CENTRO DE INVESTIGACIONES BIOLÓGICAS
DEL NOROESTE, S.C.

Programa de Estudios de Posgrado

EVIDENCIA EXPERIMENTAL COMO CRITERIO DE EVALUACIÓN Y SELECCIÓN DE ENSAMBLE *DE NOVO* DE TRANSCRIPTOMA

T E S I S

Que para obtener el grado de

Doctor en Ciencias

Uso, Manejo y Preservación de los Recursos Naturales
(Orientación en Biotecnología)

P r e s e n t a

Patricia Carvajal López

La Paz, Baja California Sur, abril de 2018

ACTA DE LIBERACIÓN DE TESIS

En la Ciudad de La Paz, B. C. S., siendo las 12 horas del día 11 del Mes de abril del 2018, se procedió por los abajo firmantes, miembros de la Comisión Revisora de Tesis avalada por la Dirección de Estudios de Posgrado y Formación de Recursos Humanos del Centro de Investigaciones Biológicas del Noroeste, S. C., a liberar la Tesis de Grado titulada:

"Evidencia experimental como criterio de evaluación y selección de ensamble *de novo* de transcriptoma"

Presentada por el alumno:

Patricia Carvajal López

Aspirante al Grado de DOCTOR EN CIENCIAS EN EL USO, MANEJO Y PRESERVACIÓN DE LOS RECURSOS NATURALES CON ORIENTACIÓN EN **Biotecnología**

Después de intercambiar opiniones los miembros de la Comisión manifestaron su **APROBACIÓN DE LA TESIS**, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Dr. Eduardo Romero Vivas
Co-Director de Tesis

Dr. Joaquín Gutiérrez Jagüey
Co-Director

Dr. Fernando Daniel Von Borstel Luna
Co-Tutor

Dr. Claudio Humberto Mejía Ruiz
Co-Tutor

Dra. Gabriella Rustici
Co-Tutor

Dra. Norma Yolanda Hernández Saavedra,
Directora de Estudios de Posgrado y
Formación de Recursos Humanos

Comité Tutorial

Dr. Eduardo Romero Vivas
Co-Director de Tesis
Centro de Investigaciones Biológicas del Noroeste, S.C.

Dr. Joaquín Gutiérrez Jagüey
Co-Director de Tesis
Centro de Investigaciones Biológicas del Noroeste, S.C.

Dr. Fernando Daniel Von Borstel Luna
Co-Tutor
Centro de Investigaciones Biológicas del Noroeste, S.C.

Dr. Claudio Humberto Mejía Ruiz
Co-Tutor
Centro de Investigaciones Biológicas de Noroeste, S.C.

Dra. Gabriella Rustici
Co-Tutora
University of Cambridge, Cambridge, Reino Unido

Comité Revisor de Tesis

Dr. Eduardo Romero Vivas
Dr. Joaquín Gutiérrez Jagüey
Dr. Fernando Daniel Von Borstel Luna
Dr. Claudio Humberto Mejía Ruiz
Dra. Gabriella Rustici

Jurado de Examen

Dr. Eduardo Romero Vivas
Dr. Joaquín Gutiérrez Jagüey
Dr. Fernando Daniel Von Borstel Luna
Dr. Claudio Humberto Mejía Ruiz
Dra. Gabriella Rustici

Suplentes

Dr. Carlos Eliud Angulo Valadez
Dra. Gracia Alicia Gómez Anduro

Resumen
El contenido de ARN se obtiene con fragmentación aleatoria (RNA-Seq), generando millones de lecturas, que en ausencia de referencias se reconstruyen como ensamblajes *de novo*, basándose en algoritmos computacionales. Diversos factores biológicos y técnicos inducen errores en el ensamblaje. Establecer un criterio que permita seleccionar ensambles de calidad se vuelve aún más crítico en el caso de ensamblaje *de novo*, debido a que las métricas cuantitativas han demostrado no corresponder a referencias de prueba. Sin embargo, existen fuentes de información de la expresión génica que podrían aprovecharse. Esta investigación propone el uso de esta evidencia en la evaluación de calidad y selección de un ensamble. Para llevar a cabo este análisis se utilizó la información de microarreglos, UniGenes y lecturas RNA-Seq de tres organismos modelo: Ratón *Mus musculus*, Mosca de la Fruta *Drosophila melanogaster* y Pulga de Agua *Daphnia pulex*; y un organismo no modelo, Camarón Blanco *Litopenaeus vannamei* para esta investigación. Los resultados fueron verificados por medio de mapeos a referencias de transcriptoma y bases de datos de proteínas UniProt/Swiss-Prot. Primero se analizó la variabilidad de ensamble en la *D. melanogaster*, *D. pulex* y *L. vannamei* en términos del contenido de los *contigs* obtenidos al repetir un ensamblaje bajo las mismas condiciones. Esto permitió analizar la influencia de los recursos computacionales en el proceso. Posteriormente, se generaron ensambles de *M. musculus* y *D. melanogaster* cambiando parámetros de ensamblaje, permitiendo la evaluación de las métricas convencionales de calidad y el mapeo a transcriptomas de referencia. Una vez determinados los ensambles de mayor calidad, se evaluó la estrategia propuesta de uso de sondas de microarreglos como criterio de selección directo. Esta evidencia experimental se generalizó usando Modelos Markovianos Ocultos (HMM) que permitieron la selección del mejor ensamble a través de un criterio probabilístico. Este criterio se aplicó a ensambles de *D. melanogaster* y se extendió al ensamblaje de *L. vannamei* haciendo uso de UniGenes. *In silico,* se encontró que las variaciones cuantitativas en cinco repeticiones de un ensamble originadas bajo las mismas condiciones fueron menores al 0.001%; sin embargo, las variaciones cualitativas alcanzaron un 22%. Por otro lado, se obtuvo hasta 7 veces mayor variabilidad al usar equipos con baja memoria (~24 GB), comparados con los de mayor memoria (128 GB), mostrando la importancia de la elección del equipo de cómputo. Contrario a las métricas cuantitativas, las estrategias propuestas sí permitieron identificar a los ensambles que coinciden con los mapeos a las referencias correspondientes: los de mayor calidad. Finalmente, se proponen estrategias de ensamblaje múltiple que aprovechen la variabilidad para la prospección de *contigs*; y el uso de evidencia experimental para la evaluación de calidad y selección de ensambles *de novo*.
Palabras clave: Repetibilidad RNA-Seq; Microarreglos; HMM; calidad.

Vo. Bo. Co-Directores de Tesis

Dr. Eduardo Romero Vivas

Dr. Joaquín Gutiérrez Jagüey

Summary

RNA sequence content is deciphered by random fragmentation of biomolecules (RNA-Seq), which generates millions of reads. In lack of references, these reads are reconstructed relying on computational algorithms by *de novo* assembly. Multiple biological and technical factors induce errors on assembly. In the *de novo* context, it is even more critical to establish quality criteria for assembly selection because quantitative metrics have not shown correspondence to test references. Nonetheless, there are sources of genetic expression information that could be exploited. This research study proposes to use this evidence for quality evaluation and assembly selection. Microarray, UniGene and RNA-Seq data from three model organisms, Mouse *Mus musculus*, Fruit Fly *Drosophila melanogaster* and Water Flea *Daphnia pulex*, and one non-model organism Whiteleg Shrimp *Litopenaeus vannamei* was used to accomplish this investigation. Results were verified by means of mapping assemblies to transcriptome references and the UniProt/Swiss-Prot protein data base. First in *D. melanogaster*, *D. pulex* and *L. vannamei*, assembly variability was analyzed in terms of the content of contigs obtained by assembling multiple times a read set under identical conditions. This allowed the analysis of the influence of computational resources on assembly. Then, *M. musculus* and *D. melanogaster* assemblies were generated through parameter variation, permitting the evaluation of conventional quality metrics and mappings of transcriptome references. Once the highest-quality assemblies were identified, the proposed strategy of microarray-probe usage as a direct selection criterion was evaluated. This experimental evidence was generalized by means of Hidden Markovian Models (HMM). The models enabled assembly evaluation and selection through a probabilistic criterion. The Microarray-based HMM criterion was applied in *D. melanogaster*, and the generalizations were extended to *L. vannamei* assembly selection through UniGene-Based models. In silico, quantitative variation from five assembly repetitions originated from the same initial conditions were measured in less than 0.001%; however, qualitative variation measurements were up to 22%. Moreover, compared to large-memory computing platforms (128 GB), low-memory platforms (~24 GB) generated up to 7 times more variability, revealing the importance of the selection of computing equipment for assembly. In contrast to quantitative metrics, the proposed evaluation strategies, based on microarrays and HMMs usage, did identify the assemblies with the highest reference mappings. Thus, multiple assembly strategies are proposed to take advantage of variability for contig prospecting; and, for the usage of experimental evidence for quality evaluation and selection of *de novo* assemblies.

Key words: RNA-Seq Repeatability; Microarrays; HMM; quality.

Vo. Bo. Thesis Co-Directors

Dr. Eduardo Romero Vivas

Dr. Joaquín Gutiérrez Jagüey

Dedicatoria

A José, Amparo, Maribel, Marco, Hannia y Keira: mi familia.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología por la beca otorgada (No. 256634) para la realización de mis estudios de posgrado.

Al Centro de Investigaciones Biológicas del Noroeste, S.C., por las facilidades recibidas para mi formación doctoral.

Al Grupo de Investigación en Bioinformática, por la oportunidad brindada para la realización de este proyecto.

A mi comité tutorial, Dr. Eduardo Romero Vivas, Dr. Joaquín Gutiérrez Jagüey, Dr. Fernando Daniel Von Borstel Luna, Dr. Humberto Mejía Ruiz y Dra. Gabriella Rustici, por su inmensa dedicación y paciencia.

A los asesores e instructores que contribuyeron a mi proyecto y formación doctoral, en especial a la Dra. Amada Torres Salazar.

A la Dra. Norma Yolanda Hernández Saavedra, Lic. Osvelia Ibarra Morales y Tania Verónica Núñez Valdez de la Dirección de Estudios de Posgrado y Formación de Recursos Humanos.

A la Lic. Leticia González Rubio Rivera del Departamento de Becas y Apoyo Estudiantil y a la Lic. Claudia Elizabeth Olachea León del Departamento de Formación de Recursos Humanos y Educación Continua.

Al Ing. Horacio Sandoval Gómez, del Laboratorio de Cómputo de Posgrado, por el apoyo y soporte técnico en diversos eventos académicos.

Al la Lic. Ana María Talamantes Cota, Susana Luna García, Lic. María Esther Ojeda Castro y Elizabeth Guadalupe Sánchez Vázquez del área de Biblioteca.

A los integrantes del departamento de Redes, Ing. Jorge Mario Rodríguez Meza, Ing. Roberto González Castellanos, Ing. Luis Carlos Moreno Galván, y Lic. María Isabel Castro Hernández.

A la Lic. Cinthya Castro Iglesias, Lic. Daniela Núñez García y la Lic. Guillermina Verdugo Apodaca por su apoyo para la divulgación de este proyecto.

Al Dr. Miguel Córdoba y la Dra. Claudia Ivette Maytorena Verdugo del CIBNOR, y al Dr. Javier Romero Vivas y el Dr. Raymond Wolfe de la Universidad de Cork por su apoyo en la revisión de diversos escritos.

Al Laboratorio Nacional de Supercómputo del Sureste de México (LNS), perteneciente al padrón de laboratorios nacionales CONACYT, por los recursos computacionales, el apoyo, y la asistencia técnica brindados, a través del proyecto No. O-2016/028.

Al personal de soporte técnico de la empresa *Penguin Computing* por su apoyo, en especial a Massimo Malagoli, Terry L. Smith y Xian Su.

A los compañeros, amigos y agregados que recorrieron junto a mí el camino del doctorado; Ana Ruth Álvarez, Aldo Valadez, Casandra Vera, Carlos Romo, Claudia Maytorena, Diana Martínez, Elvia Pérez, Esteban Velázquez, Federico Soto, Fernando Pio, Teresa Sandoval, Lilian Arzeta, Mirella Romero y Zuami Villagrán.

A todos los que cerca o a distancia me han apoyado; Aleida Peláez, Alejandro Rivera, Benjamín Pacheco, Bianca Colio, Berenice Ponce, Carla Magdaleno, Carlos Argueta, Edgar Esquivel, Gracia Gómez, Gustavo Diosdado, Gustavo Woo, Juan García, Liza Reyes, la familia Madrid Herrera, Patricia Cortez, Rocío Navarro, Rafael Villa, Rosa Escobedo y Ulises Castro.

A mi incondicional familia...

¡GRACIAS!

Contenido

Lista de figuras

Lista de tablas

Abreviaturas

BLAST – Herramienta básica de búsqueda de alineamiento local (*Basic Local Alignment Search Tool*)

BLASTX – Algoritmo BLAST para alineamiento de secuencias proteicas a partir de secuencias nucleotídicas

DBG – Grafo De Bruijn

EST - Marcadores de secuencia expresada (*Expressed Sequence Tags)*

GB – GigaByte (unidad de memoria)

HPC – Cómputo de alto rendimiento (*High Performance Computing*)

HMM – Modelo Markoviano Oculto (*Hidden Markov Model*)

*k*21, *k*25 y *k*31 – Ensambles procesados con longitudes de *k*-mero = 21, 25 y 31

LNS – Laboratorio Nacional de Supercómputo del Sureste de México

MIAME – Información mínima sobre un experimento de microarreglos (*Minimum Information About a Microarray Experiment*)

NA – No aplicable

NGS – Secuenciación de próxima generación (*Next Generation Sequencing*)

PCR – Reacción en cadena de la polimerasa (*Polymerase Chain Reaction*)

PE – Lecturas pareadas (*Paired-end*)

POD – *Penguin on demand* (Servicio de cómputo HPC en la nube)

RAM – Memoria de acceso aleatorio (*Random Access Memory*)

RNA-Seq – Secuenciación de transcriptoma por medio de técnicas de secuenciación de próxima generación. En ocasiones este término también se emplea para hacer alusión a los flujos de trabajo por los que pasa un conjunto de lecturas de secuenciación NGS de transcriptoma

Sonda MH – Sonda con hibridación mutua

1. INTRODUCCIÓN

A nivel biológico-molecular, el escrutinio del contenido de ácidos nucleicos ha permitido explorar interacciones de estas biomoléculas con respecto a susceptibilidad y la resistencia a enfermedades, estructuras genéticas, relaciones poblacionales y patrones evolutivos, entre otras (Atanur *et al.*, 2013).

Uno de los enfoques biológico-moleculares más abordados es el análisis de transcriptomas por medio de la técnica de secuenciación de próxima generación de transcriptoma (RNA-Seq), cuyo principal objetivo es el identificar y descubrir transcritos, detectar su nivel de expresión y su relación con respecto a estadios específicos en los tejidos de estudio. Este enfoque abarca diversos flujos de trabajo[1] que comprenden desde el diseño experimental, el desciframiento del contenido (secuenciación) y reconstrucción de las secuencias de los transcriptomas a partir de los datos de secuenciación (ensamblaje), la asignación de funciones, ya sea por homologías o modelos (anotación), hasta el análisis de expresión de transcritos (abundancia) y la detección de variantes, entre otros (Conesa *et al.*, 2016).

La secuenciación requiere de la fragmentación de la muestra. Dichos fragmentos pasan por complejas interacciones de reacciones químicas y enzimáticas, que son finalmente analizados con *hardware* y *software* especializado. Según el *hardware* utilizado, la plataforma de secuenciación escanea estas reacciones con sistemas ópticos de alta resolución o sistemas de detección de iones de hidrógeno; por ejemplo las plataformas Roche 454 (Margulies *et al.*, 2005), Illumina-Solexa (Bentley *et al.*, 2008) e Ion Torrent/Proton PGM (Rothberg *et al.*, 2011). La salida de datos de dichas plataformas consta de conjuntos de millones de lecturas cortas con los valores de contenido y secuencias de nucleótidos adenina, timina, citosina y guanina (A, T, C y G respectivamente). La reconstrucción de datos de secuenciación se puede llevar a cabo por medio de algoritmos que toman como guía ensambles previos de alta calidad (referencias) o basándose exclusivamente

---

[1] Flujo de trabajo se refiere a la transformación sucesiva de archivos seguida en un análisis bioinformático (Leipzig, 2017).

en métodos matemáticos. La primera estrategia es conocida como ensamblaje por referencia, mientras que la reconstrucción en ausencia de una referencia es llamada ensamblaje *de novo*. Uno de los algoritmos matemáticos más utilizados para el ensamblaje *de novo* son los grafos De Bruijn (Compeau *et al.*, 2011).

Un proceso de ensamblaje tiene como principal objetivo obtener las secuencias de los transcritos originales. Sin embargo, diversos factores técnicos y características propias de los transcritos originan errores de ensamblaje (Miller *et al.*, 2010). Aunado a estos errores, las métricas[2] de calidad disponibles no corresponden con la calidad de los transcriptomas originales (Steijger *et al.*, 2013).

Dada la importancia del análisis de transcriptomas por medio del RNA-Seq, se requiere de algoritmos y metodologías que permitan la reducción de errores en el ensamblaje *de novo* de transcriptoma, a través de establecer condiciones favorables para aprovechar los datos y conocimientos existentes sobre el transcriptoma o para evaluar el resultado de dichas metodologías.

Técnicas anteriores al RNA-Seq, como microarreglos de expresión génica permiten la identificación a menor escala de transcritos por medio de sondas de muestreo prediseñadas con base en transcritos conocidos (Stekel, 2003). En la actualidad, el RNA-Seq se prioriza en el análisis de transcriptoma debido a su capacidad para descubrir nuevos transcritos. Sin embargo, ambas técnicas, microarreglos y secuenciación, reflejan el mismo conjunto de transcritos cuando son utilizadas para el análisis de transcriptomas de organismos y tejidos en condiciones similares. Cabe mencionar que millones de bases de datos de microarreglos se encuentran disponibles en repositorios públicos y estas constituyen fuentes de datos confiables y poseen el potencial de ser utilizadas en el desarrollo y evaluación de nuevos métodos de análisis (Rung y Brazma, 2013).

Así mismo, existen bases de datos de alta calidad, ya sea de secuencias o modelos, que son potenciales auxiliares en el análisis de transcriptoma dado que provienen de evidencia experimental o modelada de expresión génica o proteica,

---

[2] Métrica: Valor numérico o nominal asignado a características o atributos de un ente computado a partir de un conjunto de datos observables (Olsina, 1999).

tal es el caso de UniGenes, ESTs, familias PFam, etcétera (Finn *et al.*, 2016; Marinković *et al.*, 2012; Pontius *et al.*, 2002).

En este trabajo de investigación se propusieron flujos de trabajo y directivas favorables para la evaluación de calidad, uniformidad y prospección en ensamblaje *de novo* de transcriptoma utilizando como sustento datos de evidencia experimental de expresión génica o la variabilidad del mismo proceso de ensamblaje, contribuyendo así a la mejora del análisis de datos RNA-Seq por medio de ensamblaje *de novo* de transcriptoma.

Como etapa inicial del proyecto se caracterizó el desempeño del proceso de ensamblaje *de novo* de transcriptoma en términos de la respuesta del ensamblador y del equipo de cómputo.

Se generaron conjuntos de ensambles en condiciones iniciales idénticas, pero en diversos equipos de cómputo, para mosca de la fruta, pulga de agua y camarón blanco. De esta forma se logró cuantificar repetibilidad y variabilidad en ensamblaje.

Posteriormente, se generaron conjuntos de ensambles de ratón, mosca de la fruta y camarón blanco aplicando condiciones iniciales distintas, pero en el mismo equipo. En este caso se varió la longitud de *k*-mero, que está ligada a la sensibilidad y especificidad del algoritmo de ensamblaje.

Una vez realizados los conjuntos iniciales de ensambles, se detectó la calidad de éstos por medio de distintas metodologías basadas en mapeos de referencias. De esta manera se identificó la semejanza de los conjuntos de *contigs*[3] con respecto a transcriptomas de referencia de las especies estudiadas, y en casos especiales, a una base de datos de alta calidad de proteínas.

Los grupos de ensambles generados por medio de la variación de *k*-mero, en conjunto con las evaluaciones por mapeos de transcriptomas de referencia, que es considerado como el estándar de evaluación de calidad en ensamble *de novo* de transcriptoma, dieron pie al análisis del desempeño de métricas tradicionales.

---

[3] *Contig*: “Secuencia contigua”, el término se refiere a una secuencia generada en el proceso de ensamblaje (Brown, 2013b).

Posteriormente se propuso un criterio basado en el uso directo de evidencia experimental, sondas de microarreglos de expresión génica, para la evaluación de calidad y selección de ensamble *de novo* de transcriptoma dentro de la estrategia de ensamblaje múltiple. El criterio se empleó en ratón y mosca de la fruta.

Subsecuentemente, se realizó la generalización de las sondas de microarreglos de mosca de la fruta por medio de Modelos Markovianos Ocultos (*Hidden Markov Models*, HMM), y los modelos se emplearon para realizar la evaluación y selección de ensamble.

Finalmente, se extendió la generalización al uso de UniGenes para entrenamiento de modelos. Dichos modelos fueron empleados para evaluar ensambles de camarón blanco, que es una especie de alto interés comercial.

2. ANTECEDENTES

El término “transcriptoma” se define como “el conjunto completo de genes expresados bajo condiciones específicas, conteniendo los ácidos ribonucleicos (ARN) codificantes y no codificantes presentes en un tipo particular de célula, tejido u organismo” (Krebs *et al.*, 2014). Los ARNs del transcriptoma (transcritos) provienen de una serie de procesos celulares, como la transcripción de segmentos a partir de ácido desoxirribonucleico (ADN) y su posterior corte y empalme génico (*splicing*). Posteriormente, dichos transcritos cumplirán con diversas funciones, ya sea como moléculas informativas intermediarias en la generación de moléculas de trabajo (proteínas), u otras funciones celulares de similar importancia como la regulación de la misma transcripción.

2.1 Análisis de transcriptoma

A diferencia del genoma, cuyo contenido y secuencia es idéntica para todo tipo de células en un organismo, el transcriptoma varía según el tejido y la condición de estudio. Así entonces, estudios transcriptómicos comparativos han permitido el descubrimiento de biomoléculas relacionadas con condiciones de control y experimentales específicas (Conesa *et al.*, 2016).

El estudio del contenido transcriptómico de un sistema biológico, ya sea cultivos celulares, tejidos u organismos completos en distintas condiciones, ha sido abordado por medio de diversas técnicas. Estas varían desde las enfocadas en el estudio de uno o un conjunto pequeño de ARNs, como las técnicas de PCR (*Polimerase Chain Reaction*) o RACE (*Rapid amplification of cDNA ends*), hasta las técnicas que exploran el transcriptoma completo, como microarreglos o secuenciación.

2.1.1 Microarreglos de expresión génica

La técnica de microarreglos de expresión génica surgió a mediados de los 90’s (Schena *et al.*, 1995), ha sido utilizada en cientos de miles de proyectos y generado millones de datos para cientos de organismos modelo y no modelo (Rustici *et al.*, 2013). Los microarreglos representan la expresión de miles de genes de manera

paralela. Éstos consisten en un conjunto de sondas de exploración fijas a una base sólida, que después de ser sometidos a un experimento de hibridación de muestras con marcajes (fluoróforos) y posteriormente procesados en lectores láser, proveen indicaciones lumínicas de los transcritos presentes en la muestra (Stekel, 2003). Dichas señales lumínicas provienen de los transcritos marcados, que al hacer hibridación Watson-Creek con las sondas del microarreglo, emiten luz al ser excitados por un láser, evidenciando la presencia de las moléculas en la muestra. Las sondas provienen del análisis y selección de secuencias de alta calidad y son exclusivas a las secuencias de los transcritos de interés.

El experimento típico de microarreglos consiste en hibridar distintos paneles con muestras en condiciones control y experimental y realizar la lectura de éstos. Subsecuentemente, por medio de la comparación de dichas lecturas, se logra la medición de expresión de los transcritos presentes en una condición con respecto a la otra. Debido a la presencia de miles de transcritos, al procesamiento de imágenes implícito y las distintas réplicas técnicas y biológicas utilizadas comúnmente en experimentos de microarreglos, esta técnica genera una gran cantidad de datos, cuyo manejo e interpretación requirieron optimización de procesos tanto de análisis como computacionales (Romero-Vivas *et al.*, 2013). Con el tiempo, los protocolos de microarreglos para hibridación, análisis y almacenamiento de datos se han regularizado y estandarizado (Brazma, 2009). Adicionalmente, el costo de esta técnica ha disminuido considerablemente haciéndola accesible. Sin embargo, una de las limitantes de los microarreglos es su inhabilidad de descubrir nuevos transcritos. Mientras que la secuenciación de próxima generación sí permite el descubrimiento de nuevas secuencias.

2.1.2 Secuenciación

Las técnicas de secuenciación consisten en descifrar la información del contenido de ácidos nucleicos en una muestra de ADN. Éstas iniciaron en los 70´s con complejos métodos manuales que determinaban pocas decenas de bases (Gilbert y Maxam, 1973; Min Jou *et al.*, 1972; Wu y Taylor, 1971). Técnicas posteriores

extenderían la cantidad de nucleótidos que podían ser secuenciados (Maxam y Gilbert, 1977; Sanger y Coulson, 1975). En 1977 Sanger *et al.* (1977) sentaron las bases de la secuenciación moderna y automatizada (Chial, 2008), que fue esencial en el Proyecto del Genoma Humano (International Human Genome Sequencing Consortium, 2004). La secuenciación por electroforesis capilar, como también se le conoce a la secuenciación Sanger automatizada, provee secuencias largas (500-600 pares base, o pb) de alta calidad (Smith *et al.*, 1986), pero requiere altas cantidades de muestra, gran cantidad de reactivos, extensos tiempos de procesamiento y es de alto costo.

A mediados de la década de los 2000 surgen las plataformas de secuenciación de próxima generación (*Next Generation Sequencing*, NGS). Éstas presentan salidas masivas, procesos completamente automatizados, bajo requerimiento de muestra, ciclos de procesamiento relativamente cortos y bajos costos (Mardis, 2008). Existen varias plataformas para realizar NGS, siendo las plataformas Illumina una de las más utilizadas al momento. La preparación de una librería de secuenciación comienza en la fragmentación de la muestra. Para librerías que son procesadas en plataformas Illumina, estos fragmentos son ligados a adaptadores y subsecuentemente añadidos a una celda de flujo, donde hibridan su superficie. A continuación, se lleva a cabo un proceso de PCR de puente por cada fragmento, creando grupos de fragmentos con múltiples copias. Posteriormente la celda es ingresada a la plataforma de secuenciación, donde se añadirán nucleótidos con etiquetas fluorescentes (que emiten distintas longitudes de onda para cada nucleótido A, T, C, G) y estos se irán incorporando de uno en uno a los fragmentos, siendo capturada por medio de sistemas ópticos la luz que emiten en su incorporación sucesiva. Las imágenes sucesivas son procesadas y su resultado final es un archivo de texto que contiene millones de lecturas cortas con los valores de contenido y secuencias de nucleótidos A, T, C y G (Bentley *et al.*, 2008).

Independientemente de la plataforma utilizada, el proceso de secuenciación de ácidos nucleicos requiere la fragmentación de éstos debido a las limitantes del

equipo. En general, las plataformas NGS proveen un conjunto de millones de secuencias cortas (35-400 pb) con poco traslape entre sí.

La NGS de transcriptoma es conocida como RNA-Seq. Los pasos típicos de un experimento de RNA-Seq son: extracción de la muestra, remoción de contaminantes (ADN y muy frecuentemente de ARN ribosómico), fragmentación de los ARN obtenidos, el cambio de éstos fragmentos a ADN complementario por medio de transcripción inversa, la adhesión de adaptadores y la selección de tamaños de fragmento (Martin y Wang, 2011). Finalmente, la secuenciación es llevada a cabo produciendo millones de lecturas con el contenido y secuencia de millones de fragmentos, que en procedimientos computacionales posteriores deben ser reconstruidas para inferir las secuencias de las biomoléculas originales. Este proceso de reconstrucción es conocido como “ensamblaje”.

A diferencia de la secuenciación en genomas, donde después de secuenciar se pretende reconstruir una sola secuencia grande uniforme, el ensamblaje de datos de una biblioteca de datos RNA-Seq apunta a obtener miles de secuencias que pertenecen a los miles de ARNs del transcriptoma y que éstos varían en abundancia (Korf, 2013).

2.2 Ensamblaje de bibliotecas NGS

Un proceso de ensamblaje puede ser realizado alineando las lecturas RNA-Seq a un ensamble previo de alta calidad (referencia), o *de novo*, donde a falta de una referencia, la reconstrucción de secuencias recae exclusivamente en algoritmos matemáticos y computacionales (Moreton *et al.*, 2016).

Dadas las características de las lecturas de las bibliotecas NGS, se requirió de un algoritmo de ensamblaje capaz de manejar millones de secuencias de una manera eficiente. Así, se encontró en los grafos De Bruijn (DBG) una opción adecuada para el ensamblaje *de novo* (Miller *et al.*, 2010).

Un DBG consiste en la descomposición de los datos de entrada en unidades fundamentales de reconstrucción de longitud fija k (k-meros), ordenar estas unidades y separarlas en segmentos iniciales y finales de longitud k-1, asignar un

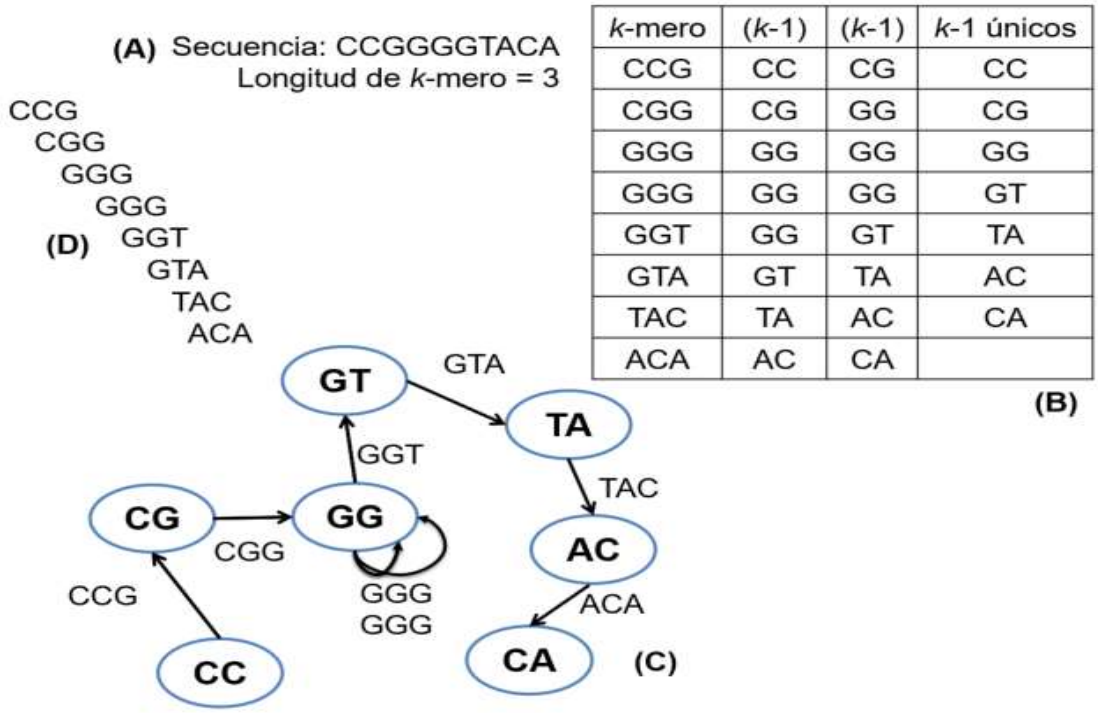
nodo en el grafo para cada segmento único *k*-1 y finalmente, trazar una trayectoria (conocida como trayectoria Euleriana) por medio de aristas que unirán la secuencia de un nodo con respecto al otro. Al unir por medio de los aristas los nodos, el DBG provee la secuencia objetivo (Compeau *et al.*, 2011) como se observa en la figura 1. Cabe mencionar que la longitud de *k*-mero está directamente relacionada con sensibilidad y especificidad de ensamblaje, al influir directamente sobre el DBG (Duan *et al.*, 2012; Zerbino y Birney, 2008).

La implementación de ensamblaje *de novo* de transcriptoma es una tarea algorítmica y computacionalmente compleja. En términos generales un ensamblador obtiene un catálogo de *k*-meros a partir de los datos de entrada, genera grafos De Bruijn con base en este catálogo para finalmente resolver ambigüedades en los DBGs y crear el conjunto final de *contigs* del ensamble (Moreton *et al.*, 2016). El ensamblador realiza agrupamientos iniciales de datos de entrada (*clustering*) para posteriormente generar DBGs y con base en estos construir *contigs*. Esto genera la misma secuencia cíclica de la secuencia original optimizando recursos computacionales.

El *software Trinity Transcriptome Assembler* es uno de los ensambladores de última generación más utilizado en estudios transcriptómicos debido a que ha sido reportado como una de las mejores opciones para ensambladores de *k*-mero sencillo (Zhao *et al.*, 2011), es parte de flujos de trabajo más complejos que comprenden desde preprocesamiento hasta anotación funcional, tiene un fuerte equipo de soporte, es de libre acceso y código abierto, permitiendo así modificaciones en su funcionamiento. Este ensamblador realiza su flujo de trabajo por medio de tres módulos: *Inchworm*, donde realiza la agrupación inicial de lecturas pertenecientes al mismo gen (*clustering*) y una construcción extendida de secuencias con base en dichos *clusters*; *Chrysalis*: que es el módulo donde construye los grafos De Bruijn con base en los *clusters* de lecturas y los *contigs* extendidos; para finalmente pasar al módulo *Butterfly*, donde resuelve ambigüedades en los grafos con base en la cantidad de lecturas que respaldan la trayectoria de análisis (Haas *et al.*, 2013b). El módulo *Inchworm* es el primero en

ser ejecutado. Posteriormente, los módulos *Chrysalis y Butterfly* se ejecutan de forma alterna (Ver Anexo A para detalles de los módulos y la salida del proceso).

Los procesos de ensamblaje *de novo* de transcriptoma demandan el uso de cómputo intensivo. La cantidad de lecturas de entrada y los requerimientos de corrida del *software* dan una aproximación inicial de los recursos computacionales esenciales para llevar a cabo un ensamblaje *de novo*. Según instrucciones del ensamblador Trinity, estos requerimientos son: dos o más núcleos de procesamiento y ~1GB de memoria RAM por cada millón de lecturas de entrada (Haas *et al.*, 2013b).



k-mero	(*k*-1)	(*k*-1)	*k*-1 únicos
CCG	CC	CG	CC
CGG	CG	GG	CG
GGG	GG	GG	GG
GGG	GG	GG	GT
GGT	GG	GT	TA
GTA	GT	TA	AC
TAC	TA	AC	CA
ACA	AC	CA	

Figura 1. Grafo De Bruijn. A) Secuencia de ejemplo. B) Creación del catálogo de *k*-meros de *k* = 3. C) Ensamblador de datos NGS: construye un DBG por medio de la representación de todos los prefijos y sufijos de los *k*-meros como nodos; posteriormente se dibujan aristas que representan a los *k*-meros con determinado prefijo y sufijo. Por ejemplo, el *k*-mero CCG tiene el prefijo CC y el sufijo CG. D) El encontrar el ciclo Euleriano permite que se reconstruya la secuencia por medio de la formación de alineamientos en el cual cada *k*-mero sucesivo (provenientes de aristas sucesivas) es recorrido por una sola posición. Esto genera la misma secuencia original optimizando recursos computacionales. Modificado de Compeau *et al.* (2011).

2.2.1 Errores y variabilidad en ensamblaje *de novo* de transcriptoma

En condiciones ideales el algoritmo de ensamblaje recae en datos libres de errores, pero en la realidad esto no sucede. Problemas con la preparación y manejo de librerías y problemas técnicos intrínsecos a las plataformas de secuenciación, aunados a características propias de los transcritos inducen errores y ambigüedades en la construcción de DBGs y en la determinación de sus trayectorias (Miller *et al.*, 2010).

Errores de secuenciación y determinaciones incorrectas de base (*base calling*) en los extremos de las lecturas, así como polimorfismos cercanos a una repetición causan ramificaciones en el DBG. Errores de secuenciación en la mitad de las lecturas y polimorfismos dentro de estas originan burbujas en el grafo. Repeticiones en los transcritos originales causan convergencias y bifurcaciones en una trayectoria, llamados patrones *frayed rope* (Fig. 2). Además, características propias del transcriptoma se suman a estas vicisitudes, tales como miles de ARNs y transcritos con ligeras variantes debido al corte y empalme génico (isoformas) con distintos niveles de expresión.

Los problemas técnicos y características biológicas descritas hacen del proceso de ensamblaje una tarea compleja, en especial en el caso *de novo*, ya que no se cuenta con una referencia con la cual se pudiese guiar el ensamblaje. Se han tomado diversas medidas y estrategias para abordar estos retos y contrarrestar potenciales errores. Estas medidas varían desde metodologías aplicadas dentro del algoritmo de ensamblaje hasta estrategias posteriores al proceso.

Dentro de las metodologías intrínsecas al ensamblador, se emplean diversos umbrales estáticos y dinámicos aplicados a la resolución de trayectorias. Por ejemplo, el ensamblador *Trinity* recurre a los datos de entrada en su etapa de resolución de ambigüedades; así, en el caso de encontrarse con una bifurcación, el ensamblador cuantifica la cantidad de lecturas que alinean en ambas trayectorias resolviendo eliminar la trayectoria en la cual se hayan alineado menos lecturas (Grabherr *et al.*, 2011).

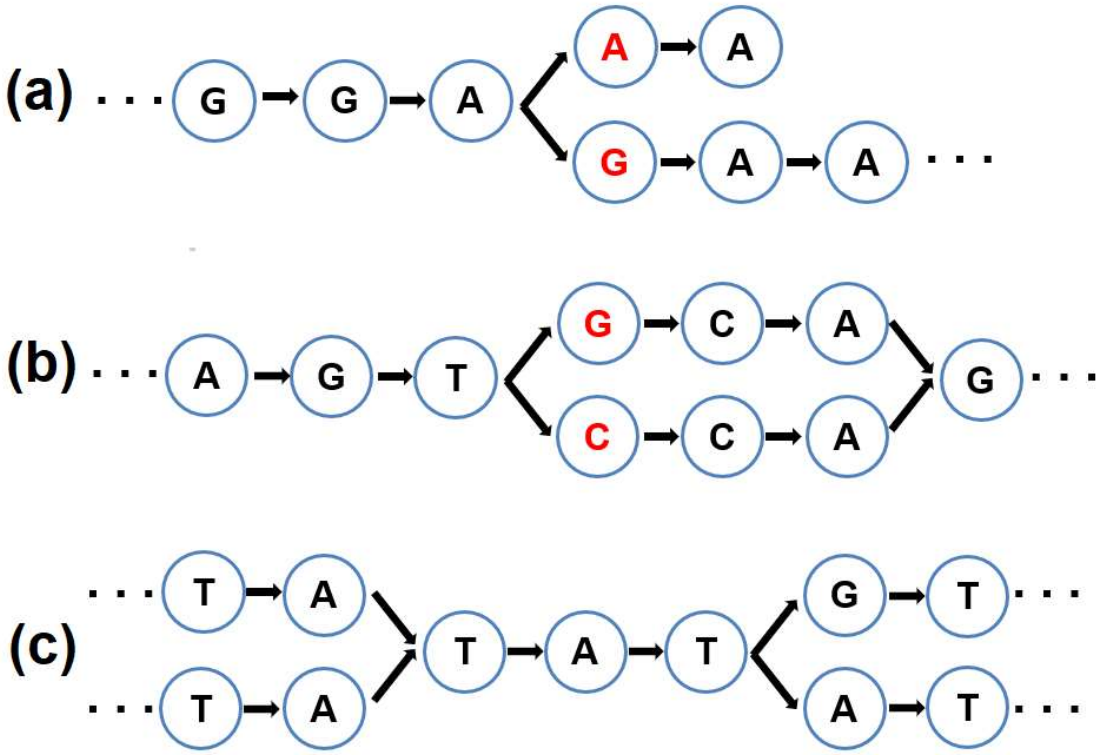


Figura 2. Tipos de errores de ensamblaje. A) Errores en los extremos de las lecturas y/o polimorfismos cercanos a una repetición causan ramificaciones. B) Errores de secuenciación en la mitad de las lecturas y/o polimorfismos dentro de estas originan burbujas. C) Repeticiones en los transcritos originales causan convergencias y bifurcaciones en una trayectoria (patrones frayed rope). Modificado de (Miller *et al.*, 2010).

Se puede decir que no existe un ensamble único para cada conjunto de lecturas NGS. Factores técnicos y biológicos como los descritos en esta sección tienen efecto en el ensamblaje *de novo* de transcriptoma y han sido previamente analizados (Bradnam *et al.*, 2013; Earl *et al.*, 2011; Salzberg *et al.*, 2012). Cabe destacar que la influencia tiene la disposición y asignación de recursos computacionales sobre el ensamblaje no se ha explorado a detalle. Los reportes actuales sobre ensamblaje *de novo* típicamente muestran los efectos que la cantidad de memoria y núcleos tienen sobre los tiempos de procesamiento o sobre

la viabilidad de la plataforma para la ejecución del proceso (Henschel *et al.*, 2012; Lin *et al.*, 2011; Zhao *et al.*, 2011). No obstante, apenas se tienen precedentes de la existencia de una ligera variación en la salida del ensamblaje a causa de correr el proceso en distintas ocasiones o en *hardware* diferente (Haas *et al.*, 2013a). Asimismo, se ha reportado que existe aleatoriedad en los resultados (de ensamblaje) debido al uso de *multi-threading* en combinación con la utilización de estructuras probabilísticas de datos (Quiagen, 2014).

2.2.2 Estrategia de ensamblaje múltiple

Una de las estrategias más utilizadas es la creación de diversos ensambles variando las condiciones iniciales del proceso (Schulz *et al.*, 2012; Zerbino y Birney, 2008). Posteriormente, se crea un consenso de todas las reconstrucciones o se elige un ensamble con base en criterios definidos por el usuario (Clarke *et al.*, 2013). Ya sea por medio de intersección y/o unión de *contigs* en condiciones idénticas o cambiando condiciones de inicio, las variaciones dan paso a la combinación de múltiples ensambles. Un ejemplo común de variación en condiciones iniciales es construir un conjunto de *contigs* a partir de múltiples ensambles procesados con distintas longitudes de *k*-mero. Dicha estrategia es reportada como beneficiosa dado que la variación en *k*-mero captura de mejor manera los transcritos con niveles distintos de expresión; *k*-meros menores tienden a identificar transcritos con bajos nivel de expresión y viceversa (Surget-Groba y Montoya-Burgos, 2010). Sin embargo, este tipo de estrategias de consolidación tienden a la acumulación de errores en el consenso final derivado de los errores los ensambles individuales (Durai y Schulz, 2016). Cabe mencionar que en estudios de evaluación esta estrategia presentó niveles menores de alineamiento a genomas de referencia comparándola con las estrategias de *k*-mero sencillo (Zhao *et al.*, 2011).

Un caso distinto son las estrategias de ensamblaje múltiple que no consolidan las reconstrucciones; en esta estrategia se exploran los múltiples ensambles y se selecciona la mejor opción posible. Dicha selección se basa en criterios definidos por el usuario; consecuentemente dichos criterios deben ser minuciosos. Sólo un

criterio ha sido reportado como factible para esta selección, la métrica N50 (ver sección 2.3 para definición de la métrica) (Zerbino y Birney, 2008). Esta métrica es considerada como una medida de contigüidad de ensamble, por lo tanto, como un indicador de fragmentación. Aun así, la métrica N50 ha sido utilizada como criterio de selección, pero bajo la condicionante de tener conocimiento *a priori* de la longitud media de la métrica (Clarke *et al.*, 2013).

2.3 Calidad de ensamble *de novo* de transcriptoma

Independientemente de la estrategia de ensamblaje o *software* utilizado para esta tarea, el objetivo principal de un ensamblaje *de novo* es que la reconstrucción de la biblioteca NGS asemeje lo más posible a los transcritos originales de la muestra. De este modo, se utiliza el término calidad como indicador de la semejanza y validez de un ensamble.

Los criterios de evaluación de calidad más utilizados se basan en estadísticas y características cuantitativas del ensamble; la longitud de las secuencias reconstruidas (llamadas *contigs*) y ensamble completo, los promedios y porcentajes de utilización de los datos de entrada. En especial la métrica N50, que indica la longitud del *contig* donde el 50% del ensamble está representado. N50 es tomada como una medida de contigüidad de un ensamble; consecuentemente, como un indicador de la fragmentación de este. Sin embargo, se han encontrado inconsistencias entre éstas métricas y la calidad de ensamble al ser evaluadas con bases en transcriptomas de referencia (Steijger *et al.*, 2013). Más aún, las métricas cuantitativas fueron diseñadas para la evaluación de ensamble de genoma, no de transcriptoma cuyo principal objetivo es crear un conjunto de secuencias de longitudes heterogéneas (Vijay *et al.*, 2013). Haciendo mención de la estrategia múltiple de ensamblaje, éstas métricas están enfocadas en la evaluación de un solo ensamble, no en la selección de una reconstrucción dentro de un grupo de ensambles.

La determinación de calidad de un ensamble puede estar enfocada en estimar la similitud de la reconstrucción con respecto a transcriptomas de referencia

de la misma especie o especies cercanas (Duan *et al.*, 2012; Mundry *et al.*, 2012; O'Neil y Emrich, 2013). Este tipo de estrategias también comprende el evaluar calidad basándose en secuencias conservadas (Parra *et al.*, 2007; Simão *et al.*, 2015; Smith-unna *et al.*, 2015). Sin embargo, a la fecha pocas especies cuentan con referencias, menos de 173 según la versión 85 de repositorio Ensembl (Flicek *et al.*, 2014). Considérese también que la confiabilidad de estas estrategias depende de la distancia evolutiva del organismo de interés con respecto al organismo de referencia, lo cual origina que las métricas sean conservadoras. No obstante, se sigue explorando el uso de datos alternativos al transcriptoma para su inclusión en estrategias de ensamblaje *de novo*.

2.3.1 Referencias de transcriptoma

En términos de ensamblaje, una referencia es la mejor aproximación disponible a determinada fecha y revisión del genoma o transcriptoma de determinada especie (Earl *et al.*, 2011). Ya que las biomoléculas pertenecientes a un transcriptoma contienen ARNs codificantes y no codificantes, es común encontrar dos tipos de referencias (Flicek *et al.*, 2014):

- Referencia codificante: Secuencias provenientes de ARNs que serán traducidas a proteínas. A pesar de que los genes que generan dichos transcritos representar una pequeña parte del genoma en mamíferos con respecto a genes no codificantes (relación de ~40 a 1), y la mitad en organismos como mosca de la fruta (relación 2.4 a 1), estas secuencias han sido el foco de atención en estudios de composición de genomas y transcriptomas (Frith *et al.*, 2005; Shabalina y Spiridonov, 2004).
- Referencia no codificante: Secuencias provenientes de ARNs que no serán traducidas a proteínas. No obstante, dichas secuencias pudiesen intervenir de manera indirecta en su generación o intervienen en la regulación de expresión génica a nivel transcripcional y post-transcripcional, por ejemplo, ribosómicos, pequeños no condicionales, nucleares, nucleolares, micro ARNs y no codificantes largos, entre otros. Se tiene menos conocimiento sobre secuencias

no codificantes y solo una pequeña proporción de estas ha sido examinadas para comprobar su función; sin embargo, durante la última década los enfoques de investigación han cambiado al estudio de estas biomoléculas (Ponting *et al.*, 2009; Quek *et al.*, 2015; Shabalina y Spiridonov, 2004).

2.4 Datos y modelos auxiliares para ensamblaje *de novo* de transcriptoma

Se puede hacer uso de datos auxiliares para ensamble o de evaluación de calidad en ensamblaje *de novo* de transcriptoma, en especial aquellos datos que proveen evidencia de los transcritos presentes en condiciones de experimentación. Dichas fuentes de evidencia pueden provenir desde el uso de transcritos hasta modelos que generalizan patrones de expresión en un sistema. Estos datos y modelos pueden incluir:

2.4.1 Bases de datos de proteínas

Existen varias bases de datos de proteínas con distintos enfoques. Entre las más utilizadas y confiables esta la base de datos UniProtKB (Boutet *et al.*, 2016), la cual contiene bases de datos anotadas manualmente o automatizadas provenientes de distintos organismos:

- UniProt/Swiss-Prot: contiene secuencias extraídas de literatura con anotaciones manuales evaluadas por curadores. A la fecha esta base de datos contiene 556,568 secuencias.
- TrEMBL: Secuencias con anotación automática en espera de anotación manual. A la fecha esta base de datos contiene 107,627,435 secuencias.

En la práctica, se opta por ensamblaje *de novo* cuando no se cuenta con una referencia aplicable al transcriptoma de estudio, por lo tanto, las evaluaciones con base en bases de datos de proteínas constituyen una estrategia para organismos que no cuenten con transcriptomas de referencia, comúnmente no modelo.

Las bases de datos de proteínas constituyen una fuente confiable de datos para la evaluación de ensamblajes y se emplea comúnmente para la valoración de calidad (Huang *et al.*, 2016). Sin embargo, la comparación de un ensamble con

respecto a estas secuencias pudiese introducir cierto nivel de incertidumbre a la evaluación, ya que la búsqueda de similitud entre *contigs* (secuencias de nucleótidos) y proteínas (secuencias de aminoácidos) involucra la traducción de las primeras, lo cual conlleva que una sola secuencia nucleotídica sea interpretada de seis maneras distintas al traducirla a sus seis marcos de referencia (se pueden leer de seis formas distintas). Adicionalmente la evaluación pudiese ser sesgada por la representación del organismo de estudio en la base de datos de proteínas.

2.4.2 Secuencias EST

Los marcadores de secuencia expresada (*Expressed Sequence Tags*, EST) son secuencias cortas de ARN mensajero expresadas en secuencias de ADN complementario, menores a 1,000 pb, algunas codificantes otras no, que representan porciones de genes expresados y que provienen de secuenciación de clonas de bibliotecas de ADN complementario (ADNc). Sin embargo, estas secuencias suelen ser de baja calidad y representar principalmente transcripciones abundantes y altamente expresadas, lo que hace que los genes expresados débilmente estén menos representados en el conjunto de secuencias (Boguski *et al.*, 1993).

Los ESTs han sido utilizados con anterioridad en ensamblaje *de novo* de transcriptoma como auxiliares de ensamblaje (Velázquez-Lizarraga, 2016) y para la valoración de calidad de ensambles de camarón blanco (Ghaffari *et al.*, 2015).

2.4.3 UniGenes

Bases de datos de grupos (*clusters*) de secuencias tipo EST, RNA mensajero (RNAm) y secuencias codificantes. Cada conjunto provienen de procesamiento computacional utilizando distintos algoritmos y alineamientos de varias secuencias que presumiblemente provienen del mismo gen, de tal manera que cada UniGene representa un transcrito (Wheeler, 2003).

Los UniGenes no han sido utilizados como auxiliares de ensamble. Sin embargo, la organización sistematizada de datos de expresión por medio de estos *clusters* permite que estos sean ideales candidatos para la inclusión generalizada

de datos de evidencia de expresión como auxiliares de ensamblaje o evaluación de calidad, lo cual se ha explorado en esta tesis.

2.4.4 Microarreglos de expresión génica

Los microarreglos de expresión génica son una fuente de secuencias de evidencia de expresión de genes. Estos son una técnica consolidada con protocolos y algoritmos de análisis optimizados; consecuentemente estos constituyen una fuente confiable de datos de expresión (Rung y Brazma, 2013). Los distintos repositorios transcriptómicos contienen millones de bases de datos de microarreglos, más de 1.5 millones de bases de datos públicos que fueron obtenidos en diversas condiciones de aproximadamente 1,600 especies (Kolesnikov *et al.*, 2015). Un microarreglo cuenta con sondas de prueba y de control; las últimas tienen función de controles positivos o negativos. Las sondas están usualmente organizadas en conjuntos (*probe sets*), que cubren cada uno distintas regiones de un mismo transcrito.

Las aproximaciones por microarreglo y RNA-Seq son complementarias; se ha tomado ventaja de ésta característica en el pasado al utilizar lecturas RNA-Seq para el diseño de sondas de microarreglos (Marinković *et al.*, 2012). Sin embargo, los microarreglos no han sido utilizados para dar soporte a ensamble o evaluar su calidad. Por ello, en este proyecto de tesis se propone esta fuente de datos como auxiliares al ensamblaje *de novo* de transcriptoma. Esto se plantea con base en que la hibridación positiva de las sondas de microarreglo, provenientes de ensayos de expresión similares a la experimentación RNA-Seq, pueden validar y dar evidencia del estado de un transcriptoma en tiempos y condiciones específicas. Esto es debido a que ambas técnicas apuntan hacia los mismos transcritos.

En la actualidad se le da preferencia al RNA-Seq. Aun así, los datos de microarreglos representan potenciales fuentes de información auxiliar en el análisis de RNA-Seq.

2.5 Uso de datos de evidencia experimental para soporte en ensamblaje *de novo* de transcriptoma

Un conjunto de secuencias que proveen evidencia experimental del estado de un transcriptoma puede ser utilizado ya sea de manera directa o generalizada. En el primer caso, el uso directo de los datos se realiza por medio de la incorporación directa dentro o fuera del proceso de ensamblaje *de novo*, ya sea como soporte del proceso o como auxiliar de evaluación de calidad.

Es posible que un conjunto de datos experimentales no pueda ser incorporados de manera directa al ensamblaje o a su evaluación de calidad, ya sea porque no es compatible con el proceso o porque no contiene suficientes datos como para representar al sistema al cual se quiere incorporar. Consecuentemente, se recurre al uso de modelos matemáticos o estadísticos, que con formalizaciones matemáticas representarán dichos datos de una forma simplificada, se aproximarán al transcriptoma que generó los datos, y de esta forma, se podrán emplear en ensamblaje *de novo* para la toma de decisiones o para hacer evaluaciones y predicciones a partir de dichas aproximaciones.

Existen diversidad de modelos que pueden ser empleados para modelar secuencias de observaciones (correlaciones entre símbolos, dominios o eventos adyacentes), como Redes Neuronales Artificiales (Murvai *et al.*, 2001), Máquinas de Vectores de Soporte (Liu *et al.*, 2006), Modelos de Covariancia (Nawrocki y Eddy, 2013), etc. Pero entre todos, uno de los más destacados son los Modelos Markovianos Ocultos (Rabiner, 1989); (ver Anexo B para una breve descripción de éstos). Un Modelo Markoviano Oculto (*Hidden Markovian Model*, o HMM) es un modelo estadístico que puede ser usado para describir la evolución de eventos observables que dependen de factores internos, que no son directamente observables (Yoon, 2009). Los HMM han sido exitosamente empleados en bioinformática en aplicaciones como predicción de genes (Munch y Krogh, 2006), alineamiento múltiple o por pares (Durbin *et al.*, 1998; Pachter *et al.*, 2002), determinación de bases *(base-calling*) (Liang *et al.*, 2007), modelado de errores de secuenciación de DNA (Lottaz *et al.*, 2003), predicción de estructuras secundarias

de proteínas (Won *et al.*, 2007), identificación de RNA no codificante (Zhang *et al.*, 2006) y alineamientos estructurales de RNA (Yoon, 2009), entre otras.

2.6 Modelos Markovianos Ocultos para representar secuencias heterogéneas de ADN

Los HMM han sido descritos con anterioridad para representar secuencias heterogéneas de ADN y aplicados a secuencias de ADN mitocondrial de humanos, ratones y levaduras, para un fragmento del cromosoma humano X, así como para el genoma completo del bacteriófago lambda (Churchill, 1989). Se describe que, debido a su composición, los ácidos desoxirribonucleicos pueden ser representados por secuencias de bases de hebra sencilla escritas en sentido 5' a 3'. Adicionalmente a esta representación de cuatro letras (A, T, C, G), pudiese ser de interés considerar los ADNs como representaciones binarias, por ejemplo, las relaciones purina-pirimidina (AG-TC), enlaces de hidrogeno fuertes-débiles (GC-AT), o las clasificaciones ceto-amina (GT-AC).

Según la descripción del mismo autor (Churchill, 1989), también se asume que las secuencias de ADN tienen estructura de mosaico, es decir, que están compuestas por segmentos de composición homogénea pero que dichos segmentos difieren el uno del otro. Cada segmento puede ser clasificado en uno de un número finito de estados. Estos estados representan una estructura subyacente a la secuencia y se asume que evolucionan lentamente según un proceso Markoviano oculto. De tal manera que se cuenta con un conjunto discreto de estados y de salidas. La secuencia de estados representa una estructura subyacente, la cual no es directamente observable, pero puede ser inferida a partir de sus observaciones.

Aplicando esta estructura al caso de ADN, las observaciones de salida corresponden a la secuencia de bases. Se asume que los estados son fijos y finitos en número, y que corresponden a las diferentes regiones o segmentos del ADN.

El modelo de Churchill (1989) logra entonces, por medio estos conjuntos discretos de estados y salidas, y de una representación binaria del ADN, generalizar

por medio de un HMM secuencias heterogéneas de ADN. La topografía de dicho modelo consiste en HMMs de primer orden, ergódicos y de dos estados (Fig. 3). La emisión de los estados emula transiciones y transversiones en las secuencias por medio de la siguiente dinámica: Estado uno (S1), mayor probabilidad de emisión de las bases A y G; estado dos (S2) mayor probabilidad de emisión de las bases C y T; de tal manera que los cambios de estados $S1 \rightarrow S1$ y $S2 \rightarrow S2$ representarán sucesiones AG, GA, CT y TC (transiciones), y los cambios de estados $S1 \rightarrow S2$ y $S2 \rightarrow S1$ representarán a todas las demás sucesiones (transversiones).

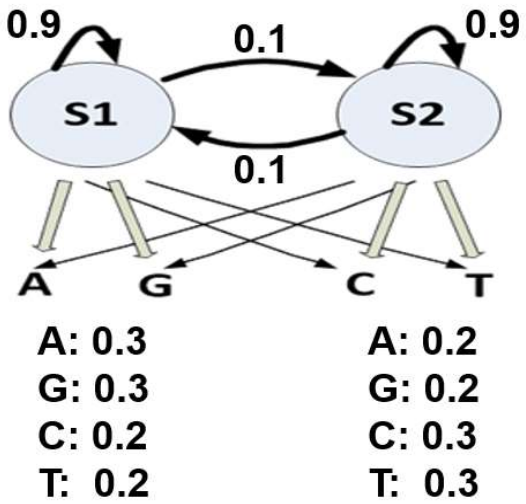


Figura 3. Modelo Markoviano Oculto para representar secuencias heterogéneas de ADN (Churchill, 1989). Transiciones de estados y sus probabilidades asociadas se indican por medio de flechas. Probabilidades de emisión de símbolos A, C, G y T para cada estado se indican debajo de ellos.

A la fecha, se sigue buscando disminuir la incertidumbre en los procesos, flujos de trabajo y evaluaciones de ensamblaje *de novo* de transcriptoma. Se encuentra entonces en la inclusión de secuencias de evidencia del estado de un transcriptoma, ya sea de manera directa o generalizada, opciones viables para fortalecer el análisis de datos de organismos que carecen de transcriptomas de referencia. Se propone, por tanto, usar microarreglos de expresión génica de forma directa o generalizada,

y UniGenes de manera generalizada, para la evaluación y selección del ensamble de mayor calidad en el marco de ensamblaje múltiple.

3. JUSTIFICACIÓN

Siendo la finalidad de un ensamblaje de transcriptoma el recrear de manera fiable las secuencias de los transcritos de una muestra, existe la necesidad de reducir la incertidumbre en este proceso para obtener reconstrucciones de mayor calidad. Este es un aspecto especialmente crítico en el caso del ensamblaje *de novo*, ya que no se cuenta con transcriptomas de referencia con los cuales se pudiese orientar el proceso o evaluar los resultados de éste, pues las métricas tradicionales de calidad han mostrado no tener correspondencia con el contenido original del transcriptoma; además, no existe un claro precedente de variabilidad en este proceso y de la influencia que tienen los equipos de cómputo sobre él.

Por tanto, se requiere contar con criterios confiables de calidad en el ensamblaje de especies que no cuenten con referencias de transcriptoma, en específico, ensamblaje *de novo*, este proyecto de tesis implementa la inclusión de datos que proporcionan evidencia experimental del estado del transcriptoma de una muestra, microarreglos y UniGenes, a diversas estrategias de evaluación de calidad y selección de ensamble con base en el uso directo y generalizado de dichos datos.

4. HIPÓTESIS

El uso de evidencia experimental dentro de un conjunto de ensambles *de novo* es un mejor indicador de calidad, en comparación con las métricas tradicionales, al identificar el ensamble que mejor mapea a secuencias de referencia.

5. OBJETIVOS

5.1 General

Evaluar el uso de evidencia experimental en comparación con las métricas tradicionales como indicativo del ensamble *de novo* de transcriptoma con mejor calidad dentro de un conjunto de reconstrucciones por medio del uso de secuencias de referencia.

5.2 Particulares

I. Determinar el grado de variabilidad en ensamblaje *de novo* de transcriptoma dadas condiciones iniciales idénticas (datos y parámetros) e identificar las condiciones que incrementen repetibilidad y variabilidad en conjuntos de ensambles.
II. Generar conjuntos de ensamblajes *de novo* de transcriptoma dadas distintas condiciones iniciales al cambiar los parámetros de ensamblaje.
III. Verificación de calidad de los ensambles *de novo* de transcriptoma por medio de secuencias de referencia, específicamente transcriptomas de referencia o bases de datos de proteínas.
IV. Análisis de desempeño de métricas tradicionales de calidad en comparación con métricas referenciadas.
V. Generar un criterio de calidad para ensamble *de novo* de transcriptoma, usando de manera directa sondas hibridadas de microarreglos de expresión génica para la selección de ensamble dentro de un conjunto de reconstrucciones.
VI. Evaluación de calidad de ensamble *de novo* de transcriptoma mediante la generalización de datos experimentales, sondas hibridadas de microarreglo y UniGenes, por medio de modelos estadísticos.

6. MATERIAL Y MÉTODOS

El ensamblaje *de novo* de transcriptoma se realiza por lo general en datos de especies que no cuentan con referencias. Teniendo la optimización en evaluación y selección de ensamble *de novo* de transcriptoma por medio de datos auxiliares como principal objetivo, toda propuesta y análisis requiere de verificación. Dadas estas condiciones, se utilizaron en su mayoría datos de organismos modelo, ya que éstos permiten las evaluaciones de metodologías dada su disponibilidad de transcriptomas referencias. Los organismos modelo seleccionados fueron ratón (*Mus musculus*), mosca de la fruta (*Drosophila melanogaster*) y pulga de agua (*Daphnia pulex*). Ya que los organismos no modelo son comúnmente especies de interés comercial, se incluyó en el proyecto de investigación una especie de interés del sector de acuicultura, camarón blanco (*Litopenaeus vannamei*), realizando evaluaciones de calidad por medio de bases de datos de proteínas. Cabe mencionar que las propuestas emitidas aplican a todo tipo de organismos y que todas las bases de datos utilizadas en el proyecto son de libre acceso.

Un análisis previo de calidad es común en lecturas RNA-Seq, y según sus resultados, éstas son preprocesadas para mejorar la calidad de los datos de entrada del ensamblador. Los análisis de calidad de lecturas usadas en el proyecto se realizaron con el *software* FastQC (Andrews, 2015) y los preprocesamientos con Trimmomatic (Bolger *et al.*, 2014). Los detalles de las estrategias de preprocesamiento, así como las fuentes y generalidades de las bibliotecas NGS de los distintos organismos involucrados en el estudio se describen en el Anexo C.

No existe un ensamble único para un conjunto de lecturas RNA-Seq. Hay múltiples condiciones de inicio que pueden producir distintos ensambles, llámese niveles de preprocesamiento, selección de *software* y configuración de parámetros de *software*, entre otras. Adicionalmente, correr el mismo proceso de ensamblaje bajo las mismas condiciones iniciales y en la misma plataforma de cómputo produce ensambles con variaciones (Haas *et al.*, 2013b). De esta manera, se exploraron dos fuentes de variación; la primera, correr ensamblajes en múltiples ocasiones en la misma plataforma computacional bajo las mismas condiciones iniciales y; la

segunda, se exploraron cambios en ensamblaje debido a la asignación de distintos valores del parámetro del ensamblador más relacionado con sensibilidad y especificidad, la longitud de *k*-mero (Chapman *et al.*, 2011; Surget-Groba y Montoya-Burgos, 2010; Zerbino y Birney, 2008).

Dado que los flujos de trabajo implementados por distintos ensambladores *de novo* son similares, y que el objetivo del proyecto es la inclusión de evidencia experimental como indicador de calidad, se hizo uso de un solo ensamblador para descartar variación de salida por efecto de *software*. Los ensamblajes *de novo* de transcriptoma en este proyecto se efectuaron con el ensamblador *Trinity Transcriptome Assembler* en su versión 2.1.1 (Grabherr *et al.*, 2011).

6.1 Ensamblaje bajo condiciones iniciales idénticas

Se exploraron ensamblajes bajo condiciones iniciales idénticas en tres organismos, dos modelo y uno no modelo con el objetivo de determinar repetibilidad y variabilidad en ensamblaje *de novo* de transcriptoma. Los datos preprocesados RNA-Seq de mosca de la fruta contienen 7,564,138 lecturas pareadas (*Paired-End*, PE) de 32 a 65 bases de longitud; las lecturas de pulga de agua contienen 7,168,393 secuencias PE de 32 a 90 bases de longitud. Los datos del organismo no modelo, camarón blanco contienen 12,907,027 lecturas PE de 80 bases de longitud.

Dado el requerimiento inicial de memoria de la plataforma de cómputo para realizar ensambles, ~1GB de memoria RAM por cada millón de lecturas de entrada (Haas *et al.*, 2013b), y debido a los antecedentes del variación de salida de ensamble según el equipo de cómputo o la ocasión de corrida de proceso (sección 2.2.1), se utilizaron tres sistemas de cómputo de mediana a alta capacidad; un sistema mínimo, representado por una estación de trabajo y dos plataformas de Cómputo de Alto Rendimiento (*High Performance Computing*, HPC). Todas las plataformas corren bajo sistemas operativos Linux de 64 Bits.

Los ensamblajes con equipos HPC se efectuaron en dos distintos centros de cómputo. El primero, el Laboratorio Nacional de Supercómputo del Sureste de México (LNS) de la Benemérita Universidad Autónoma de Puebla (BUAP, 2017); los

servidores disponibles en este centro comparten recursos entre todos los usuarios, en otras palabras, ofrecen servidores compartidos o “virtuales”. El segundo, el proveedor *Penguin Computing*, por medio de su servicio de HPC en la nube *Penguin on Demand* (POD, 2017); este proveedor ofrece servidores dedicados, es decir, una vez que una tarea es asignada a un nodo de computo dicho nodo no compartirá recursos con ningún otro usuario (ver detalles técnicos de las tres plataformas en el Anexo D).

Según los recursos mínimos de ensamblaje, se utilizaron las tres plataformas de cómputo en distintas configuraciones. La memoria RAM en la estación de trabajo se varió de manera física. Ninguna modificación física de recursos es posible en las plataformas HPC, de tal manera que solo se establecieron límites de memoria y núcleos de procesamiento por medio de los comandos de ensamblaje y códigos de proceso en plataformas HPC (*Job Scripts)*. Los parámetros de ensamblaje y asignación de recursos por medio de *Job Script,* y los comandos de ensamblaje Trinity fueron modificados de tal manera que se asignaron las configuraciones memoria/núcleo mencionadas en la Tabla I. Todos los demás parámetros de ensamblaje se especificaron a su valor por defecto.

Se obtuvieron cinco ensambles para los organismos modelo, mosca de la fruta y pulga de agua, bajo todas las configuraciones mencionadas en la tabla I. De la misma forma, se obtuvieron cinco ensambles para camarón blanco (organismo no modelo), pero en este caso solo se trabajó en las configuraciones W_2 y V_2. Una vez terminados los procesos se obtuvieron promedios de conteos de *contigs*, desviaciones estándar y se inició el análisis de contenido.

Se trabajó bajo la conjetura de que la asignación adecuada de recursos computacionales, en especial memoria RAM, es de especial importancia en el caso de ensamblaje *de novo* de transcriptoma, dado que el proceso de agrupamiento (*cluster*) inicial de datos del cual se determinan las isoformas de los *contigs*, depende de la disponibilidad de secuencias al inicio del proceso y la incorporación sucesiva de secuencias candidatas al *cluster*. La distribución inicial de secuencias en las localidades de memoria disponibles para cada procesador influirá, por tanto,

en la formación de estos *clusters*, y asimismo en el resultado final del ensamblaje. Dada esta conjetura, se realizaron monitorizaciones de memoria en dos de las tres plataformas.

Tabla I. Configuraciones de las plataformas de cómputo para realizar ensamblajes.

Plataforma de cómputo				**Parámetros Trinity**	
Configuración	Plataforma	Memoria RAM (GB)	Núcleos	Máxima Memoria	CPU
W_1	Estación de trabajo	20	6	20	6
W_2		24	6	24	6
V_1	HPC, Servidores Virtuales	128/nodo	24/nodo	24	6
V_2		128/nodo	24/nodo	64	12
H_1	HPC, Servidores dedicados	128/nodo	20/nodo	24	6
H_2		128/nodo	20/nodo	64	10

6.1.1 Repetibilidad y variabilidad

Diversas métricas se evaluaron en el análisis de variabilidad bajo condiciones iniciales idénticas en ensamblaje *de novo* de transcriptoma. La formalización de dichas mediciones se presenta en el Anexo E.

La medición de repetibilidad de ensamblaje *de novo* de transcriptoma se realizó con base en la intersección de cinco ensambles bajo condiciones iniciales idénticas en las configuraciones computacionales descritas en la sección anterior. De esta manera se obtuvieron conjuntos de *contigs* intersectados ($I_{(p,m,n)}$), y la repetibilidad por configuración se calculó como el porcentaje representado por este conjunto con respecto a la unión de los *contigs* de los cinco ensambles ($Ctotal_{(p,m,n)}$).

La medición de variabilidad de ensamblaje *de novo* de transcriptoma se realizó con base en la unión de los *contigs* no intersectados de los cinco ensambles bajo condiciones iniciales idénticas en las configuraciones computacionales. De esta manera se obtuvieron conjuntos de *contigs* no intersectados ($\bar{I}_{(p,m,n)}$), y la variabilidad por configuración se calculó como el porcentaje representado por este conjunto con respecto a la unión de los *contigs* de los cinco ensambles ($Ctotal_{(p,m,n)}$).

La generación de ensambles de camarón blanco se realizó solo en dos configuraciones, W_2 y V_2.

Finalmente, la ganancia por variabilidad entre plataformas se cuantificó tomando en cuenta la relación de la variabilidad máxima de las configuraciones de la estación de trabajo entre la variabilidad máxima de las configuraciones en las plataformas basadas en HPC. En camarón blanco, la ganancia por variabilidad se cuantificó por la relación ente variabilidad de las plataformas W_2 y V_2.

Los análisis de conjuntos intersectados y no intersectados se realizaron con Matlab (r2013a) (The MathWorks, 2013).

6.1.2 Monitorización de memoria

El uso de memoria durante procesos de ensamblaje se cuantificó por medio de la monitorización con el comando *top* de Linux; este provee una visualización dinámica del sistema en tiempo real, lo cual incluye el total de memoria RAM que está siendo utilizada por los diversos procesos que se están llevando a cabo (Rankin y Hill, 2013). La monitorización por medio de *top* se realizó en modo *batch*, el cual habilita que la salida por defecto del comando, la pantalla, sea redireccionada a un archivo de texto, realizando actualizaciones cada diez segundos. Posteriormente, se analizaron los datos con Matlab.

El registro generado a partir del muestreo y el registro general de tiempos proveniente del ensamblador (archivo *Trinity.timing*) permitieron identificar las etapas del proceso de ensamblaje que hacen uso más extensivo de memoria. La

monitorización solo se llevó a cabo en el procesamiento de organismos modelo, mosca de la fruta y pulga de agua. Las plataformas monitorizadas fueron la estación de trabajo, y los nodos de cómputo del proveedor *Penguin Computing*, ambas configuraciones. No se llevaron a cabo monitorizaciones en los servidores del LNS ya que funcionan de forma virtual.

6.2 Ensamblaje bajo condiciones iniciales distintas

Esta etapa del proyecto exploró cambios en ensambles *de novo* debido a la variación del parámetro longitud de *k*-mero, pero haciendo uso de una sola plataforma de cómputo para la generación de ensambles. De esta manera se generaron conjuntos de ensambles para las etapas posteriores de evaluación de métricas. Se utilizaron las lecturas de dos organismos modelo, ratón y mosca de la fruta, para la generación de ensambles. Se usaron el conjunto de lecturas NGS de mosca descrito en la sección 6.1 y los datos RNA-Seq de ratón que contienen 20,949,267 lecturas PE de 32 a 66 bases de longitud (Ver anexo C para más información).

6.2.1 Ensamblaje *de novo* variando longitud de *k*-mero

Utilizando la configuración H_2, se obtuvieron ensambles para ratón y mosca de la fruta utilizando los parámetros por defecto del ensamblador, con excepción de la longitud de *k*-mero, la cual se cambió utilizando longitudes mínimas (21), estándar (25) y máximas (31), cinco réplicas de ensamble por longitud de *k*-mero. Estas reconstrucciones son citadas como ensambles *k*21, *k*25 y *k*31 y fueron usadas en el análisis de métricas tradicionales, y adicionalmente para mosca de la fruta, en la inclusión de evidencia experimental para evaluación de calidad.

También se generaron tres ensambles de camarón blanco en la plataforma V_2, uno solo por longitud de *k*-mero *k*21, *k*25 y *k*31, utilizando las lecturas RNA-Seq descritas en la sección 6.1. Este conjunto de ensambles se utilizó en la etapa de inclusión de evidencia experimental para evaluación de calidad.

6.2.2 Adquisición de métricas tradicionales

Por cada grupo de ensamblaje obtenido en la sección 6.2.1 se obtuvieron promedios y desviaciones estándar de métricas tradicionales (también conocidas como cuantitativas).

Se obtuvieron cinco métricas cuantitativas de los conjuntos de ensambles. Primera: cantidad total de *contigs* ensamblados; nombrados como “Transcritos Trinity” por el equipo desarrollador de *software*, estos comprenden todos los transcritos putativos y sus isoformas. Segunda: la cantidad de transcritos putativos sin isoformas (Genes Trinity). Tercera: longitud de ensamble; indica la sumatoria de longitudes de todos los *contigs* formados. Cuarta: N50; denota la longitud del *contig* donde se refleja el 50% del ensamble. Las primeras cuatro métricas se adquirieron por medio de la ejecución de la aplicación *TrinityStats* del mismo ensamblador.

La quinta métrica cuantitativa es la cobertura de lecturas en el ensamble, esta es un indicativo del aprovechamiento de los datos de entrada por parte del ensamblador. Esta métrica solo se obtuvo para ratón y mosca de la fruta, y se calculó por medio del mapeo de las lecturas de entrada con respecto a los distintos ensambles *de novo*. El mapeo se realizó en modo local con el alineador Bowtie2 versión 2.1.0 usando sus parámetros por defecto (Langmead y Salzberg, 2012).

6.3 Calidad de ensamble

Es necesario detectar si los *contigs,* resultado del proceso de ensamblaje *de novo,* fueron representados en el transcriptoma del organismo de estudio, o bien detectar si estos *contigs* fueron artefactos del proceso matemático-computacional. De esta manera se verificó la calidad de los ensambles *de novo* de transcriptoma por medio de secuencias de referencia.

Ya que por el momento se tiene mayor conocimiento sobre secuencias codificantes, se usaron referencias codificantes para evaluar la calidad de ensamble en el caso de organismos modelo, ratón, mosca de la fruta y pulga de agua. Esta evaluación de calidad se realizó por medio de mapeos a las respectivas referencias obtenidas del repositorio Ensembl (Flicek *et al.*, 2014). La referencia de ratón

contiene 103,734 transcritos (versión GRCm38). El transcriptoma de referencia de mosca de la fruta contiene 30,651 transcritos (*release* 86). Para el caso de pulga de agua (versión GCA_000187875.1), la referencia contiene 30,590 transcritos.

Para el organismo no modelo, camarón blanco, la evaluación de calidad se realizó por medio de mapeos a la base de datos de proteínas con curación manual UniProt/Swiss-Prot (Bateman *et al.*, 2015); al momento de su acceso esta contenía 555,426 secuencias.

6.3.1 Calidad con respecto a transcriptomas de referencia en ensambles bajo condiciones iniciales idénticas

Se analizó la validez de *contigs* ensamblados por medio de mapeos al transcriptoma de referencia de los organismos modelo, mosca de la fruta y pulga de agua, detectando de esta manera calidad en ensamble (transcriptomas de referencia descritos en la sección 6.3). Los conjuntos de *contigs* obtenidos del proceso de ensamblaje descritos en la sección 6.1.1 se mapearon a la referencia codificante más próxima a los organismos usando el algoritmo BLASTN (Camacho *et al.*, 2009), de tal manera que por medio de este procedimiento se detectó la calidad en los *contigs* originados por la variación en las plataformas de cómputo.

Se pudiese dar el caso que los *contigs* intersectados mapeados ($Im_{(p,m)}$) y los *contigs* no intersectados mapeados ($\bar{I}m_{(p,m)}$) apuntasen a transcritos comunes en la referencia (Fig. 4). Se les llamaron *contigs* no intersectados compartidos ($\bar{I}m^{*}_{(p,m)}$) a todos los *contigs* que presenten mapeos en común con *contigs* intersectados. El caso contrario fueron los *contigs* no intersectados exclusivos a determinada configuración de una plataforma $\bar{I}m^{+}_{(p,m)}$.

La evaluación de calidad considera como información válida originada por variabilidad de plataforma a todos aquellos *contigs* contenidos en el subconjunto $\bar{I}_{(p,m,n)}$. Esta evaluación se expresó como el porcentaje representado por los *contigs* mapeados provenientes del conjunto no intersectado mapeado ($\bar{I}m_{(p,m)}$) con respecto al conjunto Intersectado ($\bar{I}_{(p,m,n)}$).

Asimismo, solo se consideró como información nueva originada por la variabilidad de plataforma a los *contigs* mapeados exclusivos al conjunto no intersectado ($\bar{I}m^{+}_{(p,m)}$). Esta evaluación se expresó en el porcentaje representado por los *contigs* mapeados exclusivos provenientes del conjunto no intersectado ($\bar{I}m^{+}_{(p,m)}$) con respecto al conjunto no intersectado ($\bar{I}_{(p,m,n)}$).

La ganancia en calidad se cuantificó tomando en cuenta la relación del número máximo de *contigs* representados por el conjunto $\bar{I}m_{(p,m)}$ de una de las configuraciones basadas en la estación de trabajo, entre el número máximo de *contigs* representados en $\bar{I}m_{(p,m)}$ de una de las configuraciones basadas en HPC.

De la misma forma, la ganancia en información nueva se cuantificó tomando en cuenta la relación del número máximo de *contigs* representados por el conjunto $\bar{I}m^{+}_{(p,m)}$de una de las configuraciones basadas en la estación de trabajo, entre el número máximo de *contigs* representados en $\bar{I}m_{(p,m)}$ de una de las configuraciones basadas en HPC.

Se establecieron los parámetros del *software* BLAST de tal manera que se obtuvo un *hit* (alineamiento positivo) por secuencia de entrada y alta similitud entre secuencias alineadas, pero con bajos valores de expectación. Umbrales: valor de expectación *e-value* $1x10^{-9}$; porcentaje de identidad: 95%; alineamientos máximos por secuencia de entrada *max_hps*: 1; cantidad de secuencias alineadas *max_target_seqs*:1; cantidad de núcleos: 1.

6.3.2 Calidad con respecto a transcriptomas de referencia en ensambles con variación de longitud de *k*-mero

Se realizaron dos distintas verificaciones de calidad con base en la referencia de transcriptoma para los distintos grupos de ensamble con variación de *k*-mero. La primera, aplicable a los grupos completos de ratón y mosca de la fruta, consiste en mapeos con altos umbrales de alineamiento (mapeos estrictos) debido a que estos grupos de ensambles fueron utilizados para el análisis de métricas tradicionales. La segunda verificación fue aplicable a los primeros ensambles de los grupos *k*21, *k*25 y *k*31 de mosca de la fruta y consta de mapeos BLASTN a la referencia.

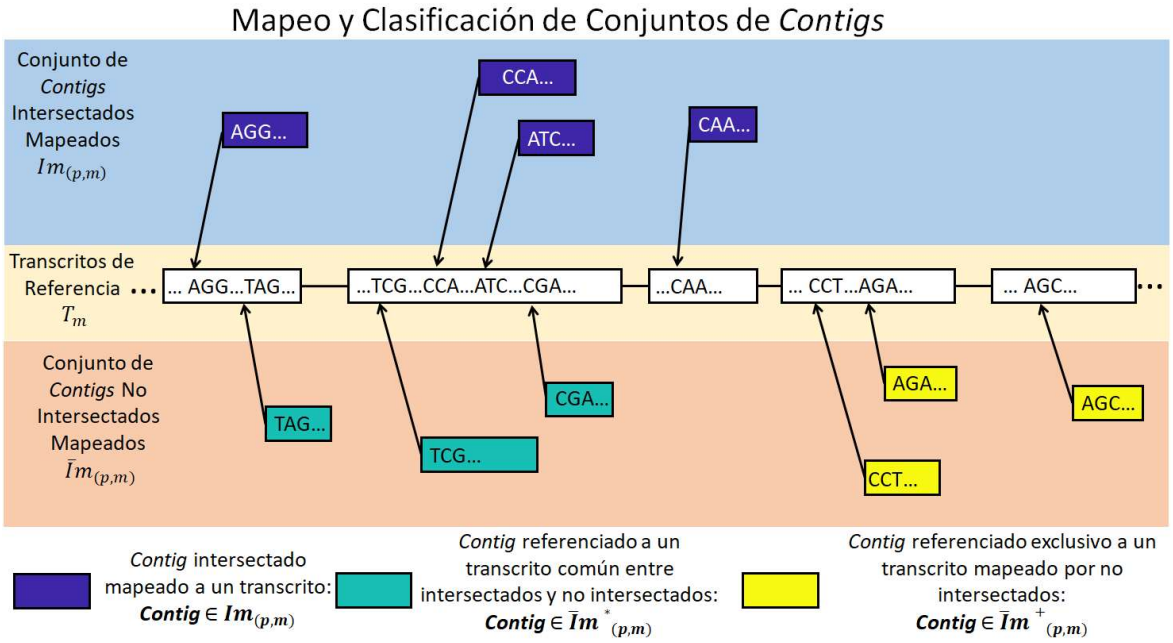


Figura 4. Mapeo de *contigs* intersectados y no intersectados. Se muestra la clasificación dependiendo del transcripto al que hayan sido referenciados.

En la primera verificación los mapeos fueron llevados a cabo en el modo local del alineador Bowtie2, configurando los parámetros de búsqueda para usar una longitud de palabra base (*seed length*) de 21 bases, permitiendo una no coincidencia en la base (*mismatch*), buscando 3 veces el mismo *seed* en intervalos que pueden variar desde 8 bases para *contigs* de ~200 bases de longitud hasta 46 bases para *contigs* de ~8200 bases de longitud según la formula $i = 1 + 0.5 * sqr(L)$, siendo L la longitud del *contig* (Langmead y Salzberg, 2012).

La segunda verificación de calidad consistió en realizar alineamientos BLASTN de los primeros ensambles de los grupos *k*21, *k*25 y *k*31 de mosca de la fruta (parámetros de alineamiento según el procedimiento de la sección 6.3.1). Estos mapeos fueron utilizados en los análisis de generalizaciones de datos de evidencia experimental (sección 6.6.2).

Los ensambles con la mayor semejanza a la referencia mostraron el mayor porcentaje de alineamiento, comprobando estos mapeos cuales fueron los mejores ensambles (transcriptomas de referencia descritos en la sección 6.3).

Para ambas verificaciones, los ensambles con la mayor semejanza a la referencia mostraron el mayor porcentaje de alineamiento, comprobando estos cuales son los mejores ensambles (transcriptomas de referencia descritos en la sección 6.3). Si bien ambas verificaciones, Bowtie2 y BLASTN, dan indicativos de similitud entre ensambles y referencias, el uso de algoritmos más estrictos de alineamiento, como Bowtie2, permite una mayor especificidad en la búsqueda; sin embargo, BLAST que es menos específico, es usado más frecuentemente con propósitos de búsquedas de similitud.

6.3.3 Calidad con respecto a la base de datos de proteínas UniProt/Swiss-Prot

Se realizaron tres verificaciones de calidad con base en la referencia la base de datos de proteína UniProt/Swiss-Prot. La primera es aplicable a los grupos de ensamble de camarón blanco generados en la etapa del proyecto descrita en la sección 6.1.1. La segunda, a los primeros ensambles de los grupos *k*21, *k*25 y *k*31 de mosca de la fruta, y los mapeos serán utilizados en los análisis de generalizaciones de datos de evidencia experimental (sección 6.6.2). La segunda y tercera verificaciones son aplicables a los primeros ensambles *k*21, *k*25 y *k*31 de mosca de la fruta y a los ensambles *k*21, *k*25 y *k*31 de camarón blanco respectivamente, y los mapeos serán utilizados en los análisis de generalizaciones de datos de evidencia experimental (sección 6.6.3).

La forma de evaluar la calidad de ensamble depende de la disponibilidad de información del organismo de estudio. Para el caso del organismo no modelo, camarón blanco, no se cuenta con un transcriptoma de referencia para poder analizar la validez de los *contigs* generados. De tal manera que se utilizó una base de datos auxiliar de alta calidad que permitirá la búsqueda de secuencias de proteínas revisadas manualmente dentro de los *contigs* de ensamble. Es común

seguir esta estrategia cuando no se cuenta con una referencia aplicable al transcriptoma de estudio, consecuentemente es apropiada para este organismo (Ghaffari *et al.*, 2015).

La evaluación de calidad consistió en el mapeo de *contigs* a la base curada de datos de proteínas UniProt/Swiss-Prot, que a la fecha del estudio (13 de septiembre del 2017) consistía en 555,426 secuencias aminoacídicas (The UniProt Consortium, 2017). Se analizó la calidad de los *contigs* originados por la variabilidad de cada plataforma de cómputo, por lo tanto se mapearán los conjuntos de *contigs* no intersectados ($\bar{I}_{(p,m,n)}$) con respecto la base de datos de proteínas utilizando el algoritmo BLASTX del *software* BLAST (Camacho *et al.*, 2009). Únicamente se seleccionaron los mapeos de mayor calidad dentro de los seis posibles marcos de lectura por *contig*.

Los parámetros utilizados en el mapeo BLASTX se configuraron de tal manera que se obtuvo un *hit* (alineamiento positivo) por secuencia de entrada, bajos valores de expectación y la retención de los segmentos alineados de las secuencias. Umbrales: valor de expectación *e-value* $1x10^{-6}$; alineamientos máximos por secuencia de entrada *max_hps*: 1; cantidad de secuencias alineadas *max_target_seqs*:1; cantidad de núcleos: 22; retención de segmentos alineados de entrada y base de datos activados *qseq* y *sseq* dentro de las especificaciones de salida de formato *outfmt*.

En la primera verificación, la evaluación de calidad se expresó en el porcentaje representado por los *contigs* no intersectados mapeados ($\bar{I}mp_{(p,m)}$) con respecto al conjunto de *contigs* no intersectados ($\bar{I}_{(p,m,n)}$).

La ganancia máxima en *contigs* no intersectados mapeados a la base de datos de proteínas será igual a la división de la cantidad máxima de *contigs* $\bar{I}mp_{(p,m)}$ de la estación de trabajo entre la cantidad máxima de estos *contigs* obtenidos en HPC.

Para el análisis de camarón blanco, el cual no se basa en transcriptoma de referencia para evaluar calidad, se tomó la consideración de que el conjunto

completo de *contigs* no intersectados mapeados a la base de datos de proteínas $\bar{I}mp_{(p,m)}$ constituyen el descubrimiento de nueva información por plataforma.

La segunda verificación de calidad consistió en realizar alineamientos BLASTX de los primeros ensambles de los grupos *k*21, *k*25 y *k*31 de mosca de la fruta (parámetros de alineamiento según el procedimiento esta misma sección). Estos mapeos fueron utilizados en los análisis de generalizaciones de datos de evidencia experimental (sección 6.6.2).

La tercera verificación de calidad consistió en realizar alineamientos BLASTX de los ensambles *k*21, *k*25 y *k*31 de camarón blanco utilizando los parámetros de alineamiento dictados esta misma sección. Los mapeos fueron utilizados en los análisis de generalizaciones de datos de evidencia experimental (sección 6.6.3).

6.4 Análisis de desempeño de métricas tradicionales de calidad

Una vez obtenidas las métricas tradicionales y realizadas las evaluaciones de calidad por medio de mapeo de transcriptomas de referencia, se pudo analizar el desempeño de las métricas tradicionales de calidad para ensamblaje *de novo* de transcriptoma.

Se compararon en esta etapa del proyecto las métricas y mapeos de ratón y mosca de la fruta obtenidas en las secciones 6.2.2 y 6.3.2.

Una vez que se analizó el desempeño de las métricas existentes, se caracterizaron distintas fuentes de variabilidad en ensamblaje *de novo* y se evaluaron los diversos conjuntos de ensambles por medio de mapeos a sus respectivas referencias, se sentaron las bases para hacer uso de evidencia experimental para el apoyo de los flujos de trabajo de ensamblaje *de novo* de transcriptoma. El uso de evidencia experimental se puede realizar de manera directa o de manera generalizada por medio del uso de modelos estadísticos con base en dicha evidencia.

6.5 Uso directo de datos experimentales para evaluación y selección de ensamble

En esta etapa del proyecto se propuso un criterio de calidad para la selección de ensamble *de novo* de transcriptoma dentro de la estrategia de ensamblaje múltiple. Dicho criterio se basa en la evaluación de calidad por medio del uso directo de sondas hibridadas de microarreglos de expresión génica, al hacer mapeos de dichas sondas a distintos ensambles y seleccionar el ensamble de mayor calidad dentro del conjunto.

Se verificó el criterio propuesto, no obstante, esta verificación no es en sí parte del flujo de trabajo. La verificación consistió en el mapeo de los conjuntos de ensambles a la referencia del organismo en cuestión.

El criterio propuesto se aplicó por medio de un flujo de trabajo que consta de cuatro fases:

i) Selección de bases de datos de microarreglos de condiciones de experimentación similares al RNA-Seq.

ii) Ensamblaje *de novo* de transcriptoma por medio de procedimientos estándar (generación de conjuntos de ensambles), lo cual comprende reportes de calidad y preprocesamiento, ensamblaje *de novo* de un conjunto de ensambles con longitudes de *k*-mero distintas, obtención de métricas tradicionales y análisis de cobertura.

iii) Análisis de los datos de microarreglos (detección de sondas hibridadas).

iv) Evaluación de calidad basada en microarreglos y selección de ensamble dentro del conjunto. Se indica como el ensamble de mayor calidad dentro del conjunto aquel donde, por medio de mapeo de sondas hibridadas a ensambles, se encuentre la mayor cantidad de sondas, sugiriendo dicho ensamble para ser seleccionado.

6.5.1 Bases de datos de microarreglos

La selección de datos de microarreglos con respecto a las lecturas RNA-Seq se fijó a que las condiciones experimentales de secuenciación y microarreglos fueran las

mismas, sin embargo, organismos cercanos y condiciones similares fueron suficientes para la validación. Se utilizaron microarreglos de dos organismos, ratón y mosca de la fruta, dos replicas por organismo (Tabla II).

Tabla II. Microarreglos de expresión génica.

Organismo	**Ratón**	**Mosca**
Repositorio	GEO[1]	GEO[1]
Fuente	(Su *et al.*, 2004)	(Doroszuk *et al.*, 2012)
No. ID.	GSM258635 y GSM258636	GSM897165 y GSM897166
Condición de estudio	Corteza cerebral de ratones C57BL/6 macho de 8-10 semanas	Organismos completos de hembras adultas alimentadas con dietas óptimas
Microarreglo	*Affimetrix Mouse Genome 430 2.0 array*	*Affimetrix GeneChip Drosophila Genome 2.0 array*
Cantidad de sondas del panel	495,374 y 1,094 sondas de prueba y control respectivamente	263,272 sondas de prueba y 2,128 de control
Longitud	25-meros[2]	25-meros[2]
Transcritos representados en el panel	34,000	18,550
Conjuntos de prueba[3]	45,000, 11 oligonucleótidos por conjunto	18,880, 14 oligonucleótidos por conjunto

Repositorio: [1]Ómnibus de Expresión (GEO) perteneciente al NCBI (Barrett *et al.*, 2013). [2] *n*-meros se refiere a la longitud de secuencia de la sonda. [3] Conjunto de prueba o *probe-set:* Conjunto de sondas que apuntan a un mismo transcrito de prueba.

6.5.2 Ensamblaje *de novo* de transcriptoma (generación de conjuntos de ensambles)

La segunda fase del flujo de trabajo comprendió la creación de conjuntos de ensambles por medio de técnicas estándar. Para ambos organismos, ratón y mosca de la fruta, se utilizaron los conjuntos de ensamblajes generados por medio de la variación del tamaño de *k*-mero descrita en la sección 6.2.1; sus métricas tradicionales están descritas en la sección 6.2.2.

El flujo de trabajo sólo requiere de un ensamble por condición inicial; sin embargo, se utilizaron las cinco réplicas de ensamblaje por longitud de *k*-mero con el propósito de verificación de diferencias en resultados por medio de las pruebas estadísticas entre grupos descritas en la sección 6.5.5.

6.5.3 Análisis de microarreglos

El análisis de datos de microarreglos se realizó en dos pasos. 1) Establecer los umbrales de hibridación de las bases de datos y; 2) detectar las sondas de prueba con hibridación positiva. Ambos pasos se realizaron por medio de códigos escritos en la plataforma de cómputo científico Matlab.

1. Umbrales de hibridación: Los umbrales de hibridación se establecieron con base en las sondas de controles negativos de las bases de datos (Causton *et al.*, 2003). Se descartaron aquellas sondas de control negativo que reportaron valores de intensidad por encima de la media más una desviación estándar. Los umbrales se establecieron en el valor medio de intensidad de los controles negativos no descartados.

2. Detección de sondas de prueba con hibridación positiva: Para toda base de datos de microarreglo se declaró con hibridación positiva a toda aquella sonda de prueba cuyo valor de intensidad estuviese por encima del umbral de hibridación. Cada organismo cuenta con dos bases de datos de microarreglos con condiciones idénticas de experimentación. Una vez que realizaron las pruebas de hibridación se efectuaron validaciones cruzadas de resultados. Se tomaron como sondas de verificación a aquellas sondas que presentaron hibridación positiva en ambas bases

de datos (que la misma sonda resultó positiva en ambas bases de datos). Las sondas de validación son referidas como sondas de hibridación mutua (sondas MH).

6.5.4 Evaluación de calidad basada en microarreglos y selección de ensamble

La evaluación de calidad se realizó mapeando las sondas MH a los conjuntos de ensambles de cada organismo. Los ensambles con la mayor cantidad de sondas MH indicaron mayor calidad, por lo tanto, se sugirieron para su selección dentro del conjunto. Este paso se realizó con el alineador Bowtie2 en modo local utilizando sus parámetros por defecto.

Este paso finalizó el flujo de trabajo para la determinación del criterio propuesto. La siguiente sección describe la evaluación de éste; cabe destacar que no es parte de la evaluación de calidad y selección de ensamble con base en microarreglos.

6.5.5 Verificación del criterio propuesto

La verificación se llevó a cabo según el procedimiento dictado en la sección 6.3.2. Los ensambles con la mayor semejanza a la referencia mostraron el mayor porcentaje de alineamiento, comprobando estos cuáles son los mejores ensambles y verificando la selección por medio del criterio de selección por microarreglos.

Los cinco ensambles por grupo permitieron la ejecución de análisis estadístico y la detección de diferencias significativas de la cantidad de sondas MH mapeadas a los grupos de ensamblaje *k*21, *k*25 y *k*31. Este análisis se llevó a cabo mediante pruebas ANOVA de una vía y consecutivamente pruebas Tukey con la plataforma Matlab.

6.6 Generalización de datos experimentales para evaluación

El segundo enfoque de uso de evidencia experimental al RNA-Seq como criterio de evaluación de calidad y selección del ensamble *de novo* de transcriptoma consistió en la generalización de secuencias por medio de Modelos Markovianos Ocultos. El uso de modelos permitió la evaluación de ensambles completos. Posteriormente,

permitió determinar cuál de los ensambles dentro de un conjunto tiene mayor calidad. El criterio se aplicó en ensambles de mosca de la fruta y en ensambles de camarón blanco, utilizando HMMs obtenidos a partir de microarreglos para el primer organismo y secuencias de UniGenes para el segundo.

6.6.1 Modelos Markovianos Ocultos (HMM) con base en evidencia experimental

Para ambas fuentes de datos de evidencia experimental, microarreglos y UniGenes, se inició con un modelo similar al descrito por Churchill (1989); HMMs de primer orden, ergódicos y de dos estados. Se representaron en las secuencias de evidencia regiones ricas en transiciones, a partir del estado 1, y en transversiones, a partir del estado 2 (Fig. 5). El conjunto de parámetros finales de los modelos se definió una vez que se ejecutaron los entrenamientos de los HMM.

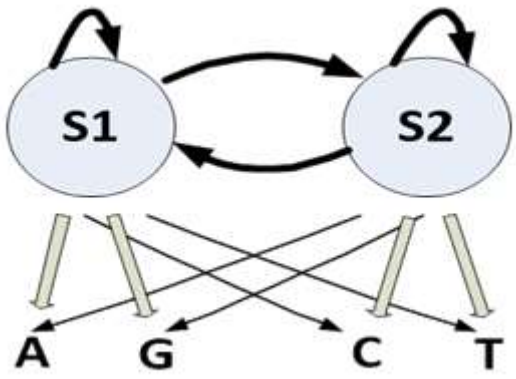


Figura 5. HMMs con base en secuencias experimentales. Basado en Churchill (1989).

6.6.2 Empleo de HMMs con base en microarreglos para la evaluación de ensambles de mosca de la fruta

El uso generalizado de sondas hibridadas de microarreglo para la evaluación y selección de ensamble se realizó en tres pasos: la obtención del conjunto de ensambles a ser evaluados, el entrenamiento de HMMs con base en microarreglos,

y, por último, la evaluación de los ensambles. Se realizó una verificación de la evaluación por medio de HMMs, sin embargo, esta verificación no forma parte del flujo de trabajo propuesto para la evaluación de calidad y la selección de ensamble.

6.6.2.1 Conjunto de ensambles *de novo* de transcriptoma

Se usaron los primeros ensambles de los grupos de ensamblaje *k*21, *k*25 y *k*31 de mosca de la fruta generados en la sección 6.2.1.

6.6.2.2 Entrenamiento de los HMM

Se utilizaron como datos experimentales los datos de microarreglo de expresión génica descritos en la sección 6.5.1. Dadas las características del microarreglo y sus sondas, la selección de secuencias de entrenamiento y el entrenamiento de los modelos se realizaron de la siguiente forma:

1. Detección de sondas hibridadas en el experimento de expresión: consistió en detectar las sondas hibridadas de la base de datos de microarreglos. Se hizo uso de las sondas MH descritas en la sección 6.5.3.
2. Para cada sonda MH se distinguieron sus correspondientes sondas de cobertura de transcritos (*probe sets)*. Independientemente de la cantidad de sondas encontradas con hibridación positiva por *probe set*, si al menos una de las sondas por conjunto resultó positiva, se tomó el conjunto entero de sondas para entrenamientos de HMMs, un modelo por conjunto con evidencia de hibridación.
3. Entrenamiento de modelos por medio del algoritmo *Baum-Welch* (Rabiner, 1989).

La identificación y selección de *probe sets*, así como el entrenamiento de los HMM fueron realizados por medio Matlab.

6.6.2.3 Evaluación de ensamble por medio de HMMs

La evaluación de los distintos ensamblajes dados los modelos entrenados en el paso anterior se realizó siguiendo el método *Forward* (Rabiner, 1989).

Se evaluó por medio del método *forward* la probabilidad de cada uno de los *contigs* de los ensambles dados los distintos modelos ($\Pr(S_i|\lambda_n)$) y para cada uno

de los *contigs,* se tomó la probabilidad máxima detectada por los modelos ($maxPr(S_i|\lambda)$).

Una vez calculadas las probabilidades máximas de cada *contig* se normalizaron según su longitud de secuencia (descrita como $maxPrnor$), multiplicando la probabilidad por un factor de $1x10^{ls}$, donde $ls = longitud\ del\ contig - 200$.

La evaluación final de cada ensamble se calculó haciendo la sumatoria de las probabilidades máximas normalizadas de los *contigs* del conjunto dividida entre la cantidad de *contigs* del ensamble:

$$(Pnor_{ensamble} = \sum_{i=1}^{n} \frac{maxPrnor_i}{n}) \qquad (1)$$

Donde $n = cantidad\ de\ contigs\ del\ ensamble$. Se tomó el conjunto completo de HMMs para procesar la evaluación.

Las evaluaciones y cálculos finales de probabilidad de ensamble se efectuaron por medio de la plataforma Matlab y el *toolbox* de Matlab de análisis de HMMs de Murphy (Murphy, 2005).

6.6.2.4 Verificación de la evaluación de ensamble por medio mapeos

Se realizaron tres distintas verificaciones de la evaluación de los ensambles. La primera se realizó por medio de los mapeos Bowtie2 con respecto al transcriptoma de referencia descritos en la sección 6.3.2. La segunda por medio de los mapeos BLASTN de ensamblaje al transcriptoma de referencia (descritos en la sección 6.3.2). La tercera verificación consistió en los mapeos de ensamblajes a la base de datos UniProt/Swiss-Prot descritos en la sección 6.3.3.

6.6.3 Empleo de HMMs con base en UniGenes para evaluación de ensambles de camarón blanco

El uso generalizado de *clusters* UniGene para la evaluación y selección de ensamble se realizó en tres pasos: la obtención del conjunto de ensambles a ser evaluados, el entrenamiento de HMMs con base en secuencias de los *cluster* UniGene, y, por último, la evaluación de los ensambles. Se realizó una verificación

de la evaluación por medio de HMMs, sin embargo, esta verificación no forma parte del flujo de trabajo propuesto para la evaluación de calidad y la selección de ensamble.

6.6.3.1 Conjunto de ensambles *de novo* de transcriptoma

Se usaron los ensambles de camarón *k*21, *k*25 y *k*31 y sus respectivas métricas tradicionales que fueron generados según los procedimientos descritos en las secciones 6.2.1 y 6.2.2.

6.6.3.2 Entrenamiento de los HMM

El entrenamiento de modelos dadas las secuencias experimentales se realizó siguiendo el método iterativo Baum-Welch (Rabiner, 1989).

Para la evaluación de camarón blanco se tomaron UniGenes del repositorio NCBI como secuencias de evidencia experimental (Pontius *et al.*, 2002), ya que dichos *clusters* contienen ESTs. Las condiciones de selección fueron: Todos los UniGenes del organismo *Litopenaeus vannamei* almacenados en el repositorio NCBI que contengan 20 ESTs (Coordinators, 2013). Algunos de los *clusters* seleccionados pudiesen contener un poco más de 20 secuencias, ya que aparte de 20 ESTs, también contienen algún otro tipo de secuencias.

Las secuencias de los UniGenes se utilizaron de manera similar a las condiciones encontradas en *probe sets* de microarreglos comerciales, es decir, pocas decenas de secuencias de longitud corta (por ejemplo, los microarreglos Affimetrix contienen 20 o menos sondas por *probe set* con una longitud de 25 bases por secuencia).

El entrenamiento de HMMs se realizó bajo el siguiente procedimiento y consideraciones:

1. Se verificó que los UniGenes contuvieran al menos 20 secuencias sin bases indeterminadas, eliminando aquellos *clusters* que no cumpliesen con esta condición.
2. Se extrajo una porción aleatoria de 25 bases consecutivas para cada una de las 20 secuencias de los UniGenes seleccionados.

3. Se entrenaron los HMMs por medio del algoritmo *Baum-Welch*.

La identificación y selección de *clusters*, así como el entrenamiento de los HMM se realizaron con Matlab (2013a).

6.6.3.3 Evaluación de ensamblaje por medio de HMMs

La evaluación de los distintos ensamblajes dados los modelos entrenados en el paso anterior se realizó siguiendo el método *Forward* (Rabiner, 1989). Por medio de este método se evaluó la probabilidad de cada uno de los *contigs* de los ensambles dados los distintos modelos ($\Pr(S_i|\lambda_n)$) y para cada uno de los *contigs* se tomó la probabilidad máxima detectada por los modelos ($maxPr(S_i|\lambda)$).

Una vez calculadas las probabilidades máximas de cada *contig* se normalizaron según su longitud de secuencia (descrita como $maxPrnor$), multiplicando la probabilidad por un factor de $1x10^{ls}$, donde $ls = longitud\ del\ contig - 200$.

La evaluación final de cada ensamble se calculó haciendo la sumatoria de las probabilidades máximas normalizadas de los *contigs* del conjunto dividida entre la cantidad de *contigs* del ensamble:

$$(Pnor_{ensamble} = \sum_{i=1}^{n} \frac{maxPrnor_i}{n}) \qquad (2)$$

Donde $n = contigs\ del\ ensamble$.

Las evaluaciones y cálculos finales de probabilidad de ensamble se efectuaron por medio de la plataforma Matlab 2013a y el *toolbox* de Matlab de análisis de HMMs de Murphy (Murphy, 2005).

6.6.3.4 Verificación de la evaluación de ensamblaje por medio mapeos

Los tres ensambles se mapearon a la base de datos de proteínas UniProt/Swiss-Prot siguiendo el procedimiento descrito en la sección 6.3.3 como método de verificación de calidad.

7. RESULTADOS

7.1 Ensamblaje bajo condiciones iniciales idénticas

La tabla III muestra el promedio de *contigs* generados después de 5 repeticiones del ensamblaje efectuados en cada configuración. De los promedios y desviaciones se puede apreciar que el número de *contigs* generados en cada repetición es muy similar; las desviaciones estándar representan menos del 0.08% de los conjuntos completos de *contigs*, y la diferencias ente mínimos y máximos de cantidad de *contigs* fueron 16.2, 99.2 y 52.4 secuencias para mosca de agua, pulga de agua y camarón blanco respectivamente.

Tabla III. Cantidad de *contigs* ensamblados por plataforma computacional.

	Mosca de la fruta	
Plataforma Computacional	**Promedio**	**Desviación Estándar**
$\boldsymbol{W_1}$	25,994.80	6.72
$\boldsymbol{W_2}$	25,988.80	5.81
$\boldsymbol{H_1}$	25,981.60	3.44
$\boldsymbol{H_2}$	25,984.40	5.68
$\boldsymbol{V_1}$	25,988.40	3.65
$\boldsymbol{V_2}$	25,989.40	2.30
	Pulga de agua	
Plataforma Computacional	**Promedio**	**Desviación Estándar**
$\boldsymbol{W_1}$	53,280.4	37.63
$\boldsymbol{W_2}$	53,276.8	39.93
$\boldsymbol{H_1}$	53,220.8	20.32
$\boldsymbol{H_2}$	53,247.8	15.08
$\boldsymbol{V_1}$	53,321.6	23.06
$\boldsymbol{V_2}$	53,320.0	23.98
	Camarón blanco	
Plataforma Computacional	**Promedio**	**Desviación Estándar**
$\boldsymbol{W_2}$	68,063.80	29.87
$\boldsymbol{V_2}$	68,116.20	26.74

7.1.1 Repetibilidad y variabilidad

Las figuras 6, 7 y 8 muestran el resultado de la intersección de los conjuntos de 5 ensambles por configuración de cómputo para mosca de la fruta, pulga de agua y camarón blanco. Las intersecciones son los *contigs* comunes a los ensambles (contienen exactamente las mismas secuencias). Basta un cambio de base, inserción o perdida entre dos *contigs* para que se consideren ambos distintos y se envíen al conjunto de no intersectados. Nótese la diferencia de escalas, y que los *contigs* no intersectados constituyen menos del 5% en el caso de la mosca de la fruta, menos del 23% para pulga de agua y menos del 18% para camarón blanco.

Se muestran en la tabla IV los porcentajes de repetibilidad y variabilidad por plataforma para los tres organismos, encontrándose mayor repetibilidad en las plataformas HPC (H_1, H_2, V_1 y V_2), pero mayor variabilidad en las plataformas con menor memoria. Las ganancias máximas por variabilidad en la estación de trabajo se presentan en la tabla V.

7.1.2 Monitorización de memoria

La figura 9 muestra el uso de memoria RAM durante los procesos de ensamblaje *de novo* de transcriptoma para mosca de la fruta y pulga de agua en las plataformas H_2 y W_2. Se delimitó con una línea vertical intermitente la duración del módulo *Inchworm*.

Nótese que el manejo de los millones de secuencias de entrada se ven reflejadas en el manejo de memoria por parte del ensamblador, sobre todo en el primer módulo, *Inchworm*, donde ambos organismos hicieron uso intensivo de memoria, ~80 GB en la plataforma H_2 (Fig. 9c y 9d). La duración del primer módulo en las configuraciones H_2 fue de ~24 minutos en procesos de mosca de la fruta y ~5 minutos en ensamblajes de pulga de agua. Asimismo, se puede observar que el uso de memoria en los módulos posteriores fue mayor en los procesos de pulga de agua teniendo un pico de uso en 34.3 GB (Tabla VI). El uso de memoria en los módulos alternados *Chrysalis* y *Butterfly* en el caso de mosca de la fruta no excedió los 24 GB en ambas plataformas.

Tabla IV. Repetibilidad y variabilidad.

Mosca de la Fruta					
Plataforma Computacional	$Ctotal_{(p,mosca,5)}$	$I_{(p,mosca,5)}$	Repetibilidad (%)	$\bar{I}_{(p,mosca,5)}$	Variabilidad (%)
$\boldsymbol{W_1}$	26,618	25,422	95.51	1,196	4.49
$\boldsymbol{W_2}$	26,544	25,474	95.97	1,070	4.03
$\boldsymbol{H_1}$	26,286	25,693	97.74	593	2.26
$\boldsymbol{H_2}$	26,286	25,692	97.74	594	2.26
$\boldsymbol{V_1}$	26,202	25,784	98.40	418	1.60
$\boldsymbol{V_2}$	26,180	25,807	98.58	373	1.42
Pulga de Agua					
Plataforma Computacional	$Ctotal_{(p,pulga,5)}$	$I_{(p,pulga,5)}$	Repetibilidad (%)	$\bar{I}_{(p,pulga,5)}$	Variabilidad (%)
$\boldsymbol{W_1}$	60,632	47,510	78.36	13,122	21.64
$\boldsymbol{W_2}$	60,943	47,261	77.55	13,682	22.45
$\boldsymbol{H_1}$	56,617	50,590	89.35	6,027	10.65
$\boldsymbol{H_2}$	56,630	50,580	89.32	6,050	10.68
$\boldsymbol{V_1}$	55,545	51,627	92.95	3,918	7.05
$\boldsymbol{V_2}$	55,176	51,783	93.85	3,393	6.15
Camarón blanco					
Plataforma Computacional	$Ctotal_{(p,cam,5)}$	$I_{(p,cam,5)}$	Repetibilidad (%)	$\bar{I}_{(p,cam,5)}$	Variabilidad (%)
$\boldsymbol{W_2}$	75,304	62,211	82.91	13,093	17.34
$\boldsymbol{V_2}$	70,505	66,184	93.87	4,321	6.12

Repetibilidad: $I_{(p,m,n)}$ / $Ctotal_{(p,m,n)}$. Variabilidad: $\bar{I}_{(p,m,n)}$ / $Ctotal_{(p,m,n)}$. Máximos por organismo y plataformas locales y HPC marcados en gris.

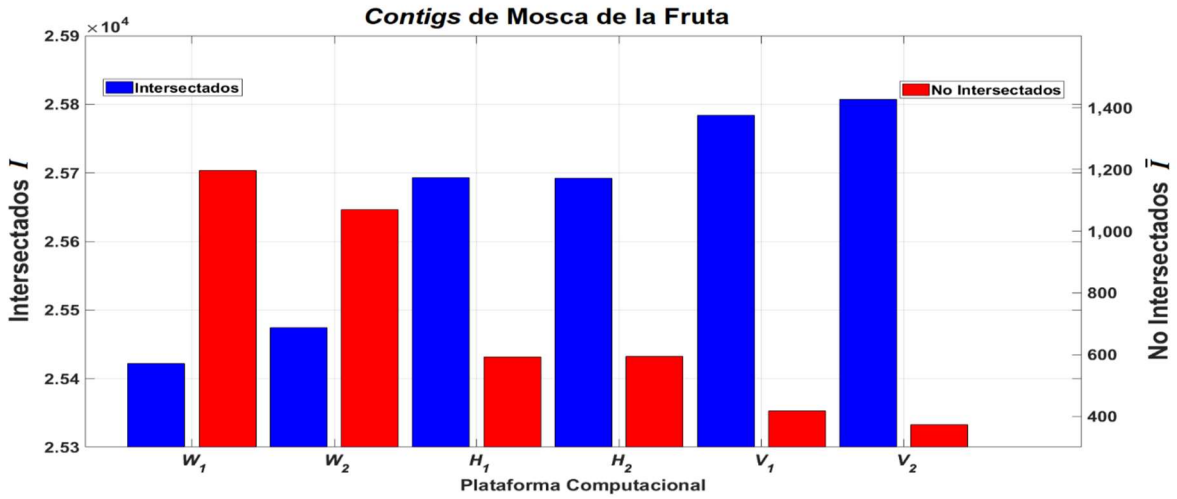


Figura 6. Comparación de los conjuntos de *contigs* intersectados de mosca de la fruta. Conjuntos de *contigs* intersectados $I_{(p,mosca,5)}$ y los conjuntos de *contigs* no intersectados $\bar{I}_{(p,mosca,5)}$ obtenidos después de 5 repeticiones de ensambles $E_{(p,mosca,5)}$ para el organismo Mosca de la Fruta, por cada plataforma computacional p de la Tabla I.

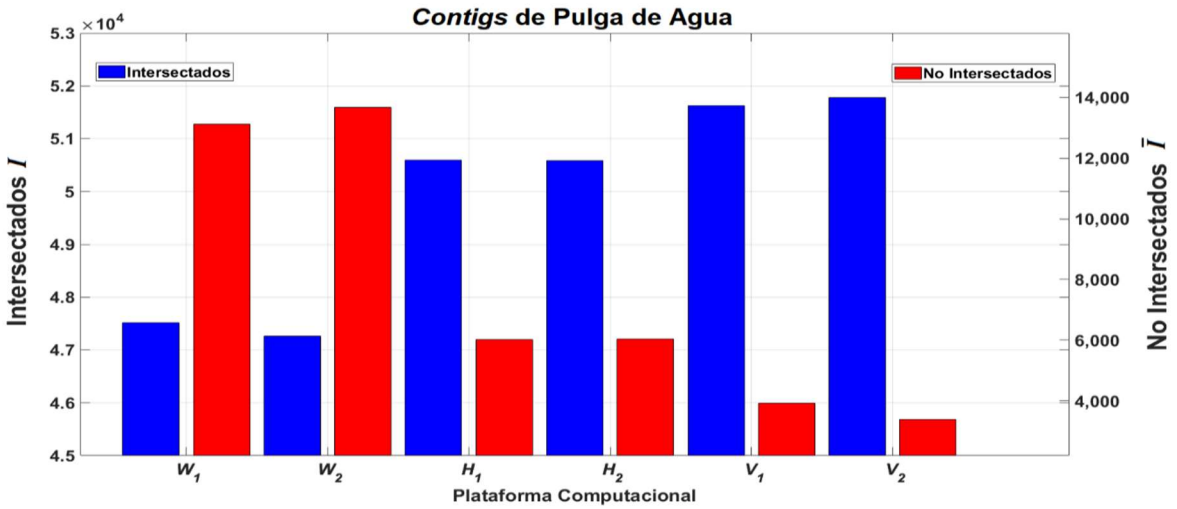


Figura 7. Comparación de los conjuntos de *contigs* intersectados de Pulga de agua. Conjuntos de *contigs* intersectados $I_{(p,pulga,5)}$ y los conjuntos de *contigs* no intersectados $\bar{I}_{(p,pulga,5)}$, obtenidos después de 5 repeticiones de ensambles $E_{(p,pulga,5)}$ para el organismo Pulga de Agua, por cada plataforma computacional p de la Tabla I.

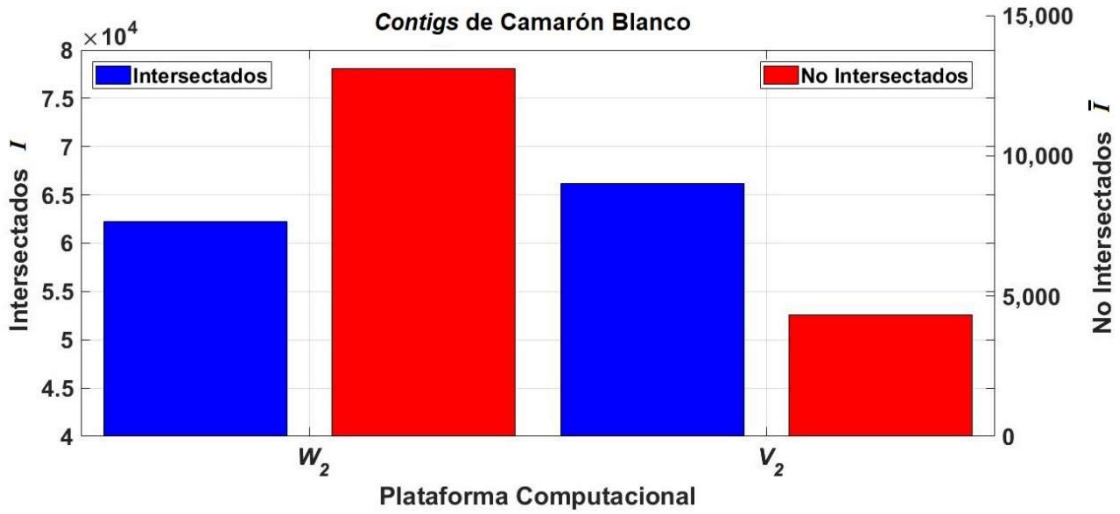


Figura 8. Comparación de los conjuntos de *contigs* intersectados de camarón blanco. Conjuntos de *contigs* intersectados $I_{(p,cam,5)}$ y los conjuntos de *contigs* no intersectados $\bar{I}_{(p,cam,5)}$, obtenidos después de 5 repeticiones de ensambles $E_{(p,cam,5)}$ para el organismo camarón blanco, por cada plataforma computacional p de la Tabla I.

Tabla V. Ganancias por variabilidad.

Organismo	% de variabilidad máxima en la estación de trabajo / % de variabilidad máxima en HPC	Ganancia máxima por variabilidad de la estación de trabajo
Mosca de la fruta	4.49/2.26	1.98
Pulga de agua	22.45/10.68	2.10
Camarón blanco	17.34/6.12	2.83

El uso de memoria en las plataformas W_1 y W_2 se vio limitado por la capacidad física de memoria de las plataformas, inclusive los ensamblajes de mosca de la fruta tendieron a saturar el primer módulo (Fig. 9a), y se observaron picos máximos de memoria en al menos 2 módulos al procesar la pulga de agua

(Fig. 9b). En la tabla VI se muestra la utilización máxima de memoria por configuración, especie y módulo de Trinity.

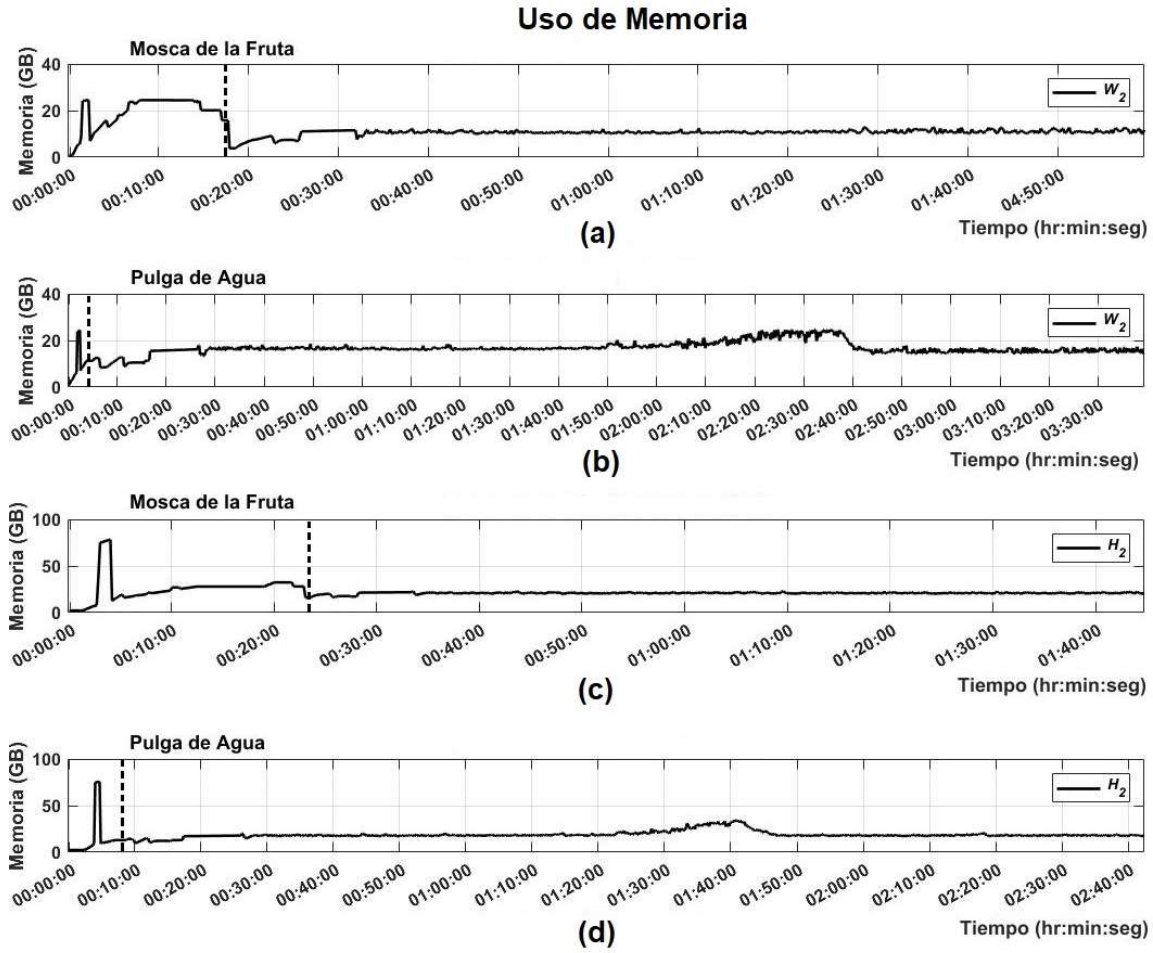


Figura 9. Uso de memoria RAM. Memoria usada durante el ensamblaje en la plataforma V_2 de mosca de la fruta en las configuraciones (a) y pulga de agua (b); ensamblaje en la plataforma H_2 de mosca de la fruta (c) y pulga de agua (d). Línea vertical intermitente indica fin del módulo *Inchworm*.

7.2 Ensamblaje bajo condiciones iniciales distintas

Se describen en esta sección los resultados de ensamblaje de lecturas de ratón y mosca de la fruta generados con distintas longitudes de *k*-mero. Posteriormente, se reportan sus métricas tradicionales y los resultados de los mapeos de los ensambles

a los transcriptomas de referencia. Los resultados de los ensambles de camarón blanco generados con variación de *k*-mero se abordan en la sección 7.6.2.1.

Tabla VI. Uso máximo de memoria RAM por plataforma computacional, organismo y módulo de ensamblaje de Trinity.

	Mosca de la fruta		**Pulga de agua**	
Plataforma Computacional	Inchworm[1]	*Chr y Btf*[2]	Inchworm[1]	*Chr y Btf*[2]
$\boldsymbol{W_1}$	20.4	9.7	20.4	20.4
$\boldsymbol{W_2}$	24.4	12.8	24.5	24.5
$\boldsymbol{H_1}$	21.1	13.0	26.0	27.8
$\boldsymbol{H_2}$	78.3	23.3	75.7	34.3

[1]*Inchworm*: Primer módulo del ensamblador. [2]Chr y Btf: Segundo y tercer módulos del ensamblador Trinity, *Chrysalis y Butterfly* respectivamente. Unidades de memoria en gigabytes (GB).

7.2.1 Ensamblaje *de novo* variando longitud de *k*-mero y valores de métricas tradicionales

Se generaron quince ensambles para ratón y mosca de la fruta, cinco por cada longitud de *k*-mero de 21, 25 y 31 bases, por medio de metodologías estándar de ensamblaje *de novo* de transcriptoma. En esta sección se indican los resultados de ensamblaje y mapeo de lecturas de entrada, lo cual proporciono los valores de métricas tradicionales de ensamblaje (Tabla VII). Nótese que las métricas tradicionales no coinciden entre sí. Por ejemplo, haber usado una mayor cantidad de lecturas RNA-Seq (Cobertura) no siempre coincide con tener el ensamble de mayor longitud, así como tener la mayor longitud de ensamble no asegura que se hayan generado una mayor cantidad de *contigs*; por su parte N50 no presenta algún patrón de coincidencia con otras métricas.

Considerando la recomendación de los desarrolladores del ensamblador de ~1 GB de memoria por cada millón de lecturas de entrada, la configuración de cómputo

H_2 se asumió adecuada para las necesidades del proceso. Los análisis en esta sección no toman en cuenta efectos de los tiempos de procesamiento o uso de memoria, por lo tanto, en este punto no son reportados.

Tabla VII. Métricas tradicionales.

	Ensamblaje	*Contigs* (Transcritos Trinity)	*Contigs* – isoformas (Genes Trinity)	Longitud de ensamble	N50	Cobertura (%)
RATÓN	***k*21**	105,674.60 14.88	105,650.40 15.81	57,711,214.60 40,237.09	832.60 0.89	76.26 0.03
	***k*25**	101,955.20 21.43	99,398.40 17.53	62,511,141.80 95,437.16	1,094.60 3.36	76.16 0.02
	***k*31**	78,754.00 58.73	77,749.00 59.23	45,511,396.80 991,328.55	1,009.80 2.58	74.39 0.05
	Promedio	95,461.27	94.265.93	55,911,336.07	979.00	75.60
MOSCA	***k*21**	23,488.00 4.89	23,473.60 3.97	13,392,836.60 4,384.17	734.40 0.54	63.83 0.01
	***k*25**	25,984.40 5.687	25,453.60 4.09	14,301,722.80 4,786.49	686.60 0.54	64.29 0.004
	***k*31**	23,467.40 7.12	23,297.40 4.92	11,643,141.00 6,662.42	581.80 0.44	62.62 0.01
	Promedio	24,313.27	24,073.80	13,112,566.53	667.60	63.58

Promedios y desviaciones estándar. Valores máximos indicados en gris.

7.3 Calidad de ensamble

La calidad de los organismos modelo analizados fueron evaluados con respecto a transcriptomas de referencia. El organismo no modelo fue evaluado con respecto a la base de datos de proteínas UniProt/Swiss-Prot. Se describen estos resultados en las secciones posteriores.

7.3.1 Calidad con respecto a transcriptomas de referencia en ensambles bajo condiciones iniciales idénticas

El mapeo a transcriptoma de referencia de mosca de la fruta y pulga de agua de los *contigs* de los conjuntos no intersectados, que representan la variabilidad por plataforma, fue mayor para los conjuntos provenientes de plataformas con menor memoria (W_1 y W_2). Las figuras 10 y 11 muestran el mapeo por plataforma

computacional para los organismos mosca de la fruta y la pulga de agua, respectivamente. También, se observan en la tabla VIII los mapeos de *contigs* intersectados con respecto a las referencias y los porcentajes que estos representan.

Las ganancias máximas de *contigs* no intersectados mapeados y ganancias máximas de *contigs* no intersectados mapeados exclusivos se presentan en la tabla IX. Obsérvese que la estación de trabajo produjo aproximadamente 2 veces más variabilidad que las plataformas HPC.

Tabla VIII. Mapeos de *contigs* con respecto a los transcriptomas de referencia.

Mosca de la fruta					
Plataforma Computacional	$\bar{I}m_{(p,mosca)}$	$\bar{I}m^{+}_{(p,mosca)}$	$\bar{I}_{(p,mosca,5)}$	$\bar{I}m_{(p,mosca)}/\bar{I}_{(p,mosca,5)}$ (%)	$\bar{I}m^{+}_{(p,mosca)}/\bar{I}_{(p,mosca,5)}$ (%)
$\boldsymbol{W_1}$	1,186	315	1,196	99.16	26.33
$\boldsymbol{W_2}$	1,054	312	1,070	98.50	29.15
$\boldsymbol{H_1}$	584	160	593	98.48	26.93
$\boldsymbol{H_2}$	583	182	594	98.15	30.63
$\boldsymbol{V_1}$	409	87	418	97.85	20.81
$\boldsymbol{V_2}$	364	86	373	97.59	23.05
Pulga de agua					
Plataforma Computacional	$\bar{I}m_{(p,pulga)}$	$\bar{I}m^{+}_{(p,pulga)}$	$\bar{I}_{(p,pulga,5)}$	$\bar{I}m_{(p,pulga)}/\bar{I}_{(p,pulga,5)}$ (%)	$\bar{I}m^{+}_{(p,pulga)}/\bar{I}_{(p,pulga,5)}$ (%)
$\boldsymbol{W_1}$	10,057	3,831	13,122	76.64	29.19
$\boldsymbol{W_2}$	10,369	4,017	13,682	75.79	29.35
$\boldsymbol{H_1}$	4,245	1,628	6,027	70.43	27.01
$\boldsymbol{H_2}$	4,327	1,567	6,050	71.52	25.90
$\boldsymbol{V_1}$	2,385	1,323	3,918	60.87	33.76
$\boldsymbol{V_2}$	2,074	1,119	3,393	61.13	32.97

$\bar{I}m_{(p,m)}/\bar{I}_{(p,m,n)}$ e $\bar{I}m^{+}_{(p,m)}/\bar{I}_{(p,m,n)}$ Expresado en porcentaje. Máximos entre plataformas W_1, W_2 y en HPC remarcados en gris.

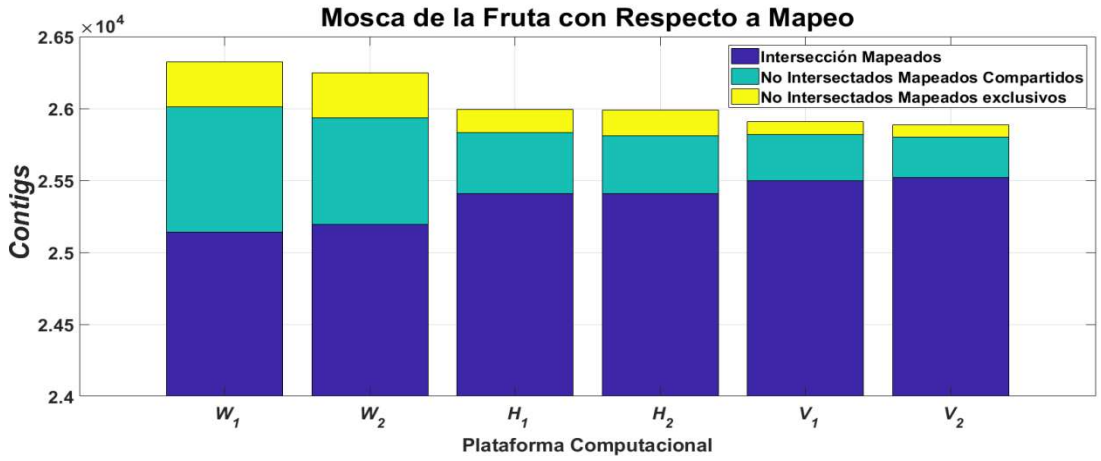


Figura 10. Comparación de los conjuntos de *contigs* intersectados mapeados de mosca de la fruta. conjuntos de *contigs* intersectados mapeados $Im_{(p,mosca)}$ al transcriptoma de referencia de la mosca de la fruta T_{mosca}, *contigs* no intersectados mapeados exclusivos $\bar{I}m^{+}_{(p,mosca)}$, y *contigs* no intersectados compartidos $\bar{I}m^{*}_{(p,mosca)}$, haciendo 5 ensamblajes $E_{(p,mosca,5)}$, por cada plataforma computacional p. Nótese que la escala inicia en 2.4x10^4.

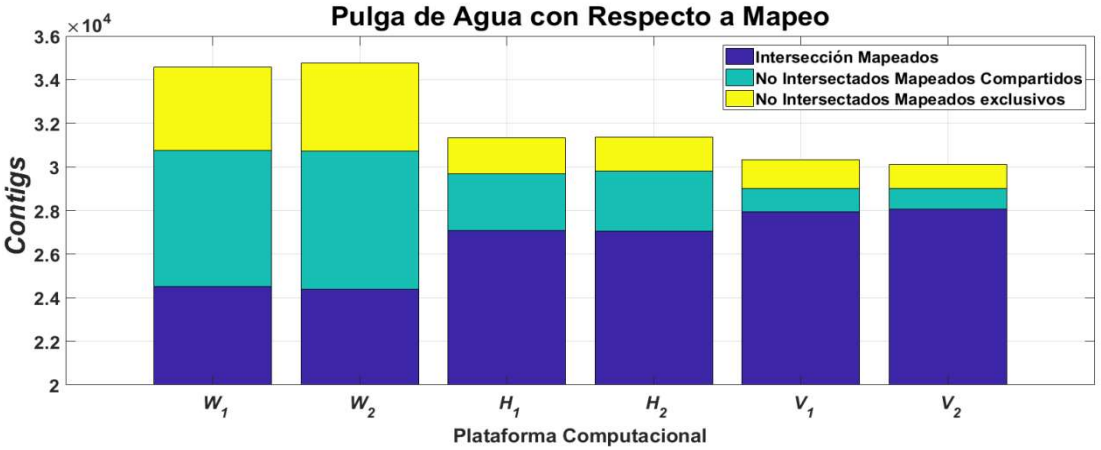


Figura 11. Comparación de los conjuntos de *contigs* intersectados mapeados de pulga de agua. Conjuntos de *contigs* intersectados mapeados $Im_{(p,pulga)}$ al transcriptoma de referencia de la pulga de agua T_{pulga}, *contigs* no intersectados mapeados exclusivos $\bar{I}m^{+}_{(p,pulga)}$ y *contigs* no intersectados compartidos $\bar{I}m^{*}_{(p,pulga)}$, haciendo 5 ensamblajes $E_{(p,pulga,5)}$, por cada plataforma computacional p. Nótese que la escala inicia en 2x10^4.

Tabla IX. Ganancias máximas de *contigs* no intersectados mapeados y ganancias máximas de *contigs* no intersectados mapeados exclusivos.

	Mosca de la fruta		**Pulga de agua**	
	Contigs	**Ganancia por *contigs* no intersectados**	***Contigs***	**Ganancia por *contigs* no intersectados exclusivos**
Máximo de *contigs* no intersectados mapeados de en estación de trabajo / Máximo de *contigs* no intersectados mapeados en HPC = Ganancia	1,186 / 584	= 2.03	10,057 / 4,327	=2.39
Máximo de *contigs* no intersectados mapeados exclusivos en la estación de trabajo / Máximo de *contigs* no intersectados mapeados en HPC = Ganancia	315 / 182	= 1.73	4,017 / 1,628	= 2.46

7.3.2 Calidad con respecto a transcriptomas de referencia en ensambles con variación de *k*-mero

Los alineamientos Bowtie2 ensamble/referencia mapearon a los grupos *k*21 de ratón y mosca con porcentajes más altos de alineamiento, siendo estos identificados como los ensambles de mayor calidad (Tabla X).

Los resultados de los mapeos BLASTN a la referencia de transcriptoma de los primeros ensambles por grupo generados con variación de *k*-mero en mosca de la fruta se abordarán en la sección 7.6.1.3.

Tabla X. Evaluación de calidad por mapeos estrictos del ensamble a la referencia.

Organismo	Ensambles	*Contigs* mapeados	Porcentaje (%)
Ratón	***k*21**	65,512.20 20.41	74.28 0.02
	***k*25**	64,163.80 13.16	72.74 0.01
	***k*31**	61,861.40 24.50	70.14 0.02
Mosca de la fruta	***k*21**	21,672.80 10.54	71.17 0.04
	***k*25**	21,114.40 2.40	69.34 0.01
	***k*31**	19,926.80 1.92	65.44 0.01

Cinco replicas por grupo de ensamble: promedio/desviación estándar. Mapeos realizados con Bowtie2. Valores máximos indicados en gris.

7.3.3 Calidad con respecto a la base de datos de proteínas UniProt/Swiss-Prot

El mapeo de los *contigs* de los conjuntos no intersectados de camarón blanco, que representan la variabilidad por plataforma, fue mayor para el conjunto proveniente de la plataforma con menor memoria (W_2). Se observan en la tabla XI los mapeos de *contigs* intersectados con respecto a la base de datos UniProt/Swiss-Prot y los porcentajes que estos representan (Fig. 12). Se observa que en la estación de trabajo es mayor la cantidad total de transcritos mapeados, al igual que la cantidad de *contigs* mapeados variables por plataforma.

La ganancia máxima de *contigs* no intersectados mapeados se presenta en la tabla XII, generando la estación de trabajo 7.20 más *contigs* variables con similitudes a la referencia de proteínas.

Los resultados de los mapeos BLASTP a la base de datos de proteínas de los ensambles generados con variación de *k*-mero en mosca de la fruta se abordarán en la sección 7.6.1.3; para camarón blanco los resultados se abordan en la sección 7.6.2.3.

Tabla XI. Mapeos de *contigs* con respecto a la base de datos UniProt/Swiss-Prot.

Camarón blanco			
Plataforma Computacional	$\bar{I}mp_{(p,cam)}$	$\bar{I}_{(p,cam,5)}$	$\bar{I}mp_{(p,cam)}$ $/\bar{I}_{(p,cam,5)}$ (%)
$\boldsymbol{W_2}$	4,604	13,093	35.16
$\boldsymbol{V_2}$	639	4,321	14.78

Tabla XII. Ganancia máxima de *contigs* no intersectados mapeados.

	Camarón Blanco	
	Contigs	**Ganancia**
Contigs no intersectados mapeados a la base de datos de proteínas provenientes de la estación de trabajo / *contigs* no intersectados mapeados a la base de datos de proteínas provenientes de la estación de trabajo	4,604 / 639	= 7.20

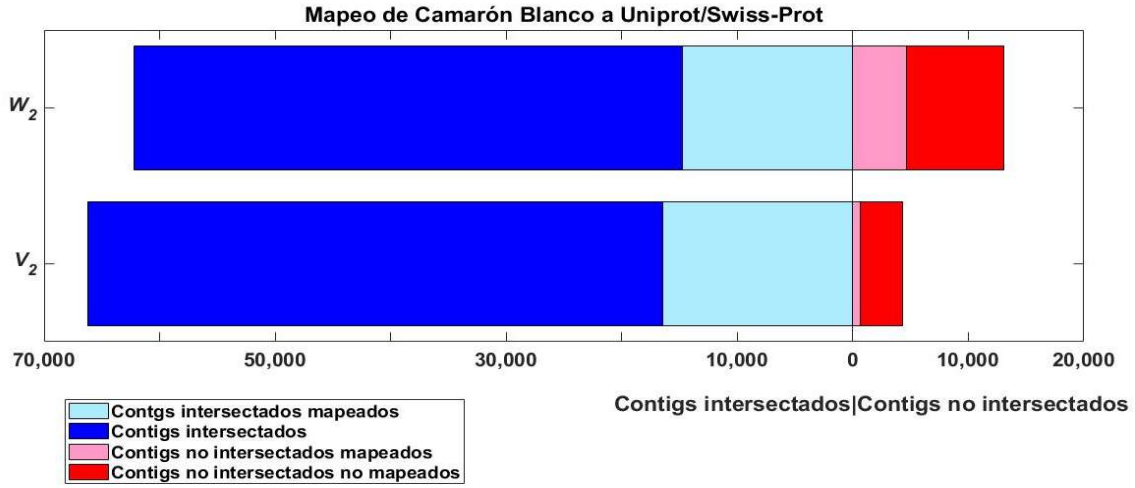


Figura 12. Mapeos de *contigs* de camarón blanco a la base de datos UniProt/Swiss-Prot.

7.4 Análisis de desempeño de métricas tradicionales de calidad

El análisis de métricas tradicionales aquí presentado se basa en las métricas de los ensambles de ratón y mosca de la fruta generados por medio de la variación de longitud de *k*-mero. Asimismo, las comparaciones de dichas métricas se hacen con respecto a los mapeos referencia/ensamble realizados con al alineador Bowtie2, configurado por medio de sus parámetros para hacer mapeos estrictos.

La métrica cuantitativa más consultada en ensamblaje es N50. Para ratón N50 indicó los ensambles *k*25 como los de mayor calidad, mientras que en mosca la métrica apuntó a los ensambles *k*21 (Tabla VII). Según la cantidad de *contigs* formados (Transcritos Trinity) y los Genes Trinity, los ensambles de mayor calidad en ratón son los *k*21 y en mosca son los *k*25. La longitud de ensamble fue mayor para los grupos *k*25 en ambos organismos. La cobertura de lecturas fue mayor para los ensambles *k*21 y *k*25 de ratón y mosca respectivamente, coincidiendo con la cantidad de transcritos y genes Trinity (Tabla VII).

El mapeo de ensamblajes a la referencia, que es la mejor evaluación posible a la fecha, indicó en ambos organismos, a los ensambles *k*21 como los de mayor calidad (Tabla X). Sin embargo, los criterios tradicionales indicaron resultados diversos, como N50, que indicó a los *k*25 como los mejores ensambles de ratón, o la longitud de ensamble, que también señala a los *k*25 como los ensambles de mayor calidad en ambos organismos. Por tanto, los criterios tradicionales no indicaron de manera congruente mayor calidad de ensamble entre sus métricas.

7.5 Uso directo de datos experimentales para evaluación de calidad y selección de ensamble

Se propone un flujo de trabajo para la selección de ensamble que consta de la selección de bases de datos de microarreglos en condiciones de experimentación similares al RNA-Seq. El flujo de trabajo con: i) la selección de bases de datos de microarreglos de condiciones de experimentación similares al RNA-Seq; ii) para continuar con ensamblaje *de novo* de transcriptoma por medio de procedimientos estándar (generación de conjuntos de ensambles) que comprendió la generación

de reportes de calidad y preprocesamiento, ensamblaje *de novo* de un conjunto de ensambles con longitudes de *k*-mero distintas, obtención de métricas tradicionales y análisis de cobertura; iii) el análisis de los datos de microarreglos (detección de sondas hibridadas); y finalmente; iv) la evaluación de calidad basada en microarreglos y selección de ensamble dentro del conjunto

Se hizo uso de los ensambles previamente generados para esta etapa del proyecto. La metodología de ensamblaje esta descrita en la sección 6.2.1 y los valores de sus métricas tradicionales en la sección 7.2.1. Se describen entonces en esta sección los resultados de los pasos iii y iv del flujo de trabajo propuesto.

7.5.1 Análisis de datos de microarreglos

Se identificaron 35,080 sondas MH a partir de los dos microarreglos de ratón. Estas provienen de la prueba de simultaneidad en 36,004 y 39,745 sondas con hibridación positiva en las dos bases de datos de microarreglos previamente seleccionadas. Los umbrales de hibridación de esas bases fueron 1,243.97 y 1,028.96 (valores de intensidad).

Los valores de intensidad que marcaron el umbral de hibridación en los datos de mosca fueron 737.00 y 748.91. Con base en estos umbrales se encontraron 62,850 y 53,555 sondas con hibridación positiva; la prueba de simultaneidad en estos datos arrojó 51,538 sondas MH.

7.5.2 Evaluación de calidad basada en microarreglos y selección de ensamble

El criterio propuesto, fundamentado en mapeos sonda MH a los ensambles, identificó a los ensambles *k*21 de ambos organismos como los de mayor calidad, y consecuentemente, respaldó su selección dentro de los conjuntos de ensambles (Tabla XIII).

Tabla XIII. Evaluación de calidad y selección de ensamble basados en microarreglos.

Organismo	Ensambles	*Contigs* Mapeados	Porcentaje (%)
Ratón	**k21**	31,537.20 12.07	89.90 0.03
	k25	31,449.60 6.14	89.65 0.02
	k31	30,918.20 15.44	88.14 0.04
Mosca de la fruta	**k21**	32,759.40 6.58	63.56 0.01
	k25	31,579.00 5.61	61.27 0.01
	k31	28,696.00 2.82	55.68 0.01

Cinco replicas por grupo de ensamblaje: promedio/desviación estándar. Valores máximos indicados en gris.

7.5.3 Verificación del criterio propuesto

Los resultados del criterio de microarreglos para la evaluación de calidad y selección de ensamble dentro de la estrategia de ensamblaje múltiple apuntaron a los ensambles *k*21 como los ensambles de mayor calidad; por lo tanto, dichos ensambles fueron indicados para selección dentro de su conjunto. La selección por medio del criterio propuesto coincide con la evaluación de calidad por referencias (mapeos estrictos Bowtie2), por tanto, respalda la selección y prueba la validez del criterio.

Las pruebas ANOVA de una vía y las pruebas consecutivas Tukey indicaron la existencia de diferencias significativas de sondas MH mapeadas entre los tres grupos de ensamblaje para ambos organismos. Ratón: $F_{2,12}$= 3,986.39, $p<0.05$. Mosca de la fruta: $F_{2,12}$= 791,553.52, $p<0.05$ (asumiendo distribuciones normales).

7.6 Generalización de datos experimentales para evaluación

Se usó evidencia experimental para el entrenamiento de Modelos Markovianos Ocultos para la evaluación de ensambles de mosca de la fruta y camarón blanco. Se describen sus resultados en las siguientes secciones.

7.6.1 Empleo de HMMs con base en microarreglos para evaluación de ensambles de mosca de la fruta

Se tomaron *probe sets* de microarreglos de expresión génica con evidencia de hibridación como evidencia experimental. Posterior a su entrenamiento con base en dichos *probe sets*, se evaluaron ensambles completos por medio de estos HMMs y se determinó cuál de los ensambles dentro de un conjunto tuvo mayor calidad. El procedimiento realizado fue verificado a través de los mapeos ensamble/transcriptoma de referencia por medio del alineador Bowtie2.

7.6.1.1 Conjunto de ensambles *de novo* de transcriptoma

Se usaron los primeros ensambles de mosca de la fruta provenientes de los grupos de ensamble con variación de *k*-mero. Las métricas tradicionales cantidad de *contigs* y cobertura apuntan al ensamble *k*25 como el mejor del conjunto (Tabla VII).

7.6.1.2 Evaluación de ensambles por medio de HMMs

Se detectaron 8,136 *probe sets* con evidencia de hibridación, los cuales fueron tomados para el entrenamiento de 8,136 HMMs. El promedio de evaluaciones máximas por ensamblaje y las evaluaciones máximas normalizadas se muestran en la tabla XIV. La evaluación por medio del conjunto de HMMs identificó al ensamble *k*21 como el de mayor calidad.

7.6.1.3 Evaluación de ensambles por medio mapeos

Los porcentajes de mapeos Bowtie2 a la referencia de transcriptoma se muestran en la tabla XIV. La evaluación con base en los modelos coincide con los mapeos al transcriptoma de referencia y al uso directo de sondas de microarreglo, identificando al ensamble *k*21 como el de mayor calidad.

Tabla XIV. Evaluación por medio de HMMs, mapeos a la referencia de transcriptoma, y mapeos a bases de datos de proteínas UniProt/Swiss-Prot.

Ensamble	*k*21	*k*25	*k*31
Probabilidad de ensamble normalizada (8,136 HMMs) [1]	-781.44	-1584.7	-2006.2
Mapeos Bowtie2 de ensamble al transcriptoma de referencia (%)	71.17	69.34	65.44
Mapeos de sondas hibridadas de microarreglo al ensamblaje (%)	63.56	61.27	55.68
Mapeos BLASTN de ensamble al transcriptoma de referencia (%)	98.73	98.88	99.14
Mapeos BLASTX de ensamble a la base de datos de proteínas UniProt/Swiss-Prot (%)	58.72	59.48	60.41

1 Valores de probabilidad expresados en logaritmo base 10

7.6.2 Empleo de HMMs con base en UniGenes para evaluación de ensambles de camarón blanco

Se tomaron como evidencia experimental secuencias contenidas en *clusters* UniGene de camarón blanco, los cuales fueron generalizados por medio de Modelos Markovianos Ocultos. Al momento del acceso a la base de datos UniGene se encontraron 68 *clusters* bajo las condiciones dictadas en la metodología (Lista completa en el Anexo F).

El uso de modelos permitió la evaluación de ensambles completos y posteriormente determinar cuál de los ensambles dentro de un conjunto tiene mayor calidad. El procedimiento realizado fue verificado por medio de mapeos a bases de datos de proteínas, coincidiendo la evaluación con base en los HMM con el mayor porcentaje de mapeo a la base de datos UniProt/Swiss-Prot, identificando el ensamblaje *k*31 de camarón blanco como el de mayor calidad.

7.6.2.1 Conjunto de ensambles *de novo* de transcriptoma

Se generaron tres ensambles *de novo* a partir de los datos de camarón blanco variando la longitud de *k*-mero. Las métricas tradicionales N50 y cantidad de *contigs* apuntan al ensamble *k*25 como el mejor del conjunto (Tabla XV).

7.6.2.2 Evaluación de ensambles por medio de HMMs

En 32 de los 68 *clusters* UniGene se detectaron al menos 20 secuencias sin bases indeterminadas, por lo cual se usaron para entrenar modelos. El promedio de evaluaciones máximas por ensamblaje y evaluaciones máximas normalizadas se muestra en la tabla XV, encontrando la evaluación máxima por conjunto en el ensamble *k*31, lo cual coincide con el mapeo a la base de datos de proteínas.

7.6.2.3 Evaluación de ensambles por medio mapeos

Se efectuaron los mapeos de ensambles a UniProt/Swiss-Prot, identificando por medio de éstos al ensamble *k*31 como el de mayor semejanza a la base de datos de proteínas, por tanto, como el ensamble de mayor calidad dentro del conjunto (Tabla XV).

Tabla XV. Métricas y evaluaciones de los ensambles de camarón blanco por medio de HMMs y mapeos a la base de datos de proteínas UniProt/Swiss-Prot.

	Ensamble		
Métrica	***k*21**	***k*25**	***k*31**
Longitud mínima de *contig*	201	201	201
Longitud máxima de *contig*	15,875	13,490	13,424
Longitud promedio de *contig*	541.23	547.50	530.42
N50	678	696	649
Cantidad de *contigs*	56,678	68,126	63,739
Contigs mapeados	13,517	16,800	16,828
Porcentaje de mapeo (%)	23.84	24.66	26.40
Probabilidad normalizada de ensamble[1]	-658.95	-659.6003	-645.1959

1 Valores de probabilidad expresados en logaritmo base 10. Valores máximos marcados en gris.

8. DISCUSIÓN

A pesar de que los transcritos de los cual proviene una biblioteca de lecturas RNA-Seq es finito, no existe un ensamble único para dichas lecturas. Diversos ajustes de condiciones iniciales ya sean burdos o finos, pueden ser hechos al momento de generar ensambles. Un ajuste burdo puede constar desde la selección de ensamblador hasta las estrategias de preprocesamiento. Un ajuste fino conlleva la variación de parámetros internos del ensamblador que están relacionados directamente con el algoritmo de ensamblaje (DBG), como los números mínimos y máximos de tanto de entrada como de salida de vértices a los nodos, o como la variación de la longitud de *k*-mero.

8.1 Ensamblaje bajo condiciones iniciales iguales en distintas plataformas: repetibilidad, variabilidad y efectos de los equipos de cómputo

La disponibilidad de memoria de las plataformas de cómputo es fija y está dada por la cantidad de RAM del equipo. Sin embargo, el usuario puede asignar los parámetros del *software* para establecer distintas limitantes al uso de memoria para el ensamblaje. El parámetro de uso de memoria en los ensamblajes de organismos modelo se configuró para ser mayor que los límites teóricos, que fueron calculados según la relación cantidad de lecturas por GB de memoria dada por el *software Trinity* (~7.6, y ~7.2 GB, para mosca de la fruta y pulga de agua respectivamente, mostrados en la Tabla I). No obstante, se observa en la figura 9 que los procesos usaron más del doble o el triple de la memoria asignada, revelando que el ensamblador tiende a utilizar cuanta RAM esté disponible para realizar sus procesos, independientemente de los valores asignados al parámetro. Por ejemplo, según la asignación de memoria para H_2, el uso debió haber sido menor a 64 GB, pero excedió en más de 11 GB en ambos organismos (ver Tabla VI). Los límites reales fueron por tanto 20 y 24 GB para plataformas menores (estación de trabajo: W_1 y W_2) y 128 GB para HPC (HPC: V_1, V_2, H_1 *y* H_2).

Aun sin hacer cambios, ensamblar en distintas ocasiones un mismo conjunto de lecturas en el mismo equipo y condiciones iniciales generó ensambles con

distinto nivel de variabilidad (Tabla IV y las figuras 6, 7 y 8), y de igual manera al cambiar la disponibilidad de memoria de la plataforma, la variabilidad cambia. Las desviaciones estándar en la cantidad de *contigs* indican muy poca variación en el número de secuencias ensambladas ($\sigma \min - \sigma max$: 2.3-6.7, 15.05-39.9 y 26.74-29.87 para mosca de la fruta, pulga de agua y camarón blanco respectivamente), en general, pero tienden a ser mayores en la estación de trabajo (Tabla III).

Según el estudio de contenido de los *contigs* hay más repetibilidad en los ensambles procesados en HPC, que tienen mayor disponibilidad de memoria (porcentajes mínimos y máximos de repetibilidad: 95.51 - 98.58, 77.55 - 93.85% y 82.91 - 93.87 para mosca de la fruta, pulga de agua y camarón blanco respectivamente), tal como se muestra en las figuras 6 y 7, y la tabla IV. Inversamente, la estación de trabajo tuvo más variabilidad, al tener menos memoria (porcentajes mínimos y máximos de variabilidad: 2.26 - 4.49, 10.68 - 22.45 y 6.12 - 17.34 para mosca de la fruta, pulga de agua y camarón blanco respectivamente) como se indica en la tabla IV. En las configuraciones V_1 y H_1 se asignó el mismo valor (24 GB) al parámetro el uso de memoria que en configuración W_2; no obstante, al no ser este parámetro una limitante real, la repetibilidad en HPC fue mayor. Para el caso de de V_1 y H_1, donde ambas tenían 128GB de memoria disponible se esperaban repetibilidades similares; no obstante, en H_1 fue menor, sugiriendo que, aparte de la disponibilidad de memoria, otros recursos computacionales pueden tener influencia en ensamblaje, tales como características intrínsecas del procesador o la arquitectura del sistema de cómputo (monolítico[4], distribuido[5], etc.). Éste y otros aspectos, como el número de núcleos de procesamiento, necesitan ser estudiados para ampliar el conocimiento de los efectos computacionales en ensamblaje.

---

[4] Sistema monolítico: se refiere a un único equipo de cómputo.

[5] Sistema distribuido: un sistema distribuido de cómputo consta de múltiples componentes de *software* que están localizados en múltiples computadoras, pero que corren como un solo sistema.

Las plataformas con menos memoria produjeron un conjunto unión mayor de *contigs* ($Ctotal_{(p,m,n)}$). Consecuentemente, se determinó por medio de mapeos a las referencias codificantes de los organismos, si estas secuencias, producto de un algoritmo computacional, mostraron correspondencia a transcritos reales. Los mapeos de los *contigs* no intersectados de los organismos modelo, que representan la variabilidad producida por la plataforma de cómputo, fueron mayores para los conjuntos generados en la estación de trabajo; porcentajes mínimos y máximos de mapeo por variabilidad de plataforma: 97.59 – 99.16 y 60.87 – 76.64 para mosca de la fruta y pulga de agua respectivamente (Tabla VIII, y figuras 10 y 11). Los mapeos de los conjuntos no intersectados exclusivos por plataforma $\bar{I}m^{+}_{(p,m)}$ son similares para ambos organismos: ~30%, estos representan la proporción de información validada generada de manera exclusiva por una plataforma dada (Tabla VIII). En ambos casos, variabilidad total y exclusiva, los mapeos fueron aproximadamente dos veces mayores en los *contigs* variables generados en la estación de trabajo en comparación con HPC (Tabla IX).

Gran porcentaje de los *contigs* obtenidos en los ensambles de ambos organismos corresponden a transcritos reales. No obstante, algunos *contigs* no lograron ser mapeados a sus referencias, pero esto no significa que las secuencias no mapeadas sean incorrectas, simplemente dichos *contigs* pueden ser parte de transcritos que aún no han sido descubiertos por falta de conocimiento sobre la biología molecular del organismo, y, por tanto, no están contenidos en la referencia.

En el caso de camarón blanco, a falta de disponibilidad de un transcriptoma de referencia, la verificación de calidad se realizó comparando los ensambles con la base de datos de proteínas de alta calidad UniProt/Swiss-Prot (The UniProt Consortium, 2017). Los porcentajes de mapeos de *contigs* no intersectados de camarón blanco, que representan el mapeo de la variabilidad por plataforma, fueron de 35.16% y 14.78% para W_2 y V_2 respectivamente (Tabla XI). Si bien los porcentajes de mapeos son bajos, la base de datos de proteínas está relacionada a secuencias de diversas especies y organismos distintos a camarón blanco, que, siendo más estudiados, están representados con mayor proporción. En la estación

de trabajo se generaron 7.20 veces más *contigs* no interceptados mapeados $\bar{I}mp_{(p,cam)}$ en comparación con HPC (Tabla XII).

Dados estos resultados, se evidenció que la variación de un ensamblaje está dada en función de la disponibilidad de memoria del equipo de cómputo: a mayor disponibilidad de memoria hay menor variación en ensamblaje y a menor disponibilidad de memoria hay mayor variación en ensamblaje.

Una de las principales ventajas del RNA-Seq es la capacidad de descubrimiento de nuevos transcritos (Korf, 2013). Se sugiere el emplear una estrategia de ensamblaje *de novo* múltiple de transcriptoma para el descubrimiento de una mayor cantidad de transcritos, la cual consiste en la obtención de varios ensambles bajo condiciones iniciales iguales en plataformas computacionales con baja disponibilidad de memoria, pero viables para proceso. Posteriormente, realizar la unión de *contigs* de múltiples ensambles, ya que se logra obtener conjuntos más grandes de *contigs* de alta calidad, como fue realizado en los *contigs* obtenidos en la estación de trabajo.

8.2 Ensamblaje bajo condiciones iniciales distintas: análisis de métricas tradicionales

Diversas métricas estadísticas buscan dar un indicativo de la calidad de ensamble con respecto al transcriptoma original, pero éstas métricas no muestran indicios cuantitativos del desempeño del equipo de cómputo o su influencia en el conjunto de *contigs*. La siguiente etapa del proyecto se enfocó en explorar cambios en ensambles *de novo* debido a la variación de condiciones iniciales en dos organismos modelo, ratón y mosca de la fruta, y explorar con base en dichos cambios el desempeño de diversas métricas tradicionales de calidad. En este caso, el parámetro de cambio seleccionado fue la longitud de *k*-mero porque tiene influencia directa sobre la sensibilidad y especificidad de ensamblaje (Zerbino y Birney, 2008). Cabe destacar que en esta etapa no se hicieron intersecciones o uniones de *contigs*, ya que se procuró identificar exclusivamente la influencia del parámetro sobre el ensamblaje.

Se obtuvieron 5 ensambles por longitud de *k*-mero mínima, estándar y máxima (*k*21, *k*25 y *k*31) en ambos organismos. Los mapeos a transcriptomas de referencia, que es la mejor evaluación de calidad disponible al momento, mostraron que los ensambles *k*21 se asemejan más a la referencia codificante, mapeando el 74.28% y 71.17% de los *contigs* de ratón y mosca de la fruta a su respectiva referencia (Tabla X).

Las métricas a las que más se recurre para evaluar la calidad de ensamble *de novo* de transcriptoma, entre ellas N50, revelaron ser inconsistentes con respecto a la referencia, al identificar distintos grupos de ensambles en ambos organismos como los de mayor calidad (Tabla VII). En ratón, tres de las cinco métricas tradicionales cuantificadas sí coincidieron con el mapeo (*contigs* totales, *contigs* sin isoformas y cobertura). Sin embargo, la métrica más recurrida, N50, no coincidió, al igual que longitud de ensamble, ya que identificaron a los ensambles *k*25 como los de mayor calidad. Caso contrario, en mosca de la fruta, la única métrica que coincidió con el mapeo fue N50. Todas las demás identificaron incorrectamente a *k*25 como los ensambles de mayor calidad.

Longitudes mayores de *k*-mero vuelve más específico al algoritmo de ensamble (DBG), pero también implicaría que se reduciría la capacidad de construir nuevas secuencias. Caso contrario, longitudes menores de *k*-mero potencializan la capacidad del algoritmo de construir un mayor número de secuencias, pero aumenta la probabilidad de construir secuencias que sean computacionalmente correctas pero que no tengan correspondencia a un transcrito real (Compeau *et al.*, 2011; Zerbino y Birney, 2008). Se puede observar en los ensambles generados para ambos organismos, que efectivamente la longitud máxima de *k*-mero (*k*31) correspondió con la menor cantidad de *contigs* construidos. Sin embargo, la longitud mínima (*k*21) solo en ratón coincidió con la mayor cantidad de *contigs* (Tablas VII y X). Más aún, para los conjuntos de ensambles del estudio, el parámetro no mostró una correspondencia directa con los mapeos a los transcriptomas de referencia.

Las métricas cuantitativas no son consistentes con respecto a la calidad del ensamble, no siempre coinciden con la evaluación de calidad (Brown, 2013a).

Muchas de estas métricas fueron creadas con el objetivo de evaluar ensambles de genomas. Estas métricas originalmente fueron dirigidas para evaluar ADN, no miles de secuencias de ARN representadas por los transcritos. De esta manera, estadísticos como longitud de ensamble, cobertura de lecturas, y N50 se tornan ambiguas en el contexto de evaluar la reconstrucción de un conjunto de miles de secuencias de distintas longitudes (ensamblaje de transcriptoma) (O'Neil y Emrich, 2013).

Este análisis enfatiza la necesidad de tener medios para evaluar de una manera certera la calidad de un ensamble, ya que en el típico caso *de novo*, no se cuenta con transcriptomas de referencia, y como se ha demostrado las métricas tradicionales y valores específicos de parámetros no garantizan correspondencia de las reconstrucciones con los transcriptomas reales.

8.3 Uso directo de datos experimentales para la evaluación de calidad y selección de ensamble

Como se mencionó con anterioridad, la determinación de calidad de un ensamble puede estar enfocada en estimar la similitud de la reconstrucción con respecto a secuencias de referencia de la misma especie o especies cercanas (O'Neil y Emrich, 2013). De esta manera se utilizaron microarreglos de forma directa y generalizada, y UniGenes de forma generalizada, para la evaluación y selección de ensamble al ser secuencias que proveen evidencia experimental del estado de un transcriptoma.

El uso directo de evidencia experimental en la evaluación de ensamble consiste en la implementación de un criterio de calidad auxiliar en la identificación de la mejor reconstrucción posible dentro de un conjunto de ensambles *de novo* de transcriptoma. Esta identificación fue realizada por medio de sondas de microarreglos de expresión génica con hibridación positiva (sondas MH), permitiendo la selección del ensamble con mayor calidad dentro del conjunto. La aplicación de este criterio en los conjuntos de ensambles con variación de longitud

de *k*-mero conllevó a la selección de los ensambles *k*21 de mosca y ratón (Tabla XIII).

La evaluación de calidad basada en microarreglos se verificó por medio de mapeos de los conjuntos de ensambles con respecto a sus respectivos transcriptomas de referencia. En ambos casos, ratón y mosca, los ensambles *k*21 tuvieron alineamientos más altos, siendo identificados de esta manera como las reconstrucciones con mayor calidad y validados por medio de sus respectivas referencias (Tabla X).

El criterio basado en microarreglos diverge del criterio N50 en ratón, sin embargo, coincide con la cantidad máxima de *contigs* construidos. Esto es distinto para mosca donde N50 coincide con el criterio propuesto, pero no es el caso para las demás métricas.

La diferencia entre las sondas MH mapeadas a los ensambles *k*21 y *k*25 de ratón es pequeña, solo un 0.25% del mapeo (88 sondas). Aun así, este porcentaje representó una diferencia estadísticamente significativa (Ratón: $F_{2,12}$=3,986.39, $p<0.05$. Mosca de la fruta: $F_{2,12}$=791,553.52, $p<0.05$; asumiendo distribuciones normales), la cual correspondió a 3,719 *contigs*. Cualquier diferencia en mapeo de sondas, por más pequeña que sea, sugiere mayor calidad de ensamble. En este caso, los *contigs* correspondientes a la diferencia de mapeos de sondas MH pudiesen representar secuencias de alta relevancia. En este contexto, usar las métricas tradicionales de calidad para la evaluación y posterior selección de ensamble pudiese descartar potenciales transcritos de alta relevancia.

Una de las ventajas principales del RNA-Seq con respecto a su técnica predecesora, los microarreglos de expresión génica, es la capacidad de detectar nuevos transcritos. Independientemente de que las probabilidades de encontrar nuevos transcritos aumenta al obtener mayor cantidad de transcritos, los *contigs* excedentes no reflejan necesariamente mayor calidad de ensamble. Se observa esto en los ensambles de mosca de la fruta, donde los conteos más altos de *contigs* fueron para el grupo *k*25, pero el grupo que mejor refleja al transcriptoma, indicado por el mapeo, es el *k*21 (Tabla VII).

El uso del criterio con base en evidencia experimental para la evaluación y selección de ensamble permitió la reducción de incertidumbre en comparación con los criterios cuantitativos, pero esta estrategia depende de las secuencias utilizadas para la evaluación de calidad. Esto también es aplicable al utilizar cualquier tipo de evidencia genómica o transcriptómica, por ejemplo, los mapeos de ensamble con respecto a un conjunto de proteínas o transcriptomas de referencia de especies cercanas. Por otra parte, criterios consistentes de evaluación de calidad de ensamble están basados exclusivamente en regiones conservadas, como los procedimientos BUSCO o CEGMA (Parra *et al.*, 2007; Simão *et al.*, 2015), pero estos no evalúan las porciones no conservadas de los ensambles.

El criterio basado en microarreglos está fundamentado exclusivamente en secuencias hibridadas a un transcriptoma. La hibridación permite detectar la expresión en condiciones iguales o similares a una experimentación RNA-Seq, permitiendo en si la evaluación de ensamble con base en la condición biológica de estudio. Así, la hibridación detecta exclusivamente segmentos de secuencias que están siendo expresadas en el tiempo y condición del experimento. Esto constituye la mejor ventaja del criterio propuesto.

La evaluación con respecto a sondas hibridadas implica una ventaja práctica ya que únicamente se evaluarían los ensambles con respecto a un subconjunto de secuencias, no con respecto a bases de datos completas, como transcriptomas de referencia o bases de datos de proteínas, que requieren largos tiempos de procesamiento para la búsqueda de similitudes. Así, el flujo de trabajo propuesto constituye una metodología simplificada para la evaluación de calidad y selección de ensamble dentro de la estrategia de ensamblaje múltiple.

Como se mencionó con anterioridad, hay escasa disponibilidad de transcriptomas de referencias de especies menos estudiadas, como organismos no modelo, los cuales suelen ser de interés comercial. Pero existen millones de bases de datos de microarreglos están disponibles y son de libre acceso. Dichas bases de datos fueron obtenidas empleando técnicas perfeccionadas y depositadas en repositorios públicos cumpliendo con estándares rigurosos de calidad, como MIAME

(Brazma, 2009; Brazma *et al.*, 2001; Rustici *et al.*, 2013). Estas características hacen del criterio de calidad con base en microarreglos una alternativa práctica y factible para la evaluación con base en evidencia experimental y la selección de ensamble.

8.4 Generalización de datos experimentales: HMMs con base en microarreglos para la evaluación de ensambles

Extendiendo el uso de datos experimentales para la evaluación de calidad de ensamble se implementó el uso generalizado de los datos de microarreglo y posteriormente se extendió la implementación a la generalización de UniGenes para la evaluación de calidad.

El uso generalizado por medio de HMMs entrenados a partir de *probe sets* de microarreglo con evidencia de hibridación permitió la evaluación y selección del ensamble con mayor calidad dentro de un conjunto de ensambles de mosca de la fruta. La evaluación por medio del conjunto completo de modelos entrenados, 8,136 en total, aplicada a los tres ensambles con variación en longitud de camero, identificó al ensamble *k*21 como el de mayor calidad, coincidiendo con los mapeos estrictos del transcriptoma de referencia (Bowtie2) y con el uso directo de las sondas hibridadas para la evaluación de calidad.

Las verificaciones de calidad de los organismos modelo utilizadas en los análisis de desempeños de métricas de calidad y en las inclusiones directa y generalizada de datos experimentales se realizaron con mapeos más específicos a la aproximación del transcriptoma, mapeos ensamble a referencia usando el alineador Bowtie2 con altos umbrales de alineamiento; esto proporcionó verificaciones estrictas de calidad.

Los mapeos BLASTN son estrategias de verificación que se usan de manera regular para evaluar calidad (Mundry *et al.*, 2012; O'Neil y Emrich, 2013). Sin embargo, comparándolos con mapeos estrictos que identifican mayores niveles de similitud entre secuencias, los mapeos BLASTN muestran no ser la mejor opción para evaluar calidad; esto se observa en la identificación del ensamble *k*31 por

medio del mapeo BLASTN, que fue contraria a la identificación del ensamble *k*21 del mapeo Bowtie2.

Los HMMs captaron la variabilidad de los datos de evidencia de expresión de tal forma que una vez que estos se utilizaron para la evaluación y selección de ensamble tuvieron un nivel de detección de calidad similar a los mapeos estrictos (Bowtie2), y superando el nivel de detección de mapeos BLAST.

8.5 Generalización de datos experimentales: HMMs con base en UniGenes para la evaluación de ensambles de camarón blanco

Litopenaeus vannamei es la principal especie de comercialización en México y su cultivo constituye una de las principales actividades económicas dentro de la acuacultura en nuestro país. En el ciclo 2011-2012 fueron reportados de 99 a 179 mil toneladas de producción, representando ~90% de los cultivos de los estados de Sonora, Sinaloa, Baja California Sur y Nayarit (Álvarez-Sánchez, 2017; CONAPESCA, 2013). Dada su importancia, se seleccionaron datos de este organismo para la implementación de generalizaciones para la evaluación de calidad de ensamblaje *de novo*, ya que el camarón blanco no cuenta con un transcriptoma de referencia. Este organismo tampoco cuenta con microarreglos, y aunque se pudiesen usar datos de microarreglos con hibridación heteróloga (Ruiz-Laguna *et al.*, 2016), en este estudio se utilizaron UniGenes para obtener HMMs.

Siendo los *clusters* UniGene grupos de secuencias con distinta procedencia (secuencias EST, ARNm, etc. de diversos tejidos), su empleo directo no es recomendable. Sin embargo, el uso de HMMs entrenados a partir de UniGenes permitió la evaluación y selección del ensamble con mayor calidad dentro de distintos ensambles. La aplicación de este criterio en un conjunto de ensambles con variación de longitud de *k*-mero identificó al ensamble *k*31 de camarón blanco como el de mayor calidad (Tabla XV).

Para el ensamblaje *de novo* de transcriptoma para camarón blanco solo se obtuvo un ensamble por longitud de *k*-mero. De la misma forma, la evaluación de

calidad por medio de modelos se basó en secuencias experimentales, *clusters* UniGene, que son de fácil acceso.

Dada la falta de disponibilidad de un transcriptoma de referencia para camarón blanco, no se pueden utilizar mapeos estrictos a secuencias nucleotídicas de referencia, como los mapeos Bowtie2. Consecuentemente, la verificación de calidad para este organismo se realizó mapeando los ensambles a la base de datos de proteínas UniProt/Swiss-Prot, ya que la evaluación apoyada en esta base de datos es una práctica común en la verificación de evaluación de calidad para organismos sin referencia (O'Neil y Emrich, 2013). Dichos mapeos identificaron al ensamble *k*31 como el de mayor semejanza a las secuencias de referencia (proteínas UniProt/Swiss-Prot), lo cual coincide con la evaluación con base en HMMs (Tabla XV).

Se observa nuevamente que las métricas tradicionales divergen tanto con los mapeos como con la evaluación por modelos. La cantidad de *contigs* y la longitud N50 identificaron al ensamble *k*25 como el de mayor calidad, mientras que la longitud máxima de *contig* fue encontrada en el ensamble *k*21 (Tabla XV).

Como se observó en camarón blanco, se pueden obtener modelos a partir de distintos tipos de secuencias que provean evidencia experimental. Adicionalmente, el entrenamiento de cada modelo no requirió de muchas secuencias. Se identificó el mejor ensamble empleando 32 HMMs entrenados con 20 secuencias de 25 bases por *cluster*, y las evaluaciones por medio de estos modelos coincidieron con mapeos BLAST, que son comúnmente empleados para evaluar ensamblaje o realizar anotaciones (Moreton *et al.*, 2016). Esto representa uno de los principales beneficios de la generalización, ya que en organismos no modelo hay poca disponibilidad de evidencia experimental y esta pudiese no ser suficiente para su empleo directo en los flujos de trabajo de ensamblaje *de novo* de transcriptoma.

El generar modelos requiere del preprocesamiento de los datos de evidencia experimental. Sin embargo, los HMM empleados para la evaluación de ensamblaje son relativamente sencillos. Éstos pueden ser obtenidos por medio de paqueterías preprogramadas para entrenamiento y evaluación, como las aplicaciones de

cómputo científico R y Matlab (Crawley, 2012; Murphy, 2005; MathWorks, 2013), o por medio de lenguajes de programación comunes.

9. CONCLUSIONES

1. El uso de evidencia experimental ya sea de manera directa o generalizada, mostró tener una mejor correspondencia a mapeos estrictos que las métricas tradicionales, e inclusive, mejor correspondencia que mapeos de referencia por métodos estándar de alineamiento (BLAST).
2. La evaluación por mapeos con alineamiento estricto a transcriptomas de referencia, demostraron que las métricas tradicionales, de carácter cuantitativo, son malos indicadores de la calidad de un ensamble.
3. Se sugiere el uso de evidencia experimental, ya sea de manera directa o indirecta, por medio de Modelos Markovianos Ocultos, para la evaluación y selección de ensamblajes *de novo* múltiples.
4. Las evaluaciones por medio de HMMs son indicadores eficientes de calidad y mostraron ser factibles y aplicables a los organismos estudiados.
5. Se demostró que dos o más ensambles *de novo* de transcriptoma generados en condiciones iniciales idénticas presentan diferencias en su contenido, y que estas diferencias no son reflejadas en las métricas cuantitativas, incluyendo número de *contigs* o N50.
6. Se demostró que el contenido de los ensambles también depende de las características del equipo de cómputo, explorándose la dependencia de la disponibilidad de memoria.
7. La variabilidad de *contigs* en los ensambles *de novo* de transcriptoma fue mayor en plataformas computacionales con menor disponibilidad de memoria. No obstante, los *contigs* extra originados por dicha variación mostraron tener correspondencia con transcriptomas de referencia. Por tanto, se proponen realizar uniones de múltiples ensambles procesados bajo las mismas condiciones iniciales en plataformas de baja, pero, suficiente memoria, para el descubrimiento de una mayor cantidad de *contigs*.
8. El ensamblaje *de novo* de transcriptoma es una etapa clave en estudios exploratorios del contenido de ARN en muestras de organismos sin referencia. Las implementaciones realizadas en este proyecto permiten mejorar criterios de

evaluación de calidad y selección ensamble para potencializar el descubrimiento de transcritos, haciendo uso de bases de datos de libre acceso y de una mejor selección de recursos computacionales.

9. Dada la eficacia de los HMMs con base en evidencia experimental, se sugiere a futuro, emplear dichos modelos dentro del algoritmo de ensamblaje, lo cual implicaría la modificación del *software* ensamblador.

9.1 Contribuciones

Este proyecto proporciona nuevos mecanismos que permiten mejorar la evaluación de calidad y selección de ensamble en el contexto de ensamblaje múltiple. Los mecanismos generan indicadores sustentados en evidencia experimental, con mejor desempeño que criterios de selección anteriores, como la métrica N50 que había sido la más recomendada para selección. De esta manera los mecanismos propuestos permiten:

1. Facilitar el uso de evidencia experimental para determinación de calidad mediante:
    - Microarreglos como evidencia experimental para la evaluación de calidad de ensamblaje *de novo* de transcriptoma.
    - Modelos Markovianos Ocultos como generalización de evidencia experimental para la evaluación de calidad de ensamblaje *de novo* de transcriptoma.
2. Provee criterios de evaluación y selección de ensamble aplicable dentro de la estrategia de ensamblaje múltiple.

Este proyecto también proporciona por primera vez reportes detallados sobre la influencia de los equipos de cómputo en los resultados finales de un proceso de ensamblaje *de novo* de transcriptoma y da directivas para su aprovechamiento. De esta manera:

1. Se da evidencia que no hay un ensamble único para un conjunto de lecturas de secuenciación que se presenten como datos de entrada a un ensamblador.
2. Se provee evidencia de que los ensambles dependen de la selección de equipo de cómputo, demostrando en específico la influencia de la memoria: entre más baja sea la disponibilidad de memoria, mayor será la variabilidad de ensamblaje.
3. Propone el aprovechamiento de plataformas computacionales de baja capacidad de memoria para incrementar el descubrimiento de *contigs*.

9.2 Trabajo futuro

1. Analizar el efecto de los criterios propuestos en el análisis de secuencias no codificantes.
2. Incluir los HMMs dentro del algoritmo de ensamblaje, lo cual daría una alternativa a la resolución de ambigüedades en la construcción de contigs. Consecuentemente se requeriría la modificación del código del *software* ensamblador.
3. Explorar formas adicionales de generalización.
4. Explorar los efectos sobre el proceso de ensamble de otras variables de la plataforma de cómputo. Por ejemplo: cantidad de núcleos de procesamiento, arquitectura del procesador, virtualización, administración de procesos, entre otros.
5. Reducir el tiempo de procesamiento de las evaluaciones por medio de HMMs empleando estrategias de procesamiento de cómputo paralelo.
6. Analizar el potencial beneficio del uso de evidencia experimental para evaluación de calidad en el contexto de ensamblaje por referencia.
7. Explorar en laboratorio los productos finales de los procesos computacionales.

10. LITERATURA CITADA

Álvarez-Sánchez, A. R. 2017. Construcción de una cepa de levadura productora de dsRNA específico conta el virus de la mancha blanca en camarón *Lithopenaeus vannamei* (Doctorado en Ciencias en el Uso Manejo y Preservación de los Recursos Naturales). La Paz, B.C.S., México. Centro de Investigaciones Biológicas del Noroeste, S.C. 3 p.

Andrews, S. 2015. FastQC A quality control tool for high-throughput sequence data. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc. Último acceso: Abril 04, 2018.

Atanur, S. S., A. G. Diaz, K. Maratou, A. Sarkis, M. Rotival, L. Game, M. R. Tschannen, P. J. Kaisaki, G. W. Otto, M. C. J. Ma, T. M. Keane, O. Hummel, K. Saar, W. Chen, V. Guryev, K. Gopalakrishnan, M. R. Garrett, B. Joe, L. Citterio, G. Bianchi, M. McBride, A. Dominiczak, D. J. Adams, T. Serikawa, P. Flicek, E. Cuppen, N. Hubner, E. Petretto, D. Gauguier, A. Kwitek, H. Jacob, T. J. Aitman. 2013. Genome sequencing reveals *loci* under artificial selection that underlie disease phenotypes in the laboratory rat. Cell 154:691-703.

Barrett, T., S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, A. Soboleva. 2013. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res. 41:D991-D995.

Bateman, A., M. J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, L. G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. Macdougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, A. Cowley, L. Figueira, W. Li, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, B. E. Suzek, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, M. S. Yerramalla, J. Zhang. 2015. UniProt: A hub for protein information. Nucleic Acids Res. 43:D204-212.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, A. J. Smith. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59.

Boguski, M. S., T. M. J. Lowe, C. M. Tolstoshev. 1993. dbEST - database for "expressed sequence tags". Nat Genet 4:332-333.

Bolger, A. M. A., M. Lohse, B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Boutet, E., D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A. J. Bridge, S. Poux, L. Bougueleret, I. Xenarios. 2016. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View BT. En: Edwards, D. Plant Bioinformatics: Methods and Protocols. Springer New York, New York, NY. E.U.A. pp 23-54.

Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. a Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.-C. Chou, J. Corbeil, C. Del Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. a Fonseca, G. Ganapathy, R. a Gibbs, S. Gnerre, E. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T.-W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. Maccallum, M. D. Macmanes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, I. F. Korf. 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. Gigascience 2:10.

Brazma, A. 2009. Minimum Information About a Microarray Experiment (MIAME) - Successes, Failures, Challenges. Sci. World J. 9:420-423.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. a Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, a Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, M. Vingron. 2001. Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. Nat. Genet. 29:365-371.

Brown, S. M. 2013a. *De novo* assembly of bacterial genomes from short sequence reads. En: Brown, S.M. (1ra edición). Next-Generation DNA Sequencing Informatics. Cold Spring Harbor Laboratory Press, New York, New York, E.U.A. pp 97-109.

Brown, S. M. 2013b. History of sequencing informatics. En: Brown, S.M. (1ra edición). Next-Generation DNA Sequencing Informatics. Cold Spring Harbor Laboratory Press, New York, New York, E.U.A. pp 27-44.

BUAP. 2017. Laboratorio Nacional de Supercómputo del Sureste de México. http://www.lns.buap.mx. Último Acceso: Abril 04, 2017

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:1.

Causton, H., J. Quackenbush, A. Brazma. 2003. Microarray Gene Expression Data Analysis: A Beginner's Guide. Wiley-Blackwell. Primera edición. Malden, MA., E.U.A. 160p.

Chapman, J. A., I. Ho, S. Sunkara, S. Luo, G. P. Schroth, D. S. Rokhsar. 2011. Meraculous: *De novo* genome assembly with short paired-end reads. PLoS One. 6:e23501.

Chial, H. 2008. DNA sequencing technologies key to the Human Genome Project. Nat. Educ. 1:219.

Churchill, G. A. 1989. Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. 51:79-94.

Clarke, K., Y. Yang, R. Marsh, L. Xie, K. K. Zhang. 2013. Comparative analysis of *de novo* transcriptome assembly. Sci. China Life Sci. 56:156-162.

Compeau, P. E. C., P. a Pevzner, G. Tesler. 2011. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 29:987-991.

CONAPESCA. 2013. CONAPESCA. www.conapesca.sagarpa.gob.mx. Último acceso: Marzo 30, 2018.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, A. Mortazavi. 2016. A survey of best practices for RNA-Seq data analysis. Genome Biol. 17:13.

Coordinators, N. R. 2013. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 41:D8-D20.

Crawley, M. J. 2012. The R Book. Wiley Publishing. Segunda edición. West Sussex, Inglaterra. 1076p.

Daines, B., H. Wang, L. Wang, Y. Li, Y. Han, D. Emmert, W. Gelbart, X. Wang, W. Li, R. Gibbs, R. Chen. 2011. The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing. Genome Res. 21:315-324.

Doroszuk, A., M. J. Jonker, N. Pul, T. M. Breit, B. J. Zwaan. 2012. Transcriptome analysis of a long-lived natural *Drosophila* variant: a prominent role of stress and reproduction genes in lifespan extension. BMC Genomics. 13:167.

Duan, J., C. Xia, G. Zhao, J. Jia, X. Kong. 2012. Optimizing *de novo* common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genomics. 13:392.

Durai, D. A., M. H. Schulz. 2016. Informed *k*-mer selection for *de novo* transcriptome assembly. Bioinformatics 32:1670-1677.

Durbin, R., S. R. Eddy, A. Krogh, G. Mitchison. 1998. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press. Primera edición. Cambridge, UK. 357p.

Earl, D., K. Bradnam, J. St John, A. Darling, D. Lin, J. Fass, H. O. K. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, N. Nguyen, P. N. Ariyaratne, W.-K. Sung, Z. Ning, M. Haimel, J. T. Simpson, N. a Fonseca, İ. Birol, T. R. Docking, I. Y. Ho, D. S. Rokhsar, R. Chikhi, D. Lavenier, G. Chapuis, D. Naquin, N. Maillet, M. C. Schatz, D. R. Kelley, A. M. Phillippy, S. Koren, S.-P. Yang, W. Wu, W.-C. Chou, A. Srivastava, T. I. Shaw, J. G. Ruby, P. Skewes-Cox, M. Betegon, M. T. Dimon, V. Solovyev, I. Seledtsov, P. Kosarev, D. Vorobyev, R. Ramirez-Gonzalez, R. Leggett, D. MacLean, F. Xia, R.

Luo, Z. Li, Y. Xie, B. Liu, S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, S. Yin, T. Sharpe, G. Hall, P. J. Kersey, R. Durbin, S. D. Jackman, J. a Chapman, X. Huang, J. L. DeRisi, M. Caccamo, Y. Li, D. B. Jaffe, R. E. Green, D. Haussler, I. Korf, B. Paten. 2011. Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. Genome Res. 21:2224-2241.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, A. Bateman. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279-D285.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-silva, P. Clapham, G. Coates, L. Gordon, T. Hourlier, S. Fitzgerald, L. Gil, C. Garcı, S. Hunt, N. Johnson, T. Juettemann, A. K. Ka, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. Mclaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. P. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino, S. M. J. Searle. 2014. Ensembl 2014. Nucleic Acids Res. 42:749-755.

Frith, M. C., M. Pheasant, J. S. Mattick. 2005. Genomics: The amazing complexity of the human transcriptome. Eur J Hum Genet 13:894-897.

Ghaffari, N., O. Arshad, H. Jeong, J. Thiltges, M. Criscitiello, B.-J. Yoon, A. Datta, y C. Johnson. 2015. Examining *de novo* transcriptome assemblies via a quality assessment pipeline. IEEE/ACM Trans. Comput. Biol. Bioinforma. 15(2):749-755.

Gilbert, W., A. Maxam. 1973. The nucleotide sequence of the lac operator. Proc. Natl. Acad. Sci. U. S. A. 70:3581-3584.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644-652.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman, A. Regev. 2013a. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494-1512.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman, A. Regev. 2013b. *De novo* transcript sequence reconstruction

from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494-1512.

Henschel, R., P. M. Nista, M. Lieber, B. J. Haas, L.-S. Wu, R. D. LeDuc. 2012. Trinity RNA-Seq assembler performance optimization. Proc. 1st Conf. Extrem. Sci. Eng. Discov. Environ. Bridg. from Extrem. to campus beyond - XSEDE '12:8.

Huang, X., X.-G. Chen, P. A. Armbruster. 2016. Comparative performance of transcriptome assembly methods for non-model organisms. BMC Genomics 17:523.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931-945.

Kolesnikov, N., E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, A. Brazma. 2015. ArrayExpress update-simplifying data submissions. Nucleic Acids Res. 43:D1113-D1116.

Korf, I. 2013. Genomics: the state of the art in RNA-Seq analysis. Nat. Methods 10:1165-1166.

Krebs, J. E., E. S. Goldstein, S. T. Kilpatrick. 2014. Lewin's Genes. Jones & Bartlett Learning. Decimoprimera edición. Burlington, Mass. E.U.A. 940p.

Langmead, B., S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat Meth 9:357-359.

Leinonen, R., R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, G. Hoad, M. Jang, N. Pakseresht, S. Plaister, R. Radhakrishnan, K. Reddy, S. Sobhany, P. Ten Hoopen, R. Vaughan, V. Zalunin, G. Cochrane. 2011a. The European Nucleotide Archive. Nucleic Acids Res. 39:D28-D31.

Leinonen, R., H. Sugawara, M. Shumway, Collaboration on behalf of the I. N. S. D. 2011b. The Sequence Read Archive. Nucleic Acids Res. 39:D19-D21.

Leipzig, J. 2017. A review of bioinformatic pipeline frameworks. Brief. Bioinform. 18:530-536.

Liang, K., X. Wang, D. Anastassiou. 2007. Bayesian Basecalling for DNA Sequence Analysis Using Hidden Markov Models. IEEE/ACM Trans. Comput. Biol. Bioinforma. 4:430-440.

Lin, Y., J. Li, H. Shen, L. Zhang, C. J. Papasian, H.-W. Deng. 2011. Comparative studies of *de novo* assembly tools for next-generation sequencing technologies. Bioinformatics. 27:2031-2037.

Liu, J., J. Gough, B. Rost. 2006. Distinguishing protein-coding from non-coding RNAs through support vector machines (J Blake, J Hancock, B Pavan, L Stubbs, y PlGEICW Frankel, Eds.). PLoS Genet. 2:e29.

Lottaz, C., C. Iseli, C. V Jongeneel, P. Bucher. 2003. Modeling sequencing errors by combining Hidden Markov models. Bioinformatics. 19(2):ii103-112.

Mardis, E. R. 2008. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9:387-402.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Marinković, M., W. C. de Leeuw, M. de Jong, M. H. S. Kraak, W. Admiraal, T. M. Breit, M. J. Jonker. 2012. Combining next-generation sequencing and microarray technology into a transcriptomics approach for the non-model organism *Chironomus riparius*. PLoS One 7:e48096.

Martin, J. A., Z. Wang. 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12:671-682.

Maxam, A. M., W. Gilbert. 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. E.U.A. 74:560-564.

Miller, J. R., S. Koren, G. Sutton. 2010. Assembly algorithms for next-generation sequencing data. Genomics 95:315-327.

Min Jou, W., G. Haegeman, M. Ysebaert, y W. Fiers. 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature 237:82-88.

Moreton, J., A. Izquierdo, y R. D. Emes. 2016. Assembly, assessment, and availability of *de novo* generated eukaryotic transcriptomes. Front. Gent. 6:1-9.

Munch, K., A. Krogh. 2006. Automatic generation of gene finders for eukaryotic species. BMC Bioinformatics. 7:263.

Mundry, M., E. Bornberg-Bauer, M. Sammeth, P. G. D. Feulner. 2012. Evaluating characteristics of *de novo* assembly software on 454 transcriptome data: a simulation approach. PLoS One. 7:e31410.

Murphy, K. 2005. Hidden Markov Model toolbox for Matlab. https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html. Último acceso: Abril 04, 2018.

Murvai, J., K. Vlahoviček, C. Szepesvári, S. Pongor. 2001. Prediction of Protein Functional Domains from Sequences Using Artificial Neural Networks. Genome Res. 11:1410-1417.

Nawrocki, E. P., S. R. Eddy. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 29:2933-2935.

O'Neil, S. T., S. J. Emrich. 2013. Assessing *de novo* transcriptome assembly metrics for consistency and utility. BMC Genomics. 14:465.

Olsina, L.A. 1999. Metodología Cuantitativa para la Evaluación y Comparación de calidad de Sitios Web. (Doctorado). La Plata, Argentina. Universidad Nacional de La Plata. 193 p.

Pachter, L., M. Alexandersson, S. Cawley. 2002. Applications of Generalized Pair Hidden Markov Models to Alignment and Gene Finding Problems. J. Comput. Biol. 9:389-399.

Parra, G., K. Bradnam, I. Korf. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23:1061-1067.

POD. 2017. Penguin Computing on Demand. https://pod.penguincomputing.com. Último acceso: Abril 04, 2018.

Ponting, C. P., P. L. Oliver, W. Reik. 2009. Evolution and functions of long non-coding RNAs. Cell. 136:629-641.

Pontius, J., L. Wagner, G. Schuler. 2002. UniGene: A unified view of the transcriptome. En: McEntyre, J., Ostell, J. The NCBI Handbook [Internet]. National Center for Biotechnology Information, Bethesda, MD. E.U.A.

Quek, X. C., D. W. Thomson, J. L. V. Maag, N. Bartonicek, B. Signal, M. B. Clark, B. S. Gloss, M. E. Dinger. 2015. lncRNAdb v2.0: expanding the reference database for functional long non-coding RNAs. Nucleic Acids Res. 43:D168-D173.

Quiagen. 2014. Manual for CLC Genomics Workbench 7.0. CLC bio, a Quiagen Company. Dinamarca. 847p.

Rabiner, L. R. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE. 77:257-286.

Rankin, K., B. M. Hill. 2013. The official Ubuntu Server book. Prentice Hall Press. Tercera edición. Upper Saddle River, NJ, E.U.A. 542p.

Romero-Vivas, E., F. Von Borstel, I. Villa-Medina. 2013. Analysis of Genetic Expression with Microarrays using GPU Implemented Algorithms. Comput. y Sist. 17:357-364.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, J. Hoon, J. F. Simons, D. Marran,

J. W. Myers, J. F. Davidson, A. Branting, J. R. Nobile, B. P. Puc, D. Light, T. A. Clark, M. Huber, J. T. Branciforte, I. B. Stoner, S. E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J. A. Fidanza, E. Namsaraev, K. J. McKernan, A. Williams, G. T. Roth, J. Bustillo. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348-352.

Rozenberg, A., M. Parida, F. Leese, L. C. Weiss, R. Tollrian, J. R. Manak. 2015. Transcriptional profiling of predator-induced phenotypic plasticity in *Daphnia pulex*. Front. Zool. 12:18.

Ruiz-Laguna, J., J. M. Vélez, C. Pueyo, N. Abril. 2016. Global gene expression profiling using heterologous DNA microarrays to analyze alterations in the transcriptome of *Mus spretus* mice living in a heavily polluted environment. Environ. Sci. Pollut. Res. 23:5853-5867.

Rung, J., A. Brazma. 2013. Reuse of public genome-wide gene expression data. Nat. Rev. Genet. 14:89-99.

Rustici, G., N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. P. Pereira, E. Pilicheva, J. Rung, A. Sharma, Y. A. Tang, T. Ternent, A. Tikhonov, D. Welter, E. Williams, A. Brazma, H. Parkinson, U. Sarkans. 2013. ArrayExpress update-trends in database growth and links to data analysis tools. Nucleic Acids Res. 41:987-990.

Salzberg, S. L., A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marçais, M. Pop, J. a Yorke. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22:557-567.

Sanger, F., A. R. Coulson. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94:441-448.

Sanger, F., S. Nicklen, A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74:5463-5467.

Schena, M., D. Shalon, R. W. Davis, y P. O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA Microarray. Science. 270:467-470.

Schulz, M. H., D. R. Zerbino, M. Vingron, E. Birney. 2012. Oases: robust *de novo* RNA-Seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086-1092.

Shabalina, S. A., N. A. Spiridonov. 2004. The mammalian transcriptome and the function of non-coding DNA sequences. Genome Biol. 5:105.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210-3212.

Smith-unna, R., C. Boursnell, R. Patro, J. M. Hibberd, S. Kelly, D. Street, S. Brook, S. P. Road. 2015. TransRate: reference free quality assessment of *de novo* transcriptome assemblies. bioRxiv:1-25.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, L. E. Hood. 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321:674-679.

Staples, G. 2006. TORQUE Resource Manager. En: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing. SC ’06. ACM, New York, NY, E.U.A.

Steijger, T., J. F. Abril, P. G. Engström, F. Kokocinski, M. Akerman, T. Alioto, G. Ambrosini, S. E. Antonarakis, J. Behr, P. Bertone. 2013. Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods 10:1177-1184.

Stekel, D. 2003. Microarray Bioinformatics. Cambridge University Press. Primera edición. Cambridge, Reino Unido. 263p.

Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. a Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, J. B. Hogenesch. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. U. S. A. 101:6062-6067.

Surget-Groba, Y., J. I. Montoya-Burgos. 2010. Optimization of *de novo* transcriptome assembly from next-generation sequencing data. Genome Res. 20:1432-1440.

MathWorks Inc. 2013. MATLAB and Bioinformatics Toolbox Release 2013a. https://www.mathworks.com/products/matlab.html. Último acceso: Abril 04, 2018.

UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45:D158-D169.

Velázquez-Lizarraga, A. E. 2016. Identificación de genes relacionados con las rutas de señalización del sistema inmune de *Litopenaeus vannamei* (Boone, 1931) expuesto a un fertilizante enriquecido con silicio orgánico. (Maestría en Ciencias en el Uso Manejo y Preservación de los Recursos Naturales). La Paz, B.C.S., México. Centro de Investigaciones Biológicas del Noroeste, S.C. 25 p.

Vijay, N., J. W. Poelstra, A. Künstner, J. B. W. Wolf. 2013. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-Seq experiments. Mol. Ecol. 22:620-634.

Wheeler, D. L. 2003. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 31:28-33.

Won, K.-J., T. Hamelryck, A. Prugel-Bennett, A. Krogh. 2007. An evolutionary method for learning HMM structure: prediction of protein secondary structure. BMC Bioinformatics. 8:357.

Wu, R., E. Taylor. 1971. Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. J. Mol. Biol. 57:491-511.

Xue, S., Y. Liu, Y. Zhang, Y. Sun, X. Geng, J. Sun. 2013. Sequencing and *de novo* analysis of the hemocytes transcriptome in *Litopenaeus vannamei* response to white spot syndrome virus Infection. PLoS One. 8.

Yalcin, B., D. J. Adams, J. Flint, T. M. Keane. 2012. Next-generation sequencing of experimental mouse strains. Mamm. Genome 23:490-498.

Yoo, A. B., M. A. Jette, M. Grondona. 2003. SLURM: Simple Linux Utility for Resource Management. En: Feitelson, D., Rudolph, L., Schwiegelshohn. Job Scheduling Strategies for Parallel Processing. Springer. Berlin, Alemania. Pp 44-60.

Yoon, B. J. 2009. Hidden Markov models and their applications in biological sequence analysis. Curr. Genomics 10:402-415.

Zerbino, D. R., E. Birney. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18:821-829.

Zhang, S., I. Borovok, Y. Aharonowitz, R. Sharan, V. Bafna. 2006. A sequence-based filtering method for ncRNA identification and its application to searching for riboswitch elements. Bioinformatics. 22:e557-565.

Zhao, Q.-Y., Y. Wang, Y.-M. Kong, D. Luo, X. Li, P. Hao. 2011. Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics. 12(1):S2.

11. ANEXOS

Anexo A. Ensamblaje *de novo* de transcriptoma, el ensamblador Trinity

El ensamblador Trinity hace uso de distintos programas para llevar a cabos los procesos de sus tres módulos, *Inchworm, Chrysalis y Butterfly*. Su funcionamiento requiere de equipos de cómputo de mediana a alta capacidad, es decir, plataformas con múltiples núcleos de procesamiento y alta disponibilidad de memoria RAM con sistemas operativos Linux de 64 bits. Según especificaciones del desarrollador se requieren como mínimo dos núcleos de procesamiento y aproximadamente 1 GB de memoria RAM por cada millón de secuencias de entrada.

El módulo *Inchworm* obtiene el catálogo de *k*-meros y con base en éste se forman colecciones únicas de *contigs* lineares. *Inchworm* utiliza construcciones extendidas de secuencias (*greedy extensions*) con base en *k*-meros, así recupera solo un (mejor) *contig* para un conjunto de variantes que comparten *k*-meros provenientes del mismo *cluster* de lecturas (que pueden ser causados por *splicing* alternativo, duplicación de genes o variación alélica). Este módulo es el más demandante de los tres en términos de memoria y corre en un máximo de seis núcleos de procesamiento independientemente si la asignación inicial es mayor (Haas *et al.*, 2013a).

El módulo *Chrisalys* agrupa los *contigs* que corresponden a porciones de transcritos con *splicing* alternativo o porciones únicas de genes parálogos. Después construye un DBG para cada grupo de *contigs* relacionados, reflejando cada grafo la complejidad de traslapes entre variantes.

El ultimo modulo, *Butterfly*, analiza las trayectorias de los DBGs con respecto a sus respectivas lecturas y pareamientos de lecturas y reporta todas las secuencias de transcritos plausibles, resolviendo también isoformas con *splicing* alternativo y transcritos provenientes de genes parálogos.

Una corrida de Trinity involucra muchos *clusters* de lecturas, cada uno ensamblado separadamente, con los cuales se generan ‘genes’ e ‘isoformas’ que son depositadas en directorios individuales. Finalmente, se unen todos los *contigs* construidos en el proceso en un archivo fasta.

Cada *contig* tiene un número de identificación que comprende, el número de ‘gen’ del cual proviene, el número de isoforma, la longitud del *contig* y la trayectoria recorrida en el grafo que lo originó. Así también, la trayectoria es descrita como una serie de números de nodos y su posición en el *contig*. Por ejemplo, los dos *contigs* mostrados a continuación describen dos transcritos procedentes del *cluster* 111.01, por lo tanto, del mismo DBG, y provienen de los genes cDN111.01_g1 y cDN111.01_g2, ambos tienen solo una isoforma y ambos tienen una longitud de 1247 bases. El DBG tiene forma de burbuja, lo cual podemos ver en un archivo distinto al documento fasta de salida.

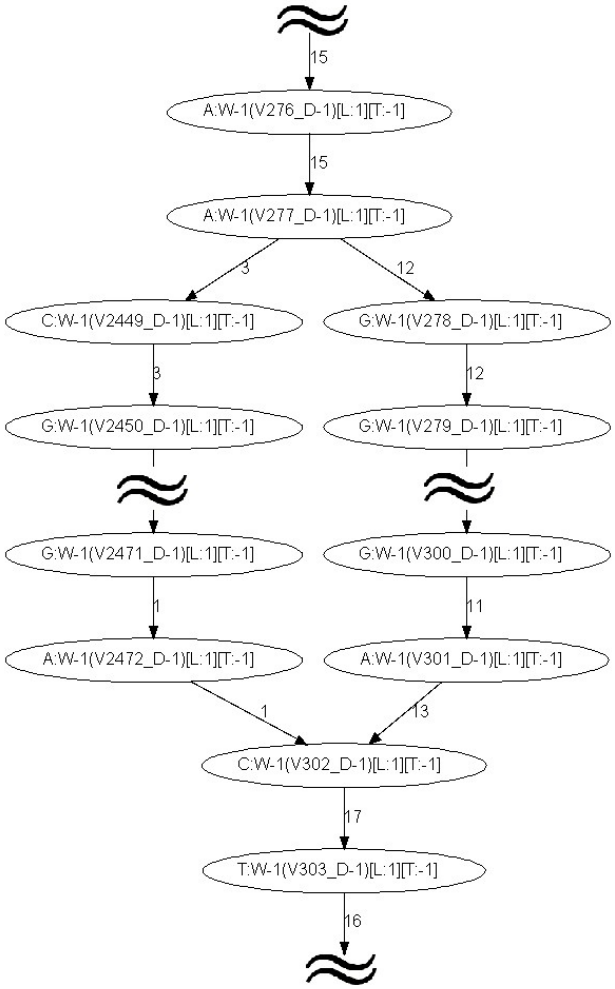
Para el primer *contig* el nodo ‘2476’ corresponde al rango de secuencia 0-299, el ‘2477’ del 300-323 y el ‘2478’ del 324-1246 de la secuencia del *contig*. Los números de nodo son únicos al identificador de gen, así los nodos en el DBG pueden ser comparados entre isoformas provenientes del mismo *cluster*.

>cDN111.01_g1_i1 len=1247 path=[2476:0-299 2477:300-323 2478:324-1246] [-1, 2476, 2477, 2478, -2]

>cDN111.01_g2_i1 len=1247 path=[2473:0-299 2474:300-323 2475:324-1246] [-1, 2473, 2474, 2475, -2]

Contig	**Posición**	**Secuencia**
cDN111.01_g1_i1	298	AA**G**GGAAGGTTGTAGTTGTAGATGGA**C**TGCA
cDN111.01_g2_i1	298	AA**C**GGAAGGTTGTAGTTGTAGATGGA**C**TGCA

Sección del grafo proveniente del *cluster* 111.01



Anexo B. Modelos Markovianos Ocultos

Un Modelo Markovianos Oculto (*Hidden Markovian Model*, o HMM) es un modelo estadístico que puede ser usado para describir la evolución de eventos observables que dependen de factores internos, que no son directamente observables (Yoon, 2009). Los HMM han sido usados en los campos de reconocimiento de voz (Rabiner, 1989) y comunicaciones digitales. También han sido empleados exitosamente en bioinformática en predicción de genes, alineamiento múltiple y por pares, *base-calling*, modelado de errores de secuenciación, predicción de estructuras secundarias de proteínas, identificación de ARN no codificante, alineamientos estructurales de ARN y muchas otra aplicaciones (Durbin *et al.*, 1998; Liang *et al.*, 2007; Munch y Krogh, 2006; Pachter *et al.*, 2002; Won *et al.*, 2007; Yoon, 2009; Zhang *et al.*, 2006).

Las definiciones formales de un HMM son explicadas a conciencia en (Rabiner, 1989); con base en este artículo este anexo provee un breve resumen de las variable y algoritmos utilizados en la sección 6.6 del documento principal de tesis.

En la definición formal de un HMM denotamos: N, el número de estados de un modelo; M, el número de distintos símbolos de observación por estado; A, la distribución de probabilidad de transición de estados (matriz de transición de estados); B, la distribución de probabilidad de símbolos de observación (matriz de confusión); y finalmente π, da distribución inicial de estados (matriz de estados inicial). Denotamos al HMM con tres de sus elementos para indicar su conjunto de parámetros, entonces la notación de un HMM es $\lambda = (A, B, \pi)$.

Con base en los elementos de los HMM, hay tres problemas básicos que resuelven los modelos. Problema 1: Dados una secuencia de observación $O = O_1, O_2 \ldots O_T$ y un modelo $\lambda = (A, B, \pi)$, determinar la probabilidad de la secuencia de observación dado el modelo $P(O|\lambda)$; Problema 2: dados la secuencia de observación $O = O_1, O_2 \ldots O_n$ y el modelo $\lambda = (A, B, \pi)$, escoger una secuencia de estados $Q = q_1, q_2 \ldots q_T$ que sea óptima en algún sentido significativo, es decir, que explique de mejor manera las observaciones; Problema 3: ajustar los parámetros

del modelo $\lambda = (A, B, \pi)$ para maximizar la $P(O|\lambda)$. Estos problemas son también llamados evaluación, decodificación y entrenamiento. En la publicación (Rabiner, 1989) se puede encontrar información en extenso sobre los algoritmos usados para resolverlos.

El entrenamiento de modelos puede ser realizado por medio del método iterativo *Baum-Welch* (Rabiner, 1989). Dado un número finito de secuencias de observación como datos de entrenamiento no existe una manera óptima de estimar los parámetros del modelo. Sin embargo, se puede escoger $\lambda = (A, B, \pi)$ de tal manera que $P(O|\lambda)$ sea maximizada localmente por medio de un algoritmo iterativo. La reestimación involucra la evaluación de cada secuencia de observación dado el modelo $P(O|\lambda)$ (por medio del procedimiento *Forward-Backward*) y las matrices de probabilidades intermedias $\alpha, \beta, \gamma\ y\ \xi$ para cada secuencia de entrenamiento, generando así un conjunto de parámetros reestimados $\bar{\lambda}$. Finalmente se realiza un paso de maximización donde se calcula si $P(O|\bar{\lambda}) > P(O|\lambda)$ y sustituyendo λ por $\bar{\lambda}$ si esta condición es satisfecha.

Anexo C. Preprocesamientos de lecturas RNA-Seq

Una estrategia de preprocesamiento regularmente comprende cortes de calidad, remoción de adaptadores o secuencias sobrerrepresentadas, y descartar secciones de lecturas con contenido de base no uniforme. Una vez ejecutados los pasos anteriores se eliminan en el preprocesamiento aquellas lecturas que resultaron demasiado cortas; el criterio de rechazo de estas lecturas está dado por el usuario y comúnmente toma como base la longitud mínima de *k*-mero utilizada por el ensamblador. A continuación, se presentan los parámetros de preprocesamiento para los datos RNA-Seq de los organismos involucrados en este estudio y descritos en la tabla C1.

Ratón: Las lecturas presentaron calidad aceptable, dando cabida a trazar una estrategia estándar de preprocesamiento con el objetivo de mejorar los datos que serían alimentados al ensamblador. El preprocesamiento en ratón comprendió: cortes de calidad por medio de ventanas deslizantes con umbral de puntaje de calidad 25; recorte de las primeras diez bases en todas las lecturas debido a contenido no uniforme de secuencias; remoción de adaptadores en las lecturas conservando complementariedad de las secuencias pareadas (modo palíndromo); las secuencias de adaptadores fueron tomadas del catálogo TrueSeq2 ya que contienen los adaptadores específicos de la plataforma de secuenciación de la base de datos RNA-Seq seleccionada. Finalmente, fueron descartadas todas aquellas secuencias que resultaron menores a 32 bases. El seguimiento de esta estrategia conllevó a descartar aproximadamente el 19% de los datos crudos, dejando alrededor de 20.9 millones de lecturas de 32 a 66 bases de longitud con calidad mínima de 25.

Mosca de la fruta: Las lecturas observaron baja calidad, lo cual requirió que se siguiera una estrategia distinta de preprocesamiento. Se ejecutó una remoción de adaptadores en modo palíndromo, se eliminaron las primeras diez bases de todas las lecturas debido a contenido no uniforme y finalmente se descartaron todas

aquellas lecturas que resultaron menores de 32 bases. Como resultado de este procedimiento se descartó aproximadamente el 40% de los datos crudos, permaneciendo aproximadamente 7.6 millones de lecturas pareadas de 32 a 65 bases de longitud y puntajes de calidad mínima de 5.

Pulga de agua: La estrategia a seguir en este organismo consistió en corte de primeras 10 bases, remoción de adaptadores en modo palíndromo, cortes de calidad por medio de ventanas deslizantes con umbral de puntaje de calidad de 25, y descartar todas aquellas lecturas que resultaron menores de 32 bases; posterior al preprocesamiento los datos seguían conteniendo secuencias sobrerrepresentadas, ribosómicas según la búsqueda en la base de datos del NCBI (Wheeler, 2003), consecuentemente se realizó un segundo pre-procesamiento para remover dichas secuencias. Posterior al preprocesamiento quedaron alrededor de 7.2 millones de lecturas pareadas de 32 a 90 bases de longitud con calidad mínima de 25.

Camarón blanco: Las lecturas crudas de este organismo tienen buena calidad (promedios de puntajes de calidad > 28), con excepción de la uniformidad de contenido de bases en las primeras diez bases de las secuencias. Por tanto, sólo se removieron las primeras diez bases de todas las lecturas dejando el 100% de las lecturas, las cuales contienen 80 bases de longitud y con calidad mínima de 18.

Tabla C1. Lecturas RNA-Seq: especificaciones y preprocesamiento.

Organismo	**Ratón**	**Mosca**	**Pulga**	**Camarón**
Repositorio	ENA[2]	SRA[3]	ENA[2]	ENA[2]
Fuente	(Yalcin *et al.*, 2012)	(Daines *et al.*, 2011)	(Rozenberg *et al.*, 2015)	(Xue *et al.*, 2013)
No. ID.	ERX012425	SRR042489	SRR2075894	ERR313300
Condición de estudio	Cerebro completo ratones C57BL/6NJ machos, de 8 semanas	Organismos completos, hembras adultas de 3 días	RNA total de juveniles R9	Hemolinfa de camarones de 15.2 gramos
Datos crudos	25,912,031 lecturas PE[1], longitud=76	12,468,019 lecturas PE[1], longitud=75	10,934,751 lecturas PE[1], longitud=100	12,907,027 lecturas PE[1], longitud=90
Datos preprocesados	20,949,267 lecturas PE, longitud = 32-66	7,564,138 lecturas PE, longitud = 32-65	7,168,393 lecturas PE, longitud = 32-90	12,907,027 lecturas PE, longitud = 80

[1]PE se refiere a lecturas pareadas (*Paired-End*). Repositorios: [2]El Archivo Europeo de Nucleótidos (ENA) perteneciente al Instituto Europeo de Bioinformática (EBI) del Laboratorio Europeo de Biología Molecular (EMBL) (Leinonen *et al.*, 2011a). [3]El Archivo de Secuencias de Lecturas (SRA) (Leinonen *et al.*, 2011b) pertenece al Centro Nacional de Información Biotecnológica (NCBI) (Coordinators, 2013).

Anexo D. Especificaciones de las plataformas de cómputo

El desarrollo de los procesos requeridos en este proyecto de tesis se realizó en distintas plataformas de cómputo con las especificaciones descritas en la tabla D1.

1) Estación de trabajo (W_1 y W_2):

Se utilizó una estación de trabajo *Dell Precision* T7500, que cuenta con un procesador *Intel Xeon* X5680 3.3 GHz de 6 núcleos, con capacidad en discos duros de 2.5 TB, a la cual se le restringió la memoria RAM para este proyecto, cambiando de 24 GB a 20 GB (baja cantidad de memoria), de acuerdo a las configuraciones W_2 y W_1, respectivamente.

Detalles adicionales del procesador:

- 6 núcleos, 12 hilos
- Velocidad: 3.33GHz, RAM:4GB por núcleo
- Ancho de banda: 32GB/s
- Memoria Cache
    - 32KB nivel 1 cache de instrucción por núcleo
    - 32KB nivel 1 cache de datos por núcleo
    - 256KB nivel 2 cache por núcleo
    - 12MB nivel 3 cache compartido

2) HPC (V_1 y V_2):

El primer recurso de HPC está conformado por la supercomputadora Cuetlaxcoapan del LNS, compuesta de un *cluster* estándar de cálculo con procesadores Intel Xeon y un *cluster* con procesadores *Intel Xeon Phi Knights* Landing. El *cluster* estándar está compuesto de 228 nodos de cálculo *Thin* y otros 42 nodos de cálculo más robustos (*fat, semi-fat, ultra-fat*). Para los procesos de ensamblaje realizados en este estudio se utilizó el *cluster* estándar, el cual funciona en modo virtual (recursos compartidos) y donde los nodos de cálculo *Thin*, tienen 2 procesadores Intel Xeon E5-2680 v3 (*Haswell*) a 2.5 GHz, con 24 núcleos en total y 128 GB de memoria RAM. Los nodos están intercomunicados con una red *Ethernet Gigabit* y una red *Infiniband FDR* a 56 Gbps.

Detalles adicionales del procesador:

- 12 núcleos, 24 hilos
- 2.5GHz, RAM: 5.33GB por núcleo

- Ancho de banda: 68GB/s
- Memoria Cache
    - 32KB nivel 1 cache de instrucción por núcleo
    - 32KB nivel 1 cache de datos por núcleo
    - 256KB nivel 2 cache por núcleo
    - 30MB nivel 3 cache compartido

Este *cluster* utiliza el administrador de carga de trabajos SLURM (Yoo *et al.*, 2003), que es libre y puede manejar un *cluster* Linux de cualquier dimensión. La especificación de los recursos computacionales a utilizar en cada ensamblaje se definió en el *Job Script*, a través de los parámetros de SLURM:

#SBATCH -n 24 # number of MPI tasks (cores) requested
#SBATCH --ntasks-per-node=24 # task (cores) per node (maximum 24)

Este ejemplo especifica que se ejecute el trabajo con 24 núcleos (1 nodo *Thin*) del *cluster* estándar, donde cada núcleo obtiene 5.3 GB de RAM (BUAP, 2017). Esta plataforma computacional, se utilizó para realizar los ensambles en las configuraciones V_1 y V_2, especificando el uso de 1 nodo en el *Job Script*, pero variando la cantidad de núcleos en concordancia con los parámetros de entrada de Trinity.

3) HPC (H_1 *y* H_2):

El segundo recurso de HPC utilizó el servicio en la nube POD de *Penguin Computing* con la cola T30, que especifica nodos con un procesador *Intel Xeon E5-2660 v3 (Haswell)* a 2.6 GHz con 20 núcleos y 128 GB de RAM. Todos los nodos están intercomunicados con una red *Ethernet Gigabit* de 10 Gbps y una red *Infiniband QDR* a 40 Gbps.

Detalles adicionales del procesador:

- 10 núcleos, 20 hilos
- 2.60GHz, RAM:12.8GB/núcleo
- Memoria Cache
    - 32KB nivel 1 cache de instrucción por núcleo
    - 32KB nivel 1 cache de datos por núcleo
    - 256KB nivel 2 cache por núcleo
    - 25MB nivel 3 cache compartido

El servicio utiliza el planificador PBS TORQUE (Staples, 2006) para introducir trabajos al *cluster* computacional. Los servidores son brindados de manera dedicada. Sin embargo, es necesario seleccionar una cola del planificador. Cada cola provee diferentes tipos de nodo de cómputo, y por lo tanto tienen diferentes precios.

La especificación de los recursos se definió en el *Job Script*, a través de los parámetros de TORQUE:

#PBS -q T30
#PBS -l nodes=1:ppn=20

Este ejemplo especifica que se ejecute el trabajo con 1 nodo de 20 núcleos de la cola T30, donde cada núcleo tiene 6.4 GB de RAM (POD, 2017). Se utilizó esta plataforma computacional, especificando en el *Job Script* el uso de 1 nodo, pero variando el número de núcleos en concordancia con los parámetros de entrada de Trinity; estas disposiciones se utilizaron para realizar los ensambles en las configuraciones H_1 *y* H_2.

Tabla D1. Especificaciones de las plataformas de cómputo.

Plataforma de cómputo				**Parámetros Trinity**		**Planificador**	
						SLURM	TORQUE
Nombre	Plataforma	Memoria RAM (GB)	Núcleos	Máxima Memoria	CPU	Núcleos	Nodo/ Núcleos
W_1	Estación de trabajo	20	6	20	6	-	-
W_2		24	6	24	6	-	-
V_1	HPC, servidores virtuales	128/nodo	24/nodo	24	6	6	-
V_2		128/nodo	24/nodo	64	12	12	-
H_1	HPC, servidores dedicados	128/nodo	20/nodo	24	6	-	1/6
H_2		128/nodo	20/nodo	64	10	-	1/10

Anexo E. Definiciones formales

Se presentan en este anexo la formalización de diversas mediciones de repetibilidad, variabilidad y calidad realizadas en este proyecto de tesis.

Repetibilidad y variabilidad

Dado un conjunto de lecturas de secuenciación L_m de una especie m obtenidas de una base de datos pública, sea $E_{(p,m,i)}$ un ensamble (constituido por *contigs*) realizado con configuración de plataforma computacional p, utilizando L_m como entrada al proceso de ensamblaje *de novo*, donde i es el número correspondiente a la repetición del proceso de ensamblaje utilizando las mismas condiciones iniciales, se tiene que:

$$I_{(p,m,n)} = \bigcap_{i=1}^{i=n} E_{(p,m,i)} \qquad (1)$$

donde $I_{(p,m,n)}$ representa el conjunto de *contigs* resultantes de la intersección entre los n ensambles *de novo*, con las mismas condiciones.

Asimismo, se puede decir que el conjunto de *contigs* no intersectados $\bar{I}_{(p,m,n)}$ es tal, que:

$$\bar{I}_{(p,m,n)} \cap I_{(p,m,n)} = \{\} \qquad (2)$$

Por lo tanto, el conjunto $\bar{I}_{(p,m,n)}$ representa el conjunto de *contigs* que no aparecen en todos los ensambles $E_{(p,m,i)}$ que se generaron durante algún ensamblaje en particular.

De tal manera que la cantidad total de *contigs* obtenidos por plataforma p para un organismo m en n repeticiones está dada por la unión de sus conjuntos intersectados y no intersectados

$$Ctotal_{(p,m,n)} = I_{(p,m,n)} \cup \bar{I}_{(p,m,n)} \qquad (3)$$

La cuantificación de la repetibilidad se da con base al porcentaje que representa el subconjunto $I_{(p,m,n)}$ del conjunto $Ctotal_{(p,m,n)}$ y la variabilidad se da con base al porcentaje que representa el subconjunto $\bar{I}_{(p,m,n)}$ del conjunto $Ctotal_{(p,m,n)}$, encontrados por plataforma p para el organismo m en n repeticiones.

Finalmente, la ganancia por variabilidad entre plataformas se cuantifica tomando en cuenta la relación de la variabilidad máxima de las configuraciones de la estación de trabajo entre la variabilidad máxima de las configuraciones en las plataformas basadas en HPC.

Calidad de ensamblaje con respecto a transcriptomas de referencia

Se define como referencia codificante a la mejor aproximación de ensamble de transcriptoma disponible en cierta fecha o versión, disponible en las bases públicas de los organismos involucrados o en repositorios de referencias como Ensembl (Flicek *et al.*, 2014; Zhao *et al.*, 2011).

Formalmente, sean los conjuntos $I_{(p,m,n)}$ e $\bar{I}_{(p,m,n)}$, se procede a analizarlos con respecto al conjunto referencia codificante de la especie, llamado Transcriptoma de referencia T_m, mediante un proceso de identidad, utilizando el *software* BLAST (Camacho *et al.*, 2009); de tal manera que:

$$\{(c,t): c \in Im_{(p,m)}, t \in T_{(p,m)}\} \quad = f : I_{(p,m)} \rightarrow T_m \qquad (4)$$

donde (c,t) representa un par (*contig*, transcrito) y $Im_{(p,m)}$ es el conjunto de *contigs* intersectados generados con la plataforma computacional p, que mapearon en el conjunto referencia T_m. $T_{(p,m)}$ es el subconjunto de transcritos de T_m a los que mapearon los *contigs* intersectados del conjunto $Im_{(p,m)}$. Nótese que, para simplificar, se omitió el subíndice n. Así mismo:

$$\{(c,t): c \in \bar{I}m_{(p,m)}, t \in \bar{T}_{(p,m)}\} \ = f : \ \bar{I}_{(p,m)} \rightarrow T_m \qquad (5)$$

donde (c,t) representa un par (*contig*, transcrito) e $\bar{I}m_{(p,m)}$ es el conjunto de *contigs* no intersectados generados con la plataforma computacional p, que mapearon en el conjunto referencia T_m. $\bar{T}_{(p,m)}$ es el subconjunto de transcritos de T_m a los que mapearon los *contigs* no intersectados del conjunto $\bar{I}m_{(p,m)}$.

Ya que pueden existir *contigs* intersectados y no intersectados que mapean a un transcrito común, se puede realizar la siguiente operación:

$$T_{(p,m)} \cap \bar{T}_{(p,m)} = \ T^*_{(p,m)} \qquad (6)$$

donde $T^*_{(p,m)}$ es el subconjunto de transcritos mapeados compartidos por los conjuntos de *contigs* intersectados y no intersectados. Para obtener los *contigs* no intersectados compartidos, se realiza:

$$\{(c,t): c \in \ \bar{I}m^*_{(p,m)}, t \in T^*_{(p,m)}\} = f : \ \bar{I}_{(p,m)} \rightarrow T^*_{(p,m)} \qquad (7)$$

donde $\bar{I}m^*_{(p,m)}$ es el conjunto de *contigs* mapeados no intersectados compartidos.

Asimismo, podemos realizar la siguiente operación:

$$\bar{I}m_{(p,m)} - \bar{I}m^*_{(p,m)} = \bar{I}m^+_{(p,m)} \qquad (8)$$

donde $\bar{I}m^+_{(p,m)}$ es el conjunto de *contigs* no intersectados que exclusivamente mapean a transcritos en T_m que no son compartidos con el conjunto de transcritos mapeados inicialmente por $Im_{(p,m)}$.

La representación de estos conjuntos con respecto al transcriptoma de referencia se muestra en la figura 4 del documento principal.

Para decidir que *contigs* se compartían entre ensambles se utilizó un criterio estricto de coincidencia única, es decir, ambas cadenas deberían ser exactamente iguales en tamaño y composición. No obstante, los algoritmos de mapeo tienen un criterio más laxo, permitiendo reconocer secciones similares aun cuando no sean las cadenas exactamente iguales. Este criterio permite variar la longitud y composición del *contig*. Debido a este criterio pudiese haber *contigs* que siendo ligeramente distintos mapean a la misma porción del transcriptoma. Esta situación puede ocurrir para ambos conjuntos (comunes y no compartidos) y entre conjuntos.

Por ello, la evaluación de calidad considera como información válida originada por variabilidad de plataforma a todos aquellos *contigs* contenidos en el subconjunto $\bar{I}_{(p,m,n)}$. Esta evaluación se expresa en el porcentaje representado por los *contigs* mapeados provenientes del conjunto no intersectado $\bar{I}m_{(p,m)}$, por plataforma p para un organismo m con respecto al $\bar{I}_{(p,m,n)}$, dados n ensambles.

Asimismo, considera como información nueva originada por la variabilidad de plataforma solo a los *contigs* mapeados exclusivos al conjunto no intersectado $\bar{I}m^{+}_{(p,m)}$. Esta evaluación se expresa en el porcentaje representado por los *contigs* mapeados exclusivos provenientes del conjunto no intersectado $\bar{I}m^{+}_{(p,m)}$, por plataforma p para un organismo m con respecto al $\bar{I}_{(p,m,n)}$, dados n ensambles.

Calidad de ensamblaje con respecto a la base de datos de proteínas UniProt/Swiss-Prot

La evaluación de calidad consiste en el mapeo de *contigs* a la base curada de datos de proteínas UniProt/Swiss-Prot (The UniProt Consortium, 2017). Se analizó la calidad de los *contigs* originados por la variabilidad de cada plataforma de cómputo, por lo tanto se mapearon los conjuntos de *contigs* no intersectados ($\bar{I}_{(p,m,n)}$) con respecto la base de datos de proteínas utilizando el algoritmo BLASTX del *software* BLAST (Camacho *et al.*, 2009). Únicamente se seleccionaron los mapeos de mayor calidad dentro de las seis posibles traducciones por *contig*.

Formalmente, el conjunto de *contig* no intersectados $\bar{I}_{(p,m,n)}$ que observaron similitud (*hit* positivo) con alguna de las secuencias aminoacídicas de la base de datos UniProt/Swiss-Prot es referido como

$$\bar{I}mp_{(p,m)} \qquad (9)$$

De tal manera que, la evaluación de calidad se expresa en el porcentaje representados por los *contigs* $\bar{I}mp_{(p,m)}$ por plataforma p para un organismo m con respecto al $\bar{I}_{(p,m,n)}$, dados n ensambles.

Asimismo, la ganancia máxima en *contigs* no intersectados mapeados a la base de datos de proteínas se estableció por la división de la cantidad máxima de *contigs* $\bar{I}mp_{(p,m)}$, obtenidos en una configuración de la estación de trabajo entre la cantidad máxima de estos *contigs* obtenidos en la configuración basada en HPC.

Anexo F. Bases de datos utilizadas en el proyecto de tesis

Transcriptomas de referencia

Organismo	Repositorio	Versión	Cantidad de transcritos
Ratón	Ensembl[1]	GRCm38	103,734
Mosca de la fruta	Ensembl[1]	*Release* 86	30,651
Pulga de agua	Ensembl[1]	GCA_000187875.1	30,590

Repositorio: [1]Ensembl (Flicek *et al.*, 2014) perteneciente al instituto EMBL-EBI.

Base de datos de proteínas

El conjunto de secuencias de proteínas utilizada en la evaluación de camarón blanco fueron las secuencias de la base de datos UniProtKB. En específico, la base de datos con curación manual UniProt/Swiss-Prot (Bateman *et al.*, 2015); al momento de su acceso esta contenía 555,426 secuencias (13 de septiembre del 2017).

Bases de datos de microarreglos de expresión génica

Organismo	Ratón	Mosca
Repositorio	GEO[1]	GEO[1]
Fuente	(Su *et al.*, 2004)	(Doroszuk *et al.*, 2012)
No. ID.	GSM258635 y GSM258636	GSM897165 y GSM897166
Condición de estudio	Corteza cerebral de ratones C57BL/6 macho de 8-10 semanas	Organismos completos de hembras adultas alimentadas con dietas óptimas
Microarreglo	*Affimetrix Mouse Genome 430 2.0 array*	*Affimetrix GeneChip Drosophila Genome 2.0 array*
Cantidad de sondas del panel	495,374 y 1,094 sondas de prueba y control respectivamente	263,272 sondas de prueba y 2,128 de control
Longitud	25-meros[2]	25-meros[2]
Transcritos representados en el panel	34,000	18,550
Conjuntos de prueba[3]	45,000, 11 oligonucleótidos por conjunto	18,880, 14 oligonucleótidos por conjunto

Repositorio: [1]Ómnibus de Expresión (GEO) perteneciente al NCBI (Barrett *et al.*, 2013). [2] *n*-meros se refiere a la longitud de secuencia de la sonda. [3] Conjunto de prueba o *probe-set:* Conjunto de sondas que apuntan a un mismo transcrito de prueba.

Bases de datos de UniGenes de camarón blanco

UniGenes de camarón blanco utilizados en para entrenamiento de HMMs empleados en la evaluación de calidad y selección de ensamblaje.

ID	Cantidad de Secuencias	ID	Cantidad de Secuencias
Lva.915	25	Lva.1305	21
Lva.2240	23	Lva.3882	21
Lva.7852	23	Lva.579	21
Lva.2443	26	Lva.855	21
Lva.162	20	Lva.20545	21
Lva.3348	20	Lva.18779	20
Lva.5310	23	Lva.18744	21
Lva.674	20	Lva.14761	21
Lva.5531	22	Lva.14748	20
Lva.3699	26	Lva.7912	21
Lva.3490	21	Lva.4473	21
Lva.3711	23	Lva.3695	20
Lva.530	22	Lva.2930	20
Lva.5265	22	Lva.2456	22
Lva.3327	22	Lva.980	21
Lva.3267	23	Lva.323	21
Lva.9485	21	Lva.4154	21
Lva.4508	23	Lva.3320	21
Lva.235	20	Lva.2642	21
Lva.6146	22	Lva.300	22
Lva.1713	22	Lva.23719	23
Lva.2190	22	Lva.28201	21
Lva.5479	22	Lva.25153	21
Lva.4713	22	Lva.24812	21
Lva.3892	23	Lva.24777	21
Lva.3799	21	Lva.24598	21
Lva.1501	24	Lva.24509	21
Lva.3931	21	Lva.24346	21
Lva.6500	21	Lva.24202	21
Lva.1279	23	Lva.24131	21
Lva.3743	21	Lva.24050	21
Lva.15256	21	Lva.24048	21
Lva.419	21	Lva.24045	21
Lva.5219	21	Lva.24032	21