29. Ciencia, Tecnología, e Innovación

Secuenciación y análisis del transcriptoma de *Dalbulusmaidis*

Palacio, Victorio Gabriel; Andrés Lavore; María Inés Catalano; Rolando Rivera Pomar
victoriopalacio@gmail.com; alavore@gmail.com; mariainescatalano@gmail.com;
rrivera@unnoba.edu.ar
Centro de Bioinvestigaciones
Universidad Nacional del Noroeste de la Provincia de Buenos Aires

Resumen

Los auquenorrincos (chicharritas o cotorritas) son insectos exclusivamente fitófagos, que pueden causar importantes daños económicos sobre los cultivos. Una de las enfermedades vectorizadas por ellos es el achaparramiento del maíz o *Corn Stunt Disease,* potencialmente una de las enfermedades más serias del cultivo de maíz, capaz de causar pérdidas parciales o totales en la producción en las zonas afectadas. En Argentina, *Dalbulusmaidis* (Hemiptera: Auchenorrhyncha) es el único vector a campo conocido como transmisor del *Spiroplasmakunkelii*, patógeno causal del *Corn Stunt*. Dada su importancia como plaga en la agricultura, se secuenció el transcriptoma de todos los estadios del ciclo de vida de este insecto (huevos, 5 estadios ninfales y dos muestras de adultos). Se utilizó un *pool* de insectos para abarcar la mayor cantidad de genes expresados. Como la información genómica de *Dalbulusmaidis* no está disponible, se realizó el ensamblado *de novo*. Se compararon los ensambles realizados con 3 programas: VELVET – OASES, ABySS y Trinity. Se evaluaron utilizando métricas (N50, longitud de *contig*) y medidas de cobertura (CEG, BUSCO). En base a estosanálisis, se decidió buscar genes del desarrollo en los ensambles de VELVET – OASES y Trinity. El porcentaje total de genes encontrado fue mayor para el ensamble de Trinity. Teniendo en cuenta los resultados previos, se ensamblaron el resto de las muestras con Trinity, obteniendo valores de métricas y coberturas muy buenos. Además se compararon los transcriptomas con proteomas publicados como medida de homología entre especies. En este trabajo se compararon distintos métodos de ensamble *de novo*y se seleccionó el que mejor se adaptó a nuestros datos y experimentos.

Palabras clave: Dalbulusmaidis, transcriptoma, bioinformática

Introducción

Los auquenorrincos (chicharritas o cotorritas) comprenden un grupo de insectos exclusivamente fitófagos, que representan un importante componente de la biodiversidad en casi todos los hábitats terrestres, tanto naturales como agroecosistemas. Varias especies de los mismos son

severas plagas de la agricultura por la transmisión de patógenos causales de enfermedades, las cuales pueden provocar importantes pérdidas económicas por el daño ocasionado durante la alimentación o la transmisión de patógenos (Maramorosch, K. Harris, 1979; Nault&Ammar, 1989; Virla, E.; Díaz, C.; Carpane, P.; Laguna, I.; Ramallo, J.; Gerónimo Gómez, 2004).

En Argentina, una de las enfermedades causadas por patógenos transmitidos por auquenorrincos es el achaparramiento del maíz o *Corn Stunt Disease*. Ésta es potencialmente una de las enfermedades más serias del cultivo de maíz y puede ser un factor limitante al causar pérdidas parciales o totales en la producción en las zonas afectadas (Bajet&Renfro, 1989; Bradfute, 1981). La prevalencia y el impacto económico de esta enfermedad han aumentado en los últimos años (Carpane*et al.*, 2013), desde que la enfermedad fue inicialmente descripta (Alstatt, 1945; Frazier, 1945).

Dalbulusmaidis (DeLong&Wolcott, 1923) (Hemiptera: Cicadellidae: Deltocephalinae) es uno de los cicadélidos más importantes hallados sobre el maíz (*Zea mays*) en América y el principal vector (Kunkel, 1946) a nivel mundial del *Spiroplasmakunkelii*, patógeno causal del Corn Stunt

El transcriptoma es el *set* completo de transcriptos de una célula, para un estado de desarrollo o condición fisiológica específica. Entender el transcriptoma, y cómo éste varía en diferentes condiciones, es esencial para interpretar los elementos funcionales del genoma, los cuales revelan los constituyentes moleculares de las células y tejidos. Además, sirve para comprender procesos tales como el desarrollo embrionario y distintos tipos de enfermedades (Z. Wang, Gerstein, &Snyder, 2009).

Antes de la aparición de nuevas tecnologías, los métodos más utilizados para analizar variantes génicas y de expresión eran los de *microarray* (Boguski*et al*., 1994; Gerhard *et al*., 2004). En contraste, los métodos basados en secuenciación, llamados RNA-Seq, determinan directamente la secuencia del ADN copia (ADNc). La secuenciación paralela masiva de millones de secuencias de ADNc nos brinda un método efectivo y de relativo 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 permiten identificar todos los transcriptos expresados, como una secuencia continua de ARN mensajero (ARNm) desde el principio de la transcripción hasta el final, aun para los genes que tienen múltiples isoformas debido al *splicing* alternativo (Guttman*et al*., 2010). A raíz de su creciente fácil acceso, los estudios que utilizan un enfoque genómico o transcriptómico trascendieron su nicho inicial en ciencias biomédicas o para un puñado de organismos modelo, y se tornó fundamental en los avances de otros campos y disciplinas (Ekblom& Wolf, 2014).

Sin embargo, el RNA-seq tiene diversas limitaciones; como la dependencia del conocimiento previo de secuencias genómicas, presencia de altos niveles de ruido debido a la hibridación cruzada (Okoniewski& Miller, 2006; Royce, Rozowsky, &Gerstein, 2007) y también un bajo nivel de detección debido al *background* y a la saturación de señales. Además, comparar niveles de expresión en distintos experimentos es difícil y requiere métodos de normalización complejos.
En los últimos 15 años, se desarrollaron nuevas tecnologías de secuenciación (Metzker, 2005), por ejemplo,pirosecuenciación (454 Sequencing) (Margulies*et al.*, 2005) y secuenciación por síntesis (Solexa) (Bentley, 2006), que revolucionaron y expandieron el campo de la genómica. Comparado con los métodos tradicionales de Sanger, estas tecnologías generan *reads* cortos de alta calidad (Goodwin, McPherson, &McCombie, 2016), los cuales van desde 25 pb a 500 pb en longitud, lo cual es sustancialmente más corto que las secuencias obtenidas por la tecnología de capilaridad. Éstas, aplicadas a la expresión de ARNm (RNA-seq) transformaron el campo de la transcriptómica (Blencowe, Ahmad, & Lee, 2009; Z. Wang *et al.,* 2009).
Entre estas plataformas, Illumina se convirtió en la más extendida a nivel mundial (Goodwin*et al.*, 2016), principalmente debido a una combinación de madurez como tecnología, junto con un bajo costo y alta precisión por base secuenciada.

Sin embargo, la reconstrucción de un transcripto de longitud completa desde *reads* cortos con considerables errores de secuenciación tiene diversos desafíos computacionales (Haas&Zody, 2010). Entre ellos, algunas de las situaciones a enfrentar son: algunos transcriptos tienen poca cobertura, mientras que otros están altamente expresados; la cobertura de un transcripto puede que no sea igual a lo largo de toda su longitud, debido a errores de secuenciación; *reads* con errores de secuenciación derivados de un transcripto altamente expresado pueden ser más abundantes que *reads* correctos de un transcripto con baja expresión; transcriptos codificados por *loci* adyacentes pueden solaparse y formar erróneamente un transcripto quimérico; las estructuras de datos tienen que acomodar múltiples transcriptos por *locus*, debido al *splicing* alternativo; secuencias repetidas de genes diferentes introducen ambigüedad.
Un método eficiente debería poder afrontar y superar los distintos desafíos, siendo útil tanto para genomas complejos de animales o pequeñosgenomas bacterianos. También, tener la capacidad de reconstruir transcriptos de tamaños, niveles de expresión e isoformas variables.
Existen dos alternativas computacionales para la reconstrucción de un transcriptoma: mapeando a una referencia o ensamblando *de novo*. La primera consiste en alinear los *reads* a un genoma de referencia y unir las secuencias que se solapen, mientras que la segunda manera no requiere una referencia. Esta última es la que se utiliza cuando no hay un genoma disponible, el mismo se encuentra altamente fragmentado o con partes faltantes (Conesa*et al.*, 2016).

Dado que en la mayoría de los organismos no existe un genoma publicado para utilizar como referencia, el ensamble *de novo* es la única alternativa viable para el estudio del transcriptoma. Incluso en casos en los que se dispone de genomas de especies emparentadas, no se puede confiar completamente en esas secuencias para ensamblar los transcriptos. Por ejemplo, a pesar de la diversidad que presenta la clase Insecta y que tiene varios organismos modelo clave, no hay publicada gran cantidad de secuencias genómicas completas de alta calidad. Sumado a eso, los transcriptomas de insectos muestran patrones de *splicing* complejos (Graveley, 2001). Para sumar variabilidad a la cuestión, el ARN utilizado en la mayoría de los estudios transcriptómicos, proviene de un *pool* de insectos de una población.

Una solución computacional para la estrategia de reconstrucción sin referencia, es la utilización de los grafos de De Brujin (De Brujin, 1946; Good, 1946). 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, solamente 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, llamadas *contigs*. 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 del *set* total de datos previo al ensamble, ya que *reads* con errores pueden resultar en falsos nodos, llevando el grafo por caminos erróneos.

Trabajos previos mostraron que no hay un ensamblador definitivamente mejor que los demás, sino que depende también del *set* de datos y de qué se busca en cada experimento en particular (Schulz, Zerbino, Vingron, &Birney, 2012; S. Wang &Gribskov, 2016).

Los auquenorrincos (chicharritas o cotorritas) comprenden un grupo de insectos exclusivamente fitófagos, que representan un importante componente de la biodiversidad en casi todos los hábitats terrestres, tanto naturales como agroecosistemas. Varias especies de los mismos son severas plagas de la agricultura por la transmisión de patógenos causales de enfermedades, las cuales pueden provocar importantes pérdidas económicas por el daño ocasionado durante la alimentación o la transmisión de patógenos (Maramorosch, K. Harris, 1979; Nault&Ammar, 1989; Virla, E.; Díaz, C.; Carpane, P.; Laguna, I.; Ramallo, J.; Gerónimo Gómez, 2004)□.

En Argentina, una de las enfermedades causadas por patógenos transmitidos por auquenorrincos es el achaparramiento del maíz o *Corn Stunt Disease*. Ésta es potencialmente una de las enfermedades más serias del cultivo de maíz y puede ser un factor limitante al causar pérdidas parciales o totales en la producción en las zonas afectadas (Bajet&Renfro, 1989; Bradfute, 1981). La prevalencia y el impacto económico de esta enfermedad han aumentado en los últimos años (Carpane*et al.*, 2013), desde que la enfermedad fue inicialmente descripta (Alstatt, 1945; Frazier, 1945).

Dalbulusmaidis (DeLong&Wolcott, 1923) (Hemiptera: Cicadellidae: Deltocephalinae) es uno de los cicadélidos más importantes hallados sobre el maíz (*Zea mays*) en América.
*D. maidis*es el principal vector (Kunkel, 1946) a nivel mundial del *Spiroplasmakunkelii*,

Objetivos

Dado que es un insecto sumamente relevante para la agricultura y, ante la falta de datos del tipotranscriptómicos del mismo, planteamos como objetivo del trabajo establecer un protocolo para ensamblar los transcriptomas de los distintos estadios del ciclo de vida de *Dalbulus maidis*, y, a partir de los ensambles, evaluar la calidad de los mismos, identificar un *set* de genes del desarrollo y compararlos con proteomas de otras especies.

Materiales y Métodos

Limpieza y calidad de datos

Los datos crudos fueron pre-procesados con el programa FastQCtoolkit (www.bioinformatics.babraham.ac.uk/projects/fastqc/) para evaluar la calidad de los *reads*generados. Porotrolado, utilizando este mismo paquete de programas, se eliminaron secuencias adaptadoras y secuencias de bajacalidad.
Para evidenciar contaminación en los datos, se utilizó la herramienta BLAST (Altschul, Gish, Miller, Myers, &Lipman, 1990) para, de esta forma, comparar el total de los *reads*secuenciados contra la base de datosVecScreen del NCBI (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec) y buscar secuencias contaminantes como vectores y primers. Los parámetros utilizados fueron: *evalue*= 0,00001 y max _target_seq=1. Los *reads* que dieron *hit* fueron eliminados de nuestro *set* de datos, tanto en las secuencias generadas durante la primera lectura así como las generadas en la segunda, utilizando un *script* en el lenguaje python. De esta forma se logramantener la simetría en la secuenciación *pair end*. Además, se realizó una búsqueda similar contra una base de datos de adaptadores de Illumina.

Ensamble transcriptómico

Dado que el genoma de *D. maidis* no estásecuenciadoaún, se realizó un ensambe*de novo*. El mismo se llevó a cabo a través de accesoremoto a un *cluster* computacional del Instituto de SaludPública de México, en Cuernavaca y otro*cluster*computacional en Göttingen, Alemania. Se utilizaron los *reads* de la muestra de huevos de *D. maidis* para realizarensambles con distintosprogramas y determinar cuál es el másadecuado para ensamblar el resto de lasmuestras. Los softwares utilizadosfueron VELVET – OASES, ABySS y Trinity.

En el caso del ensamblerealizado con VELVET y OASES, primero se utilizó la herramienta velveth para partir e indexar los datos en los distintos valores de kmerosespecificados, que en este caso fueron de 47, 49 y 51 bases de longitud. Además, se especificó que eran *reads* cortos y apareados, dados en dos archivos separados de formato fastq.

Con los archivosgenerados en el paso anterior, se utilizó la herramientavelvetg, la cual crea los gráficos de De Brujin y, a partir de corridas de simplificación y corrección de errores, forma los *contigs*. Se utilizó en este paso una opción llamada *read tracking*, la cual tiene un alto costo de memoria y tiempo de cálculo, pero da como resultado una descripción más detallada del ensamble y nos permite continuar con el software OASES.

A cada conjunto de archivos generados pertenecientes a los distintos kmeros iniciales se le aplica el programa OASES, el cualnos genera un archivo llamado transcripts.fa para cada uno de estos valores.

Ya con los transcripts.fa para 47, 49 y 51 pb, se procedió a repetir los pasos anteriormente mencionados, pero ahora usando los *contigs* como *input* en vez de los *reads* originales. La diferencia es que, en este caso, se unieron estos tres ensambles en uno, utilizando kmeros de 27, como el programa lo indica.

Se realizó el paso de velveth, luegovelvetg y por último el de OASES, obteniendo así nuestro archivo transcripts.fa final.

El ensamble con ABySS se realizó con las opciones estándares del programa, especificando que nuestros *reads* son *pair end*. Los kmeros utilizados fueron 31, 47 y 63. Para cada uno de ellos se generaron 3 ensambles: *unitigs*, *contigs* y *scaffolds*. Los primeros se establecen con información de un solo *end* de corrida, los *contigs* con los dos *end* y los *scaffolds*utilizando los *contigs*comomateria prima.

Por último, se utilizó el ensamblador Trinity. Éste se corrió en el modo *pairend*, con las opciones de memoria máxima limitada a 30 Gigas y limitando la salida final a *contigs* de por lo menos 200 o 100 pares de bases de longitud.

CEG y BUSCO

Para evaluar el porcentaje de representatividad de los genes codificantes para proteínas se realizóunabúsquedautilizando la herramientaHmmer. Los transcriptomas ensamblados se tradujeron a los seis marcos de lectura usando la herramienta transeq (http://emboss.sourceforge.net/apps/cvs/emboss/apps/transeq.html) para luego hacer una búsqueda por HMMER3 (Eddy, 1995), con el comando hmmscan y los filtros -T 40 y –domT40, de los transcriptos traducidos con el perfil de CEG (core eukariotyc genome) (Parra, Bradnam, &Korf,

2007), como se describió en 2012por Martínez-Barnetche y colaboradores. El CEG estácomprendidopor un grupo de 458 proteínas.
Como otra medida de cobertura, se realizó el mismo procedimiento con el perfil de BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão, Waterhouse, Ioannidis, Kriventseva, &Zdobnov, 2015) de artrópodos, el cual está comprendido por 2.676 proteínas.

Búsqueda de ortólogos

Se contrastaron los transcriptomas ensamblados de huevos realizados con VELVET – OASES y Trinity contra una base de datos de los principales genes del desarrollo quea ctúan durante la embriogénesis de insectos. Para determinar qué genes se expresaban en el transcriptoma ensamblado, se utilizó la herramienta NCBI - BLASTx (Altschul*et al.*, 1990), filtrando la búsqueda con un *E-value* de $1x10^{-5}$ y mostrando el mejor *hit* para cada uno de nuestros transcriptos. En el caso de que el *E-value* fuera igual en distintos transcriptos para una misma proteína, se tuvo en cuenta la longitud del *hit* y el porcentaje de cobertura de la secuencia proteica, de modo de identificar el mejor *hit* posible para cada transcripto.

Mejor recíproco

Se realizaron BLAST bidireccionales entre los transcriptomas ensamblados y los proteomas de *Halyomorphahalys*(https://i5k.nal.usda.gov/Halyomorpha_halys), *Oncopeltusfasciatus*(https://data.nal.usda.gov/dataset/oncopeltus-fasciatus-genome-assembly-10) y *Rhodniusprolixus*(httpd://www.vectorbase.org/organisms/rhodnius-prolixus/cdc/rproc3). Luego, se utilizó un *script* (https://github.com/toritori5/RecipABCD) para determinar los mejores *hits* recíprocos entre dos, tres y cuatro especies.

Resultados y Discusión

Comparación de ensambladores

Existen distintos programas para ensamblar transcriptomas. Trabajos previos muestran que la mejor estrategia está relacionada directamente con el *set* de datos particular y con el fin de la investigación (Schulz *et al.*, 2012; S. Wang & Gribskov, 2016). Es decir, no existe un ensamblador mejor que los otros, sino que cada uno tiene virtudes y defectos. Es por eso que se probaron tres programas distintos para ver cuál se adecuaba mejor a nuestras necesidades. Para evaluar los ensambles fueron utilizadas tanto métricas como N50, longitud de *contig* promedio, cantidad de transcriptos ensamblados,así como medidas de cobertura. Para identificar la proporción del genoma eucariótico núcleo que cubre el transcriptoma se utilizaron perfiles de HMM correspondientes a las 458 proteínas núcleo de los eucariotas que provee el algorítmo CEGMA

(Au, Underwood, Lee, & Wong, 2012). Otra medida fue una búsqueda similar contra 2676 perfiles de proteínas comunes de artrópodos, conocido como BUSCO (Robasky, Lewis, & Church, 2013); de esta manera,búsquedas locales con HMMER3 permitieron determinar la cobertura de nuestros transcriptomas.

Se iniciaron las pruebas de ensambles con el *set* de datos de huevos porque en ellos se buscarían los ortólogos para distintos genes del desarrollo. El primer ensamble se realizó con los programas VELVET y OASES. Utilizando valores de kmeros altos (47, 49 y 51), obtuvimos 116951 transcriptos pertenecientes a 39552 *loci*. El CEG y BUSCO dieron valores altos, 97.81% y 91.67% respectivamente. Además, las longitudes de la mayor cantidad de transcriptos estuvieron alrededor de los 400 pb.

Luego se utilizó ABySS, se seleccionaron valores de kmeros en distintos rangos para poder comparar los resultados. Se utilizó 31 como valor inferior, 47 intermedio y 63 mayor. Claramente, a medida que el valor del kmero aumentaba, la cantidad de transcriptos totales decrecía. Otro detalle a tener en cuenta, es que este ensamblador no reporta las diferentes isoformas, sino el total de transcriptos ensamblados. Se analizó la distribución de longitudes para los distintos ensambles, con y sin los transcriptos menores a 100 pb, ya que el programa ensambla transcriptos a partir del tamaño de kmeroutilizado. Las longitudes son mayores a medida que aumenta el kmero, pero no llega a los valores de los otros ensambladores. Se buscó el CEG para cada uno de ellos, obteniéndose valores altos pero no comparables con los otros ensambladores.

Por último, se realizaron los ensambles con Trinity,donde se compararon dos opciones, ensamblando transcriptos mayores a 100 pb o mayores a 200 pb. En ambos casos, el kmero utilizado fue 25, valor asignado por *default* por el programa. La mayor diferencia entre los dos transcriptomas, se observó en la cantidad de transcriptos ensamblados, siendo mucho mayor la cantidad cuando el mínimo de longitud es 100 pb. Proporcional a esa diferencia, se evidenció un aumento en lo que el programa nombra como ‘genes’. Aunque la proporción de isoformas se mantiene a lo largo de los dos ensambles (Tabla 1).

Las longitudes medias de los *contigs* tuvieron valores más altos que en ABySS y ligeramentemenoresqueen VELVET – OASES. En contraste, los valores de CEG y BUSCO fueron los más elevados, siendo de 99.56% y 98.23% respectivamente, para el ensamble de secuencias mayores a 200 pb.

A partir de los resultados obtenidos en las distintas métricas de los ensambles, se llevó a cabo la búsqueda de genes del desarrollo en el ensamble de VELVET – OASES y en el de Trinity de mayores a 200 pb. Los resultados mostraron una mayor cantidad de *hits* en el ensamble de Trinity. Entonces, teniendo en cuenta los distintos análisis, se determinó realizar el ensamble de los otros estadios de *D. maidis* con el programa Trinity.

Al analizar los ensambles de los distintos estadios, se observó en las métricas que los valores obtenidos continuaban siendo buenos, al igual que en el ensamble de Huevos. Por otro lado, la cobertura, estimada por medio del CEG y BUSCO tuvo valores muy altos. Esto estaría indicando que los ensambles tienen una buena cobertura de los transcriptomas, y que son aptos para continuar con otros estudios.

Tabla 1. Métricas para los ensambles de todos los estadios realizados con el software Trinity

	Huevos	DMI	DMII	DMIII	DMIV	DMV	Adultos2	Adultos3
Total 'genes'	50218	39126	45685	55294	65720	59366	72430	64728
Total transcripts	55588	43993	51829	62316	77757	67958	88952	76444
Menor longitud	201	201	201	201	201	201	201	201
Mayor longitud	8162	8646	12611	12636	12647	12649	25970	13294
CEG	99.56	96.73	98.69	99.13	99.78	99.78	99.78	99.78
BUSCO	98.32	87.37	91.78	97.83	98.57	98.58	99.48	98.88
	Estadísticas basadas en todos los transcriptos							
N10	2822	1895	2134	2839	3060	3116	4819	3008
N20	2142	1500	1656	2119	2280	2301	3511	2238
N30	1701	1229	1362	1678	1795	1799	2686	1779
N40	1358	1021	1110	1337	1414	1423	2086	1413
N50	1058	835	885	1045	1092	1099	1618	1097
Mediana	427	416	410	413	406	409	438	413
Promedio	699,7	605,3	625,38	687,78	700,35	704,8	880,63	704,27
Bases ensambladas	38894652	26628753	32412787	42859782	54457019	47896923	78333356	53837107
	Estadísticas basadas en la isoforma más larga por 'gen'							
N10	2728	1851	2082	2702	2917	2957	4407	2835
N20	2041	1454	1603	1999	2142	2151	3083	2102
N30	1623	1191	1306	1565	1659	1670	2327	1648
N40	1247	983	1048	1225	1258	1286	1761	1263
N50	975	793	824	935	936	965	1305	950
Mediana	405	395	387	387	373	381	385	380
Promedio	663,15	581,84	594,63	640,21	635,90	648,39	751,32	641,59
Bases ensambladas	333	22764923	27165494	35399838	41791190	38492181	54418037	41528533

Búsqueda de ortólogos

Los métodos de secuenciación de nueva generación (NGS) permiten encontrar genes del desarrollo en órdenes de magnitud mayores a décadas anteriores, a una fracción de su costo. Cuando se realizan las búsquedas, hay que tener en cuenta que el nivel de conservación entre los genomas de insectos es muy bajo (Parra *et al.*, 2007) por lo que no encontrar un gen no implica su ausencia.

El porcentaje de genes del desarrollo encontrados en el transcriptoma ensamblado con Trinity fue de 89,3%.Trabajos en organismos como *Oncopeltus fasciatus* (Simão *et al.*, 2015) y el crustáceo *Parhyale hawaiensis* (Zdobnov & Bork, 2007) también hicieron búsquedas de ortólogos entre

genes del desarrollo, mayormente con bases de datos de genes de *Drosophila melanogaster* para comparar. En ambos casos, los porcentajes de genes encontrados fueron menores al nuestro.
En el ensamble de Trinity se encontraron más genes que en el de VELVET – OASES, 89,32% y 82,05%, respectivamente (Figura 1). Por lo cual el resto de las muestras se ensamblaron con Trinity.

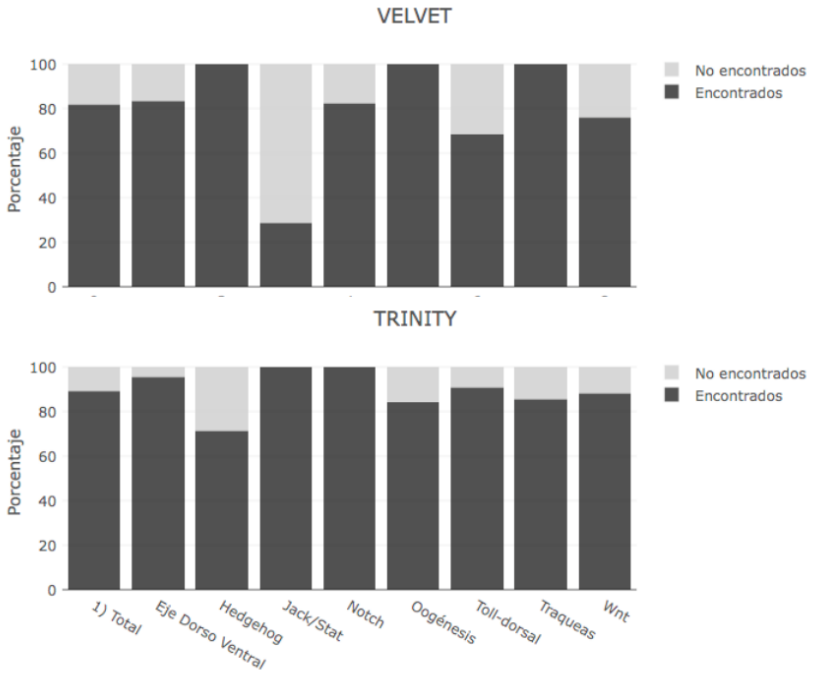

Figura 1. Genes del desarrollo encontrados en cada uno de los ensambles.

Recíprocos

Para estimar la proporción del transcriptoma que es homóloga con los proteomas de *Halyomorphahalys*, *Oncopeltusfasciatus*y*Rhodniusprolixus* se utilizó la técnica del mejor recíproco. Se realizó el análisis con todos los transcriptomas ensamblados, pero la muestra que presentómayor homología fue la de Adultos2 (Tabla2).

Tabla 2. Mejores recíprocos entre *D. maidis (Adultos2)*, *H. halys*, *O. fasciatus* y *R. prolixus*.

Dalbulus maidis	Halyomorpha halys			6619
Dalbulus maidis		Oncopeltus fasciatus		7925
Dalbulus maidis			Rhodnius prolixus	7311
Dalbulus maidis	Halyomorpha halys	Oncopeltus fasciatus		3851
Dalbulus maidis	Halyomorpha halys		Rhodnius prolixus	3982
Dalbulus maidis		Oncopeltus fasciatus	Rhodnius prolixus	3896
	Halyomorpha halys	Oncopeltus fasciatus	Rhodnius prolixus	5818
Dalbulus maidis	Halyomorpha halys	Oncopeltus fasciatus	Rhodnius prolixus	3526

Un total de 3526 genes ortólogos son compartidos por las 4 especies, mientras que *D. maidis* muestra una mayor homología con*O.fasciatus*, con el cual comparte 7925 genes.

Conclusiones

En este trabajo se evaluaron distintos programas para el ensamble *de novo*de millones de *reads* de *Dalbulusmaidis* pertenecientes a distintos estadios del insecto. A razón de distintas pruebas con el *set* de datos de los huevos con cada uno de los programas, se eligió Trinity. ,Utilizando este programa se ensamblaron los transcriptomas de todos los estadios. Todas las métricas y análisis de cobertura dieron valores alentadores y permiten continuar con los estudios. Además, se buscaron una serie de genes del desarrollo utilizando una base de datos de *D.melanogaster*, hallandoen nuestros datos la mayoría de los mismos. Como dato final, se compararon los transcriptomas con los proteomas publicados de especies emparentadas y se obtuvieron datos de homología.

Bibliografía

Alstatt, G. E. (1945). A new corn disease in the Rio Grande Valley. Plant Disease, *29*, 533–534.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., &Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Au, K. F., Underwood, J. G., Lee, L., & Wong, W. H. (2012). Improving PacBio Long Read Accuracy by Short Read Alignment. *PLoS ONE*, *7*(10), e46679. https://doi.org/10.1371/journal.pone.0046679

Bajet, N. B., & Renfro, B. L. (1989). Occurence of corn stunt spiroplasma at different elevations in Mexico.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, *16*(6), 545–552. https://doi.org/10.1016/j.gde.2006.10.009

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., … Jones, S. J. M. (2009). De novo transcriptome assembly with ABySS. *Bioinformatics*, *25*(21), 2872–2877. https://doi.org/10.1093/bioinformatics/btp367

Blencowe, B. J., Ahmad, S., & Lee, L. J. (2009). Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & Development*, *23*(12), 1379–86. https://doi.org/10.1101/gad.1788009

Boguski, M. S., Tolstoshev, C. M., & Bassett, D. E. (1994). Gene discovery in dbEST. *Science (New York, N.Y.)*, *265*(5181), 1993–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8091218

Bradfute, O. E. (1981). Corn Stunt Spiroplasma and Viruses Associated with a Maize Disease Epidemic in Southern Florida. *Plant Disease*, *65*(10), 837. https://doi.org/10.1094/PD-65-837

Carloni, E., Virla, E., Paradell, S., Carpane, P., Nome, C., Laguna, I., &GiménezPecci, M. P. (2011). Exitianusobscurinervis (Hemiptera: Cicadellidae), a New Experimental Vector of Spiroplasmakunkelii. *Journal of Economic Entomology*, *104*(6), 1793–1799. https://doi.org/10.1603/EC11156

Carpane, P., Melcher, U., Wayadande, A., de la Paz GimenezPecci, M., Laguna, G., Dolezal, W., & Fletcher, J. (2013). An Analysis of the Genomic Variability of the PhytopathogenicMollicute*Spiroplasmakunkelii*. *Phytopathology*, *103*(2), 129–134. https://doi.org/10.1094/PHYTO-07-12-0158-R

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. https://doi.org/10.1186/s13059-016-0881-8

Davis, R. (1966). Biology of the Leafhopper Dalbulusmaidis at Selected Temperatures12. *Journal of Economic Entomology*, *59*(3), 766–766. https://doi.org/10.1093/jee/59.3.766

Davis R. E. (1974). Spiroplasma in corn stunt-infected individuals of the vector leafhopper Dalbulusmaidis. *Institut Technique DesCereales et Des Fourrages*. Retrieved from http://agris.fao.org/agris-search/search.do;jsessionid=43BD6BFEFE0858EB196A7EB0F6D3FE99?request_locale=zh_CN&recordID=US19750033580&query=&sourceQuery=&sortField=&sortOrder=&agrovocString=&advQuery=&centerString=&enableField=

De Brujin, N. (1946). A Combinatorial Problem. *KoninklijkeNederlandseAkademie v Wetenschappen*, *46*, 758–764.

Eddy, S. R. (1995). Multiple alignment using hidden Markov models. Retrieved from http://eddylab.org/publications/Eddy95b/Eddy95b-reprint.pdf

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–42. https://doi.org/10.1111/eva.12178

Frazier, N. W. (1945). A streak disease of corn in California.

Gerhard, D. S., Wagner, L., Feingold, E. A., Shenmen, C. M., Grouse, L. H., Schuler, G., … MGC Project Team. (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Research*, *14*(10B), 2121–7. https://doi.org/10.1101/gr.2596504

Good, I. J. (1946). Normal Recurring Decimals. *Journal of the London Mathematical Society*, *s1-21*(3), 167–169. https://doi.org/10.1112/jlms/s1-21.3.167

Goodwin, S., McPherson, J. D., &McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics : TIG*, *17*(2), 100–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11173120

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., … 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), 503–10. https://doi.org/10.1038/nbt.1633

Haas, B. J., &Zody, M. C. (2010). Advancing RNA-Seq analysis. *Nature Biotechnology*, *28*(5), 421–423. https://doi.org/10.1038/nbt0510-421

Heady, S. E., Nault, L. R., Shambaugh, G. F., & Fairchild, L. (1986). Acoustic and Mating Behavior of Dalbulus Leaf hoppers (Homoptera: Cicadellidae). *Annals of the Entomological Society of America*, *79*(4), 727–736. https://doi.org/10.1093/aesa/79.4.727

Kwon, M.-O., Wayadande, A. C., & Fletcher, J. (1999). Spiroplasmacitri Movement into the Intestines and Salivary Glands of Its Leafhopper Vector, Circulifertenellus. *Phytopathology*, *89*(12), 1144–1151. https://doi.org/10.1094/PHYTO.1999.89.12.1144

Maramorosch, K. Harris, K. (1979). *Leafhopper Vectors and Plant Disease Agents. London: Academic*.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376. https://doi.org/10.1038/nature03959

Markham, &Alivizatos. (1983). Proc. Int. Maize Virus. https://doi.org/10.1093/database/bar009

Martínez-Barnetche, J., Gómez-Barreto, R. E., Ovilla-Muñoz, M., Téllez-Sosa, J., López, D. E., Dinglasan, R. R., … López, M. H. (2012). Transcriptome of the adult female malaria mosquito vector Anopheles albimanus. *BMC Genomics*, *13*(1), 207. https://doi.org/10.1186/1471-2164-13-207

Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome Research*, *15*(12), 1767–1776. https://doi.org/10.1101/gr.3770505

Nault, L. R. (1980). Maize Bushy Stunt and Corn Stunt: A Comparison of Disease Symptoms, Pathogens Host Ranges, and Vectors. *Phytopatology*, *70*, 659–662. Retrieved from https://www.apsnet.org/publications/phytopathology/backissues/Documents/1980Articles/Phyto70n07_659.PDF

Nault, L. R., &Ammar, E. D. (1989). Leafhopper and Planthopper Transmission of Plant Viruses. *Annual Review of Entomology*, *34*(1), 503–529. https://doi.org/10.1146/annurev.en.34.010189.002443

Nault, L. R., & Madden, L. V. (1985). Ecological strategies of Dalbulus leafhoppers. *Ecological Entomology*, *10*(1), 57–63. https://doi.org/10.1111/j.1365-2311.1985.tb00534.x

Okoniewski, M. J., & Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, *7*(1), 276. https://doi.org/10.1186/1471-2105-7-276

Ozbek, E., Miller, S. A., Meulia, T., &Hogenhout, S. A. (2003). Infection and replication sites of Spiroplasmakunkelii (Class: Mollicutes) in midgut and Malpighian tubules of the leafhopper Dalbulusmaidis. *Journal of Invertebrate Pathology*, *82*(3), 167–75. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12676553

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–1067. https://doi.org/10.1093/bioinformatics/btm071

Ramirez-C, J. L., Leon-G, C. de., Garcia-M, C., & Granados-R, G. (1975). Dalbulus guevarai (Del.) nuevo vector del achaparramiento del maiz en Mexico: incidencia de la enfermedad y su relacion con el vector Dalbulus maidis (Del. & W.) en Muna, Yucatan [1975]. *Agrociencia*. Retrieved from http://agris.fao.org/agris-search/search.do?recordID=US201302491570

Robasky, K., Lewis, N. E., & Church, G. M. (2013). The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, *15*(1), 56–62. https://doi.org/10.1038/nrg3655

Royce, T. E., Rozowsky, J. S., & Gerstein, M. B. (2007). Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Research*, *35*(15), e99–e99. https://doi.org/10.1093/nar/gkm549

Schulz, M. H.,Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, *28*(8), 1086–92. https://doi.org/10.1093/bioinformatics/bts094

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V, &Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, *31*(19), 3210–2. https://doi.org/10.1093/bioinformatics/btv351

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … 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. https://doi.org/10.1038/nbt.1621

Virla, E.; Díaz, C.; Carpane, P.; Laguna, I.; Ramallo, J.; Gerónimo Gómez, L. . G. P. (2004). Evaluación preliminar de la disminución en la producción de maíz causada por el “Corn Stunt Spiroplasma” (CSS) en Tucumán, Argentina.

Wang, S., &Gribskov, M. (2016). Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics*, btw625. https://doi.org/10.1093/BIOINFORMATICS/BTW625

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wayadande, A. C., & Fletcher, J. (1995). Transmission of Spiroplasmacitri lines and their ability to cross gut and salivary gland barriers within the leafhopper vector Circulifertenellus. *Phytopathology (USA)*. Retrieved from http://agris.fao.org/agris-search/search.do?recordID=US9625059

Zdobnov, E. M., & Bork, P. (2007). Quantification of insect genome divergence. *Trends in Genetics*, *23*(1), 16–20. https://doi.org/10.1016/j.tig.2006.10.004