ANÁLISIS TRANCRIPTÓMICO DE LOS GENES IMPLICADOS EN LA SUPERVIVENCIA DE *Nezara viridula,* PLAGA DE CULTIVOS AGRÍCOLAS EN ARGENTINA.

Trabajo Final de Grado
De la alumna

LUCILA MAITÉ PÉREZ GIANMARCO

Este trabajo ha sido presentado como requisito
para la obtención del título de

Licenciado en Genética

Carrera

Licenciatura en Genética

(Sheila Ons)
Director

(María Inés Catalano)
Co-Director

Escuela de Ciencias Agrarias, Naturales y Ambientales.
Universidad Nacional del Noroeste de la Provincia de Buenos Aires.

Pergamino, 2 de marzo de 2018

ANÁLISIS TRANCRIPTÓMICO DE LOS GENES IMPLICADOS EN LA SUPERVIVENCIA DE *Nezara viridula,* PLAGA DE CULTIVOS AGRÍCOLAS EN ARGENTINA.

Trabajo Final de Grado

del alumno

LUCILA MAITÉ PÉREZ GIANMARCO

Aprobada por el Tribunal Evaluador

(Rolando Rivera Pomar)	(Agustina Pascual)	(Andrés Lavore)
Evaluador	**Evaluador**	**Evaluador**

(Sheila Ons)	(María Inés Catalano)
Director	**Co-Director**

Escuela de Ciencias Agrarias, Naturales y Ambientales,
Universidad Nacional del Noroeste de la Provincia de Buenos Aires

Pergamino,12 de marzo de 2018

RESUMEN

Nezara viridula es un hemíptero fitófago, plaga de cultivos agrícolas de importancia a nivel nacional e internacional. Para su control, como complemento o reemplazo de insecticidas neurotóxicos, se plantea el diseño de mecanismos especie-específicos, de manera de disminuir sus consecuencias negativas en el ecosistema y prevenir la aparición de resistencias. Para ello, un primer paso es la obtención de información genética que aporte conocimiento de secuencia de genes involucrados en la supervivencia de este insecto. En este trabajo se propuso obtener un transcriptoma de *N. viridula*, evaluar su cobertura e identificar en el mismo genes involucrados en su supervivencia, como neuropéptidos y genes neurales. Para ello, se hicieron comparaciones de los ensambles aquí obtenidos con bases de datos de especies evolutivamente cercanas, buscando ortología. Habiendo obtenido dos ensambles, éstos se compararon según el número de perfiles CEG y BUSCO para evaluar su completitud, así como también se contrastó la relación entre el número de bases ensambladas y genes de Uniprot encontrados. Tras elegir el transcriptoma que consideramos mejor para este trabajo, se hizo una búsqueda para genes implicados en la supervivencia de esta chinche, generando así conocimiento genético que resulta valioso para la futura investigación sobre este insecto plaga importante para la región.

INTRODUCCIÓN

Nezara viridula

Los hemípteros fitófagos, conocidos comúnmente como chinches, constituyen una amenaza para varios cultivos de la región productora de Argentina, principalmente para la soja, dado que se alimentan de vainas durante todo el ciclo de vida post-eclosión (Bimboni 1978).

Las chinches fitófagas causan daños en los cultivos al alimentarse por medio de sus estiletes e inyectar enzimas digestivas que licuan las células para poder ingerirlas (Todd 1982). Son capaces de afectar a la semilla de soja desde los inicios de su crecimiento dentro de su cavidad individual en la vaina y hasta el momento mismo de la cosecha, donde la gravedad del daño va decreciendo en la medida que nos acercamos al último estadio. Esto se debe a que la semilla va incrementando su dureza como consecuencia del proceso de deshidratación luego de alcanzar su mayor contenido de materia seca. Debido a esto, las posibilidades de penetración del estilete en el interior de las semillas son menores, aunque si la picadura de la chinche se produce sobre el eje embrionario, expuesto en este cultivo, el daño puede ser igualmente letal para la futura simiente. En los últimos años, la ocurrencia de precipitaciones ininterrumpidas resultó en nuevas oportunidades de generar daños internos a las semillas, por un nuevo proceso de imbibición (Fuentes *et al.* 2016).

La presencia de chinches durante el período de maduración de la semilla conduce a un menor poder germinativo y a la reducción del vigor de plántulas y de semillas, así como a la deformación de los granos y a la susceptibilidad frente a microorganismos patogénicos, entre otros daños de gravedad variable (Gamundi & Sosa 2007). Se observa un alto grado de asociación entre la incidencia de picaduras de chinches y la abundante producción de micelios de patógenos de distintos géneros, que son introducidos en los tejidos hidratados de la semilla por medio del estilete (Fuentes *et al.* 2016). Las modalidades de tecnología agrícola aplicadas en los últimos años han provocado un aumento en la abundancia de poblaciones de chinches, que son la plaga más importante en soja en la actualidad (Luna & Iannone 2013). De esta forma, los daños causados por chinches conducen a una significativa pérdida económica.

Nezara viridula (Hemiptera: Pentatomidae), conocida comúnmente como “chinche verde”, es la principal causante de pérdidas económicas dentro de un grupo de chinches en el que también se hallan *Dichelops furcatus, Piezodorus guildinii* y *Edessa meditabunda* (Aragón *et al*. 1997, Gamundi & Sosa 2007). El impacto de las distintas especies está relacionado a su densidad poblacional y al período de permanencia en el cultivo, la susceptibilidad del mismo y características del cultivar (Fuentes *et al.* 2016).
N. viridula es considerada el mayor pentatómido plaga del mundo. Es cosmopolita y marcadamente polífaga, afecta numerosos cultivos; la secuencia trigo-soja le ofrece un puente para el aumento de sus poblaciones. Luego de pasar por un período de inactividad en el invierno, se traslada desde la corteza de los árboles, arbustos y hojarasca hacia cultivos como alfalfa y soja para alimentarse. La disponibilidad y accesibilidad de estos hábitats son factores cruciales en la abundancia y dinámica temporal de las poblaciones en la primavera siguiente, más allá de los factores de mortalidad natural (Fuentes *et al.* 2016).

Los adultos de *N. viridula* miden entre 13 y 15 mm de largo y tienen un color verde intenso. Las hembras efectúan de 2 a 3 posturas hexagonales, de 80 a 100 huevos cada una. Los huevos tienen una coloración amarillenta y se enrojecen hacia el nacimiento. Pasan por cinco estadios ninfales hasta alcanzar la adultez unos 35-40 días post-eclosión. Las ninfas varían de un color rojizo en el primer estadio a negras y verdes con manchas blancas en estados avanzados de desarrollo. Permanecen agrupadas sin provocar mayores daños hasta el tercer estadio ninfal, cuando comienzan a dispersarse gradualmente. Los adultos pueden vivir hasta dos meses (http://www.agromeat.com/34440/chinches-fitofagas-en-el-cultivo-de-soja).
Dado que no hay en el mercado, productos químicos específicos para chinches, el control de estos hemípteros fitófagos se realiza principalmente con insecticidas neurotóxicos. Así, se eliminan no solo los insectos plaga sino también benéficos, como enemigos naturales y polinizadores (Perotti *et al.* 2010). Por otra parte, se han reportado casos de fallas en los tratamientos químicos, lo que podría atribuirse a la aparición de resistencias, si bien esto no ha sido estudiado en profundidad (Toledo *et al*. 2005).

Con el fin de retrasar o eludir la resistencia es necesario alternar insecticidas con diferentes modos de acción. Por lo tanto, resulta importante contar con una diversidad de opciones y de sitios de acción para nuevos productos, así como aumentar su especificidad a fin de preservar a las especies benéficas (Verlinden *et al.* 2014).

Se ha propuesto que la intervención en la fisiología de los insectos a través de manipulaciones en la expresión de genes determinados, como por ejemplo los sistemas de regulación hormonal, puede resultar una dirección productiva hacia la obtención de una nueva generación de insecticidas mediados por silenciamiento génico (Verlinden *et al.* 2014). Un primer paso para esto es conocer la secuencia nucleotídica de los genes de interés. Llamativamente, a pesar de su importancia económica, hasta el momento no se han reportado estudios destinados a conocer la secuencia de genes en la especie *N. viridula*. Las tecnologías de alto rendimiento, como la transcriptómica, pueden permitir superar esta falta de información génica en tiempos acelerados, en comparación con tecnologías previas.

Transcriptómica

Hasta la década pasada, los estudios de expresión génica estaban enfocados en el análisis de genes candidatos, mediante la reacción en cadena de la polimerasa en tiempo real o cuantitativa (qPCR), o dependían de la hibridación inter-específica en los *microarrays* (Boguski *et al.* 1994; Gerhard *et al.* 2004). Estos estudios se encontraban limitados a la investigación de organismos modelo (Lorenz *et al.* 1989) y a la genética médica (Korinek *et al.* 1973).

El desarrollo de la secuenciación de próxima generación (NGS) (Margulies *et al*. 2005) y de herramientas bioinformáticas para el análisis de la gran cantidad de datos generados, han permitido cambiar la situación en los últimos años, de forma tal que los análisis genómicos y transcriptómicos se han tornado accesibles para estudiar organismos no modelo (Ellegren *et al.* 2012; Lamichhaney *et al.* 2012; Ekblom & Wolf 2014).

La secuenciación de un transcriptoma, o “transcriptómica”, se refiere al uso de tecnologías de secuenciación de alto rendimiento, para la caracterización del contenido y la composición del ARN de una muestra particular. A diferencia del genoma, que es una entidad esencialmente estática, el transcriptoma puede ser modulado tanto por

factores internos como externos (Velculescu *et al.* 1997). El conocimiento del mismo es esencial para la interpretación de los elementos funcionales de un genoma, así como para revelar los constituyentes moleculares de las células, los tejidos y órganos en condiciones específicas (Wang *et al*. 2009).

La utilización de tecnologías de secuenciación de alto rendimiento para la caracterización del contenido y de la cantidad de ARN en una muestra dada se denomina ARNseq. La secuenciación paralela masiva de millones de secuencias de ADNc brinda un método efectivo y de bajo costo para obtener grandes cantidades de datos transcriptómicos de muchos organismos y tipos de tejidos (Birol *et al.* 2009; Trapnell *et al.* 2010). En principio, esos datos nos permitirían identificar todos los transcriptos expresados, como una secuencia de ARN mensajero (ARNm) desde el principio de la transcripción hasta el final, aún para los genes que tienen múltiples isoformas debido al *splicing* alternativo (Guttman *et al.* 2010).

En el año 2005 comenzaron a desarrollarse estas nuevas tecnologías de secuenciación, con la aparición de 454 Sequencing (Margulies *et al.* 2005) y la secuenciación por síntesis (Solexa) (Bentley 2006). A diferencia de los métodos de secuenciación por Sanger, éstas generan lecturas cortas (25pb-500pb) de buena calidad. De esta forma fueron apareciendo varias plataformas que, con distinta metodología, generaban lecturas cortas. Entre ellas, Illumina es la que logró imponerse, dado que ofrece diferentes plataformas, mayor cobertura y costos relativamente bajos (Goodwin *et al.* 2016).

Actualmente, por limitaciones tecnológicas, los sistemas de secuenciación utilizados masivamente no permiten obtener la información de secuencias en la forma de transcriptos completos, sino que antes de la secuenciación debe fragmentarse al azar el ADN para generar una genoteca de ADN simple cadena. A los fragmentos obtenidos, deben agregárseles adaptadores universales antes de dispersarlos en un vidrio inmovilizado (Fedurco *et al.* 2006), donde se lleva a cabo la amplificación en forma de puente (*bridge* PCR). El proceso genera grupos de fragmentos idénticos alrededor del fragmento original que se amplificó. La secuencia nucleotídica de los fragmentos se determina utilizando terminadores reversibles; éstos son nucleótidos bloqueados y marcados con un fluoróforo. De esta manera, cuando se agregan los nucleótidos

marcados, detienen la síntesis y producen una señal que es detectada por una cámara. Luego de la detección, se clivan los grupos terminales y fluoróforos, permitiendo que continúen la síntesis y la lectura (Turcatti *et al.* 2008).

La reconstrucción de un transcripto completo a partir de las lecturas cortas generadas durante la secuenciación es un desafío técnico, teniendo en cuenta que ocurren errores en el proceso de síntesis. Por ejemplo, no todos los transcriptos tienen un mismo nivel de cobertura, dadas las diferencias de expresión. Incluso dentro de un mismo transcripto, puede haber regiones con distinta cobertura por errores de secuenciación. Las lecturas con errores de secuenciación correspondientes a transcriptos altamente expresados pueden ser más abundantes que lecturas correctamente sintetizadas de transcriptos con baja expresión. Los mensajeros codificados por *loci* adyacentes pueden solaparse, formando transcriptos quiméricos. Las secuencias repetidas provenientes de distintos genes, introducen ambigüedad. Todas estas dificultades deberían ser superadas por un método eficiente, capaz de reconstruir transcriptos variables en tamaño, nivel de expresión e isoformas (Haas & Zody 2010).

La reconstrucción de un transcriptoma puede hacerse utilizando un genoma como referencia al que se mapean las lecturas, o bien mediante un ensamble *de novo*. Esta última alternativa permite ensamblar cuando no se cuenta con un genoma, o el mismo se encuentra fragmentado (Conesa *et al.* 2016), por lo que es altamente utilizada en el estudio transcriptómico de organismos no modelo. El ensamble sin referencia es también importante para realizar anotación génica. A pesar de que en los últimos años se multiplicaron los esfuerzos de secuenciación de genoma, para la mayoría de las especies la secuenciación genómica no ha sido completada. Incluso muchos genomas que se categorizan como finalizados contienen sitios sin secuencia, errores de ensamble y regiones no asignadas a cromosomas (Salzberg & Yorke 2005). Por lo anterior, las lecturas transcriptómicas pueden no mapear al genoma de referencia, por más que pertenezcan a un transcripto particular.

Una importante herramienta computacional para la estrategia de reconstrucción sin referencia es la utilización de los grafos de *De Brujin*. En este tipo de gráficos, se define un nodo por una secuencia de un largo fijo de *k* nucleótidos ("*kmero"*, siendo *k*

considerablemente más corto que el largo de la secuencia). Los nodos se conectan por sus bordes, si éstos se solapan perfectamente con un largo de *k-1* y los datos de secuencia avalan esta unión, se van construyendo las secuencias más largas (Figura 1). En el caso de construcción de transcriptomas, cada camino en el gráfico representa un posible transcripto. En este punto se puede apreciar la importancia de la limpieza por calidad del conjunto total de datos previo al ensamble, ya que lecturas con errores pueden resultar en falsos nodos, llevando al gráfico por caminos falsos (De Brujin 1946; Good 1946).

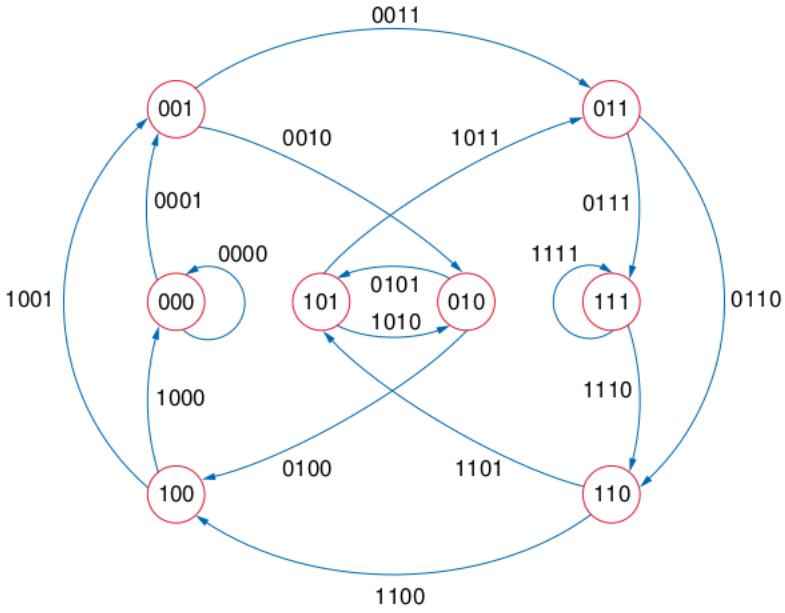

Figura 1: Gráfico de De Brujin con un alfabeto compuesto por los caracteres 0 y 1 (Pevsner 2009).

Hay varios programas para generar ensambles *de novo*, de los cuales, el más utilizado actualmente es Trinity (Grabherr *et al.* 2011). El mismo está compuesto por tres partes: *Inchworm, Chrysalis* y *Butterfly:*

- En una primera instancia, *Inchworm* construye un diccionario de *kmeros* de todas las lecturas *(k=25)* y elimina del diccionario los que pueden tener errores. Del diccionario se selecciona el *kmero* más frecuente y se utiliza como iniciador de un fragmento más largo, llamado *contig,* que luego se extiende hacia ambos extremos con *kmeros* que se solapan en *k-1* bases. Cada secuencia utilizada se

elimina del diccionario, de forma que no pueda volver a usarse. Este proceso se lleva a cabo hasta que ya no hay más *kmeros* en el diccionario que solapen con el *contig* en extensión, momento en el que se reporta el *contig* lineal. Con los *kmeros* restantes en el diccionario se inicia nuevamente el proceso hasta finalizar con el diccionario vacío (Figura 2a).

- A continuación, *Chrysalis* toma grupos de *contigs* solapantes y los convierte en componentes conectados, construyendo gráficos de *De Brujin* para los componentes. Esta tarea se realiza en tres etapas, donde la primera es la conexión de los *contigs*, la segunda es la construcción de los grafos y la última consta de la asignación de lecturas a cada componente (Figura 2b).
- En última instancia, *Butterfly* media la reconstrucción de los transcriptos completos, considerando los gráficos formados en la etapa anterior, así como la información *pair end* de las lecturas cuando la misma está disponible. De esta forma, se resuelven isoformas y genes parálogos en el transcriptoma (Figura 2c).

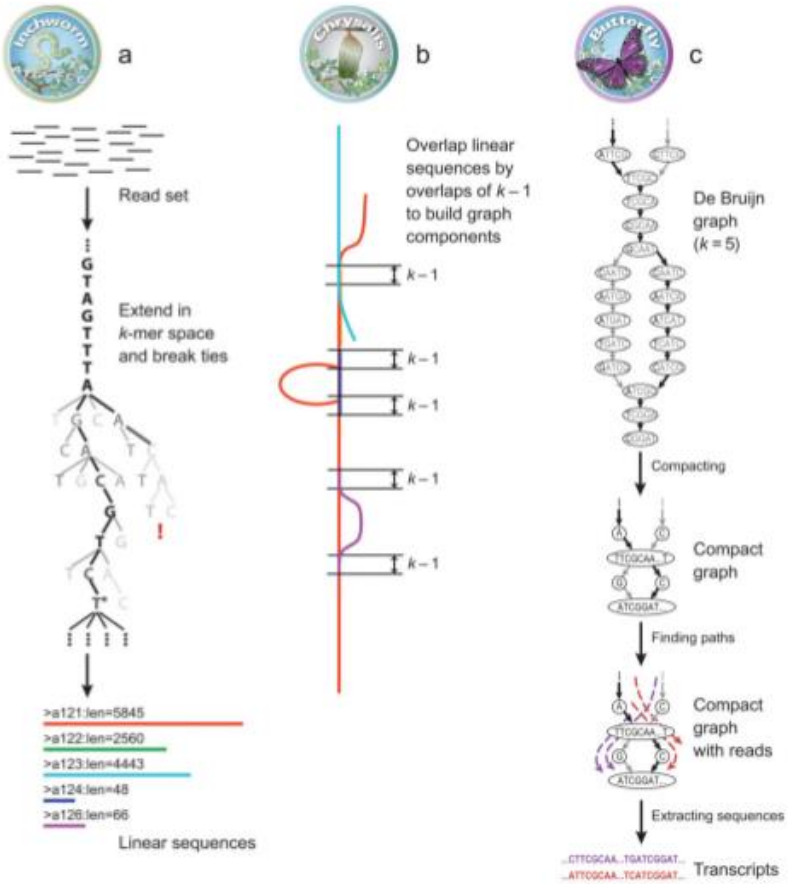

Figura 2: (a) *Inchworm* ensambla el set de datos de *reads* (líneas cortas, arriba) buscando distintos caminos en un grafo *k-mero* con un algorítmo voraz. Esto resulta en una colección de *contigs* (líneas de distintas longitudes, abajo) con cada *k-mero* presente una sola vez en los *contigs*. (b) *Chrysalis* junta los *contigs* si éstos comparten al menos un *k-mero* y lee a lo largo de la unión, a partir de eso construye grafos de de Brujin individuales de cada grupo. (c) *Butterfly* toma cada uno de los grafos formados en *Chrysalis*, y corta los ejes y compacta los caminos lineales (medio). Luego, completa los grafos utilizando los *reads* limpios. La salida de de ésto es una secuencia lineal para cada forma de *splicing* y/o transcripto parálogo reflejado en el grafo. (Grabherr *et al.* 2011).

En este trabajo final de grado, se propuso el uso de herramientas bioinformáticas y del ensamblador Trinity para obtener el transcriptoma de una especie que, aunque posee una gran importancia agronómica y económica, no ha sido prácticamente estudiada en cuanto a su genética y fisiología. De esta manera, generamos información valiosa para su futuro estudio genómico e información de secuencia de genes de interés, lo que constituye una herramienta imprescindible para estudios genéticos y moleculares en la especie, aplicables a un eventual desarrollo de mecanismos de control.

HIPÓTESIS

N. viridula tiene genes conservados que codifican péptidos y receptores involucrados en procesos neurohormonales.

OBJETIVO GENERAL

Aportar información de secuencia génica de utilidad en futuros estudios genéticos, moleculares y fisiológicos en *N. viridula*, con aplicabilidad en estudios de fisiología de insectos, y en el desarrollo de mecanismos de control de próxima generación.

OBJETIVOS ESPECÍFICOS

a) Secuenciar el transcriptoma de adultos de *N. viridula*
b) Evaluar distintas aproximaciones de ensamblado a fin de seleccionar la mejor para los datos obtenidos.
c) Identificar y caracterizar familias génicas implicadas en la supervivencia de chinches fitófagas, en base a ortologías con otras especies.

Palabras clave: plagas, pentatómidos, soja, bioinformática.

MATERIALES Y MÉTODOS

Obtención y mantenimiento de la colonia de N. viridula

La colonia se estableció a partir de individuos de *N. viridula* recolectados a campo por medio de paño, red entomológica y muestreo manual, en cultivos de soja y arveja, en el partido de Pergamino, Provincia de Buenos Aires. Con los insectos obtenidos, se estableció una colonia mantenida en el insectario del Centro de Bioinvestigaciones (CeBio) a 25 °C, con un fotoperiodo de 16h luz: 8h oscuridad. Se les proporcionó papel absorbente para el desove, vainas como alimento y algodones húmedos como fuente de agua (Marco *et al.* 2014). Los huevos fueron mantenidos por separado en las mismas condiciones que los adultos y al alcanzar el estado adulto, los individuos se incorporaban a la colonia.

Extracción de ARN

Para obtener ARN total, se realizaron extracciones de cuerpo completo a partir de 10 adultos (5 machos y 5 hembras) mediante el protocolo del tiocianato de guanidinio disuelto en fenol 50%, con el reactivo comercial TRIZOL (Thermofisher). El mismo fue modificado de manera que mantuvimos el ARN a 4º C durante todo el procedimiento y todos los pasos de centrifugación se llevaron a cabo con la velocidad máxima. La precipitación en isopropanol se hizo durante toda la noche a -20º C.

El ARN obtenido se resuspendió en 100 µl de agua ultrapura estéril (Sartorius). Para determinar la calidad y cantidad del ARN, se sembró 1 µl en gel de agarosa 1% con 2 µl del agente intercalante, Bromuro de etidio (EtBr). Se sembró utilizando Formamida para minimizar las estructuras secundarias del ácido ribonucleico. Se realizó una corrida electroforética a 80V durante 40 minutos. La cuantificación se hizo a partir del gel, utilizando el programa Image J (Rueden *et al.* 2017).

Secuenciación

La genoteca fue preparada en el servicio de secuenciación, fragmentando el ARN y sintetizando cADN a partir de los fragmentos. Previo a la PCR, los fragmentos de cADN sufrieron la ligación de adaptadores (Figura 3).

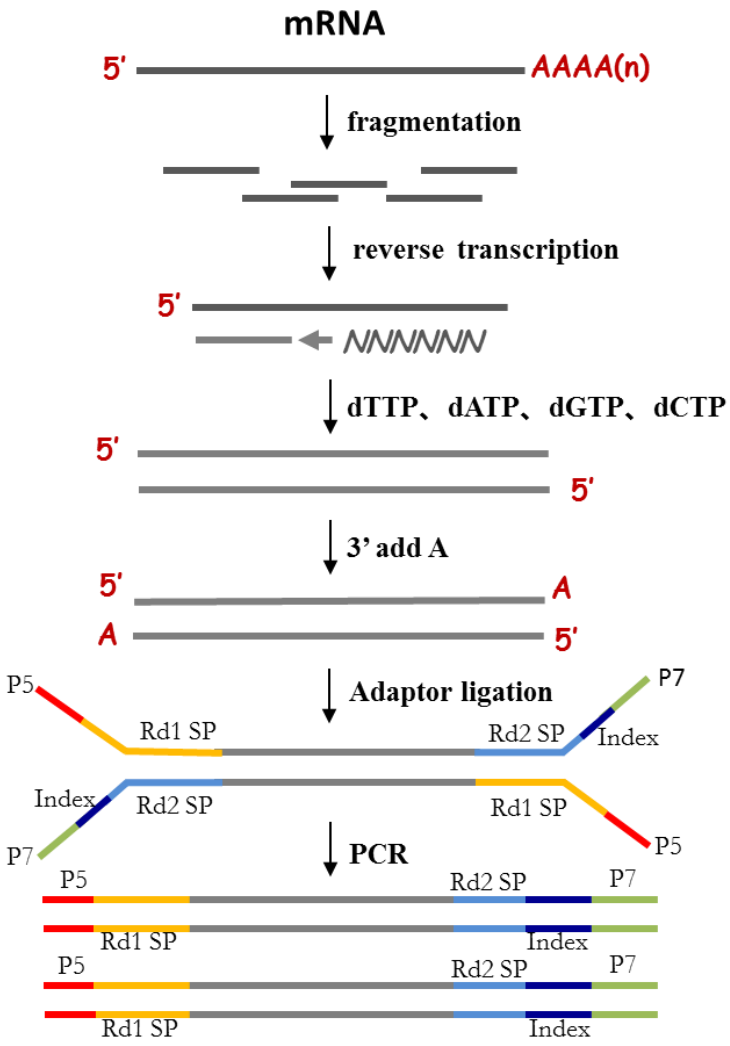

Figura 3: El ARN mensajero se fragmenta para realizar la síntesis del cADN por retrotranscripción. Se ligan los adaptadores en los extremos de los fragmentos para iniciar la PCR (Novogene).

Las muestras fueron secuenciadas con una plataforma HiSeq-2000 Ilumina (*paired end reads*, 300 millones de *reads*, 100bp de largo) en el servicio de secuenciación Novogene (China), el cual presentó las mejores condiciones de calidad y precio del mercado.

Análisis bioinformático

Limpieza de datos

La calidad de las lecturas crudas se determinó utilizando el programa FastQCtoolkit (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Con el mismo programa, se eliminaron todas las secuencias de calidad menor a Q30 y adaptadores.

La búsqueda de secuencias contaminantes, como vectores y *primers*, se realizó mediante una comparación de los datos con la base de datos de UniVec del NCBI (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec), utilizando la herramienta BLAST (Altschul *et al.* 1990) para llevarlo a cabo. Las lecturas que dieron *hit* utilizando un *E-value* menor o igual a 1×10^{-5} fueron eliminados del conjunto de datos por considerarse secuencias contaminantes. Debido a que los datos eran *paired end,* todas las secuencias eliminadas de una corrida fueron eliminadas también de la otra, a fin de favorecer el correcto ensamble de las lecturas.

Se utilizó un *script* codficado en el lenguaje *python* para eliminar las secuencias no deseadas del total de *reads* (Anexo I).

Ensamble

Los *reads* limpios se ensamblaron con el programa Trinity (Grabherr *et al.* 2011), especificando que la información a utilizar era *pair end* y utilizando un máximo de 500 Gb de memoria por medio de acceso remoto a un servidor presente en el instituto Max Planck de Göettingen, Alemania. Se ensamblaron dos transcriptomas con este ensamblador, modificando el parámetro de cobertura mínima de *k-mer* (min_kmer_cov). Por *default*, el programa utiliza un valor de 1 para este parámetro. En el segundo ensamble, se utilizó un valor de 2, recomendado en Grabherr y colaboradores (2011) para los conjuntos de datos grandes. Esto evitaría que se agreguen los *kmers* únicos, ricos en errores de secuenciación.

Métricas

-CEG y BUSCO. Para evaluar el grado de completitud de los ensambles generados, se calculó el porcentaje de representatividad de los genes codificantes de proteínas. Para esto, se tradujo el ensamble en los seis posibles marcos de lectura, con la herramienta transeq (http://emboss.sourceforge.net/apps/cvs/emboss/apps/transeq.html). Se realizó una búsqueda por Hmmer de los transcriptos traducidos con el perfil de CEG (core eukaryotic genome) (Parra *et al.* 2007), compuesto por 458 proteínas, con el comando *hmmscan* y eliminando aquellas secuencias con puntaje menor a 40, tal como se describe en Martínez-Bernetche y colaboradores (2012).

El mismo procedimiento se llevó a cabo con el perfil de BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão *et al.* 2015) de artrópodos, el cual está compuesto por 2676 proteínas. De acuerdo a los resultados obtenidos en base a la completitud y al número de genes de cada transcriptoma, se eligió uno de los dos ensambles para continuar el análisis.

-Uniprot: A fin de evaluar la calidad del ensamble, se comparó el transcriptoma ensamblado con la base de datos de proteínas de Uniprot-Swissprot (Suzek *et al.* 2007). Se realizó una búsqueda en la base de datos UNIPROT mediante el programa BLASTx (Altschul *et al*. 1990) con parámetro de *E-value* de $1x10^{-5}$ y obteniendo el mejor *hit* para cada transcripto.

-Búsqueda de ortólogos. El transcriptoma se contrastó con bases de datos de genes neurales de *Rhodnius prolixus* y de neuropéptidos de *R. prolixus* y otros insectos relacionados, como *D. melanogaster, Bombyx mori* y *Tribolium castaneum* (Mezquita *et al.* 2015; Ons *et al.* 2011; Ons 2017) para observar presencia o ausencia de expresión de estos grupos génicos. Esto se hizo utilizando la herramienta BLASTx del NCBI (Altschul *et al.* 1990) y con un *E-value* de $1x10^{-5}$, obteniendo el mejor *hit* únicamente. 3 neuropéptidos conocidos, no fueron hallados en Uniprot para los organismos aquí utilizados y por lo tanto, no se añadieron a la base de datos generada.
Los neuropéptidos hallados en el transcriptoma fueron clasificados de acuerdo a la función que cumplen, siguiendo el trabajo de Ons (2015).

RESULTADOS

Extracción de ARN

A partir de 10 individuos adultos, 5 hembras y 5 machos, se realizó eficientemente la extracción de ARN. Las muestras fueron sometidas a electroforesis en gel de agarosa y en base al resultado del gel (Figura 4), se utilizaron las muestras sembradas en las calles 1 y 4, las cuales tenían una concentración de 0,267 µg/µl de ARN y mostraban el menor grado de degradación. Ambas muestras fueron enviadas al servicio de secuenciación, donde después de un análisis de calidad se seleccionó la muestra correspondiente a la calle 4.

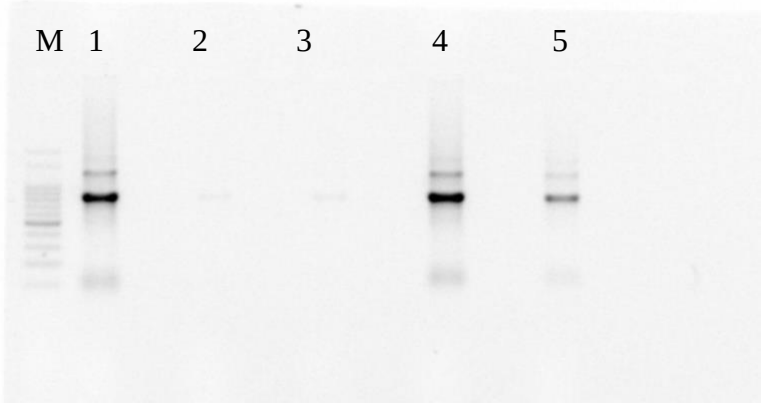

Figura 4: Gel de agarosa 1%, revelado con BrEt. Las calles 1 y 4 son las que muestran una mayor concentración.

Secuenciación

Cada corrida de secuenciación (*paired end*) resultó en 280443227 *reads*, donde el 96,72% de las secuencias tuvieron una calidad igual o mayor a Q30. La secuenciación se realizó con un 0,01% de error y una efectividad del 99,69%. El porcentaje de GC fue del 36,34% (Tabla 1).

Limpieza de datos

Se eliminaron 2009838 secuencias en total de ambas corridas, teniendo en cuenta las secuencias de baja calidad, los adaptadores y las secuencias contaminantes, correspondientes a vectores y *primers*, para ambos extremos de la corrida *paired end*.

Ensamble

A partir de los datos limpios, se generaron dos ensambles por medio del programa Trinity, cada uno de ellos con un *kmer* mínimo distinto, donde se denominó DN2 al transcriptoma generado con un min_kmer_cov=1 y DN3 al realizado con el parámetro min_kmer_cov=2. De esta manera, se obtuvieron transcriptomas diferentes, con número variable de genes (260.318 para DN2 y 221.206 para DN3), bases ensambladas (219.003.449 para DN2 y 193.263.233 para DN3), total de transcriptos (343.745 para DN2 y 299.148 para DN3) y longitud promedio de *contig* (324 para DN2 y 329 para DN3), entre otros valores (Tabla 2).

Métricas

La completitud observada en ambos ensambles para el CEG (458 para DN2 y 457 para DN3) y el BUSCO (2.652 para DN2 y 2.527 para DN3) nos demuestra que los dos transcriptomas generados tienen una cobertura alta, siendo el ensamble DN2 el que presenta una mayor cobertura, con una diferencia de 4,7% en genes BUSCO hallados (Tabla 2).

Entre ambos transcriptomas, hay una diferencia de 34.000 transcriptos (343.745 para DN2 y 299.148 para DN3), que se traduce solamente en una pérdida de 1500 hits de Uniprot (15.880 para DN2 y 14.375 para DN3) y de pocos perfiles BUSCO (2.652 para DN2 y 2.527 para DN3). Por el otro lado, DN3 presenta un número superior de *contigs* N50 (1.093 contra 1.061 de DN2), así como un *contig* promedio (646,05 contra 637,11 en DN2) mayor (Tabla 2).

En base a los resultados, consideramos que el ensamble más apropiado para estudios genéticos posteriores es DN3, puesto que si bien el grado de cobertura de ambos es alto, consideramos de mayor importancia la eliminación de *kmers* únicos (que presentan frecuentemente errores de secuenciación) y de secuencias redundantes, así como la incorporación de *contigs* más largos.

Los genes de interés y de baja expresión involucrados en la supervivencia de *N. viridula* se encuentran en ambos ensambles generados, lo cual es un indicio de la calidad de ambos.

Tabla1: métricas de secuenciación

Muestra	NV1_2
reads crudos	280.443.227
reads limpios	279.564.558
Datos crudos(G)	56.088.645.400
Datos limpios(G)	55.912.911.600
Efectividad(%)	99,69
Error(%)	0,01
Q20(%)	98,63
Q30(%)	96,72
GC(%)	36,34

Tabla2: métricas de los transcriptomas obtenidos

Ensamble	DN2 Trinity (min_kmer_cov=1)	DN3 Trinity (min_kmer_cov=2)
Porcentaje de GC	32,47	32,53
Longitud prom. Contig	324	329
Contig promedio	637,11	646,05
N50	1.061	1.093
Total de genes trinity	2603.18	221206
Total de transcriptos	343.745	299.148
Bases ensambladas	219.003.449	193.263.233
Contig más largo	16.931	15.451
Perfiles CEG	458/459	457/459
Perfiles BUSCO	2652/2676	2527/2676
BLASTx vs Uniprot	15.880 hits únicos	14.375 hits únicos
Neuropéptidos	39/40	39/40
Genes neurales	70/103	69/103

En el caso de los genes precursores de neuropéptidos (Tabla 3), hemos podido inferir la identidad de todos los genes precursores presentes en *R. prolixus*, con excepción de la

hormona de la eclosión. Además, se identificaron precursores de PTTH, de la hormona adipokinética II y de IMF-amida, los cuáles no se encontraron en *R. prolixus.*

La trisina no fue hallada en el transcriptoma generado, mientras que el resto de los neuropéptidos ausentes (Inotocina, *Long Neuropeptide F2* y *Neuropeptide-like precursor 2*), no se encontraban en la base de datos, dado que sus secuencias no se encuentran disponibles en Uniprot.

Tabla 3: Neuropéptidos presentes en el transcriptoma de *N. viridula* y en otras especies

Neuropéptidos	***Bombyx mori***	***Drosophila melanogaster***	***Rhodnius prolixus***	***Tribolium castaneum***	***Nezara viridula***
ACP	+	-	+	+	+
Hormona adipokinética 1	+	+	+	+	+
Hormona adipokinética 2	+	-	-	+	-
Alatotropina	+	-	+	+	+
AST-CC	+	+	+	+	+
Hormona diurética simil-calcitonina	+	+	+	+	+
CAPA	+	+	+	+	+
Péptido CCH-amida	+	+	+	+	+
Péptido CNM-amida	-	+	+	+	+
Corazonina	+	+	+	-	+
Hormona diruética simil-CRF	+	+	+	+	+
Péptido cardioactivo crustaceo	+	+	+	+	+
Hormona de eclosión	+	+	+	+	+
Elevenina	-	+	+	+	+
ETH	+	+	+	+	+
FLP	+	+	+	+	+
AST FGL-amida	+	+	+	-	+
IMF-amida	+	-	-	-	-
Inotocina	-	-	-	+	-
Kinina de insectos	+	+	+	-	+
Péptidos simil-Insulina	+	+	+	+	+
Transporte Iónico. Péptido A	+	+	+	+	+
Símil-ITG	+	-	+	+	+
Neuropéptido Largo F1	+	+	+	-	+
Neuropéptido Largo F2	+	-	-	-	-
Péptido mioinhibidor	+	+	+	+	+
Miosupresina	+	+	+	+	+
Natalisina	+	+	+	+	+

Neuroparsina A	+	-	+	+	+
Precursor símil-neuropéptido 1	+	+	+	+	+
Precursor símil-neuropéptido 2 to 4	-	+	-	-	-
Símil-NVP (Prohormona 2)	-	-	+	+	+
Orcokinina A	+	+	+	+	+
Orcokinina B	+	+	+	+	+
PBAN	+	+	+	+	+
Factor dispersante de pigmento	+	+	+	-	+
PISCF AST	+	+	+	+	+
Proctolina	+	+	+	+	+
PTTH	+	+	-	+	-
RYamida	+	+	+	+	+
Neuropéptido corto F	+	+	+	+	+
SIF-amida	+	+	+	+	+
Sulfakininas	+	+	+	+	+
Taquikininas	+	+	+	+	+
Trisina	+	+	-	+	-

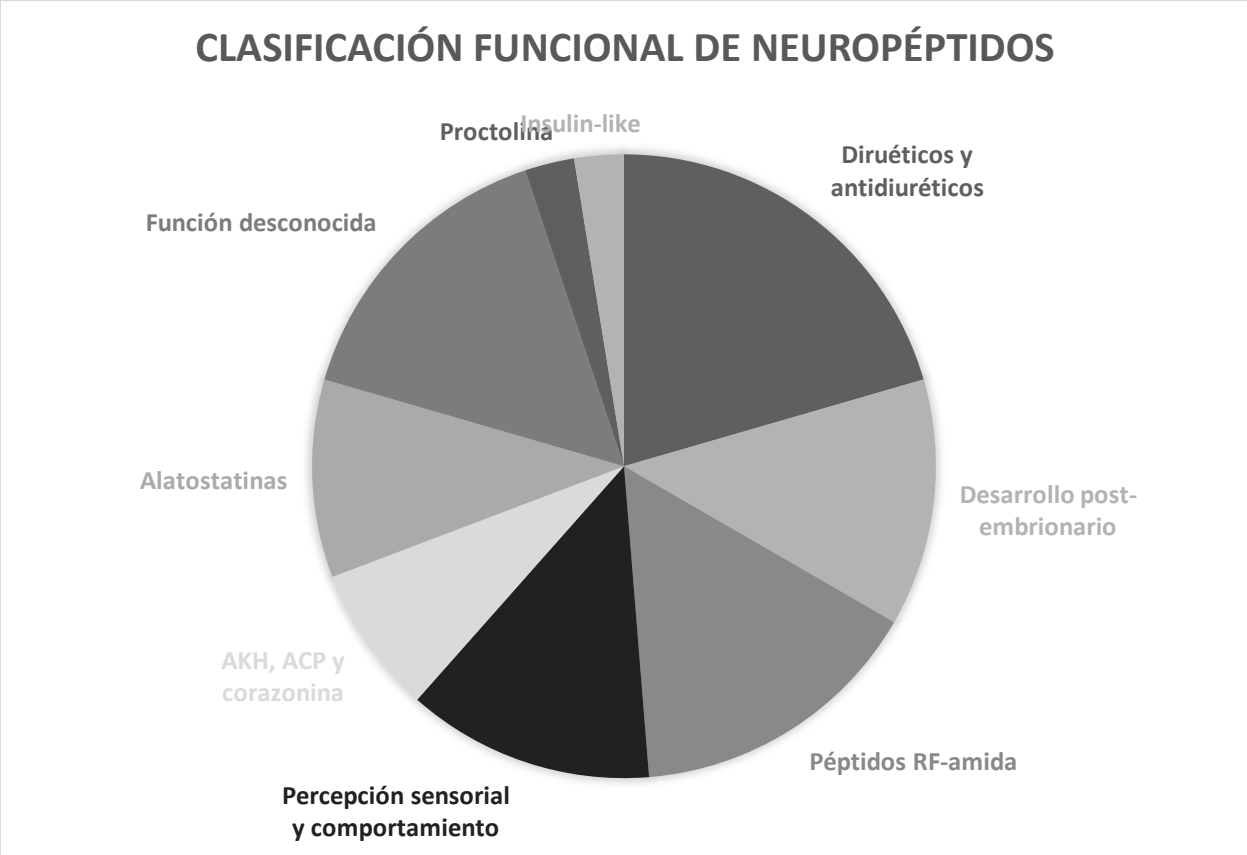

Figura 5: clasificación funcional de los neuropéptidos hallados en el ensamble

En cambio, para los genes nerviosos(Tabla 4), hallamos una menor cantidad que la reportada en *R. prolixus.* El gen hallado en el ensamble DN2, ausente en DN3 (_FMRFamide_receptor_RPRC001551) se encuentra anotado en *R. prolixus* como un receptor acoplado a proteína G (GPCR), similar al GPCR de la rodopsina.

Se observó también la presencia del receptor del péptido sexual (sex_peptide_receptor__RPRC000605), no así la del péptido sexual.

Tabla 4: genes neurales en los transcriptomas de *N. viridula*

Genes neurales de Rp	***DN2***	***DN3***
>5hydroxytryptamine_receptorlike__RPRC008923	+	+
>ACP_I_receptor_Armado_(RPRC000057+RPRC004783)	+	+
>ADIPORlike_receptor_(RPRC012657)	-	-

>AKH	+	+
>CCAP_receptor_Armado_(RPRC000969+RPRC012063)	-	-
>CCAP_receptor_RPRC001248	+	+
>CCHa/_gastrinreleasing_peptide_receptor__RPRC000608	+	+
>CCHa__gastrinreleasing_peptide_receptorlike_II_[Nasonia_vitripennis]	-	-
>ETHr_(RPRC008652)	-	-
>ETHr__RPRC000848	-	-
>Gprotein_coupled_receptor_moodylike_RPRC011268	+	+
>Octopamine_receptor_beta1R_armado_(RPRC001507+_RPRC005349)	+	+
>Opsin_armado_(RPRC001049+RPRC001048)	+	+
>PDF_receptor_RPRC009680	+	+
>Proctolin/FMRFa_receptor_(rp_asb64986)	+	+
>Prolactinreleasing_peptide_receptor_RPRC002266+RPRC002268+RPRC002269	+	+
>RPRC003273	+	+
>RPRC010865	-	-
>RPRC012816	-	-
>SK_receptor_Armado	+	+
>TK2__(RPRC001687)	-	-
>TK_receptor_armado	+	+
>TRHlike_receptor_Armado	-	-
>Trace_amineassociated_receptor_4_RPRC002007	-	-
>UVopsin2__RPRC002621	+	+
>_5hydroxytryptamine_receptor_1A_RPRC010656	+	+
>_ACP_II/Gonadotropinreleasing_hormone_II_receptor_RPRC000523	+	+
>_ALS_C/somatostatin_receptor_type_2like_RPRC013486	+	+
>_Activin_receptor_type2B_rp_asb18166_(RPRC008684)	+	+
>_CG13597like_amine_receptor_RPRC004409_rp_asb31126	+	+
>_FMRFamide_receptor_RPRC001551	+	-
>_FMRFamide_receptor_rp_asb64986	-	-
>_Gprotein_coupled_receptor_52__RPRC014528	+	+
>_Gprotein_coupled_receptor_moody_RPRC004128	+	+
>_Lowdensity_lipoprotein_receptorrelated_protein_2_RPRC000281	+	+
>_Lutropinchoriogonadotropic_hormone_receptor_RPRC001663	-	-
>_Neuropeptide_FF/SIFa_receptor_2_RPRC013738	-	-
>_Neuropeptide_FF_receptor_2_RPRC004565	-	-
>_Octopamine_receptor_beta2R_RPRC011545	+	+
>_Parathyroid_hormone/parathyroid_hormonerelated_peptide_receptor_RPRC011083	-	-
>_RPRC015285_diuretic_hormone_receptor_(DH44)_	-	-

>_Sortilinrelated_receptor_RPRC000270	+	+
>_adenosine_receptor_A2b_RPRC006842	+	+
>_alpha1D_adrenergic_receptorlike_RPRC011623	+	+
>_class_A_rhodopsinlike_Gprotein_coupled_receptor_GPR5ht3_RPRC007788 +RPRC001792	+	+
>_class_A_rhodopsinlike_Gprotein_coupled_receptor_GPRadr1__RPRC01545 6	-	-
>_class_B_secretinlike_Gprotein_coupled_receptor_GPRmth1_rp_asb12126	+	+
>_diuretic_hormone_receptor_(DH44)_RPRC000578	+	+
>_dopamine_D2like_receptor_RPRC000473	+	+
>_dopamine_receptor_2_RPRC013708	+	+
>_dopamine_receptor__RPRC011175	-	-
>_growth_hormone_secretagogue_receptor_type_1_RPRC001428	+	+
>_insulinlike_receptor_(Falta_Aminoterminal)	+	+
>_lowdensity_lipoprotein_receptor_RPRC000060	+	+
>_muscarinic_acetylcholine_receptor_gar2like_RPRC010907	+	+
>_myosuppressin_receptor_GL563092	-	-
>_neuropeptide_FF_receptor_(SIFa)(_RPRC000835)	+	+
>_neuropeptide_GPCR_A23_RPRC000494	+	+
>_neuropeptide_GPCR_A37_RPRC004793	+	+
>_neuropeptide_GPCR_A47_RPRC014721	+	+
>_neuropeptide_Y_receptor_type_2__RPRC008364_	+	+
>_neuropeptide_Y_receptor_type_2like_RPRC00203	+	+
>_neuropeptide_Y_receptor_type_2like_RPRC008894	+	+
>_nicotinic_acetylcholine_receptor_alpha9_subunit_precursor_rp_asb25893	+	+
>_octopamine_receptor_RPRC001341	+	+
>_opsin,_ultravioletsensitivelike_RPRC015283	+	+
>_oxidized_lowdensity_lipoprotein_receptor_1__rp_asb62172	+	+
>_ryanodine_receptor__RPRC013986	+	+
>_sex_peptide_receptor__RPRC000605	+	+
>_thyrotropin_receptorlike_rp_asb14748	+	+
>_tyramine_receptor_2_RPRC008712	-	-
>_vitellogenin_receptor_RPRC000551	+	+
>allatostatin_A_receptor_RPRC004705	+	+
>allatostatin_A_receptor_RPRC004706	-	-
>allatostatin_A_receptor_RPRC004708	-	-
>calcitoninlike_diuretic_hormone_receptor_DH31_GL561032	-	-
>dopamine_receptor_1___RPRC014093	+	+
>gi|295684930|gb|ADG27752.1|_CAPA_receptor_variant_A_[Rhodnius_proli xus]	+	+

>gi|295684932|gb|ADG27753.1|_CAPA_receptor_variant_B_[Rhodnius_proli xus]	+	+
>gi|398314269|gb|AFO73269.1|_pyrokinin1_receptor_variant_A_[Rhodnius_ prolixus]	+	+
>gi|398314271|gb|AFO73270.1|_pyrokinin1_receptor_variant_B_[Rhodnius_ prolixus]	+	+
>gi|398314273|gb|AFO73271.1|_pyrokinin1_receptor_variant_C_[Rhodnius_ prolixus]	-	-
>gi|564583205|gb|AHB86317.1|_calcitoninlike_diuretic_hormone_receptor_ variant_B_[Rhodnius_prolixus]	+	+
>gi|564583207|gb|AHB86318.1|_calcitoninlike_diuretic_hormone_receptor_ variant_C_[Rhodnius_prolixus]	+	+
>gi|564583824|gb|AHB86571.1|_calcitoninlike_diuretic_hormone_receptor_ 2_[Rhodnius_prolixus]	+	+
>hormone/parathyroid_hormonerelated_peptide_receptor_like_isoform_X2_ rp_asb27633_RPRC011086	+	+
>lowdensity_lipoprotein_receptor__RPRC000138	+	+
>muscarinic_acetylcholine_receptor_DM1_armado_(RPRC001750+RPRC0017 51+RPRC007566)	+	+
>neuropeptide_F_receptor_RPRC008140	-	-
>neuropeptide_GPCR_A27_GL548735	-	-
>neuropeptide_GPCR_A6a_GL559380	+	+
>neuropeptide_receptor_(neuropeptide_Y_o_neuromedinK_receptor_o_LK)_ RPRC008570	-	-
>octopamine_alpha_receptor_GL561185	+	+
>octopamine_receptor_beta3R_armado_(RPRC014610+RPRC001054)	+	+
>oxidized_lowdensity_lipoprotein_receptor_1	-	-
>pheromone_biosynthesisactivating_neuropeptide_RPRC008528	-	-
>putative_diuretic_hormone_receptor_II_(DH44)_GL563066	-	-
>putative_diuretic_hormone_receptor_I_(DH44)_GL562334	-	-
>pyroglutamylated_RFamide_peptide_receptor_(rp_asb67882)	+	+
>serotonin_receptor_2_alpha_armado_(RPRC005858+RPRC001892)	+	+
>serotonin_receptor_RPRC010931	+	+
>sulfakinin_receptor_GL545664	-	-
_>_Opsin1_RPRC010623	-	-

DISCUSIÓN

Secuenciación

La utilización de RNA-Seq como herramienta para obtener conocimiento genético de *N. viridula* fue particularmente beneficiosa, principalmente porque esta tecnología no requiere conocimiento genómico previo. Dado que la especie aquí estudiada no constituye un organismo modelo, sus secuencias genómicas son actualmente desconocidas. Sin embargo, los datos que se generaron en el transcruso de este trabajo pueden proveer información sobre variaciones de secuencia, localización de límites de transcripción y relaciones entre exones (Wang *et al.* 2009).

Una decisión clave a tomar es la elección de la plataforma de secuenciación a utilizar. Al tratar con un organismo cuyo genoma no se encuentra disponible, se debió recurrir a un ensamble *de novo*, teniendo las consideraciones necesarias para elegir la modalidad de secuenciación. La divergencia genética dentro de un mismo orden de la clase Insecta es demasiado grande (Zdobnov & Bork 2007) para que un genoma específico pueda ser usado como referencia para otros del mismo orden (Ewen-Campen *et al.* 2011).

Por lo tanto, dado que en los últimos tiempo el costo por base secuenciada ha decaído a valores accesibles, la tecnología de elección debe ser la que mejor se adapte a los resultados que se busca obtener. En este trabajo se utilizó un equipo Illumina HiSeq 2000, que permitió obtener cerca de 600 millones de *reads pair end* de 100 pb de longitud. De esta manera, la profundidad de secuenciación alta y la longitud de los *reads*, sumado a la secuenciación *pair end,* fue suficiente para poder realizar un ensamble *de novo* de buena calidad (Garber *et al.* 2011*;* Katz *et al.* 2010). Además, en el caso de otras plataformas, como PacBio, que producen *reads* más largos, el error de secuenciación es mayor y deben completarse los experimentos con *reads* más cortos de Illumina (Au *et al.* 2012).

De esta forma, consideramos las condiciones óptimas de secuenciación en función de alcanzar los objetivos particulares propuestos. Los resultados obtenidos confirmaron que la metodología de secuenciación seleccionada fue adecuada.

Ensamble

Múltiples trabajos previos demuestran que la mejor estrategia o el mejor programa de ensamblado, está relacionada directamente con el conjunto de datos particular y con el

objetivo de la investigación (Schulz *et al.* 2012; Wang & Gribskov 2016). Es decir, ningún programa de ensamblado es superior al resto, sino que cada uno tiene ventajas y desventajas a considerar según los objetivos del experimento. En nuestro laboratorio se han realizado trabajos con objetivos similares a los propuestos aquí con otras especies de insectos, que han indicado que, en nuestras condiciones, Trinity resultó superior a otros programas tales como VELVET-Oasis o AbySS, tanto por el grado de cobertura como por el número de transcriptos ensamblados (Palacio, no publicado). En base a esa experiencia, en lugar de analizar distintos programas de ensamblado, para el transcriptoma de *N. viridula* presentado se decidió ensayar distintos parámetros de Trinity, a fin de seleccionar el que diera mejores resultados.

Métricas

En los ensambles que se realizaron, se decidió atribuir un mayor peso a la eliminación de errores de secuenciación, a costa de resignar la aparición de algunos pocos genes más, considerando que ambos transcriptomas generados presentan un alto grado de cobertura. De esta manera, se tomó la decisión de seleccionar el ensamble con menos bases ensambladas, donde se eliminan los *kmers* únicos, los cuales frecuentemente contienen errores de secuenciación (Grabherr *et al.* 2011); considerando además, la obtención de mayores *contigs* que podrían resultar más informativos. Los resultados obtenidos al comparar los transcriptomas, confirman que la estrategia utilizada fue acertada para el análisis que se llevó a cabo.

-Búsqueda de ortólogos

En la búsqueda de neuropéptidos, comparando con la base de datos de *R. prolixus*, observamos que el único ausente corresponde a la hormona de la eclosión. Una posible explicación para esto, es que habiendo utilizado adultos, la misma ya no se encuentre expresada. Lo mismo sucede con los genes neurales, de los cuales encontramos únicamente el 67%. Teniendo en cuenta que parte de estos genes realizan su función principal durante el desarrollo embrionario, creemos que es posible que estos genes hallados en *R. prolixus* se encuentren también en *N. viridula,* aunque no los hayamos identificado en los ensambles realizados a partir de individuos adultos. Para demostrar

lo anterior debería repetirse todo el trabajo aquí descrito utilizando estadios tempranos, tanto huevos como ninfas.

Adicionalmente, se identificó al receptor del péptido sexual aun sin haber hallado al péptido sexual. Se ha reportado en *D. melanogaster* que este receptor puede estar, tanto en hembras como en machos, cumpliendo roles adicionales a su función canónica sobre el comportamiento post-cópula de las hembras(Yapici *et al.* 2008). Se ha demostrado que este receptor es promiscuo, tanto en *D. melanogaster* como en *B. mori*, y que los péptidos mioinhibidores funcionan como ligando para el mismo (Kim *et al.* 2010; Yamanaka *et al.* 2010; Poels *et al.* 2010). De esta manera, podemos pensar que los péptidos mioinhibidores de *N. viridula* hallados en este transcriptoma son los responsables de activar al receptor en este organismo.

Por último, debemos considerar que los resultados *in silico* no deben reemplazar a la experimentación *in vivo*. Por lo tanto, los resultados aquí obtenidos por medio de programas bioinformáticos, serán complementados con la validación biológica correspondiente.

CONCLUSIÓN

En este trabajo se obtuvieron dos transciptomas de *N. viridula,* los cuales fueron analizados de acuerdo a su grado de cobertura y para los cuales se evaluó la presencia de genes involucrados en la supervivencia, como lo son neuropéptidos y genes neurales. De esta manera, se generó información genética valiosa para continuar los estudios sobre este organismo de gran importancia agronómica.

AGRADECIMIENTOS

El proyecto “Identificación de genes blanco para el control de chinches *Nezara viridula* y *Dichelops furcatus*” (PICT-2015-0468/Resolución ANPCyT 240/16), dirigido por la doctora Sheila Ons financió este trabajo.

Además, me fue otorgada una beca de estímulo a las vocaciones científicas por el Consejo Interuniversitario Nacional (CIN) con este mismo proyecto. Ambos financiamientos hicieron posible la realización de esta tesis de grado que es un punto de partida para un proyecto mayor.

Agradezco además a todos los que hacen al CeBIO y particularmente, a mi grupo de trabajo, por ayudarme a solucionar todos los problemas que aparecieron en el transcurso de este tiempo y me acompañaron tanto desde lo laboral como desde lo personal.

Por último, fue el apoyo de mi familia y de mis amigos cercanos, el que me permitió llegar hasta este lugar. No solamente aseguraron mi desarrollo académico, pero también me acompañaron paso a paso e hicieron del trayecto una experiencia inolvidable.

BIBLIOGRAFÍA

Altschul, SF; Gish, W; Miller, W; Myers, EW; Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology. 215(3):403-10.

Aragón, JR; Molinari, A; Lorenzati De Diez, S. 1997. Manejo Integrado de Plagas. El cultivo desoja en Argentina. Giorda, L. M. y Baigorri, H (eds). Marcos Juarez. INTA. EEA Marcos Juarez. Agro de Córdoba Nº4: 248-288.

Au, KF; Underwood, JG; Lee, L; Wong WH. 2012. Improving PacBio Long Read Accuracy by Short Read Alignment. PLoS ONE. *7*(10): e46679.

Bentley, DR. 2006. Whole-genome re-sequencing. Current Opinion in Genetics & Development. 16(6): 545–552.

Bimboni, HG. 1978. Daños producidos por distintas densidades de población de chinche verde Nezara viridula (L.). IDIA nº 361-366. Enero – Junio: 76-82.

Birol, I; Jackman, SD; Nielsen, CB; Qian, JQ; Varhol, R; Stazyk, G; Moryn, RD; Zhao, Y; Hirst, M; Schein, JE; Horsman, SDE; Connors, JM; Gascoyne, RD; Marra, MA; Jones, SJM. 2009. De novo transcriptome assembly with ABySS. Bioinformatics. 25(21): 2872–2877.

Boguski, MS; Tolstoshev, CM; Bassett, DE. 1994. Gene discovery in dbEST. Science (New York, N.Y.). 265(5181): 1993–4.

Conesa, A; Madrigal, P; Tarazona, S; Gomez-Cabrero, D; Cervera, A; McPherson, A; Mortazavi, A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology. 17(1):13.

De Brujin, N. 1946. A Combinatorial Problem. Koninklijke Nederlandse Akademie v Wetenschappen. 46: 758–764.

Ekblom, R; Wolf, JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications. *7*(9): 1026–42.

Ellegren, H; Smeds, L; Burri, R; Olason, PI; Backström, N; Kawakami, T; Künstner, A; Mäkinen, H; Nadachowska-Brzyska, K; Qvarnström, A; Uebbing, S; Wolf, JB. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature. 491, 756–760.

Ewen-Campen, B; Shaner, N; Panfilio, KA; Suzuki, Y; Roth, S; Extavour, CG. 2011. The maternal and early embryonic transcriptome of the milkweed bug Oncopeltus fasciatus. BMC Genomics. 12(1): 61.

Fedurco, M; Romieu, A; Williams, S; Lawrence, I; Turcatti, G. 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Research. 34(3): e22.

Fuentes, FH; Distéfano, S; Lenzi, L; Pesaresi, E; Cuniberti, M; Herrero, R; Muñoz, S; Gadbán, L; Craviotto, RM; Gallo, C; Perearnau, AM; Fared, M; Montero, M; Belluccini, P; Francioni, F; Flores, F; Balbi, E; Nicollier, M; Le Roux, M; Gentili, OA; Murgio, M; Vissani, C; Gilli, J; Bernardi, C; Rojas, E; Ferraris, G; Mousegne, F; Díaz, ZM; Ghida DC; Conde, MB; Lizondo, M; Giménez, F; Berra, O; Macagno, S; Pronotti, M; Chialvo, E; Muñoz, S; Pagnan, L; Cottura, G; Errasquin, L; Giordano, M; Verdelli, D; Carrio, A; Heredia, A; Salines, L. 2016. SOJA, Actualización 2016: Informe de Actualización Técnica en línea Nº6. INTA Ediciones.

Gamundi, JC; Sosa, MA. 2007. Caracterización de daños de chinches en soja y criterios para la toma de decisiones de manejo. E.V. TRUMPER & J.D. EDELSTEIN (eds), Chinches fitófagas en soja. Revisión y avances en el estudio de su ecología y manejo, Ediciones INTA, Manfredi.

Garber, M; Grabherr, MG; Guttman, M; Trapnell, C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods. 8(6): 469–477.

Gerhard, DS; Wagner, L; Feingold, EA; Shenmen, CM; Grouse, LH; Schuler, G. 2004. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Research. 14(10B): 2121–2127.

Good, IJ. 1946. Normal Recurring Decimals. Journal of the London Mathematical Society. s1-21(3): 167–169.

Goodwin, S; McPherson, JD; McCombie, WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6): 333–351.

Grabherr, MG; Haas, BJ; Yassour, M; Levin, JZ; Thompson, DA; Amit, I; Adiconis, X; Fan, L; Raychowdhury, R; Zeng, Q; Chen, Z; Mauceli, E; Hacohen, N; Gnirke, A; Rhind, N; di Palma, F; Birren, BW; Nusbaum, C; Lindblad-Toh, K; Friedman, N; Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7), 644–52.

Guttman, M; Garber, M; Levin, JZ; Donaghey, J; Robinson, J; Adiconis, X; Fan, L; Koziol, MJ; Gnirke, A; Nusbaum, C; Rinn, JL; Lander, ES; Regev, A. 2010. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature Biotechnology, 28(5): 503–510.

Haas, BJ; Zody, MC. 2010. Advancing RNA-Seq analysis. Nature Biotechnology. 28(5): 421–423.

Katz, Y; Wang, ET; Airoldi, EM; Burge, CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods. *7*(12): 1009–1015.

Kim, YJ; Bartalska, K; Audsley, N; Yamanaka, N; Yapici, N; Lee, JY; Kim, YC; Markovic, M; Isaac, E; Tanaja, Y; Dickson, BJ. 2010. MIPs are ancestral ligands for the sex peptide receptor. Proc Natl Acad Sci U S A. 107: 6520–6525.

Korinek, J; Spelsberg, TC; Mithcell, WM. 1973. mRNA transcription linked to the morphological and plasma membrane changes induced by cyclic AMP in tumour cells. Nature. 246(5434): 455-8.

Lamichhaney, S; Barrio, AM; Rafati, N; Sundström, G; Rubin, CJ; Gilbert, ER; Berglund, J; Wetterbom, A; Laikre, L; Webster, MT; Grabherr, M; Ryman, N; Andersson, L. 2012. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences.109: 19345–19350.

Lorenz, LJ; Hall, JC; Rosbash, M. 1989. Expression of a Drosophila mRNA is under circadian clock control during pupation. Development. 107(4): 869-80.

Luna, MJ; Iannone, N. 2013. Efecto de la chinche de los cuernos “Dichelops furcatus” (F.) sobre la calidad de la semilla de soja. Revista de la Facultad de Agronomía La Plata. 112: 141-145.

Marco, HG; Anders, L; Gade, G. 2014. cDNA cloning and transcript distribution of two novel members of the neuroparsin peptide family in a hemipteran insect (Nezara viridula) and a decapod crustacean (Jasus lalandii). Peptides. 53: 97-105.

Margulies, M; Egholm, M; Altman, WE; Attiya, S; Bader, JS; Bemben, LA; Berka, J; Braverman, MS; Chen, YJ, Chen, Z; Dewell, SB; Du, L; Fierro, JM; Gomes, XV; Godwin, BC; He, W; Helgesen, S; Ho, CH; Irzyk, GP, Jando, SC; Alenquer, ML; Jarvie, TP; Jirage, KB; Kim, JB, Knight, JR; Lanza, JR; Leamon, JH; Lefkowitz, SM; Lei, M; Li, J; Lohman, KL; Lu, H; Makhijani, VB; McDade, KE; McKenna, MP; Myers, EW; Nickerson, E; Nobile, JR; Plant, R;Puc, BP; Ronan, MT; Roth, GT, Sarkis, GJ; Simons, JF; Simpson, JW; Srinivasan, M; Tartaro, KR; Tomasz, A; Vogt, KA; Volkmer, GA; Wang, SH; Wang, Y;

Weiner, MP; Yu, P; Begley, RF, Rothberg, JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437: 376–380.

Martínez-Barnetche, J; Gómez-Barreto, RE; Ovilla-Muñoz, M; Téllez-Sosa, J; López, DE; Dinglasan, RR; Mohien, CU; MacCallum, RM; Redmond, SN; Gibbons, JG; Rokas, A; Machado, CA; Cazares-Raga, FE; González-Cerón, L; Hernández-Martínez, S; López, MH. 2012. Transcriptome of the adult female malaria mosquito vector Anopheles albimanus. BMC Genomics. 13(1): 207.

Mesquita, RD; Vionette-Amaral, R; Lowenberger, C; Rivera-Pomar, R; Monteiro, FA; Minx, P; Spieth, J; Carvalho, AB; Panzera, F; Lawson, D; Torres, ALQ; Ribeiro, JMC; Sorgine, MHF; Waterhouse, RM; Montague, MJ; Abad-Franch, F; Alves-Bezerra, M; Amaral, LR; Araujo, HM; Araujo, RN; Aravind, L; Atella, GC; Azambuja, P; Berni, M; Bittencourt-Cunha, PR; Braz, GRC; Calderón-Fernández, GM; Carareto, CMA; Christensen, MB; Costa, IR; Costa, SG; Dansa, M; Daumas-Filho, CRO; De-Paula, IF; Dias, FA; Dimopoulos, G; Emrich, SJ; Esponda-Behrens, N; Fampa, P; Fernández-Medina, RD; Fonseca, RN; Fontenele, M; Fronick, C; Fulton, LA; Gandara, ACP; Garcia, ES; Genta, FA; Giraldo-Calderón, GI; Gomes, B; Gondim, KC; Granzotto, A; Guarneri, AA; Guigó, R; Harry, M; Hughes, DST; Jablonka, W; Jacquin-Joly, E; Juárez, MP; Koerich, LB; Latorre-Estivalis, JM; Lavore, AE; Lawrence, GG; Lazoski, C; Lazzari, CR; Lopes, RR; Lorenzo, MG; Lugon, MD; Majerowicz, D; Marcet, PL; Mariotti, M; Masuda, H; Megy, K; Melo, ACA; Missirlis, F; Mota, T; Noriega, FG; Nouzova, M; Nunes, RD; Oliveira, RLL; Oliveira-Silveira, G; Ons, S; Pagola, L; Paiva-Silva, GO; Pascual, A; Pavan, MG; Pedrini, N; Peixoto, AA; Pereira, MH; Pike, A; Polycarpo, C; Prosdocimi, F; Ribeiro-Rodrigues, R; Robertson, HM; Salerno, AP; Salmon, D; Santesmasses, D; Schama, R; Seabra-Junior, ES; Silva-Cardoso, L; Silva-Neto, MAC; Souza-Gomes, M; Sterkel, M; Taracena, ML; Tojo, M; Tu, ZJ; Tubio, JMC; Ursic-Bedoya, R; Venancio, TM; Walter-Nuno, AB; Wilson, D; Warren, WC; Wilson, RK; Huebner, E; Dotson, EM; Oliveira, PL. 2015. The genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations yo hematophagy and parasite infection. Proceedings of the National Academy of Sciences of the United States of America. 112(48): 14936–41.

National Center of Biotechnology Information. 25/08/17 <ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec.

Ons, S; Sterkel, M; Diambra, L; Urlaub, H; Rivera-Pomar, R. 2011. Neuropeptide precursor gene discovery in the Chagas disease vector *Rhodnius prolixus.* Insect Molecular Biology. 20(1): 29-44.

Ons, S. 2015. Neuropeptides in the regulation of *Rhodnius prolixus* physiology. Journal of Insect Physiology. 97: 77-92

Parra, G; Bradnam, K; Korf, I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23(9): 1061-7

Perotti, E; Gamundi, JC; Russo, R. 2010. Control de Piezodorus guildinii (Westwood) en el cultivo soja. Para mejorar la producción 45- Estación Experimental Agropecuaria Oliveros. INTA.

Pevsner, J. (2009). *Bioinformatics and Functional Genomics*. New Jersey, USA: Wiley Blackwell

Poels, J; Van Loy, T; Vandersmissen, HP; Van Hiel, B; Van Soest, S; Nachman, RJ; Vanden-Broeck, J. 2010. Myoinhibiting peptides are the ancestral ligands of the promiscuous *Drosophila* sex peptide receptor. Cell Mol Life Sci 67: 3511–3522.

Rueden, TC.; Schindelin, J; Hiner, MC; DeZonia, BE; Walter, AE; Arena, ET; Eliceiri, KW. 2017. ImageJ2: ImageJ for the next generation of scientific image data. BMC Bioinformatics. 18: 529.

Salzberg, SL; Yorke, JA. 2005. Beware of mis-assembled genomes. Bioinformatics. 21(24): 4320–4321.

Schulz, MH; Zerbino, DR; Vingron, M; Birney, E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics (Oxford, England). 28(8): 1086–1092.

Simao, FA; Waterhouse, RM; Ioannidis, P; Kriventseva, EV; Zdobnov, EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19): 3210-2

Suzek, BE; Huang, H; McGarvey, P; Mazumder, R; Wu, CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 23(10): 1282-8.

Todd, JW. 1982. Effects of stinkbug damage on soybeans quality. Soybean seed quality and stand establishment. Sinclair, JB; Jackobs, JA. Proc. Internatl. Congr. INTSOY. Series number 22: 46-51.

Toledo, C; Anglada; Salusso, A. 2005. Productos fitosanitarios utilizados en las últimas campañas agrícolas para el control de plagas insectiles en soja. INTA EEA Paraná. Actualización Técnica SOJA. Serie Extensión n° 34. Septiembre. 4 pág.

Trapnell, C; Williams, BA; Pertea, G; Mortazavi, A; Kwan, G; van Baren, MJ; Salzberg, L; Wold, BJ; Pachter, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 28(5): 511–515.

Turcatti, G; Romieu, A; Fedurco, M; Tairi, AP. 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis †. Nucleic Acids Research. 36(4): e25–e25.

Velculescu, VE; Zhang, L; Zhow, W; Vogelstein, J; Basrai, MA; Bassett, DE Jr; Hieter, P; Vogelstein, B; Kinzler, KW. 1997. Characterization of the yeast transcriptome. Cell. 88(2): 243-51.

Verlinden, H; Vleugels, R; Zels, S; Dillen, S; Lenaerts, C; Crabbé, K; Spit, J; Vanden Broeck, J. 2014. Receptors for Neuronal or Endocrine Signalling Molecules as Potential Targets for the Control of Insect Pests. Advances in insect physiology 4: 167-303.

Wang, S; Gribskov, M. 2016. Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics, btw625.

Wang, Z; Gerstein, M; Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10(1): 57-63.

Yapici, N; Kim, YJ; Ribeiro, C; Dickson, BJ. 2008. A receptor that mediates the post-mating switch in Drosophila reproductive behaviour. Nature. 451: 33–37.

Yamanaka, N; Hua, YJ; Roller, L; Spalovská-Valachová, I; Mizoguchi, A; Kataoka,H; Tanaka,Y. 2010. *Bombyx* prothoracicostatic peptides activate the sex peptide receptor to regulate ecdysteroid biosynthesis. Proc Natl Acad Sci U S A. 107: 2060–2065.

Zdobnov, EM; Bork, P. 2007. Quantification of insect genome divergence. Trends in Genetics. 23(1): 16–20.

<http://www.agromeat.com/34440/chinches-fitofagas-en-el-cultivo-de-soja. 20/07/17

<http://emboss.sourceforge.net/apps/cvs/emboss/apps/transeq.html. 15/11/17

Anexo I

```
#!/usr/bin/python

from Bio import SeqIO

fasta_file    = "r958sDm0_xx_fq_pe_Rivera_A_1_ATCACGAT_L008_R3.fastq"  # Input fast[a|q]file
wanted_file   = "hits_blast_huevos"  # Input interesting sequence IDs, one per line
result_file   = "Huevos_R3_sucias.fastq"  # Output fast[a|q] file. Aca van las sucias
result_file_2 = "Huevos_R3_limpias.fastq"  # Output fast[a|q] file. Aca van las limpias


wanted = set()
with open(wanted_file) as f:
    for line in f:
        line = line.strip()
        if line != "":
            wanted.add(line)

fasta_sequences = SeqIO.parse(open(fasta_file),"fastq")
file = open(result_file_2, "w")

with open(result_file, "w") as f:
    for seq in fasta_sequences:
        if seq.id in wanted:
            SeqIO.write([seq], f, "fastq")
        else:
                SeqIO.write([seq], file, "fastq")

file.close()
```