el contexto de los requerimientos de *Wolbachia* hacia la N-acetilglucosamina, el monómero constituyente de la quitina. La N-acetilglucosamina activada por UDP (uridina difosfato) es un intermediario en la síntesis del lípido II, un compuesto que formaría parte de una estructura reminiscente del péptidoglucano en *Wolbachia* y que ha

mostrado ser fundamental para la coordinación de su división celular [93, 94]. A pesar de su esencialidad, no está claro si *Wolbachia* tiene la capacidad de sintetizar N-acetilglucosamina *de novo* a partir de fructosa 6-P (como sí la tienen otras bacterias extracelulares [95, 96]), o si debe conseguirla desde su hospedero. Las referencias halladas sobre el uso de UDP-N-acetilglucosamina por parte de la bacteria no especifican la manera en que esta la obtiene inicialmente [93, 94, 97]. Por otra parte, todos los modelos metabólicos de cepas *Wolbachia* hallados en la base de datos KEGG carecen de genes necesarios tanto para la síntesis *de novo* (EC 2.3.1.4 y EC 5.4.2.3, o bien EC 5.4.2.10) como para la importación de N-acetilglucosamina [96]. Un antecedente sugestivo proviene de *Sodalis glossinidius*, un endosimbionte de la mosca Tse-tse que libera N-acetilglucosamina hospedera mediante quitinasas y la importa para sus propias actividades metabólicas [98]. Determinar si *Wolbachia* posee o no los genes necesarios para la síntesis y/o importación de N-acetilglucosamina se propone como una primera tarea para comprender las bases de la interacción entre *Wolbachia* y la quitina.

3.4.2. Unión a iones de calcio

En la Figura 3.7 se presenta el módulo titulado *Unión a iones de calcio*.

Figura 3.7: Módulo *Unión a iones de calcio*.

La unión a iones de calcio es una función molecular comúnmente afectada por *Wolbachia*, hallándose regulada en las tres especies estudiadas; en cabezas, abdómenes y cuerpos completos; e incluyendo a machos. No se distingue una tendencia clara con respecto a los sentidos de regulación.

Se hallaron algunos indicios adicionales de la afectación de funciones asociadas con el calcio en sistemas nativos. Testículos larvales de *D. melanogaster*-wMel presentan subexpresión de *cg31958* (con dominio predicho de unión a calcio) y la sobreexpresión de *cg4535* (regulación del transporte de calcio) [99]. También, según Baiao *et al.* (2019), un gen codificante de un canal de calcio (*ryr*) fue el único con ortólogos sobreexpresados en las tres semiespecies *D. paulistorum* naturalmente infectadas [31].

La afectación de funciones asociadas con iones de calcio también se ha observado en sistemas transfectados, nuevamente mostrando patrones de regulación complejos. En el experimento proteómico de Geoghegan *et al.* (2017) sobre intestinos de hembras *A. aegypti*-wMel, se detectó una sobreexpresión de tres proteínas con actividad de unión a calcio predicha (AAEL006095, AAEL005463 y AAEL001020) y la subexpresión de un transportador de calcio (AAEL005561) [67]. En un estudio transcriptómico independiente sobre hembras completas *A. aegypti*-wMel no se halló mayor efecto sobre funciones asociadas con el calcio [68], lo cual podría responder a diferencias en las metodologías de ambos estudios (en particular, la tejido-especificidad), o bien reflejar una diferencia real entre los estados funcionales. Tam-

bién, en *A. aegypti*-wMelPop se halló sobre- y subexpresión de distintos canales y sensores de calcio, así como de intercambiadores calcio/potasio [68].

Interesantemente, la manifestación de efectos asociados con calcio por parte de *Wolbachia* no parece requerir de un contexto orgánico. Por ejemplo, en el estudio proteómico de Baldridge *et al.* (2017) sobre una línea celular de *A. albopictus* transfectada con wStr, se detectó la sobre- y subexpresión de dos ATPasas asociadas con transporte de calcio (AAEL006582 y AAEL005561, respectivamente), así como la sobreexpresión de una proteína con función hipotética de captación de calcio en mitocondria (AAEL010116) [72]. En células de *Anopheles gambiae* con infecciones artificiales wAlbB y wRi se hallaron consistentemente subexpresados cinco genes con actividad de unión a calcio (AGAP005032, AGAP009528, AGAP012067, AGAP001177 y AGAP008822) [71]. Finalmente, en células de *Aedes albopictus* transfectadas con la cepa wStr se halló sub- y sobreabundancia de dos y una bomba de calcio, respectivamente [72].

Sorprendentemente, en las publicaciones consultadas se presta poca o ninguna atención a los efectos de *Wolbachia* asociados al calcio, tanto en sistemas transfectados como nativos [17, 29, 30, 31, 67, 68, 71, 72, 99]. La señalización por calcio está involucrada en procesos fundamentales para la vida de los mosquitos, incluyendo la proliferación celular, la secreción de sustancias, la apoptosis, la respuesta inmune innata, el vuelo y la percepción sensorial [100, 101, 102]. Así, es fácil percibir la importancia de desarrollar conocimiento mecanístico sobre la interacción entre *Wolbachia* y el calcio en sistemas transfectados, así como posibles cambios de esta interacción luego de un proceso de coevolución.

Sin considerar la dirección específica de la regulación, el que las actividades de unión de calcio sean comúnmente afectadas en sistemas tan diversos (infecciones nativas y transfecciones *in vivo* e *in vitro*) podría ser reflejo de interacciones necesarias para la vida intracelular de *Wolbachia*. Se postula que uno de estos podría ser la interacción entre *Wolbachia* y el retículo endoplásmico. Se ha descrito que *Wolbachia* obtiene su membrana vacuolar externa desde el retículo endoplásmico, proceso que implica una interacción directa con el organelo y deriva en la modificación de su composición y organización [67, 103, 104]. Dado que el retículo endoplásmico es un reservorio intracelular de iones de calcio, es razonable suponer que esta interacción perturba en alguna medida los gradientes naturales de este elemento. A su vez, dado lo crítico de la señalización por calcio para los procesos celulares, la perturbación al retículo endoplasmático podría provocar una respuesta transcripcional compensatoria. La existencia de tal respuesta puede hallar respaldo en cuanto se ha visto que la subversión del retículo endoplasmático por parte de *Wolbachia* puede ocurrir sin gatillar las respuestas típicas del estrés en este organelo, como la respuesta a proteínas mal plegadas o la apoptosis [103]. No sería extraño que efectores *Wolbachia* estuvieran involucrados también en el control de los efectos causados por la perturbación del retículo endoplásmico, añadiendo complejidad a la interpretación de los patrones de regulación transcripcional.

Una última observación puede servir como pivote para comenzar a desentrañar el fenómeno a nivel transcripcional. El *heatmap* anexo (Figura F.3) muestra que el gen *regucalcin* fue diferencialmente expresado en *D. paulistorum* y *D. melanogaster*. Según la misma Figura, *regucalcin* aparece simultáneamente sobre- y subexpresado en *A. fluviatilis*, lo cual refleja la sobre- y subexpresión de los ortólogos AAEL000757 y AAEL001020, respectivamente. In-

teresantemente, Geoghegan et al. (2017) también halló una sobreabundancia de AAEL001020 [67]. *Regucalcin* es un agente clave en la regulación de la concentración de calcio intracelular, controlando la actividad de canales y bombas de calcio en distintas membranas, así como de enzimas dependientes de calcio [101, 102]. Importantemente, la transcripción de *regucalcin* es inducida por la presencia de calcio [101], lo cual es consistente con que la expresión diferencial tenga un origen común en la perturbación de los gradientes de calcio. Dados los indicios de que *regucalcin* tiene una participación común en la interacción entre *Wolbachia* y el calcio hospedero, se propone como un blanco de futuras investigaciones.

3.4.3. Serina- y metalopeptidasas

En las Figuras 3.8 y 3.9 se presentan los módulos funcionales titulados *Actividad metalopeptidasa* y *Actividad serina peptidasa*, respectivamente.

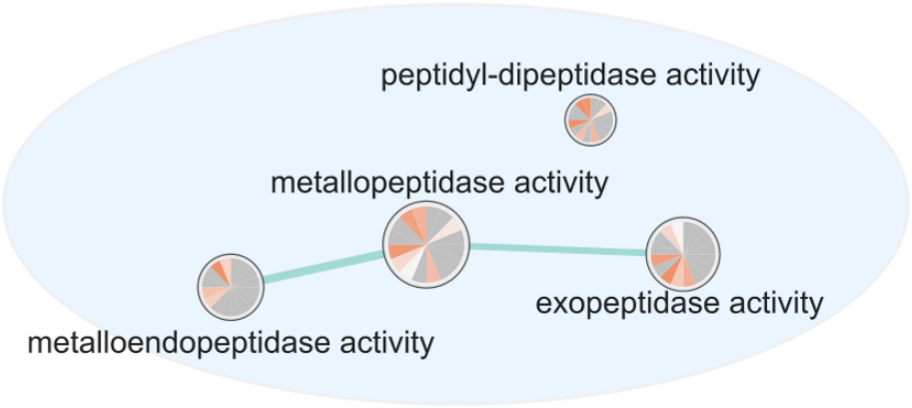

Figura 3.8: Módulo *Actividad metalopeptidasa*.

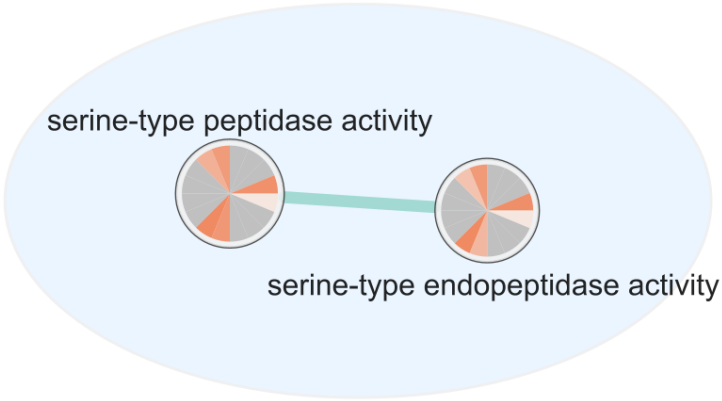

Figura 3.9: Módulo *Actividad serina peptidasa*.

Las actividades serina- y metalopeptidasas fueron reguladas por infecciones nativas en las tres especies estudiadas; en cabezas, abdómenes, ovarios (sólo metalopeptidasa) y cuerpos completos; e incluyendo a machos. La regulación de la actividad metalopeptidasa es principalmente en sentido negativo, mientras que la actividad serina peptidasa se halla subregulada en tejidos de *D. paulistorum* y sobrerregulada en cuerpos completos de *D. melanogaster* y *A. fluviatilis*. En un estudio adicional sobre testículos larvales *D. melanogaster*-wMel se halló

sobre- y subexpresión de las metalopeptidasas *cg4408* y *cg32454*, respectivamente; además de sobreexpresión de *cg30091*, *cg8586*, *cg18563* y *cg3589*, y subexpresión de *cg17571*, *cg13527* y *cg30375* (todas serina-peptidasas) [99]. Lo anterior respalda la afectación de ambas actividades peptidasas en machos *Drosophila* y sugiere que también podrían manifestarse en etapas tempranas. Interesantemente, en el mismo estudio se halló sobreexpresión de dos inhibidores de serina peptidasas (*cg31778* y *cg16704*) [99], plantando una duda sobre la medida en que la sobreexpresión de serina-peptidasas se traduce en un aumento efectivo de la actividad. En total puede extraerse que, si bien la regulación de las actividades peptidasa ocurren en ambos sentidos en los sistemas nativos estudiados, hay cierta tendencia hacia la subregulación, especialmente para la actividad metalopeptidasa.

La regulación de ambas actividades peptidasas también se observa en sistemas transfectados, aunque en un sentido marcadamente positivo. Específicamente, en intestinos de hembras *A. aegypti*-wMel, Geoghegan *et al.* (2017) detectó una clara sobrerregulación de las actividades peptidasas, reflejada en la sobreabundancia de tres metalopeptidasas y siete serina peptidasas [67]. Este resultado es secundado por un análisis transcriptómico sobre hembras completas *A. aegypti*-wMel, donde se halló una sobreexpresión de dos metalopeptidasas y más de diez serina-peptidasas [68]. La sobrerregulación de metalo- y serina-peptidasas en *A. aegypti*-wMelPop resultó ser aún más acentuada según el mismo estudio [68].

Los patrones de expresión de las actividades metalo- y serina-peptidasa en sistemas nativos sugieren que las sobrerregulaciones observadas en transfecciones pueden mutar durante la coevolución *Wolbachia*-hospedero. Una característica general de las cepas *Wolbachia* es que no tienen la capacidad metabólica para sintetizar ciertos aminoácidos, por lo que dependen de la importación de aminoácidos hospederos para sostener su crecimiento [105, 106]. La sobrerregulación de las actividades peptidasas hospederas en sistemas transfectados podrían servir al crecimiento de *Wolbachia*, mas este no sería un mecanismo necesario pues la bacteria dispone de peptidasas propias [105, 106]. La aparente inversión del patrón de regulación de metalopeptidasas en sistemas nativos con respecto a transfectados, podría reflejar el desarrollo de una respuesta hospedera tendiente a contrarrestar la actividad peptidasa de origen bacteriano.

Los efectos de *Wolbachia* sobre la proteólisis han recibido importante interés científico, habiéndose asociado con fenómenos cruciales para las estrategias de control de vectores, como el bloqueo patogénico [104] y la disminución de la supervivencia de huevos *A. aegypti* [90]. Aquí se propone, por lo tanto, la importancia de prever un posible cambio en los efectos proteolíticos observados en transfecciones.

3.4.4. Monooxigenasas

En la Figura 3.10 se presenta el módulo funcional titulado *Actividad monooxigenasa*.

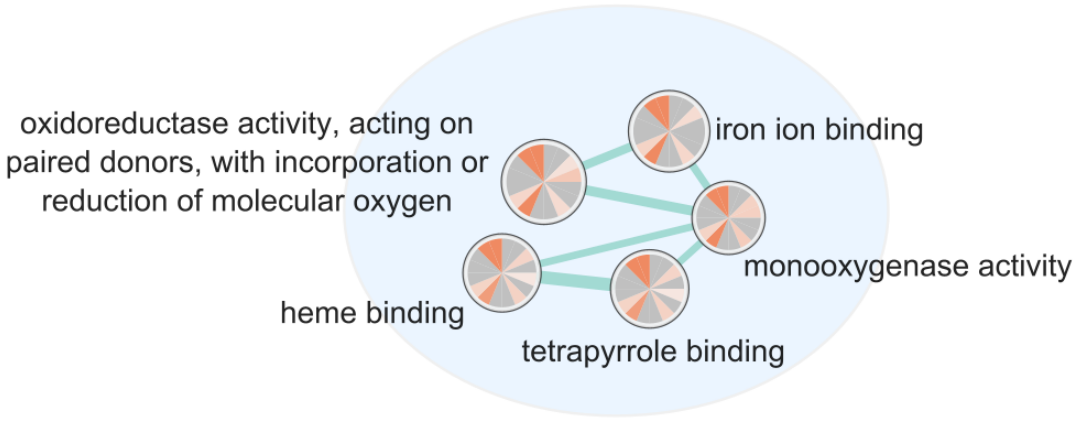

Figura 3.10: Módulo *Actividad monooxigenasa*.

Existió una común regulación de las actividades de unión a iones de hierro, grupo hemo y tetrapirrol, así como de la actividad monooxigenasa (óxidorreductasa que implica la incorporación de un átomo de oxígeno proveniente de O_2 a un compuesto [107]).

Que todos los términos estén conectados con aristas y posean patrones de enriquecimiento parecidos, sugiere que estos son invocados por conjuntos de genes similares. Lo anterior se explica fácilmente para las funciones de unión a hemo, tetrapirrol y hierro: el grupo hemo es un cofactor conformado por un tetrapirrol cíclico con un ión de hierro en su centro [108]. Por otra parte, el *heatmap* anexo (Figura F.6) sugiere que los conjuntos de genes que producen simultáneamente enriquecimiento de las funciones de unión y de la actividad monooxigenasa están constituidos en buena parte por citocromos P450 (identificables por el prefijo *cyp*). Los citocromos P450 son una amplia familia de proteínas dependientes de hemo, con actividad monooxigenasa e implicadas en diversas funciones biológicas como transporte de oxígeno, señalización, metabolismo de xenobióticos, transducción energética, biosíntesis de lípidos y regulación génica [108]. Aunque no fue suficiente para provocar un enriquecimiento de la actividad monooxigenasa, el *heatmap* anexo (Figura F.6) muestra que hubo una sobreexpresión de tres citocromos P450 en hembras *A. fluviatilis*, dos de ellos con $log_2FC > 1$.

En testículos larvales de *D. melanogaster* con infección nativa wMel, Zheng *et al.* (2011) detectó la sobreexpresión de cinco citocromos P450 (*cyp9b2*, *cyp4s3*, *cyp6d5*, *cyp12d1-p*, *cyp6g1*) [99], apoyando la existencia de un efecto en machos y sugiriendo que este se manifiesta en etapas tempranas del desarrollo hospedero. Más aún, este último resultado refuerza la tendencia en *D. melanogaster*-wMel hacia la sobrerregulación, reflejada en la Figura 3.10 con el enriquecimiento de las listas Lind_up y Det_up. Una excepción a dicha tendencia parece estar dada por los resultados de ovarios *D. melanogaster*-wMel, pues para estos la Figura 3.10 indica subregulación (enriquecimiento de lista He_down). Sin embargo, hay razones para cuestionar este último resultado. En efecto, el *heatmap* de la Figura F.6 muestra que el enriquecimiento fue gatillado por la subexpresión de un sólo gen, *cyp6a13*. Además, si bien el análisis de expresión diferencial original por He *et al.* (2019) concuerda con la subexpresión de *cyp6a13*, también reporta la sobreexpresión de otros cuatro citocromos P450 [30]. Así,

las tendencias de regulación de citocromos P450 parecen ser eminentemente positivas en *D. melanogaster* y *A. fluviatilis*, y negativas en *D. paulistorum*.

Las transfecciones también exhiben regulación de citocromos P450. Por ejemplo, un efecto dramático de la transfección con wMelPop en hembras *A. aegypti* fue detectado por Rances *et al.* (2012), con la sobre- y subexpresión de 45 y 20 citocromos P450, respectivamente [68]. En el mismo estudio se detectó una regulación más moderada en hembras *A. aegypti*-wMel, con la sobre- y subexpresión de 7 y 2 citocromos P450, respectivamente [68]. En contraste con este último resultado, en intestinos de hembras *A. aegypti*-wMel, Geoghegan *et al.* (2017) encontró una regulación principalmente negativa, con la sub- y sobreeexpresión de 8 y 1 citocromos P450, respectivamente [67].

La alteración de la expresión de citocromos P450 no parece requerir de un contexto orgánico. Por ejemplo, Baldridge *et al.* (2012) detectó la sobre- y subabundancia de 3 y 1 citocromo P450 en células de *A. albopictus* transfectadas con wStr, respectivamente [72]. Por otra parte, Geoghegan *et al.* (2017) encontró sobreabundancia de siete citocromos P450 en células de *A. aegypti* transfectadas con wMelPop [67], sobrerregulación consistente con los resultados de Rances *et al.* (2012) en hembras *A. aegypti*-wMelPop mencionados previamente.

En total, puede extraerse que la regulación de citocromos P450 por *Wolbachia* es común a sistemas nativos y transfecciones *in vivo* e *in vitro*. En sistemas nativos, se observan dos tendencias opuestas: subregulación en *D. paulistorum* y sobrerregulación en *A. fluviatilis* y *D. melanogaster*. Por otra parte, en transfecciones la tendencia es, casi exclusivamente, a la sobrerregulación.

Dada la amplitud de las funciones cubiertas por los citocromos P450, es difícil asociar con certeza la alteración de su transcripción a procesos biológicos específicos. Así, el mejor intento para generar hipótesis sobre posibles rasgos universales de *Wolbachia* implicados en los efectos transcripcionales observados, debe acudir a una generalidad de los citocromos P450: su calidad de marcadores del estrés oxidativo [109]. Específicamente, durante la exposición a xenobióticos −sustancias extrínsecas al metabolismo normal de un organismo− la transcripción de los citocromos P450 puede ser inducida mediante una variedad de receptores de xenobióticos, generando una respuesta tendiente a facilitar su excreción [109]. No puede descartarse que existan compuestos de origen bacteriano, reconocidos como xenobióticos en las células hospederas, que induzcan la comúnmente observada sobreexpresión de citocromos P450. La existencia de tal interacción sería consistente con el estrés oxidativo que frecuentemente se ha reportado como consecuencia de la infección *Wolbachia* pues, durante la exposición a xenobióticos, la actividad de los citocromos P450 se puede constituir en una fuente principal de especies oxígeno reactivas (ROS, por sus siglas en inglés) [109]. Se ha propuesto previamente que la producción de ROS por efecto de *Wolbachia* sería mediada, además de por su metabolismo central, por la inducción de la expresión de dual oxidasas hospederas [110]. Si se hace bien en asumir que hay un mecanismo adicional a la fosforilación oxidativa comúnmente implicado en la sobreproducción de ROS, es cuestionable que este sea mediante dual oxidasas, pues estas no aparecieron entre las listas de genes D.E. revisadas en el presente estudio.

Nuevamente asumiendo que existe una correlación entre la abundancia de citocromos P450 y el estrés oxidativo, la subregulación en *D. paulistorum* puede entenderse como una adaptación hospedera a *Wolbachia*. En *D. paulistorum* la presencia de la bacteria es necesaria para su supervivencia (mutualismo obligado) [31]; por lo que es esperable que existan presiones selectivas que favorezcan a los rasgos de tolerancia (en oposición a los de resistencia). Es difícil creer que presiones selectivas igualmente drásticas operen sobre las poblaciones transfectadas que deben interactuar con poblaciones sin la bacteria pues, por ejemplo, los machos infectados tienen menos chance de dejar descendencia que los no infectados (la IC implica que sólo pueden tener descendencia con hembras infectadas). Aún así, dado que el estrés oxidativo es uno de los fenómenos que se creen implicados en el bloqueo patogénico [104], la posibilidad de que este decaiga en el tiempo debe ser considerada en las estrategias de control de vectores. En vista de lo anterior, se destaca la importancia de comprender las interacciones específicas que inducen la producción de ROS por parte de *Wolbachia* y, en particular, develar si la común regulación de citocromos P450 está implicada de alguna forma.

Para terminar cabe decir que, aunque las discusiones se orientaron en torno a los citocromos P450 por cuanto son el principal tipo elemento D.E. que invoca a los términos funcionales presentados, el *heatmap* de la Figura F.6 muestra que hay más genes diferencialmente expresados. Entre tales genes, hay uno que sobresale por aparecer diferencialmente expresado en listas provenientes de las tres especies estudiadas: *fmo-1* (con ortólogo AAEL000797 en *Aedes*). Notablemente, ortólogos a *fmo-1* también se han visto sobreexpresados en hembras *A. aegypti* transfectadas con wMel y wMelPop [68], así como en células *A. gambiae* transfectadas con wAlbB [71]. Las *fmo* (del inglés *flavin-containing monooxygenases*), son enzimas con actividad monooxigenasa que no dependen de hemo sino de flavina, y que en mamíferos tienen un rol en el metabolismo de xenobióticos, catalizando preferentemente la oxigenación de nucleófilos débiles como el nitrógeno y el azufre [111]. Aunque el estudio del rol de las *fmo* en la biología de los insectos es incipiente, también se han asociado al metabolismo de xenobióticos, habiéndose mostrado que confieren resistencia a algunos insecticidas [111, 112]. La regulación de *fmo-1* o sus ortólogos en infecciones nativas (*D. melanogaster*-wMel, *D. paulistorum*-wPau y *A. fluviatilis*-wFlu), transfecciones *in vivo* (*A. aegypti*-wMel y *A. aegypti*-wMelPop) y transfecciones *in vitro* (*A. gambiae*-wAlbB) lo convierten en un claro candidato para futuros estudios.

3.5. Discusiones sobre la metodología

Una de las motivaciones de este trabajo fue aportar a la caracterización de los efectos funcionales *Wolbachia*-Diptera, controlando la incertidumbre asociada con distintos métodos de análisis en distintas publicaciones. En ese sentido, la metodología aquí seguida para la obtención de las listas D.E. es perfectible. Siguiendo el ejemplo del meta-análisis de Chung *et al.* (2020), las lecturas RNA-Seq provenientes de cada publicación seleccionada podrían haber sido tomadas en bruto y sometidas a un análisis RNA-Seq propio; en cambio, sólo las lecturas provenientes de He *et al.* (2019) y Detcharoen *et al.* (2021) lo fueron. Obtener todas las listas D.E. a partir de análisis RNA-Seq propios fue descartado, principalmente, porque habría requerido la reconstrucción de transcriptomas *de novo*; una actividad que demandaría mayores recursos de tiempo y computacionales que los disponibles para la ejecucion de este trabajo. Así, la obtención de listas D.E. pudo ser ″mediada por análisis RNA-Seq propio″ o ″desde tablas″.

La ″Obtención mediada por análisis RNA-Seq propio″, que comprendió la Perfilación transcriptómica y Análisis de expresión diferencial sobre las lecturas de He *et al.* (2019) y Detcharoen *et al.* (2021), se ajustó a las prácticas estándar expuestas en el Anexo A, incluyendo etapas de control de calidad de lecturas y de su alineamiento (los resultados intermedios se presentan y discuten en el Anexo D). Como se expone en el Anexo A, una de las decisiones críticas en el análisis RNA-Seq es qué hacer con las lecturas múltiplemente mapeadas al genoma de referencia (*multireads*). Dado que la cantidad de *multireads* fue baja en ambos casos, la estrategia de ignorar tales alineamientos se consideró bien justificada.

La ″Obtención desde tablas″ fue directa en el caso de las listas provenientes de Lindsey *et al.* (2021), y no amerita discusión. Para obtener listas D.E. a partir de las tablas publicadas por Caragata *et al.* (2017) y Baiao *et al.* (2019), sin embargo, fue necesario hacer un paso intermedio de búsqueda de ortólogos en *A. aegypti* y *D. melanogaster*, respectivamente. Tal como se precisa en el Anexo B.1, la búsqueda ortóloga se hizo mediante el módulo g:Orth de g:Profiler (que acude a mapeos precomputados en Ensembl [46]) o mediante una búsqueda BLAST con los parámetros de corte predefinidos para declarar ortología según STRING. Aunque los métodos aquí usados para inferir ortología son comunes en la práctica, cabe destacar un problema general que los aqueja y es que se basan en homología de secuencias, lo cual no necesariamente implica ortología. Esto introduce incertidumbre sobre la caracterización funcional de las listas D.E., pues no hay garantía de que los supuestos ortólogos cumplan funciones equivalentes en sus respectivos organismos de origen. Aceptar dicha incertidumbre es necesario ante la falta de anotaciones funcionales para genes de organismos no-modelo (en este caso, *A. fluviatilis* y *D. paulistorum*) [17, 31].

Desde la caracterización funcional en adelante, todas las listas fueron sometidas al mismo procedimiento. Una decisión inicial a considerar fue la separación de las listas en genes D.E. según la dirección de la expresión diferencial. Este enfoque no es inédito, habiéndose propuesto que posee ciertas ventajas con respecto a analizar listas completas, como un aumento del poder estadístico [31, 45]. La forma de definir el universo de genes es otra decisión importante. En este trabajo se definió como el conjunto de todos los genes de un organismo de referencia con anotación GO. Por su cercanía filogenética, los organismos de referencia definidos fueron *D. melanogaster* para los datos provenientes de organismos *Drosophila* y *A. aegypti* para los provenientes de *A. fluviatilis*. Cabe notar que la necesidad de que la Obtención desde tablas fuera mediada por búsqueda ortóloga en el caso de Caragata *et al.* (2017) y Baiao *et al.* (2019) surge de esta decisión sobre la manera de definir el universo de genes. En este sentido, la justificación de esta decisión justifica también el haber corrido el riesgo de equivocarse sobre la isofuncionalidad de genes presuntamente ortólogos.

Para las listas provenientes de He *et al.* (2019), Lindsey *et al.* (2021) y Detcharoen *et al.* (2021), el organismo de origen correspondió a un organismo modelo (*D. melanogaster*), en cuyo caso la definición del universo de genes aquí usada ha sido recomendada explícitamente [43]. La pertinencia de definir así el universo de genes para el análisis de enriquecimiento funcional sobre los datos de Caragata *et al.* (2017) y Baiao *et al.* (2019), que provienen de organismos no-modelo, no es tan evidente. En particular, Baiao *et al.* (2019) utilizó una estrategia distinta para su propio análisis, definiendo un universo de genes personalizado, con base en los transcritos identificados en su reconstrucción de transcriptoma [31].

En teoría, el universo debe contener todos los elementos monitoreados, es decir, aquellos que podrían aparecer diferencialmente expresados mediante el respectivo experimento ómico [43]. Es claro que ni la definición propia ni en la de Baiao *et al.* (2019), representa exactamente el conjunto de los genes monitoreados. En la definición de los autores, basada en los transcritos expresados, se excluyen genes que sí fueron monitoreados, en cuanto el RNA-Seq podría identificarlos eventualmente en repeticiones del experimento. Por otra parte, en la definición del universo basada en todos los genes anotados de *D. melanogaster*, se pudieron incluir injustamente genes de *D. melanogaster* sin ortólogos en *D. paulistorum* y se puedieron excluir injustamente genes de *D. paulistorum* sin ortólogos en *D. melanogaster*. Dado que ninguna tercera forma razonable de definir el universo de genes fue vislumbrada, y considerando que la estrategia de personalizar el universo de genes fue inhabilitada por la carencia de las reconstrucciones de transcriptoma, la metodología aquí seguida fue la única posible. En cualquier caso, tal como se expone en el Anexo E, los análisis propios y originales de Baiao *et al.* (2019) son bastante consistentes (una comparación similar no fue posible para el análisis propio sobre listas de Caragata *et al.* (2017), pues los autores no realizaron un análisis originalmente).

Otro punto de consideración es que los parámetros preferidos en g:Profiler fueron improductivos para varias de las listas de genes D.E., lo cual obligó a realizar un segundo análisis con parámetros relajados, que fue el que entregó la mayoría de los resultados expuestos. Especialmente en las listas de genes D.E. pequeñas, como He_up y He_down, ciertos términos funcionales fueron considerados enriquecidos en virtud de la expresión diferencial de muy pocos genes, suscitando dudas razonables sobre si realmente se está ante un fenómeno de relevancia biológica. Aún así, aceptando las suposiciones que justifican al análisis de enriquecimiento de términos funcionales, se puede argumentar que las funciones determinadas bajo los parámetros relajados siguen siendo las mejores candidatas a reflejar efectos relevantes. Sin ir más lejos, opiniones expertas han considerado el uso de parámetros relajados como una medida adecuada frente a listas que se resisten a un análisis más estricto [43].

Las listas de Baiao *et al.* (2019) fueron las que mostraron los patrones de enriquecimiento más contundentes, tanto con respecto al tamaño de los *adjusted p-value* como de la cantidad de términos GO enriquecidos. Una de las principales lecciones de Baiao *et al.* (2019) es que hay efectos funcionales que se manifiestan con especificidad de tejido, y que la secuenciación de cuerpo completo puede significar la pérdida o deterioro de señales locales. Es posible que esta sea una de las razones para la resistencia de las listas de Caragata *et al.* (2017) y Lindsey *et al.* (2021) al análisis de enriquecimiento de términos funcionales: señales de regulación transcripcional con sentido biológico podrían estar siendo reflejados de forma fragmentada en los datos RNA-Seq de cuerpo completo. Esta no puede ser, sin embargo, la única explicación, pues bajo parámetros estrictos hubo términos enriquecidos para los datos de Detcharoen *et al.* (2021), donde las muestras fueron de cuerpo completo, así como ausencia de términos enriquecidos bajo los mismos parámetros para datos de He *et al.* (2019), donde las muestras fueron específicas de ovarios. Las mismas razones propuestas para explicar los tamaños de las listas provenientes de ambos estudios, relativas a los niveles de replicación y variabilidad intra-condición, podrían haber sido determinantes sobre la integridad de las señales transcriptómicas y, por lo tanto, de la facilidad de enriquecimiento de las listas.

Capítulo 4

Conclusiones

Mediante un meta-análisis de datos transcriptómicos provenientes de *Drosophila melanogaster*-wMel, *Drosophila paulistorum*-wPau y *Aedes fluviatilis*-wFlu, el trabajo realizado permitió identificar ámbitos funcionales comúnmente afectados por infecciones *Wolbachia* nativas, lográndose los objetivos propuestos. Para sacar utilidad de los resultados, posteriores esfuerzos se hicieron por comparar las alteraciones comunes en sistemas nativos, con lo reportado para transfecciones. Dependiendo de los resultados de tales comparaciones, se propusieron actividades de investigación o monitoreo relevantes para el presente y futuro de las estrategias de reemplazo de poblaciones basadas en *Wolbachia*. En general, las funciones comúnmente afectadas en sistemas nativos apuntaron a alteraciones igualmente comunes en sistemas transfectados, en cuyo caso los resultados mediaron la generación hipótesis con respecto a rasgos fundamentales de *Wolbachia*.

Las infecciones nativas afectaron la unión e hidrólisis de quitina, un polímero que recientemente ha sido propuesto como punto de interacción entre *Wolbachia*, hospedero y patógenos. La comparación con efectos asociados a quitina en sistemas transfectados sugirió la posibilidad de adaptación de estos, motivando el monitoreo de este ámbito de interacción en el tiempo. Además, se vincularon los efectos asociados a quitina, con pérdidas de resistencia a la desecación observados actualmente en huevos *A. aegypti*-wMel, motivando un estudio inmediato de estos. Finalmente se postuló que, siendo comunes en sistemas nativos y transfectados, estos efectos podrían reflejar un requerimiento metabólico de *Wolbachia* por el monómero constituyente de la quitina, una hipótesis cuyo testeo podría contribuir a la comprensión del fenómeno.

Las infecciones nativas afectaron la unión a iones de calcio. En transfecciones también se vieron efectos asociados, aunque no pudieron identificarse diferencias claras con respecto a los efectos nativos. Se postuló que la afectación común de estas funciones podrían reflejar perturbaciones en los gradientes de calcio debidos a la interacción directa de *Wolbachia* con el retículo endoplásmico. Se identificó una expresión diferencial común de ortólogos a *regucalcin* en infecciones nativas y transfecciones, proponiéndose como blanco de futuras investigaciones tendientes a caracterizar la interacción de *Wolbachia* con el calcio, que hasta ahora ha sido desatendida.

Las actividades serina- y metalopeptidasa fueron comúnmente reguladas en infecciones nativas, la segunda en un sentido negativo. En transfecciones las mismas funciones son so-

brerreguladas; se propuso que la inversión de la dirección de regulación de metalopeptidasas podría reflejar una adaptación hospedera para compensar la actividad peptidasa bacteriana. Dado que los efectos proteolíticos se creen implicados en fenotipos críticos para el control de vectores, los resultados sugirieron la importancia de prever y monitorear posibles cambios de estos en el tiempo.

En infecciones nativas se vio una común regulación de la actividad monooxigenasa, principalmente de aquella dependiente del grupo hemo y ejercida por los citocromos P450. De la comparación con los efectos en transfecciones se derivó nuevamente una posibilidad de adaptación en el tiempo. También, se postuló que la usual sobrerregulación de citocromos P450 podría constituir un mecanismo para la usual producción de especies oxígeno reactivas por efecto de *Wolbachia*, que previamente se ha relacionado con la sobreexpresión de dual oxidasas (para la que no se halló mayor evidencia en este trabajo). También se identificó la expresión diferencial de ortólogos de una monooxigenasa dependiente de flavina (*fmo-1*) en los tres sistemas nativos y en sistemas transfectados, sugiriendo un posible blanco para futuras investigaciones.

Se espera que este trabajo contribuya a la generación de conocimiento sobre las interacciones *Wolbachia*-hospedero a nivel funcional, e indique direcciones de investigación relevantes para el éxito de las estrategias de control basadas en *Wolbachia* en el tiempo.

Bibliografía

[1] JONES, R., et al. 2020. Arbovirus vectors of epidemiological concern in the Americas: A scoping review of entomological studies on Zika, dengue and chikungunya virus vectors. PLOS ONE 15(2).

[2] ARREDONDO, J., MÉNDEZ, A., y MEDINA, H. 2016. Arbovirus en Latinoamérica. Acta Pediátrica Mexicana 37(2): 111-131.

[3] SILVA, J., et al. 2017. Wolbachia and dengue virus infection in the mosquito Aedes fluviatilis (Diptera: Culicidae). PloS one 12(7): e0181678.

[4] ORGANIZACIÓN MUNDIAL DE LA SALUD. 2020. Dengue y Dengue Grave [En línea]<https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue> [Fecha consulta: 03 de julio de 2021].

[5] ORGANIZACIÓN MUNDIAL DE LA SALUD. 2016. La Historia del Virus Zika [En línea] <https://www.who.int/news-room/feature-stories/detail/the-history-of-zika-virus> [Fecha consulta: 03 de julio de 2021].

[6] PATTNAIK, A., y SAHOO, B. 2020. Current Status of Zika Virus Vaccines: Successes and Challenges. Vaccines 8(2): 266.

[7] ORGANIZACIÓN PANAMERICANA DE LA SALUD. 2013. Fiebre Amarilla [En línea] <https://www3.paho.org/hq/index.php?option=com_contentview=articleid=9476:yellow-feverItemid=40721lang=es> [Fecha consulta: 07 de julio de 2021].

[8] ORGANIZACIÓN MUNDIAL DE LA SALUD. 2019. Fiebre Amarilla [En línea] <https://www.who.int/es/news-room/fact-sheets/detail/yellow-fever> [Fecha consulta: 07 de julio de 2021].

[9] YEN, P. y FAILLOUX, A. 2020. A Review: Wolbachia-Based Population Replacement for Mosquito Control Shares Common Points with Genetically Modified Control Approaches. Pathogens 9(5): 404.

[10] GIRARD, M., et al. 2020. Arboviruses: A global public health threat. Vaccine 38(24): 3989–3994.

[11] REDONI, M., et al. 2020. Dengue: Status of current and under-development vaccines. Reviews in Medical Virology, e2101.

[12] BECKHAM, J., y TYLER, K. 2015. Arbovirus Infections. CONTINUUM: Lifelong Learning in Neurology 21: 1599–1611.

[13] SIGLE, L., et al. 2022. Assessing Aedes aegypti candidate genes during viral infection and Wolbachia-mediated pathogen blocking. Insect molecular biology.

[14] DA SILVA, I., et al. 2021. Systematic Review of Wolbachia Symbiont Detection in Mosquitoes: An Entangled Topic about Methodological Power and True Symbiosis. Pathogens 10 (39).

[15] LINDSEY, A., et al. 2021. Wolbachia and Virus Alter the Host Transcriptome at the Interface of Nucleotide Metabolism Pathways. ASM Journals 12(1).

[16] TERAMOTO, T. et al. 2019. Infection of Aedes albopictus Mosquito C6/36 Cells with the wMelpop Strain of Wolbachia Modulates Dengue Virus-Induced Host Cellular Transcripts and Induces Critical Sequence Alterations in the Dengue Viral Genome. ASM Journals, Journal of Virology 93(15).

[17] CARAGATA, E., et al. 2017. The transcriptome of the mosquito Aedes fluviatilis (Diptera: Culicidae), and transcriptional changes associated with its native Wolbachia infection. BMC Genomics 18(1).

[18] HOFFMANN, A., et al. 2015. Wolbachia strains for disease control: ecological and evolutionary considerations. Evolutionary applications 8(8): 751–768.

[19] ZUG, R., y HAMMERSTEIN, P. 2015. Bad guys turned nice? A critical assessment of Wolbachia mutualisms in arthropod hosts. Biological reviews of the Cambridge Philosophical Society 90(1): 89–111.

[20] ROSS, P., et al. 2020. An elusive endosymbiont: Does Wolbachia occur naturally in Aedes aegypti? Ecol Evol 10: 1581– 1591.

[21] SEGOLI, M., et al. 2014. The effect of virus-blocking Wolbachia on male competitiveness of the dengue vector mosquito, Aedes aegypti. PLoS neglected tropical diseases 8(12): e3294.

[22] NAZNI, W., et al. 2019. Establishment of Wolbachia Strain wAlbB in Malaysian Populations of Aedes aegypti for Dengue Control. Current biology 29(24): 4241–4248.

[23] RAINEY, S., et al. 2016. Wolbachia Blocks Viral Genome Replication Early in Infection without a Transcriptional Response by the Endosymbiont or Host Small RNA Pathways. PLOS Pathogens 12(4): e1005536.

[24] RYAN, P., et al. 2020. Establishment of wMel Wolbachia in Aedes aegypti mosquitoes and reduction of local dengue transmission in Cairns and surrounding locations in northern Queensland, Australia. Gates open research 3: 1547.

[25] CARAGATA, E., et al. 2019. Pathogen blocking in Wolbachia-infected Aedes aegypti is not affected by Zika and dengue virus co-infection. PLoS neglected tropical diseases 13(5): e0007443.

[26] UTARINI, A., et al. 2021. Efficacy of Wolbachia-Infected Mosquito Deployments for the Control of Dengue. The New England Journal of Medicine 384: 2177-2186.

[27] EDENBOROUGH, K. et al. 2021. Using Wolbachia to eliminate Dengue: Will the Virus Fight Back? ASM Journals, Journal of Virology 95(13).

[28] ROSS, P., et al. 2022. A decade of stability for wMel Wolbachia in natural Aedes aegypti populations. PLOS Pathogens 18(2): e1010256.

[29] DETCHAROEN, M. et al. 2021. Differential gene expression in Drosophila melanogaster and D. nigrosparsa infected with the same Wolbachia strain. Scientific Reports 11(1).

[30] HE, Z., et al. 2019. How do Wolbachia modify the Drosophila ovary? New evidences support the "titration-restitution" model for the mechanisms of Wolbachia-induced CI. BMC Genomics 20: 608.

[31] BAIAO, G. et al. 2019. The effect of Wolbachia on gene expression in Drosophila paulistorum and its implications for symbiont-induced host speciation. BMC genomics 20(465).

[32] CHUNG, M., et al. 2020. A Meta-Analysis of Wolbachia Transcriptomics Reveals a Stage-Specific Wolbachia Transcriptional Response Shared Across Different Hosts. G3 10(9): 3243–3260.

[33] LAROSSA, R. 2013. Brenner's Encyclopedia of Genetics. 2a edición. Academic Press.

[34] ZHANG, H. 2019. The review of transcriptome sequencing: principles, history and advances. IOP Conference Series: Earth and Environmental Science 332.

[35] VAILATI-RIBONI M., et al. Palombo V., Loor J.J. (2017) What Are Omics Sciences?. In: Ametaj B. (eds) Periparturient Diseases of Dairy Cows. Springer, Cham.

[36] CONESA, A., et al. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol 17: 13.

[37] PAREKH, S., et al. 2016. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep 6.

[38] WANG, Z., et al. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63.

[39] DARBY, A., et al. 2012. Analysis of gene expression from the Wolbachia genome of a filarial nematode supports both metabolic and defensive roles within the symbiosis. Genome research 22(12): 2467–2477.

[40] ZHAO, S., et al. 2016. Bioinformatics for RNA-Seq Data Analysis. In (Ed.), Bioinformatics - Updated Features and Applications. IntechOpen.

[41] TOMCZAK, A. et al. 2018. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Scientific Reports 8 5115.

[42] GENE ONTOLOGY CONSORTIUM. Gene Ontology Resource [En línea] <www.geneontology.org> [Consulta: 10 de enero de 2022]

[43] REIMAND, J., et al. 2019. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc 14: 482–517.

[44] RIVALS, I., et al. 2007. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23(4): 401–407.

[45] HONG, G., et al. 2013. Separate enrichment analysis of pathways for up- and downregulated genes. Journal of the Royal Society, Interface 11(92).

[46] REIMAND, J., et al. 2007. g:Profiler, a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic acids research 35.

[47] BIOINFORMATICS AND INFORMATION TECHNOLOGY RESEARCH GROUP. g:Profiler, a web-based toolset for functional profiling of gene lists [En línea] <https://biit.cs.ut.ee/gprofiler/>[Consulta: 12 de diciembre 2021]

[48] SHANNON, P., et al. 2003. Cytoscape: a software environment for integrated models of

biomolecular interaction networks. Genome research 13(11): 2498–2504.

[49] MERICO, D., et al. 2010. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. PLOS ONE 5(11): e13984.

[50] MORRIS, J., et al. 2011. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics 12.

[51] OESPER, L., et al. 2011. WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. Source Code Biol Med 6.

[52] ENRICHMENTMAP. EnrichmentMap Cytoscape App 3.3 [En línea]<https://enrichmentmap.readthedocs.io/en/latest/index.html>[Consulta: 12 de octubre de 2021]

[53] STIJN, D. 2002. A New Cluster Algorithm for Graphs. Inf Syst 1.

[54] SZKLARCZYK, D., et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research 47(1): D607–D613.

[55] NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. Home - SRA - NCBI. [En línea] <https://www.ncbi.nlm.nih.gov/sra> [Consulta: 22 de junio de 2022]

[56] ANDREWS, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [En línea] <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> [Consulta: 02 de febrero 2022]

[57] MULTIQC. Documentation: MultiQC [En línea] <https://multiqc.info/docs/> [Consulta: 02 de febrero de 2022]

[58] ANTHONY, M., et al. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30(15): 2114–2120.

[59] DOS SANTOS, G., et al. 2015. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic acids research 43: D690–D697.

[60] FLYBASE. Index of /genomes/Drosophila_melanogaster/dmel_r6.32_FB2020_01/gtf [En línea] <http://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.32_FB2020_01/gtf/> [Consulta: 22 de junio de 2022]

[61] GRAMATES, L., et al. 2022. FlyBase: a guided tour of highlighted features. Genetics 220(4).

[62] DOBIN, A., y GINGERAS, T. 2015. Mapping RNA-seq Reads with STAR. Current protocols in bioinformatics 51.

[63] LIAO, Y., et al. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30(7): 923–930

[64] LOVE, M., et al. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15.

[65] R CORE TEAM. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[66] UNIPROT. Proteomes - Drosophila melanogaster (Fruit fly). [En línea]

<https://www.uniprot.org/proteomes/UP000000803> [Consulta: 10 de abril de 2022]

[67] GEOGHEGAN, V., et al. 2017. Perturbed cholesterol and vesicular trafficking associated with dengue blocking in Wolbachia-infected Aedes aegypti cells. Nat Commun 8: 526.

[68] RANCES, E., et al. 2012. The relative importance of innate immune priming in Wolbachia-mediated dengue interference. PLoS pathogens 8(2): e1002548.

[69] XI, Z., et al. 2008. Genome-wide analysis of the interaction between the endosymbiotic bacterium Wolbachia and its Drosophila host. BMC Genomics 9(1).

[70] KAMBRIS, Z., et al. 2009. Immune activation by life-shortening Wolbachia and reduced filarial competence in mosquitoes. Science 326(5949): 134–136.

[71] HUGHES, G., et al. 2011. Wolbachia infections in Anopheles gambiae cells: transcriptomic characterization of a novel host-symbiont interaction. PLoS pathogens 7(2): e1001296.

[72] BALDRIDGE, G., et al. 2017. Proteomic analysis of a mosquito host cell response to persistent Wolbachia infection. Research in microbiology 168(7): 609–625.

[73] PIETRI, J., et al. 2016. The rich somatic life of Wolbachia. MicrobiologyOpen 5(6): 923–936.

[74] PENG, Y., et al. 2008. Wolbachia infection alters olfactory-cued locomotion in Drosophila spp. Applied and environmental microbiology 74(13): 3943–3948.

[75] SUH, E. y DOBSON, S. 2013. Reduced competitiveness of Wolbachia infected Aedes aegypti larvae in intra- and inter-specific immature interactions. Journal of invertebrate pathology 114(2): 173–177.

[76] GIRALDO-CALDERÓN, G., et al. 2017. Retention of duplicated long-wavelength opsins in mosquito lineages by positive selection and differential expression. BMC Evol Biol 17: 84.

[77] ROCHA, M., et al. 2015. Expression and light-triggered movement of rhodopsins in the larval visual system of mosquitoes. The Journal of experimental biology 218(9): 1386–1392.

[78] LAU, M. et al. 2020. Measuring the Host-Seeking Ability of Aedes aegypti Destined for Field Release. The American journal of tropical medicine and hygiene 102(1): 223–231.

[79] ZHAN, et al. 2021. Elimination of vision-guided target attraction in Aedes aegypti using CRISPR. Current Biology 31(18): 4180-4187.

[80] KACZMAREK, A., et al. 2021. The type of blood used to feed Aedes aegypti females affects their cuticular and internal free fatty acid (FFA) profiles. PloS one 16(4): e0251100.

[81] WATANABE, K., et al. 2018. Light is required for proper female mate choice between winged and wingless males in Drosophila. Genes & genetic systems 93(3): 119–123.

[82] RODRÍGUEZ, B., et al. 2019. The Role of Preferential Mating and Wolbachia Infection on the Aedes Aegypti Population Dynamics. USA, Arizona. Arizona State University, Mathematical and Theoretical Biology Institute. 38p.

[83] YEAP, H., et al. 2018. The Effect of Nonrandom Mating on Wolbachia Dynamics: Implications for Population Replacement and Sterile Releases in Aedes Mosquitoes. The American journal of tropical medicine and hygiene 99(3): 608–617.

[84] BHAGAVAN, N. y HA. C. 2011. Chapter 9 - Heteropolysaccharides: Glycoconjugates,

Glycoproteins, and Glycolipids. En: Essentials of Medical Biochemistry: With Clinical Cases. San Diego, Academic Press. pp.75-83

[85] SHUKLA, A. 2020. Chemical and Synthetic Biology Approaches to Understand Cellular Functions. 3a edición. Amsterdam, Países Bajos. Elsevier.

[86] UDAYA PRAKASH, N., et al. 2010. Evolution, homology conservation, and identification of unique sequence signatures in GH19 family chitinases. Journal of molecular evolution 70(5): 466–478.

[87] ZHANG, X., et al. 2012. Identification and characterization of two chitin synthase genes in African malaria mosquito, Anopheles gambiae. Insect biochemistry and molecular biology 42(9): 674–682.

[88] SIGLE, L. y McGRAW, E. 2019. Expanding the canon: Non-classical mosquito genes at the interface of arboviral infection. Insect biochemistry and molecular biology 109: 72–80.

[89] ZHENG, Y., et al. 2011. Differentially expressed profiles in the larval testes of Wolbachia infected and uninfected Drosophila. BMC genomics 12: 595.

[90] ALLMAN, M., et al. 2020. Wolbachia's Deleterious Impact on Aedes aegypti Egg Development: The Potential Role of Nutritional Parasitism. Insects 11(11): 735.

[91] REZENDE, G., et al. 2008. Embryonic desiccation resistance in Aedes aegypti: presumptive role of the chitinized serosal cuticle. BMC developmental biology 8: 82.

[92] FARNESI, L., et al. 2015. Physical features and chitin content of eggs from the mosquito vectors Aedes aegypti, Anopheles aquasalis and Culex quinquefasciatus: Connection with distinct levels of resistance to desiccation. Journal of Insect Physiology 83: 43-52.

[93] VOLLMER, J., et al. 2013. Requirement of lipid II biosynthesis for cell division in cell wall-less Wolbachia, endobacteria of arthropods and filarial nematodes. International journal of medical microbiology 303(3): 140–149.

[94] HENRICHFREISE, B., et al. 2009. Functional conservation of the lipid II biosynthesis pathway in the cell wall-less bacteria Chlamydia and Wolbachia: why is lipid II needed?. Molecular microbiology 73(5): 913–923.

[95] SACOMAN, J. y HOLLINGSWORTH, R. 2011. Synthesis and evaluation of an N-acetylglucosamine biosynthesis inhibitor. Carbohydrate research 346(14): 2294–2299.

[96] KYOTO ENCYCLOPEDIA OF GENES AND GENOMES. KEGG PATHWAY Database. [En línea] <https://www.genome.jp/kegg/pathway.html> [Consulta: 20 de abril de 2022]

[97] WILMES, A., et al. 2017. AmiD Is a Novel Peptidoglycan Amidase in Wolbachia Endosymbionts of Drosophila melanogaster. Frontiers in Cellular and Infection Microbiology 7.

[98] DALE, C. y WELBURN, S. 2001. The endosymbionts of tsetse flies: manipulating host parasite interactions. International Journal for Parasitology 31: 628-631.

[99] ZHENG, Y., et al. 2011. Differentially expressed profiles in the larval testes of Wolbachia infected and uninfected Drosophila. BMC genomics 12: 595.

[100] BERRIDGE, M., et al. 2000. The versatility and universality of calcium signalling. Nat

Rev Mol Cell Biol 1: 11–21.

[101] LAURENTINO, S., et al. 2012. Regucalcin, a calcium-binding protein with a role in male reproduction?. Molecular human reproduction 18(4): 161–170.

[102] YAMAGUCHI, M. 2005. Role of regucalcin in maintaining cell homeostasis and function (review). International journal of molecular medicine 15(3): 371–389.

[103] FATTOUH, N., et al. 2019. Wolbachia endosymbionts subvert the endoplasmic reticulum to acquire host membranes without triggering ER stress. PLOS Neglected Tropical Diseases 13(3): e0007218.

[104] LINDSEY, A., et al. 2018. Conflict in the Intracellular Lives of Endosymbionts and Viruses: A Mechanistic Look at Wolbachia-Mediated Pathogen-blocking. Viruses 10(4): 141.

[105] CARAGATA, E., et al. 2014. Competition for amino acids between Wolbachia and the mosquito host, Aedes aegypti. Microbial ecology 67(1): 205–218.

[106] JIMÉNEZ, N., et al. 2019. A systems biology approach for studying Wolbachia metabolism reveals points of interaction with its host in the context of arboviral infection. PLOS Neglected Tropical Diseases 13(8): e0007678.

[107] TORRES, D., et al. 2010. Monooxygenases as biocatalysts: Classification, mechanistic aspects and biotechnological applications. Journal of biotechnology 146(1): 9–24.

[108] BUSCH, A. y MONTGOMERY, B. 2015. Interdependence of tetrapyrrole metabolism, the generation of oxidative stress and the mitigative oxidative stress response. Redox biology 4: 260–271.

[109] HE, L., et al. 2017. Antioxidants Maintain Cellular Redox Homeostasis by Elimination of Reactive Oxygen Species. Cell Physiol Biochem 44: 532-553.

[110] ZUG, R. y HAMMERSTEIN, P. 2015. Wolbachia and the insect immune system: what reactive oxygen species can tell us about the mechanisms of Wolbachia-host interactions. Frontiers in Microbiology 6.

[111] SEHLMEYER, S., et al. 2010. Flavin-dependent monooxygenases as a detoxification mechanism in insects: new insights from the arctiids (lepidoptera). PloS one 5(5): e10435.

[112] MALLOTT, M., et al. 2019. A flavin-dependent monooxgenase confers resistance to chlorantraniliprole in the diamondback moth, Plutella xylostella. Insect Biochemistry and Molecular Biology 115.

[113] LANDMANN F. 2019. The Wolbachia Endosymbionts. Microbiology spectrum 7(2).

[114] PETER, J., et al. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research 38(6): 1767–1771.

[115] BANSAL, V. 2017. A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. BMC bioinformatics 18(3): 43.

[116] DESCHAMPS-FRANCOEUR, G., et al. 2020. Handling multi-mapped reads in RNA-seq. Computational and structural biotechnology journal 18: 1569–1576.

[117] LI, B. y DEWEY, C. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12.

[118] CONSIGLIO, A., et al. 2016. A fuzzy method for RNA-Seq differential expression analysis in presence of multireads. BMC Bioinformatics 17.

[119] CHEN, Y., et al. Bioconductor - EdgeR [En línea] <https://bioconductor.org/packages/release/bioc/html/edgeR.html> [Consulta: 02 de noviembre de 2021]

[120] HUANG, H., et al. 2015. Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software. Cancer informatics 14: 57–67.

[121] WINER, B. 1962. Statistical principles in experimental design. Nueva York, Estados Unidos. McGraw-Hill Book Company.

[122] STOREY, J., y TIBSHIRANI, R. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100(16): 9440–9445.

[123] LI, H., et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16): 2078–2079.

Anexo A

Antecedentes complementarios

En este Anexo se precisan las etapas del Análisis de datos RNA-Seq que fueron omitidas en el Marco teórico, esto es: Perfilación transcriptómica y Análisis de expresión diferencial. Estos antecedentes son útiles para justificar la metodología seguida para el análisis RNA-Seq propio realizado sobre los datos en bruto publicados por He *et al.* (2019) y Detcharoen *et al.* (2021), así como para comprender los resultados intermedios de estos análisis, presentados en el Anexo D

A.1. Perfilación transcriptómica

Los tipos de análisis relevantes para este trabajo inician con la generación de un perfil de expresión para cada muestra secuenciada, esto es, un conjunto de cuentas de lecturas mapeadas a cada gen o transcrito de interés. Dependiendo del algoritmo a utilizar para el posterior análisis de expresión diferencial, las cuentas podrán ser brutas o normalizadas para eliminar sesgos asociados con el tamaño de cada librería de ADNc secuenciada, el largo de los elementos, su contenido de GC, u otros (algunos métodos de normalización se mencionan más adelante) [36]. Para obtener los perfiles de expresión, los archivos con lecturas son sometidos a un proceso que, en general, puede descomponerse en tres subetapas revisadas a continuación: Control de calidad, Alineamiento de lecturas y Conteo de expresión.

A.1.1. Control de calidad

Un formato prevalente para la transferencia de los resultados de la secuenciación es FASTQ [114]. En dicho formato, cada lectura se representa por una entrada que incluye un identificador, la secuencia nucleotídica en sentido 5' - 3', y una medida de la certeza con que cada base de la secuencia fue identificada, usualmente en términos del *Phred Quality Score* ($Q_{Phred} = -10 \cdot log_{10}(p_e)$, donde p_e estima la probabilidad de que se haya cometido un error al determinar la base) [114]. En el caso de las lecturas *paired-end*, los resultados se distribuyen en dos archivos FASTQ, cuyas entradas i-ésimas se corresponden entre sí como lecturas de ambos extremos de un mismo fragmento.

El control de calidad inicia con la evaluación de las lecturas en términos de su Q_{Phred}, su longitud, su contenido de *GC* (guanina-citosina), el nivel de duplicación de lecturas, y la presencia de secuencias de adaptadores y de *k-mers* (subsecuencias nucleotídicas de tamaño k) sobrerrepresentados, entre otros. FastQC es una herramienta ampliamente utilizada, que permite evaluar dichos aspectos y reunir los resultados en un reporte dividido en módulos, con respecto a los cuales el archivo puede ser *normal*, *ligeramente anormal* o *muy anormal* [56]. Tres de los módulos que componen a un reporte FastQC se ilustran en la Figura A.1 y se describen posteriormente.

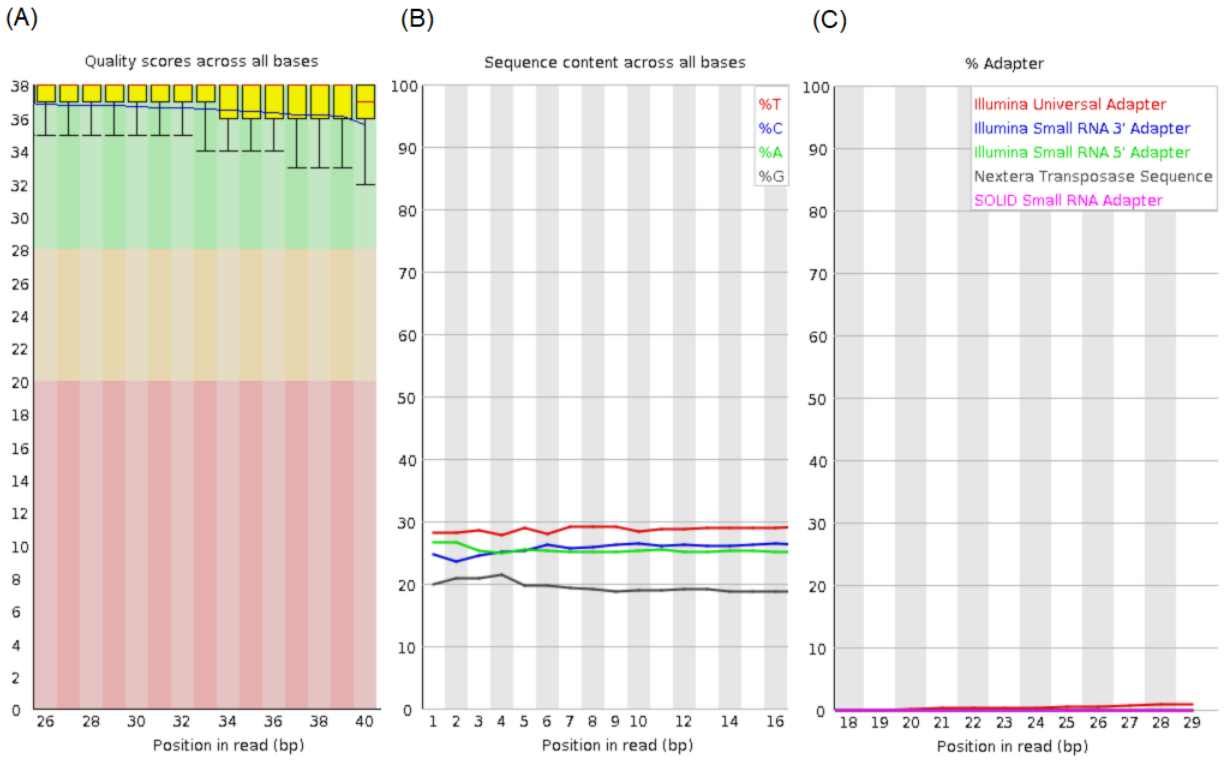

Figura A.1: Ejemplo de evaluación FastQC de un archivo de lecturas de 40 bases, según los módulos: (a) *Per Base Sequence Quality*, (b) *Per Base Sequence Content*, y (c) *Adapter Content*. En cada caso se muestra sólo un rango reducido de las posiciones nucleotídicas, suficiente para detectar cambios típicos (comentados en el texto del informe). Imágenes modificadas de las originales en la documentación de FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

El módulo *Per Base Sequence Quality* resume, para cada posición, la distribución de los Q_{Phred} entre las lecturas, mediante un diagrama de caja donde el cuadro amarillo demarca el rango intercuartílico (25 %-75 %), y las líneas rojas y azules denotan la mediana y la media, respectivamente. Tal como se aprecia en la Figura A.1a, es común que la calidad disminuya hacia el extremo 3' de las lecturas, fenómeno que tiene base en la química del proceso de secuenciación [56]. Si la magnitud de la caída de calidad la convierte en un riesgo para los análisis subsecuentes, se puede solucionar mediante aplicaciones como Trimmomatic o Cutadapt, que permiten remover lecturas o extremos de lecturas que no cumplan con cierto criterio de calidad definido por el usuario [40]. Por ejemplo, el

módulo SLIDINGWINDOW de Trimmomatic detecta lecturas con ventanas (*i.e.*, bases consecutivas) que promedian un Q_{Phred} por debajo de un valor determinado, removiendo un extremo de la lectura o la lectura completa, dependiendo de la ubicación de la ventana problemática [40].

Per Base Sequence Content es un módulo que muestra, para cada posición, la proporción de lecturas que han sido asignadas con cada base nitrogenada. Para una librería aleatoria se esperaría que dichas proporciones no variaran considerablemente con la posición, sino que se mantuvieran en un nivel estable determinado por el sistema biológico de origen, visualizándose como líneas horizontales paralelas (tal como se observa en la Figura A.1b, aproximadamente desde la posición 7 en adelante) [56]. Las librerías RNA-Seq, sin embargo, suelen presentar sesgos en la composición de las posiciones iniciales de las lecturas, introducidos durante la etapa de amplificación del ADNc [56]. Esta anomalía se puede observar en la misma Figura A.1b (aproximadamente entre las posiciones 1 y 6) y, aunque puede ocasionar que el archivo sea considerado anormal según el módulo *Per Base Sequence Content*, se suele considerar como un sesgo tolerable [56].

El módulo *Adapter Content* muestra, para cada posición, el porcentaje de lecturas en las que se han detectado secuencias de adaptadores que incluyen a dicha posición. En la Figura A.1c se observa una leve presencia de adaptadores hacia el extremo 3' de la lectura. Al igual que las caídas de calidad, la presencia de adaptadores puede ser solucionada removiendo los extremos 3' problemáticos mediante Trimmomatic, Cutadapt o similares [56].

Como se mencionó en el caso del módulo *Per Base Sequence Content*, una anormalidad declarada por FastQC no significa necesariamente que el archivo evaluado sea defectuoso: los resultados deben ser interpretados en el contexto de lo que se espera de la librería secuenciada, incluyendo posibles sesgos [56]. Niveles aceptables de duplicación, de sobrerrepresentación de *k-mers* o de contenido *GC* son experimento- y organismo-específicos (por ejemplo, en RNA-Seq se suele hallar considerable duplicación de lecturas debido a la existencia de genes con alto nivel de expresión y, en menor medida, de la duplicación artificial de fragmentos de ADNc mediante PCR [115]). En general se espera, sin embargo, que estos valores sean relativamente similares entre archivos provenientes del mismo experimento [36].

Una herramienta útil para comparar los resultados de FastQC sobre múltiples archivos con lecturas es MultiQC [57], que reúne esta información en un único reporte. Importantemente, uno de los módulos de MultiQC resume la evaluación de cada archivo de lecturas en cada módulo FastQC, mediante una tabla de doble entrada. En dicha tabla, cada columna se asocia a un módulo FastQC y cada fila se asigna a un archivo de lecturas; para cada combinación módulo-archivo el color de la entrada representa el resultado de la evaluación (verde equivale a *normal*, amarillo a *ligeramente anormal* y rojo a *muy anormal*) [57].

A.1.2. Alineamiento de lecturas

En general, la cuantificación de la expresión de genes o transcritos debe ser precedida por un mapeo o *alineamiento* de las lecturas RNA-Seq limpias a un conjunto de secuencias nucleotídicas de referencia, esto es, un genoma o un transcriptoma [36, 63].

El alineamiento de lecturas a genomas eucariontes se ve dificultado por la existencia de intrones, esto es, regiones genómicas inicialmente transcritas pero posteriormente removidas del ARNm maduro mediante el proceso de *splicing* [40, 62]. Para sortear dicho obstáculo, herramientas como TopHat2 o STAR permiten realizar alineamientos espaciados (o conscientes de *splicing*) al genoma de referencia [40, 62], como se representa en la Figura G.44a. Si bien el alineamiento espaciado puede realizarse usando sólo las lecturas y el genoma de referencia, es altamente recomendable proveer a los algoritmos una anotación de los elementos genómicos de este último, para guiar el mapeo de las lecturas que atraviesan intrones [40]. Este tipo de anotaciones se codifica en formatos derivados del GFF (*General Feature Format*), como GFF3 o GTF.

En contraste, el alineamiento a un transcriptoma eucarionte no requiere ser espaciado, ya que en RNA-Seq cada lectura se origina de la secuenciación de un fragmento contiguo de un transcrito, como se observa en la Figura G.44b. A cambio, una desventaja de este método es que no permite la identificación y cuantificación de isoformas no incluidas en el transcriptoma de referencia, dependiendo críticamente de la completitud y calidad de este [36]. Herramientas como el módulo *rsem-prepare-reference* del paquete RSEM permiten extraer transcriptomas de referencia a partir de genomas y sus anotaciones [117].

Cuando no se dispone de una referencia *a priori*, como en el caso de especies nuevas o poco estudiadas, el alineamiento debe ser precedido por una reconstrucción de transcriptoma *de novo* a partir de las lecturas limpias. Cada una de las secuencias ensambladas en esta operación se denomina *contig* e, idealmente, se corresponde con un transcrito íntegro (aunque en la práctica se generan *contigs* correspondientes a transcritos truncados o artefactos del ensamble, como uniones quiméricas). La reconstrucción de transcriptoma se puede realizar mediante herramientas basadas en grafos *De Bruijn*, como Trans-ABySS y Trinity, mientras que herramientas como BUSCO, TransDecoder y CD-HIT-EST permiten evaluar el ensamble según su inclusión de genes conservados, anotar potenciales marcos abiertos de lectura y colapsar transcritos redundantes, respectivamente [36].

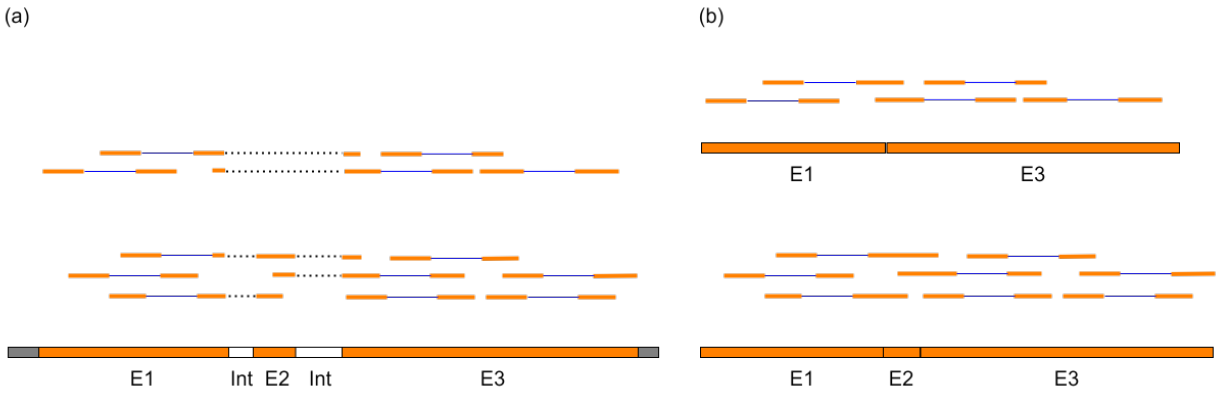

Figura A.2: Esquematización del alineamiento ideal de lecturas *paired-end* a: (a) el gen eucarionte del que provienen, compuesto por tres exones (E1, E2 y E3) y dos intrones (Int); o (b) los dos transcritos específicos (E1-E3 y E1-E2-E3) de los cuales provienen. Se ilustra como el mapeo a genes debe tomar en cuenta la existencia de secuencias que son removidas de los transcritos maduros mediante *splicing* y que, por lo tanto, no son capturadas por las lecturas RNA-Seq. Cada par de líneas naranjas unidas por una línea azul (y eventualmente interrumpidas por una línea punteada que denotan un espaciado) simbolizan un par de lecturas de ambos extremos de un mismo fragmento de ADNc. Elaboración propia.

Cabe destacar que la Figura G.44 representa un escenario ideal, en que las lecturas pueden ser perfectamente alineadas a su origen en la referencia. En la práctica, sin embargo, las lecturas contienen discordancias (en inglés, *mismatches*), inserciones y deleciones con respecto a la referencia, que pueden deberse a variaciones genómicas, errores de secuenciación o errores en la referencia [62]. Para lidiar con estas dificultades, el proceso de alineamiento involucra el cálculo de puntajes que son favorecidos por las coincidencias exactas, y penalizados por discordancias, inserciones y deleciones [116].

Otra dificultad práctica radica en que algunas lecturas pueden ser alineadas a múltiples ubicaciones de la referencia, debido a la presencia de secuencias genómicas repetitivas, genes parálogos, pseudogenes o múltiples variantes de *splicing* (siendo este último un factor exclusivo para el caso del alineamiento a transcriptoma) [36, 116]. La forma en que las herramientas deciden qué alineamientos reportar para dichas lecturas, denominadas *multireads*, varía. Por ejemplo, STAR reporta directamente el alineamiento de mayor puntaje, junto con todos aquellos cuyo puntaje se encuentre a una distancia dada de éste [62]. Por otra parte, antes de reportar los alineamientos de los *multireads*, TopHat2 recalcula los puntajes en cada región candidata, tomando en cuenta información derivada del resto de los mapeos [116].

Los resultados de los alineamientos se almacenan en formatos tabulares SAM (*Sequence Alignment/Map*) o BAM (*Binary Alignment/Map*), donde cada mapeo se expresa en una entrada con 11 campos fijos, incluyendo: QNAME (nombre de la lectura), FLAG (código que indica la hebra a la que corresponde la lectura actual, si se trata de una lectura pareada o no, entre otros), RNAME (nombre de la secuencia de referencia), POS (primera posición de la referencia que se alinea con la lectura), MAPQ (puntaje del alineamiento) y CIGAR (código que detalla el alineamiento base a base, en términos de coincidencias, discordancias,

inserciones, deleciones, saltos asociados a intrones, etc) [123].

A.1.3. Conteo de expresión

En esta etapa se ocupan los alineamientos previos para contar lecturas a nivel de genes o transcritos. Por ser atingente a los métodos adoptados en este trabajo, las siguientes descripciones se centrarán en el conteo de la expresión a nivel de genes, en cuyo caso los insumos necesarios son un archivo de alineamientos (en formato SAM/BAM) y una anotación de elementos genómicos (en formato GFF o derivados) [116].

Tal como en la etapa anterior, la presencia de *multireads* impone dificultades y exige la toma de decisiones por parte de los algoritmos a cargo. La proporción de *multireads* varía dependiendo del sistema biológico, la muestra específica estudiada, el tipo de ARN enriquecido, el algoritmo de alineamiento y la referencia utilizada, reportándose valores típicos de entre 5 y 40 % [116]. Dada su potencial abundancia, la manera de contar los *multireads* puede afectar considerablemente los resultados de los análisis posteriores, razón por la cual se han implementado distintas estrategias [116].

En la cuantificación de la expresión a nivel de genes, herramientas como HTSeq-count y featureCounts implementan métodos básicos para tratar los grupos de *multireads*, esto es: ignorarlos, contarlos en cada gen candidato, o repartirlos equitativamente entre los genes candidatos [116]. Ignorar los *multireads* es el método implementado por defecto en featureCounts y se considera útil por cuanto elimina la incertidumbre con respecto al origen de las lecturas. Sin perjuicio de lo anterior, tanto ignorar los *multireads* como contarlos en cada alineamiento candidato alteran artificialmente el tamaño de las muestras, prefiriéndose cuando la proporción de *multireads* es más bien baja [116, 118]. Por otra parte, si bien distribuir los *multireads* equitativamente asegura que cada uno sea contado una única vez, se torna impreciso cuando varios de los genes candidatos tienen una expresión real baja, en cuyo caso esta estrategia sobreestima su expresión en desmedro de aquellos genes candidatos realmente activos [116].

Métodos más elaborados consisten en repartir los *multireads* de forma proporcional a la cantidad de lecturas mapeadas en la vecindad de cada alineamiento candidato (métodos de cobertura), o basándose en modelos estadísticos [116]. Las estrategias de cobertura, como la adoptada por el programa Rcount, se consideran apropiadas para ARN con cobertura de lecturas relativamente homogénea y con regiones únicas flanqueando a las regiones compartidas, mientras que subóptimas para el análisis de sRNA-Seq (secuenciación de ARN pequeño), donde dichas condiciones son menos comunes [116]. Alternativamente, herramientas como IsoEM y el módulo *rsem-calculate-expression* del paquete RSEM plantean modelos estadísticos para obtener estimadores de máxima verosimilitud de los niveles de expresión de genes/transcritos mediante el algoritmo EM (*Expectation-Maximization*), manejando la incertidumbre asociada con los *multireads* en el marco de dichos modelos [116, 117].

Otra decisión importante en esta etapa es el método de normalización que se emplea

para corregir efectos no-biológicos sobre las cuentas brutas, debido a factores como el largo de los elementos genómicos, su contenido de GC, la profundidad de secuenciación en cada muestra, y la proporción de genes alta- y diferencialmente expresados en cada muestra (el efecto asociado con este último factor se denomina sesgo de *composición*) [36]. Los métodos de normalización no necesariamente se hacen cargo de todos los factores no-biológicos; algunos se obvian bajo la suposición de que sus efectos se cancelan en las comparaciones inter-muestrales de un mismo elemento genómico, sin afectar el análisis de expresión diferencial [36].

Un efecto que se controla en general es el de la profundidad de secuenciación, reescalando las cuentas brutas por factores relacionados con esta. Por ejemplo, herramientas como Cufflinks, RSEM y Salmon multiplican cada cuenta bruta por un factor que depende del largo de la porción exónica del elemento y de la cantidad de lecturas de la muestra de origen efectivamente mapeadas a la referencia, generando valores RPKM (*Reads Per Kilobase of exon model per Million mapped reads*), FPKM (*Fragments Per Kilobase of exon model per Million mapped reads*), u otros relacionados [116]. Métodos más elaborados como TMM (*Trimmed Mean of M-values*) y otros basados en cuantiles, abordan el sesgo de composición al calcular el factor de normalización [116, 64].

En la Tabla A.1 se resumen, para distintas herramientas que cuentan lecturas a nivel de genes, las estrategias para tratar los *multireads* y los métodos de normalización disponibles.

Tabla A.1: Métodos de normalización y manejo de *multireads* de distintas herramientas para el conteo de lecturas a nivel de genes [116].

Herramienta	Normalización	Estrategia para tratar multireads
HTSeq-count	Cuentas brutas	Ignorar
featureCounts	Cuentas brutas	Ignorar, contar en cada alineamiento, distribuir equitativamente
CoCo	Cuentas brutas, CPM, TPM	Distribuir según cantidad de lecturas mapeadas sin ambigüedad
IsoEM2	FPKM, TPM	Distribuir según modelo estadístico
RSEM	Cuentas brutas, TPM, FPKM	Distribuir según modelo estadístico
Rcount	Cuentas brutas	Distribuir según cobertura en la vecindad de los alineamientos

Como se aprecia en la Tabla A.1, varios programas dan la opción de entregar cuentas brutas, lo cual resulta adecuado cuando la herramienta para el análisis de expresión diferencial incluye su propio método de normalización (como es el caso de DESeq2 y edgeR) [64, 119]. En cualquier caso, un formato común para proveer los perfiles de expresión a la siguiente etapa es una tabla (o matriz) de expresión Y, como la representada en la Tabla A.2.

Tabla A.2: Estructura de una matriz de expresión Y que resume la cuenta (total o normalizada) de lecturas mapeadas a cada elemento genómico de interés, en cada muestra. El valor y_{gi} denota la cuenta para el elemento $g \in \{1, ..., G\}$, en la muestra $i \in \{1, ..., N\}$.

Elemento	*Muestra*				
	1	...	i	...	N
1	y_{11}	...	y_{1i}	...	y_{1N}
⋮	⋮		⋮		⋮
g	y_{g1}	...	y_{gi}	...	y_{gN}
⋮	⋮		⋮		⋮
G	y_{G1}	...	y_{Gi}	...	y_{GN}

Una vez obtenidas las cuentas normalizadas para las distintas muestras, una práctica común es visualizar su similitud global mediante métodos de reducción de dimensionalidad como PCA (*Principal Component Analysis*) o MDS (*Multi Dimensional Scaling*) [15, 31]. Tal ejercicio permite observar patrones de agrupamiento (en particular, corroborar si las muestras de condiciones equivalentes se agrupan entre sí), detectar factores que pudieran confundir el posterior análisis de expresión diferencial y detectar eventuales muestras aisladas cuya remoción del análisis pudiera ser conveniente [43].

A.2. Análisis de expresión diferencial

Los experimentos RNA-Seq relevantes para el presente trabajo evalúan la expresión génica −global o por tejido− de individuos en distintas condiciones experimentales, distinguibles entre sí en términos de niveles (valores) con respecto a factores (variables).

En la mayoría de los trabajos considerados se comparan dos condiciones experimentales inducidas por dos tratamientos sobre individuos con infección *Wolbachia* nativa, a saber: el tratamiento nulo y la eliminación total o parcial de la respectiva infección mediante antibióticos, seguida de una recuperación de la flora intestinal [17, 29, 30, 31]. En otras palabras, tales experimentos constan de un factor (tratamiento *Wolbachia*) con dos niveles (eliminación de la infección o tratamiento nulo). Por ejemplo, en la primera parte del estudio de He *et al.* (2019) introducido en la Sección 1.2.2, la respuesta transcriptómica en ovarios de hembras adultas *D. melanogaster* con infección nativa wMel se propuso comparar frente a los dos tratamientos mencionados [30]. Adoptando la nomenclatura de algunos autores, los niveles del factor tratamiento *Wolbachia* asociados con el tratamiento nulo y con la eliminación de la infección se denominarán, respectivamente, WT (*Wild Type*) y GFR (*Gut Flora Restored*).

En otros diseños experimentales se evalúan simultáneamente las respuestas a niveles de múltiples factores, dando la oportunidad de estudiar la manera en que estas se moderan entre sí. Un ejemplo está dado por el trabajo de Lindsey *et al.* (2021) también introducido en la Sección 1.2.2, donde se estudió, simultáneamente con las respuestas a los tratamientos

Wolbachia, la respuesta a dos tipos de inyección (una inocua y una con el virus Sindbis) y a tres tiempos post-inyección (8hpi, 24hpi y 48hpi) [15].

Como se ilustra en la Tabla A.3, las condiciones representadas por cada muestra pueden informarse a los programas que realizan el análisis de expresión diferencial mediante una matriz de diseño X, donde la entrada x_{ip} denota el nivel de la condición que origina a la muestra i, con respecto al factor p [120].

Tabla A.3: Estructura de una matriz de diseño X que indica la condición representada por cada muestra, en términos de niveles de factores. El valor x_{ip} denota el nivel de la muestra $i \in \{1, ..., N\}$ con respecto al factor $p \in \{1, ..., P\}$. Muestras de una misma condición se denominan *réplicas*.

$Muestra$	$Factor$				
	1	...	p	...	P
1	x_{11}	...	x_{1p}	...	x_{1P}
⋮	⋮		⋮		⋮
i	x_{i1}	...	x_{ip}	...	x_{iP}
⋮	⋮		⋮		⋮
N	x_{N1}	...	x_{Np}	...	x_{NP}

En un análisis de expresión diferencial, las condiciones definidas en el diseño experimental (representadas por las muestras) se separan en grupos, donde cada grupo k determina a una población que comprende, virtualmente, todas las muestras que podrían obtenerse en las condiciones pertenecientes a k [120]. Determinar que un elemento g se encuentra diferencialmente expresado entre los grupos, equivale a rechazar una hipótesis nula referida a las poblaciones asociadas a estos, mediante una prueba de hipótesis. Usualmente, la hipótesis nula H_0 se plantea como una igualdad entre los valores esperados de las cuentas normalizadas del elemento g en cada grupo, o de las proporciones de lecturas asociadas al elemento g en cada grupo [120].

Por ejemplo, en edgeR se asume que Y_{gi} sigue una distribución binomial negativa con valor esperado $\mu_{gi} = m_i \lambda^g_{k(i)}$, donde m_i es un factor de normalización que depende del tamaño de la muestra i, y $\lambda^g_{k(i)}$ es el valor esperado de la proporción de lecturas mapeadas al elemento g para muestras del grupo $k(i)$ (en DESeq2 se adopta un enfoque similar pero permitiendo que el factor de normalización sea específico para cada elemento g) [64, 119, 120]. Así, determinar que g se encuentra diferencialmente expresado entre los grupos (1,...,K) equivale a rechazar la hipótesis nula $H_0 : \lambda^g_1 = ... = \lambda^g_K$ con cierto nivel de significancia estadística previamente establecido [120].

Para ilustrar lo anterior, considérese nuevamente el diseño multifactorial de Lindsey *et al.* (2021), donde el análisis de expresión diferencial fue realizado mediante edgeR. Uno de los objetivos del análisis fue determinar genes cuya expresión fuera distinta según el tratamiento *Wolbachia*, independientemente del resto de los factores [15]. Definiendo WT y GFR como los grupos de condiciones con factor ″tratamiento *Wolbachia*″ en nivel WT

y GFR, respectivamente, la tarea descrita consiste en obtener una lista con los elementos g tales que la hipótesis $H_0 : \lambda^g_{WT} = \lambda^g_{GFR}$ puede ser rechazada. Cabe notar que la misma hipótesis nula puede ser planteada como la ausencia de un *efecto principal diferencial* entre los grupos, esto es, $H_0 : \alpha^g_{WT} - \alpha^g_{GFR} = 0$, donde $\alpha^g_k := \lambda^g_k - \lambda^g$ es el *efecto principal* del grupo k (siendo λ^g el valor esperado de la proporción de lecturas mapeadas al elemento g para muestras de la población que incluye a todas las condiciones experimentales) [121].

Como ya se ha mencionado, además de comparar efectos de niveles de un mismo factor, el análisis de expresión diferencial puede usarse para detectar efectos interactivos entre niveles de distintos factores. Por ejemplo, un segundo análisis realizado por Lindsey *et al.* (2021) tuvo como objetivo identificar genes cuya expresión fuera sensible a la interacción entre la eliminación *Wolbachia* y la inyección SINV (esto es, genes g tales que el efecto de SINV sobre la expresión de g fuera moderado por la eliminación de *Wolbachia*) [15]. En tal caso, definiendo $SINV$ como el grupo de condiciones con factor ″tipo de inyección″ en nivel SINV, lo propuesto corresponde a determinar una lista con los elementos g tales que la hipótesis $H_0 : \alpha^g_{GFR \cap SINV} = \alpha^g_{GFR} + \alpha^g_{SINV}$ puede ser rechazada [121]. En otras palabras, para dichos elementos, el efecto combinado de los tratamientos GFR y $SINV$ sobre su expresión no es simplemente la suma de los efectos principales de cada nivel.

Naturalmente, la implementación de las pruebas de hipótesis depende de los modelos estadísticos subyacentes. Aún así, todas comparten la necesidad de afrontar el problema de testear muchas hipótesis simultáneamente, que torna inadecuados los procedimientos comunes para declarar significancia mediante el uso directo de *p-values* [122]. Al realizar múltiples pruebas, aplicar el método usual de declarar significancia cada vez que el *p-value* de una prueba sea menor que un valor de corte predefinido (α, el nivel de significancia), controla el valor esperado de la razón de falsos positivos sobre hipótesis nulas ciertas (*False Positive Rate* o *FPR*), lo cual puede admitir excesivos falsos positivos cuando tales hipótesis son abundantes [122]. Una cantidad cuyo control se considera más adecuado es el valor esperado de la razón de falsos positivos sobre positivos totales (*False Discovery Rate* o *FDR*). En general, los programas para el análisis de expresión diferencial computan el conjunto de *p-values* para todas las prueba de hipótesis y definen, en función de dicho conjunto, una transformación que a cada *p-value* le asigna un nuevo valor, que será denominado *adjusted p-value*. La transformación es tal que decidir la significancia de las pruebas con base en el criterio *adjusted p-value* $\leq \alpha$, asegura que $FDR \leq \alpha$ [122].

Así, la selección de los elementos D.E. se realiza definiendo un valor de corte para los *adjusted p-value*, siendo común el uso de 0.05 o 0.01 [15, 17, 30, 31]. Para dirigir el análisis hacia los cambios de expresión más pronunciados, algunos autores imponen restricciones sobre el tamaño de los efectos diferenciales. En la comparación de la expresión del elemento g en dos grupos 1 y 2, a menudo el tamaño del efecto diferencial se mide en términos del $log_2(FC)$, donde $FC = \hat{\lambda^g_1}/\hat{\lambda^g_2}$, siendo $\hat{\lambda^g_i}$ un estimador de λ^g_i [64, 119].

Anexo B

Detalles adicionales sobre la Metodología

En este Anexo se precisan algunos aspectos de la Metodología que fueron omitidos en el cuerpo del informe.

B.1. Recuperación de identificadores a partir de tablas publicadas

Para obtener listas de elementos D.E. aptas para someter a análisis de enriquecimiento funcional a partir de las tablas descargadas de Caragata *et al.* (2017) y Baiao *et al.* (2019), se requirió un trabajo adicional de búsqueda de IDs de ortólogos en el respectivo organismo de referencia aquí escogido para tal análisis (esto es, *A. aegypti* para especies Aedes y *D. melanogaster* para especies Drosophila). La razón es que, en dichos trabajos, la identificación de las secuencias expresadas en los hospederos no-modelo (*Aedes fluviatilis* y *Drosophila paulistorum*) fue mediada por ensambles de transcriptoma *de novo*, y la posterior caracterización de los transcritos ensamblados (basada en búsquedas BLAST) no siempre les asignó un identificador del organismo de referencia. Por ejemplo, en el estudio de Caragata *et al.* (2017) varios de los productos génicos que resultaron mejores coincidencias BLAST de transcritos ensamblados correspondieron a organismos distintos de *Aedes aegypti*, como *Culex pipiens* [17]. Similarmente, en el estudio de Baiao *et al.* (2019) algunas de las mejores coincidencias BLAST correspondieron a *Drosophila wilistoni*, distinto del organismo de referencia aquí escogido para el análisis de enriquecimiento, *Drosophila melanogaster* [31].

Los procesos de decisión definidos para la recuperación de IDs mediada por ortología dependieron de los tipos específicos de información contenida en las tablas descargadas, los cuales se explicitarán mediante sus encabezados. A pesar de sus particularidades, ambos métodos tuvieron en común el objetivo de recuperar tantas IDs del respectivo organismo de referencia como fuera posible, manteniendo un criterio estándar para inferir ortología.

En la Tabla B.1 se presentan los principales campos del encabezado de las tablas descargadas desde el material suplementario de Caragata *et al.* (2017)

Tabla B.1: Encabezado de tablas suplementarias de Caragata *et al.* (2017)

Sequence Name	BLAST hit	RPKM - wFlu	RPKM - Tet
comp13612_c0_seq1	gi|403182639|gb|EAT44218.2|AAEL004390-PA [Aedes aegypti]	287.11	168.42
comp7819_c0_seq1	gi|157142081|ref|XP_001647805.1|Niemann-Pick Type C-2, putative [Aedes aegypti]	268.16	88.12
comp4837_c0_seq1	gi|56417510|gb|AAV90696.1|putative salivary basic peptide 4.2k-1 [Aedes albopictus]	259.44	69.40

Tal como se observa en la Tabla B.1, los elementos D.E. reportados correspondieron a productos génicos identificados mediante un campo compuesto (BLAST hit), que constituye la entrada para el proceso de recuperación de IDs representado en la Figura B.1.

(El siguiente espacio se ha dejado en blanco intencionalmente)

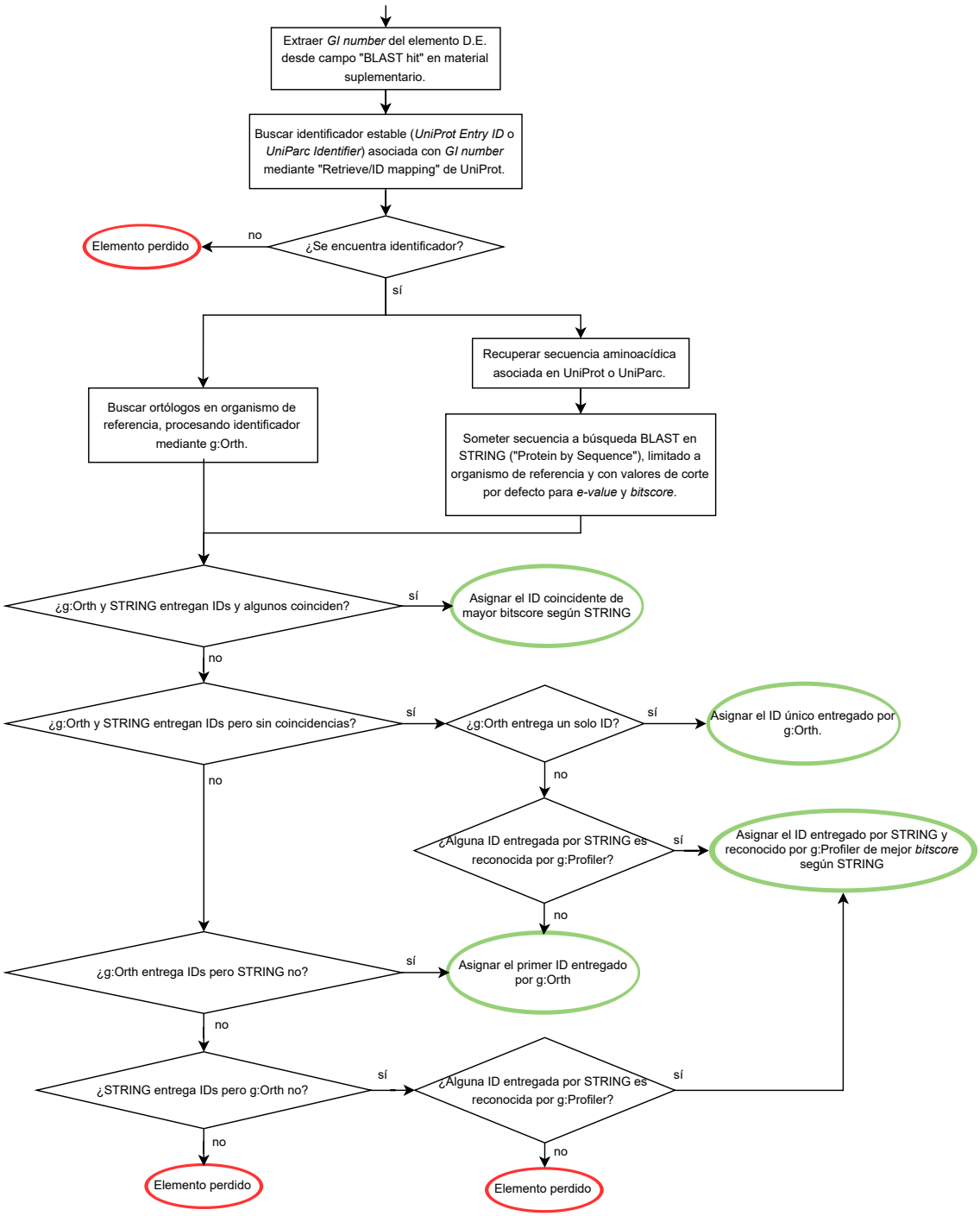

Figura B.1: Proceso de decisión seguido para la obtención de identificadores de *A. aegypti* a partir de los elementos D.E. publicados en Caragata *et al.* (2017).

En la Tabla B.2 se presentan los principales campos del encabezado de las tablas descargadas desde el material suplementario de Baiao *et al.* (2019)

Tabla B.2: Encabezado de tablas suplementarias de Baiao *et al.* (2019).

Trinity_Transcript	log2FC	qvalue	DmelFBgn	DwilFBgn	gene_name
TRINITY_DN94708_c0_g1_i1	2.65	1.511E-10	FBgn0034407	FBgn0222924	DptB
TRINITY_DN57947_c1_g1_i1	2.42	2.3229E-09			
TRINITY_DN79847_c0_g1_i1	2.08	3.7198E-09	FBgn0041579	FBgn0219821	AttC

Como se aprecia en la Tabla B.2, los elementos D.E. reportados correspondieron a genes identificados mediante dos campos opcionales (DmelFBgn y DwilFBgn) referidos a *D. melanogaster* y *D. willistoni*, respectivamente. Dichos campos fueron usados como entrada para el proceso de decisión ilustrado en la Figura B.2.

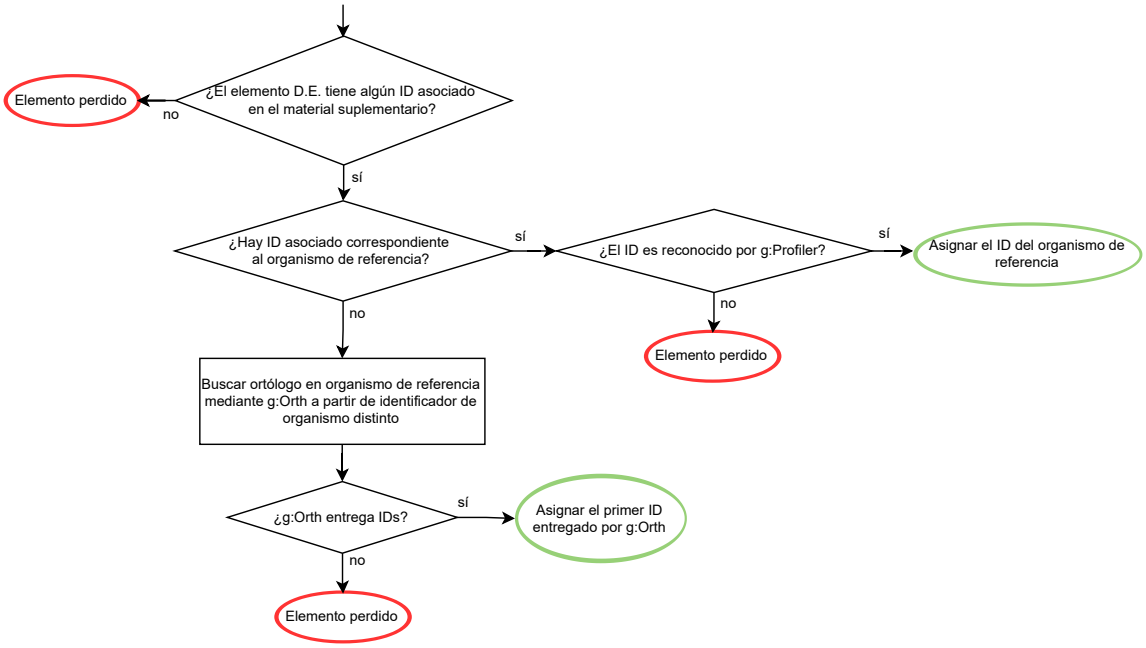

Figura B.2: Proceso de decisión seguido para la obtención de identificadores de *D. melanogaster* a partir de los elementos D.E. publicados en Baiao *et al.* (2019).

B.2. Análisis de enriquecimiento funcional y generación de *heatmaps*

En esta sección se presentan los códigos en Python y R que mediaron los análisis de enriquecimiento de términos GO y la generación de *heatmaps* con niveles de expresión.

El insumo principal para el código en Python presentado en esta sección −consistente en funciones auxiliares (Códigos B.1 a B.7) y un *script* (Código B.8)− fueron las listas de genes D.E. ordenadas decrecientemente según la magnitud del log_2FC, en el formato descrito en la Tabla B.3.

Tabla B.3: Formato de las listas D.E. sometidas a análisis de enriquecimiento de términos GO

Enrichment_ID	Expression_ID	log2FC
AAEL005377	nan	4.17
AAEL025268	FBgn0003435	3.93
...	...	...

Cabe notar que en el formato representado en la Tabla B.3, cada elemento se asocia con dos identificadores. El primer identificador, *Enrichment_ID*, es el que se usa para el análisis de enriquecimiento de términos GO, y consiste en un identificador de *D. melanogaster* en el caso de los organismos *Drosophila*, o en un identificador de *A. aegypti* en el caso de los organismos *Aedes*. Tal identificador es el que fue hallado mediante un análisis RNA-Seq propio, o bien desde tablas mediante los procesos de decisión especificados en el Anexo B.1.

El segundo identificador, *Expression_ID*, es el que se usa para la comparación de los niveles de expresión entre las distintas listas, a través de *heatmaps*. En el caso de los organismos *Drosophila*, el campo *Expression_ID* es equivalente al campo *Enrichment_ID*. En el caso de los organismos *Aedes*, el campo *Expression_ID* contiene el primer identificador de *D. melanogaster* entregado por el módulo g:Orth de g:Profiler, ejecutado sobre el identificador de *A. aegypti* que se encuentra en el campo *Enrichment_ID*.

Los productos del procedimiento en Python corresponden a: (i) los archivos GEM con términos enriquecidos para cada lista D.E., filtrados según tamaño y grado (número de listas D.E. en que un término se encuentra enriquecido); y (ii) el archivo de expresión que reúne los log_2FC de todos los ortólogos de *D. melanogaster* en cada lista D.E. (posteriormente proveido a Cytoscape). En las Tablas B.4 y B.5 se especifican los formatos de los archivos de salida mencionados.

Tabla B.4: Formato de los archivos GEM generados mediante rutina Python para cada lista D.E. Los campos que efectivamente se usan en la visualización de términos enriquecidos mediante Cytoscape son GO.ID, y FDR, que contienen el identificador GO del término y su FDR en la lista D.E., respectivamente

GO.ID	Description	p.Val	FDR	Phenotype	Genes
GO:0005201	extracellular matrix structural constituent	0.155	0.155	+1	FBGN0000000
GO:0046463	acylglycerol biosynthetic process	0.052	0.052	+1	FBGN0000000
...	...	...	...	...	...

Tabla B.5: Formato del archivo de expresión generado mediante rutina Python. La primera columna contiene los nombres de uso común de los ortólogos *D. melanogaster*. De la tercera columna en adelante los campos contienen el log_2FC de los elementos en cada lista D.E.

name	description	Lind_up	Lind_down	...	Car_up	Car_down
CG13029	None	4.31	0	...	0	0
Lsp1beta	Larval serum protein 1 beta	3.3	0	...	0	0
...	...	...	...	...	...	...

Por otra parte, el insumo principal para el código en R presentado en esta sección −consistente en un único *script* (Código B.9)− fue una carpeta con los sub-archivos de expresión exportados desde Cytoscape para cada módulo funcional. El formato de estos sub-archivos de expresión es equivalente al del archivo de expresión generado mediante Python (representado previamente en la Tabla B.5), excepto por la inclusión de una columna "rank" a la derecha de la columna "description". Los productos del código en R son los *heatmaps* en formato PNG resumiendo el log_2FC a través de las listas, para cada ortólogo de *D. melanogaster* asociado con cada módulo funcional.

A continuación se presentan los códigos en Python y R previamente anunciados, con sus respectivos comentarios.

Código B.1: Definición de función *prepare_query*

```
def prepare_query(list_name,src_dir,query_type):
    """
    Entrega una lista con identificadores a someter a consulta en g:GOSt (análisis de enriquecimiento) o en g:Convert
    (conversión entre namespaces), a partir de una lista D.E. con encabezado [Enrichment_ID , Expression_ID , log2FC]

    Parametros:
    -----------
        list_name (string): Nombre de la lista D.E.
        src_dir (string): Nombre del directorio en que se encuentra la lista D.E.
        query_type (string): Determina tipo de consulta a realizar ('enrichment' para g:GOSt o 'expression' para g:Convert).
    Salida:
    -------
        query: Lista con los identificadores a someter a consulta
    """
    id_type = ''
    if query_type == 'enrichment':
        id_type = 'Enrichment_ID'
    if query_type == 'expression':
        id_type = 'Expression_ID'
    query = []
    field_index = {}
    row_count = 0
    with open(src_dir + '/' + list_name + '.txt') as actual_list:
        csv_reader = csv.reader(actual_list, delimiter='\t')
        for row in csv_reader:
            if row_count == 0: #Primera línea, encabezado lista
                for i in range(0,len(row)):
                    field_index[row[i]]=i
                row_count+=1
                print(row)
            else:
                query.append(row[field_index[id_type]])
                row_count+=1
    return query
```

Código B.2: Definición de función *query_gOSt*

```
def query_gGOSt(list_name,src_dir,organism,ordered,sources,adjusting_method,adjpv_threshold):
    """
    Realiza analisis de enriquecimiento de terminos funcionales a partir de una lista D.E. con encabezado
    [Enrichment_ID , Expression_ID , log2FC], mediante una consulta en g:GOSt via la API de g:Profiler.

    Parametros:
    -----------
        list_name (string): Nombre de la lista D.E.
        src_dir (string): Nombre del directorio en que se encuentra la lista D.E.
        organism (string): Identificador del organismo de referencia (por ejemplo, 'dmelanogaster' para D. melanogaster)
        ordered (boolean): Indica si la consulta se trata como lista ordenada.
        sources (List): Fuentes de terminos funcionales a testear (ej.: ["GO:MF","GO:CC","GO:BP","KEGG"])
        adjusting_method (string): Metodo de ajuste de p-values
        adjpv_threshold (float): Valor de corte para los p-value ajustados.

    Salida:
    -------
        r: Resultado del analisis de enriquecimiento en g:GOSt para la lista D.E. ingresada.
    """

    query = prepare_query(list_name,src_dir,'enrichment')

    r = requests.post(
        url='https://biit.cs.ut.ee/gprofiler/api/gost/profile/',
        json={
        'organism':organism,
        'query':query,
        'sources' :sources,
        'ordered':ordered,
        'user_threshold':adjpv_threshold,
        'significance_threshold_method':adjusting_method,
        'domain_scope':'annotated', #Usar genes anotados como conjunto de genes de fondo
        },
        headers={
        'User-Agent':'FullPythonRequest'
        }
    )

    return r
```

Código B.3: Definición de función *multiple_queries_gOSt*

```
def multiple_queries_gOSt(lists_by_ref,src_dir,ordered,sources,adjusting_method,adjpv_threshold):
    """
    Realiza analisis de enriquecimiento de terminos funcionales sobre múltiples listas. Los nombres de las listas y su
    ↪ respectivo organismo de referencia para el análisis son informados a la función mediante un diccionario con la
    ↪ siguiente estructura lists_by_ref = {'ref_1': ['Lista_11', 'Lista_12', ...], 'ref_2': ['Lista_21', 'Lista_22', ...], ... },
    ↪ donde 'Lista_ij' es el nombre de la j-ésima lista D.E. cuyo organismo de referencia es 'ref_i'. Nuevamente, cada
    ↪ lista tiene encabezado [Enrichment_ID , Expression_ID , log2FC].

    Parametros:
    -----------
        lists_by_ref (Dict): Diccionario que contiene el nombre de las listas D.E. asociadas a sus organismos de referencia.
        src_dir (string): Nombre del directorio en que se encuentran las listas D.E.
        ordered (boolean): Indica si las consultas se tratan como listas ordenadas.
        sources (List): Fuentes de terminos funcionales a testear (ej.: ["GO:MF","GO:CC","GO:BP","KEGG"])
        adjusting_method (string): Metodo de ajuste de p-values.
        adjpv_threshold (float): Valor de corte para los p-value ajustados.

    Salida:
    -------
        term_dict: Diccionario donde cada identificador de término GO se asocia con su nombre, su tamaño, su grado (# de
        ↪ listas en que aparece enriquecido) y sus adj. p-values en cada lista. Un ejemplo de par llave-valor de dicho
        ↪ diccionario es el siguiente:

                term_dict['GO:0015291'] == {
                                    'term_name': 'secondary active transporter activity',
                                    'term_size': 129,
                                    'adjpvalues': {'Lind_up': 0.072, 'Lind_down':0.087, ...},
                                    'term_degree': 7
                                    }
    """

    result_dict = {} # Almacena transitoriamente los resultados de los analisis de enriquecimiento sobre cada lista DE
    for ref in lists_by_ref: #Rutina para hacer los analisis y almacenarlos en result_dict.
        for actual_list in lists_by_ref[ref]:
            result_dict[actual_list]=query_gOSt(actual_list,src_dir,ref,ordered,sources,adjusting_method,adjpv_threshold)

    term_dict = {}
    for actual_list in result_dict: #Se llena el dicc. con los ID de los terminos, tamaño y sus adjpvalues en cada lista
        actual_r = result_dict[actual_list]
        actual_result = actual_r.json()['result']
        for term in actual_result:
            term_id = term['native']
            term_name = term['name']
            term_size = term['term_size']
            term_adjpv = term['p_value']

            if term_id not in term_dict: #Si el termino no ha sido agregado
                term_dict[term_id]={ 'term_name':term_name , 'term_size':term_size , 'adjpvalues':{} }
                for aux_list in result_dict:
                    term_dict[term_id]['adjpvalues'][aux_list] = 1 #Se inicializan los adjpvalues en 1

            term_dict[term_id]['adjpvalues'][actual_list] = term_adjpv

    for term_id in term_dict: #Se calcula y anota el grado de cada termino
        term_degree = 0
        for key in term_dict[term_id]['adjpvalues']:
            if term_dict[term_id]['adjpvalues'][key]<=adjpv_threshold:
                term_degree+=1
        term_dict[term_id]['term_degree'] = term_degree

    return term_dict
```

Código B.4: Definición de función *create_GEM_files*

```
def create_GEM_files(lists_by_ref,order_in_EM,src_dir,ordered,sources,adjusting_method,adjpv_threshold,
                     min_term_size,max_term_size,min_term_degree,tgt_dir):

    """
    Guarda en disco los archivos GEM (General Enrichment Map) correspondientes a cada lista D.E. a partir de los
      ↪ resultados del analisis de enriquecimiento en gGOSt.

    Parámetros:
    -----------
        lists_by_ref (Dict): Diccionario que contiene el nombre de las listas D.E. asociadas a sus organismos de referencia.
        src_dir (string): Nombre del directorio en que se encuentran las listas D.E.
        ordered (boolean): Indica si las consultas se tratan como listas ordenadas.
        sources (List): Fuentes de terminos funcionales a testear (ej.: ["GO:MF","GO:CC","GO:BP","KEGG"])
        adjusting_method (string): Metodo de ajuste de p-values.
        adjpv_threshold (float): Valor de corte para los p-value ajustados.
        min_term_size (int): Minimo tamaño admitido para los terminos a considerar.
        max_term_size (int): Maximo tamaño admitido para los terminos a considerar.
        min_term_degree (int): Minimo grado admitido para los terminos a considerar.
        tgt_dir (string): Nombre del directorio en que se almacenaran los archivos GEM generados.

    """
    term_dict = multiple_queries_gGOSt(lists_by_ref,src_dir,ordered,sources,adjusting_method,adjpv_threshold)
    header = ['GO.ID','Description','p.Val','FDR','Phenotype','Genes']

    for ref in lists_by_ref:
        for actual_list in lists_by_ref[ref]:
            with open(tgt_dir + '/' + order_in_EM[actual_list] + '_' + actual_list + '.gem.txt','w',newline='') as f:

                # Se escribe el encabezado
                csv_writer = csv.writer(f,delimiter='\t')
                csv_writer.writerow(header)

                for term_id in term_dict:
                    term_degree = term_dict[term_id]['term_degree']
                    term_name = term_dict[term_id]['term_name']
                    term_size = term_dict[term_id]['term_size']
                    actual_adjpv = term_dict[term_id]['adjpvalues'][actual_list]

                    #Solo si el termino cumple todos los requisitos sera considerado
                    if (actual_adjpv <= adjpv_threshold and term_degree >= min_term_degree and
                        term_size >= min_term_size and term_size <= max_term_size):
                        new_row = [term_id,term_name,actual_adjpv,actual_adjpv,'+1','FBGN0000000']
                        csv_writer.writerow(new_row)
```

Código B.5: Definición de función *query_gConvert*

```
def query_gConvert(list_name,src_dir,organism,target):
    """
    Realiza una consulta en g:Convert para mapear identificadores de una lista D.E. con encabezado
    [Enrichment_ID , Expression_ID , log2FC], de un namespace a otro.

    Parámetros:
    -----------
        list_name (string): Nombre de la lista cuyos elementos se quieren mapear.
        src_dir (string): Nombre del directorio donde se ubica la lista D.E.
        organism (string): Nombre del organismo de referencia.
        target (string): Nombre del namespace objetivo (por ejemplo 'ENSG')

    Salida:
    -------
        r: Resultado de la conversión de identificadores en g:Convert para la lista D.E. ingresada.

    """

    query = prepare_query(list_name,src_dir,'expression')

    r = requests.post(
        url='https://biit.cs.ut.ee/gprofiler/api/convert/convert/',
        json={
            'organism':organism,
            'target':target,  #Ejemplo: 'ENSG' para identificadores de Ensembl
            'query':query,
        }
        )

    return r
```

Código B.6: Definición de función *create_ENSG_to_name_dict*

```
def create_ENSG_to_name_dict(lists_by_ref,src_dir,organism):

    """
    Crea un diccionario que mapea identificadores génicos Ensembl a sus nombres de uso común.
    Los nombres de las listas y su respectivo organismo de referencia para el análisis son informados
    a la función mediante un diccionario con la siguiente estructura :
    lists_by_ref = {'ref_1': ['Lista_11', 'Lista_12', ...], 'ref_2': ['Lista_21', 'Lista_22', ...], ... },
    donde 'Lista_ij' es el nombre de la j-ésima lista D.E. cuyo organismo de referencia es 'ref_i'.


    Parámetros:
    -----------
        lists_by_ref (Dict): Diccionario que contiene el nombre de las listas D.E. asociadas a sus organismos de referencia.
        src_dir (string): Nombre del directorio donde se ubica la lista D.E.
        organism (string): Nombre del organismo de referencia.

    Salida:
    -------
        ENSG_to_name_dict: Diccionario donde los identificadores de Ensembl que aparecen en las listas ingresadas se
        ↪ asocian
                        a su nombre de uso común y a su descripción. Un ejemplo de par llave-valor de dicho diccionario es
                        el siguiente:

                        ENSG_to_name_dict['FBgn0284435'] == {
                                                'name': 'tyn',
                                                'description': trynity
                                                }

    """

    ENSG_to_name_dict = {}
    for ref in lists_by_ref:
        for actual_list in lists_by_ref[ref]:
            r = query_gConvert(actual_list,src_dir,organism,'ENSG')
            query_gConvert_result = r.json()['result']

            for gene_entry in query_gConvert_result:
                ENSG_ID = gene_entry['converted']
                name = gene_entry['name']
                description = gene_entry['description']
                if ENSG_ID not in ENSG_to_name_dict: #si no hay entrada ENSG_ID en el diccionario
                    ENSG_to_name_dict[ENSG_ID] = {'name' : name , 'description' : description}
                    print('Entrada nueva con llave ' + ENSG_ID + ' y valor {name:' + name + ',description:' + description + '}
        ↪ ')
                else: #ya hay entrada con llave ENSG_ID en diccionario
                    if name != ENSG_to_name_dict[ENSG_ID]['name']: #name actual es el valor asociado con la entrada
        ↪ ENSG_ID que ya existe
                        print(name + ' es un nuevo name asociado con el ENSG_ID ' + ENSG_ID + '. Se conserva el anterior.')
                    else: #if name in ENSG_to_name_dict[ENSG_ID]
                        print(name + ' ya se encuentra asociado con el ENSG_ID ' + ENSG_ID)
    return ENSG_to_name_dict
```

Código B.7: Definición de función *create_expression_file*

```
def create_expression_file(lists_by_ref,src_dir,organism,output_filename):
    """
    Guarda en disco un archivo de expresión donde, para cada elemento génico que aparece en al menos una lista D.E.
      ↪ ingresada, se muestra el log2FC en cada una de las listas D.E. ingresadas.

    Parámetros:
    -----------
        lists_by_ref (Dict): Diccionario que contiene el nombre de las listas D.E., asociados a sus organismos de referencia.
        src_dir (string): Nombre del directorio donde se ubica la lista D.E.
        organism (string): Nombre del organismo de referencia.
        output_filename (string): Nombre asignado al archivo de expresión
    """
    #Se genera diccionario que contiene los mapeos de todos los ENSG_ID de las listas D.E., a su name y su description
    ENSG_to_name_dict = create_ENSG_to_name_dict(lists_by_ref,src_dir,organism)
    expression_dict = {} #Diccionario que contendrá los mapeos de los ENSG_ID
    header = ['name','description'] #Encabezado (incompleto) del expression file

    #Inicializacion del diccionario con log2FC = 0 para cada lista D.E., por defecto
    for ENSG_ID in ENSG_to_name_dict:
        if ENSG_ID != 'nan': #Si el ID fue reconocido por g:Convert
            name = ENSG_to_name_dict[ENSG_ID]['name']
            description = ENSG_to_name_dict[ENSG_ID]['description']
            expression_dict[ENSG_ID]={'name':name,'description':description}
            for ref in lists_by_ref:
                for actual_list in lists_by_ref[ref]:
                    expression_dict[ENSG_ID][actual_list] = 0 # Por defecto el nivel de expresion es 0

    #Rellenado de expression_dict con los log2FC de cada lista D.E.
    for ref in lists_by_ref:
        for actual_list in lists_by_ref[ref]:
            header.append(actual_list) #Se añade al encabezado un campo para la lista D.E. actual
            with open(src_dir + '/' + actual_list + '.txt') as actual_list_file:
                field_index = {}
                row_count = 0 #Se regenera el contador de filas
                csv_reader = csv.reader(actual_list_file, delimiter='\t')
                for row in csv_reader:
                    if row_count == 0: #Primera línea, encabezado lista
                        for i in range(0,len(row)):
                            field_index[row[i]]=i
                        row_count+=1
                    else:
                        row_count+=1
                        ENSG_ID = row[field_index['Expression_ID']]
                        logFC = row[field_index['log2FC']]
                        if ENSG_ID in expression_dict: #Si el ENSG_ID pudo ser mapeado a un nombre de uso común
                            #Reemplazo de valores infinitos por valores numericos grandes
                            if logFC == 'Inf':
                                logFC = 10
                            if logFC == '-Inf':
                                logFC = -10
                            expression_dict[ENSG_ID][actual_list] = logFC
                        else: #Si el ENSG_ID fue mapeado a 'name'= nan
                            print('Par [ENSG_ID,log2FC], donde ' + ENSG_ID + ' no es reconocida por g:Convert.')

    with open(output_filename,'w',newline='') as f: #Escritura del archivo de expresion
        csv_writer = csv.writer(f,delimiter='\t')
        csv_writer.writerow(header) #Escribir el encabezado
        for ENSG_ID in expression_dict:
            name = expression_dict[ENSG_ID]['name']
            description = expression_dict[ENSG_ID]['description']
            new_row = [name,description]
            for ref in lists_by_ref:
                for actual_list in lists_by_ref[ref]:
                    new_row.append(expression_dict[ENSG_ID][actual_list])
            csv_writer.writerow(new_row)
```

Código B.8: Rutina para generar archivos GEM y archivo de expresión

```
#1. IMPORTACIÓN DE MÓDULOS Y DEFINICIÓN DE CONSTANTES
import csv
import os
import pandas as pd
import requests

ordered = True # Binario que indica si las listas D.E. se procesan considerando el orden de los elementos.
adjusting_method = 'fdr' # Metodo de ajuste de p-values
adjpv_threshold = 0.20 # Valor de corte para los p-value ajustados en las pruebas de enriquecimiento sobre las listas D.E.
src_dir = 'Ordered_DE_IDs' #Directorio de origen (donde se encuentran las listas DE a procesar)
tgt_dir = 'Definitive_Filtered_GEMs' #Directorio de destino (donde se almacenan los archivos generados)
sources = ["GO:MF","GO:CC","GO:BP"] # Fuentes de términos funcionales a visualizar mediante EnrichmentMap
min_term_degree = 4 # Nro. mínimo de listas D.E. en que deben estar enriquecidos los términos para visualizarse
max_term_size = 350 # Tamaño máximo de los términos a incluirse en visualizaciones
min_term_size = 4 # Tamaño mínimo de los términos a incluirse en visualizaciones
organism = 'dmelanogaster' #Organismo de referencia para archivos de expresión

#Listas de elementos DE, asociadas via diccionario a su respectivo organismo de referencia para analisis de enriquecimiento
lists_by_ref = {'dmelanogaster':
            [
               "Lind_up",
               "Lind_down",
               "He_up",
               "He_down",
               "Bai_F_abd_up",
               "Bai_F_abd_down",
               "Bai_F_head_up",
               "Bai_F_head_down",
               "Bai_M_abd_up",
               "Bai_M_abd_down",
               "Bai_M_head_up",
               "Bai_M_head_down",
               "Det_up",
               "Det_down",
            ],
            'aalvpagwg':
            [
               "Car_up",
               "Car_down"
            ]
          }

# Orden deseado de las listas para visualizacion en EnrichmentMap
order_in_EM = {"Bai_F_head_up":'a',
            "Bai_M_head_up":'b',
            "Lind_up":'c',
            "Det_up":'d',
            "Car_up":'e',
            "He_up":'f',
            "Bai_M_abd_up":'g',
            "Bai_F_abd_up":'h',
            "Bai_F_abd_down":'i',
            "Bai_M_abd_down":'j',
            "He_down":'k',
            "Car_down":'l',
            "Det_down":'m',
            "Lind_down":'n',
            "Bai_M_head_down":'o',
            "Bai_F_head_down":'p'}

#2. CREAR ARCHIVOS GEM
create_GEM_files(lists_by_ref,order_in_EM,src_dir,ordered,sources,adjusting_method,adjpv_threshold,
                 min_term_size,max_term_size,min_term_degree,tgt_dir)

#3. CREAR ARCHIVOS DE EXPRESIÓN
create_expression_file(lists_by_ref,src_dir,organism,'Expression_File.txt')
```

Código B.9: Rutina en R para generar heatmaps a partir de sub-archivos de expresión exportados desde Cytoscape

```r
library("gplots")
library(stringi)

setwd("dir") #Reemplazar "dir" por dirección de carpeta con los sub-archivos de expresión exportados desde Cytoscape

filenames <- list.files() #Se listan los archivos del directorio actual

#A continuación se carga cada sub-archivo de expresión, y se extrae una matriz numérica desde este,
#la cual se visualiza en forma de heatmap y se exporta en formato png hacia una carpeta llamada "HeatMaps",
#ubicada en el mismo directorio actual
for (filename in filenames)
{
 if (stri_sub(filename,-3)=='txt')
 {
  imp <- read.delim2(filename , row.names=1) #Se lee el archivo actual
  df <- imp[,-c(1,2)] #Se remueven campos no numéricos
  df <- na.omit(df) #Se omiten valores nan
  matriz <- as.matrix(df) #Se transforma a matriz

  #Las siguientes dos operaciones truncan los valores log2FC demasiado extremos, para evitar que
  #la paleta de colores pierda sensibilidad con respecto a los log2FC de menor magnitud
  matriz[matriz>5] <- 5
  matriz[matriz<-5] <- -5

  g_o <- 52 #Parámetro que influirá en el tamaño vertical de las filas del heatmap (se modificará según el nro. de filas)

  if(nrow(matriz)<30){ #Si el heatmap contiene pocas filas, se aumenta g_o para que cada fila tenga más espacio
    g_o <- 62
  }

  g <- g_o -(5/7)*(nrow(matriz) - 6) #Variable que determinará el tamaño vertical de las filas (dependiente de g_o)

  rowfont_size = 0.7 #Tamaño de fuente para las etiquetas de las filas (nombres de los genes)

  if(g<=4) #Si g acusa demasiadas filas, se disminuye el tamaño de fuente
  {
    g=4
    rowfont_size = 0.5
  }

  truncated_filename <- substring(filename, 1, nchar(filename)-4) #Nombre del sub-archivo de expresión, sin el sufijo "txt"
  png(paste("HeatMaps/",truncated_filename,".png",sep=""),width=12,height=12,units='in',res=300) #Se declara el
    ↪ archivo png a exportar

  #Finalmente se genera y guarda el heatmap en el archivo png
  heatmap.2(matriz , scale='none', col=bluered(100) , dendrogram='none' , trace='none' , key='TRUE', keysize=0.5 ,
    ↪ key.title=NA,srtCol=30,density.info='none',  margins = c(g,20), lhei = c(0.7,8), lwid = c(0.5,2.0), cexRow=rowfont
    ↪ _size,cexCol=0.82)

  dev.off()
 }
}
```

Anexo C

Listas de elementos D.E. completas

En las Tablas C.1 a C.16 se presentan las listas D.E. obtenidas. Los identificadores son los utilizados para el análisis de enriquecimiento de términos GO, esto es, los correspondientes al campo *Enrichment_ID* según la nomenclatura de la Tabla B.3.

Tabla C.1: Lista de genes sobreexpresados en hembras *A. fluviatilis*-wFlu completas (Car_up), obtenida desde Caragata *et al.* (2017).

Gene ID	log2FC	Gene ID	log2FC
AAEL007668	Inf	AAEL014617	1.48
AAEL027774	Inf	AAEL004931	1.47
AAEL005924	4.43	AAEL006720	1.47
AAEL005377	4.17	AAEL012383	1.46
AAEL025268	3.93	AAEL013170	1.4
AAEL019684	3.75	AAEL002360	1.38
AAEL009209	3.49	AAEL007788	1.37
AAEL002292	3.3	AAEL003861	1.35
AAEL007536	2.93	AAEL000500	1.25
AAEL013487	2.84	AAEL007063	1.22
AAEL004712	2.65	AAEL003066	1.19
AAEL024813	2.56	AAEL002908	1.19
AAEL003190	2.43	AAEL006585	1.18
AAEL006376	2.41	AAEL009432	1.1
AAEL009625	2.25	AAEL019957	1.09
AAEL005375	2.24	AAEL005064	1.05
AAEL018189	2.24	AAEL008543	1.02
AAEL015265	2.18	AAEL026278	1.01
AAEL013547	2.17	AAEL017320	1.01
AAEL012837	2.13	AAEL002306	0.99
AAEL020585	2.03	AAEL012931	0.97
AAEL009476	1.9	AAEL004213	0.96
AAEL005607	1.88	AAEL027688	0.93
AAEL004366	1.87	AAEL007064	0.92
AAEL004719	1.86	AAEL019603	0.91
AAEL015051	1.8	AAEL007772	0.9
AAEL013856	1.76	AAEL005533	0.89
AAEL007292	1.75	AAEL015682	0.86
AAEL029062	1.71	AAEL006845	0.83
AAEL019602	1.64	AAEL015048	0.79
AAEL015136	1.61	AAEL004048	0.77
AAEL001140	1.6	AAEL004390	0.77
AAEL010381	1.6	AAEL000757	0.76
AAEL007815	1.56	AAEL010228	0.75

Tabla C.2: Lista de genes subexpresados en hembras *A. fluviatilis*-wFlu completas (Car_down), obtenida desde Caragata *et al.* (2017).

Gene ID	log2FC	Gene ID	log2FC
AAEL012873	-Inf	AAEL003035	-1.57
AAEL006498	-Inf	AAEL011908	-1.56
AAEL006128	-4.16	AAEL008108	-1.52
AAEL008527	-4.16	AAEL008753	-1.51
AAEL020954	-3.86	AAEL009615	-1.41
AAEL006946	-3.85	AAEL005621	-1.39
AAEL013971	-3.74	AAEL005014	-1.28
AAEL026431	-3.48	AAEL029025	-1.26
AAEL006563	-3.08	AAEL000028	-1.24
AAEL008553	-2.91	AAEL004042	-1.23
AAEL000797	-2.82	AAEL006568	-1.23
AAEL020442	-2.82	AAEL000669	-1.23
AAEL009052	-2.74	AAEL011979	-1.19
AAEL000419	-2.71	AAEL019616	-1.17
AAEL014451	-2.62	AAEL001020	-1.17
AAEL020040	-2.6	AAEL029058	-1.12
AAEL008543	-2.55	AAEL006259	-1.08
AAEL006232	-2.34	AAEL008485	-1.03
AAEL004603	-2.18	AAEL010180	-1.02
AAEL002367	-2.09	AAEL004023	-1.02
AAEL010506	-2.07	AAEL005768	-1.01
AAEL001494	-2.06	AAEL006406	-0.99
AAEL006423	-2.06	AAEL019550	-0.95
AAEL009528	-1.73		

Tabla C.3: Lista de genes sobreexpresados en ovarios de *D. melanogaster*-wMel (He_up), obtenida desde He *et al.* (2019).

Gene ID	log2FC	Gene ID	log2FC
FBgn0034837	3.2	FBgn0038718	1.23
FBgn0263589	3.04	FBgn0052364	1.13
FBgn0031879	2.43	FBgn0032116	1.13
FBgn0026175	1.72	FBgn0028886	1.06
FBgn0050401	1.67	FBgn0250757	1.05
FBgn0028396	1.56	FBgn0031337	1.04
FBgn0265457	1.55	FBgn0033093	1
FBgn0030999	1.55	FBgn0085428	0.87
FBgn0032084	1.54	FBgn0260005	0.87
FBgn0262104	1.5	FBgn0267728	0.85
FBgn0040370	1.49	FBgn0262895	0.85
FBgn0031523	1.48	FBgn0039350	0.77
FBgn0046776	1.31	FBgn0036690	0.67
FBgn0038658	1.3	FBgn0039993	0.58
FBgn0039452	1.3	FBgn0032783	0.58
FBgn0035673	1.26	FBgn0027070	0.55

Tabla C.4: Lista de genes subexpresados en ovarios de *D. melanogaster*-wMel (He_down), obtenida desde He *et al.* (2019).

Gene ID	log2FC	Gene ID	log2FC
FBgn0036881	-2.24	FBgn0085474	-1.05
FBgn0265267	-2.18	FBgn0033268	-1.05
FBgn0016920	-1.77	FBgn0010424	-1.02
FBgn0033304	-1.37	FBgn0020908	-1
FBgn0051619	-1.33	FBgn0002773	-0.97
FBgn0265312	-1.31	FBgn0039827	-0.93
FBgn0058298	-1.25	FBgn0037447	-0.91
FBgn0033458	-1.16	FBgn0265203	-0.84
FBgn0039343	-1.07	FBgn0034804	-0.53

Tabla C.5: Lista de genes sobreexpresados en cabezas de hembras *D. paulistorum*-wPau (Bai_F_head_up), obtenida desde Baiao *et al.* (2019).

Gene ID	log2FC	Gene ID	log2FC
FBgn0219132	1.62	FBgn0031414	0.62
FBgn0034407	1.12	FBgn0267435	0.61
FBgn0028331	0.93	FBgn0037896	0.61
FBgn0028866	0.89	FBgn0262738	0.59
FBgn0037162	0.88	FBgn0024943	0.56
FBgn0032495	0.86	FBgn0000206	0.54
FBgn0037164	0.8	FBgn0264087	0.51
FBgn0004919	0.79	FBgn0279901	0.5
FBgn0038530	0.7	FBgn0030555	0.5
FBgn0259210	0.69	FBgn0002938	0.48
FBgn0032116	0.66	FBgn0218178	0.44

Tabla C.6: Lista de genes subexpresados en cabezas de hembras *D. paulistorum*-wPau (Bai_F_head_down), obtenida desde Baiao *et al.* (2019).

Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC
FBgn0030073	-1.54	FBgn0034999	-0.8	FBgn0041194	-0.64	FBgn0002590	-0.53
FBgn0029831	-1.33	FBgn0034479	-0.8	FBgn0038516	-0.63	FBgn0037955	-0.53
FBgn0031974	-1.12	FBgn0029932	-0.8	FBgn0012034	-0.63	FBgn0058191	-0.53
FBgn0014903	-1.09	FBgn0039827	-0.79	FBgn0039359	-0.63	FBgn0027348	-0.53
FBgn0032638	-1.09	FBgn0039154	-0.79	FBgn0002579	-0.62	FBgn0019936	-0.51
FBgn0032213	-1.08	FBgn0026593	-0.77	FBgn0280195	-0.61	FBgn0027580	-0.5
FBgn0032387	-1.08	FBgn0034717	-0.77	FBgn0011693	-0.61	FBgn0016687	-0.5
FBgn0010222	-1.05	FBgn0036750	-0.76	FBgn0086254	-0.61	FBgn0032773	-0.5
FBgn0003067	-1.05	FBgn0035348	-0.75	FBgn0036825	-0.61	FBgn0030151	-0.49
FBgn0033079	-1.03	FBgn0001187	-0.75	FBgn0085453	-0.61	FBgn0032136	-0.49
FBgn0020385	-1.02	FBgn0220668	-0.75	FBgn0023507	-0.6	FBgn0015808	-0.49
FBgn0029823	-1.01	FBgn0050489	-0.74	FBgn0038105	-0.6	FBgn0010408	-0.49
FBgn0011705	-1	FBgn0033246	-0.74	FBgn0014455	-0.6	FBgn0003483	-0.48
FBgn0001089	-0.99	FBgn0027945	-0.74	FBgn0037146	-0.6	FBgn0010516	-0.48
FBgn0052687	-0.99	FBgn0034468	-0.73	FBgn0015663	-0.59	FBgn0037351	-0.48
FBgn0259998	-0.97	FBgn0037684	-0.73	FBgn0039300	-0.59	FBgn0285950	-0.48
FBgn0040732	-0.96	FBgn0035076	-0.73	FBgn0023129	-0.58	FBgn0036816	-0.47
FBgn0041337	-0.96	FBgn0027552	-0.72	FBgn0001114	-0.58	FBgn0029897	-0.47
FBgn0061356	-0.94	FBgn0030425	-0.72	FBgn0040064	-0.58	FBgn0034138	-0.47
FBgn0086691	-0.94	FBgn0033820	-0.72	FBgn0038681	-0.58	FBgn0010412	-0.47
FBgn0033124	-0.92	FBgn0032715	-0.72	FBgn0027579	-0.58	FBgn0034394	-0.47
FBgn0037387	-0.92	FBgn0266369	-0.72	FBgn0035811	-0.58	FBgn0027547	-0.47
FBgn0000140	-0.91	FBgn0040232	-0.72	FBgn0037912	-0.57	FBgn0039682	-0.47
FBgn0040256	-0.89	FBgn0025454	-0.72	FBgn0035484	-0.57	FBgn0015268	-0.46
FBgn0036727	-0.89	FBgn0038897	-0.71	FBgn0285949	-0.57	FBgn0010078	-0.46
FBgn0221155	-0.88	FBgn0003358	-0.7	FBgn0039697	-0.57	FBgn0003942	-0.45
FBgn0284252	-0.88	FBgn0034247	-0.7	FBgn0027930	-0.57	FBgn0013954	-0.45
FBgn0040252	-0.87	FBgn0011695	-0.7	FBgn0284245	-0.56	FBgn0024293	-0.45
FBgn0050280	-0.87	FBgn0000592	-0.7	FBgn0031907	-0.56	FBgn0016122	-0.45
FBgn0033271	-0.87	FBgn0053910	-0.7	FBgn0037686	-0.55	FBgn0014018	-0.45
FBgn0034511	-0.87	FBgn0027657	-0.7	FBgn0037874	-0.55	FBgn0035422	-0.45
FBgn0216594	-0.86	FBgn0000052	-0.69	FBgn0032987	-0.55	FBgn0278830	-0.45
FBgn0259101	-0.86	FBgn0021765	-0.69	FBgn0029176	-0.55	FBgn0261862	-0.44
FBgn0027106	-0.85	FBgn0035734	-0.68	FBgn0003517	-0.55	FBgn0285952	-0.44
FBgn0034276	-0.85	FBgn0035082	-0.68	FBgn0064225	-0.55	FBgn0028697	-0.43
FBgn0014031	-0.84	FBgn0263200	-0.68	FBgn0036837	-0.55	FBgn0033912	-0.43
FBgn0027793	-0.84	FBgn0033879	-0.67	FBgn0283427	-0.55	FBgn0000536	-0.42
FBgn0027611	-0.83	FBgn0039316	-0.67	FBgn0038465	-0.54	FBgn0010225	-0.42
FBgn0031182	-0.83	FBgn0038115	-0.67	FBgn0031320	-0.54	FBgn0020910	-0.41
FBgn0000406	-0.83	FBgn0034364	-0.67	FBgn0027560	-0.54	FBgn0020618	-0.41
FBgn0033397	-0.82	FBgn0035266	-0.66	FBgn0029969	-0.54	FBgn0262782	-0.4
FBgn0012036	-0.81	FBgn0031970	-0.66	FBgn0265178	-0.54	FBgn0264294	-0.39
FBgn0038194	-0.81	FBgn0039241	-0.65	FBgn0013763	-0.54	FBgn0000559	-0.39
FBgn0050281	-0.81	FBgn0225588	-0.65	FBgn0011284	-0.54	FBgn0261592	-0.39
FBgn0020513	-0.81	FBgn0036824	-0.65	FBgn0026721	-0.53	FBgn0017579	-0.37
FBgn0000639	-0.81	FBgn0039609	-0.65	FBgn0267408	-0.53	FBgn0017545	-0.37

Tabla C.7: Lista de genes sobreexpresados en abdómenes de hembras *D. paulistorum*-wPau (Bai_F_abd_up), obtenida desde Baiao *et al.* (2019).

Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC
FBgn0034407	2.65	FBgn0052521	1.11	FBgn0029830	0.93	FBgn0010424	0.76
FBgn0041579	2.08	FBgn0085491	1.11	FBgn0004117	0.93	FBgn0261955	0.76
FBgn0052185	1.98	FBgn0219132	1.11	FBgn0004580	0.93	FBgn0051352	0.75
FBgn0038530	1.83	FBgn0032601	1.11	FBgn0262508	0.92	FBgn0038652	0.75
FBgn0212326	1.65	FBgn0284247	1.11	FBgn0038820	0.92	FBgn0036454	0.75
FBgn0005633	1.53	FBgn0001216	1.1	FBgn0039358	0.92	FBgn0026415	0.73
FBgn0261565	1.5	FBgn0038653	1.1	FBgn0038612	0.91	FBgn0033504	0.73
FBgn0033027	1.45	FBgn0259246	1.09	FBgn0039257	0.91	FBgn0003137	0.73
FBgn0264489	1.44	FBgn0051374	1.09	FBgn0029791	0.91	FBgn0038953	0.72
FBgn0028866	1.44	FBgn0003386	1.09	FBgn0262579	0.91	FBgn0262870	0.72
FBgn0036044	1.42	FBgn0000246	1.07	FBgn0034094	0.9	FBgn0014454	0.72
FBgn0265045	1.41	FBgn0259209	1.07	FBgn0261258	0.89	FBgn0026061	0.72
FBgn0212323	1.41	FBgn0034612	1.07	FBgn0011828	0.88	FBgn0265191	0.71
FBgn0005666	1.41	FBgn0040505	1.07	FBgn0053556	0.87	FBgn0041181	0.71
FBgn0004169	1.41	FBgn0267339	1.07	FBgn0032666	0.87	FBgn0011722	0.71
FBgn0053519	1.4	FBgn0020294	1.06	FBgn0051821	0.87	FBgn0035280	0.71
FBgn0039299	1.39	FBgn0261836	1.06	FBgn0260463	0.87	FBgn0036923	0.71
FBgn0259697	1.37	FBgn0000667	1.06	FBgn0031908	0.87	FBgn0038922	0.7
FBgn0265991	1.37	FBgn0265998	1.05	FBgn0000320	0.87	FBgn0014863	0.7
FBgn0035798	1.36	FBgn0002772	1.05	FBgn0010425	0.87	FBgn0266446	0.69
FBgn0035293	1.35	FBgn0042696	1.04	FBgn0034438	0.87	FBgn0035710	0.69
FBgn0023550	1.32	FBgn0031264	1.04	FBgn0261999	0.86	FBgn0035359	0.69
FBgn0010226	1.31	FBgn0030706	1.03	FBgn0250819	0.85	FBgn0085397	0.68
FBgn0034497	1.31	FBgn0261611	1.02	FBgn0039328	0.85	FBgn0250815	0.67
FBgn0032143	1.31	FBgn0001090	1.01	FBgn0035917	0.85	FBgn0034693	0.67
FBgn0034647	1.28	FBgn0010015	1	FBgn0029762	0.84	FBgn0038881	0.66
FBgn0033679	1.26	FBgn0034512	1	FBgn0264273	0.84	FBgn0031879	0.66
FBgn0031869	1.25	FBgn0033917	1	FBgn0001145	0.83	FBgn0027291	0.65
FBgn0000557	1.23	FBgn0002773	1	FBgn0046874	0.83	FBgn0032494	0.65
FBgn0010482	1.21	FBgn0285958	0.99	FBgn0036116	0.83	FBgn0026059	0.64
FBgn0051140	1.21	FBgn0035410	0.99	FBgn0026077	0.82	FBgn0035985	0.64
FBgn0279040	1.2	FBgn0028371	0.99	FBgn0034420	0.82	FBgn0015766	0.63
FBgn0052333	1.2	FBgn0000527	0.98	FBgn0264907	0.82	FBgn0032422	0.63
FBgn0003961	1.2	FBgn0035452	0.98	FBgn0027585	0.81	FBgn0261260	0.62
FBgn0031857	1.19	FBgn0001250	0.98	FBgn0034470	0.8	FBgn0021906	0.61
FBgn0264695	1.18	FBgn0266801	0.97	FBgn0034723	0.8	FBgn0010100	0.57
FBgn0032536	1.18	FBgn0036891	0.97	FBgn0263258	0.8	FBgn0031037	0.56
FBgn0034151	1.18	FBgn0003149	0.97	FBgn0261015	0.8	FBgn0013733	0.54
FBgn0028704	1.16	FBgn0035868	0.97	FBgn0263236	0.78	FBgn0028509	0.53
FBgn0259680	1.16	FBgn0265356	0.96	FBgn0038294	0.78	FBgn0267348	0.52
FBgn0035490	1.15	FBgn0259736	0.96	FBgn0000064	0.78	FBgn0016075	0.52
FBgn0030562	1.15	FBgn0035358	0.96	FBgn0051973	0.78	FBgn0035600	0.51
FBgn0030929	1.15	FBgn0265180	0.95	FBgn0039048	0.78	FBgn0002526	0.5
FBgn0011286	1.14	FBgn0032685	0.95	FBgn0037796	0.77	FBgn0000299	0.49
FBgn0022160	1.14	FBgn0042201	0.95	FBgn0011695	0.77	FBgn0015221	0.47
FBgn0035929	1.13	FBgn0031327	0.94	FBgn0266758	0.77	FBgn0028325	0.46
FBgn0026255	1.13	FBgn0030102	0.94	FBgn0023540	0.77	FBgn0016081	0.39
FBgn0034860	1.12	FBgn0260942	0.94	FBgn0041180	0.76	FBgn0031734	0.34
FBgn0039897	1.12	FBgn0283471	0.94	FBgn0037447	0.76		

Tabla C.8: Lista de genes subexpresados en abdómenes de hembras *D. paulistorum*-wPau (Bai_F_abd_down), obtenida desde Baiao *et al.* (2019).

Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC
FBgn0035887	-1.98	FBgn0087040	-0.78	FBgn0034299	-0.6
FBgn0036690	-1.72	FBgn0030817	-0.78	FBgn0003041	-0.6
FBgn0261575	-1.69	FBgn0038470	-0.78	FBgn0031535	-0.6
FBgn0038702	-1.56	FBgn0036710	-0.78	FBgn0028375	-0.6
FBgn0032069	-1.41	FBgn0032189	-0.77	FBgn0031381	-0.6
FBgn0033999	-1.36	FBgn0259682	-0.76	FBgn0014859	-0.59
FBgn0032066	-1.34	FBgn0034569	-0.76	FBgn0039773	-0.58
FBgn0279385	-1.21	FBgn0032929	-0.76	FBgn0034282	-0.58
FBgn0005696	-1.16	FBgn0029733	-0.76	FBgn0037376	-0.58
FBgn0003358	-1.16	FBgn0030196	-0.76	FBgn0033871	-0.58
FBgn0034712	-1.15	FBgn0015553	-0.75	FBgn0005683	-0.58
FBgn0011761	-1.14	FBgn0000826	-0.75	FBgn0261854	-0.57
FBgn0032374	-1.09	FBgn0000182	-0.75	FBgn0036893	-0.57
FBgn0010431	-1.03	FBgn0036689	-0.75	FBgn0032726	-0.57
FBgn0001225	-1.02	FBgn0038588	-0.74	FBgn0011704	-0.56
FBgn0030813	-0.98	FBgn0267727	-0.72	FBgn0026084	-0.55
FBgn0034924	-0.98	FBgn0028683	-0.7	FBgn0033752	-0.55
FBgn0038147	-0.97	FBgn0000927	-0.7	FBgn0004698	-0.54
FBgn0002673	-0.96	FBgn0005695	-0.7	FBgn0034433	-0.54
FBgn0003655	-0.95	FBgn0038845	-0.7	FBgn0262743	-0.53
FBgn0001086	-0.95	FBgn0026143	-0.69	FBgn0019686	-0.52
FBgn0040670	-0.93	FBgn0035969	-0.69	FBgn0034009	-0.51
FBgn0031435	-0.92	FBgn0027500	-0.69	FBgn0003124	-0.51
FBgn0000615	-0.91	FBgn0032401	-0.69	FBgn0026430	-0.51
FBgn0032727	-0.89	FBgn0034403	-0.69	FBgn0261108	-0.51
FBgn0038296	-0.88	FBgn0001085	-0.68	FBgn0004649	-0.5
FBgn0033740	-0.88	FBgn0037878	-0.67	FBgn0030206	-0.5
FBgn0002962	-0.88	FBgn0003733	-0.67	FBgn0039637	-0.5
FBgn0032049	-0.86	FBgn0004359	-0.67	FBgn0086908	-0.5
FBgn0279811	-0.86	FBgn0265575	-0.67	FBgn0002719	-0.5
FBgn0030660	-0.85	FBgn0026144	-0.66	FBgn0032042	-0.5
FBgn0022772	-0.84	FBgn0032906	-0.66	FBgn0283500	-0.49
FBgn0266521	-0.84	FBgn0052579	-0.65	FBgn0032197	-0.48
FBgn0031309	-0.83	FBgn0266000	-0.65	FBgn0278978	-0.47
FBgn0222066	-0.83	FBgn0002778	-0.65	FBgn0037548	-0.47
FBgn0015625	-0.82	FBgn0022338	-0.65	FBgn0033635	-0.44
FBgn0043854	-0.82	FBgn0011705	-0.64	FBgn0037756	-0.44
FBgn0030170	-0.82	FBgn0038252	-0.64	FBgn0086384	-0.44
FBgn0022981	-0.82	FBgn0010309	-0.63	FBgn0039348	-0.43
FBgn0000996	-0.81	FBgn0036660	-0.63	FBgn0029819	-0.43
FBgn0038608	-0.8	FBgn0033773	-0.63	FBgn0037468	-0.42
FBgn0000351	-0.79	FBgn0034021	-0.62	FBgn0004864	-0.41
FBgn0040465	-0.79	FBgn0000251	-0.61	FBgn0029157	-0.4
FBgn0040732	-0.79	FBgn0026238	-0.6		
FBgn0046687	-0.78	FBgn0033486	-0.6		

Tabla C.9: Lista de genes sobreexpresados en cabezas de machos *D. paulistorum*-wPau (Bai_M_head_up), obtenida desde Baiao *et al.* (2019).

Gene ID	log2FC
FBgn0032495	0.89
FBgn0012042	0.81
FBgn0020386	0.59
FBgn0050497	0.48

(El siguiente espacio se ha dejado en blanco intencionalmente)

Tabla C.10: Lista de genes subexpresados en cabezas de machos *D. paulistorum*-wPau (Bai_M_head_down), obtenida desde Baiao *et al.* (2019).

Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC
FBgn0040349	-1.5	FBgn0030270	-0.89	FBgn0038463	-0.75	FBgn0033320	-0.62
FBgn0022359	-1.44	FBgn0023477	-0.89	FBgn0086254	-0.75	FBgn0027945	-0.61
FBgn0004045	-1.36	FBgn0027348	-0.89	FBgn0034470	-0.75	FBgn0035298	-0.61
FBgn0043783	-1.33	FBgn0033820	-0.89	FBgn0029932	-0.75	FBgn0031973	-0.61
FBgn0034276	-1.32	FBgn0053493	-0.88	FBgn0004654	-0.75	FBgn0001114	-0.61
FBgn0030737	-1.29	FBgn0032213	-0.88	FBgn0027579	-0.74	FBgn0039040	-0.61
FBgn0038257	-1.26	FBgn0280195	-0.88	FBgn0037493	-0.74	FBgn0035082	-0.6
FBgn0036750	-1.26	FBgn0036857	-0.87	FBgn0031907	-0.74	FBgn0039102	-0.6
FBgn0220668	-1.26	FBgn0037607	-0.87	FBgn0038730	-0.73	FBgn0020513	-0.6
FBgn0041337	-1.23	FBgn0279337	-0.87	FBgn0031538	-0.73	FBgn0004629	-0.6
FBgn0061356	-1.21	FBgn0034885	-0.87	FBgn0058160	-0.73	FBgn0016687	-0.59
FBgn0031974	-1.2	FBgn0013763	-0.86	FBgn0024293	-0.73	FBgn0035978	-0.59
FBgn0034468	-1.17	FBgn0030968	-0.86	FBgn0027657	-0.73	FBgn0034474	-0.58
FBgn0034511	-1.17	FBgn0032638	-0.85	FBgn0034331	-0.73	FBgn0039464	-0.58
FBgn0259715	-1.16	FBgn0038147	-0.85	FBgn0033093	-0.73	FBgn0032494	-0.58
FBgn0002939	-1.15	FBgn0034943	-0.85	FBgn0015663	-0.72	FBgn0013954	-0.57
FBgn0000406	-1.15	FBgn0005391	-0.84	FBgn0026721	-0.72	FBgn0053080	-0.57
FBgn0034247	-1.14	FBgn0032287	-0.84	FBgn0027930	-0.72	FBgn0032029	-0.57
FBgn0031970	-1.13	FBgn0035189	-0.84	FBgn0016122	-0.71	FBgn0010225	-0.57
FBgn0278798	-1.11	FBgn0219402	-0.83	FBgn0283427	-0.71	FBgn0015808	-0.57
FBgn0259713	-1.1	FBgn0033079	-0.83	FBgn0014032	-0.71	FBgn0034126	-0.57
FBgn0004047	-1.07	FBgn0039682	-0.83	FBgn0263199	-0.71	FBgn0027611	-0.57
FBgn0046876	-1.06	FBgn0038878	-0.83	FBgn0052687	-0.71	FBgn0034282	-0.56
FBgn0020385	-1.06	FBgn0259682	-0.83	FBgn0038465	-0.7	FBgn0279105	-0.56
FBgn0029765	-1.05	FBgn0034229	-0.82	FBgn0030362	-0.7	FBgn0027580	-0.56
FBgn0086691	-1.04	FBgn0030615	-0.82	FBgn0037973	-0.7	FBgn0264294	-0.55
FBgn0213515	-1.04	FBgn0011695	-0.82	FBgn0039300	-0.69	FBgn0037063	-0.55
FBgn0216594	-1.03	FBgn0034117	-0.82	FBgn0218353	-0.69	FBgn0035484	-0.54
FBgn0224728	-1.02	FBgn0014031	-0.81	FBgn0031824	-0.69	FBgn0261862	-0.54
FBgn0025454	-1.02	FBgn0040582	-0.81	FBgn0037146	-0.69	FBgn0262782	-0.53
FBgn0063491	-1.01	FBgn0027552	-0.8	FBgn0034724	-0.69	FBgn0011284	-0.53
FBgn0012036	-1	FBgn0030239	-0.8	FBgn0033728	-0.69	FBgn0051676	-0.52
FBgn0278830	-1	FBgn0039801	-0.8	FBgn0001208	-0.68	FBgn0010470	-0.52
FBgn0029823	-0.99	FBgn0013307	-0.79	FBgn0043806	-0.68	FBgn0000055	-0.52
FBgn0039685	-0.99	FBgn0036837	-0.79	FBgn0039611	-0.68	FBgn0051075	-0.52
FBgn0029828	-0.98	FBgn0031694	-0.79	FBgn0053138	-0.68	FBgn0032008	-0.52
FBgn0040256	-0.98	FBgn0038194	-0.79	FBgn0039151	-0.67	FBgn0030521	-0.5
FBgn0259716	-0.97	FBgn0000592	-0.79	FBgn0022355	-0.66	FBgn0284245	-0.5
FBgn0038865	-0.97	FBgn0037635	-0.78	FBgn0043792	-0.66	FBgn0263120	-0.49
FBgn0038105	-0.97	FBgn0031418	-0.78	FBgn0087002	-0.66	FBgn0039697	-0.49
FBgn0030425	-0.94	FBgn0085453	-0.78	FBgn0025595	-0.66	FBgn0027844	-0.49
FBgn0034475	-0.92	FBgn0014455	-0.77	FBgn0030593	-0.65	FBgn0278758	-0.49
FBgn0037975	-0.92	FBgn0012034	-0.77	FBgn0027578	-0.65	FBgn0263773	-0.48
FBgn0033397	-0.92	FBgn0033246	-0.77	FBgn0001187	-0.65	FBgn0032136	-0.48
FBgn0050503	-0.92	FBgn0011705	-0.76	FBgn0035091	-0.64	FBgn0086472	-0.47
FBgn0028526	-0.92	FBgn0032864	-0.76	FBgn0217203	-0.64	FBgn0267408	-0.47
FBgn0040322	-0.92	FBgn0266369	-0.76	FBgn0039349	-0.63	FBgn0285937	-0.42
FBgn0033170	-0.91	FBgn0032773	-0.76	FBgn0034618	-0.63	FBgn0035763	-0.4
FBgn0035076	-0.89	FBgn0011693	-0.75	FBgn0036046	-0.63	FBgn0032949	-0.36
FBgn0034364	-0.89	FBgn0040398	-0.75	FBgn0031327	-0.63		

Tabla C.11: Lista de genes sobreexpresados en abdómenes de machos *D. paulistorum*-wPau (Bai_M_abd_up), obtenida desde Baiao *et al.* (2019).

Gene ID	log2FC	Gene ID	log2FC
FBgn0264489	1.66	FBgn0032129	0.89
FBgn0034151	1.51	FBgn0001216	0.86
FBgn0265045	1.12	FBgn0027291	0.77
FBgn0265991	1.42	FBgn0259209	0.85
FBgn0005633	1.41	FBgn0052333	0.85
FBgn0015035	1.4	FBgn0039358	0.84
FBgn0004427	1.17	FBgn0032253	0.82
FBgn0034497	1.35	FBgn0036597	0.82
FBgn0085487	1.32	FBgn0262508	0.76
FBgn0261565	1.29	FBgn0267486	0.81
FBgn0213602	1.29	FBgn0265998	0.8
FBgn0004169	1.23	FBgn0026059	0.8
FBgn0005666	1.27	FBgn0030745	0.78
FBgn0000557	1.27	FBgn0038820	0.77
FBgn0283471	1.14	FBgn0051973	0.77
FBgn0264695	1.2	FBgn0262599	0.77
FBgn0053519	1.17	FBgn0263006	0.73
FBgn0031632	1.17	FBgn0261836	0.77
FBgn0002772	1.08	FBgn0010482	0.72
FBgn0033027	1.14	FBgn0026077	0.66
FBgn0003149	0.87	FBgn0265356	0.71
FBgn0004242	1.08	FBgn0010425	0.71
FBgn0026255	1.09	FBgn0250819	0.67
FBgn0031264	1.09	FBgn0004117	0.64
FBgn0051410	1.08	FBgn0265191	0.65
FBgn0216177	1.08	FBgn0261053	0.66
FBgn0031302	1.07	FBgn0000064	0.66
FBgn0032505	1.05	FBgn0270926	0.66
FBgn0031869	1.05	FBgn0003415	0.65
FBgn0022160	1.05	FBgn0034723	0.63
FBgn0000667	0.99	FBgn0283499	0.61
FBgn0002773	1.02	FBgn0034391	0.59
FBgn0225785	1.01	FBgn0032779	0.59
FBgn0038840	1.01	FBgn0261955	0.57
FBgn0036783	1.01	FBgn0259736	0.56
FBgn0015010	1	FBgn0003360	0.56
FBgn0037562	0.98	FBgn0261258	0.55
FBgn0023540	0.97	FBgn0027585	0.55
FBgn0039897	0.96	FBgn0053208	0.49
FBgn0034182	0.96	FBgn0013733	0.49
FBgn0031857	0.9	FBgn0261439	0.47
FBgn0036044	0.94	FBgn0035600	0.47
FBgn0039008	0.93	FBgn0010352	0.45
FBgn0261999	0.92	FBgn0033661	0.45
FBgn0085232	0.91	FBgn0261574	0.44
FBgn0025712	0.89		

Tabla C.12: Lista de genes subexpresados en abdómenes de machos *D. paulistorum*-wPau (Bai_M_abd_down), obtenida desde Baiao *et al.* (2019).

Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC
FBgn0038702	-2	FBgn0010052	-0.82	FBgn0039801	-0.64	FBgn0005533	-0.54
FBgn0030097	-1.65	FBgn0278721	-0.81	FBgn0032266	-0.64	FBgn0036795	-0.54
FBgn0029765	-1.62	FBgn0086691	-0.81	FBgn0051021	-0.64	FBgn0280128	-0.54
FBgn0022359	-1.57	FBgn0214058	-0.81	FBgn0226563	-0.64	FBgn0039049	-0.54
FBgn0000640	-1.48	FBgn0033397	-0.8	FBgn0033728	-0.64	FBgn0278645	-0.54
FBgn0279124	-1.47	FBgn0086708	-0.8	FBgn0034291	-0.64	FBgn0050042	-0.53
FBgn0053301	-1.42	FBgn0283437	-0.8	FBgn0040398	-0.64	FBgn0038919	-0.53
FBgn0030029	-1.39	FBgn0030425	-0.8	FBgn0030837	-0.64	FBgn0037664	-0.53
FBgn0085487	-1.33	FBgn0034943	-0.8	FBgn0050151	-0.64	FBgn0001187	-0.53
FBgn0213515	-1.32	FBgn0279337	-0.79	FBgn0283427	-0.64	FBgn0033520	-0.53
FBgn0000639	-1.29	FBgn0038115	-0.78	FBgn0004057	-0.63	FBgn0278840	-0.53
FBgn0224870	-1.24	FBgn0063491	-0.78	FBgn0051323	-0.63	FBgn0035726	-0.52
FBgn0033294	-1.19	FBgn0030880	-0.78	FBgn0040732	-0.63	FBgn0032834	-0.52
FBgn0051091	-1.16	FBgn0028517	-0.77	FBgn0038878	-0.63	FBgn0034296	-0.52
FBgn0015001	-1.14	FBgn0033170	-0.76	FBgn0043043	-0.63	FBgn0031973	-0.52
FBgn0031940	-1.12	FBgn0039774	-0.76	FBgn0035673	-0.63	FBgn0034618	-0.51
FBgn0034712	-1.11	FBgn0038257	-0.76	FBgn0038074	-0.63	FBgn0030306	-0.51
FBgn0002563	-1.07	FBgn0038337	-0.75	FBgn0262571	-0.63	FBgn0031000	-0.51
FBgn0261575	-1.06	FBgn0030737	-0.75	FBgn0034052	-0.63	FBgn0031907	-0.51
FBgn0001225	-1.06	FBgn0053532	-0.74	FBgn0259716	-0.63	FBgn0020236	-0.51
FBgn0035409	-1.05	FBgn0042104	-0.74	FBgn0000055	-0.62	FBgn0214992	-0.51
FBgn0000056	-1.05	FBgn0034331	-0.74	FBgn0033431	-0.62	FBgn0250848	-0.5
FBgn0030367	-1.05	FBgn0030270	-0.74	FBgn0278758	-0.62	FBgn0050488	-0.5
FBgn0031533	-1.04	FBgn0031032	-0.74	FBgn0038105	-0.62	FBgn0024293	-0.5
FBgn0085285	-1.03	FBgn0013953	-0.73	FBgn0038719	-0.61	FBgn0004512	-0.5
FBgn0265137	-1.02	FBgn0015583	-0.73	FBgn0038194	-0.61	FBgn0034478	-0.5
FBgn0039685	-1.02	FBgn0001208	-0.73	FBgn0035734	-0.61	FBgn0030999	-0.49
FBgn0037650	-1.01	FBgn0037386	-0.73	FBgn0039102	-0.61	FBgn0027348	-0.49
FBgn0263106	-1	FBgn0039298	-0.73	FBgn0039111	-0.61	FBgn0259979	-0.49
FBgn0051463	-1	FBgn0051659	-0.73	FBgn0030993	-0.61	FBgn0034068	-0.49
FBgn0053306	-1	FBgn0214205	-0.72	FBgn0035978	-0.61	FBgn0034662	-0.49
FBgn0051636	-0.99	FBgn0038700	-0.71	FBgn0034756	-0.61	FBgn0029093	-0.49
FBgn0020506	-0.99	FBgn0279710	-0.71	FBgn0034885	-0.61	FBgn0034405	-0.49
FBgn0001226	-0.98	FBgn0030968	-0.7	FBgn0033814	-0.61	FBgn0035471	-0.49
FBgn0022700	-0.98	FBgn0035003	-0.7	FBgn0030828	-0.6	FBgn0030521	-0.48
FBgn0034247	-0.98	FBgn0012037	-0.7	FBgn0039755	-0.6	FBgn0038419	-0.48
FBgn0036495	-0.97	FBgn0051419	-0.69	FBgn0030484	-0.6	FBgn0030593	-0.48
FBgn0031930	-0.96	FBgn0030775	-0.69	FBgn0262801	-0.59	FBgn0031260	-0.48
FBgn0032055	-0.96	FBgn0011695	-0.69	FBgn0038829	-0.59	FBgn0030859	-0.48
FBgn0033367	-0.96	FBgn0040308	-0.69	FBgn0061356	-0.59	FBgn0039562	-0.47
FBgn0032149	-0.94	FBgn0038876	-0.69	FBgn0031184	-0.59	FBgn0278830	-0.47
FBgn0033423	-0.94	FBgn0023477	-0.69	FBgn0029823	-0.59	FBgn0279493	-0.47
FBgn0039769	-0.94	FBgn0039598	-0.69	FBgn0037955	-0.58	FBgn0031741	-0.47
FBgn0017448	-0.93	FBgn0038701	-0.68	FBgn0083966	-0.58	FBgn0034383	-0.46
FBgn0030539	-0.93	FBgn0035880	-0.68	FBgn0052203	-0.58	FBgn0039754	-0.46
FBgn0053138	-0.93	FBgn0029838	-0.68	FBgn0223610	-0.58	FBgn0013772	-0.46
FBgn0003295	-0.92	FBgn0036996	-0.67	FBgn0086909	-0.58	FBgn0004066	-0.46
FBgn0037038	-0.91	FBgn0030098	-0.67	FBgn0027579	-0.57	FBgn0010611	-0.45
FBgn0040629	-0.91	FBgn0032615	-0.67	FBgn0004654	-0.57	FBgn0221174	-0.45
FBgn0053530	-0.91	FBgn0014469	-0.67	FBgn0219842	-0.57	FBgn0032923	-0.45
FBgn0037975	-0.91	FBgn0036750	-0.67	FBgn0054002	-0.57	FBgn0038819	-0.45
FBgn0031804	-0.9	FBgn0036942	-0.66	FBgn0037288	-0.56	FBgn0053303	-0.44
FBgn0083965	-0.9	FBgn0037553	-0.66	FBgn0052687	-0.56	FBgn0033820	-0.44
FBgn0038865	-0.88	FBgn0213647	-0.66	FBgn0004432	-0.56	FBgn0050008	-0.44
FBgn0039154	-0.88	FBgn0029709	-0.66	FBgn0037433	-0.56	FBgn0262782	-0.43
FBgn0032213	-0.87	FBgn0039325	-0.66	FBgn0036793	-0.55	FBgn0032271	-0.43
FBgn0259974	-0.86	FBgn0012034	-0.66	FBgn0032853	-0.55	FBgn0013763	-0.43
FBgn0034470	-0.86	FBgn0030615	-0.65	FBgn0031373	-0.55	FBgn0014868	-0.42
FBgn0052573	-0.85	FBgn0039464	-0.65	FBgn0037607	-0.55	FBgn0029821	-0.4
FBgn0034117	-0.85	FBgn0280195	-0.65	FBgn0051266	-0.55	FBgn0013954	-0.4
FBgn0036929	-0.84	FBgn0020765	-0.65	FBgn0036820	-0.54	FBgn0042627	-0.37

Tabla C.13: Lista de genes sobreexpresados en hembras *D. melanogaster*-wMel completas (Lind_up), obtenida desde Lindsey *et al.* (2021).

Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC
FBgn0036670	4.31	FBgn0001612	1.26	FBgn0037186	0.74
FBgn0031277	3.4	FBgn0036620	1.26	FBgn0051262	0.72
FBgn0263093	3.38	FBgn0040350	1.24	FBgn0027513	0.7
FBgn0002563	3.3	FBgn0001087	1.2	FBgn0038704	0.68
FBgn0039881	3.28	FBgn0085224	1.18	FBgn0050382	0.68
FBgn0035482	2.95	FBgn0002878	1.18	FBgn0037960	0.64
FBgn0010549	2.42	FBgn0265726	1.17	FBgn0261984	0.64
FBgn0284435	2.39	FBgn0039938	1.15	FBgn0036332	0.63
FBgn0262108	2.32	FBgn0037521	1.14	FBgn0250785	0.62
FBgn0036362	2.31	FBgn0033257	1.12	FBgn0051075	0.61
FBgn0004428	2.27	FBgn0030514	1.11	FBgn0037807	0.61
FBgn0262509	2.18	FBgn0032382	1.1	FBgn0259247	0.6
FBgn0030051	2.01	FBgn0015336	1.1	FBgn0003082	0.59
FBgn0033395	2.01	FBgn0030653	1.09	FBgn0266084	0.59
FBgn0262123	1.98	FBgn0039932	1.07	FBgn0038593	0.56
FBgn0034716	1.89	FBgn0033302	1.06	FBgn0266918	0.56
FBgn0010040	1.84	FBgn0029766	1.01	FBgn0263593	0.54
FBgn0037742	1.84	FBgn0051999	1.01	FBgn0003227	0.54
FBgn0030073	1.83	FBgn0033047	1	FBgn0031678	0.54
FBgn0030716	1.78	FBgn0086898	1	FBgn0030839	0.53
FBgn0015558	1.75	FBgn0030311	0.99	FBgn0038475	0.53
FBgn0267967	1.74	FBgn0051619	0.98	FBgn0086254	0.52
FBgn0033698	1.71	FBgn0031632	0.95	FBgn0032521	0.51
FBgn0263762	1.7	FBgn0038652	0.95	FBgn0032633	0.49
FBgn0051997	1.69	FBgn0083940	0.95	FBgn0262955	0.49
FBgn0003356	1.65	FBgn0039667	0.95	FBgn0011672	0.45
FBgn0030666	1.64	FBgn0038658	0.94	FBgn0035955	0.45
FBgn0034356	1.61	FBgn0026084	0.94	FBgn0052638	0.43
FBgn0003866	1.59	FBgn0037265	0.94	FBgn0039348	0.42
FBgn0015037	1.58	FBgn0039789	0.93	FBgn0261794	0.41
FBgn0039637	1.56	FBgn0030518	0.92	FBgn0037234	0.41
FBgn0030594	1.55	FBgn0038912	0.92	FBgn0032135	0.4
FBgn0031961	1.36	FBgn0030999	0.89	FBgn0000032	0.38
FBgn0035360	1.35	FBgn0030519	0.89	FBgn0011754	0.36
FBgn0261556	1.33	FBgn0261988	0.86	FBgn0002899	0.35
FBgn0031157	1.33	FBgn0034270	0.85	FBgn0035558	0.34
FBgn0264385	1.3	FBgn0036640	0.83	FBgn0032455	0.34
FBgn0030077	1.3	FBgn0035076	0.82	FBgn0038662	0.33
FBgn0050438	1.27	FBgn0030660	0.77	FBgn0034628	0.32
FBgn0042178	1.27	FBgn0063492	0.77	FBgn0032938	0.3
FBgn0025693	1.26	FBgn0051217	0.75	FBgn0052699	0.3

Tabla C.14: Lista de genes subexpresados en hembras *D. melanogaster*-wMel completas (Lind_down), obtenida desde Lindsey *et al.* (2021).

Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC
FBgn0037844	-8.21	FBgn0001224	-1.4	FBgn0030447	-0.66
FBgn0031018	-5.4	FBgn0002868	-1.38	FBgn0035461	-0.65
FBgn0002645	-5.35	FBgn0046776	-1.29	FBgn0031865	-0.64
FBgn0263001	-4.57	FBgn0033214	-1.28	FBgn0052027	-0.62
FBgn0037548	-4.55	FBgn0038744	-1.27	FBgn0037876	-0.61
FBgn0267160	-4.2	FBgn0035264	-1.27	FBgn0033428	-0.6
FBgn0036738	-4	FBgn0039321	-1.24	FBgn0035890	-0.59
FBgn0032891	-3.99	FBgn0000477	-1.23	FBgn0029858	-0.58
FBgn0264675	-3.73	FBgn0085359	-1.21	FBgn0034750	-0.57
FBgn0032869	-3.46	FBgn0001149	-1.14	FBgn0263232	-0.56
FBgn0037811	-3.29	FBgn0052243	-1.11	FBgn0004838	-0.56
FBgn0053926	-3.23	FBgn0026602	-1.08	FBgn0038869	-0.55
FBgn0034715	-3.11	FBgn0037999	-1.02	FBgn0035964	-0.53
FBgn0052282	-2.71	FBgn0028572	-1.02	FBgn0051038	-0.52
FBgn0050361	-2.67	FBgn0010504	-1	FBgn0035529	-0.51
FBgn0033392	-2.65	FBgn0038741	-0.98	FBgn0038577	-0.5
FBgn0016054	-2.64	FBgn0033356	-0.96	FBgn0030793	-0.49
FBgn0039098	-2.59	FBgn0033431	-0.95	FBgn0000100	-0.48
FBgn0003285	-2.44	FBgn0039297	-0.93	FBgn0035064	-0.48
FBgn0038613	-2.14	FBgn0034564	-0.93	FBgn0025366	-0.47
FBgn0027070	-2.12	FBgn0000246	-0.89	FBgn0033184	-0.47
FBgn0028369	-2.07	FBgn0036906	-0.86	FBgn0017429	-0.47
FBgn0035978	-2.06	FBgn0037696	-0.85	FBgn0015568	-0.47
FBgn0034292	-1.99	FBgn0038277	-0.85	FBgn0034405	-0.46
FBgn0267001	-1.95	FBgn0039459	-0.85	FBgn0052056	-0.44
FBgn0033216	-1.81	FBgn0036294	-0.84	FBgn0015801	-0.42
FBgn0026268	-1.72	FBgn0001225	-0.8	FBgn0029755	-0.41
FBgn0030357	-1.68	FBgn0051044	-0.79	FBgn0028327	-0.4
FBgn0013307	-1.65	FBgn0032727	-0.78	FBgn0021750	-0.39
FBgn0039232	-1.63	FBgn0010408	-0.71	FBgn0034351	-0.37
FBgn0038889	-1.63	FBgn0030065	-0.71	FBgn0035936	-0.37
FBgn0052834	-1.57	FBgn0261380	-0.71	FBgn0035405	-0.35
FBgn0262900	-1.48	FBgn0001296	-0.71	FBgn0086443	-0.31
FBgn0034663	-1.45	FBgn0033543	-0.7	FBgn0038989	-0.3
FBgn0030048	-1.44	FBgn0035833	-0.7	FBgn0038467	-0.3
FBgn0004400	-1.44	FBgn0034729	-0.7	FBgn0003257	-0.3
FBgn0038237	-1.43	FBgn0032329	-0.67	FBgn0030177	-0.3
FBgn0031176	-1.4	FBgn0033195	-0.66	FBgn0023514	-0.29

Tabla C.15: Lista de genes sobreexpresados en hembras *D. melanogaster*-wMel completas (Det_up), obtenida desde Detcharoen *et al.* (2021).

Gene ID	lFC	Gene ID	lFC	Gene ID	lFC	Gene ID	lFC	Gene ID	lFC	Gene ID	lFC
FBgn0011832	1.73	FBgn0038385	0.83	FBgn0037577	0.73	FBgn0022800	0.67	FBgn0051266	0.59	FBgn0036121	0.52
FBgn0051681	1.59	FBgn0035817	0.83	FBgn0036550	0.73	FBgn0028699	0.67	FBgn0037818	0.59	FBgn0037057	0.52
FBgn0262572	1.26	FBgn0031634	0.83	FBgn0261616	0.73	FBgn0031700	0.67	FBgn0011822	0.59	FBgn0085468	0.52
FBgn0085475	1.16	FBgn0259707	0.83	FBgn0034075	0.73	FBgn0039313	0.67	FBgn0022709	0.59	FBgn0036727	0.52
FBgn0264827	1.10	FBgn0051781	0.83	FBgn0020521	0.73	FBgn0243486	0.67	FBgn0029172	0.59	FBgn0037796	0.52
FBgn0040377	1.10	FBgn0035103	0.83	FBgn0031879	0.73	FBgn0260027	0.67	FBgn0039787	0.59	FBgn0052112	0.51
FBgn0261362	1.09	FBgn0036479	0.82	FBgn0027562	0.73	FBgn0033301	0.67	FBgn0029843	0.59	FBgn0001128	0.51
FBgn0262484	1.07	FBgn0036836	0.82	FBgn0031630	0.73	FBgn0037721	0.67	FBgn0029768	0.59	FBgn0010019	0.51
FBgn0033367	1.01	FBgn0030539	0.82	FBgn0016013	0.72	FBgn0038420	0.66	FBgn0038542	0.58	FBgn0046763	0.51
FBgn0261269	1.00	FBgn0033702	0.82	FBgn0037288	0.72	FBgn0037275	0.66	FBgn0038658	0.58	FBgn0010389	0.50
FBgn0260764	0.99	FBgn0046875	0.82	FBgn0053196	0.72	FBgn0031517	0.66	FBgn0051374	0.58	FBgn0034693	0.50
FBgn0020908	0.99	FBgn0264791	0.82	FBgn0030304	0.72	FBgn0053054	0.66	FBgn0039048	0.58	FBgn0259680	0.50
FBgn0035798	0.99	FBgn0261534	0.82	FBgn0031520	0.72	FBgn0262539	0.66	FBgn0035293	0.58	FBgn0031645	0.50
FBgn0050419	0.98	FBgn0039297	0.82	FBgn0026562	0.72	FBgn0265356	0.66	FBgn0037765	0.58	FBgn0263006	0.50
FBgn0033949	0.98	FBgn0265045	0.81	FBgn0010414	0.72	FBgn0051100	0.65	FBgn0033919	0.58	FBgn0062978	0.50
FBgn0036656	0.98	FBgn0033135	0.81	FBgn0027109	0.72	FBgn0266446	0.65	FBgn0035917	0.58	FBgn0031505	0.50
FBgn0260653	0.95	FBgn0001216	0.81	FBgn0050008	0.72	FBgn0063499	0.65	FBgn0085419	0.58	FBgn0050296	0.49
FBgn0032286	0.95	FBgn0051646	0.80	FBgn0052024	0.71	FBgn0052311	0.65	FBgn0250819	0.58	FBgn0038981	0.49
FBgn0039485	0.94	FBgn0030369	0.80	FBgn0031735	0.71	FBgn0265180	0.65	FBgn0039486	0.58	FBgn0026872	0.49
FBgn0040074	0.94	FBgn0261341	0.80	FBgn0030102	0.71	FBgn0052195	0.65	FBgn0039054	0.57	FBgn0265726	0.49
FBgn0031110	0.94	FBgn0051267	0.80	FBgn0033782	0.71	FBgn0259209	0.65	FBgn0039184	0.57	FBgn0032394	0.49
FBgn0052834	0.94	FBgn0037764	0.80	FBgn0043535	0.71	FBgn0026255	0.65	FBgn0259219	0.57	FBgn0032713	0.48
FBgn0010381	0.94	FBgn0031940	0.80	FBgn0034943	0.71	FBgn0035231	0.64	FBgn0039883	0.57	FBgn0031907	0.48
FBgn0035280	0.93	FBgn0051676	0.79	FBgn0034470	0.71	FBgn0260793	0.64	FBgn0028658	0.57	FBgn0034420	0.48
FBgn0039052	0.93	FBgn0035985	0.79	FBgn0267001	0.71	FBgn0039807	0.64	FBgn0036766	0.57	FBgn0030976	0.47
FBgn0033807	0.92	FBgn0031692	0.79	FBgn0002789	0.71	FBgn0051706	0.64	FBgn0032022	0.57	FBgn0037391	0.47
FBgn0011828	0.92	FBgn0036094	0.79	FBgn0022160	0.71	FBgn0034639	0.64	FBgn0039927	0.57	FBgn0034920	0.47
FBgn0032494	0.91	FBgn0038404	0.79	FBgn0039667	0.70	FBgn0014863	0.64	FBgn0011695	0.57	FBgn0030292	0.46
FBgn0034807	0.91	FBgn0037292	0.78	FBgn0032683	0.70	FBgn0030362	0.64	FBgn0033448	0.57	FBgn0033483	0.46
FBgn0265457	0.91	FBgn0037140	0.78	FBgn0033911	0.70	FBgn0000046	0.64	FBgn0027608	0.57	FBgn0039737	0.45
FBgn0031406	0.90	FBgn0267481	0.78	FBgn0051103	0.70	FBgn0036742	0.63	FBgn0035427	0.57	FBgn0037715	0.45
FBgn0040609	0.90	FBgn0041713	0.78	FBgn0039564	0.70	FBgn0259740	0.63	FBgn0051352	0.57	FBgn0020416	0.44
FBgn0032283	0.90	FBgn0261649	0.77	FBgn0052669	0.70	FBgn0038198	0.63	FBgn0033710	0.56	FBgn0039755	0.44
FBgn0031522	0.90	FBgn0029147	0.77	FBgn0038654	0.70	FBgn0039040	0.63	FBgn0051199	0.56	FBgn0031708	0.44
FBgn0039789	0.90	FBgn0005666	0.77	FBgn0029898	0.70	FBgn0040091	0.63	FBgn0053110	0.56	FBgn0052230	0.43
FBgn0034860	0.90	FBgn0037997	0.77	FBgn0039483	0.70	FBgn0031850	0.63	FBgn0264894	0.56	FBgn0034723	0.43
FBgn0031089	0.89	FBgn0262139	0.77	FBgn0040281	0.70	FBgn0261545	0.63	FBgn0031857	0.56	FBgn0034724	0.43
FBgn0039897	0.89	FBgn0085426	0.77	FBgn0005677	0.70	FBgn0034706	0.63	FBgn0037153	0.56	FBgn0004055	0.43
FBgn0052277	0.88	FBgn0050043	0.77	FBgn0261990	0.70	FBgn0038693	0.63	FBgn0031939	0.56	FBgn0025839	0.42
FBgn0037762	0.88	FBgn0033128	0.77	FBgn0010385	0.69	FBgn0030745	0.63	FBgn0265991	0.56	FBgn0014391	0.42
FBgn0051207	0.88	FBgn0029826	0.76	FBgn0263980	0.69	FBgn0037447	0.63	FBgn0035710	0.56	FBgn0262512	0.41
FBgn0040102	0.88	FBgn0051189	0.76	FBgn0037565	0.69	FBgn0265262	0.63	FBgn0029821	0.55	FBgn0023507	0.41
FBgn0032899	0.88	FBgn0038291	0.76	FBgn0036364	0.69	FBgn0040001	0.62	FBgn0037754	0.55	FBgn0000261	0.41
FBgn0038449	0.88	FBgn0053519	0.76	FBgn0031805	0.69	FBgn0035490	0.62	FBgn0025879	0.55	FBgn0032774	0.39
FBgn0031737	0.87	FBgn0041181	0.76	FBgn0033464	0.69	FBgn0011286	0.62	FBgn0050069	0.55	FBgn0015039	0.39
FBgn0053178	0.86	FBgn0038959	0.76	FBgn0038774	0.69	FBgn0041712	0.62	FBgn0032422	0.55	FBgn0026088	0.38
FBgn0051259	0.86	FBgn0035982	0.76	FBgn0035678	0.69	FBgn0038088	0.61	FBgn0032156	0.55	FBgn0027590	0.38
FBgn0036361	0.86	FBgn0039321	0.75	FBgn0039131	0.69	FBgn0001991	0.61	FBgn0024366	0.55	FBgn0023540	0.37
FBgn0013812	0.86	FBgn0265267	0.75	FBgn0032897	0.69	FBgn0045761	0.61	FBgn0019643	0.54	FBgn0038842	0.37
FBgn0025712	0.85	FBgn0262509	0.75	FBgn0031908	0.69	FBgn0038701	0.61	FBgn0037290	0.54	FBgn0035020	0.36
FBgn0041711	0.85	FBgn0033731	0.75	FBgn0037763	0.68	FBgn0265856	0.61	FBgn0026593	0.54	FBgn0051523	0.34
FBgn0036091	0.85	FBgn0264489	0.74	FBgn0051176	0.68	FBgn0034484	0.61	FBgn0004429	0.54	FBgn0039970	0.32
FBgn0054034	0.85	FBgn0032139	0.74	FBgn0027586	0.68	FBgn0034479	0.61	FBgn0030482	0.54	FBgn0010434	0.32
FBgn0261848	0.84	FBgn0031918	0.74	FBgn0085425	0.68	FBgn0016684	0.61	FBgn0085434	0.54	FBgn0031670	0.31
FBgn0029580	0.84	FBgn0000153	0.74	FBgn0033358	0.68	FBgn0039844	0.61	FBgn0039008	0.54	FBgn0260866	0.31
FBgn0030634	0.84	FBgn0033603	0.74	FBgn0034135	0.68	FBgn0264821	0.61	FBgn0028482	0.54	FBgn0025885	0.29
FBgn0034515	0.84	FBgn0053281	0.74	FBgn0015034	0.68	FBgn0086906	0.61	FBgn0037547	0.54	FBgn0030478	0.27
FBgn0033136	0.84	FBgn0085399	0.74	FBgn0086450	0.68	FBgn0001145	0.61	FBgn0015872	0.54	FBgn0030703	0.19
FBgn0016032	0.84	FBgn0003308	0.74	FBgn0262538	0.68	FBgn0034509	0.61	FBgn0027341	0.53		
FBgn0266382	0.84	FBgn0030334	0.73	FBgn0031261	0.68	FBgn0036433	0.60	FBgn0042174	0.53		
FBgn0032669	0.83	FBgn0033602	0.73	FBgn0040837	0.68	FBgn0028622	0.60	FBgn0038294	0.53		
FBgn0267758	0.83	FBgn0034493	0.73	FBgn0013467	0.68	FBgn0038181	0.60	FBgn0051997	0.52		

Tabla C.16: Lista de genes subexpresados en hembras *D. melanogaster*-wMel completas (Det_down), obtenida desde Detcharoen *et al.* (2021).

Gene ID	log2FC	Gene ID	log2FC	Gene ID	log2FC
FBgn0036382	-0.87	FBgn0001085	-0.36	FBgn0010280	-0.30
FBgn0029114	-0.76	FBgn0263490	-0.36	FBgn0039212	-0.30
FBgn0050438	-0.73	FBgn0264493	-0.36	FBgn0024833	-0.30
FBgn0000360	-0.72	FBgn0263144	-0.36	FBgn0264785	-0.30
FBgn0036715	-0.68	FBgn0027329	-0.35	FBgn0261016	-0.30
FBgn0000359	-0.66	FBgn0085430	-0.35	FBgn0032886	-0.30
FBgn0053129	-0.62	FBgn0010877	-0.35	FBgn0035425	-0.29
FBgn0038312	-0.58	FBgn0002781	-0.35	FBgn0036237	-0.29
FBgn0036646	-0.56	FBgn0003449	-0.35	FBgn0035388	-0.29
FBgn0022702	-0.55	FBgn0035987	-0.35	FBgn0021818	-0.29
FBgn0003507	-0.54	FBgn0262169	-0.34	FBgn0036661	-0.29
FBgn0036330	-0.52	FBgn0259984	-0.34	FBgn0003415	-0.29
FBgn0035760	-0.51	FBgn0027499	-0.34	FBgn0082831	-0.29
FBgn0030174	-0.46	FBgn0029905	-0.34	FBgn0000721	-0.28
FBgn0266101	-0.45	FBgn0014133	-0.34	FBgn0015019	-0.28
FBgn0013732	-0.45	FBgn0026317	-0.34	FBgn0029685	-0.28
FBgn0031739	-0.43	FBgn0016984	-0.34	FBgn0030753	-0.28
FBgn0265523	-0.42	FBgn0038649	-0.34	FBgn0002872	-0.28
FBgn0035162	-0.42	FBgn0030869	-0.34	FBgn0004876	-0.28
FBgn0003416	-0.42	FBgn0262582	-0.34	FBgn0038532	-0.28
FBgn0033349	-0.42	FBgn0037248	-0.34	FBgn0043010	-0.27
FBgn0031381	-0.41	FBgn0010825	-0.34	FBgn0030838	-0.27
FBgn0038989	-0.40	FBgn0041775	-0.33	FBgn0010278	-0.27
FBgn0004889	-0.40	FBgn0086129	-0.33	FBgn0011224	-0.27
FBgn0003044	-0.40	FBgn0030018	-0.33	FBgn0262114	-0.26
FBgn0259734	-0.40	FBgn0034118	-0.33	FBgn0039508	-0.26
FBgn0028862	-0.40	FBgn0031799	-0.33	FBgn0013759	-0.26
FBgn0028408	-0.40	FBgn0250837	-0.33	FBgn0263929	-0.26
FBgn0042134	-0.39	FBgn0030034	-0.33	FBgn0039215	-0.25
FBgn0030054	-0.39	FBgn0020445	-0.33	FBgn0035237	-0.25
FBgn0001120	-0.39	FBgn0030228	-0.33	FBgn0020443	-0.25
FBgn0014037	-0.39	FBgn0030973	-0.32	FBgn0036621	-0.25
FBgn0033988	-0.39	FBgn0034962	-0.32	FBgn0026147	-0.25
FBgn0052484	-0.38	FBgn0039633	-0.32	FBgn0031048	-0.25
FBgn0037555	-0.38	FBgn0031988	-0.32	FBgn0011207	-0.25
FBgn0037614	-0.38	FBgn0004649	-0.32	FBgn0003165	-0.24
FBgn0000251	-0.38	FBgn0052676	-0.32	FBgn0010380	-0.24
FBgn0039488	-0.38	FBgn0003015	-0.31	FBgn0029092	-0.24
FBgn0051151	-0.37	FBgn0003475	-0.31	FBgn0263102	-0.24
FBgn0035059	-0.37	FBgn0031126	-0.31	FBgn0003732	-0.23
FBgn0020622	-0.37	FBgn0011481	-0.31	FBgn0027951	-0.23
FBgn0039972	-0.37	FBgn0035983	-0.31	FBgn0032444	-0.23
FBgn0052066	-0.37	FBgn0261811	-0.31	FBgn0011763	-0.23
FBgn0041604	-0.37	FBgn0264495	-0.31	FBgn0002121	-0.23
FBgn0005640	-0.37	FBgn0261885	-0.31	FBgn0026370	-0.22
FBgn0031267	-0.36	FBgn0016641	-0.31	FBgn0014075	-0.22
FBgn0036518	-0.36	FBgn0038167	-0.31	FBgn0261458	-0.22
FBgn0004650	-0.36	FBgn0015240	-0.30	FBgn0260936	-0.20
FBgn0051163	-0.36	FBgn0011260	-0.30	FBgn0037249	-0.19
FBgn0003520	-0.36	FBgn0026375	-0.30	FBgn0032643	-0.18
FBgn0003210	-0.36	FBgn0032752	-0.30	FBgn0053303	-0.17
FBgn0031896	-0.36	FBgn0053558	-0.30		

Anexo D

Resultados adicionales de la obtención de listas D.E.

En esta sección se presentan y discuten resultados adicionales de los procesos de obtención de listas D.E., incluyendo evaluaciones FastQC, estadísticas de alineamiento y cantidad de identificadores perdidos en búsquedas ortólogas. También, las listas obtenidas se comparan en términos generales con las listas publicadas (en caso de que estas sean distintas), proponiéndose explicaciones para las discordancias y similitudes. La obtención de las listas Lind_up y Lind_down se excluye, pues consistió simplemente en separar el conjunto de genes de *D. melanogaster* diferencialmente expresados por efecto de *Wolbachia* publicada por Lindsey *et al.* (2021).

D.1. Obtención de listas desde He *et al.* (2019)

Para interpretar los siguientes resultados es preciso conocer los nombres de archivos con lecturas correspondientes a las muestras experimentales. Las tres réplicas de la condición WT se denominaron DW1, DW2 y DW3, y sus respectivos pares de archivos con lecturas se almacenaron bajo los códigos SRR6885180, SRR6885179 y SRR6885178. Por su parte, las tres réplicas de la condición GFR se denominaron DT1, DT2 y DT3, y sus respectivos pares de archivos con lecturas se almacenaron bajo los códigos SRR6885183, SRR6885182 y SRR6885181.

En la Figura D.1 se resumen las evaluaciones de todos los módulos FastQC sobre los 6 pares de archivos con lecturas *paired-end* en bruto obtenidos desde He *et al.* (2019).

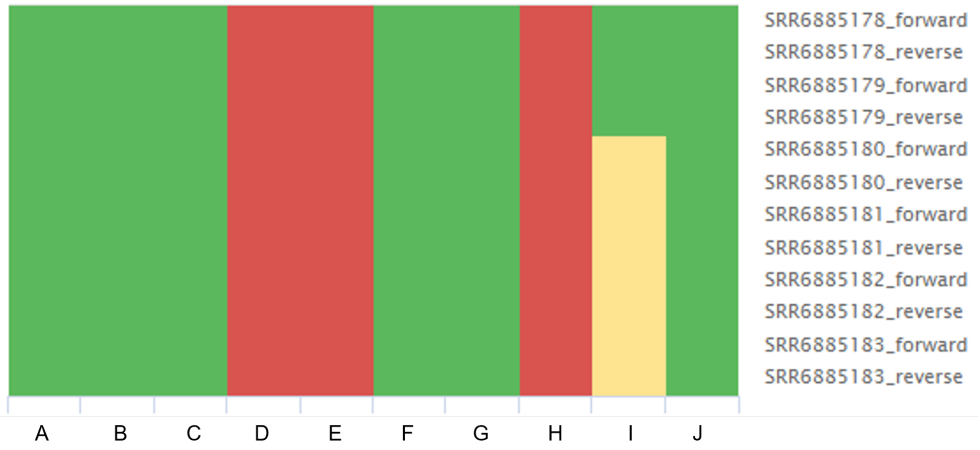

Figura D.1: Resumen FastQC de lecturas en bruto publicadas por He *et al.* (2019). A: *Basic Statistics*, B: *Per Base Sequence Quality*, C: *Per Sequence Quality Scores*, D: *Per Base Sequence Content*, E: *Per Sequence GC Content*, F: *Per Base N Content*, G: *Sequence Length Distribution*, H: *Sequence Duplication Levels*, I: *Overrepresented Sequences*, J: *Adapter Content*

En la Figura D.2 se resume la evaluación FastQC según el módulo *Adapter Content* sobre los 6 pares de archivos con lecturas *paired-end* en bruto obtenidos desde He *et al.* (2019).

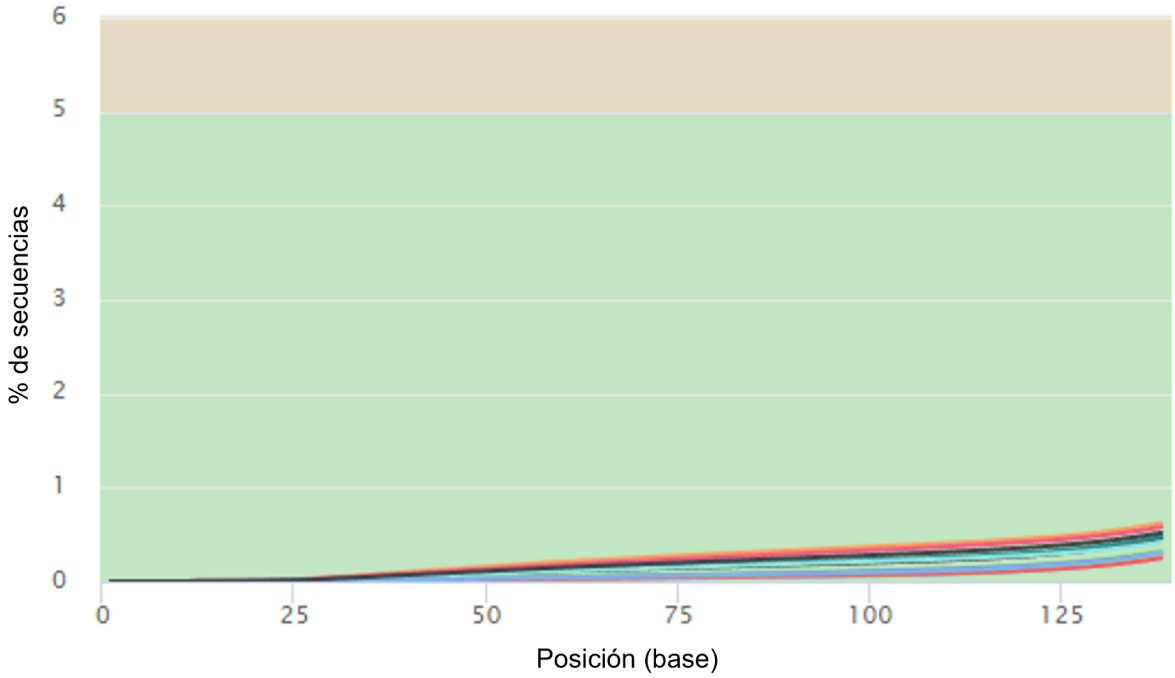

Figura D.2: Resumen de la evaluación de contenido de adaptadores en los archivos con lecturas en bruto publicados por He *et al.* (2019). Cada color se corresponde con un par de archivos (leyenda de colores omitida).

En la Figura D.3 se presenta el resumen de los informes STAR sobre la distribución de los alineamientos de las lecturas de He *et al.* (2019), en términos de su multiplicidad.

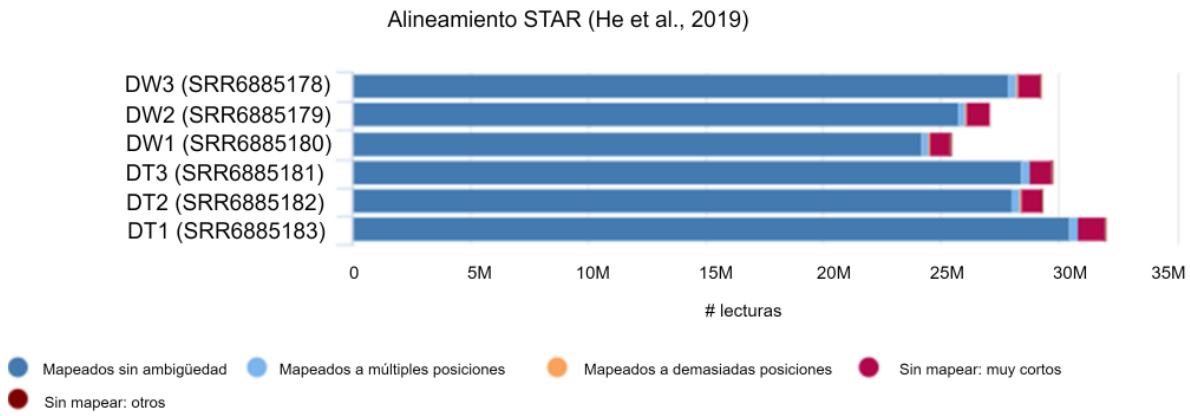

Figura D.3: Distribución de alineamientos STAR de lecturas de He *et al.* (2019).

En la Figura D.4 se muestra la distribución de los promedios de cuentas de lecturas alineadas a genes en el análisis propio. Cabe señalar que sólo se incluyeron los promedios menores a 10, que fueron considerados suficientes para mostrar la tendencia decreciente.

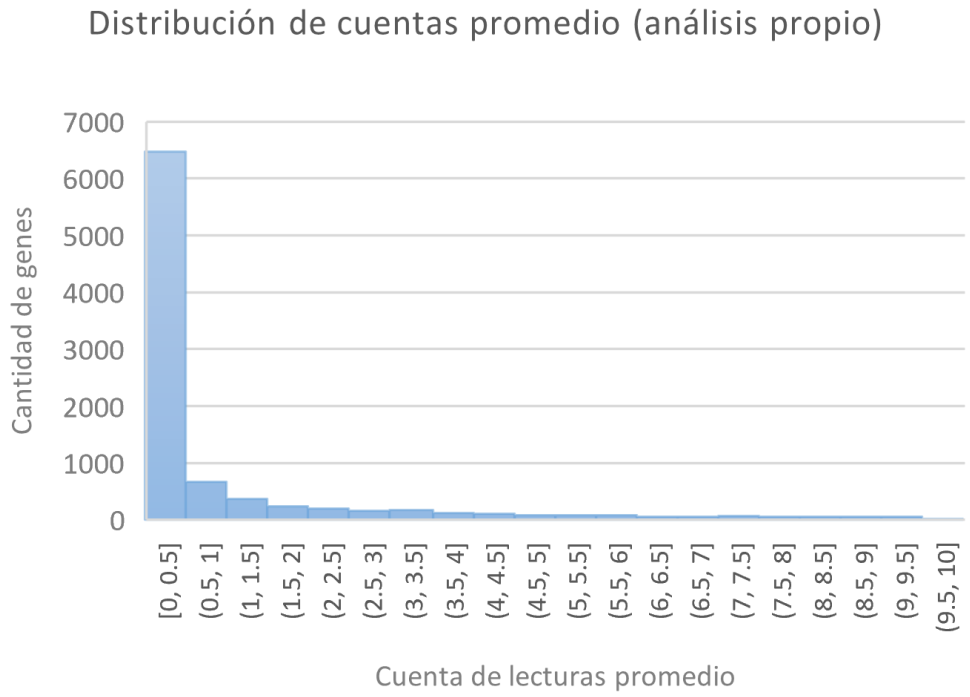

Figura D.4: Distribución de los promedios de cuentas de lecturas alineadas a genes en el análisis propio. En el eje horizontal sólo se representan los valores menores a 10.

Como se muestra en la Figura D.1, la evaluación FastQC catalogó a los archivos con lecturas en bruto como anormales sólo con respecto a los módulos *Per Base Sequence Content*, *Per Sequence GC Content* y *Sequence Duplication Levels*, lo cual se consideró tolerable según lo indicado en la Metodología. Por otra parte, aunque no fue suficiente para levantar advertencias, el resumen de las evaluaciones según el módulo *Adapter Content* señaló la presencia de adaptadores en un pequeño porcentaje de las lecturas (Figura D.2), lo cual fue corregido mediante la ejecución de Trimmomatic en los términos descritos en

la Metodología. Las evaluaciones de los alineamientos de lecturas realizados por STAR, resumidas en la Figura D.3, muestran un muy bajo porcentaje de *multireads*, lo cual justificó el uso de featureCounts para un conteo de lecturas a nivel de genes descartando los mapeos múltiples.

Como fue indicado anteriormente, una de las motivaciones para obtener las listas He_up y He_down mediante análisis propio fue que, en la publicación original, los elementos D.E. fueron filtrados según la magnitud del cambio de expresión ($|log_2FC| > 1$), descartando datos útiles según la metodología aquí adoptada. Sin realizar dicho filtro se esperaba un aumento de tamaño para las listas D.E., sin embargo, el resultado fue el opuesto: las listas de genes sobre- y subexpresados fueron reducidas en el análisis propio, en las magnitudes indicadas por la Figura D.5.

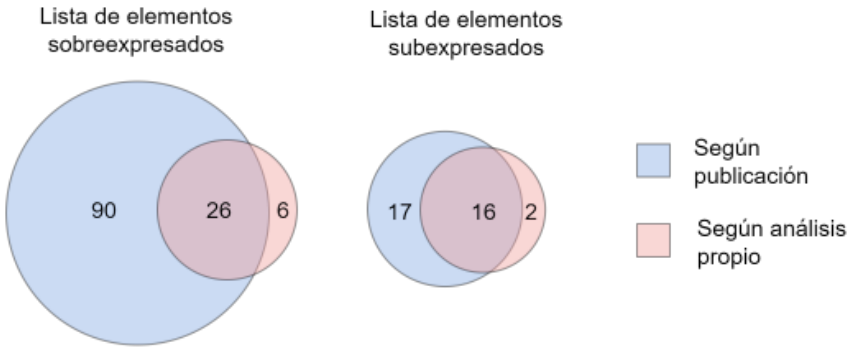

Figura D.5: Diagrama de Venn representando las intersecciones entre las listas de genes sobre- y subexpresados publicadas en He *et al.* (2019) y las obtenidas según análisis RNA-Seq propio.

La diferencia de tamaños es sustantiva, especialmente entre las listas sobreexpresadas, donde 120 genes fueron publicados y sólo 32 genes se hallaron según el análisis propio. En principio, esta discordancia podría deberse a cualquiera de las posibles diferencias metodológicas entre ambos análisis, cuya identificación se dificulta por la omisión de la descripción de una parte importante de la metodología en He *et al.* (2019) (relativa al alineamiento de lecturas, conteo y análisis de expresión diferencial) [30]. Afortunadamente, la cuantificación vía qPCR del cambio de expresión de un grupo de genes, realizado en la misma publicación, permite seguir indagando al respecto. En la Tabla D.1 se comparan los resultados del análisis RNA-Seq propio, con los de los análisis qPCR y RNA-Seq realizados por He *et al.* (2019) para un grupo de genes.

Tabla D.1: Comparación entre genes D.E. según análisis RNA-Seq propio y original de He *et al.* (2019), versus cuantificación mediante qPCR realizado por los mismos autores [30]. N.S. indica que la diferencia de expresión fue considerada no significativa. Los valores Inf o -Inf reflejan que la expresión de un gen se consideró nula exclusivamente en la condición GFR o WT, respectivamente.

Gen	log_2FC (qPCR)	log_2FC (RNA-Seq propio)	$log2FC$ (RNA-Seq original)
rpl22-like	3.3	3.2	5.3
uif	2.6	2.4	6.0
cg32054	2.1	N.S.	Inf
dany	1.9	1.7	2.2
otk2	1.4	0.9	1.1
cg10659	0.9	N.S.	Inf
nompc	N.S.	-1.8	-5.7
cpr76bd	N.S.	-2.2	-6.4
attc	-0.7	N.S.	-2.6
cg6435	-1.1	N.S.	-Inf
twdlb	-1.2	N.S.	-1.9
pgant8	-1.4	N.S.	-Inf
cg5111	-1.5	-1.1	-1.4
def	-1.9	N.S.	-2.4
cg18258	-1.9	-2.2	-3.3

Tomando los resultados del qPCR como el estándar, se observa que tanto el análisis RNA-Seq propio como el original incurren en los mismos dos falsos positivos (para *nompc* y *cpr76bd*). Por otra parte, el análisis propio deriva en siete falsos negativos, para genes cuyo cambio de expresión es declarado significativo por el análisis RNA-Seq original (*cg32054*, *cg10659*, *attc*, *cg6435*, *twdlβ*, *pgant8* y *def*). Una revisión de las cuentas brutas de lecturas alineadas a dichos genes en el análisis propio, revela que su nivel de expresión medio fue muy bajo (datos no mostrados). En particular, las cuentas brutas concuerdan con los resultados RNA-Seq originales en cuanto a la expresión nula de los genes *cg32054* y *cg10659* en la condición GFR, y de los genes *cg6435* y *pgant8* en la condición WT (que en la Tabla D.1 se expresan con valores log_2FC de magnitud infinita).

Un problema común en el análisis de expresión diferencial, es que los genes con baja expresión tienden a mostrar una dispersión muy alta, pudiendo conducir a estimaciones exageradas del tamaño de efecto ($|log_2FC|$) entre las condiciones [64]. La manera de tratar este problema puede impactar considerablemente en la cantidad de resultados significativos, pues los genes con cuentas bajas suelen ser muy abundantes en los experimentos RNA-Seq [64]. El experimento aquí estudiado no sería la excepción, tal como lo sugiere la Figura D.4, que muestra la distribución de los genes según su promedio de cuentas de lecturas entre las muestras. Es posible que el análisis de He *et al.* (2019) haya sido realizado con un algoritmo que procesara de manera distinta los genes con cuentas promedio bajas. En el algoritmo utilizado en el análisis propio, DESeq2, los tamaños de efecto para un gen g se encogen hacia cero de una manera que depende de la cantidad de muestras, y del promedio y dispersión de las cuentas de g entre las muestras [64]. La baja cantidad de muestras (3 por condición, el mínimo recomendado para análisis de expresión diferencial [36]) puede haber contribuido a que DESeq2 adoptara un enfoque

relativamente conservador hacia los genes con niveles de expresión bajos. El hecho de que, según la Tabla D.1, las estimaciones de log_2FC sean siempre de menor magnitud en el análisis RNA-Seq propio, es consistente con la influencia del procedimiento de moderación llevado a cabo por DESeq2.

La Tabla D.1 parece indicar que la metodología original aumenta el poder estadístico sin pagar con un aumento de falsos positivos. Sin embargo, es importante notar que para la validación mediante qPCR, los autores seleccionaron un conjunto no aleatorio de genes. Específicamente, varios de los genes fueron seleccionados en virtud de su asociación con fenómenos reproductivos, para los que cabía prever algún efecto en ovarios [17]. En cualquier caso, es rescatable que los genes D.E. identificados por el método propio hayan sido, en su mayor parte, identificados también por el método de los autores (Figura D.5), sugiriendo que ambas metodologías difieren principalmente en un sentido cuantitativo (en cuanto a distintos balances entre poder estadístico y tolerancias a falsos positivos) y no cualitativo.

Queda por discutir el hecho de que la cantidad de genes D.E. halladas según la metodología propia no sólo es pequeña con respecto a lo publicado por He *et al.* (2019), sino también con respecto al resto de los análisis transcriptómicos, especialmente considerando que se realizó con especificidad de tejido [30]. Además de la influencia de la etapa de moderación realizada por DESeq2 (que opera en desconocimiento de la partición de muestras según condición [64]), se pudo sufrir de un bajo poder estadístico en las pruebas estadísticas posteriores, en virtud de una alta variabilidad intra-condición. En la Figura D.6 se presenta el PCA sobre las cuentas normalizadas por DESeq2 en el análisis propio.

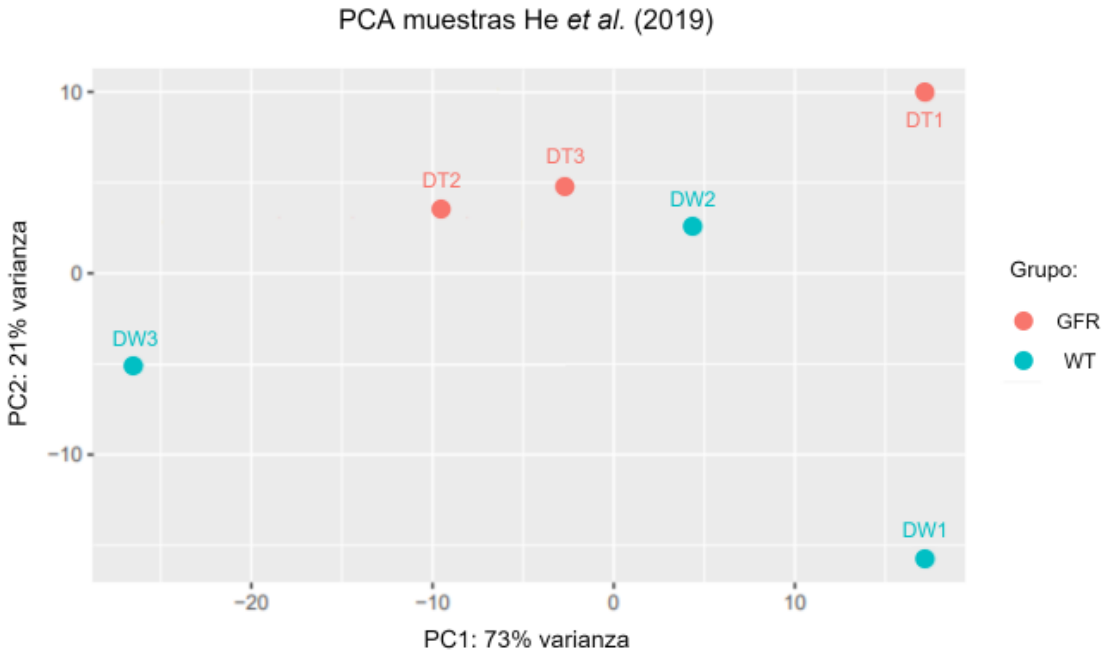

Figura D.6: PCA sobre las cuentas normalizadas de lecturas de He *et al.* (2019) mediante DESeq2, según análisis propio.

Como se observa en el gráfico PCA, si bien existe una agrupación de réplicas según la segunda componente principal (eje vertical), también existe una gran dispersión de las muestras WT con respecto a la primera componente, que es la que explica la mayor parte de la variabilidad entre las muestras, y con respecto a este eje no hay agrupación de las réplicas. Esta importante variabilidad intra-condición más un bajo nivel de replicación, podrían haber significado una pérdida de poder estadístico para detectar genes diferencialmente expresados. Las causas de dicha variabilidad son un tema de especulación muy amplio y no será abordado.

D.2. Obtención de listas desde Caragata *et al.* (2017)

Las listas Car_up y Car_down fueron obtenidas desde tablas cuya estructura se expone en la Tabla B.1, mediante el proceso de decisión detallado en la Figura B.1, resultando en 68 y 47 identificadores de genes de *A. aegypti*, respectivamente. Lo anterior implica la pérdida de 38 % y 29 % de los identificadores de productos génicos asociados a *contigs* sobre- y subexpresados por Caragata *et al.* (2017); una pérdida de datos que ocurrió durante la búsqueda de ortólogos de *A. aegypti*, a pesar de que el proceso de decisión asociado se definió a conciencia para evitarla.

Se estima que un análisis RNA-Seq propio de las lecturas podría haber beneficiado la recuperación de datos de diversas formas. Por ejemplo, el análisis propio permitiría disponer de una lista completa de identificadores asociados a *contigs* (no sólo los D.E.), la cual podría usarse para personalizar el universo del análisis de enriquecimiento sin depender de la elección de un organismo de referencia, haciendo innecesario el descarte de elementos sin ortólogos en tal organismo. También, se especula que un nuevo análisis RNA-Seq podría haber ayudado a recuperar *contigs* que fueron perdidos originalmente por Caragata *et al.* (2017) por no tener resultados BLAST, en cuanto nuevas anotaciones van surgiendo constantemente en la literatura.

A pesar de las motivaciones anteriores, el análisis RNA-Seq propio de los datos generados en Caragata *et al.* (2017), así como el de todos los datos que requirieran la reconstrucción de transcriptomas *de novo*, fue descartado. La razón es que, luego de investigar sobre la teoría, implementación y validación de las reconstrucciones de transcriptomas, se consideró un ejercicio cuya correcta ejecución podría ser difícil de abordar en el tiempo disponible para el desarrollo de este trabajo.

Dado que los identificadores en las listas Car_up y Car_down son de *A. aegypti*, estos fueron mapeados a ortólogos de *D. melanogaster* para poder comparar directamente sus valores log_2FC con los del resto de las listas, mediante los *heatmaps* asociados a módulos funcionales. Dicho mapeo entregó 47 ortólogos a genes contenidos en Car_up y 33 ortólogos a genes contenidos en Car_down, respectivamente.

D.3. Obtención de listas desde Baiao *et al.* (2019)

Las listas de Baiao *et al.* (2019) fueron obtenidas desde tablas cuya estructura se expone en la Tabla B.2, mediante el proceso de decisión detallado en la Figura B.2, siendo conformadas por identificadores de genes de *D. melanogaster*. Las listas Bai_F_abd_up y Bai_F_abd_down obtenidas fueron de 195 y 133 identificadores, respectivamente. Las listas Bai_F_head_up y Bai_F_head_down obtenidas fueron de 22 y 184 identificadores, respectivamente. Las listas Bai_M_abd_up y Bai_M_abd_down obtenidas fueron de 91 y 244 identificadores, respectivamente. Las listas Bai_M_head_up y Bai_M_head_down obtenidas fueron de 4 y 199 identificadores, respectivamente. Para comparar los tamaños de las listas obtenidas entre sí y obtener una perspectiva del nivel de pérdida de datos en su obtención, en la Figura D.7 se integra la información anterior con la presentada en la sección 3.1.3, sobre las cantidades de *contigs* D.E. e identificadores *Drosophila* originalmente asociados a estos.

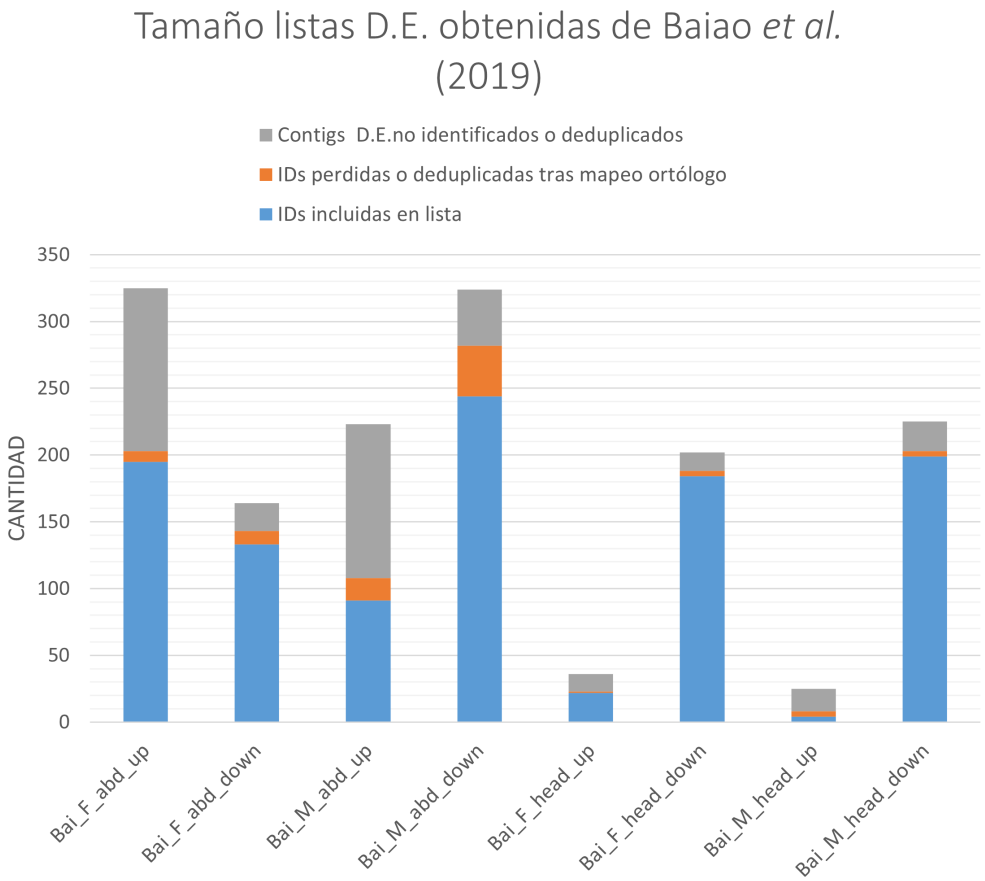

Figura D.7: Cantidad de genes en cada lista D.E. obtenida desde Baiao *et al.* (2019) y número de elementos perdidos por los autores (en la asociación de *contigs* D.E. a identificadores *Drosophila*) o en el presente trabajo (al buscar ortología con genes de *D. melanogaster*). En ambos casos la pérdida pudo darse al no encontrar asociaciones o por deduplicación, esto es, el hecho de que un identificador se cuenta una sola vez aunque múltiples elementos se asocien con el.

En la Figura D.7 se observa que las cantidades de identificadores perdidos mediante la metodología propia (representadas en rojo) es relativamente baja, tanto en comparación con los tamaños de las listas de genes D.E. obtenidas (representados en azul), como en comparación con las cantidades de *contigs* D.E. que ya se habían descartado por parte de los autores (representadas en gris). La razón es que, entre los *contigs* que pudieron ser asociados a genes *Drosophila* en la publicación de Baiao *et al.* (2019), la mayoría lo fueron en particular a *D. melanogaster* [31], sin sufrir el riesgo de ser descartados en la búsqueda de ortólogos incluida en la presente metodología. Los pocos identificadores descartados corresponden a genes *D. willistoni* sin ortólogos hallados en *D. melanogaster* o, en menor medida, deduplicados tras el mapeo.

En términos porcentuales, la única lista afectada considerablemente por la pérdida de identificadores debida al procedimiento propio fue Bai_M_head_up, correspondiente a elementos sobreexpresados en cabezas de machos *D. paulistorum* con infección nativa intacta. La razón es que la cantidad de *contigs* sobreexpresados derivados de dicha comparación fue baja, y más aún lo fue la cantidad de genes Drosophila que pudieron ser originalmente asociados a estos (8), tornando sensible la pérdida de 4 identificadores de *D. willistoni* en que se incurrió en el presente trabajo. Aunque más notorio en el caso anterior, la Figura D.7 muestra que la relativa escasez de genes sobreexpresados en comparaciones de cabezas también se constató en hembras (Bai_F_head_up). Esta característica no fue discutida en la publicación de Baiao *et al.* (2019) [31] y, dado que no se hallaron otros estudios ómicos que hubieran comparado cabezas de individuos Diptera con y sin *Wolbachia*, no fue posible determinar si este constituye un fenómeno común ni hipotetizar sobre sus causas.

D.4. Obtención de listas desde Detcharoen *et al.* (2021)

Para interpretar los siguientes resultados es preciso conocer los nombres de archivos con lecturas correspondientes a las distintas muestras experimentales, resumidas en la Tabla D.2.

Tabla D.2: Asociación entre nombres de muestras estudiadas en Detcharoen *et al.* (2021) y los códigos de sus respectivos archivos con lecturas en el repositorio SRA.

Condición	Línea	Nombre de muestra	Código de archivo de lecturas
GFR	0	mu_0.1	SRR10915899
		mu_0.2	SRR10915898
		mu_0.3	SRR10915897
		mu_0.4	SRR10915896
		mu_0.5	SRR10915895
WT	1	mi_1.1	SRR10915906
		mi_1.2	SRR10915905
		mi_1.3	SRR10915894
		mi_1.4	SRR10915883
		mi_1.5	SRR10915872
	2	mi_2.1	SRR10915871
		mi_2.2	SRR10915870
		mi_2.3	SRR10915869
		mi_2.4	SRR10915868
		mi_2.5	SRR10915867
	3	mi_3.1	SRR10915904
		mi_3.2	SRR10915903
		mi_3.3	SRR10915902
		mi_3.4	SRR10915901
		mi_3.5	SRR10915900

En la Figura D.8 se muestra el resumen de las evaluaciones de todos los módulos FastQC sobre los 19 archivos con lecturas *single-end* en bruto obtenidos desde Detcharoen *et al.* (2021).

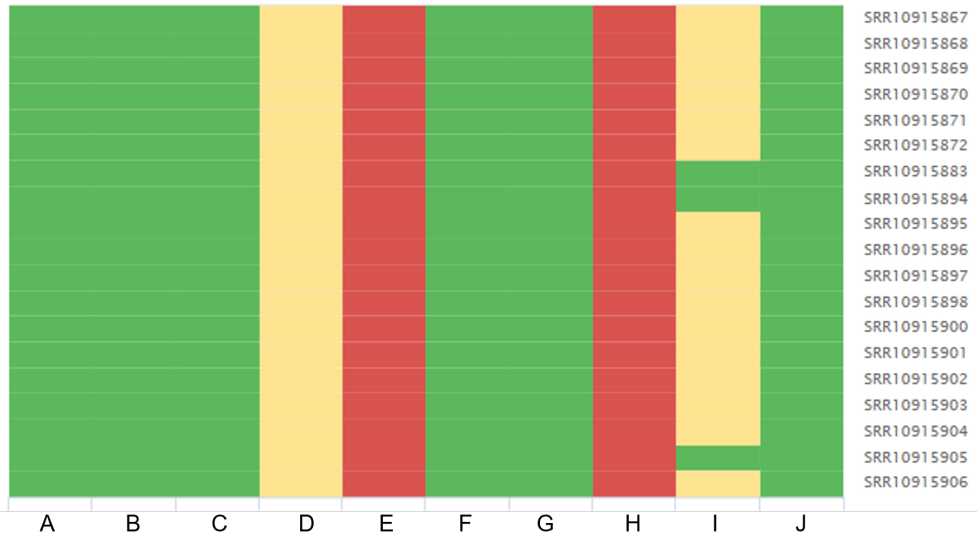

Figura D.8: Resumen de la evaluación FastQC de los archivos con lecturas en bruto publicados por Detcharoen *et al.* (2021). La asociación entre letras y módulos FastQC es equivalente a la expuesta en la Figura D.1.

En la Figura D.9 se muestra el resumen de las evaluaciones FastQC según el módulo *Adapter Content* sobre los archivos con lecturas en bruto obtenidos desde Detcharoen *et al.* (2021).

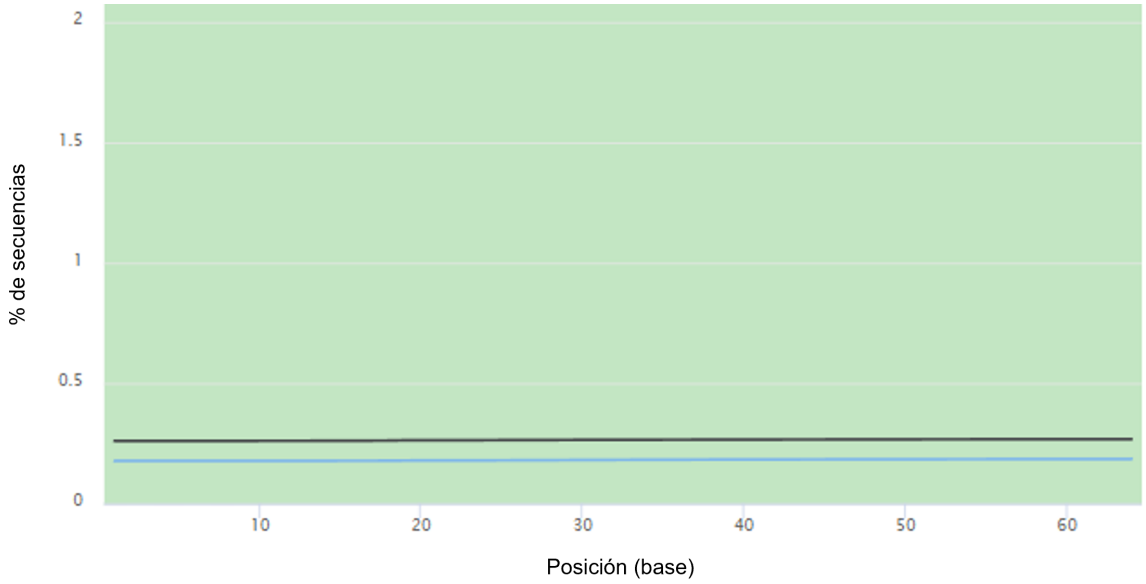

Figura D.9: Resumen de evaluación FastQC según módulo *Adapter Content* sobre archivos con lecturas en bruto publicadas por Detcharoen *et al.* (2021). Sólo se muestran las curvas correspondientes a los archivos SRR10915901 (negro) y SRR10915867 (celeste), omitiéndose todas las demás curvas, en cuanto no sobrepasan el 0.1 %

En la Figura D.10 se muestra el resumen de las evaluaciones de todos los módulos FastQC sobre los 19 archivos con lecturas *single-end* obtenidos desde Detcharoen *et al.* (2021), luego de su corrección mediante Trimmomatic.

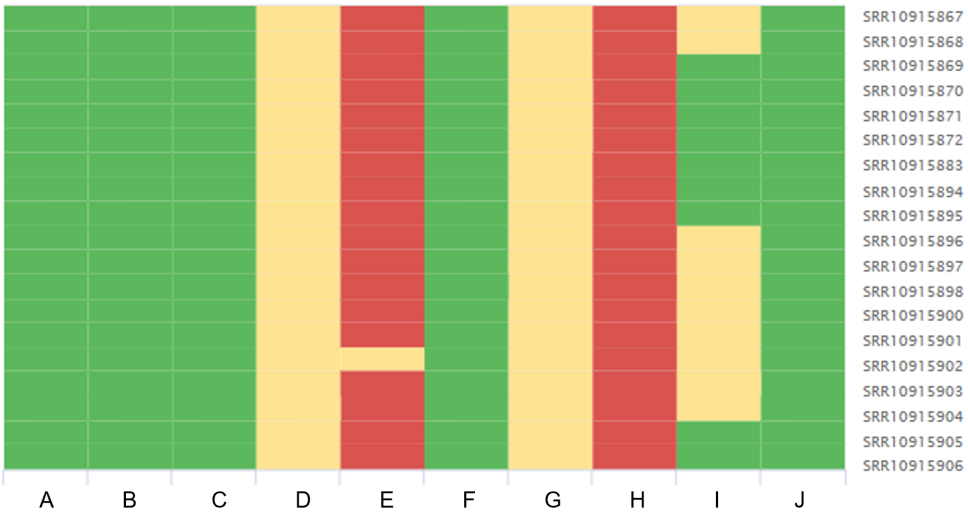

Figura D.10: Resumen de la evaluación FastQC de los archivos con lecturas luego del control de calidad mediante Trimmomatic. La asociación entre letras y módulos FastQC es equivalente a la expuesta en la Figura D.1.

En la Figura D.11 se muestra el resumen de las evaluaciones FastQC según el módulo *Sequence Length Distribution* sobre los 19 archivos con lecturas *single-end* obtenidos desde Detcharoen *et al.* (2021), luego de su corrección mediante Trimmomatic.

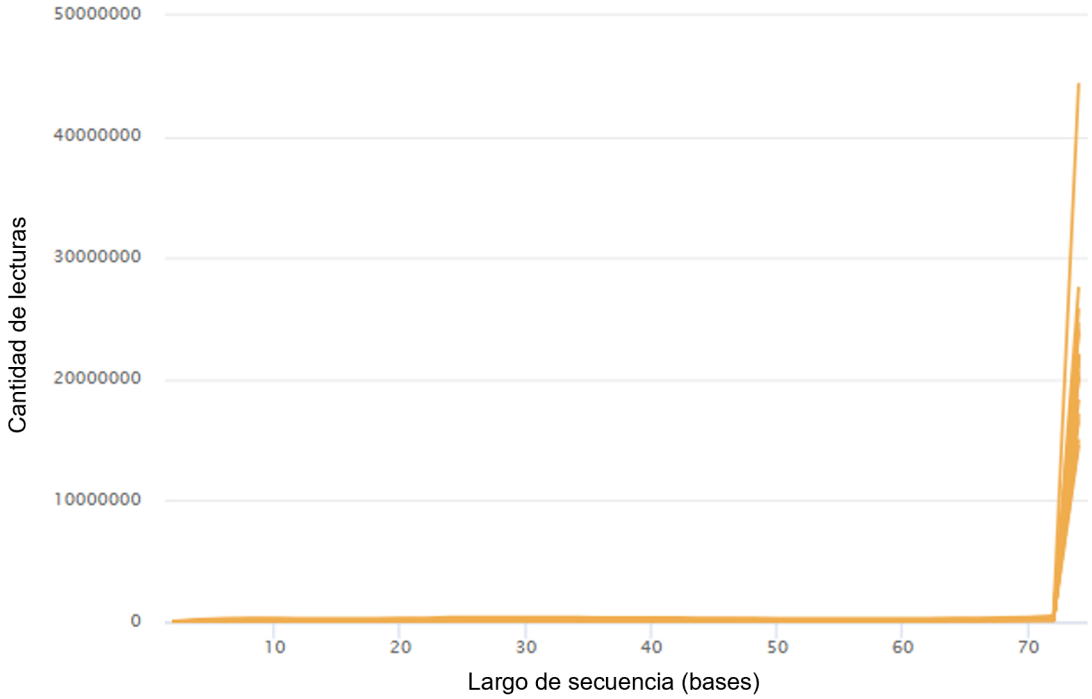

Figura D.11: Resumen de evaluación FastQC según módulo *Sequence Length Distribution* sobre archivos con lecturas de Detcharoen *et al.* (2021) luego del control de calidad.

En la Figura D.12 se presenta el resumen de los informes STAR sobre la distribución de los alineamientos de las lecturas de Detcharoen *et al.* (2021), en términos de su multiplicidad.

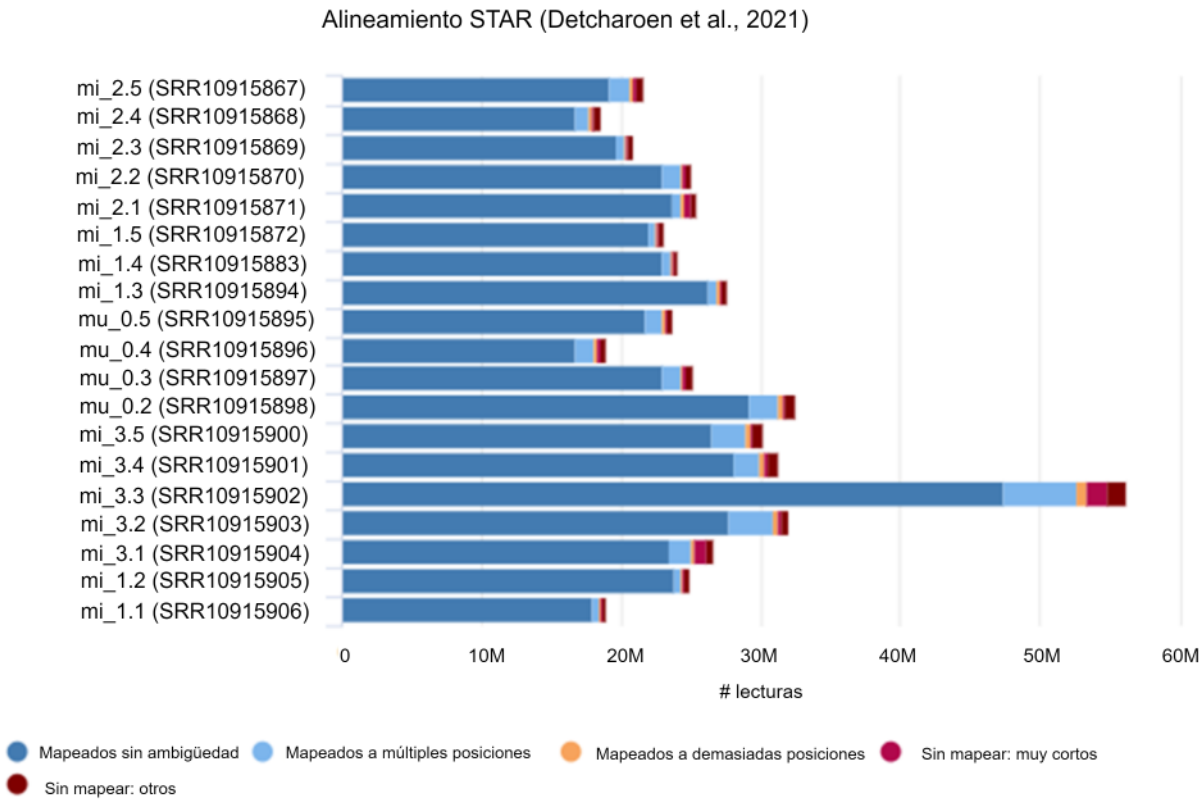

Figura D.12: Resumen de calidad de alineamientos STAR de lecturas de Detcharoen *et al.* (2021).

En la Figura D.13 se presenta el gráfico PCA sobre las cuentas normalizadas calculadas a partir de las lecturas de Detcharoen *et al.* (2021), antes de remover las muestras aisladas.

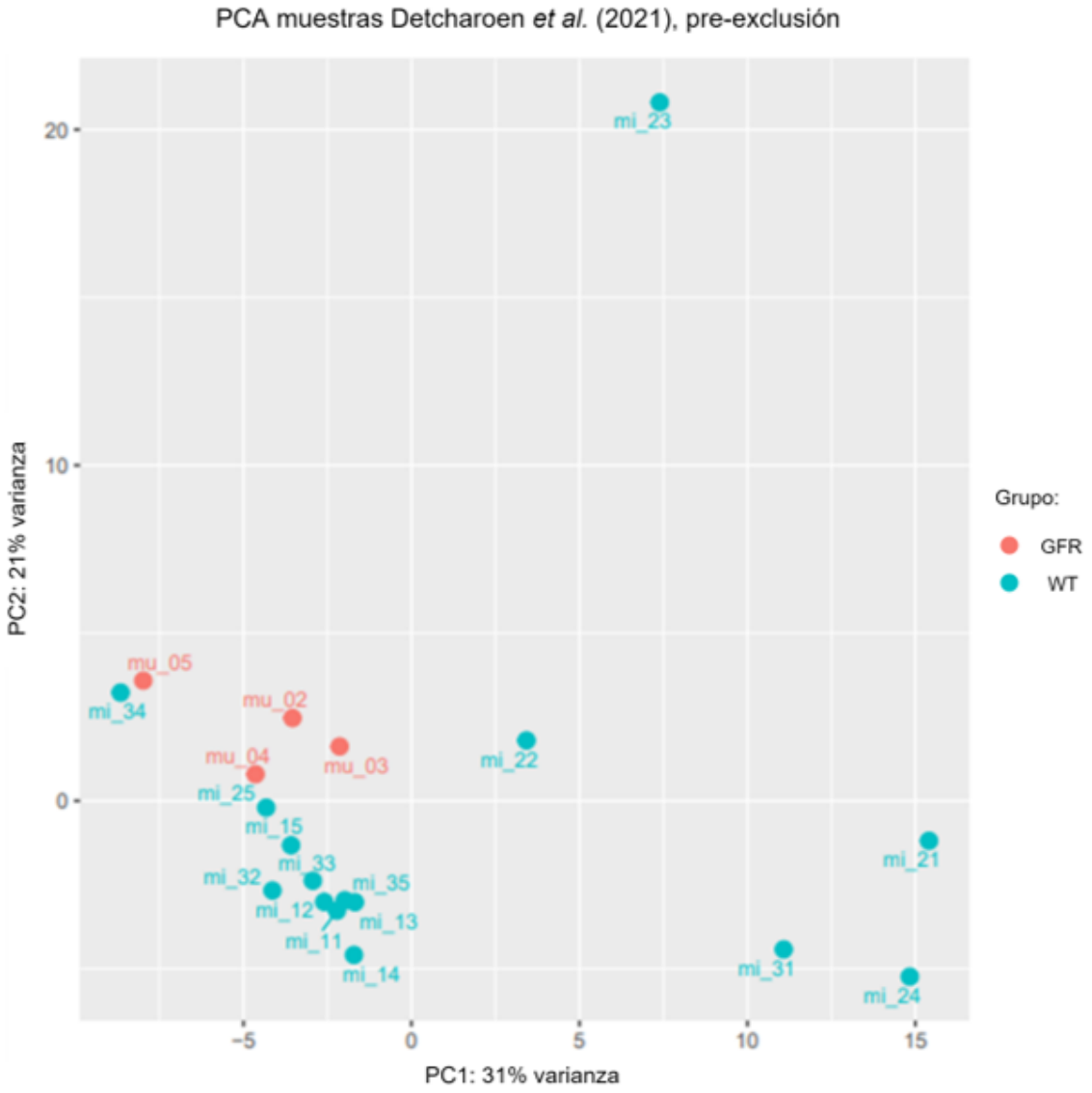

Figura D.13: PCA sobre las cuentas normalizadas de lecturas de Detcharoen *et al.* (2021) mediante DESeq2, antes de la exclusión de la muestra aislada, según análisis propio.

En la Figura D.14 se presenta el gráfico PCA sobre las cuentas normalizadas calculadas a partir de las lecturas de Detcharoen *et al.* (2021), después de remover las muestras aisladas.

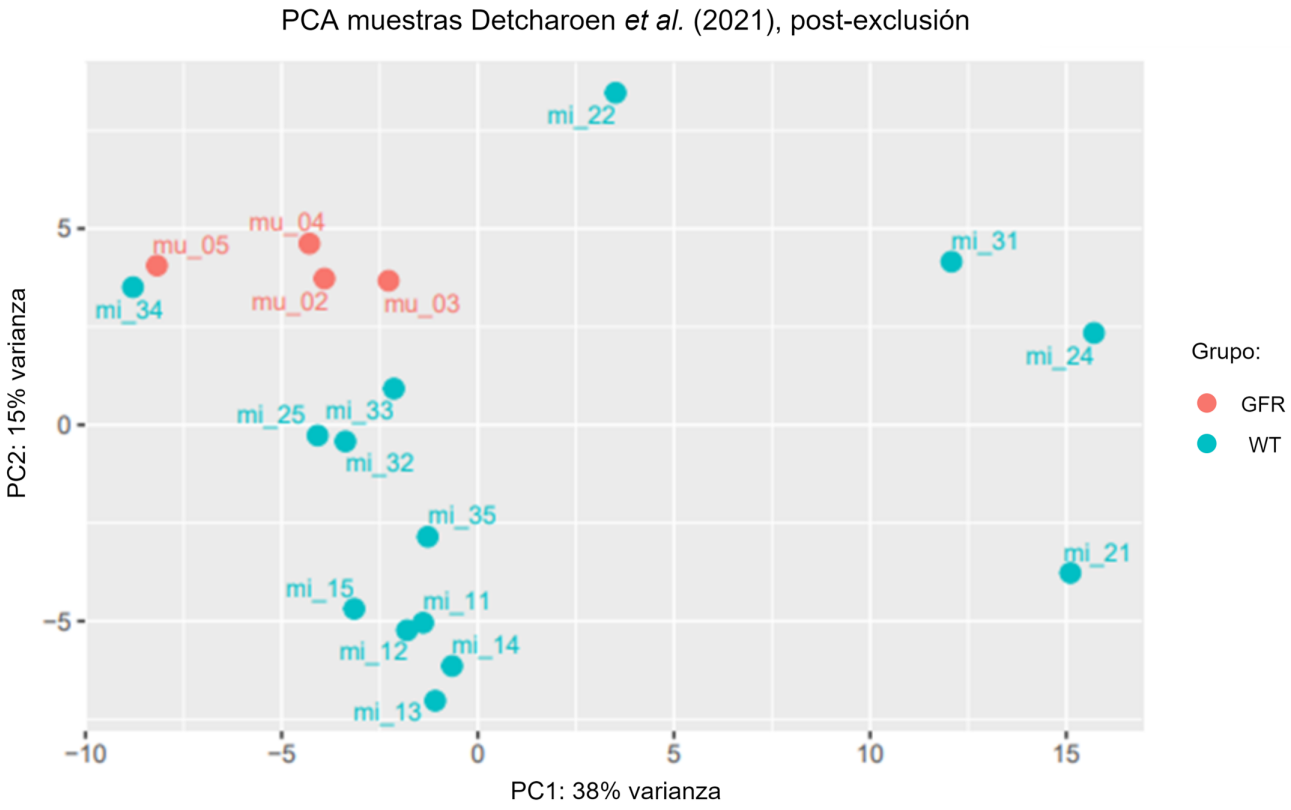

Figura D.14: PCA sobre las cuentas normalizadas de lecturas de Detcharoen *et al.* (2021) mediante DESeq2, después de la exclusión de la muestra aislada, según análisis propio.

Como se ve en la Figura D.8, la evaluación FastQC catalogó a los archivos con lecturas en bruto como anormales sólo con respecto a los módulos *Per Sequence GC Content* y *Sequence Duplication Levels*, lo cual se consideró tolerable según los criterios indicados en la Metodología. De todas formas, las lecturas fueron procesadas mediante Trimmomatic según los términos indicados en la Metodología, principalmente para remover cierto contenido de adaptadores que se puede identificar en el resumen de evaluaciones según módulo *Adapter Content* (Figura D.9).

Como se puede observar en el resumen de una segunda ronda de evaluaciones FastQC, presentado en la Figura D.10, luego del tratamiento mediante Trimmomatic las lecturas fueron catalogadas como ligeramente anormales según el módulo *Sequence Length Distribution*, lo cual no ocurrió con las lecturas brutas. La inspección del resumen específico de dicho módulo, expuesto en la Figura D.11, muestra que las lecturas siguen siendo predominantemente de 75 pares de bases o tamaños bastante cercanos. Cierto grupo de lecturas debió haber visto su tamaño reducido por la remoción de sus últimas bases en el extremo 3' que, como se vio en el Marco Teórico, suele presentar caídas de calidad.

Las evaluaciones de los alineamientos de lecturas realizados por STAR, resumidas en la Figura D.12, muestran porcentajes de *multireads* comparables a la cota inferior del rango típicamente observado, mencionado en el Marco Teórico. Lo anterior respaldó el uso de featureCounts con descarte de *multireads*.

Dos muestras fueron excluidas del análisis de expresión diferencial. La muestra mu_01, representante de la condición GFR, fue excluida atendiendo a las sospechas de los autores de que fue contaminada con ARN de machos [29]. En tanto, la muestra mi_23, representante de la condición WT, fue excluida porque el PCA basado en las cuentas normalizadas por DESeq2, presentado en la Figura D.13, la ubicó como una muestra aislada. Una nueva ejecución de DESeq2 fue realizada sin la muestra, generándose el PCA final de la Figura D.14. Se puede observar en dicho gráfico que las muestras representantes de la condición GFR (mu_02, mu_03, mu_04 y mu_05) se agrupan entre sí, al igual que 9 de las 14 muestras representantes de la condición WT (mi_11, mi_12, mi_13, mi_14, mi_15, mi_25, mi_32, mi_33 y mi_35), distinguiéndose ambos grupos por sus coordenadas según la segunda componente principal.

Anexo E

Comparación de análisis de enriquecimiento propios y publicados

El propósito de esta sección es validar en grueso la metodología propia para el análisis de enriquecimiento funcional, identificando y explicando consensos o discrepancias con respecto a los resultados presentados en las publicaciones rescatadas. Las listas obtenidas desde Baiao *et al.* (2019) mostraron los patrones de enriquecimiento más contundentes, tanto con respecto a la abundancia de términos funcionales como a los tamaños de los *adjusted p-value*. Como se mencionó en la Sección 3.1, los autores de dicho estudio también realizaron análisis de enriquecimiento de términos GO sobre las listas D.E. utilizando un *adjusted p-value* de corte de 0.05, proveyendo un buen punto de comparación para la metodología utilizada en este trabajo. En las Tablas E.1 a E.4 se comparan resultados de los análisis de enriquecimiento original y propio, para cuatro listas D.E. provenientes de Baiao *et al.* (2019).

Tabla E.1: Comparación de resultados para los 10 términos GO más enriquecidos por genes sobreexpresados en cabezas de hembras *D. paulistorum OR* según análisis original. n_T denota el tamaño de T (genes del universo anotados con el respectivo término). $n_{T,Q}$ denota el tamaño de $T \cap Q$ (genes de la lista D.E. anotados con el respectivo término). N.S.: no significativo.

Nombre	ID	Análisis original		Análisis propio	
		Adj.p-val.	$n_{T,Q}/n_T$	*Adj. p-val.*	$n_{T,Q}/n_T$
Deactivation of rhodopsin mediated signaling	GO:0016059	$6.6 \cdot 10^{-6}$	3/16	$2.0 \cdot 10^{-3}$	3/17
Rhabdomere development	GO:0042052	$8.1 \cdot 10^{-5}$	3/36	$9.7 \cdot 10^{-5}$	4/44
Visual perception	GO:0007601	$8.3 \cdot 10^{-4}$	2/18	$4.3 \cdot 10^{-5}$	5/66
Photoreceptor cell maintenance	GO:0045494	$2.2 \cdot 10^{-3}$	2/29	$2.0 \cdot 10^{-3}$	3/17
Defense response to Gram-positive bacterium	GO:0050830	$3.3 \cdot 10^{-3}$	2/36	N.S.	2/47
Positive regulation of clathrin-mediated endocytosis	GO:2000370	0.012	1/5	N.S.	1/6
Regulation of sequestering of calcium ion	GO:0051282	0.012	1/5	-	-
Regulation of synaptic transmission, glutamatergic	GO:0051966	0.012	1/5	-	-
Sensory perception of bitter taste	GO:0050913	0.012	1/5	-	-
Phototransduction, UV	GO:0007604	0.012	1/5	N.S.	1/5

Tabla E.2: Comparación de resultados para los 10 términos GO más enriquecidos por genes sobreexpresados en abdómenes de hembras *D. paulistorum OR* según análisis original. n_T denota el tamaño de T (genes del universo anotados con el respectivo término). $n_{T,Q}$ denota el tamaño de $T \cap Q$ (genes de la lista D.E. anotados con el respectivo término). N.S.: no significativo.

Nombre	GO ID	Análisis original		Análisis propio	
		Adj. p-val.	$n_{T,Q}/n_T$	Adj. p-val.	$n_{T,Q}/n_T$
Sarcomere organization	GO:0045214	$9.2 \cdot 10^{-14}$	12/27	$5.5 \cdot 10^{-12}$	11/38
Myofibril assembly	GO:0030239	$4.9 \cdot 10^{-9}$	18/38	$2.5 \cdot 10^{-19}$	17/50
Mesoderm development	GO:0007498	$4.8 \cdot 10^{-8}$	12/74	$1.5 \cdot 10^{-5}$	11/99
Muscle contraction	GO:0006936	$4.5 \cdot 10^{-7}$	10/23	$1.3 \cdot 10^{-9}$	9/30
Striated muscle myosin thick filament assembly	GO:0071688	$1.0 \cdot 10^{-6}$	4/5	$2.2 \cdot 10^{-2}$	2/3
Actin filament organization	GO:0007015	$2.3 \cdot 10^{-6}$	12/111	$1.9 \cdot 10^{-3}$	10/181
Substrate adhesion-dependent cell spreading	GO:0034446	$7.0 \cdot 10^{-6}$	4/7	$2.7 \cdot 10^{-2}$	4/12
Striated muscle contraction	GO:0006941	$1.4 \cdot 10^{-5}$	4/8	$2.9 \cdot 10^{-2}$	3/9
Flight	GO:0060361	$4.0 \cdot 10^{-5}$	4/10	$2.0 \cdot 10^{-6}$	5/10
Skeletal muscle tissue development	GO:0007519	$3.3 \cdot 10^{-4}$	3/7	-	-

.

Tabla E.3: Comparación de resultados para los 10 términos GO más enriquecidos por genes sobreexpresados en abdómenes de machos *D. paulistorum OR* según análisis original. n_T denota el tamaño de T (genes del universo anotados con el respectivo término). $n_{T,Q}$ denota el tamaño de $T \cap Q$ (genes de la lista D.E. anotados con el respectivo término). N.S.: no significativo.

Nombre	ID	Análisis original		Análisis propio	
		Adj. p-val.	$n_{T,Q}/n_T$	Adj. p-val.	$n_{T,Q}/n_T$
Sarcomere organization	GO:0045214	$3.0 \cdot 10^{-12}$	9/27	$3.4 \cdot 10^{-9}$	7/38
Myofibril assembly	GO:0030239	$6.5 \cdot 10^{-9}$	14/38	$1.2 \cdot 10^{-16}$	14/50
Skeletal myofibril assembly	GO:0014866	$3.4 \cdot 10^{-7}$	4/7	$1.3 \cdot 10^{-2}$	3/7
Muscle contraction	GO:0006936	$9.8 \cdot 10^{-7}$	8/23	$3.3 \cdot 10^{-9}$	6/30
Flight	GO:0060361	$2.0 \cdot 10^{-6}$	4/10	$1.5 \cdot 10^{-4}$	4/10
Flight behavior	GO:0007629	$4.3 \cdot 10^{-6}$	5/25	-	-
Mesorderm development	GO:0007498	$9.0 \cdot 10^{-6}$	7/74	$6.0 \cdot 10^{-4}$	5/99
Striated muscle myosin thick filament assembly	GO:0071688	$1.0 \cdot 10^{-5}$	3/5	N.S.	2/3
Actin filament organization	GO:0007015	$1.6 \cdot 10^{-5}$	8/111	$4.1 \cdot 10^{-2}$	8/181
Tricarboxylic acid cycle	GO:0006099	$2.1 \cdot 10^{-5}$	5/34	$9.9 \cdot 10^{-3}$	5/39

Tabla E.4: Comparación de resultados para los 10 términos GO más enriquecidos por genes subexpresados en cabezas de machos *D. paulistorum OR* según análisis original. n_T denota el tamaño de T (genes del universo anotados con el respectivo término). $n_{T,Q}$ denota el tamaño de $T \cap Q$ (genes de la lista D.E. anotados con el respectivo término). N.S.: no significativo.

Nombre	ID	Análisis original		Análisis propio	
		Adj. p-val.	$n_{T,Q}/n_T$	Adj. p-val.	$n_{T,Q}/n_T$
Oxidation-reduction process	GO:0055114	$4.4 \cdot 10^{-19}$	45/472	-	-
Alpha-amino acid catabolic process	GO:1901606	$2.1 \cdot 10^{-5}$	7/35	$2.8 \cdot 10^{-4}$	8/55
Tetrahydrofolate metabolic process	GO:0046653	$9.4 \cdot 10^{-5}$	3/5	$1.9 \cdot 10^{-2}$	3/8
Response to fungus	GO:0009620	$1.3 \cdot 10^{-4}$	5/54	-	-
One-carbon metabolic process	GO:0006730	$1.7 \cdot 10^{-4}$	4/14	N.S.	4/32
Response to pheromone	GO:0019236	$2.3 \cdot 10^{-4}$	4/15	N.S.	4/30
Pteridine-containing compound biosynthetic process	GO:0042559	$3.2 \cdot 10^{-4}$	3/7	-	-
Monocarboxylic acid catabolic process	GO:0072329	$4.6 \cdot 10^{-4}$	3/40	N.S.	6/54
Glycogen metabolic process	GO:0005977	$4.9 \cdot 10^{-4}$	5/15	$8.4 \cdot 10^{-4}$	7/41
Pentose-phosphate shunt	GO:0006098	$5.0 \cdot 10^{-4}$	3/8	$3.1 \cdot 10^{-2}$	3/8

Si bien se observa un consenso mayoritario entre ambos análisis con respecto a los términos enriquecidos, también se observa una cantidad considerable de discrepancias, atribuibles a múltiples factores. Aunque ambos análisis basaron el cálculo de los *p-value* en la prueba hipergeométrica, tanto los parámetros de la distribución nula (n_T, n_Q y n_U) como el valor de la realización $n_{T,Q}$ pudieron variar en virtud de distintas definiciones del universo de genes, nuevas anotaciones GO, y pérdidas de identificadores mediante la metodología propia (resumidas previamente en la Figura D.7). Por ejemplo, las Tablas E.1 a E.4 muestran que los n_T −cantidad de genes del universo anotados con T− son generalmente mayores en el análisis propio. Esto puede deberse a que, en el análisis original, el universo se definió con base en los transcritos reconstruidos a partir de los datos RNA-Seq [31], posiblemente ignorando genes con ortólogos en *D. melanogaster* pero sin expresión en las muestras. Adicionalmente, las Tablas E.1 a E.4 revelan cambios en $n_{T,Q}$, esto es, cantidad de genes de la lista D.E. analizada que están anotados con T. Las disminuciones de $n_{T,Q}$ para algunos de los términos en el análisis propio podrían deberse a la pérdida de identificadores que no pudieron ser asociados a ortólogos de *D. melanogaster* (y que por lo tanto se excluyeron de la lista D.E. obtenida), mientras que aumentos de $n_{T,Q}$ podrían explicarse por nuevas anotaciones GO introducidas entre 2019 y el presente.

Además de diferencias en el cálculo inicial de los *p-value*, los métodos de ajuste por testeo múltiple pueden haber influido en la disparidad de los *adjusted p-value*. En efecto, aunque tanto g:Profiler como el programa utilizado originalmente por los autores (TopGO) pueden tomar en cuenta la dependencia que existe entre las pruebas estadísticas para términos relacionados, g:Profiler lo hace mediante su propio método, g:SCS [46]. También cabe considerar que, en el análisis propio, se impuso tamaños máximos y mínimos para los T a testear (350 y 4, respectivamente), mientras que una restricción análoga no fue informada por Baiao *et al.* (2019) [31]. Mediante la restricción al tamaño de los T se altera la cantidad total de pruebas de hipótesis a realizar, lo cual podría haber influido en el resultado de las correcciones por testeo múltiple.

En contraste con lo descrito para las listas obtenidas desde Baiao *et al.* (2019), el análisis de enriquecimiento con parámetros estrictos fue infructuoso para las listas Car_up, Lind_up, Lind_down, He_up y He_down, encontrándose pocos o ningún término enriquecido. Como se mencionó en la Sección 3.1, un análisis de enriquecimiento análogo no fue realizado originalmente en Caragata *et al.* (2017) ni en Lindsey *et al.* (2021), por lo que no se dispone de una comparación para la falta de enriquecimiento aquí hallado [17, 15]. En el caso de He *et al.* (2019) sí se realizó un análisis de enriquecimiento de términos GO aunque, como se discutió previamente, la lista D.E. que ahí se obtuvo fue sustancialmente más abundante que la propia (además, el análisis se hizo sin separación entre genes sobre- y subexpresados). A pesar de tales diferencias, tanto el análisis propio como el original coinciden en la ausencia de términos GO enriquecidos con *adjusted p-value* < 0.05 [30]. Finalmente, en Detcharoen *et al.* (2021), el análisis de enriquecimiento original fue realizado luego del descarte de genes D.E. sin ortología entre *D. melanogaster* y *D. nigrosparsa*, y sin separación entre genes sobre- y subexpresados, resultando en sólo dos términos GO enriquecidos en hembras *D. melanogaster* infectadas: *Iron ion binding* y *Oxidation-reduction process* [29]. Las diferencias mencionadas dificultan una comparación directa de los resultados, sin embargo, se rescata el hecho de que el análisis propio también identifica la regulación de la actividad de unión a iones de hierro (por otro lado, el término *Oxidation-reduction process* puede haber sido ignorado por la restricción impuesta sobre el tamaño de los términos en el análisis propio).

Anexo F

Heatmaps complementarios

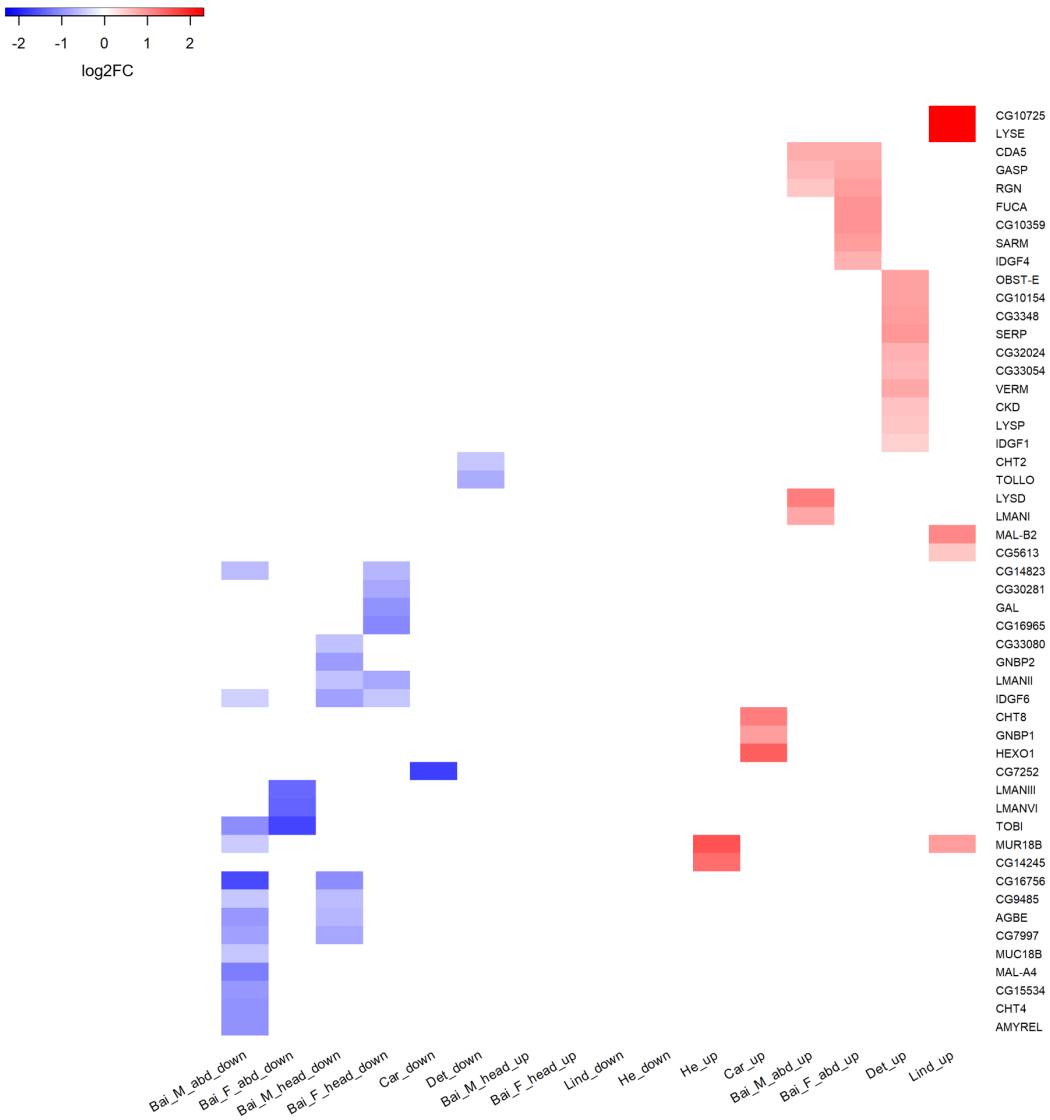

Figura F.1: *Heatmap* asociado con el Módulo *Actividad glicosil hidrolasa e hidrólisis/unión de quitina.*

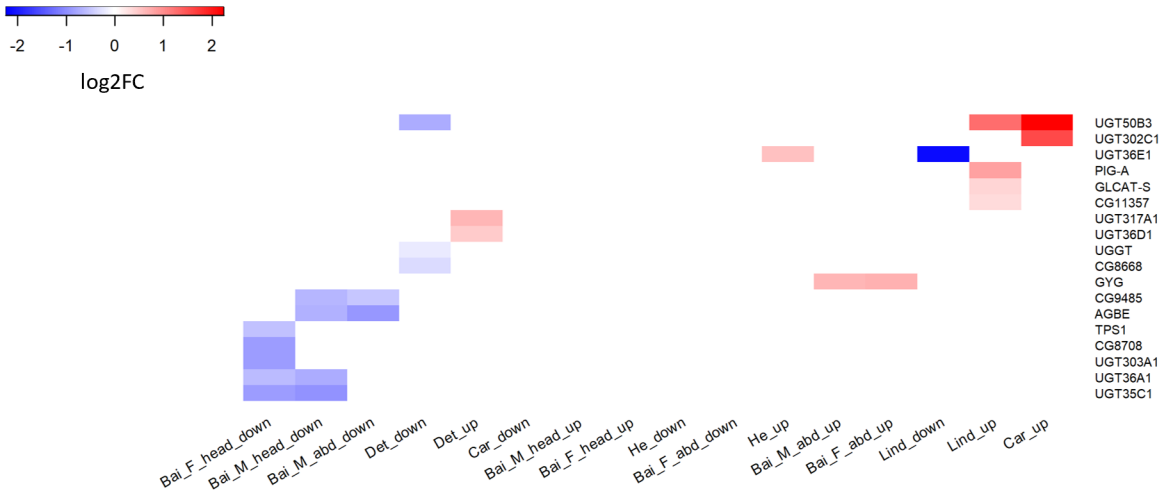

Figura F.2: *Heatmap* asociado con el Módulo *Actividad glicosil transferasa.*

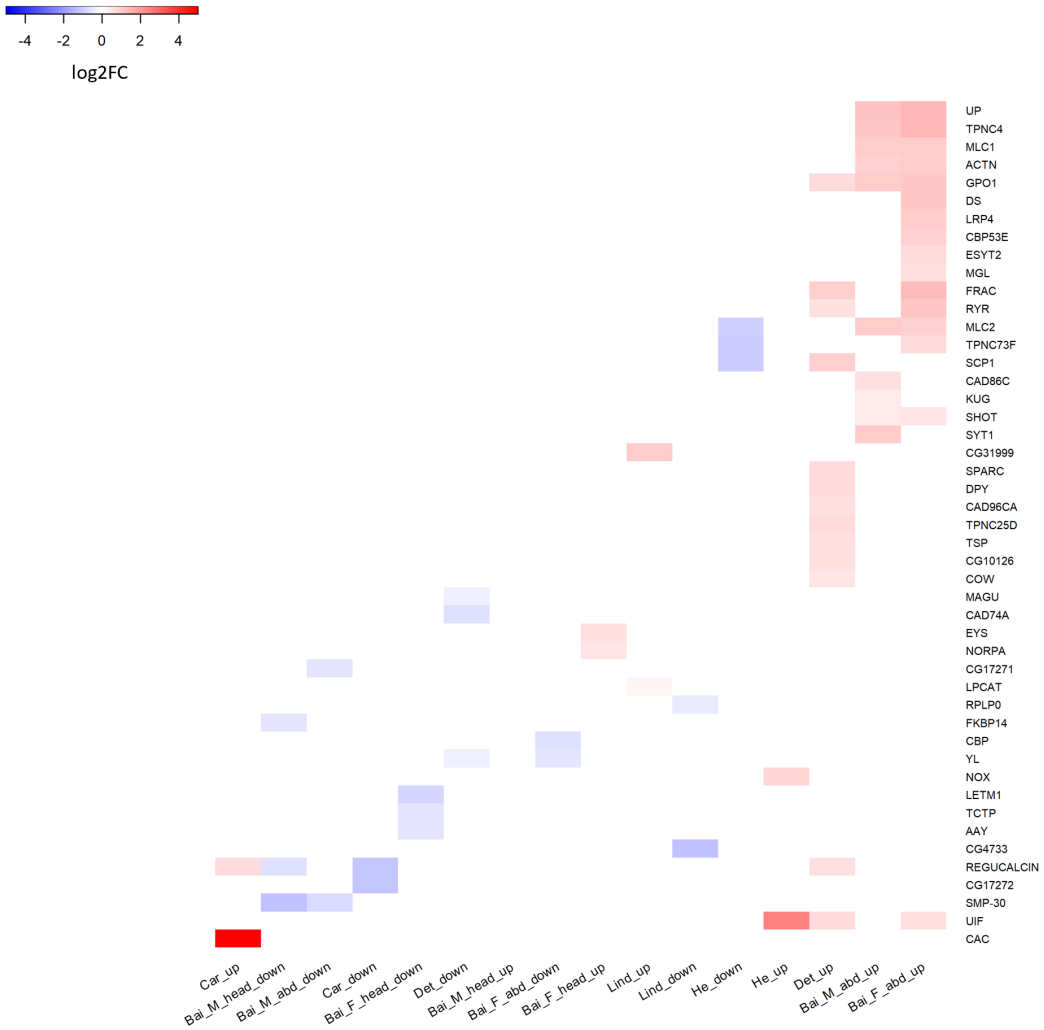

Figura F.3: *Heatmap* asociado con el Módulo *Unión a iones de calcio.*

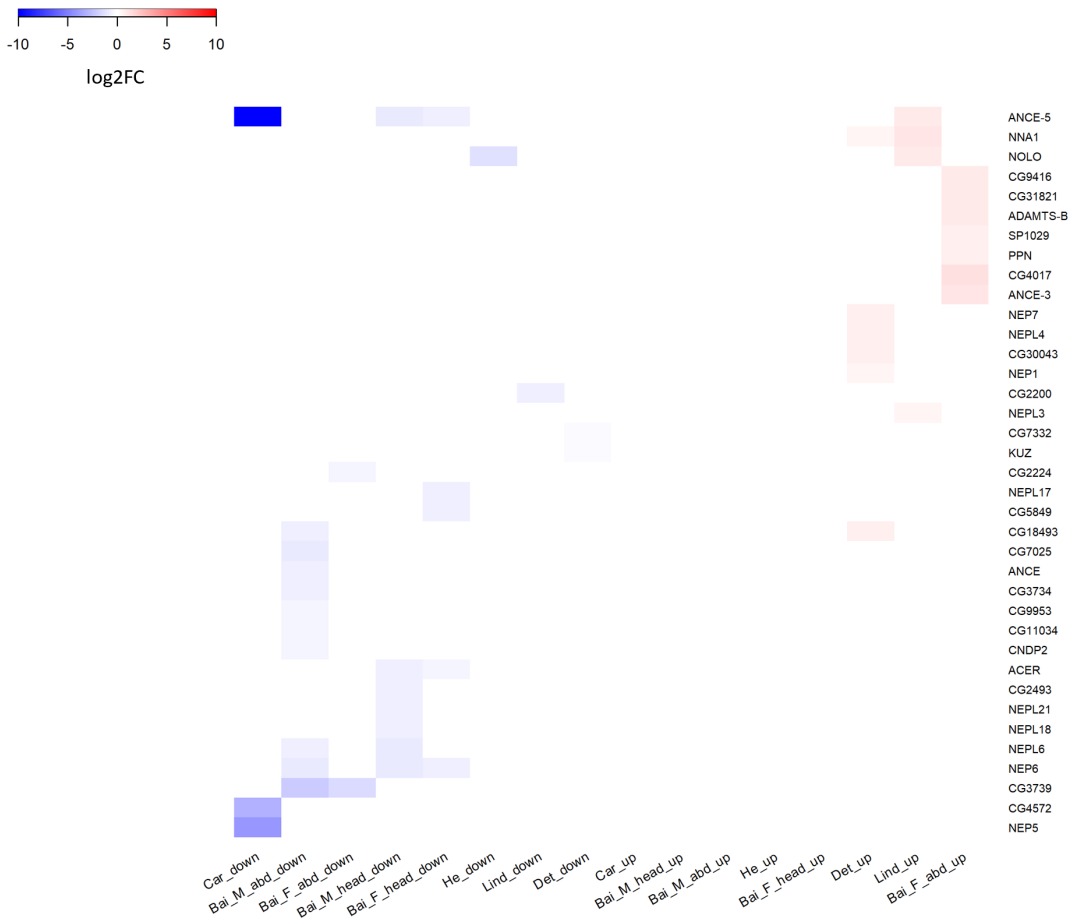

Figura F.4: *Heatmap* asociado con el Módulo *Actividad metalopeptidasa*.

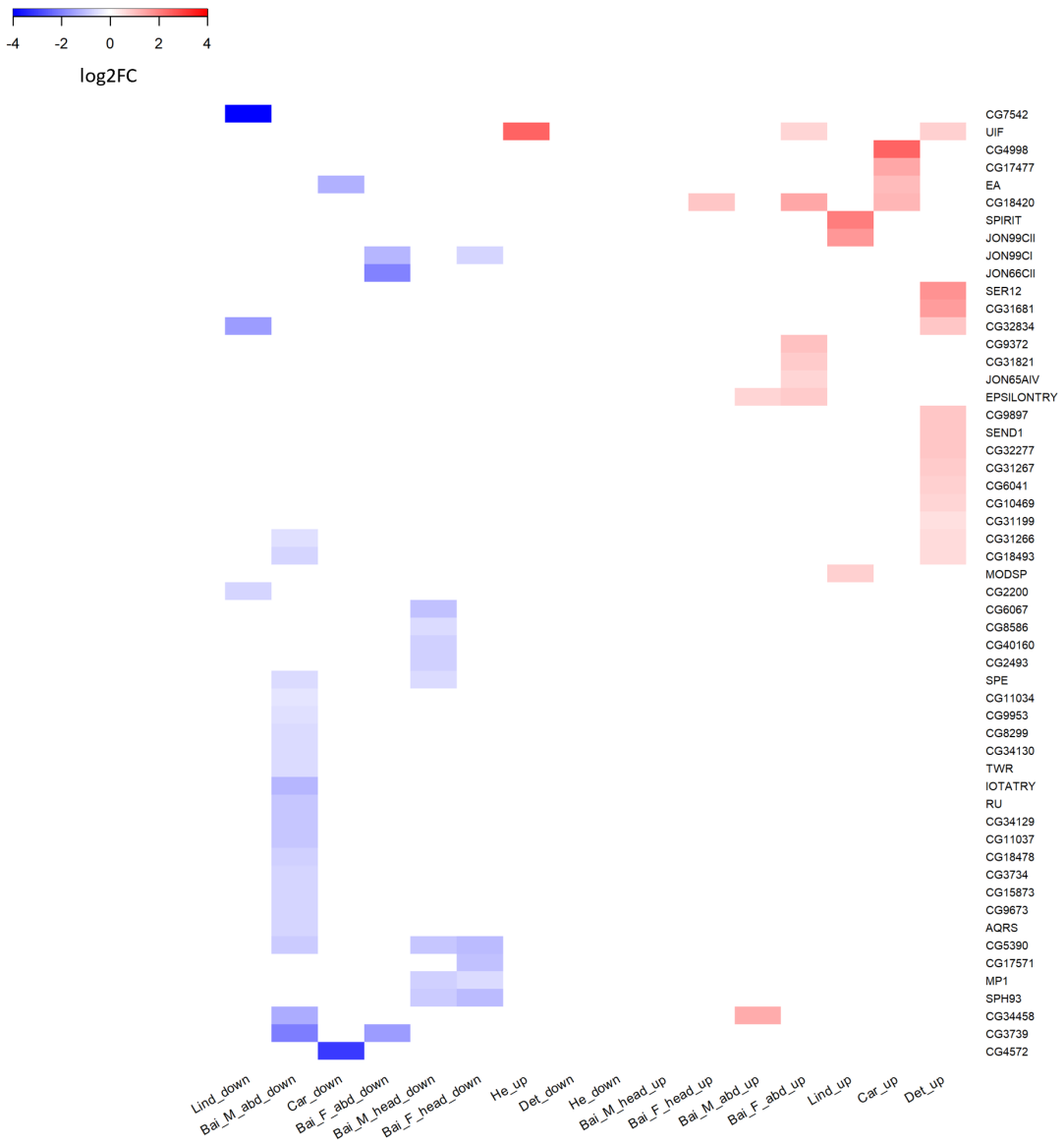

Figura F.5: *Heatmap* asociado con el Módulo *Actividad serina peptidasa.*

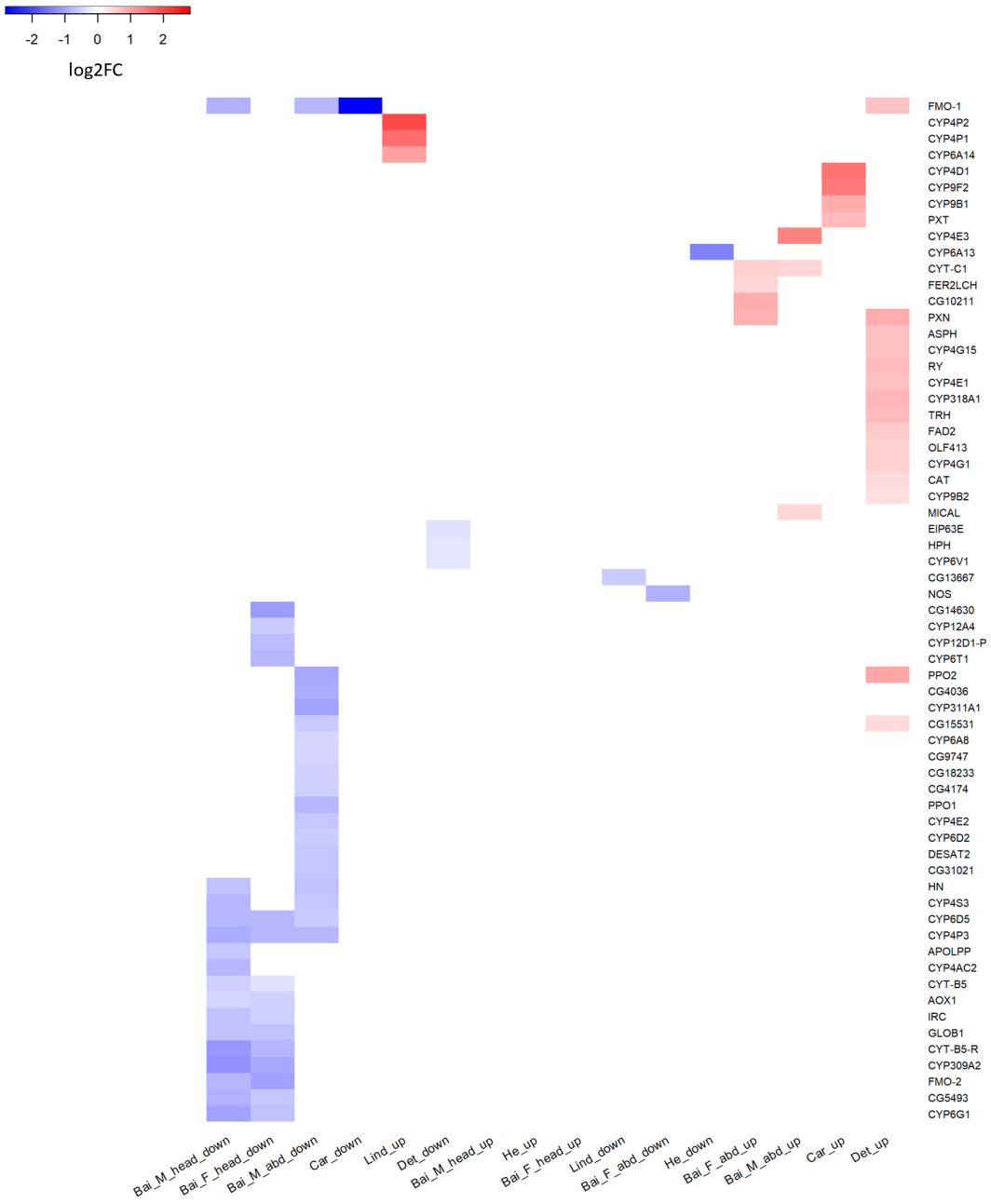

Figura F.6: *Heatmap* asociado con el Módulo *Actividad monooxigenasa*.

Anexo G

Módulos funcionales adicionales

En los siguientes apartados se adjuntan los módulos funcionales GO:MF, GO:BP y GO:CC no rescatados en el cuerpo del informe. Se advierte que los nombres de los siguientes módulos no fueron curados manualmente, sino que son el resultado directo de la ejecución de WordCloud.

G.1. Módulos de términos GO:MF

Las Figuras G.1 a G.12 muestran módulos GO:MF no rescatados en el cuerpo del informe.

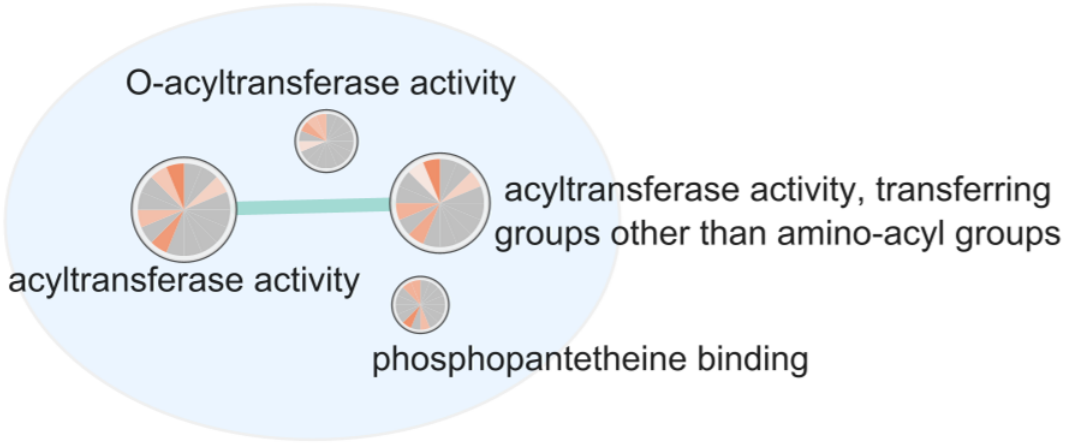

Figura G.1: Módulo *Acyltransferase activity*

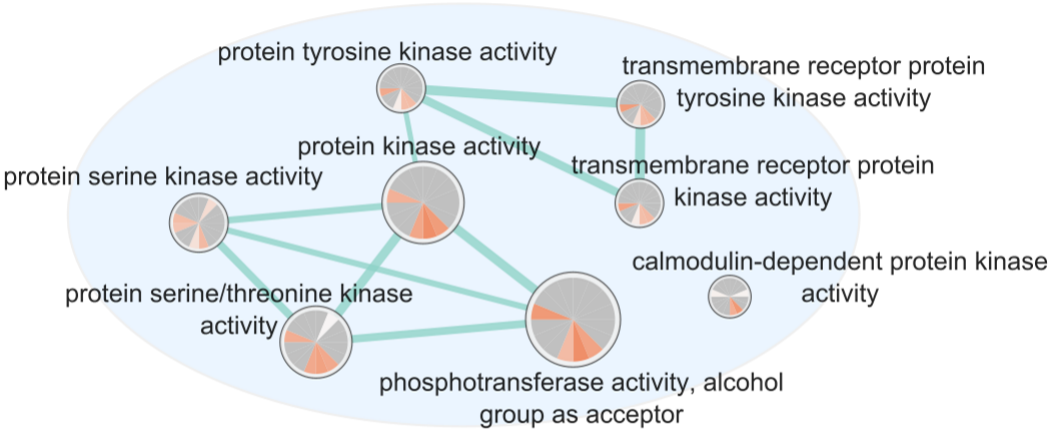

Figura G.2: Módulo *Calmodulin protein kinase*

Figura G.3: Módulo *Catalytic activity acting on DNA*

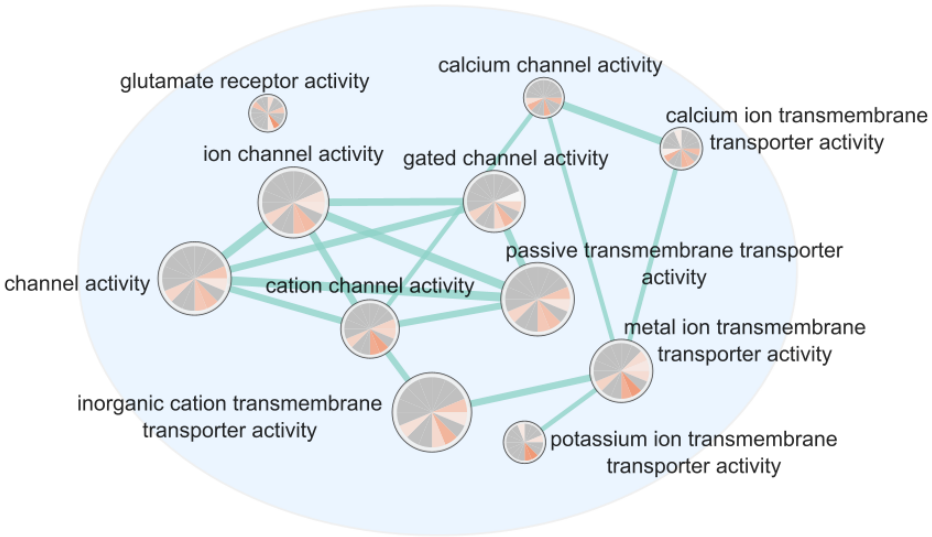

Figura G.4: Módulo *Cation channel transporter*

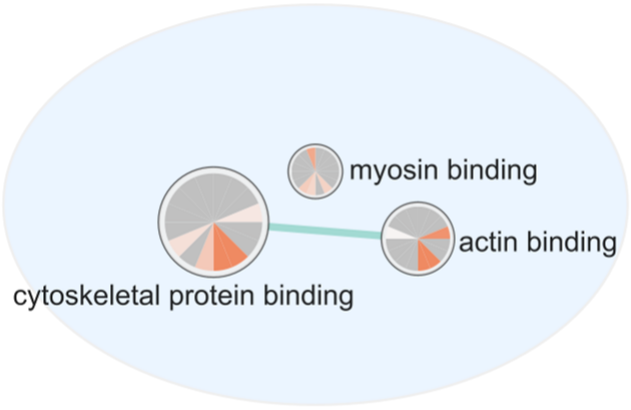

Figura G.5: Módulo *Cytoskeletal myosin actin*

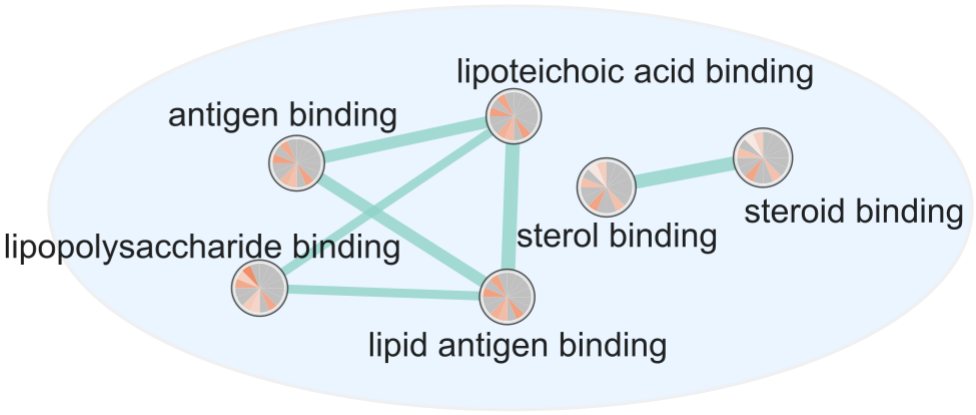

Figura G.6: Módulo *Lipid antigen binding*

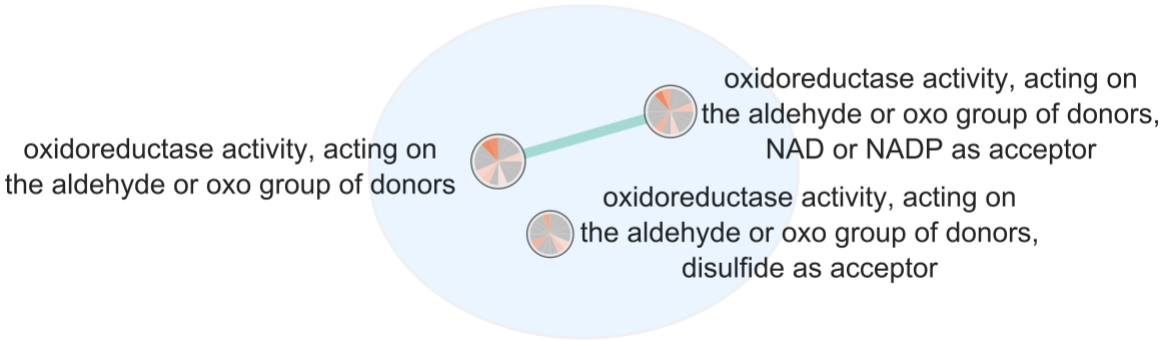

Figura G.7: Módulo *Oxidoreductase aldehyde oxo*

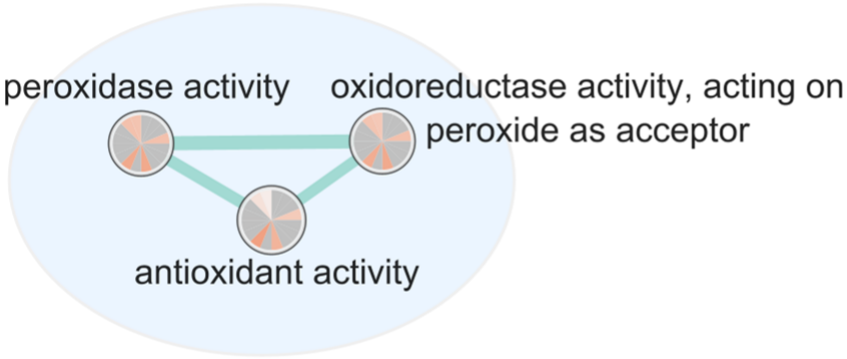

Figura G.8: Módulo *Peroxidase peroxide antioxidant*

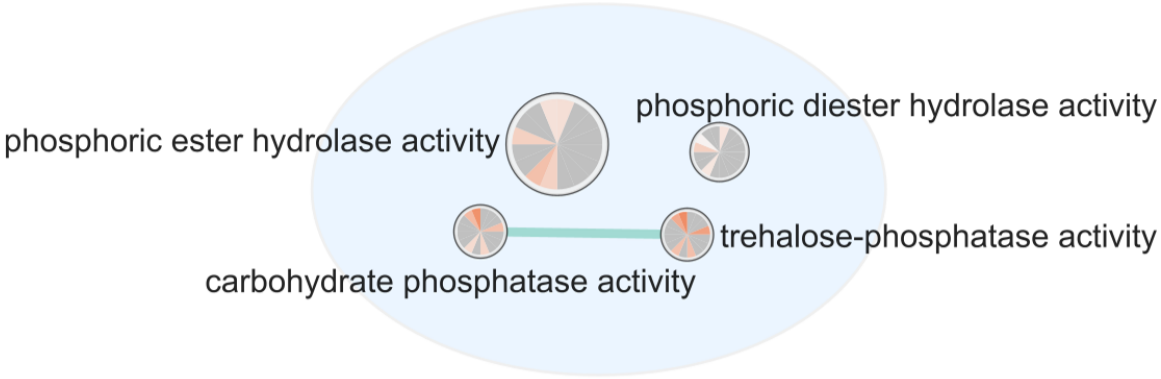

Figura G.9: Módulo *Phosphatase phosphoric hydrolase*

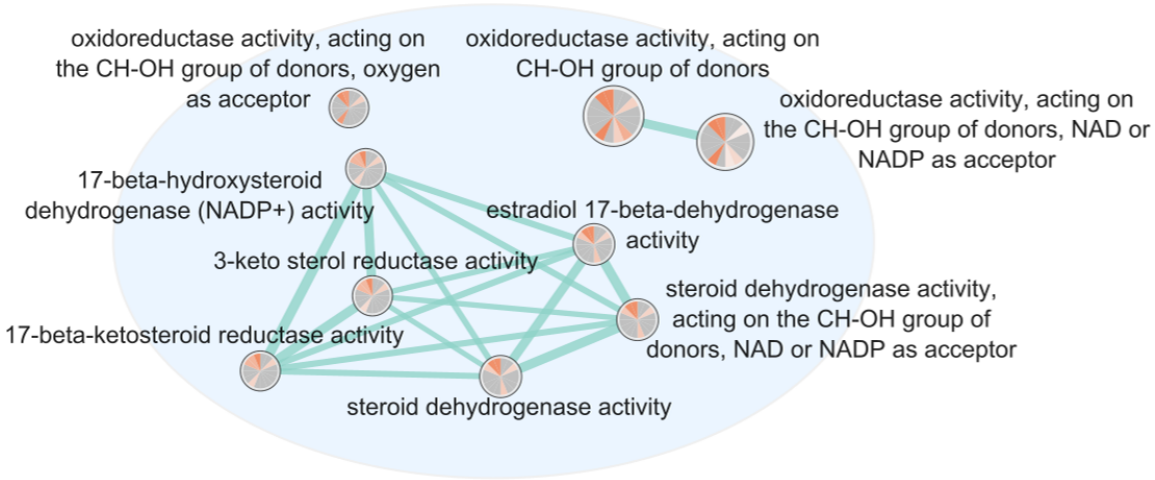

Figura G.10: Módulo *Reductase CH group*

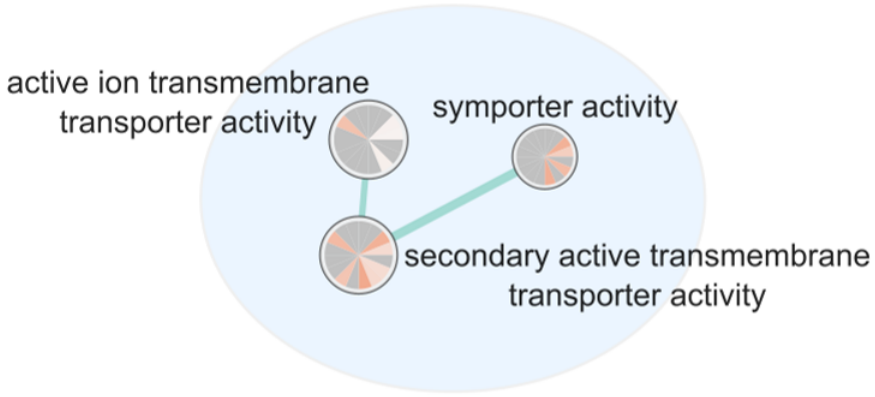

Figura G.11: Módulo *Secondary active symporter*

Figura G.12: Módulo *Unfolded protein binding*

G.2. Módulos de términos GO:BP

Las Figuras G.13 a G.37 muestran módulos GO:BP no rescatados en el cuerpo del informe.

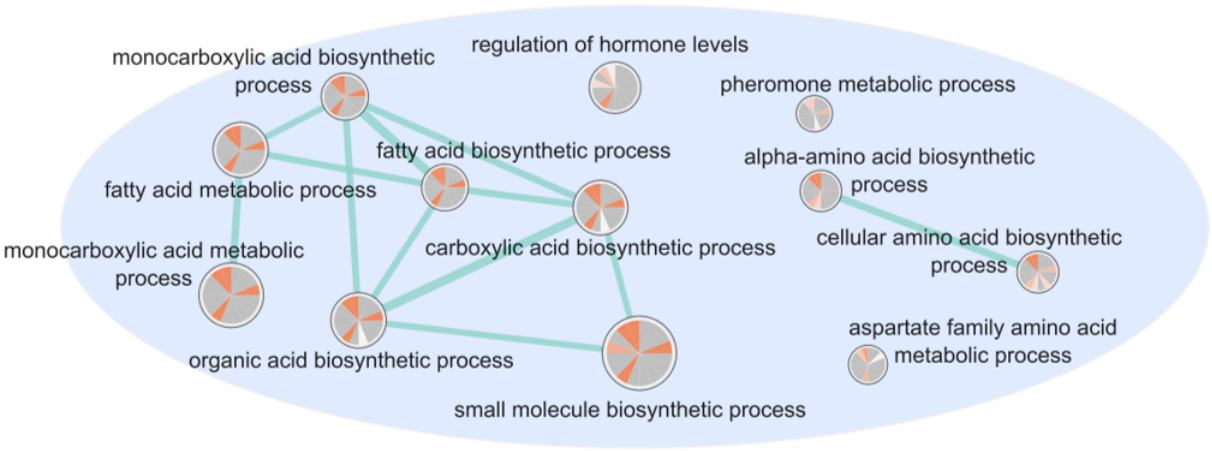

Figura G.13: Módulo *Acid biosynthetic process.*

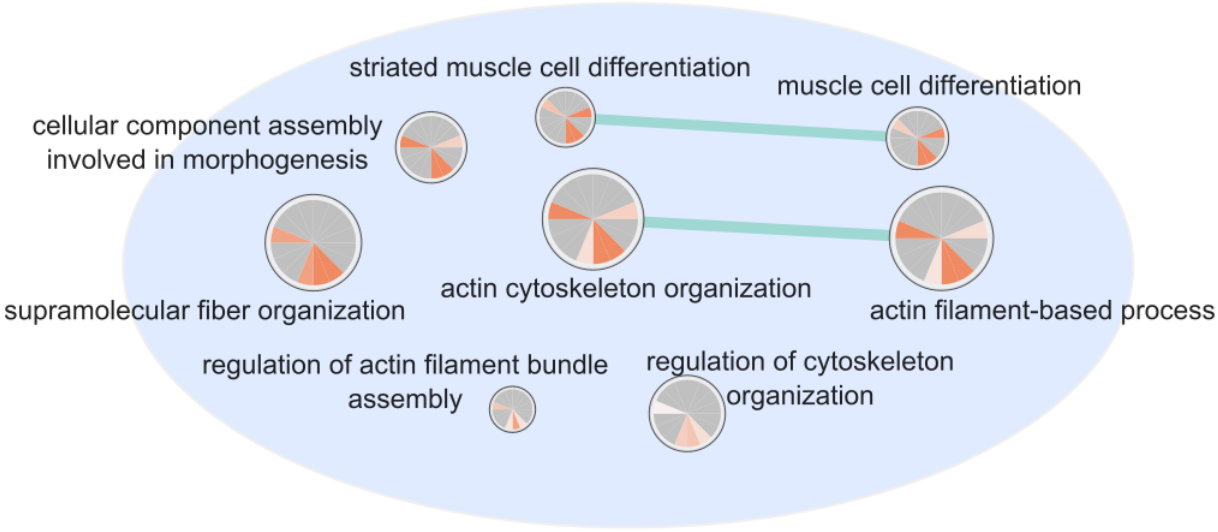

Figura G.14: Módulo *Actin filament organization*

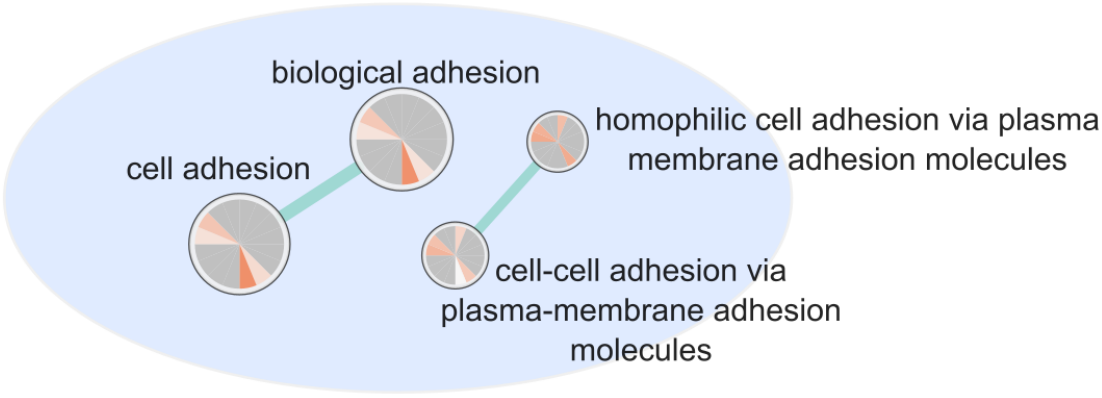

Figura G.15: Módulo *Cell adhesion*

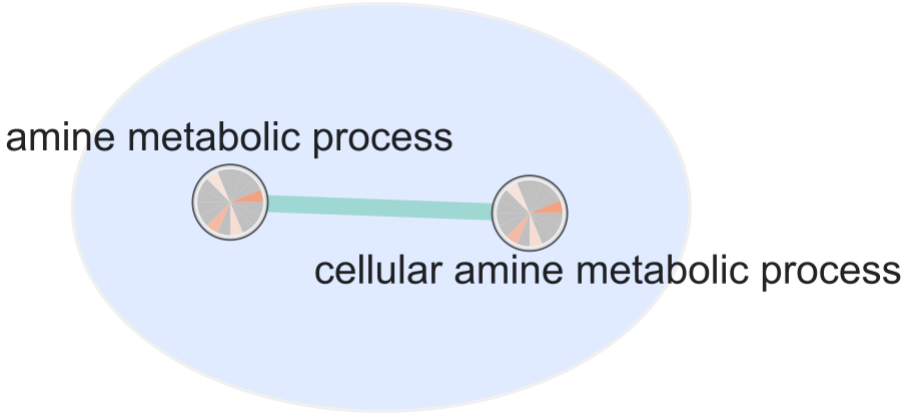

Figura G.16: Módulo *Amine metabolic process*

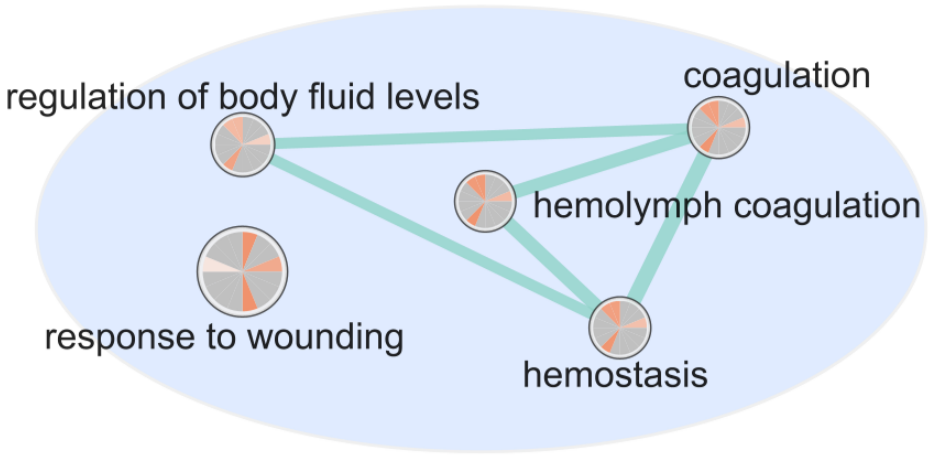

Figura G.17: Módulo *Hemolymph coagulation*

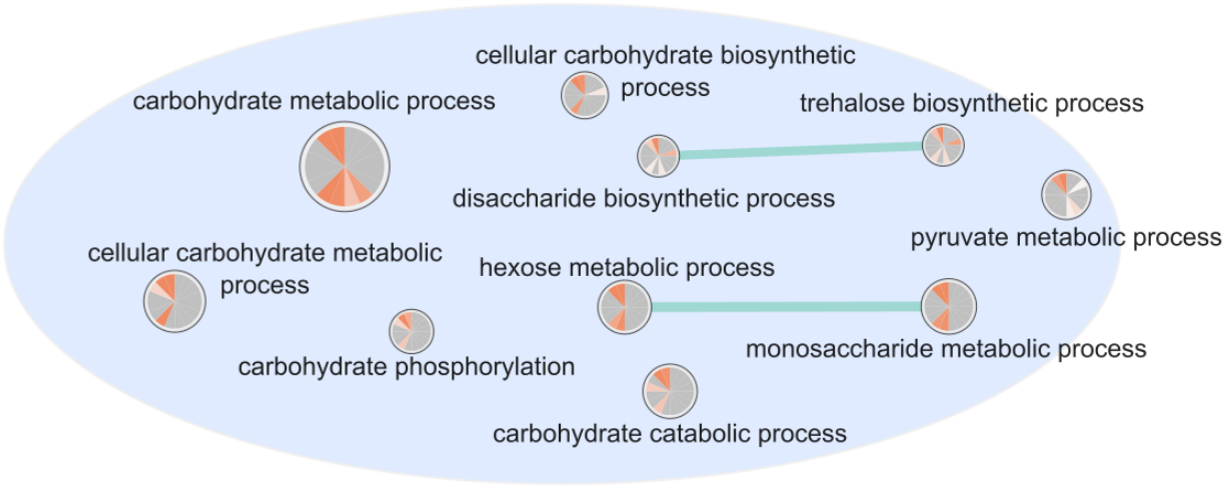

Figura G.18: Módulo *Carbohydrate metabolic process*

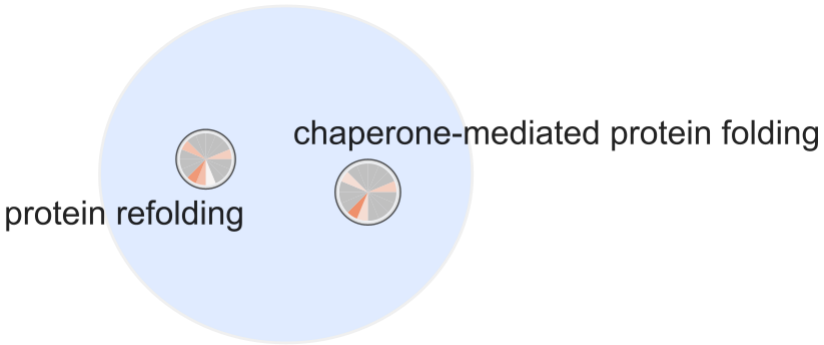

Figura G.19: Módulo *Chaperone mediated refolding*

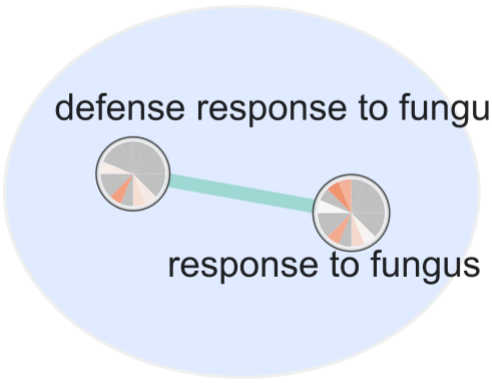

Figura G.20: Módulo *Defense response fungus*

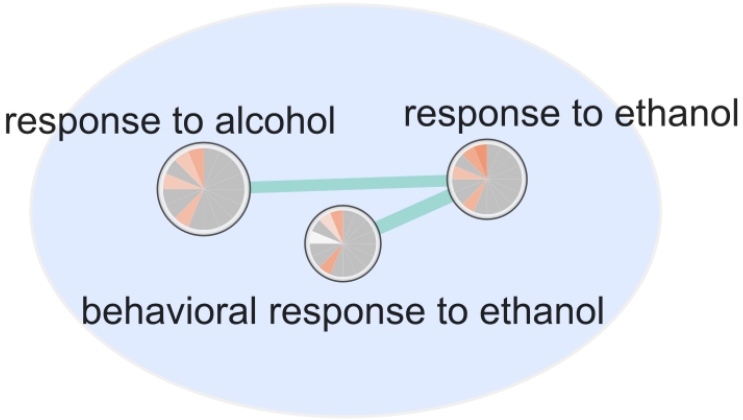

Figura G.21: Módulo *Ethanol behavioral response*

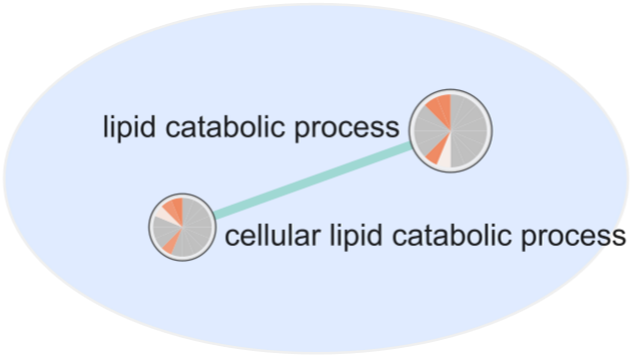

Figura G.22: Módulo *Lipid catabolic process*

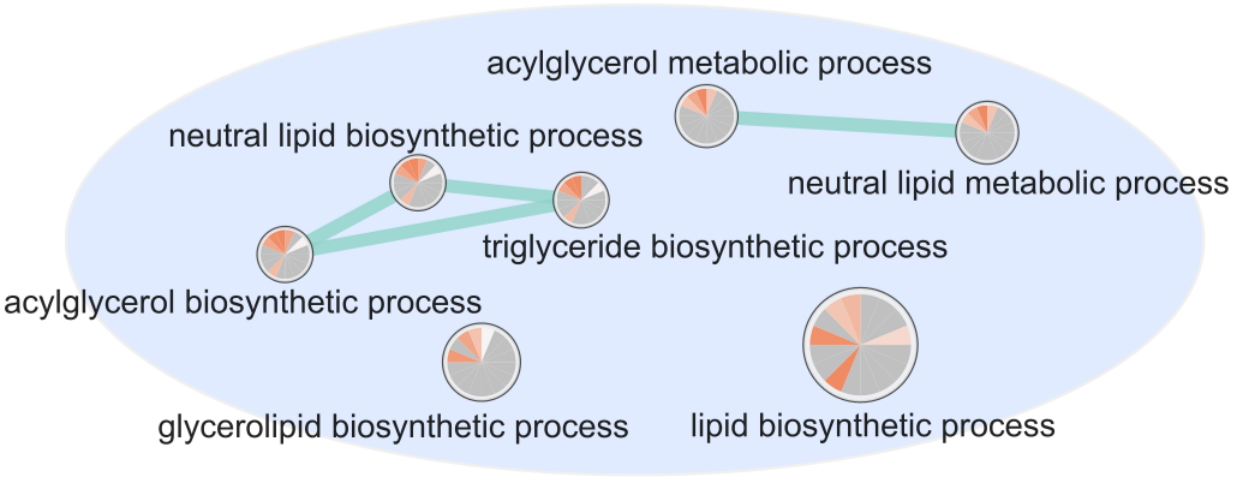

Figura G.23: Módulo *Neutral lipid biosynthetic process*

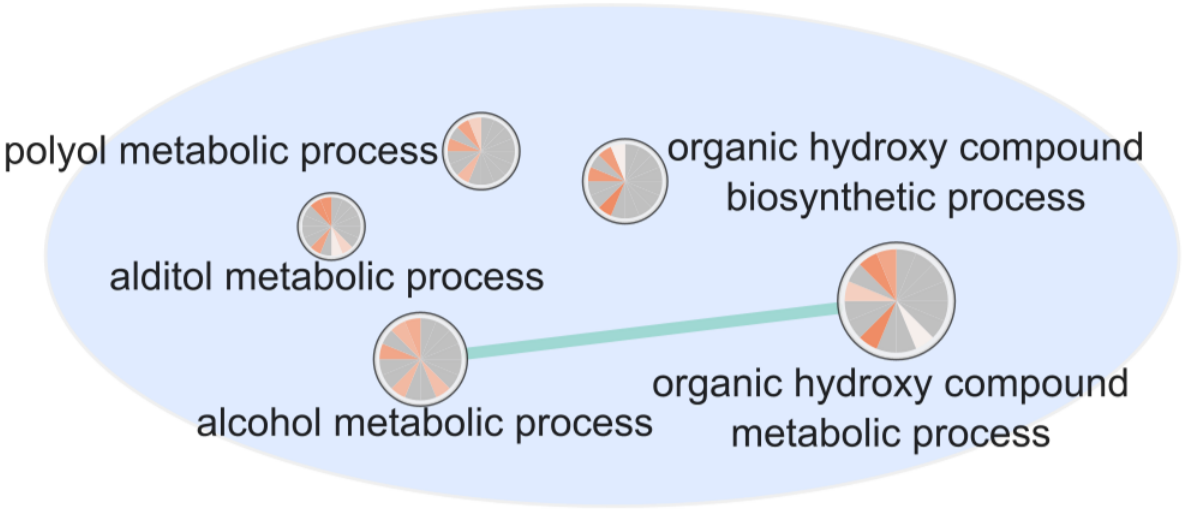

Figura G.24: Módulo *Organic hydroxy compound*

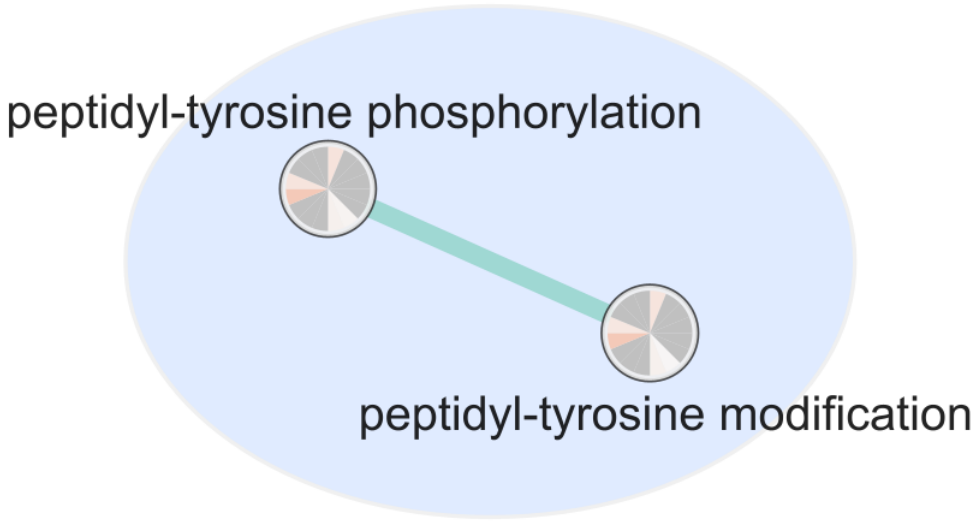

Figura G.25: Módulo *Peptidyl tyrosine phosphorylation*

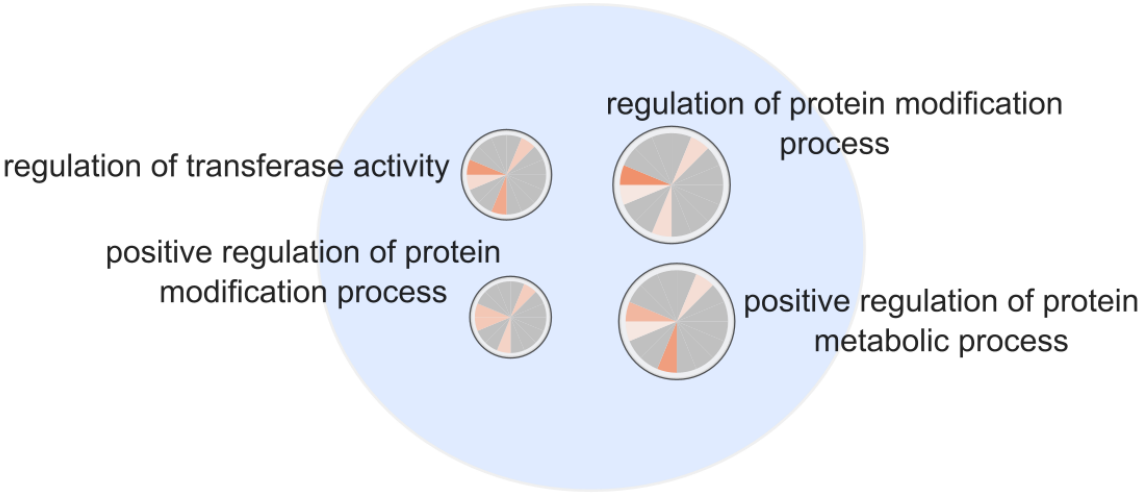

Figura G.26: Módulo *Positive regulation protein*

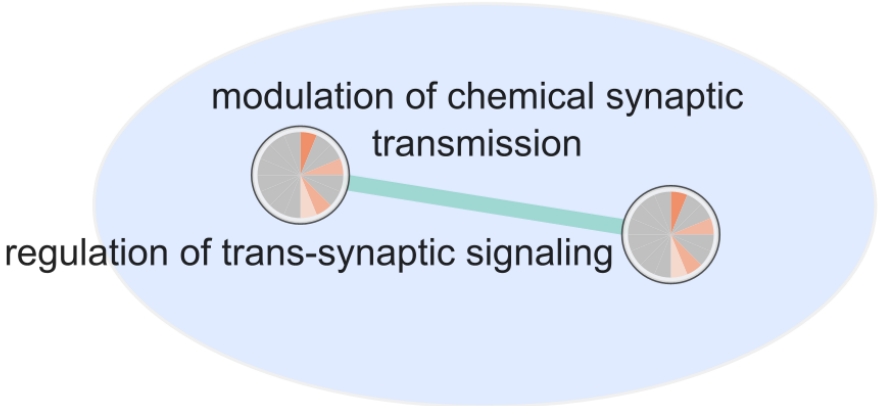

Figura G.27: Módulo *Regulation trans synaptic*

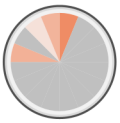

Figura G.28: Módulo *Regulation of circadian rhythm*

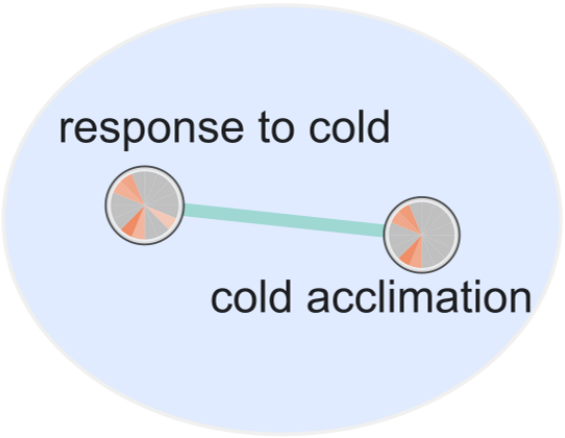

Figura G.29: Módulo *Response cold acclimation*

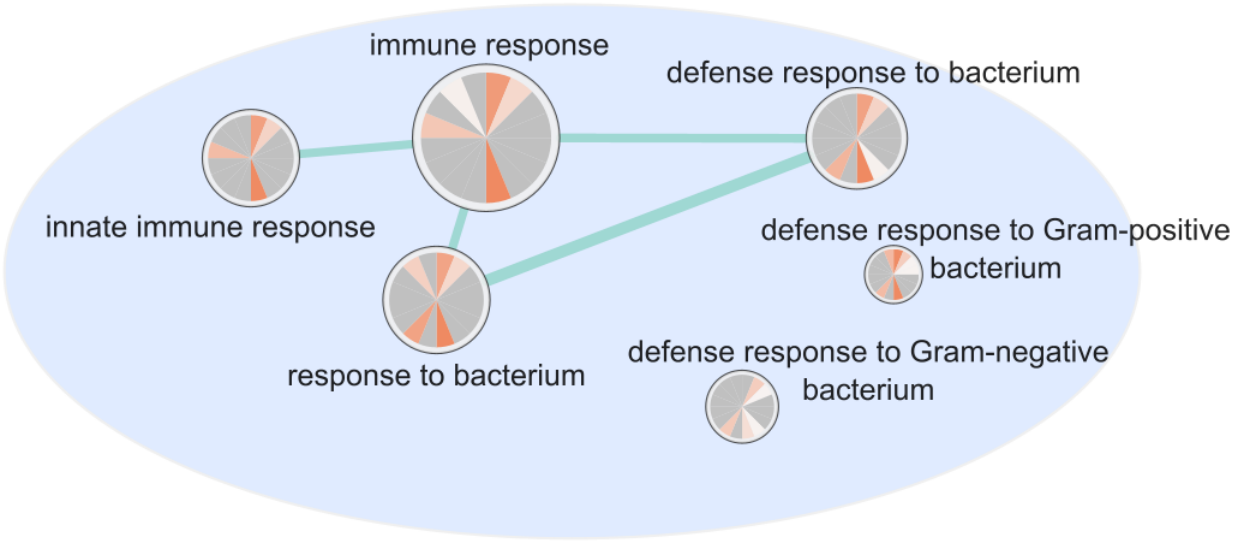

Figura G.30: Módulo *Response gram bacterium*

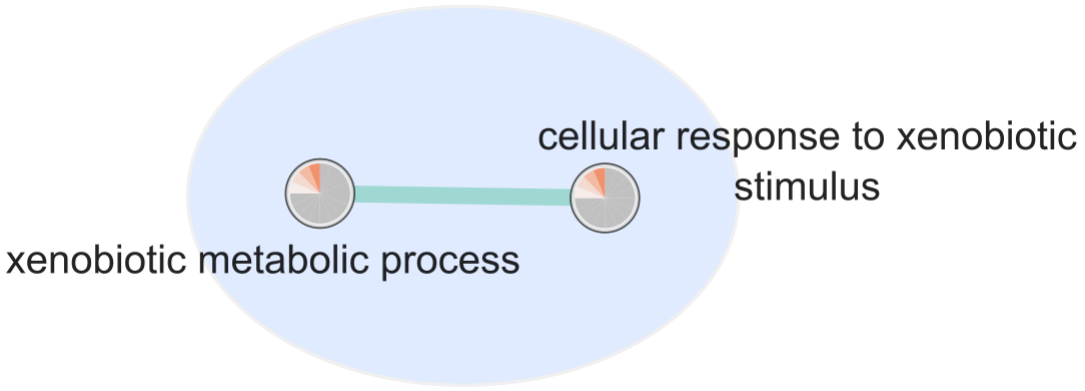

Figura G.31: Módulo *Response xenobiotic stimulus*

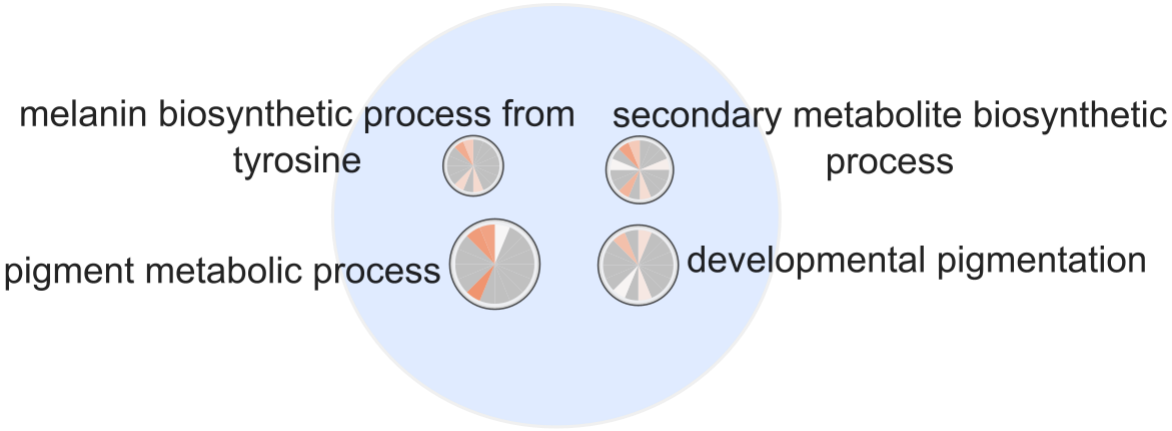

Figura G.32: Módulo *Secondary metabolite pigment*

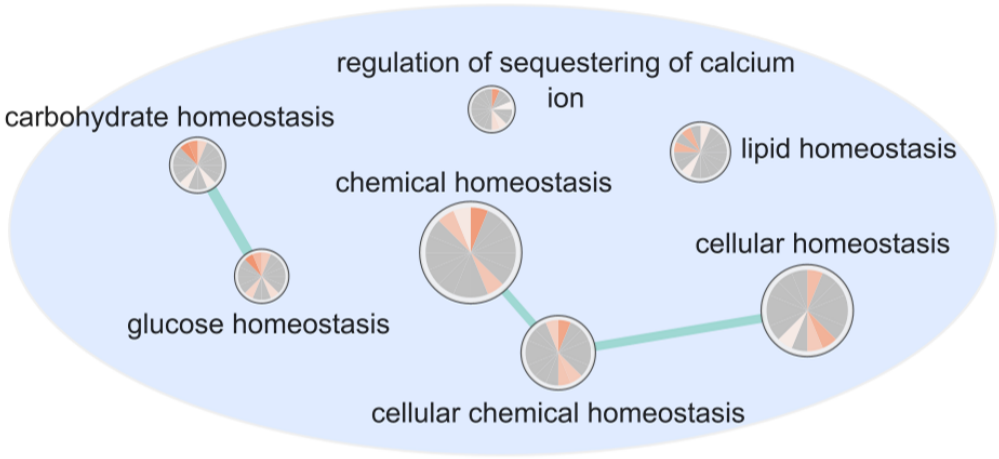

Figura G.33: Módulo *Sequestering homeostasis chemical*

Figura G.34: Módulo *Sleep*

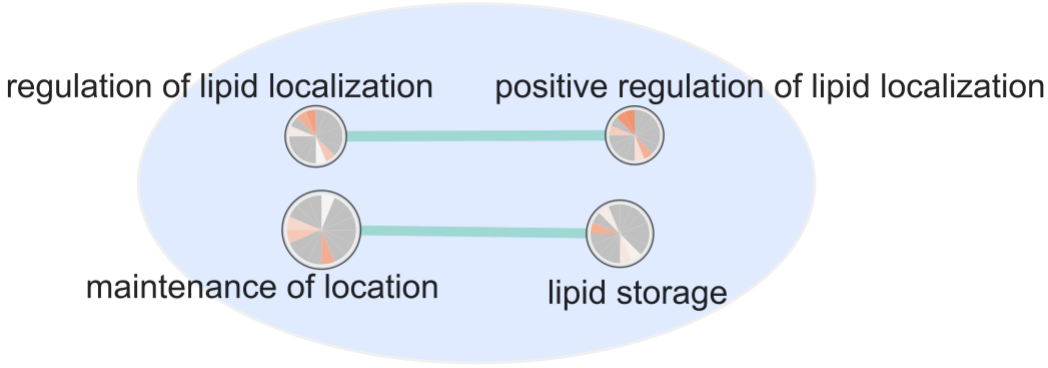

Figura G.35: Módulo *Storage maitenance localization*

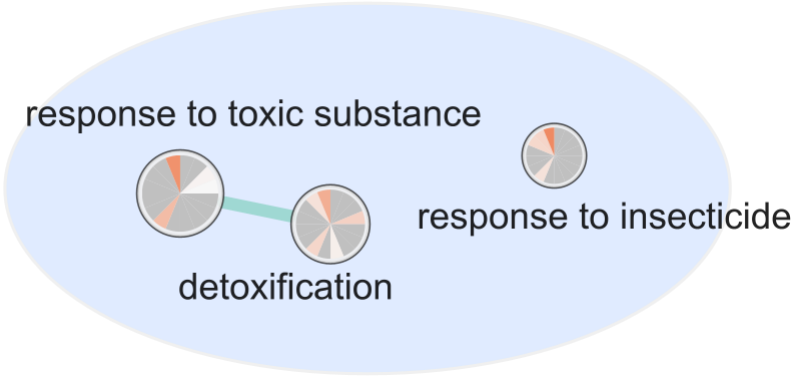

Figura G.36: Módulo *Toxic substance detoxification*

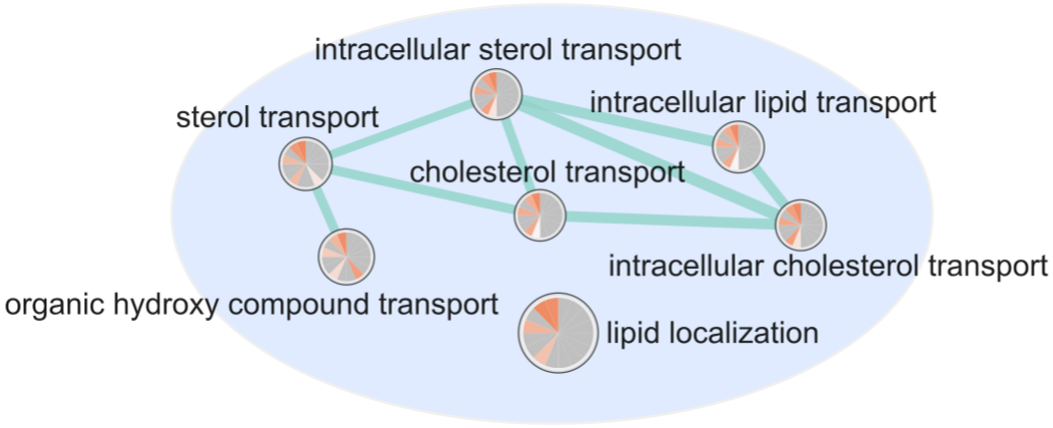

Figura G.37: Módulo *Transport intracellular sterol*

G.3. Módulos de términos GO:CC

Las Figuras G.38 a G.45 muestran módulos GO:CC no rescatados en el cuerpo del informe.

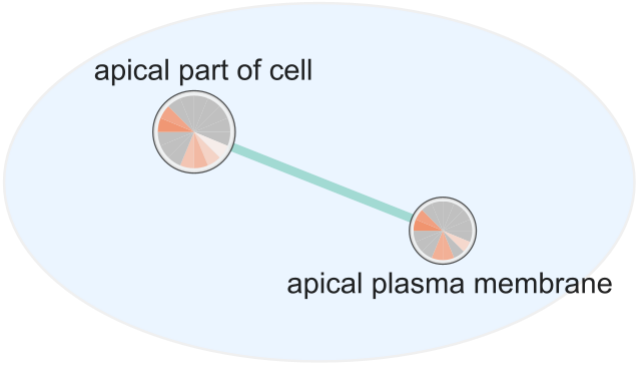

Figura G.38: Módulo *Apical membrane part*

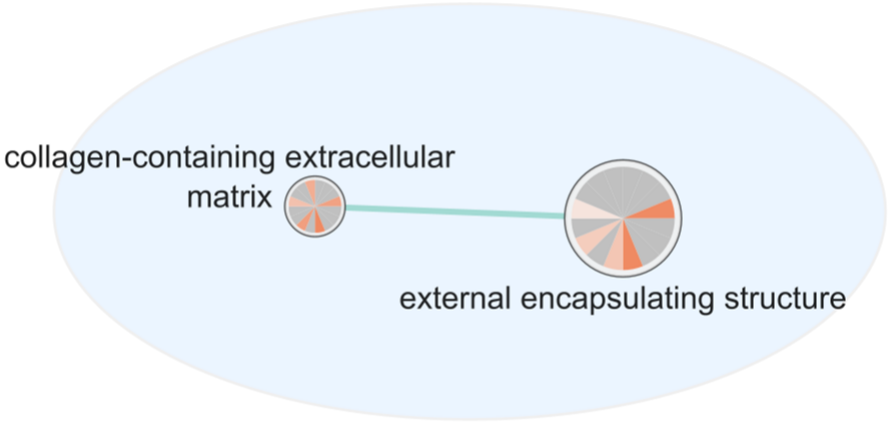

Figura G.39: Módulo *Collagen extracellular matrix*

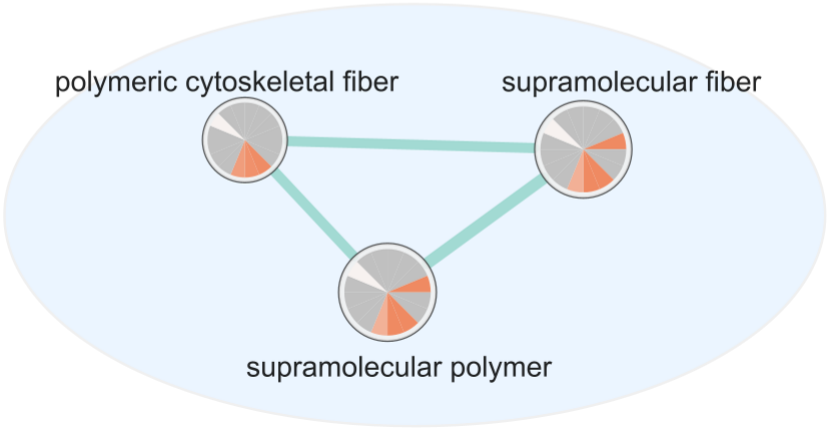

Figura G.40: Módulo *Cytoskeletal fiber supramolecular*

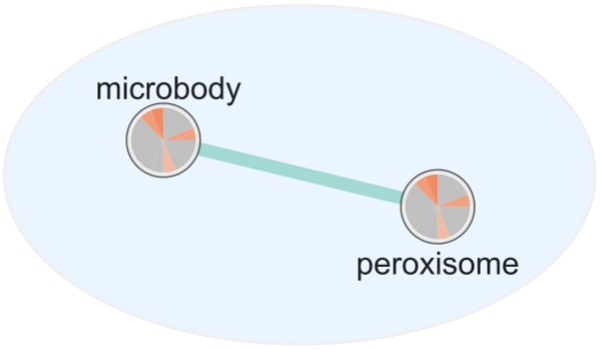

Figura G.41: Módulo *Microbody peroxisome*

Figura G.42: Módulo *Oxidoreductase complex*

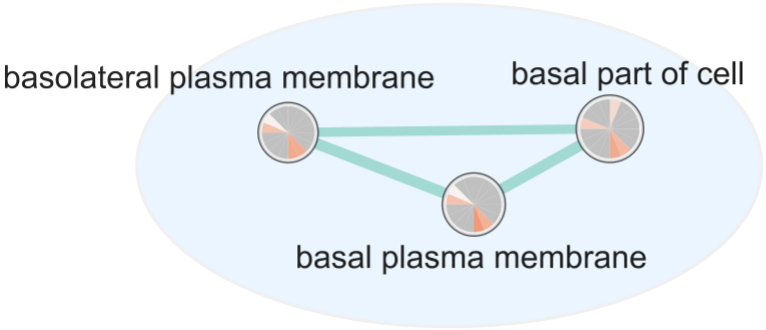

Figura G.43: Módulo *Plasma membrane basal*

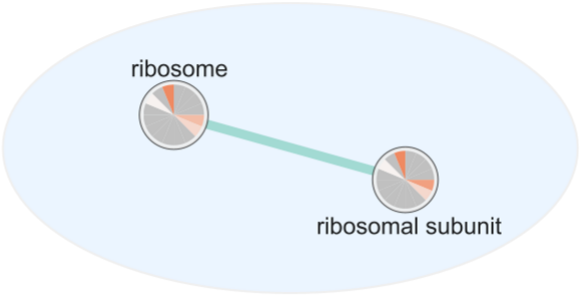

Figura G.44: Módulo *Ribosomal subunit ribosome*

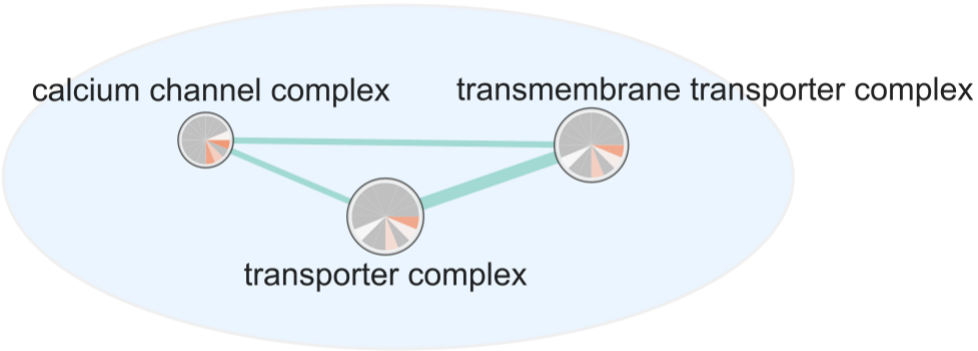

Figura G.45: Módulo *Transporter calcium complex*