



CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL
INSTITUTO POLITÉCNICO NACIONAL

UNIDAD ZACATENCO
DEPARTAMENTO DE BIOQUÍMICA

Caracterización, cuantificación y ensamblado *de novo* de circRNAs a partir de datos de
RNA-seq de cepas virulentas y no virulentas de *Entamoeba histolytica*.

TESIS

Que presenta:

IBT. Cristian Julio César Padrón Manrique

Para obtener el grado de
Maestro en Ciencias

En la especialidad de
Bioquímica

Directores de tesis:

Dr. Jesús Valdés Flores.

Dr. Alfonso Méndez Tenorio.

ESTE TRABAJO SE REALIZÓ EN EL DEPARTAMENTO DE BIOQUÍMICA DEL CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL, BAJO LA DIRECCIÓN DEL Dr. JESÚS VALDÉS FLORES y DEL Dr. ALFONSO MÉNDEZ TENORIO.

DURANTE LA REALIZACIÓN DE ESTE TRABAJO SE CONTRÓ CON LA BECA 484770 OTORGADA POR CONACyT y BECARIO 632497, POR LO CUAL SE AGRADECE A DICHA INSTITUCIÓN.

Abstract

Within classification of non-coding RNAs there is a new participant, circular RNAs (circRNA), circRNAs are single-stranded RNA molecules that differentiate from linear RNAs, because form a continuous loop covalently closed. These circles are present in eukaryotic organisms. CircRNAs are produced from coding regions, intergenic regions, introns, coding regions with intronic regions and UTRs by mechanisms still under investigation. Circular RNAs were first discovered in the 1970s, but with the advent of mass sequencing (NGS) and bioinformatic methods, their study advanced enormously. *Entamoeba histolytica* is a human pathogenic anaerobic parasite that infects human and other species, causing amebiasis including amoebic colitis and liver abscess. *E. histolytica* infection may develop asymptotically or may cause dysentery. As mentioned above, with modern methods of study, it has been possible to determine different genes involved in virulence. In this project it was possible to characterize, quantify and assemble *in silico* circRNAs in *E. histolytica*, HM-1: IMSS (virulent strain) and Rahman (non-virulent strain) using NGS data. Using different computational algorithms, a total of 599 circRNAs in *E. histolytica* were characterized and assembled from the database reported by Hon *et al.* 2013. We found that there was a positive monotonic correlation of the expression of the circular transcripts with their linear counterparts, the correlation increased when the expression of the transcripts of the differentially expressed circRNAs was compared with their linear counterparts. We performed analysis of protein interaction networks, and we observed the participation of different biochemical pathways. Thus, the most conspicuous cluster corresponded to ribosomal proteins and the ribosome biogenesis pathway. With computationally methods, we were able to identify an overexpressed exonic circRNA in non-virulent amoebas whose parental gene occupies the locus EHI_169670. This results was validated by RT-PCR.

Resumen

Dentro de la clasificación de los RNAs no codificantes hay un nuevo participante, los RNAs circulares (circRNAs), moléculas de RNA monocatenarios que a diferencia de los RNAs lineales, forman un bucle continuo cerrado covalentemente. Estos circRNAs están presentes en eucariontes. Los circRNAs son producidos a partir de regiones codificantes, regiones intergénicas, intrones, regiones codificantes con regiones intrónicas y UTRs por mecanismos aún no lucidados por completo estando aún en investigación. Los RNA circulares se descubrieron por primera vez en la década de los 70, pero con el advenimiento de la secuenciación masiva (NGS) y métodos bioinformáticos es que su estudio floreció. *E. histolytica* es un protozoo parásito anaerobio patógeno para el humano entre otras especies, causando amebiasis incluyendo colitis amébrica y absceso hepático. La infección por *E. histolytica* puede desarrollarse de forma asintomática o puede producir disentería. Como mencionamos arriba, con métodos modernos de estudio es que se han podido determinar diferentes genes involucrados en la virulencia. En este proyecto se logró caracterizar, cuantificar y ensamblar *in silico* circRNAs en *E. histolytica*, HM-1:IMSS (cepa virulenta) y Rahman (cepa no virulenta) utilizando datos de NGS. Utilizando diferentes algoritmos computacionales se caracterizaron y ensamblaron un total de 599 circRNAs en *E. histolytica* a partir de la base de datos reportada por Hon y cols. 2013. Encontramos que hubo una correlación monotónica positiva de la expresión de los transcritos circulares con sus contrapartes lineales, la correlación aumentó cuando se comparó la expresión de los transcritos de los circRNAs expresados diferencialmente con sus contrapartes lineales. Se realizó un análisis de redes de interacción entre proteínas en las que se observó la participación de diferentes rutas bioquímicas, aunque el *cluster* más conspicuo correspondió a las proteínas ribosomales y a la ruta de la biogénesis del ribosoma. Computacionalmente pudimos identificar un circRNA exónico sobreexpresado en amebas no virulentas cuyo gen parental ocupa el locus EHI_169670. El circRNA fue validado por RT-PCR.

Agradecimientos

Agradezco primeramente a mi familia en general, pero sobretodo a mi padre y madre. Sobretodo a mi madre Magda, por darme todas las herramientas y darme en mis manos la oportunidad de ser músico y además ser lo que soy ahora, muy pocos tienen esa oportunidad más aún del lugar de origen que soy gracias mamá. Sin ellos esto no hubiera sido posible, por darme todas las armas para luchar en este mundo.

A mis grandes amigos de la maestría a Vicente, Nicole y Jesús sin ustedes hubiera sido muy solitaria esta experiencia y muy poco divertida los amo. Ha Jesús Alberto García Lerena muchas gracias por tu aportación en el proyecto en la validación experimental.

Muchas gracias a José Manuel Galindo, por todo el apoyo en el trabajo experimental siempre ha estado guiándonos. Tea por ayudarnos, dándonos sus valiosos consejos y compartiendo sus experiencias. A Martín por ayudarnos siempre en lo que necesitemos.

A otros amigos, Tamyko, Chema, Gretter, Diego, Manu, Roy, Adrián y Filisola muchas gracias. A los amigos de Cancún Víctor, Julio y Andrea. Por estar en el proceso de inicio de la maestría y en los tiempos oscuros que pasé ustedes estaban allí muchas gracias de corazón. Y por supuesto aunque seas nuevo muchas gracias Luismi.

A la Dra. Rosaura Hernández Rivas y a Dani por su ayuda y sobretodo por la donación del transcriptoma que fue importante para este proyecto. Muchas gracias a los Dr. Edgar Morales Ríos y Dr. Luis Marat Alvarez Salas por asesorarme en este trabajo de tesis y por las correcciones y aclaraciones hechas.

Muchas gracias al Dr. Alfonso Méndez Tenorio, por sus sabios consejos y su ayuda en el área que me gusta que es la bioinformática. Y por último al Dr. Jesús Valdés Flores por tenerme fe y confianza en permitirme realizar este proyecto en esta área increíble y por sus aportaciones. También por permitirme ser parte de su equipo y estar en su laboratorio.

El jurado designado por el Departamento de Bioquímica del Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional aprobó la tesis titulada "Caracterización, cuantificación y ensamblado *de novo* de circRNAs a partir de datos de RNA-seq de cepas virulentas y no virulentas de *Entamoeba histolytica*", presentado por el IBT Cristian Julio César Padrón Manrique el día 30 de agosto de 2019.

Dr. Jesús Valdés Flores

Dr. Alfonso Méndez Tenorio

Dr. Edgar Morales Ríos

Dr. Luis Marat Alvarez Salas

Glosario

BSJ: Una unión entre dos exones relacionados en el orden opuesto con respecto a sus posiciones en la secuencia de referencia. Una unión circular es un tipo típico de BAM.

BWA-MEM: Burrows-Wheeler Aligner es un algoritmo para mapear lecturas de secuenciación tipo Illumina utilizando que utiliza la transformada Burrows-Wheeler.

cDNA: DNA complementario, copia del RNA mensajero sintetizada en el laboratorio por medio de la transcriptasa inversa.

circRNA: Un tipo de molécula de RNA que forma un bucle cerrado covalentemente.

EBI: European Bioinformatics Institute.

Elemento de un circRNA: son los componentes que conforman la parte interna de un circRNA pueden ser exónicos, intrónicos e intergénicos.

Ensamble de circRNA: es un método que permite tomar varias lecturas de secuenciación y ensamblarlas con el fin de lucidar los elementos internos o la parte interior del circRNA.

etsRNA: external transcribed spacer RNA

Extensión: La extensión de un archivo es la parte de su nombre que indica de qué tipo es. El nombre completo de cualquier archivo consta siempre de dos partes separadas por un punto (por ejemplo "Windows.exe" o "read.fq"). Lo que está a la izquierda del punto es el nombre en sí del archivo

FASTA: En bioinformática, el formato FASTA es un formato de fichero informático basado en texto, utilizado para representar secuencias bien de ácidos nucleicos, bien de péptido, y en el que los pares de bases o los aminoácidos se representan usando códigos de una única letra. El formato también permite incluir nombres de secuencias y comentarios que preceden a las secuencias en sí.

FASTQ: El formato FASTQ es un formato basado en texto para almacenar tanto una secuencia biológica (generalmente una secuencia de nucleótidos) como sus puntajes de calidad correspondientes. Tanto la letra de secuencia como el puntaje de calidad están codificados con un solo carácter ASCII por brevedad.

flicRNA: Full-Length Intron Circular RNA. En español, RNA circular intrónico de longitud completa.

FSJ: Una unión entre dos exones relacionados en el mismo orden en relación con sus posiciones en la secuencia de referencia. Las uniones de corte y empalme de un mRNA son un tipo típico de FSJ.

Genoma de referencia: Un genoma de referencia (también conocido como ensamblaje de referencia) es una base de datos digital de secuencias de ácido nucleico, ensamblada por científicos como un ejemplo representativo del conjunto de genes de una especie. Como a menudo se ensamblan a partir de la secuenciación de ADN de varios donantes u organismos, los genomas de referencia no representan con precisión el conjunto de genes de una sola persona u organismo.

Genoteca: Una genoteca de DNA copia, es un tipo de genoteca que contiene copias de DNA copia de la población de RNA mensajero presente en un determinado tejido u organismo.

IRES: internal ribosome entry site.

KEGG: Kyoto Encyclopedia of Genes and Genomes

Linux: es un sistema operativo libre tipo Unix POSIX; multiplataforma, multiusuario y multitarea.

lncRNA: long non coding RNA.

m⁶A: La N⁶-metiladenosina (m⁶A) es una modificación abundante en el ARNm.

MBL proteins: Mannan-binding lectin proteins.

miRNA: micro RNA.

ncRNA: no coding RNA.

Single y Paired-ends: En la lectura de single-end, el secuenciador, genera la secuencia de pares de bases de solo uno de los extremos de los fragmentos del DNA a secuenciar. En las lectura de paired-ends, se obtienen las secuencias de ambos extremos de cada fragmento secuenciado.

PCR: Polymerase Chain Reaction.

Pipeline: Consiste en una cadena de procesos conectados de forma tal que la salida de cada elemento de la cadena es la entrada del próximo. Permiten la comunicación y sincronización entre procesos.

Pol II: Polimerasa II.

poly(A)⁺: RNA enriquecido con mRNA poliadenilado.

RBP: RNA Binding Protein.

RNA-seq: (“secuenciación de RNA”), también llamado Secuenciación del Transcriptoma Entero para Clonación al Azar 1 (del inglés Whole Transcriptome Shotgun Sequencing), utiliza la secuenciación masiva (NGS) para revelar la presencia y cantidad de ARN en una muestra biológica en un momento dado.

Script: En la programación de computadoras, un script es un programa o secuencia de instrucciones que es interpretado o ejecutado por otro programa en lugar de por el procesador de la computadora (como lo es un programa compilado).

SRA: Sequence Read Archive. Base de datos de lecturas de secuenciación perteneciente al NCBI.

NCBI: Base de datos del National Center for Biotechnology Information

NGS: Secuenciación de Nueva Generación (del inglés Next Generation Sequencing).

Terminal: En informática, una terminal o consola (hardware) es un dispositivo electrónico o electromecánico que se utiliza para interactuar con un computador. Suele confundirse con su homónimo virtual, programado para emular las especificaciones de un terminal estándar.

UTR : Untranslated region

Lista de tablas

Tabla	Página
Tabla 1. Información general de la base de datos.	21
Tabla 2. Oligonucleótidos y condiciones de amplificación.	33
Tabla 3. Bases de datos de secuenciación de <i>Entamoeba</i> reportados por diferentes grupos de investigación.	34
Tabla 4. CircRNAs con más lecturas de BSJ en <i>E. invadens</i> detectados por CIRI2.	49

Lista de figuras

Figuras	Página
Figura 1. Clasificación de ncRNA	2
Figura 2. Unión de corte y empalme retrógrada característica distintiva de los CircRNAs	4
Figura 3. Función de los circRNAs.	5
Figura 4. Funciones biológicas de circRNAs descubiertas hasta el momento	6
Figura 5. Diferentes métodos de extracción de RNA para el enriquecimiento de circRNAs en orden de crecientes cantidades relativas de circRNAs	8
Figura 6. Algoritmo de detección de circRNAs basado en un alineamiento dividido.	9
Figura 7. Algoritmo de detección de circRNAs basados en una pseudoreferencia.	10
Figura 8. Característica del solapamiento inverso para la identificación de circRNAs.	11
Figura 9. Flujo de trabajo de CIRI-FULL.	12
Figura 10. Ciclo de vida de <i>E. histolytica</i>	13
Figura 11. Estrategia experimental	20
Figura 12. Compendio de circRNA en cepas virulentas y no virulentas de <i>E. histolytica</i>	35
Figura 13. Histograma del número de circRNAs ensamblados <i>E. histolytica</i> en función de su longitud.	36
Figura 14. Coordenadas del circRNA cuyo ID es DS571186:1621 1788 con dos elementos exónicos, expresado diferencialmente en <i>E. histolytica</i> .	37
Figura 15. Correlación entre los promedios de los TPM lineales y TPM circulares en <i>E. histolytica</i> .	38
Figura 16. Correlación entre los promedios de los TPM circulares expresados diferencialmente y los promedios de los TPM lineales en <i>E. histolytica</i> .	39
Figura 17. Comparación de TPM de circRNAs de cepas virulentas y no virulentas.	40
Figura 18. Comparación de fracciones de transcritos circulares categorizado por cepas virulentas y no virulentas.	41
Figura 19. Análisis de los Principales Componentes de las genotecas de cepas virulentas y no virulentas.	42
Figura 20. Gráfica de Volcán de la expresión diferencial de los circRNAs.	43

Figura 21. Heatmap representando los niveles de expresión normalizada por TPM de los circRNAs expresados diferencialmente.	44
Figura 22. Heatmaps representando los niveles de expresión normalizada por TPM de los genes parentales de los circRNAs de cepas virulentas y no virulentas.	45
Figura 23. Heatmap de las tres isoformas circRNAs más expresados que contienen un mismo BSJ cuyo gene parental es EHI_169670	46
Figura 24. Validación del circRNA DS571377:17778 18245_A con locus EHI_169670 por RT-PCR divergente utilizando primers divergentes	46
Figura 25. Alineamiento de los circRNAs más expresados en <i>E. invadens</i> y <i>E. histolytica</i> .	48
Figura 26. Principio y final del circRNA con ID scaff_1105074726403:279553 280008 tipo intrónico/exónico.	49
Figura 27. Información de enriquecimientos funcionales KEGG y del estado de la red de interacciones de los genes parentales de los circRNAs detectados en <i>E. histolytica</i> .	51
Figura 28. Análisis de enriquecimientos funcionales KEGG en la red de interacciones de los genes parentales de los circRNAs en <i>E. histolytica</i>	52
Figura 29. Información de enriquecimientos funcionales KEGG y del estado de la red de interacciones de los genes parentales ortólogos de <i>E. histolytica</i> de los circRNAs identificados en <i>E. invadens</i> .	53
Figura 30. Networking de interacciones de los genes parentales ortólogos en <i>E. histolytica</i> de los circRNAs encontrados en <i>E. invadens</i> y el Enriquecimientos funcionales con un mínimo score requerido con un nivel de confianza medio de 0.400.	54
Figura 31. Network de las interacciones de los genes parentales de los circRNAs expresados diferencialmente en <i>E. histolytica</i> y su análisis de enriquecimiento funcional	55
Figura 32. El "árbol de la vida" que muestra los tres dominios principales de los organismos vivos: las bacterias, las arqueas y la eucariota.	59

Índice

Abstract	I
Resumen	II
Agradecimientos.....	III
Glosario	V
Lista de tablas	VIII
Lista de figuras	IX
Índice.....	XI
1. Introducción.....	1
1.1 RNA no codificantes	1
1.2 Biogénesis de los circRNAs.....	2
1.3 Función de los circRNAs.....	5
1.4 Enriquecimiento de circRNAs en experimentos para RNA-seq	7
1.5 Detección computacional de circRNA utilizando datos de secuenciación masiva de experimentos de RNA-seq.....	8
1.6 <i>Entamoeba histolytica</i> y <i>Entamoeba invadens</i>	12
1.7 Factores relacionados a virulencia y ncRNAs en <i>E. histolytica</i> e <i>E. invadens</i>	14

2. Antecedentes	16
3. Justificación.....	17
4. Hipótesis.....	17
5. Objetivos	18
5.1 Objetivo general.....	18
5.2 Objetivos específicos	18
6. Estrategía experimental.....	19
7. Materiales y métodos	21
7.1 Hardware	21
7.2 Descarga de bases de datos	21
7.3 Bioconda.....	23
7.4 Control de Calidad de datos de secuenciación masiva	23
7.5 CIRI-full pipelines.....	24
7.6 Transcritos lineales	29
7.7 Cuantificación de circRNAs y transcritos lineales	30
7.8 Expresión diferencial.....	30
7.9 Análisis de las redes de asociación de proteínas	31
7.10 Análisis estadísticos.....	31
7.11 Cultivo de trofozoitos de <i>Entamoeba histolytica</i>	31

7.12 Extracción de RNA	32
7.13 RT-PCR	32
8. Resultados	34
8.1 Selección de genotecas.....	34
8.2 Caracterización de circRNAs en cepas virulentas y no virulentas de <i>E. histolytica</i>	35
8.3 CircRNAs formados por diferentes elementos.....	36
8.4 Relación entre la expresión de los transcritos circulares y lineales	38
8.5 Expresión diferencial de circRNAs en cepas virulentas y no virulentas	42
8.6 Los circRNAs del locus EHI_169670	46
8.7 Los CircRNAs en <i>Entamoeba invades</i>	49
8.8 Análisis STRING de los circRNAs identificados de las cepas <i>E. histolytica</i> y <i>E. invadens</i>	50
8.8.1 Análisis STRING de los genes parentales de los circRNAs identificados en <i>E. histolytica</i>	51
8.8.2 Análisis STRING en de los genes parentales ortólogos en <i>E. histolytica</i> de los circRNAs identificados en <i>E. invadens</i>	53
8.8.3 Análisis STRING de los circRNAs expresados diferencialmente de <i>E. histolytica</i> de las cepas virulentas y no virulentas	55
9. Discusión.....	57
10. Conclusiones.....	65
11. Perspectivas.....	66
12. Referencias	69

13. Apéndice	77
Apéndice A. Lista de circRNAs con más de un elemento interno	77
Apéndice B. Lista de los 39 circRNAs expresados diferencialmente	79
Apéndice C. Gráfica de volcán abarcando los 39 circRNAs expresados diferencialmente	81
Apéndice D. Red e información de otros tipos de enriquecimiento funcional diferentes a KEGG en la red de interacciones entre proteínas/genes parentales de circRNAs detectados en <i>E. histolytica</i> con una confianza del 0.150	82
Apéndice E. Red e información de otros tipos de enriquecimiento funcional diferentes a KEGG en la red de interacciones entre proteínas/genes parentales ortólogos de <i>E.</i> <i>histolytica</i> de los circRNAs identificados en <i>E. invadens</i> con una red de interacciones con una confianza baja del 0.150	84
Apéndice F. Lista de circRNAs detectados en el trabajo de Weber y cols. (2016).	86
Apéndice G. IDs de CircRNAs detectados en las genotécas de <i>Entamoeba Histolytica</i> HM1:IMSS	87
Apéndice H. IDs de CircRNAs detectados en las genotécas de <i>Entamoeba Histolytica</i> Rahman	93
Apéndice I. <i>Script</i> en R de la expresión diferencial utilizando DESeq2.	97
Apéndice J. Mápa genómico de algunos circRNAs conformados por dos elementos internos	99
Apéndice K. Genes ortólogos de <i>E. histolytica</i> de los genes parentales de los circRNAs detectados en <i>E. invadens</i>	101

1. Introducción

1.1 RNAs no codificantes

El término RNA no codificante (ncRNA) es comúnmente empleado para RNA que no es codificado a proteína, esto no quiere decir que tales RNAs no contengan información para ser codificado a proteínas o no tengan funciones. Aunque es asumido generalmente que la mayoría de la información genética es traducida a proteínas, evidencia reciente sugiere lo contrario. La mayoría de los genomas de los metazoos es transcrito mayoritariamente en ncRNAs (Long y cols., 2017). Los ncRNAs en mayor proporción son productos de *splicing* alternativo y/o procesados en productos pequeños (Mattick y cols., 2006).

Hay diferentes clasificaciones de los ncRNAs, una de ellas es si son moléculas lineales o circulares como se puede observar en la figura 1, dentro de los ncRNA lineales encontramos los de *housekeeping* (como los rRNAs, tRNA y los snoRNAs) y los regulatorios (sRNA y lncRNA).

Existe un nuevo participante en los ncRNAs, los RNAs circulares (circRNA) que son moléculas de RNA monocatenarios que a diferencia de los RNAs lineales, forman un bucle continuo cerrado covalentemente. Se clasifican en 5 tipos según Liu y cols. (2017): los exónicos, intrónicos, UTR, intragénicos y otros circRNAs que contienen en su estructura interna una combinación de diferentes tipos de circRNA.

El tipo de circRNAs que se identificaron en este proyecto son de tipo exónico. Para la formación del circRNA es necesario que ocurra el fenómeno de corte y empalme retrógrado o *backsplicing* por sus siglas en inglés.

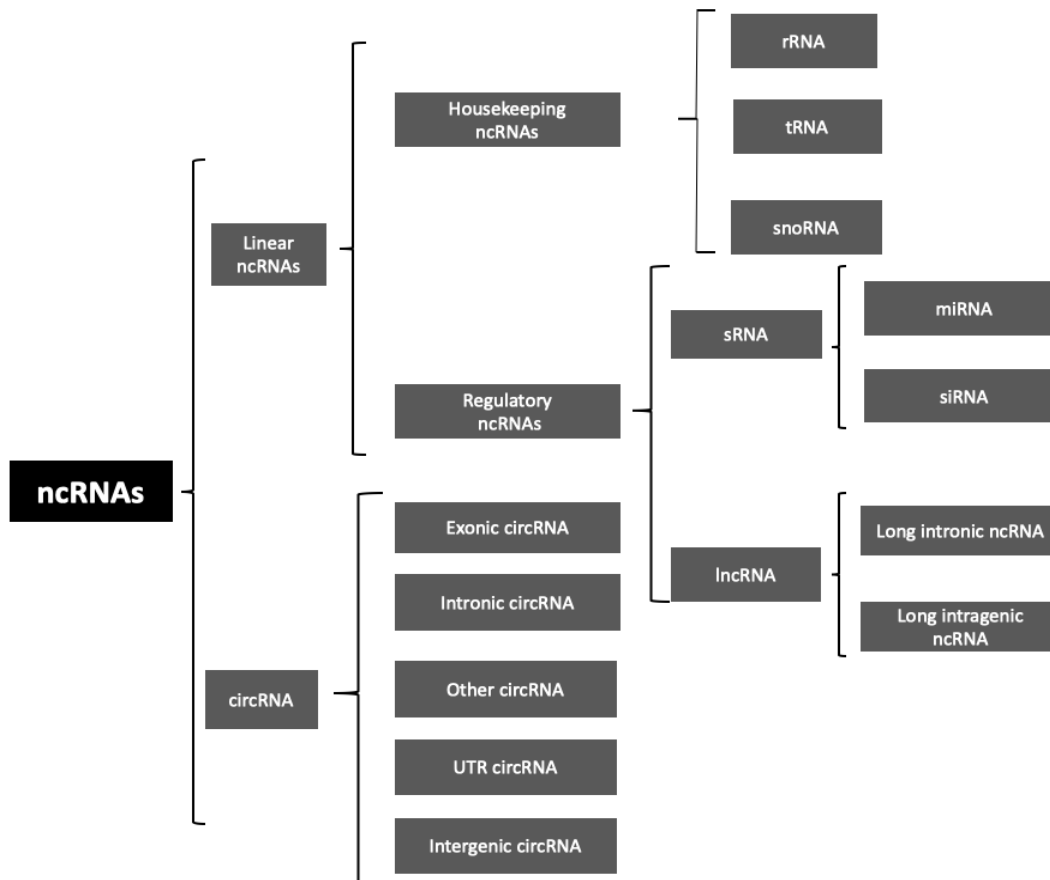


Figura 1. Clasificación de ncRNA. Los ncRNAs organizados en si son circulares o lineales (Liu y cols., 2017).

1.2 Biogénesis de los circRNAs

En el fenómeno de corte y empalme progresivo (*forward-splicing*) o *splicing* lineal, se unen dos exones progresivamente, permitiendo la maduración del pre-mRNA a mRNA. Para formarse la unión de corte y empalme progresivo (FSJ, *Forward-Splice Junction*), se requiere que la secuencia intrónica del pre-mRNA contenga dos sitios de *splicing*, uno 5' río abajo y el otro 3' río arriba (5'ss y 3'ss respectivamente). Esta unión de exones es progresiva debido a la ligación del nucleótido exónico adyacente al 5'ss con el nucleótido exónico adyacente al 3'ss. La formación del FSJ es mediada por la maquinaria del *splicing*. Por lo tanto, ocurriendo dos reacciones de transesterificación, generándose la

remoción del intrón en un producto intrón lariat, la unión de exones progresivamente y la producción del mRNA lineal (Shi, 2017).

Por otro lado, el fenómeno de corte y empalme retrógrado (*backsplicing*) o *splicing* circular une dos exones retrógradamente, permitiendo la formación del circRNA de tipo exónico a partir de un pre-mRNA. Para formarse la unión de corte y empalme retrógrado (BSJ, *Back-Splice-Junction*) se requieren de dos señales de *backsplicing*, una 5' río abajo y otra 3' río arriba (5' AG y 3' GT respectivamente). Esta unión de exones es retrógrada debido a la ligación del nucleótido exónico adyacente a la señal del *backsplicing* 3'GT con el nucleótido exónico adyacente a la señal de *backsplicing* 5' AG. La unión es de orden inverso al *forward-splicing*, generándose un bucle que une los extremos 5' y 3' del mRNA mensajero (Chen y Yang, 2015). En (figura 2) se ilustra la estructura del pre-mRNA para explicar los fenómenos del *back-splicing* y del *forward-splicing* descritos anteriormente.

Debido al fenómeno de *backsplicing*, los circRNAs exónicos son estables. Lo anterior es debido a que forman un bucle y evitan ser degradados por las exonucleasas teniendo una vida media de incluso días. Las contrapartes lineales de lo circRNAs tienen una vida media mucho menor (Eneka y cols., 2016).

En nuestro laboratorio se descubrieron los flicRNAs (*full-length intronic circular RNA*), un nuevo tipo de circRNA, los cuales son estables con el paso del tiempo (Mendoza-Figueroa y cols., 2018). Por otro lado, el mecanismo mediante el cual los circRNAs exónicos son eventualmente degradados es un misterio, pero se ha propuesto un mecanismo de degradación mediante exosomas. Lo anterior es importante debido a que la degradación de los circRNAs es crucial para la expresión de los mismos (Conn y cols., 2017). A pesar de que el *backsplicing* es relativamente ineficiente (poca expresión a comparación de sus contrapartes lineales), los circRNAs pueden llegar a acumularse (Lasda y Parker, 2016).

Con respecto a los circRNA de tipo exónico, estos se originan a partir de sus genes parentales. Es el tipo más abundante de circRNA y es ubicuo en eucariotas. Aproximadamente contienen un 88.8% señales canónicas de *backsplicing* 5' AG y 3' GT del total de circRNAs exónicos (Ye y cols., 2017).

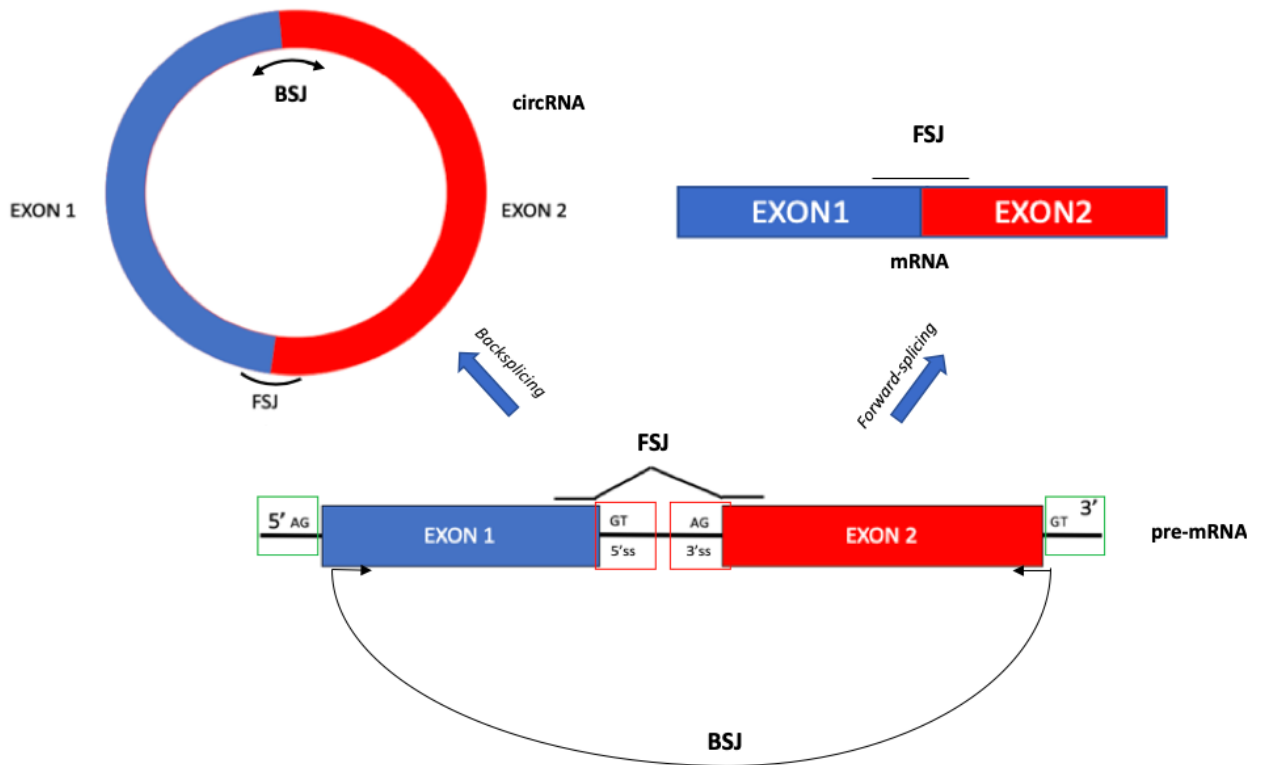


Figura 2. Unión de corte y empalme retrógrada característica distintiva de los circRNAs. El fenómeno de *backsplicing* o *splicing* circular forma la unión de *backsplicing* retrógrada (BSJ). En cambio en su contraparte lineal la unión de corte y empalme progresiva también conocida como *forward-splicing* o *splicing* lineal, forma la unión de *forward-splicing* progresiva (FSJ). Para la formación del FSJ se requiere en la secuencia intrónica del pre-mRNA contenga dos sitios de *splicing*, uno 5' río abajo y el otro 3' río arriba (5'ss y 3'ss respectivamente). Para la formación del BSJ requiere de dos señales de *backsplicing*, una 5' río abajo y otra 3' río arriba (5' AG y 3' GT respectivamente) en el pre-mRNA. El BSJ se forma debido a la ligación de los nucleótidos exónicos adyacentes a las señales de *splicing* 5'AG y 3'GT. Si un circRNA contiene dos exones va a contener un FSJ y un BSJ. En cambio, si un circRNA en su estructura interna contiene sólo un exón este tendrá solamente un BSJ, sin tener un FSJ. En los recuadros verdes son las señales de *backsplicing*. En los recuadros rojos son los sitios de *splicing*.

1.3 Función de los circRNAs

Desde la llegada de la secuenciación masiva se inició el análisis del estudio de la prevalencia, biogénesis y sus posibles funciones de los circRNAs. Se pensaba que los circRNAs podían ser errores de la maquinaria de *splicing* siendo estos productos secundarios intrascendentes de la maquinaria del *splicing*, debido a que sólo se conocía la función de los circRNAs como esponjas de miRNAs (Huang y cols., 2015). Con el paso del tiempo se descubrieron mayores funciones. A continuación, se describen brevemente las funciones más representativas.

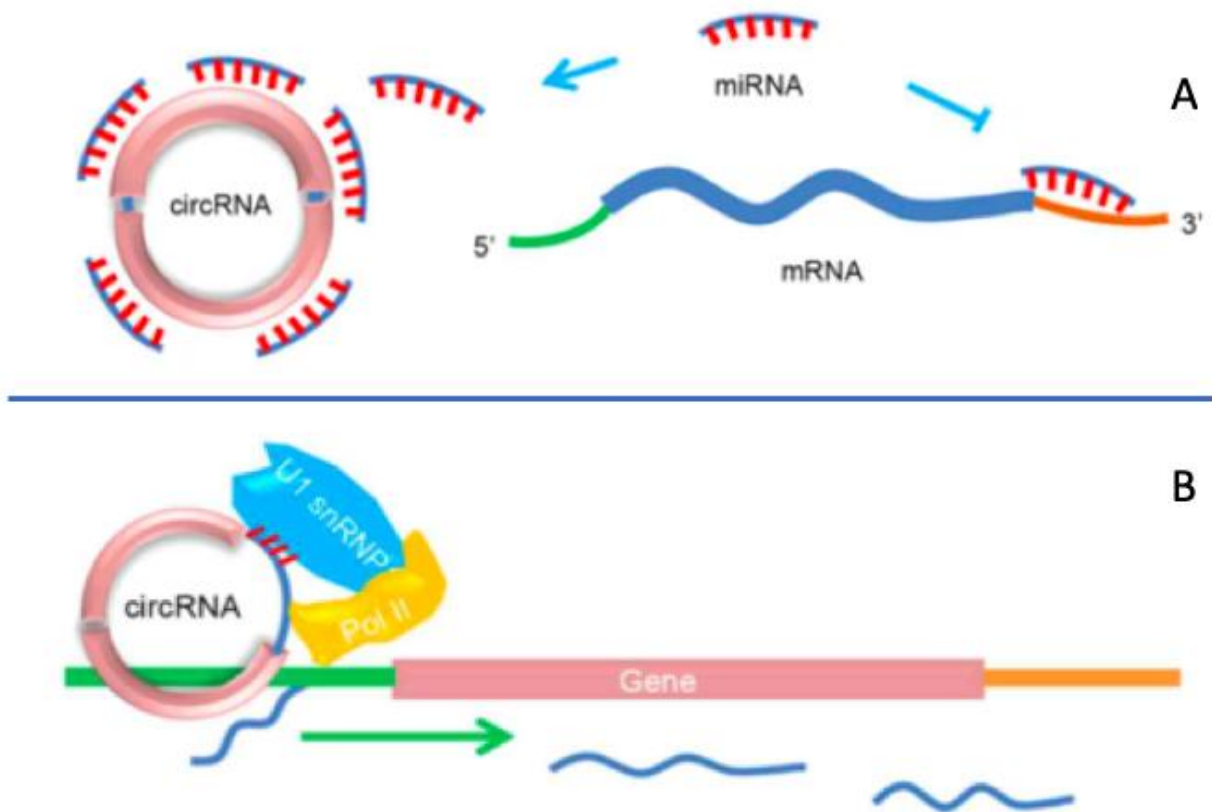


Figura 3. Función de los circRNAs. En (A) es un circRNA con función de esponja de miRNAs, así evitando que los miRNAs se unan a su objetivo diana (región 3' UTR del mRNA), el circRNA está codificado por exones y está localizado en el citoplasma. En (B) es un circRNA cuya función es la de aumentar la transcripción de su gen parental, apareándose con el snRNA U1 del U1snRNP que interactúan con Pol II en el complejo del inicio de la transcripción del gen parental. Así, estimulando su transcripción. El circRNA está codificado por exones y un intrón (Ren y cols., 2017).

Una de ellas es la de esponjas de miRNAs citoplásmicas codificadas por exones que contienen sitios de unión complementarios a miRNAs. Por lo tanto, suprimen la habilidad de los miRNAs a unirse a sus objetivos en la 3'UTR del mRNA. Regulando así a nivel postranscripcional (figura 3A). Por otro lado, tenemos al círculo exónico intrónico apareado con el snRNA U1 del U1 snRNP que interactúan con Pol II en el complejo de inicio de transcripción del gen parental estimulando la transcripción del mismo. Cuando este círculo es silenciado mediante siRNAs, la transcripción del gen parental disminuye (Li y cols., 2015) (figura 3B). Con el paso del tiempo es que se han descubierto aún más funciones que las anteriores descritas (figura 4).

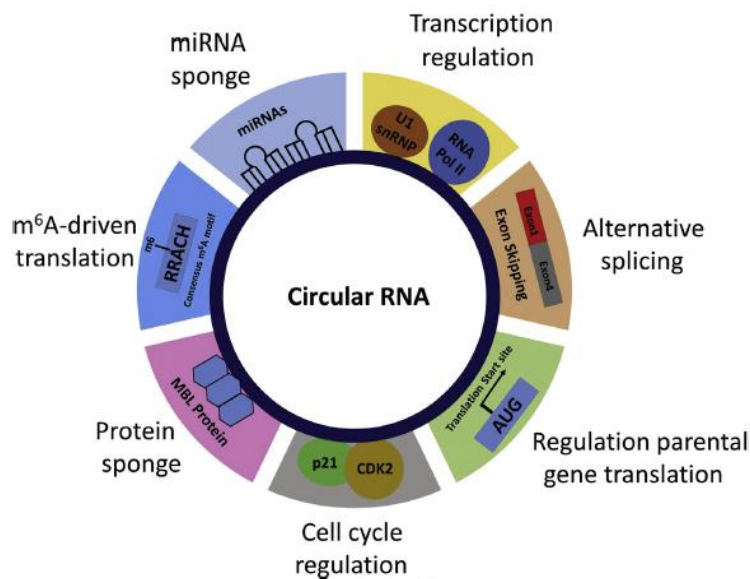


Figura 4. Funciones biológicas de circRNAs descubiertas hasta el momento. (Guanqun Huang y cols., 2017). **1. Splicing alternativo**, en la biogénesis de los circRNA puede competir contra el *splicing* canónico del pre-mRNA para facilitar el *splicing* alternativo (Kelly y cols., 2015). **2. Regulación de la traducción del gen parental**, si el circRNA contiene un sitio de inicio de la traducción, el mRNA trunco no será traducido, por lo tanto reduciendo la traducción del gen parental de manera indirecta (Chao y cols., 1998). **3. Regulación del ciclo celular**, el circRNA Circ-Foxo3 puede unirse a p21 y CDK2 para regular la progresión del ciclo celular (Du y cols., 2016). **4. Esponja de proteínas**, está reportado que el circRNA CircMbl se une directamente a proteínas *muscleblind* y así, disminuyendo la actividad de las proteínas MBL (Ashwal-Fluss y cols., 2014). **5. Traducción llevado por m⁶A**, un solo N⁶-metiladenosina en un motivo consenso RRACH en circRNAs es más que suficiente para llevar a cabo el inicio de la traducción (Yang y cols., 2017). **6. Esponja de miRNAs**, está reportado que el circRNA has_circ_0078710 es una esponja del miRNA oncoprotector miRNA-31 (Xie y cols., 2019). **7. Regulación de la transcripción**, el circRNA exónico

intrónico apareado con el snRNA U1 del U1 snRNP que interactúan con Pol II en el complejo del inicio de la transcripción del gen parental, estimula la transcripción del gen parental (Y. Zhang y cols., 2013).

1.4 Enriquecimiento de circRNAs en experimentos para RNA-seq

Un diseño eficiente de las genotecas para la identificación de circRNAs mediante experimentos de RNA-seq, se debe empezar con la construcción de estas a partir de un RNA enriquecido con las moléculas de circRNAs. Hay diferentes métodos para la extracción de RNA para su uso en RNA-seq. Las genotecas construidas a partir de RNA total son las que generan menor rendimiento para la detección de circRNAs ya que incluyen los mRNAs poliadenilados y no poliadenilados, todo el ncRNAs, circRNAs y sobre todo rRNA que es mayoritario e interfiere con la detección de cualquier forma de RNA que no sea ribosomal.

La modificación cotranscripcional de la poliadenilación de mRNA es una ventaja para la detección de éste. Aunque el enriquecimiento de circRNAs es mucho mayor en una preparación enriquecida en poly(A)⁺ comparado con una extracción de RNA total, en una preparación de poly(A)⁺ todavía contiene una buena proporción de RNA ribosomal, además de otros ncRNAs y del propio mRNA que interfieren con la detección de circRNAs. Sin embargo, la detección de circRNA en este método se puede compensar con una mayor profundidad de secuenciación.

La depleción del rRNA con RiboMinus™ por selección del híbrido es un procedimiento que elimina una fracción significativa del rRNA ya que tiene sondas dirigidas contra estos. En estas preparaciones hay mayor enriquecimiento de circRNAs, pero hay algunas sondas que no reconocen el rRNA de algunas especies. Por lo tanto, no logran tal eliminación significativa del rRNA. Entre estas especies se encuentran *E. histolytica* y *E. invadens* (Nozaki y Bhattacharya, 2015).

Por otro lado, tenemos a la combinación de la depleción de rRNA y de mRNA poliadenilados, lo cual es un método eficiente para encontrar miRNAs y circRNAs. Con

bajas cantidades de mRNA poliadenilados. Este método es efectivo para el enriquecimiento de miRNAs y circRNAs toda vez que haya una eliminación significativa del rRNA (Szabo y Salzman, 2016).

El método óptimo para construir genotecas con fines de detección de circRNAs es a partir de RNA tratado con Ribominus™ y RNasa R (figura 5), exonucleasa que degrada RNA de cadena sencilla en dirección 3'-5' y que se ha demostrado degrada selectivamente mRNA (Venkataraman y cols., 2014). Una vez que las genotecas están construidas estas se utilizan en la secuenciación masiva para la detección de los circRNAs.



Figura 5. Diferentes métodos de extracción de RNA para el enriquecimiento de circRNAs en orden de crecientes cantidades relativas de circRNAs. Los métodos son los siguientes en orden de enriquecimiento de circRNAs; poly(A)⁺, enriquecimiento de mRNA poliadenilados; rRNA⁻, depleción del rRNA; rRNA⁻ y poly(A)⁻, depleción del rRNA y mRNA poliadenilado; rRNA⁻ y RNaseR⁺, depleción del rRNA y tratamiento con RNasa R para la eliminación de transcritos lineales (Szabo y Salzman, 2016).

1.5 Detección computacional de circRNA utilizando datos de secuenciación masiva de experimentos de RNA-seq

Un genoma de referencia es necesario para todos los algoritmos de detección de circRNAs, pero puede ser utilizado en diferentes maneras en el flujo de trabajo de la detección de circRNAs. El uso más común implica la alineación directa de todas las lecturas de secuenciación contra el genoma de referencia. Debido a que los circRNA son

diferentes a otros RNA por su circularidad. Una característica obvia que puede ser capturada por un alineamiento es la unión circular, llamada BSJ. En contraste al las FSJs en el mRNA que generan lecturas de secuenciación alineadas co-linealmente en el genoma, las lecturas que abarcan las BSJs están divididos en dos segmentos y se alinean al genoma en orden reverso o en quiasma. Por lo tanto, los algoritmos de detección en esta categoría se denominan enfoques basados en alineamientos divididos (*split-alignment-based approaches*) (figura 6). La mayoría de los algoritmos de detección como find_circ (Memczak y cols., 2013), CIRCexplorer (Zhang y cols., 2014), CIRI2 (Gao, Wang, y Zhao, 2015) y UROBUS (Song y cols., 2016), son clasificados en esta categoría.

Además, dado que las señales de *backsplicing* son mayoritariamente canónicas, estas se pueden usar como un método de filtración en la detección de circRNAs exónicos, en la búsqueda de circRNA con señales de *backsplicing* canónicos. También, si al algoritmo (e.g el algoritmo de CIRI2) se le provee de anotaciones del genoma de referencia, se pueden detectar circRNAs con señales de *splicing* no canónicas (e.g. 5' AT y 3' AC), mismas que aparecen minoritariamente en humano o rata.

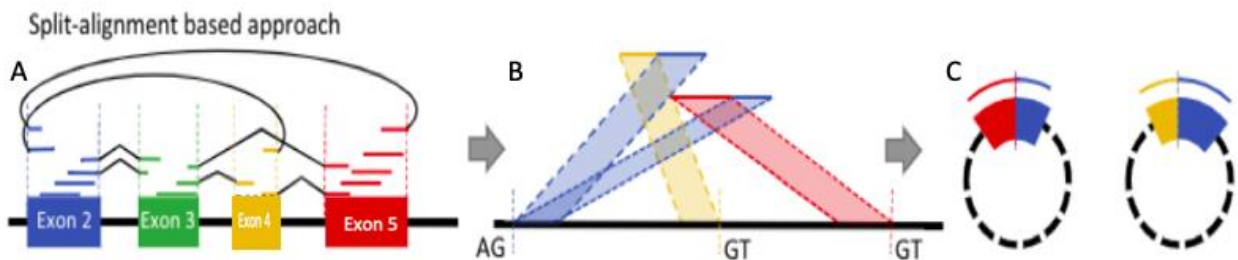


Figura 6. Algoritmo de detección de circRNAs basado en un alineamiento dividido. En (A), hay tres tipos de lecturas de secuenciación cuando se alinean contra un genoma de referencia: el primero son lecturas que son colinealmente alineadas a un exón, el segundo son lecturas que son co-linealmente alineadas a una FSJ y el tercero son lecturas que no son co-linealmente alineadas. Dentro de las lecturas que no son colinealmente alineadas, se encuentran las lecturas de BSJs. Las lecturas de BSJs son las utilizadas por algoritmos para la identificación de circRNAs. Por lo tanto, el algoritmo utiliza las lecturas no colinealmente alineadas como candidatos a lecturas que abarcan BSJs de circRNAs. En (B) el algoritmo hace un alineamiento de las lecturas no colinealmente alineadas contra el genoma de referencia con el fin de identificar lecturas que tengan una alineación dividida y en orden inverso. Lecturas que no tengan alineación dividida y en orden inverso son descartadas. Se pueden utilizar las señales de *splicing* canónicas

como un proceso de filtración. En (C), son lecturas identificadas que abarcan BSJs provenientes de circRNAs.

Por otro lado, existen algoritmos como KNIFE (Szabo y cols., 2015) y NCLscan (Chuang y cols., 2016), en los cuales el genoma de referencia está combinado con su anotación correspondiente para construir pseudosecuencias de las uniones de BSJ putativa. Los algoritmos dentro de esta categoría se les denomina enfoques basados en pseudoreferencias (*pseudoreference-based approaches*) (figura 7).

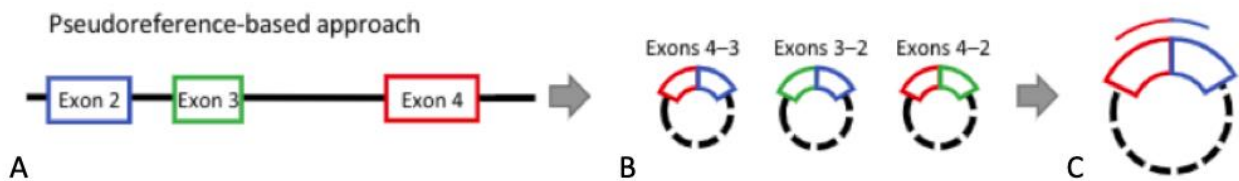


Figura 7. Algoritmo de detección de circRNAs basados en una pseudoreferencia. En (A), se utiliza un genoma de referencia con su respectiva anotación del genoma, con el fin de crear pseudosecuencias de las uniones de BSJs. En (B), creación de la pseudoreferencia que contiene pseudosecuencias de las uniones de BSJ. Posteriormente en (C), todas las lecturas de secuenciación se alinean a las pseudosecuencias de las BSJs. Lo anterior es con el fin identificar lecturas que abarquen un BSJ, las lecturas identificadas que abarquen una BSJ son indicativo que provienen de circRNAs. Sin embargo, la desventaja de este tipo de algoritmos es que sólo se detectarán uniones de BSJs de circRNAs proveídas por la pseudoreferencia.

En el caso de los algoritmos basado en alineamientos divididos, se tiene que mencionar que para la identificación de circRNAs, el algoritmo no reconoce lecturas de BSJs que estén desbalanceadas. Una lectura de BSJ desbalanceada, es una lectura de BSJ en donde un segmento que flanquea la BSJ es mucho más corto que el otro segmento flanqueante. Por lo tanto, el segmento más corto puede ser alineado en diferentes regiones del genoma generando falsos positivos en la identificación de circRNAs. Para compensar la problemática se utiliza una característica nueva en la identificación de circRNAs, el *reverse overlap* o solapamiento de reversos extremos.

CIRI-FULL es un algoritmo que utiliza el sobrelapamiento reverse de los extremos 5' y 3' (5' RO y 3' RO, *reverse overlap* del inglés). El sobrelapamiento reverse del extremo 5' consiste en utilizar dos lecturas tipo *paired-end* que contengan en sus extremos 5' secuencias en común. Los extremos 5' con secuencias en común son fusionados convirtiéndose en una lectura más grande de tipo *single-end*. La lectura fusionada es alineada al genoma para encontrar nuevos alineamientos divididos en orden inverso, con el fin de encontrar nuevas BSJs (Figura 8). Aquellas lecturas fusionadas que no contengan alineamientos divididos en orden inverso son descartadas debido a que no pertenecen a BSJs. El sobrelapamiento reverse compensa la problemática de lecturas de BSJs desbalanceadas, ya que pueden ser utilizadas en la búsqueda de circRNAs con este método.

Por último, sí la lectura fusionada con un sobrelapamiento reverse 5' presenta un sobrelapamiento reverse 3' es indicativo de un ensamble completo de los elementos internos de un circRNA. Las características 5' RO y 3' RO son importantes para el ensamble y detección de circRNAs de longitudes pequeñas. (Zheng y cols., 2019). El flujo de trabajo de CIRI-FULL se ilustra en (figura 9).

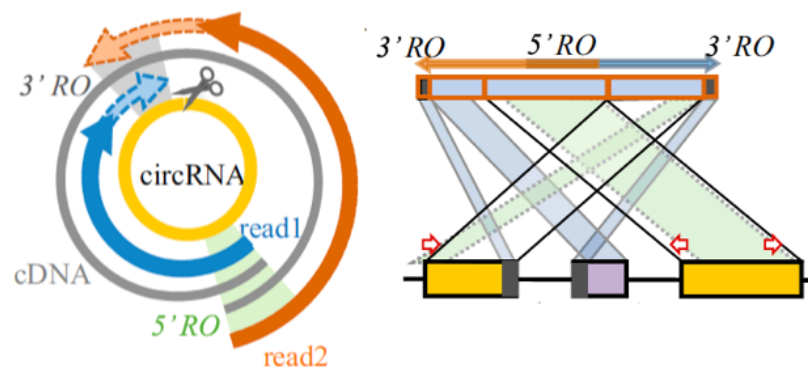


Figura 8. Característica del sobrelapamiento reverse para la identificación de circRNAs. En la figura de la izquierda, las flechas de color azul y naranja punteadas indican lecturas que además de tener un sobrelapamiento reverse 5' presentan un sobrelapamiento reverse 3', las flechas no punteadas indican lecturas que sólo tienen un sobrelapamiento reverse 5'. La figura de la derecha es el esquema de la alineación con el genoma de las lecturas fusionadas o sobrelapadas reversamente en el 5', en la búsqueda de lecturas divididas en orden inverso representativas de BSJs y en la búsqueda del sobrelapamiento reverse 3'. Modificado de (Zheng y cols., 2019).

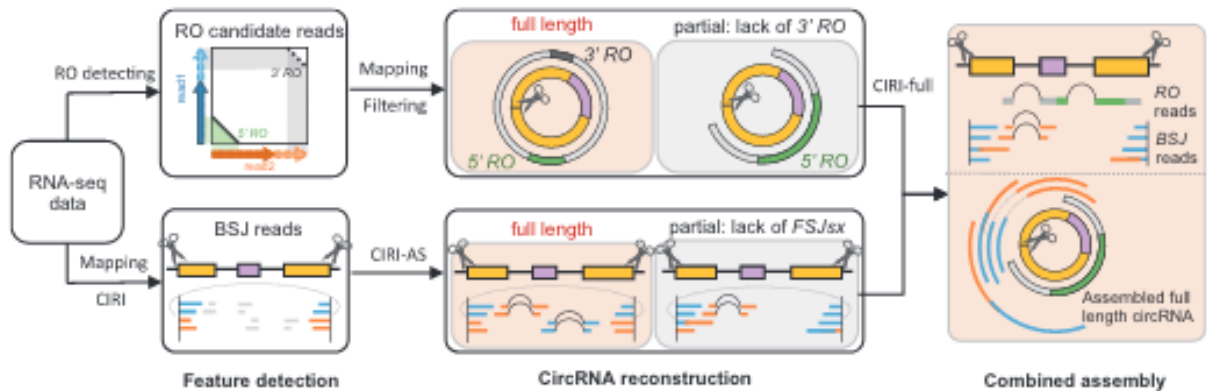


Figura 9. Flujo de trabajo de CIRI-FULL. Consiste en los siguientes pasos: **1.** A partir de datos de RNA-seq se detectan BSJs utilizando el *script* de CIRI2. **2.** Se identifican isoformas alternativas que contengan el mismo BSJ y se hace una reconstrucción completa de circRNAs utilizando lecturas de secuenciación que abarquen FSJ y elementos internos exónicos utilizando el *script* de CIRI-AS. La reconstrucción de los elementos internos de un circRNA puede no estar completa debido a la falta de una FSJ, esta información de los circRNAs no completamente construidos es utilizada para un posterior ensamble combinado. **3.** Por otro lado, CIRIFULL detecta características RO. Determina que lecturas fusionadas 5' RO y 3' RO pueden ser utilizadas para la construcción completa de circRNAs pequeños. También determinan que lecturas fusionadas 5' RO carentes de 3' RO son pertenecientes a BSJ, éstas puede que abarquen o no FSJs. **4.** CIRI-FULL hace un ensamblado combinado, utilizando las lecturas fusionadas con características 5' RO carentes de 3' RO y circRNA que no fueron completamente ensamblados debido a la falta de una lectura de la FSJ. Modificado de (Zheng y cols., 2019).

1.6 Entamoeba histolytica y Entamoeba invadens

El patógeno gastrointestinal *E. histolytica* es el agente causante de la amebiasis, enfermedad que es una amenaza para la salud global ya que se produce aproximadamente un total de 100,000 muertes cada año (“Weekly Epidemiological Record”, 2017). La virulencia de *E. histolytica* se atribuye generalmente a su capacidad para destruir los tejidos a través de la adherencia, matando a la célula huésped y la proteólisis de la matriz extracelular, aunado a la expresión de un gran conjunto de factores de virulencia (Faust y Guillen, 2012). Las infecciones asintomáticas por *E. histolytica* son comunes, aproximadamente entre el 10 y 20% de los individuos infectados presentan síntomas de amebiasis invasiva (Pearson y Singh, 2010). Este organismo tiene un ciclo

de vida relativamente simple (figura 10), el cual consiste en dos estadios. El trofozoíto que es la forma patogénica móvil que pueden invadir múltiples órganos y el quiste que es la forma infectiva. Cuando el quiste es ingerido a través de agua o comida contaminada, se desenquista en el lumen del intestino y produce trofozoítos que terminan de colonizar el intestino mediante la adhesión a las mucinas en donde se alimentan de bacterias de la flora intestinal y se dividen (Faust y Guillen 2012).

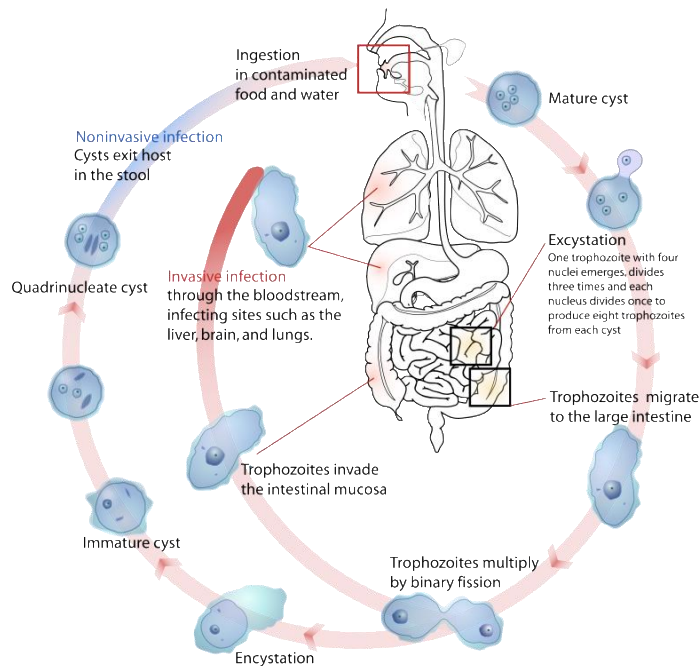


Figura 10. Ciclo de vida de *E. histolytica*. La infección por *E. histolytica* ocurre por la ingestión de comida o agua contaminadas con quistes maduros. El desenquistamiento ocurre en el intestino delgado liberando a trofozoítos, que migran al intestino grueso. Los trofozoítos se multiplican por fisión binaria y producen quistes, los cuales son excretados por las heces fecales. Por la protección que confiere la pared del quiste, este puede sobrevivir días en ambientes extremos y ser la responsable de la transmisión (los trofozoítos se excretan en las heces diarreicas, pero se destruyen rápidamente fuera del cuerpo y si fueran ingeridos no sobrevive al ser expuestos al ambiente gástrico). En muchos casos, los trofozoítos se mantienen confinados al lumen intestinal (infección no invasiva) de los individuos que se convierten en portadores asintomáticos, que excretan los quistes en heces. En algunos pacientes los trofozoítos invaden la mucosa intestinal (infección intestinal), o a través del torrente sanguíneo, en sitios extraintestinales como son hígado, cerebro y pulmones (infección extraintestinal) (Kean, 2017).

Por otro lado, tenemos a *E. invadens*, el cual es un parásito protista de reptiles. Este parásito está cercanamente relacionado con el parásito humano *E. histolytica* causando una invasión similar en reptiles (Ehrenkaufer y cols., 2013), además de su similitud en su morfología y ciclo de vida (Sanchez y Eichinger, 1994). Esta especie se ha usado como sistema modelo en el estudio del desarrollo y el enquistamiento *in vitro*. Particularmente por las dificultades asociadas al estudio del enquistamiento de *E. histolytica* (Ehrenkaufer, Singh y cols. 2013). Diferentes grupos a lo largo del tiempo con diferentes enfoques han estudiado los factores que están relacionados a la virulencia de Entamoeba.

1.7 Factores relacionados a virulencia y ncRNAs en *E. histolytica* e *E. invadens*

Antes de la llegada de la secuenciación masiva, distintos grupos utilizaron diferentes técnicas de biología molecular. Martínez-Palomo y cols. (1973) usaron lectinas con el fin de investigar los monosacáridos en la superficie celular de *E. histolytica* crecidas *in vitro*. Ellos observaron el efecto de la concanavilina A en la aglutinación de cepas de *E. histolytica*. Observaron aglutinación de amebas aisladas de un paciente con enfermedad, mientras en amebas de pacientes asintomáticos no hubo presencia de aglutinación. Lo anterior fue interpretado como un marcador de virulencia. Después de unos años Sargeant y Williams (1978) empezaron a usar un método analítico de isoenzima, originalmente desarrollado para bacterias, para investigar la variación en Entamoeba. Ellos mostraron inicialmente que este método podía distinguir diferentes especies de Entamoeba descubriendo patrones de variación intraespecífica. Más tarde encontraron que los aislamientos de *E. histolytica* se dividieron en dos grupos (más tarde llamados “patógenos” y “no patógenos”) que se correlacionaron con el estado de la enfermedad del paciente del que se aisló la ameba (Sargeant, Williams y Grene, 1978). Otros investigadores como Strachan y cols. (1988) produjeron anticuerpos dirigidos a una isoforma de hexoquinasa de una cepa virulenta utilizada como marcador de virulencia. Pero aun así con los métodos de biología molecular, fueron descubiertos muy pocos factores relacionados con virulencia. Después, con el advenimiento de las técnicas de secuenciación masiva es que estos factores o genes relacionados a virulencia, se pudieron identificar con mayor abundancia.

Después de la secuenciación completa del genoma de varias especies de *Entamoeba* es que se optó por realizar ensayos de microarreglos y de RNA-seq, con el fin de dilucidar genes relacionados con virulencia. Varios grupos de investigación tuvieron sus propios enfoques obteniendo diferentes perfiles de expresión de transcritos en diferentes situaciones como son: privando de glucosa a la ameba y posteriormente reincorporándosela (Tovy y cols., 2011); comparando perfiles de expresión entre cepas virulentas y no virulentas en *steady-state* (Hon y cols., 2013); induciendo ameba a virulencia inoculándola en hámster (Weber y cols., 2016); también utilizando clonas aisladas de ameba patogénicas y no patogénicas (Meyer y cols., 2016).

Por otro lado, se sabe que los miRNAs son un gran grupo de ncRNAs y estos tienen una función importante en la regulación de la expresión génica y en la traducción de proteínas. Diferentes grupos (H. Zhang y cols., 2013; De y cols., 2006; Mar-Aguilar y cols., 2013) se han encargado de encontrar diversos miRNAs en cepas amebianas virulentas y no virulentas utilizando pirosecuenciación y secuenciación masiva.

2. Antecedentes

En el trabajo de Hon y cols. (2013) a partir de cepas virulentas y no virulentas en *steady-state* se realizó una secuenciación masiva de alta profundidad (10^9 de lecturas) cuya biblioteca fue diseñada a partir de RNA poly(A)⁺ para detectar las isoformas de *splicing* alternativo y determinar si funcionalmente eran relevantes respecto al ruido estocástico del procesamiento del RNA. Concluyeron que la generación de las isoformas de *splicing* alternativo son producto principalmente del ruido estocástico de la maquinaria de *splicing*. Dado a que no era su objetivo, los autores no contemplaron analizar las isoformas circulares con BSJs. Dentro de las 10^9 lecturas totales de secuenciación aproximadamente un 5 % no fueron mapeadas/alineadas colinealmente con el genoma, por lo tanto, es muy probable que ese 5 % contengan secuencias de BSJ mismas que pueden usarse para detectar circRNAs mediante los algoritmos antes descritos.

Como ya se ha reportado en metazoo, los circRNAs tienen una gama de funciones y se sabe que son moléculas estables que escapan a la degradación por exonucleasas más que sus contrapartes lineales. Típicamente los circRNAs se expresan menos que su contraparte lineal, sin embargo, se acumulan e incluso tienen modificaciones postraduccionales que probablemente les proveen funciones por descubrir. Es por esto por lo que, se ha descartado la posibilidad de que estas formas circulares no sean productos estocásticos del *splicing* sino productos con funciones aun no descritas, abriendo así un nuevo camino para el entendimiento de la función de estas moléculas en *E. histolytica*. Por otro lado, pocos grupos de investigación se han encargado de la búsqueda de moléculas circulares de RNA en Entamoeba, los cuales se conoce que tienen importantes funciones en metazoos. Gupta y cols. (2012) descubrieron un espaciador transcrito externo al 5' de genes de rRNA o etsRNAs (*external transcribed spacer RNAs*) conteniendo sitios importantes para el procesamiento del rRNA y que tienen la particularidad de circularizarse *in vivo* en respuesta a estrés y pueden autocircularizarse espontáneamente *in vitro*. En nuestro laboratorio (Mendoza-Figueroa y cols., 2018) se descubrió un nuevo tipo de molécula circular de RNA, los flicRNAs (*full-length intronic RNAs*) son moléculas de RNA circulares intrónicas de longitud completa,

algunas de estas provenientes de loci de genes de virulencia. Sin embargo, hasta el momento en *E. histolytica* no hay información reportada de circRNAs de tipo exónico.

3. Justificación

Para identificar, cuantificar la expresión de circRNAs y evaluar el compendio de circRNAs a partir de un conjunto de datos de secuenciación masiva de transcriptomas, se han desarrollado varios *pipelines* bioinformáticos (Gao y Zhao, 2018). En la actualidad con la llegada de la secuenciación masiva está disponible una gran cantidad de datos de RNA-seq con libre acceso. Hay varios datos de secuenciación de Entamoeba disponibles en reservorios como EBI y NCBI-SRA. Entre esas genotecas algunas fueron diseñadas con el fin de comparar perfiles de ncRNAs entre diferentes especies de Entamoeba (H. Zhang y cols., 2015) y otras con el fin de comparar los perfiles de expresión y detección de formas alternativas de *splicing* de transcritos codificantes de cepas virulentas y no virulentas en *steady-state* (Hon y cols., 2013). Sin embargo, en *E. histolytica* hasta el momento no hay moléculas de circRNAs exónicos reportadas por lo tanto seríamos los primeros en abordar la caracterización y ensamblado de circRNAs exónicos utilizando datos de NGS, también los primeros en analizar la expresión diferencial de los perfiles de las cepas virulentas y no virulentas en *steady-state* además de información acerca de los genes parentales de estos circRNAs.

4. Hipótesis

Usando genotecas no óptimas podremos detectar circRNAs exónicos en *E. histolytica* y obtener diferentes perfiles de expresión de circRNAs entre cepas amebianas.

5. Objetivos

5.1 Objetivo general

Caracterizar, cuantificar y ensamblar *in silico* circRNAs en *Entamoeba histolytica*, HM-1:IMSS y Rahman utilizando datos de NGS.

5.2 Objetivos específicos

- Caracterizar circRNAs de *E. histolytica* HM-1:IMSS y Rahman.
- Detectar isoformas de circRNAs de *E. histolytica* HM-1:IMSS y Rahman.
- Ensamblar circRNAs en *E. histolytica* HM-1:IMSS y Rahman.
- Realizar la expresión diferencial de circRNAs en *E. histolytica* HM-1:IMSS y Rahman.
- Realizar el análisis STRING de los genes parentales de los circRNAs detectados.
- Realizar la RT-PCR circular divergente dirigido a circRNAs de *E. histolytica* H1:IMSS.

6. Estrategia experimental

Con el fin de cumplir el objetivo general se utilizó una base de datos de transcriptomas de *E. histolytica* de cepas virulentas (HM-1:IMSS) y no virulentas (Rahman) descargadas en reservorios (**SRA** y **EBI**) publicados por Hon y cols, 2013, y un transcriptoma de *E. invadens* donado amablemente por la Dra. Rosaura Hernández. Las herramientas y *pipelines* utilizados fueron:

BWA es una herramienta de mapeado de lecturas en un genoma, la cual genera un archivo SAM que contiene información de las coordenadas en el genoma del alineamiento de las lecturas que contienen *split-alignments* (lecturas con una sección mapeada alineada) que son candidatos a circRNA. Esta información es usada posteriormente por CIRI2.

CIRI2 se utilizó con el fin de identificar lecturas con uniones de BSJs, es decir, los candidatos a circRNAs. También arroja información de la cuantificación relativa de circRNAs utilizando la información del archivo SAM generado por BWA.

CIRI AS (*Alternative Splicing events*), fue utilizado para el ensamblado de lecturas para la construcción interna de los circRNAs e identificación de isoformas de circRNAs que contengan el mismo BSJ.

CIRI FULL, fue utilizado para la integración de los resultados obtenidos de CIRI y CIRI AS, además de la integración de la característica RO (*Reverse Overlap*), una forma alterna para el ensamblado y detección de circRNAs. Este *pipeline* también permite la cuantificación de diferentes isoformas de circRNAs que contengan el mismo BSJ. Los datos arrojados por CIRI FULL contienen la información de los circRNAs caracterizados, cuantificados y ensamblados.

Se prosiguió con el análisis de esta información para darle un sentido biológico:

- Se hizo un *script* (en **R**) utilizando la paquetería de **DEseq2** para obtener perfiles de expresión diferencial entre cepas virulentas y no virulentas de *E. histolytica*.
- Se utilizó **STRING** para hacer un análisis de las redes de interacción de proteínas de los genes parentales de los circRNAs en *E. histolytica* y *E. invadens*. Así como el análisis del enriquecimiento funcional.

Se realizó la cuantificación de la expresión de los transcritos lineales con el fin de comparar su expresión con los circRNAs. Para ello se utilizaron los siguientes *pipelines*:

- **STAR** es una herramienta de mapeado de lecturas. Genera un archivo BAM que contiene las coordenadas de las lecturas mapeadas en el genoma.
- **HTSeq** utiliza el archivo BAM generado por STAR para hacer un conteo de las lecturas tipo *paired-end* que se mapean en las zonas codificantes de cada gen.

Finalmente se realizó **RT-PCR divergente** para la validación del circRNA del exón 2 del gen parental EHI_169670.

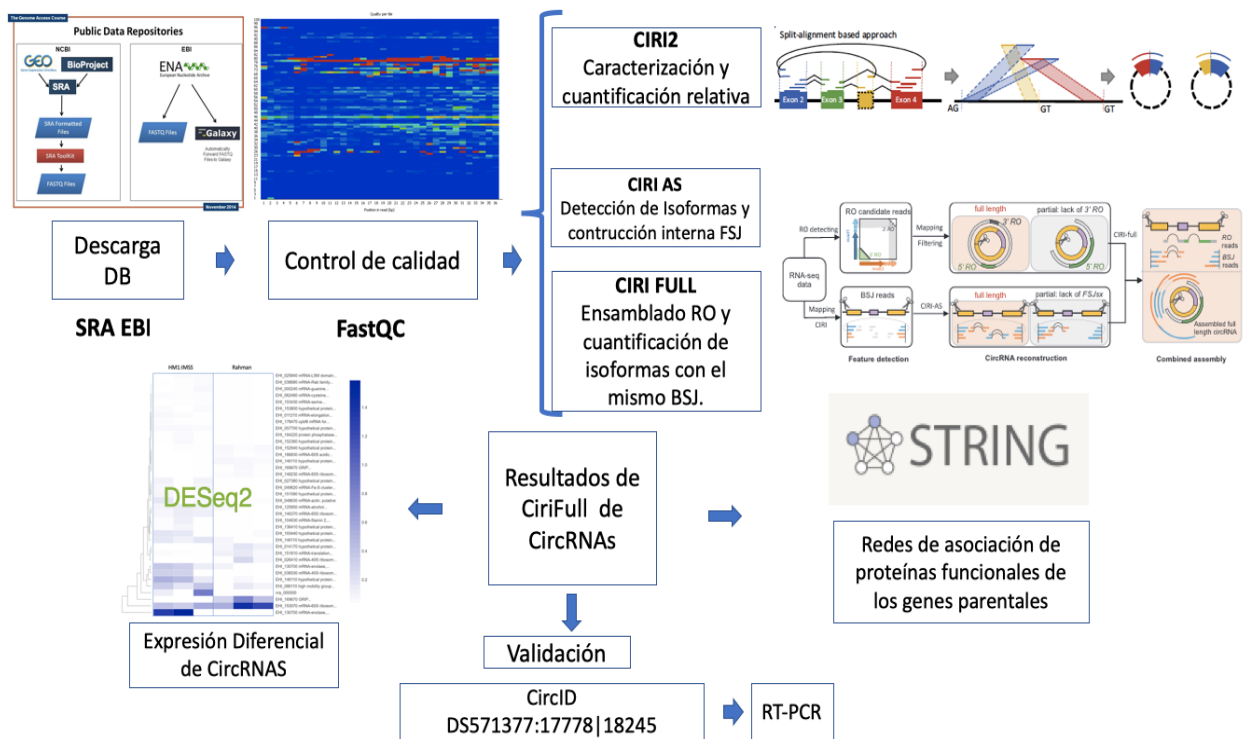


Figura 11. Estrategia experimental.

7. Materiales y métodos

7.1 Hardware

Se utilizó un *cluster* con 20 núcleos con una computadora central, sistema operativo macOS High Sierra, con memoria de 32 GB, procesador 2.7 GHz 12-Core Intel Xeon E5 Mac Pro.

7.2 Descarga de bases de datos

Se descargó la base de datos de secuenciación publicados por Hon y cols. (2013) de las cepas virulentas y no virulentas de *E. histolytica*. La referencia menciona la liga de la EBI (European Bioinformatics Institute) para su descarga. Misma que tenía las siguientes características:

- Para la secuenciación de alto rendimiento (*high-throughput sequencing*) se diseñaron genotecas de cDNA para lecturas tipo *paired-end*. Estas genotecas fueron preparadas a partir de poly(A)⁺ mRNA de acuerdo con las instrucciones del fabricante (mRNA-Seq 8-Sample Prep Kit, Illumina).
- Los fragmentos de cDNA de 200 bp fueron purificados de cada genoteca y 100 bp fueron secuenciados por ambos extremos utilizando un instrumento Illumina HiSeq2000 según las instrucciones del fabricante (Illumina).

Por muestra o corrida (*Run*) se obtuvieron dos archivos FASTQ, uno por cada extremo pareado (*paired-end*). Con extensión “.fq”.

Tabla 1. Información general de la base de datos. Contiene información de las genotecas secuenciadas. El número de *Spots* hace mención del total de las lecturas tipo *paired-end* secuenciadas.

Run	# of Spots	# of Bases	Size	Nombre de librería	SRA Experiment
ERR058005	113,997,619	22.8G	16.5Gb	lib2A_HM1_Rep1	ERX035851
ERR058006	91,291,028	18.3G	12.4Gb	lib2A_HM1_Rep2	ERX035852
ERR058007	81,271,700	16.3G	10.8Gb	lib2A_HM1_Rep3	n/a
ERR058008	70,821,013	14.2G	9.4Gb	lib2F_Rahman_Rep1	n/a
ERR058009	68,278,102	13.7G	9Gb	lib2F_Rahman_Rep2	n/a
ERR0580010	108,168,110	21.6G	15.3Gb	lib2F_Rahman_Rep3	n/a

Se descargaron las corridas de los transcriptomas ERR058007, ERR058008, ERR058009, ERR058010 utilizando la interfaz web del EBI obteniendo dos archivos tipo FASTQ una de las lecturas sentido y la otra de las antisentido por Run.

Los archivos FASTQ de dos corridas no pudieron descargarse utilizando la interfaz web de EBI debido al tamaño de los archivos y la calidad del internet. Su descarga resultaba corrupta ydesequilibrada. No podían ser utilizados en los ensayos *in silico* debido a que los archivos FASTQ estaban desequilibrados, no tenían el mismo tamaño. Los archivos ERR058005 y ERR058006 se descargaron utilizando un *script* de la colección de herramientas SRA Toolkit del NCBI. El *script* de SRA Toolkit se puede descargar del sitio <https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/>. Se descarga el SRA Toolkit dependiendo el sistema operativo a trabajar, en este caso MacOS. El archivo SRA se guardó en una carpeta llamada “nombre_de_la_carpeta”, para su posterior conversión a FASTQ. Se utilizaron los siguientes comandos para la descarga del archivo SRA para su posterior conversión a un archivo FASTQ:

1. ~/nombre_de_la_carpeta/bin/prefetch ERX035851
2. ~/nombre_de_la_carpeta/bin/fastq-dump --split-files ERX035851.sra

La conversión del archivo SRA a FASTQ se ejecuta en la terminal en la misma ubicación de “nombre_de_la_carpeta/bin/”.

Se generan dos archivos FASTQ llamados ERX035851_1.fq (sentido) ERX035851_2.fq (antisentido).

Lo mismo se hizo con la corrida ERR058006 con un SRA “ERX035852”.

Los transcriptomas donados de *E. invades* tienen las siguientes características: fueron 2 muestras de RNA total extraídas con Reactivo TRIzol (Invitrogen, No. de catálogo 15596-

018) de las cuales se generaron 2 replicas metodológicas por muestra de RNA total, generando un total de 4 genotecas de secuenciación. Los fragmentos cDNA se sintetizaron de aproximadamente 550 pb. Se secuenciaron para lecturas de tamaño de 2x150 bp de tipo *paired-end* aproximadamente generando 10 Millones de lecturas generando un total de 8 archivos FASTQ en el transcriptoma de *E. invadens*.

7.3 Bioconda

Bioconda es un canal para el administrador de paquetes Anaconda que se especializa en software de bioinformática con el fin de instalarlos de una manera simple localizado en <http://www.ddocent.com//bioconda/>. La instalación de Bioconda depende del sistema operativo; en la misma página se puede consultar la documentación de los softwares bioinformáticos y que comandos utilizar para su instalación.

Bioconda se usó para instalar los siguientes programas:

- BWA-MEM, FastQC, STAR y HTSeq

7.4 Control de Calidad de datos de secuenciación masiva

Se ejecutó FastQC dentro de la terminal con el siguiente comando ubicándose en la misma carpeta:

- `fastqc archivo_ejemplo.fq`

Se genera un conjunto de archivos que contienen información de control de calidad de las secuencias.

7.5 CIRI-full pipelines

CIRI-FULL se descarga en esta liga <https://sourceforge.net/projects/ciri-full/>. Al ejecutar los comandos de CIRI desde la terminal, toda ejecución de comando debe realizarse en la misma carpeta, los archivos del genoma de referencia y su respectiva anotación deben estar localizados en la misma carpeta al igual que los *scripts* a utilizar. Todo comando se ejecutó con las mismas especificaciones por muestra:

Paso 1: Entrar al directorio donde se descargó CIRI-full (este contiene los siguientes *scripts* CIRI2.pl CIRI_AS_v1.2.pl CIRI-full.jar y CIRI-vis.jar).

Manualmente con la interfaz gráfica se colocó en una carpeta con los siguientes archivos y *scripts*:

a) *Scripts*:

CIRI2.pl

CIRI_AS_v1.2.pl

CIRI-full.jar

CIRI-vis.jar

b) Archivos:

1) “Entamoeba_histolytica.JCVI-ESG2-1.0.dna_sm.toplevel.fa.gz” (genoma de *E. histolytica*). Después de descomprimirlo se le cambió a un nombre menos complicado: “E.fa”. El archivo del genoma de referencia de *E. invadens* se cambió de “Entamoeba_invadens_ip1_gca_000330505.EIA2_v2.dna.toplevel.fa.gz” a “Ei.fa”

2) “Entamoeba_histolytica.JCVI-ESG2-1.0.44.gtf.gz” (archivo de anotación) contiene las coordenadas de los exones en el genoma de referencia de *E. histolytica*, después de descomprimirlo se nombró “E.gtf” y el archivo de anotación de *E.*

invadens cambio de
“Entamoeba_invadens_ip1_gca_000330505.EIA2_v2.43.gtf.gz a “Ei.gtf”.

Ambos archivos son descargados desde Ensembl protist:
<https://protists.ensembl.org/info/website/ftp/index.html>

Paso 2: Hacer un índice utilizando BWA-mem usando los siguientes comandos:

```
bwa index E.fa
```

```
bwa mem -T 19 E.fa ERR058010_1.fq ERR058010_2.fq > E.sam
```

Con este comando se genera un archivo SAM llamado E.sam

Nota: debido a que el genoma de *Entamoeba* no es grande comparándolo con humano no hubo la necesidad de utilizar la opción “-a bwtsv” del programa BWA-mem. El argumento -T 19 filtra alineamientos con un score < 19, los archivos ERR058010_1.fq y ERR058010_2.fq son los archivos de secuenciación (FASTQ) son lecturas *paired-ended*.

Paso 3: Se corrió CIRI2 para la detección de circRNAs a partir del archivo SAM usándose el comando siguiente:

```
perl CIRI2.pl -I E.sam -O E.ciri -F E.fa -A E.gtf -0 -T 20
```

Se utilizó el *clúster* de la ENCB del IPN para correr todos los *pipelines* de CIRI. El *clúster* tiene un procesador de 20 núcleos para aumentar la velocidad del proceso, se utilizó el argumento “-T 20”, es decir se usaron todos los núcleos del *clúster* .

El argumento “-0” (*no strategy*) se utilizó para obtener todos los circRNAs que sólo tuvieran al menos una lectura de BSJ que sean independientes de las señales de *splicing*.

Se generó un archivo E.ciri con la información de las coordenadas de los límites del circRNA (el inicio y fin del circRNA en el genoma), o dicho de otra forma las coordenadas de localización de la unión del BSJ en el genoma; también proporciona información de la expresión relativa de los circRNAs o número de lecturas de BSJ.

Los datos de secuenciación de *E. invades* sólo se corrieron en CIRI2 debido a una falla computacional desconocida, sin embargo, aun así, se caracterizaron circRNAs.

Paso 4: Se corrió CIRI-AS con el siguiente comando:

```
perl CIRI_AS_v1.2.pl -S E.sam -C E.ciri -F E.fa -A E.gtf -O E_AS -D yes
```

De los archivos generados el importante para su posterior análisis es “E_AS_jav.list” con información de isoformas alternativas con el mismo BSJ y ensamblado de la estructura interna de circRNAs exónicos.

Paso 5: Módulo RO1. En este módulo se detecta y fusionan lecturas 5' RO a partir de los dos archivos de FASTQ con extensión “.fq”.

```
java -jar CIRI-full.jar RO1 -1 ERR058010_1.fq -2 ERR058010_2.fq -o EHI
```

Se genera el siguiente archivo (en formato FASTQ) EHI_ro1.fq que contiene las lecturas fusionadas con características 5' RO.

Paso 6: Se corrió BWA-mem con el fin de mapear las lecturas candidatas con características 5' RO contra el genoma de referencia.

```
bwa mem -T 19 E.fa EHI_ro1.fq > EHI_RO1.sam
```

Paso 7: Módulo RO2. Se corrió con el fin de identificar lecturas fusionadas que contengan 3' RO. El comando que se ejecutó fue el siguiente:

```
java -jar CIRI-full.jar RO2 -r E.fa -s EHI_RO1.sam -l 100 -o EHIdRO2
```

el argumento “-l” es la longitud de las lecturas de secuenciación tienen que ser de un mismo tamaño, en nuestro caso de 100 bp.

Se genera un archivo con la extensión “_ro2_info.list” que contiene información detallada de la lista de RO que pasaron el filtrado y su localización en el genoma de referencia.

Este archivo que se genera es el siguiente: EHIdRO2_ro2_info.list

Paso 8: El módulo Merge combina los resultados del CIRI2, el módulo RO2 y CIRI-AS para reconstruir los elementos internos de los circRNAs además de cuantificar circRNAs que contengan un mismo BSJ.

El módulo Merge se ejecuta desde una línea de comando de la siguiente manera:

```
java -jar CIRI-full.jar Merge -c E.ciri -as E_AS_jav.list -ro EHIdRO2_ro2_info.list -a E.gtf -r E.fa -o EHI_MERGE
```

El módulo Merge generará el siguiente archivo: prefix_merge_circRNA_detail.anno

Este archivo arroja la cuantificación relativa de circRNAs con diferentes BSJs e información de sus elementos internos. Además de la cuantificación e información de los elementos interno de isoformas circRNAs con la misma BSJ e información de sus elementos internos.

El archivo prefix_merge_circRNA_detail.anno es muy difícil de interpretar es por eso que se utiliza en el siguiente paso.

Paso 9: CIRI-vis. Es una herramienta para visualizar las lecturas BSJ y RO fusionados y la abundancia relativa de las isoformas con el mismo BSJ, de acuerdo a los archivos generados por CIRI-full prefix_merge_circRNA_detail.anno; el comando que se utilizó fue el siguiente:

```
java -jar CIRI-vis.jar -i EHI_MERGE_merge_circRNA_detail.anno -l  
E_AS_library_length.list -r E.fa
```

Se generaron varios archivos:

- 1) "stout.list" muestra información de todos los circRNAs incluyendo las isoformas con el mismo BSJ como:

Columna 1: nombre del archivo del pdf

Columna 2: ID de la posición del BSJ del circRNA y/o isoforma del circRNA en la forma "chr:star|end"

Columna 3: cromosoma o contig del circRNA o isoforma del circRNA predicho

Columna 4: inicio del locus del circRNA y/o isoforma del circRNA predicho en el cromosoma o contig

Columna 5: final del locus del circRNA y/o isoforma del circRNA predicho en el cromosoma o contig.

Columna 6: conteo de lecturas de unión circular de un circRNA predicho (también llamado como lectura de BSJ)

Columna 7: número de isoformas en el circRNA con el mismo BSJ.

Columna 8: estimado de cuentas de BSJ de las isoformas predichas

Columna 9: la longitud del circRNA

Columna 10: menciona si está totalmente construido el circRNA

Columna 11: sentido

Columna 12: posición del circexon en la isoforma predicha; "0-0" representa un hueco durante la reconstrucción

- 2) Un conjunto de archivos pdf que muestra las isoformas de los circRNAs con el mismo BSJ y muestra información precisa de cada isoforma

- 3) “.fa” un archivo fasta con la secuencia completa de los circRNAs

7.6 Transcritos lineales

Paso 1: Creando un índice con STAR

Se realizó un mapeo de las lecturas de la base de datos del transcriptoma de Hon y cols, (2013) con el fin de cuantificar y para posteriormente hacer un análisis de los perfiles de expresión de los circRNAs y sus contrapartes lineales (la expresión de los genes parentales). En el directorio o carpeta con el nombre “directorio_de_trabajo” es donde se corre el programa STAR.

Utilizando la terminal y ubicando el directorio de trabajo se ejecutó el siguiente comando:

```
ejemplo@usuariolinux:~/directorio_de_trabajo$ mkdir genomeDir
```

se creó una segunda carpeta el cual contenga el genoma:

Posteriormente desde la interfaz gráfica (no por comandos) se pasó el archivo E.fa (el genoma de *E. histolytica*) a la carpeta “genomefasta”. Se utilizó el comando mkdir con el fin de tener la autorización como super-usuario para que STAR pueda trabajar sobre esa carpeta llamada genomeDir

- 1) Generación de índices:

```
STAR --runThreadN 4 --runMode genomeGenerate --genomeDir genomeDir --  
genomeFastaFiles genomefasta/E.fasta
```

Paso 2: Mapeo de las lecturas con STAR

Se creó una carpeta llamada “fastq_lecturas_pareadas” para contener los archivos FASTQ a utilizar (estos deben estar con extensión “.fq”). En este ejemplo son ERR058010_1.fq y ERR058010_2.fq. Se utilizó el siguiente comando:

```
STAR --genomeDir genomeDir --readFilesIn fastq_lecturas_pareadas/
ERR058010_1.fq fastq_lecturas_pareadas/ERR058010_2.fq --runThreadN 4 --
outSAMtype BAM SortedByCoordinate
```

El archivo BAM que se genera contiene la información para ser utilizado posteriormente el conteo de lecturas sobre transcritos es “Aligned.sortedByCoord.out.bam”.

Paso 3: Conteo de lecturas de secuenciación utilizando HTSeq

En la carpeta de trabajo se ubica el archivo “Aligned.sortedByCoord.out.bam” y se corre el siguiente comando:

```
htseq-count -f bam Aligned.sortedByCoord.out.bam E.gtf > E.txt
```

El archivo E.txt es un archivo que contiene el conteo de las lecturas que se alinean o mapean en el exón; la información de los exones de los genes es proporcionada por el archivo E.gtf

7.7 Cuantificación de circRNAs y transcritos lineales

Los niveles de expresión de los circRNAs y RNAs lineales fueron normalizados basados en el método de transcritos por millón (TPM, del inglés *Transcripts Per Million*) usando la siguiente formula:

expresión normalizada = (Lecturas Mapeadas) / Total de Lecturas) x 1,000,000.

7.8 Expresión diferencial

Después de la normalización de los conteos de las lecturas de unión circular de los circRNAs o BSJ *reads* se realizó un *script* en lenguaje R utilizando el paquete de DESeq2 para la expresión diferencial. El *script* en R está descrito en apéndice.

Los criterios estadísticos para determinar la expresión diferencial de los circRNAs fueron los siguientes: un *fold change* > 1.0, p-value con un umbral de 0.05, y un p-value ajustado (padj) con un umbral de 0.1

7.9 Análisis de las redes de asociación de proteínas

Se utilizó STRING para hacer un análisis de las redes de asociación de proteínas de los genes parentales de los circRNAs en *E. histolytica* y en *E. invadens*. Para *E. invadens* se procedió primero a identificar sus ortólogos en *E. histolytica* (la lista de ortólogos aparece en los apéndices) debido a que STRING no reconoce los identificadores EIN (*genes de E. invadens*).

7.10 Análisis estadísticos

Se utilizó R para la determinación de los análisis estadísticos: a partir de la paquetería de *DESeq2* se determinó el *p-value* con el método de *Wald* y con el mismo se ajustó con el método de Benjamini y Hochberg para obtener el valor de padj. También *DESeq2* se utilizó para el análisis de los componentes principales de las genotecas virulentas y no virulentas con el fin de medir la variación entre muestras. Para la determinación del coeficiente de correlación de Spearman de los promedios de TPMs de los circRNAs y de los transcritos lineales de sus genes parentales.

7.11 Cultivo de trofozoíto de *Entamoeba histolytica*

Los trofozoítos de la cepa HM1:IMSS de *E. histolytica* se crecieron de forma axénica en medio TYI-S-33 (*Trypticase-yeast extract-iron serum*; Diamond y cols., 1995) suplementado con Suero Bovino Adulto inactivado (Microlab Laboratorios, No. de catálogo SU140) al 10 % y Penicilina-Estreptomicina (Gibco, No. de catálogo 15140-148) al 1x en tubos de vidrio de cultivo de 10 mL. Los cultivos se incubaron a 37°C hasta que

alcanzaran la fase exponencial de crecimiento (80 % de confluencia), se contó el número de células usando la cámara de Neubauer). La suspensión celular se centrifugó a 1000 rpm durante 5 min a 4°C, se eliminó el medio de cultivo y los trofozoítos fueron utilizados inmediatamente para extraer RNA.

7.12 Extracción de RNA

Se extrajo RNA total utilizando reactivo TRIzol (Invitrogen, No. De catálogo 15596-018). Para la lisis se aplicaron 0.75 mL del reactivo Trizol^{RM} por 0.25 mL de muestra de la pastilla (aproximadamente $0.5-1 \times 10^7$ células) y se homogenizó para obtener un lisado. Posteriormente se incubó por 5 minutos para permitir una completa disociación de los complejos nucleoprotéicos. Se adicionó 0.2 mL por cada 1 mL de reactivo Trizol usado en la lisis. El tubo se incubó en hielo de 2 – 3 minutos, se centrifugó la muestra por 15 minutos a 12,000x g a 4 °C. Las mezclas se separaron en fases, para la extracción de RNA se obtuvo con precaución sólo la fase acuosa transparente, se transfirió la fase acuosa a un nuevo tubo. Posteriormente se agregó 5 µg de glicógeno libre de RNAasas, para coprecipitar el RNA. Se adicionó 0.5 mL de isopropanol al 100% a la fase acuosa por 1 mL de reactivo Trizol utilizado en la lisis. Se incubó por 10 minutos, se centrifugó por 10 minutos a 12,000 x g 4 °C, el RNA precipitado formó un pellet de color blanco al fondo del tubo, se descartó el supernadante con una micropipeta. Se resuspendió en 1 mL al 75% de etanol por 1 mL de reactivo Trizol utilizado en la lisis, se homogenizó brevemente en un vortex, se descartó el sobrenadante con una micropipeta y se dejó secar la pastilla por 10 minutos. La pastilla se resuspendió en 30 µL de agua libre de nucleasas, se cuantificó espectrofotométricamente en un NanoDrop a una absorbancia de 260/280.

7.13 RT-PCR

Las reacciones de retro-transcripción se llevaron a cabo utilizando el kit RevertAid First Strand cDNA Synthesis #1622 siguiendo el protocolo recomendado por el fabricante. Se empleó una concentración de RNA de 2.5 µg, como molde se adicionó 1 µL de hexámeros aleatorios y se incubó a 65 °C durante 5 minutos. Posteriormente, se agregaron 4 µL de

buffer de reacción 1 µL de dNTPs, 1 µL de RiboLock y 1 µl de M-MLV, completando un volumen final de 20 µL. Se agitó suavemente y se incubó la mezcla por 42 °C por una hora. La inactivación se realizó por 5 minutos a 70 °C. El cDNA obtenido se almacenó a -20 °C hasta su uso. En las PCR se empleó un volumen de 1µL de cDNA (≈ 10 ng de cDNA), se completó con cebadores específicos 10 µM, 1 µL dNTPs 10 mM, 2.5 buffer de PCR 10X (#cat: KB1004), 1.5 de MgCl₂ 25 mM (#cat:KB1001), y 0.15 µl de Taq KAPPA (#cat: KE1000) se completó con agua hasta un volumen de 25 µL con agua miliQ estéril. Para visualizar el producto de PCR, se analizó en un gel de agarosa al 2 % y se tiñeron con bromuro de etidio. Ver la tabla# para las condiciones de PCR.

Tabla 2. Oligonucleótidos y condiciones de amplificación.

circRNA	Oligonucleótido	Secuencia 5'-3'	MgCl ₂ (mM)	Condiciones de amplificación	Tamaño de amplicón
Exón 2 (EHI_169670)	EHI169670_E2as	CTTCTTTTTCTTTTT CTAATTCTTCACCC	3	35 ciclos de 94°C /45", 58 °C/45", 72 °C/45"	355
	EHI169670_E2s	AAGAAAGTTAATGAT TCTGAGAAAGAG			

8. Resultados

8.1 Selección de genotecas

Se realizó una búsqueda en varias bases de datos de diferentes grupos que investigan la biología de *Entamoeba* (Tabla 2). Entre ellas, la genoteca de (H. Zhang y cols., 2015) reportó secuencias menores de 70 nucleótidos y no resultaron adecuadas en la búsqueda de circRNAs. La genoteca de Ehrenkaufner y cols. (2013) tampoco resulta apropiada debido al método de secuenciación utilizado que difiere de Illumina. Cuando usamos la base de datos reportada por Weber y cols., (2016) detectamos circRNAs pero en cantidades muy pobres, debido a la poca profundidad de secuenciación y a que contenía lecturas de secuenciación entre 40 y 70 bp (70 bp es el mínimo necesario en la búsqueda de circRNAs utilizando CIRI2). La genoteca de *E. invades* donada amablemente por la Dra. Rosaura Hernández Rivas se usó para correr CIRI2 y encontramos un compendio interesante de circRNAs, suficiente para realizar análisis de enriquecimiento funcional y análisis de interacciones de proteínas funcionales. Sin embargo, no determinamos los elementos internos de los circRNAs detectados en *E. invadens*, esto por algún error desconocido proveniente del transcriptoma en la utilización de CIRI-FULL. Por otro lado, la genoteca de Hon y cols. (2012), la genoteca principal de este proyecto se corrió utilizando todo el conjunto de *pipelines* de CIRI-FULL y obtuvimos la información de circRNAs para los análisis que son motivos de esta tesis.

Tabla 3. Bases de datos de secuenciación de *Entamoeba* reportados por diferentes grupos de investigación.

Referencia	Características de la genoteca	Aplicabilidad del pipeline CIRIFULL
Hon y cols. (2012)	A partir de Poly(A) ⁺ , lecturas tipo paired-end y una secuenciación profunda por ILLUMINA	OK a CIRI-FULL
Donado por la Dra. Rosaura Hernandez Rivas	A partir de RNA total, lecturas tipo paired-end, y secuenciación por ILLUMINA	OK CIRI2, no funcionó con CIRIFULL
Weber y cols. (2016)	A partir de Poly(A) ⁺ , lecturas tipo single-end secuenciación por ILLUMINA	OK CIRI2, no funcionó con CIRIFULL
Ehrenkaufner y cols. (2013)	A partir de RNA total y lecturas tipo single-end secuenciación por SOLID™ 4	NO
H. Zhang y cols. (2015)	RNA extraído de un gel con el objetivo de ver RNAs de tamaño de 27 nucleótidos secuenciación por ILLUMINA	NO

8.2 Caracterización de circRNAs en cepas virulentas y no virulentas de *E. histolytica*

De las 10^9 lecturas, aproximadamente el 5 % (54 mil lecturas), correspondieron a lecturas no alineadas (mapeadas) colinealmente. Y de esas 54 mil, alrededor del 0.05 %, es decir unas 2700 lecturas de BSJ correspondieron a circRNAs. Debido a que se quería realizar un compendio de circRNAs, se hizo un proceso de filtrado para la caracterización de los circRNAs identificados. El conjunto de *pipelines* de CIRI-FULL computó un total de 958 circRNAs, de los cuales en 920 de ellos se logró la reconstrucción interna del circRNA. El programa descartó a los 38 no reconstruidos totalmente debido a no tener lecturas que completen el ensamble de los elementos internos del circRNA.

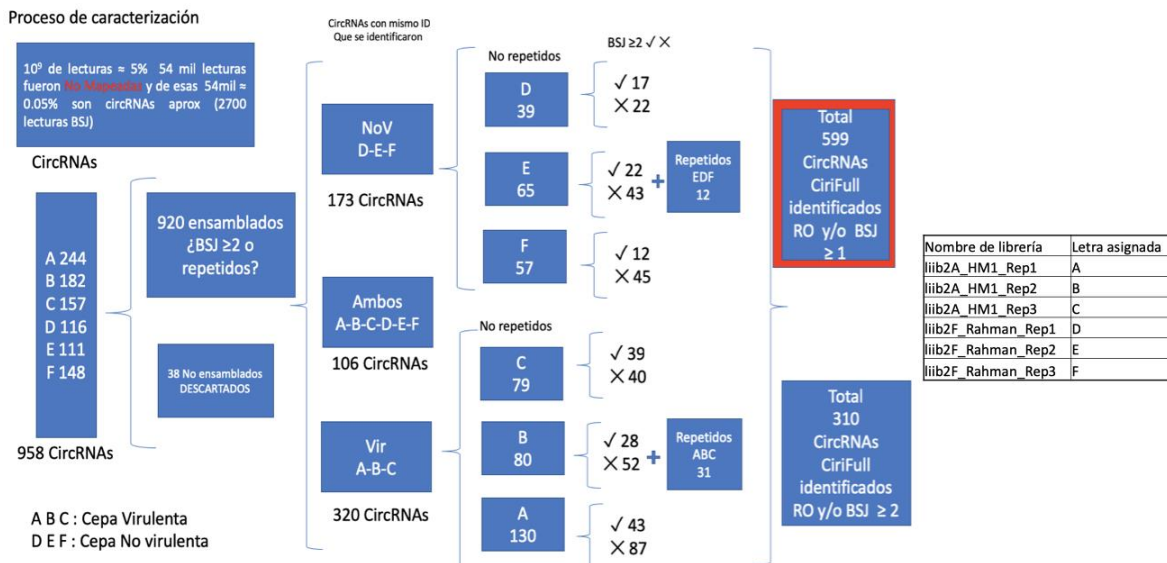


Figura 12. Compendio de circRNA en cepas virulentas y no virulentas de *E. histolytica*. Proceso de caracterización de los circRNAs de todos que se computaron. Todo con el fin de identificar cuantos circRNAs se identificaron en este trabajo. Además, un criterio que se tomo en cuenta fue el de un FDR de circRNAs con BSJ ≥ 1 . Así se obtuvo un total de 310 circRNAs con ese criterio, siendo estos circRNAs computacionalmente más probables de ser validados, pero en total se identificaron un total de 599 circRNAs con al menos una lectura de BSJ.

Posteriormente de los 920 círculos, sólo 173 aparecían en Rahman, 320 en HM1:IMSS y 106 en ambas cepas, indicando más presencia de circRNAs propios en la cepa

HM1:IMSS. En la cepa no virulenta hubo un total de 161 circRNAs sin duplicado entre las réplicas y un total de 12 circRNA con duplicado entre las réplicas. Dentro las genotecas virulentas hubo un total de 289 circRNAs que no se repetían entre las réplicas y hubo un total de 31 que se repetían entre las réplicas. La identificación de circRNAs potenciales fue restringida aún más: de los 599 circRNAs identificados con lecturas de BSJ ≥ 1 , se utilizó un criterio de selección exigiendo que los circRNAs identificados tuvieran lecturas de BSJ ≥ 2 , arrojando un total de 310 circRNAs en *E. histolytica* (Figura 12).

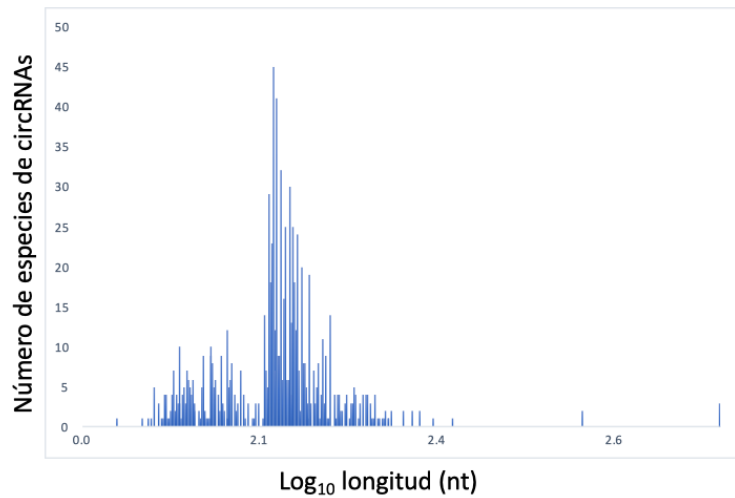


Figura 13. Histograma del número de circRNAs ensamblados *E. histolytica* en función de su longitud. (antilog₁₀ 2.15 = 141 bp).

La mayor frecuencia de la longitud interna de circRNAs computados fue de 141 bp (figura 13), de acuerdo con lo estimado con CIRI-FULL.

8.3 CircRNAs formados por diferentes elementos

Un elemento en un circRNA es una región del circRNA la cual tiene dos coordenadas con respecto al genoma que coinciden con el principio y el fin del elemento. Estos elementos pueden ser de varios tipos, ya sea elementos de tipo intergénico, elementos exónicos o elementos intrónicos. Se identificaron un total de 43 circRNAs que están constituidos por dos elementos (Apéndice A). Algunos de estos circRNA están formados por dos

elementos exónicos, otros por dos elementos exónicos pertenecientes a un mismo exón, otros contienen una combinación de un elemento exónico y con un elemento intergénico y/o intrónico. En el caso de *E. histolytica* no se encontraron circRNAs compuestos de exón-intrón, mismos que sólo se detectaron en *E. invadens*. La mayoría de los circRNAs identificados tienen dos elementos exónicos, tales elementos pertenecieron a un mismo exón. Seis circRNAs tuvieron una expresión diferencial que contenían dos elementos de diferentes exones ver (Apéndice A). No se encontraron circRNAs compuestos por tres o más elementos, el máximo fue de dos elementos. Se sabe que en organismos superiores puede haber más elementos o bien formado por más exones (Zhou y cols., 2018). Como ejemplo en la figura 14 observamos un mapa genómico con información de un circRNA expresado diferencialmente cuyo ID es DS571186:1621|1788:

- El ID es DS571186:1621|1788 el gene parental es EHI_000240.
- La unión del BSJ está determinado por el inicio del circRNA y el final, en el mismo ID se tiene esa información, es decir el círculo inicia en la posición 1621 y termina en la posición 1788, todo esto en el contig DS571186. Las posiciones adyacentes corresponden a las señales de *backsplicing*.
- Este circRNA está compuesto por dos elementos exónicos pertenecientes a diferentes exones en el mismo gen: el primer elemento tiene las coordenadas 1621-1657 y el segundo elemento tiene las coordenadas 1717-1788, que dan información de la composición interna del circRNA.
- Este circRNA es de tipo exónico debido a los elementos que contiene.

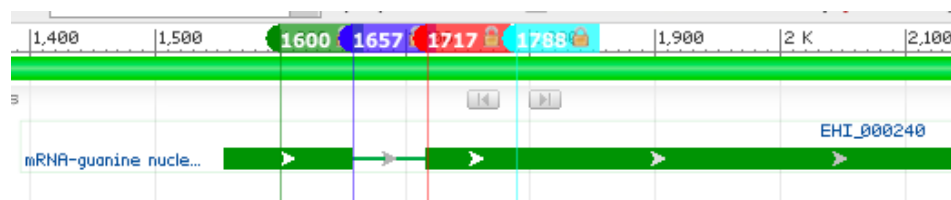


Figura 14. Coordenadas del circRNA. Cuyo ID es DS571186:1621|1788 con dos elementos exónicos, este circRNA es expresado diferencialmente en *E. histolytica*.

8.4 Relación entre la expresión de los transcritos circulares y lineales

En nuestro laboratorio observamos que los flicRNAs tienen una vida media mayor que sus contrapartes lineales (Mendoza-Figueroa y cols., 2018) resultando en una aparente abundancia de los flicRNAs. Nos preguntamos entonces si los circRNAs se expresan más que sus contrapartes lineales y para ello se abordó un enfoque estadístico. Primero se graficó el promedio de TPM de los circRNAs identificados (Promedio TPM_{circular}) con respecto al promedio de TPM de los genes parentales de los circRNAs identificados (Promedio TPM_{lineal}). Las ordenadas y las abscisas de la figura 15 están una escala logarítmica idéntica, y en la misma se observa que efectivamente que la abundancia de los circRNAs está en función de la abundancia de su mensajero parental.

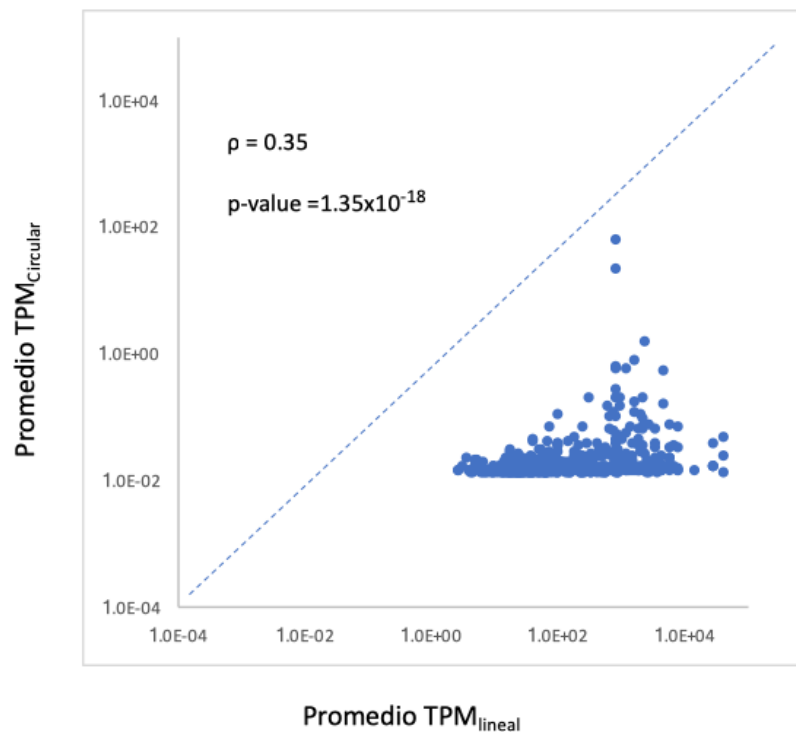


Figura 15. Correlación entre los promedios de los TPM lineales y TPM circulares en *E. histolytica*.

Relación entre la expresión de los circRNAs y sus contrapartes lineales. La correlación entre el promedio de los valores del TPM_{lineal} y el promedio de los TPM_{circ} en cepas de *E. histolytica*. Cada punto representa un gene parental. Los puntos que hubiesen aparecido arriba de la línea hubieran sido circRNAs expresados mayormente que sus contrapartes lineales. Sin embargo, no hay circRNAs más expresados que sus contrapartes lineales. El coeficiente de correlación. (ρ) y el P -value fueron calculados con la prueba de correlación de Spearman

A estos datos se les realizó el estadístico de correlación de Spearman. La correlación de Spearman entre dos variables es igual a la correlación de Pearson entre los valores de rango de esas dos variables; mientras que la correlación de Pearson evalúa las relaciones lineales, la correlación de Spearman (siendo la versión no paramétrica de Pearson, ideal para este tipo de datos biológicos) evalúa las relaciones monotónicas (ya sean lineales o no). Por lo tanto, observamos una correlación monotónica positiva, teniendo un incremento monotónico en la expresión de todos los circRNAs detectados y sus contrapartes en las cepas virulentas y no virulentas (un $\rho = 0.35$ y un $p\text{-value} = 1.35 \times 10^{-18}$) ver figura 15.

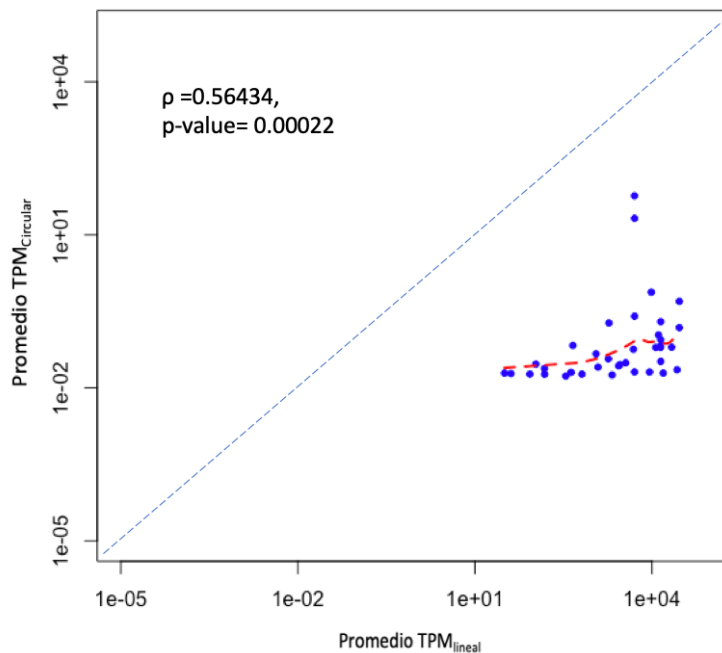


Figura 16. Correlación entre los promedios de los TPM circulares expresados diferencialmente y los promedios de los TPM lineales en *E. histolytica*. Relación entre la expresión de los circRNAs expresados diferencialmente y sus contrapartes lineales. La correlación entre el promedio de los valores del TPM_{lineal} y el promedio de los TPM_{circ} en cepas de *E. histolytica*. Cada punto representa un gene parental. Los puntos que hubiesen aparecido arriba de la línea hubieran sido circRNAs expresados mayormente que sus contrapartes lineales. Sin embargo, no hay circRNAs más expresados que sus contrapartes lineales. El coeficiente de correlación. (ρ) y el P -value fueron calculados con la prueba de correlación de Spearman. La línea punteada de color rojo representa la regresión local.

También se tenía la interrogante de si los circRNAs expresados diferencialmente presentan una correlación con sus contrapartes lineales. Efectivamente la expresión de los 39 circRNAs presentó una correlación positiva, observando un incremento monotónico en la expresión de los circRNAs y sus contrapartes lineales en todas las muestras de cepas virulentas y no virulentas (un $\rho = 0.56433$ y un $p\text{-value} = 0.00022$), ver figura 16.

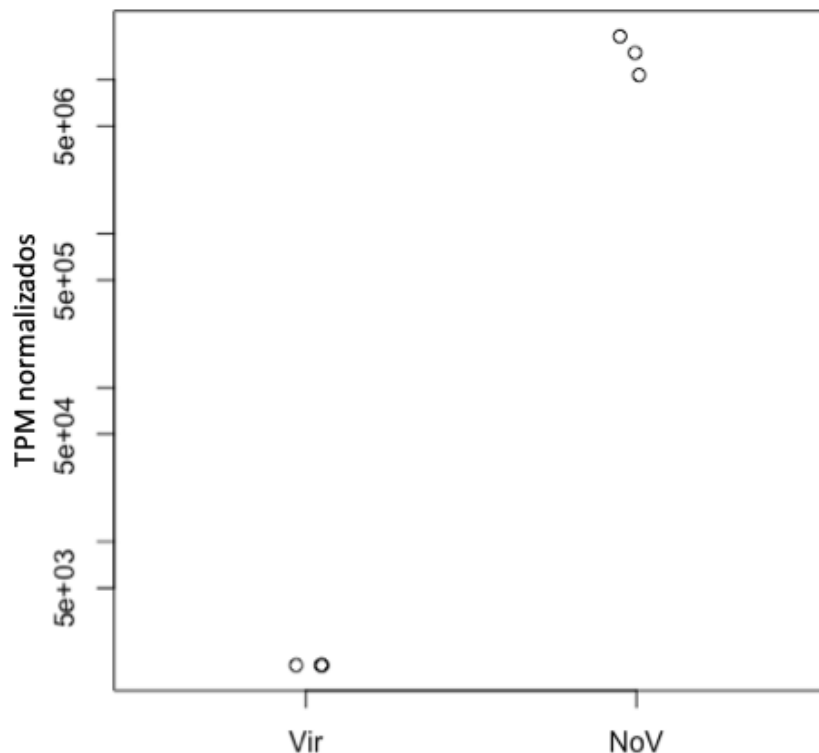


Figura 17. Comparación de TPM de circRNAs de cepas virulentas y no virulentas. Cada circulo en la figura representa la sumatoria de los TPM de las lecturas de BSJ pertenecientes a circRNAs por muestra amebiana (Vir, HM1:IMSS y NoV, Rahman)

Se realizó un análisis para comparar los TPM de circRNAs entre cepas amebianas (Figura 17) y encontramos que en la cepa avirulenta Rahman hay tres órdenes de magnitud más de TPM que en la virulenta HM1:IMSS. Estos hallazgos se confirmaron utilizando un análisis estadístico para comparar la fracción de circRNAs respecto a sus contrapartes lineales entre cepas amebianas (Figura 18) utilizando el siguiente cociente: $\text{TPM}_{\text{circ}} / (\text{TPM}_{\text{circ}} + \text{TPM}_{\text{lineal}})$. De esta manera encontramos que la cepa avirulenta

tiene más fracción de transcritos circulares que la cepa virulenta. Observamos que en el 25 % de las fracciones circulares por arriba del 75 % (tercer cuartil), se observa que hay más fracciones circulares en avirulencia que en virulencia.

Se puede observar que en avirulencia hay mayores fracciones circulares atípicas, debido a los circRNAs sobreexpresados mayormente en avirulencia que en virulencia. No se observa aumentos significativos en el 25 % de las fracciones circulares por debajo del 75 % (primer cuartil). Concluyendo que en cepas amebianas en *steady-state* hay mayor expresión de circRNAs exónicos, pero aun así no hay una mayor expresión de circRNAs con respecto a su correspondiente gen parental.

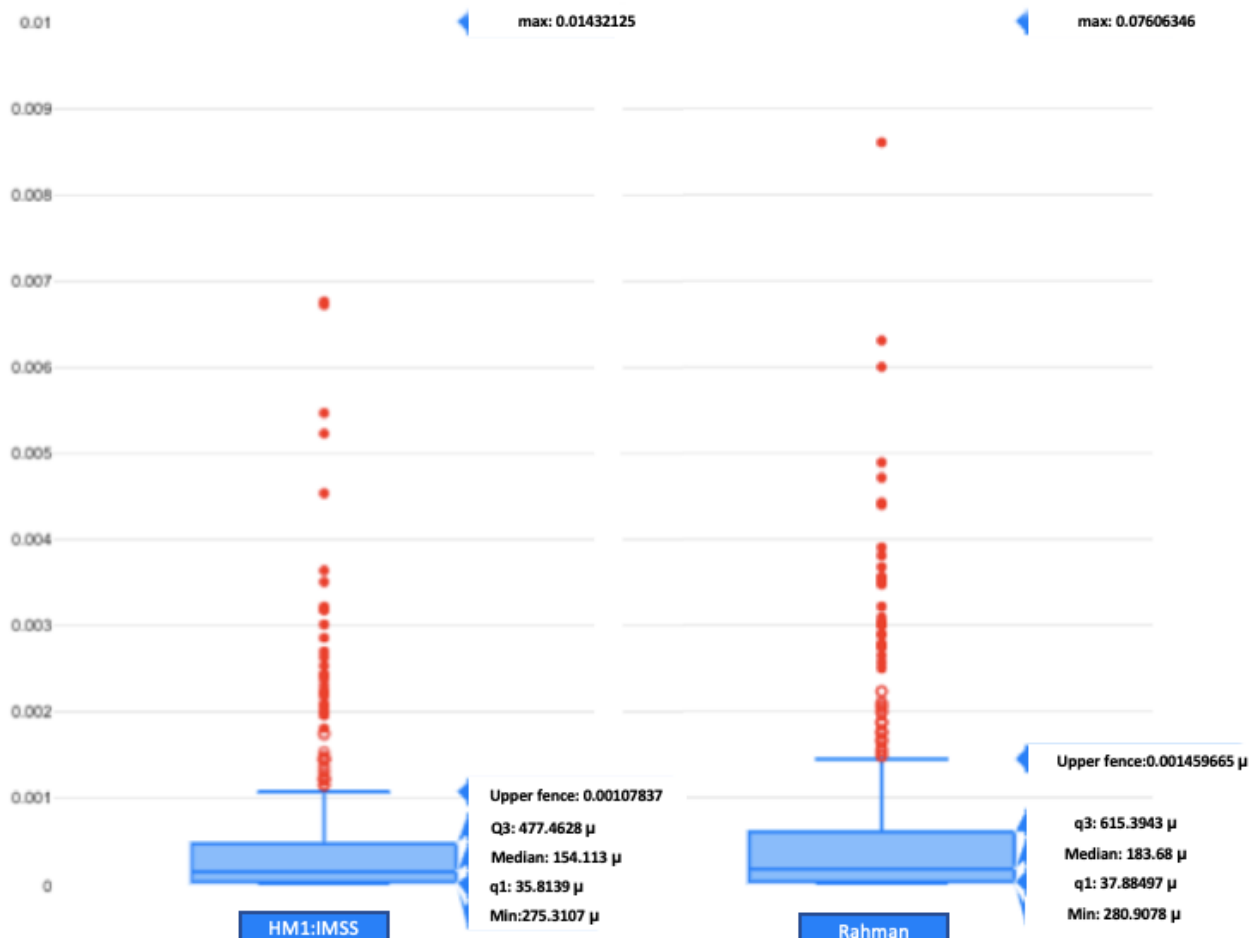


Figura 18. Comparación de fracciones de transcritos circulares categorizado por cepas virulentas y no virulentas. El gráfico *boxplot* se hizo con el promedio de la fracción de los transcritos circulares de cepas virulentas y no virulentas.

8.5 Expresión diferencial de circRNAs en cepas virulentas y no virulentas

A partir de la cuantificación normalizada de la expresión de los circRNAs se realizó un análisis de los componentes principales (PCA, del inglés **P**rincipal **C**omponent **A**nalisis) con el fin de visualizar la variación entre las muestras del análisis de expresión de los circRNAs. En el análisis PCA (Figura 19) encontramos que las muestras entre los triplicados de las cepas no virulentas tienden a formar un *cluster* definido. Por otro lado, observamos que las muestras provenientes de cepas virulentas no forman un *cluster* definido.

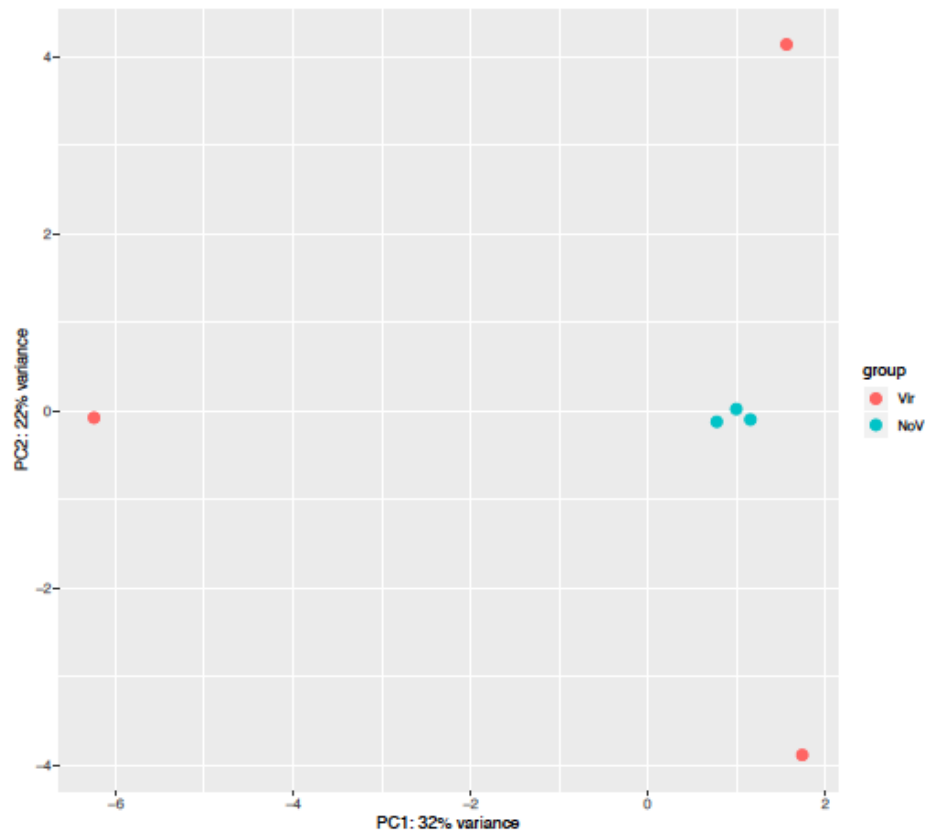


Figura 19. Análisis de los componentes principales de las genotecas de cepas virulentas y no virulentas. Se realizó un PCA de la expresión de los circRNAs. Cada punto representa una muestra amebiana en steady-state. (PC1, first principal component) primer componente principal y (PC2, second principal component) segundo componente principal. Vir, genotecas de cepas virulentas, NoV, genotecas de cepas no virulentas.

La expresión diferencial se realizó utilizando DESeq2 a partir de muestras normalizadas. Utilizando DESeq2 se determinó el p-value con el método de Wald y se ajustó con el método de Benjamini y Hochberg. Se encontró un total de 39 circRNAs los cuales estaban expresados diferencialmente (ver apéndice B y figura 20) el criterio para determinar la expresión diferencial de los circRNAs fueron los siguientes: un p-value = 0.05, un padj= 0.1 y un *fold change* mayor a 1.

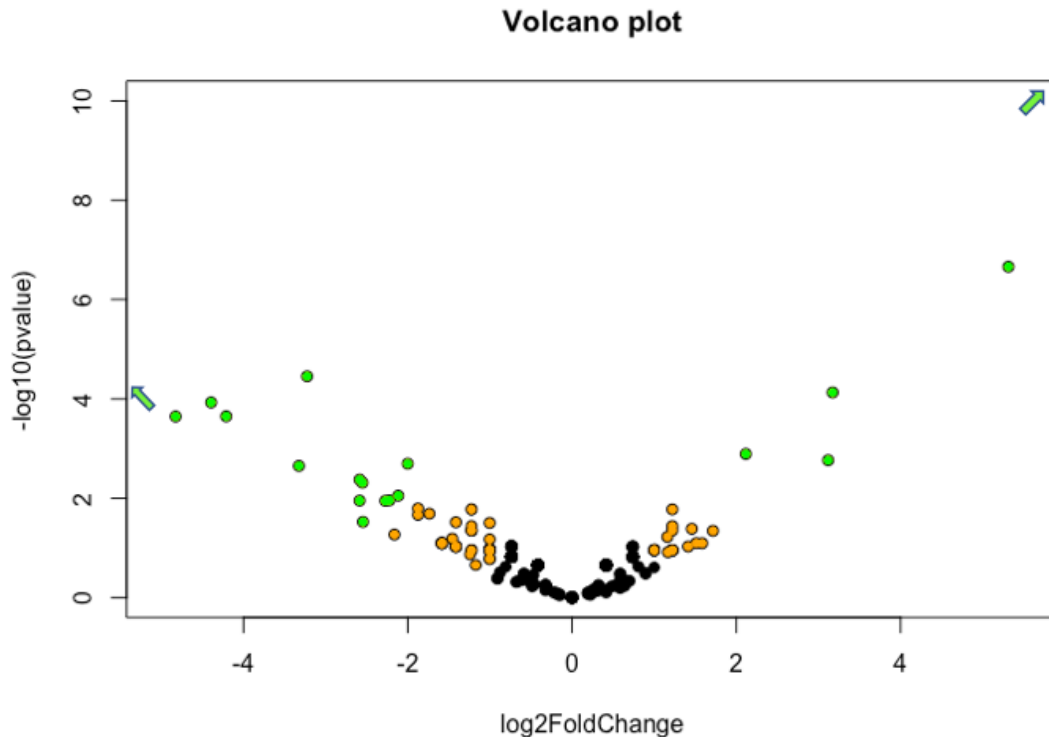


Figura 20. Gráfica de Volcán de la expresión diferencial de los circRNAs. Los puntos verdes y naranjas en la gráfica representan los circRNAs expresados diferencialmente que fueron estadísticamente significativos entre grupos. Los puntos verdes son los que tienen una mayor expresión diferencial tienen un p-value < 0.05, un padj < 0.1 y un *fold change* > 2.0, los naranjas representan los que están medianamente expresados diferencialmente tienen un p-value < 0.05, un padj < 0.1 y un *fold change* entre 1.0 y 2.0, los puntos negro representan estadísticamente que no hubo expresión diferencial. En apéndice C está la misma gráfica abarcando todos los puntos.

En la figura 21 se observa un heatmap de los 37 de los 39 circRNAs expresados diferencialmente, se utilizaron 37 circRNAs debido a que dos circRNAs contienen niveles

de expresión altos que no permiten visualizar la expresión diferencial en el gráfico de heatmap estos se expresan en Rahman.

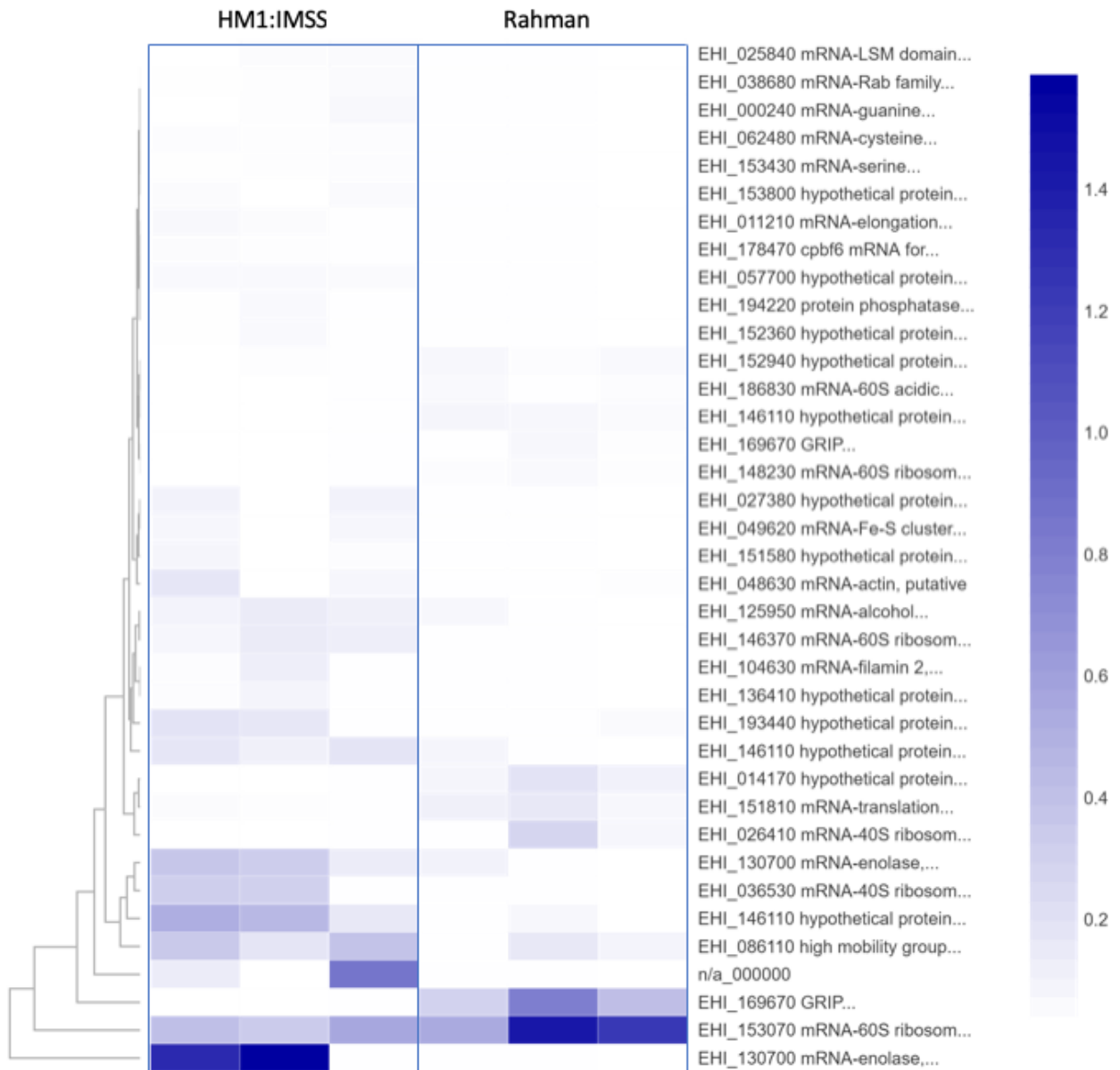


Figura 21. Heatmap representando los niveles de expresión normalizada por TPM de los circRNAs expresados diferencialmente. Con un *fold change* > 1, p-value < 0.05 de cepas virulentas y no virulentas. Los valores de la escala de color representan los niveles de expresión en TPM normalizados.

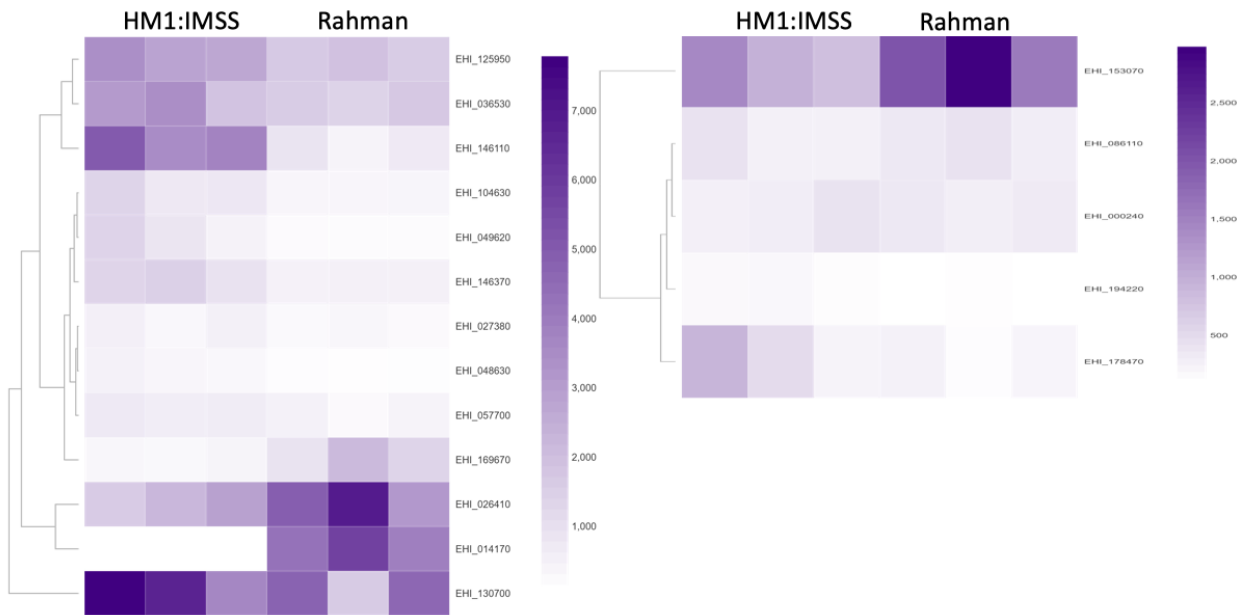


Figura 22. Heatmaps representando los niveles de expresión normalizada por TPM de los genes parentales de los circRNAs de cepas virulentas y no virulentas. Los valores de la escala de color representan los niveles de expresión en TPM normalizados. La gráfica se hizo en dos gráficas de heatmap con el fin de observar la expresión de los circRNAs con fines de visualización.

Si se comparan los niveles de expresión de los circRNAs expresados diferencialmente (figura 21) con sus contrapartes lineales (figura 22), se observa que mientras más expresión de circRNA, hay también más expresión de la contraparte lineal. Por ejemplo, en el circRNA cuyo gen parental está codificado en el locus EHI_146110 que se expresa más en virulencia, podemos observar que su contraparte lineal se expresa más en la cepa virulenta. Todos los circRNAs expresados diferencialmente tienen este comportamiento. Por lo tanto, podemos concluir que mientras haya más expresión de circRNAs expresados diferencialmente hay más expresión de sus contrapartes lineales.

8.6 Los circRNAs del locus EHI_169670

El gen EHI_169670 codifica una proteína hipotética que contiene un dominio GRIP que contiene la proteína RUD3. Este gen está compuesto por tres exones, y el segundo exón genera un total de 12 circRNAs, de las cuales tres isoformas tienen el mismo BSJ y las 9 restantes tienen diferentes BSJ. El circRNA más expresado comprende todo el exón 2.

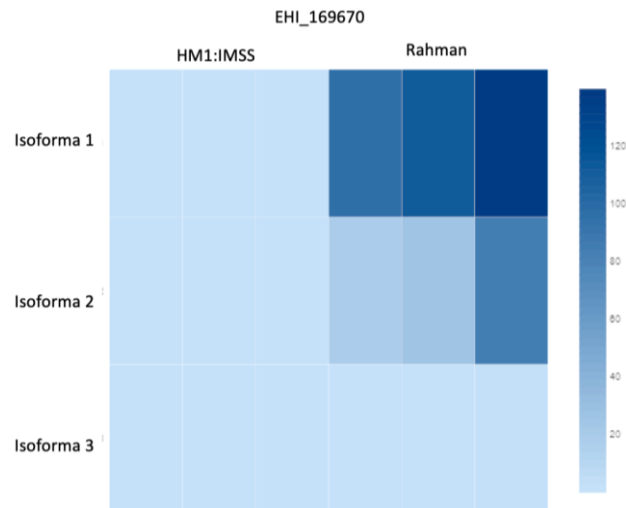


Figura 23. Heatmap de las tres isoformas circRNAs más expresados que contienen un mismo BSJ cuyo gene parental es EHI_169670. La expresión de los circRNAs está normalizada por TPM.

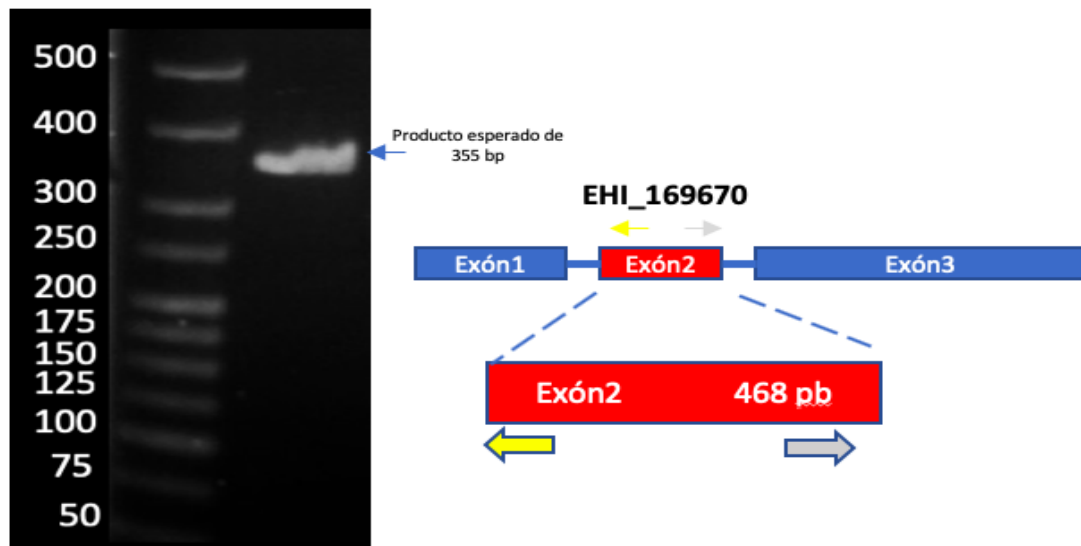


Figura 24. Validación del circRNA DS571377:17778|18245_A con locus EHI_169670 por RT-PCR divergente utilizando primers divergentes, en la figura de la derecha se observa un gel con el amplicón esperado de la reacción de PCR de la unión del BSJ del circRNA conformado del exón 2 cuyo locus es

EHI_169670. El tamaño del amplicón es de 355 bp. La flecha amarilla primer antisentido y la flecha gris es el primer sentido.

Las tres isoformas con el mismo BSJ son de mayor expresión diferencial, (sobrexpresadas en Rahman) siendo la isoforma A la más expresada (Figura 23) perteneciente al circRNA que abarca el segundo exón con locus EHI_169670 con ID DS571377:17778|18245_A. Usamos RT-PCR divergente para validar la existencia del circRNA más expresado diferencialmente (ID DS571377:17778|18245_A), a partir de RNA total de *E. histolytica* cepa HM1:IMSS usando los *primers* EHI169670_E2as y EHI169670E2s con las condiciones de PCR descritas en la tabla 1. En la figura 24 observamos que se obtuvo el amplicón de aproximadamente 355 bp, cuya secuenciación está en curso.

El circRNA DS571377:17778|18245_A del locus EHI_169670 el cual denominamos Circ169670ex2A. está sobreexpresado diferencialmente en la cepa Rahman siendo casi indetectable en la cepa HM1:IMSS. Una vez que las amebas de la cepa HM1:IMSS son inducidas a virulencia al inocularlas a los hígados de hámster, los transcritos de este locus son ampliamente sobreexpresados (Meyer y cols., 2016) aunque hasta la fecha no se ha analizado la expresión de circRNAs amebianos en estas condiciones.

De esta manera, la única cepa amebiana útil para comparar expresión de circRNAs relacionados a virulencia fue la cepa L6 de *E. invadens*. El circRNA scaff_1105074726403:279553|280008 (locus EIN_391640) de *E. invadens* mostró más lecturas BSJ, es decir que son muy abundantes en condiciones *steady-state*. El locus EIN_391640 es ortólogo al locus EHI_169670 por lo que buscamos similitudes de secuencia entre los circRNAs que resultan de ellos. El alineamiento mostró una identidad del 61.9 % entre secuencias (Figura 25), indicando que dichos exones están conservados entre especies y son utilizados como ncRNAs posiblemente con funciones regulatorias en virulencia.


```

EIN_391640 , GAAAGTGCCTTAAAGGACAAAGATGGGAAAGATCTCGGAATTGAAACAAGCTTCAGGAA 60
EHI_169670 GAAAGTGCAGTTAAAGAAAAAGATATTCAAATGAAAGAAATGAAAGTTCAAGAA 60
*** ** * * * * * * * * * * * * * * * * * * * * * * * * * * * *

EIN_391640 , GAGACAAAACAAAAGAAAATACAAAGCCGCTCTTCAGTCTGCTAACAGCTTGACTGCC 120
EHI_169670 GAAACAAAAGAAAAGAAAGAGGCTAAAGCTTCTTTAGCAATTTGAGTTGCTGCTGAAAGCT 120
** * * * * * * * * * * * * * * * * * * * * * * * * * * * *

EIN_391640 , GAGTTGAAAGGGAACCTATTTCAAACAAAAGAAAATGAAGCTGCTGAGATGAAGAAACAAT 180
EHI_169670 ACACCTTAA-----AGCAGAGTTGAAAAGAAAGATCAAGAAATTAAGAAATAAGGGT 171
* * * * * * * * * * * * * * * * * * * * * * * * * * * *

EIN_391640 , GAAGAAAAGG-----AGAAATAAGAAAGATCA-----ACTCGAAAAGGAAA 220
EHI_169670 GAAGAAATTAGAAAAGAAAAGAAACAAAGCAAAAGAAAATGAAGAAATTCAAAAAGAGA 231
***** * * * * * * * * * * * * * * * * * * * * * * * * * * * *

EIN_391640 , TTGAAGACCTCAGAAAGAGGTGACGACGAGCGAAAGCAAA-----AGAAAGCGGTAG 274
EHI_169670 AAGAAGAACAAACAAAGAGGTGAGAGATTAGAGGAGAGAAAAGAAATACGAAACAAA 291
***** * * * * * * * * * * * * * * * * * * * * * * * * * * * *

EIN_391640 , CTGGAGAAATTAGCAGTATCTGCTGCACTTGCTGTTGAGTTGAAGGCCACTGTAGCGAAGA 334
EHI_169670 AGCTAGAAGAATTAGAAAGAAAAGTTAATGATTCTGAGAAAGAGAAATA-----ATGAATT 346
* * * * * * * * * * * * * * * * * * * * * * * * * * * *

EIN_391640 , AAGAAGAAGACATAGAAAGAGCTTAAGAAAGGTTGAAGAAATCGAGAAAGATGCGCCA 394
EHI_169670 AAAAGGTCACCTT--AAAGACTTACAAAAGAAATAGAAGAAACTGAAAGAAATGCTGCTG 404
** * * * * * * * * * * * * * * * * * * * * * * * * * * * *

EIN_391640 , AAGGCTCTGAAGATCTTTTTCGACAGAAAGAAATGAAGAGATCGAGAAAATCAAGAACGAGA 454
EHI_169670 CAGGTTCTGAAGATTTTAAACAAAGAAATGAAGAAATAGCAATATTAAGAAAGAA 464
*** * * * * * * * * * * * * * * * * * * * * * * * * * * * *

EIN_391640 , AA 456
EHI_169670 AA 466
**

```

Figura 25. Alineamiento de los circRNAs más expresados en *E. invadens* y *E. histolytica*. El alineamiento tiene una identidad del 61.9%, y un gap con 9.5%

8.7 Los CircRNAs en *Entamoeba invadens*

Para el transcriptoma de *E. invadens*, a partir de 8 genotecas construidas de RNA total sin depleción de RNA ribosomal, se computaron un total de 199 circRNAs, de los cuales se caracterizaron un total de 188 circRNAs. De los datos computados según la anotación de *E. invadens* utilizada, se encontraron un total de 2 circRNAs intrónicos, 50 intergénicos y 127 exónicos.

Tabla 4. CircRNAs con más lecturas de BSJ en *E. invadens* detectados por CIRI2.

circRNA_ID	Número de lecturas de BSJ	Tipo de circRNA	Locus
scaff_1105074726403:279553 280008	56	Intronic/exonic	EIN_391640
scaff_1105074724939:262098 262511	5	exon	EIN_419420
scaff_1105074725849:128792 128935	5	intergenic_region	n/a
scaff_1105074724939:262098 262511	3	exon	EIN_419420

Para el circRNA intrónico cuyo ID es scaff_1105074726403:279553|280008 su gene parental EIN_391640 es el más expresado. Este tiene un elemento intrónico de 15bp y un elemento exónico de 440bp proveniente del segundo exón.

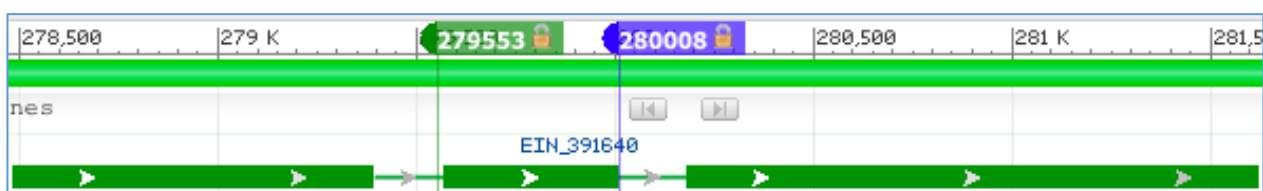


Figura 26. Principio y final del circRNA con ID scaff_1105074726403:279553|280008 tipo intrónico/exónico. Mapa génico con las coordenadas del inicio y final del circRNA exónico-intrónico, el cual abarca el segundo exón y el primer intrón del gen EIN_391640.

8.8 Análisis STRING de los circRNAs identificados de las cepas *E. histolytica* y *E. invadens*.

En biología molecular, STRING es una herramienta de búsqueda para la obtención de las interacciones de Genes / Proteínas (por sus siglas en inglés, **S**earch **T**ool for the **R**etrieval of **I**nteracting **G**enes/**P**roteins), es una base de datos biológicos y un reservorio web de interacciones proteína-proteína conocidas y/o predichas. Por fortuna, la base de datos de STRING contiene información de interacciones proteína-proteína de *E. histolytica*, sin embargo, para *E. invadens* no. Por lo tanto, para resolver esto, se buscaron los genes parentales ortólogos en *E. histolytica* (*E. histolytica* está en la base de datos de STRING) de los circRNAs identificados en *E. invades*.

Se realizó dos análisis STRING para las dos especies de Entamoeba de todos los circRNAs identificados:

El primer análisis consistió en utilizar los genes parentales de los circRNAs identificados en *E. histolytica*.

El segundo análisis consistió en utilizar los genes parentales ortólogos de *E. histolytica* de los genes parentales de los circRNAs identificados en *E. invadens*.

Las vías bioquímicas que tienen en común ambas especies en las cuales participan los genes parentales de los circRNAs identificados en ambas especies (genes parentales ortólogos en el caso de *E. invades*) se obtuvo por los resultados del análisis del enriquecimiento funcional además de las redes proporcionadas por el análisis STRING.

Los enriquecimientos funcionales en las redes que se obtuvieron en los análisis de los datos de este proyecto fueron: KEGG Pathways, UniProt Keywords, PFAM Protein Domains y SMART Protein Domains (este último sólo fue para genes parentales de circRNAs que se expresaban diferencialmente de circRNAs).

8.8.1 Análisis STRING de los genes parentales de los circRNAs identificados en *E. histolytica*

En *E. histolytica* se decidió utilizar un *score* mínimo con un nivel de confianza alto de 0.700. Por lo tanto los nodos que están unidos entre estos tienen un nivel de confianza de 0.700 o 0.900. Se decidió trabajar con este nivel de confianza alto debido a la gran cantidad de ejes que representan las conexiones entre los nodos. En el enriquecimiento funcional el valor del FDR en este tipo de análisis es un indicador el cual, mientras menor sea este valor hay mayor enriquecimiento de la vía o dicho de otra manera, hay más proteínas/genes de una vía bioquímica que están interactuando entre sí. Las rutas bioquímicas arrojadas del análisis son las siguientes, en orden de enriquecimiento descendente: Ribosomal, glucolisis/gluconeogénesis, metabolismo del almidón y sucrosa, biosíntesis de metabolitos secundarios, vías metabólicas, biosíntesis de antibióticos, amoebiasis, biogénesis del ribosoma en eucariotas y finalmente interconversiones de pentosa y glucuronato. En el enriquecimiento funcional de KEEG (figura 27) se observa que el cluster más conspicuo es el ribosomal.

Network Stats			
number of nodes:	336	expected number of edges:	562
number of edges:	625	PPI enrichment p-value:	0.00479
average node degree:	3.72	<i>your network has significantly more interactions than expected (what does that mean?)</i>	
avg. local clustering coefficient:	0.226		
KEGG Pathways			
pathway	description	count in gene set	false discovery rate
ehi03010	Ribosome	21 of 135	4.05e-05
ehi00010	Glycolysis / Gluconeogenesis	8 of 24	0.00063
ehi00500	Starch and sucrose metabolism	6 of 17	0.0034
ehi01110	Biosynthesis of secondary metabolites	12 of 86	0.0058
ehi01100	Metabolic pathways	24 of 266	0.0058
ehi01130	Biosynthesis of antibiotics	10 of 71	0.0105
ehi05146	Amoebiasis	8 of 54	0.0205
ehi03008	Ribosome biogenesis in eukaryotes	9 of 72	0.0281
ehi00040	Pentose and glucuronate interconversions	3 of 7	0.0324

Figura 27. Información de enriquecimientos funcionales KEEG y del estado de la red de interacciones de los genes parentales de los circRNAs detectados en *E. histolytica*.

Otros análisis de enriquecimiento funcional también fueron realizados con niveles de confianza más bajos en el análisis de interacciones entre proteínas (Apéndice C)

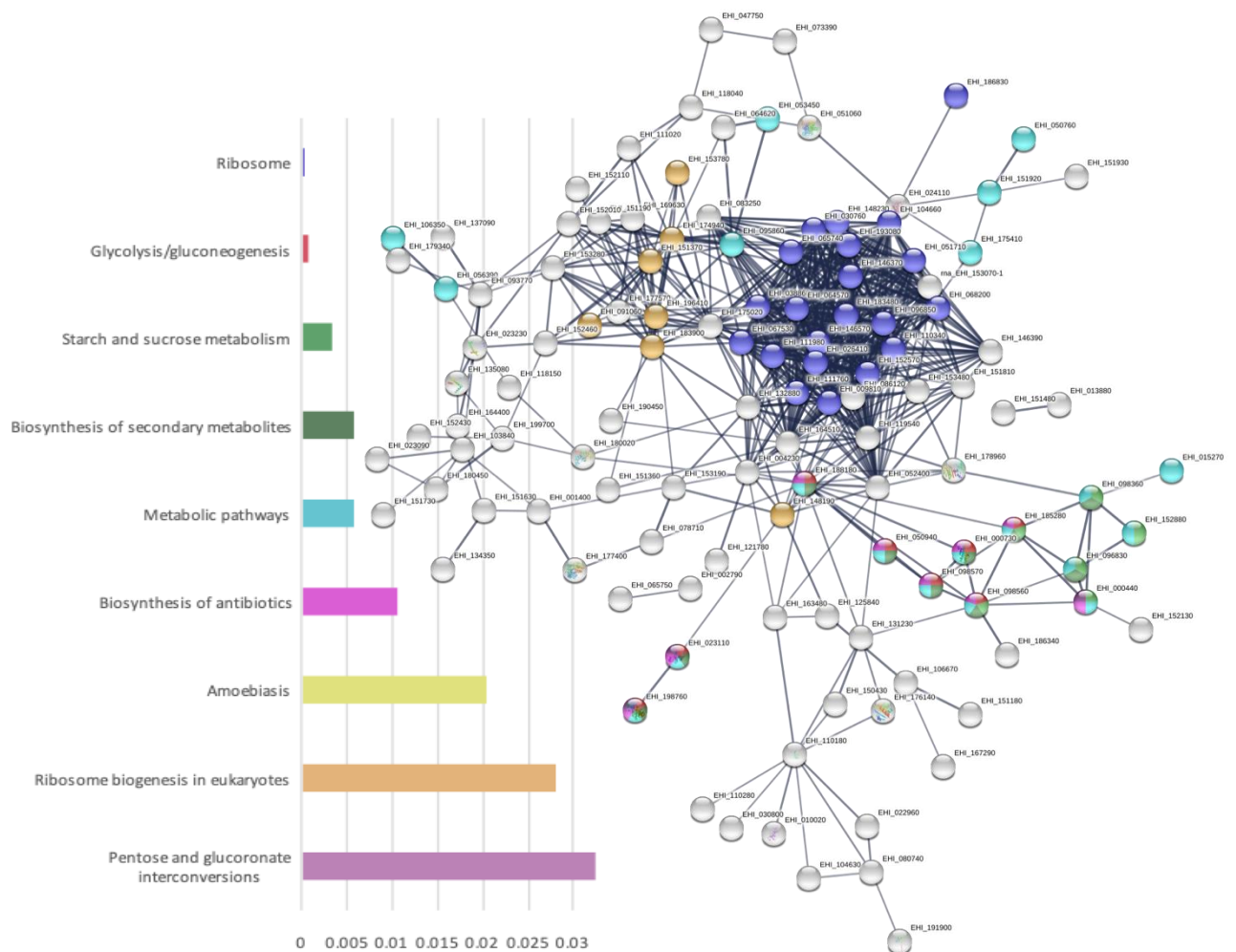


Figura 28. Análisis de enriquecimientos funcionales KEGG en la red de interacciones de los genes parentales de los circRNAs en *E. histolytica*. Los valores de la gráfica es el FDR de cada vía presentada. Los nodos están coloreados con los valores de FDR de la gráfica del enriquecimiento funcional con el fin de ser localizada la vía en la cual participa en la gráfica del enriquecimiento funcional la cual refleja el FDR(false discovery rate) con un umbral de 0.05.

8.8.2 Análisis STRING en de los genes parentales ortólogos en *E. histolytica* de los circRNAs identificados en *E. invadens*

En *E. invadens* se decidió utilizar un score mínimo requerido con un nivel de confianza media de 0.400 por lo tanto los nodos que están unidos entre estos tienen un nivel de confianza de 0.400 a 0.900, esto fue debido a la poca profundidad de secuenciación de *E. invadens*.

La única vía bioquímica que mostró el análisis de enriquecimiento funcional KEGG de los genes parentales ortólogos en *E. histolytica* de los circRNAs encontrados en *E. invadens* fue la vía de las proteínas ribosomales. La única vía bioquímica enriquecida fue los genes ribosomales con un FDR de 0.0167. Otros análisis de enriquecimiento funcional también fueron realizados con niveles de confianza más bajo (Apéndice D)

Network Stats			
number of nodes:	88	expected number of edges:	129
number of edges:	141	PPI enrichment p-value:	0.149
average node degree:	3.2	your network does not have significantly more interactions than expected	
avg. local clustering coefficient:	0.293		

KEGG Pathways			
pathway	description	count in gene set	false discovery rate
ehi03010	Ribosome	7 of 135	0.0167

Figura 29. Información de enriquecimientos funcionales KEGG y del estado de la red de interacciones de los genes parentales ortólogos de *E. histolytica* de los circRNAs identificados en *E. invadens*. La única vía enriquecida fue la de las proteínas ribosomales con un FDR de 0.0167 dato obtenido del enriquecimiento funcional KEGG.

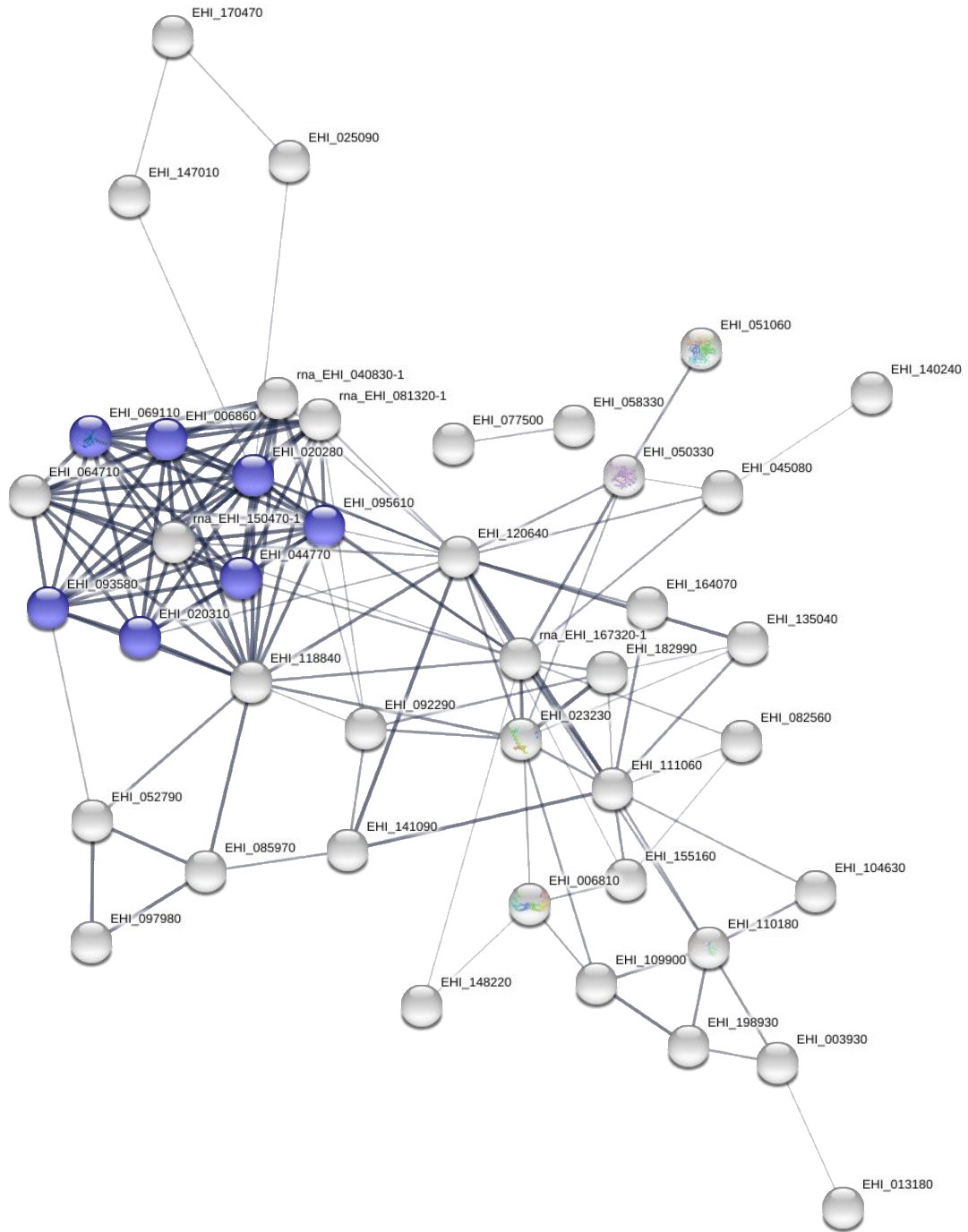


Figura 30. Networking de interacciones de los genes parentales ortólogos en *E. histolytica* de los circRNAs encontrados en *E. invadens* y el enriquecimiento funcional con un mínimo score requerido con un nivel de confianza medio de 0.400. La única vía bioquímica enriquecida fue los genes ribosomales con un FDR de 0.0167.

8.8.3 Análisis STRING de los circRNAs expresados diferencialmente de *E. histolytica* de las cepas virulentas y no virulentas

Para el análisis STRING de los genes parentales de los circRNAs expresados diferencialmente en *E. histolytica* se decidió utilizar un score mínimo con un nivel de confianza media de 0.400 por lo tanto los nodos que están unidos entre estos tienen un nivel de confianza de 0.400 a 0.900, esto con el fin de ver las vías en las que participan los genes parentales de los circRNAs expresados diferencialmente.

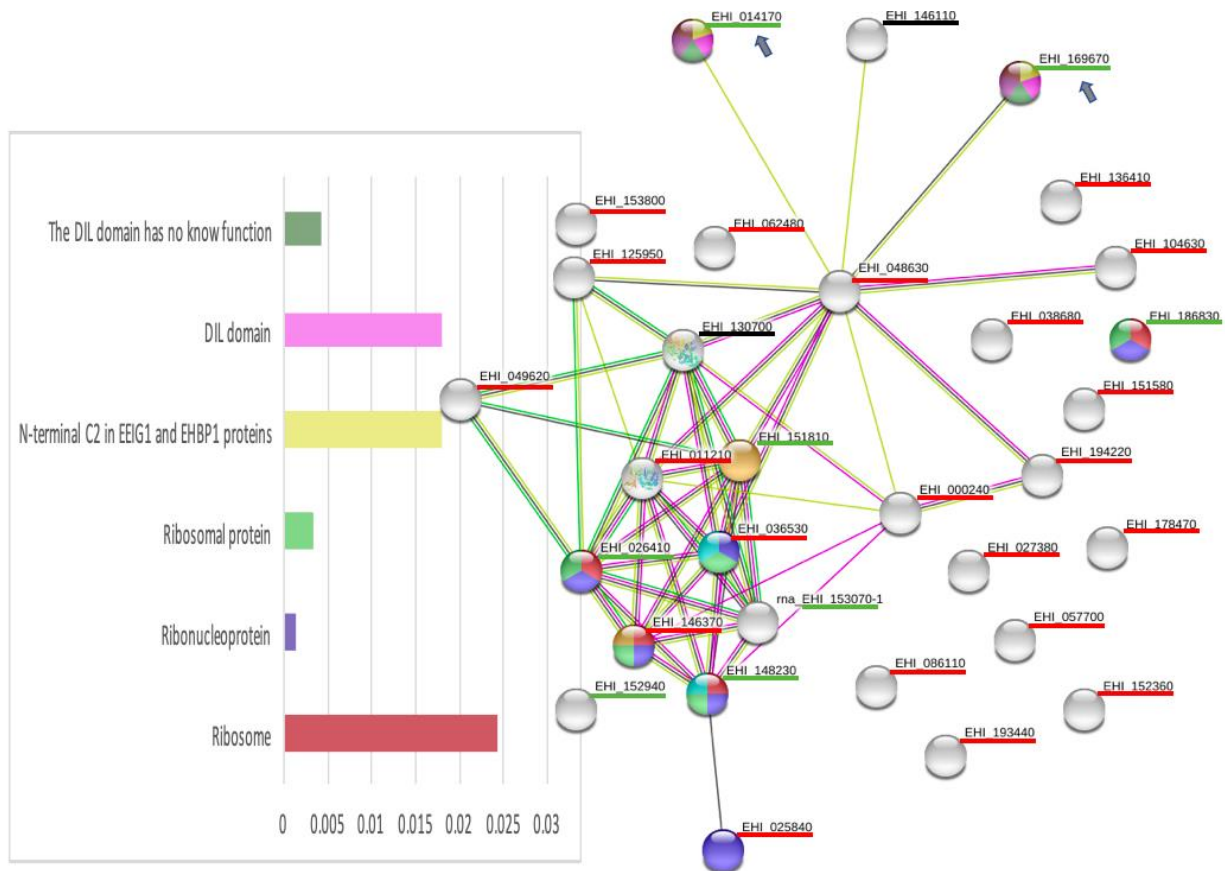


Figura 31. Network de las interacciones de los genes parentales de los circRNAs expresados diferencialmente en *E. histolytica* y su análisis de enriquecimiento funcional. Los nodos están coloreados con el fin de ser localizados en la gráfica del enriquecimiento funcional la cual refleja el FDR (false discovery rate) con un umbral de 0.05. Los edges (o conexiones) tienen un nivel de confianza entre 0.400 a 0.900. Las flechas indican que los genes pertenecen a genes de virulencia.

En los mismos análisis realizados con genes parentales de los circRNAs expresados diferencialmente se observa que las vías en las que participan son: ribosomas (KEGG), ribonucleoproteínas y proteínas ribosomales (UniProt Keywords), N-terminal C2 en EEIG1 y proteínas EHBP1 (PFAM Protein Domains) y el dominio DIL que no tiene función conocida (SMART Protein Domains). Interesantemente, el gen EHI_048630 que codifica una actina putativa pertenece al *cluster* de proteínas ribosomales y ribonucleoproteínas. Este gen también conecta con genes que codifican factores de virulencia como EHI_1609670 y EHI_014170.

9. Discusión

El presente trabajo se enfocó en detectar circRNAs mediante minería de datos de secuenciación, de cepas de *Entamoeba*. Nuestro trabajo está validado porque inicialmente corrimos los mismos análisis en bancos de datos de rata (Zhou y cols., 2018), encontrando resultados similares a los reportados (datos no mostrados). En segundo lugar, encontramos los identificadores de circRNAs esperados (BSJs) tanto cuando se minaron bases de datos de *E. histolytica* como de *E. invadens*. En tercer lugar, en la comparación de circRNAs entre especies y cepas amebianas se detectaron circRNAs de genes ortólogos. Finalmente, uno de los circRNAs detectado mediante esta estrategia fue validado por RT-PCR.

En este trabajo los circRNAs se identificaron mediante la búsqueda de lecturas conteniendo secuencias abarcando las BSJs, característica obvia de los circRNAs *per se*. Por otro lado, también se identificó circRNAs utilizando la característica del solapamiento reverso 5' y 3', encontrando así nuevos BSJ de circRNAs de longitud pequeña y BSJ que no pudieron ser detectados por CIRI2, pero sí por CIRI-FULL.

Debido a que los BSJ contienen señales de *clipping* quiásticas que les permiten a las lecturas quiásticas ser alineadas al genoma de manera quiástica, esto es una primera evidencia de que estos circRNAs amebianos son identificados *in silico*.

Con los análisis anteriores, en este trabajo detectamos un total de 599 circRNAs, de los cuales 310 circRNAs presentan más de una lectura. Esta metodología reduce la tasa de descubrimiento de falsos positivos pues computacionalmente fueron detectados al menos una vez en dos genotecas o más de una vez en la misma genoteca. Esto hace que los circRNAs detectados sean válidos computacionalmente.

Tanto los transcriptomas de *E. histolytica* y *E. invadens*, en el control de calidad todas las secuencias presentan secuencias de buena calidad. El transcriptoma de Hon y cols. (2012) tiene la particularidad de que fue obtenido mediante una secuenciación profunda

dirigida a la detección de formas alternativas de transcritos. Por lo tanto, se obtuvo más información de circRNAs de lo que haría una secuenciación de una genoteca construida a partir de poly(A)⁺ con una profundidad de secuenciación dirigida sólo a analizar la expresión de transcritos mensajeros..

Cuando se usaron datos de experimentos de RNA-seq de Weber y cols. (2016) siendo genotecas a partir de poly (A)⁺, nosotros detectamos muy pocos circRNAs (apéndice E) A pesar de eso, encontramos algunos circRNAs con el mismo BSJ en el transcriptoma de Hon y cols. (2013), dando evidencia de que en la búsqueda de circRNAs amebianos es reproducible entre diferentes genotecas de otros experimentos de RNA-seq de la misma especie de *E. histolytica* HM1:IMSS.

Por otro lado, hay reportes de conjuntos de datos de secuencia de RNA seleccionado con poly(A)⁺, que proporcionan evidencia de la expresión de circRNAs (Lamm et al. 2011). Caso contrario fue la utilización de la genoteca de Hon y cols. (2012) que, a pesar de haber sido construida a partir de RNA sin las condiciones ideales para el enriquecimiento de circRNAs, nos permitió obtener numerosas secuencias correspondientes a circRNAs.

Por la profundidad de secuenciación de la genoteca principal usada en este proyecto es que se identificaron varios circRNAs, incluso para realizar perfiles de expresión diferencias, entre otros análisis. Por lo tanto, la profundidad de secuenciación es el elemento clave con el que se pudo compensar el problema del enriquecimiento de circRNAs. Cabe aclarar que experimentos de tal profundidad son muy costosos, razón por la cual nosotros nos inclinamos a realizar nuestros experimentos de identificación de estas moléculas mediante minería de datos de datos de RNA-seq existentes.

Decidimos no considerar un filtrado por señales de *backsplicing* canónicas pues se utilizó de una estrategia que busca secuencias BSJ que presentaron señales de *clipping* quiástico (señales de BSJs). Es decir, se optó por identificar circRNAs mediante señales de *backsplicing* tanto canónicas como no canónicas. Todo con el fin de obtener el mayor

rendimiento en la identificación de circRNAs. No se pudo hacer un análisis de la proporción de los tipos de señales de *backsplicing* de los circRNAs debido a que es una tarea masiva y repetitiva con probables errores, sobre todo si se realiza manualmente. Para resolver este problema se puede recurrir al diseño de un *script* en algún lenguaje de programación como Python.

Con el fin de investigar un patrón de señales de *splicing* en circRNAs de plantas (Chucols., 2018), minaron bases de datos usando todos los tipos de señales de *backsplicing* (permutaciones y combinaciones de 4 bases nucleotídicas) en *O. sativa* y *A. thaliana* utilizando, como en este proyecto el software CIRI2. Sus resultados mostraron, con alta confianza, que los circRNAs con señales de *backsplicing* canónicos no tuvieron la mayor proporción (6% en *O. sativa* y un 9% en *A. thaliana*). Esta información es importante debido a que en ratas y en humanos (metazoa en general) las señales canónicas de *backsplicing* de los circRNAs comprenden alrededor de un 95%. El género *Entamoeba* se desprendió tempranamente en el árbol de la vida (Figura 32), incluso antes que las plantas y metazoos, por lo tanto, hacer un análisis de las señales de *backsplicing* de Entamoeba podría ayudarnos a comprender mejor la evolución de la circularización de RNAs amebianos, en comparación con la de los organismos que bifurcaron tardíamente en el árbol de la vida.

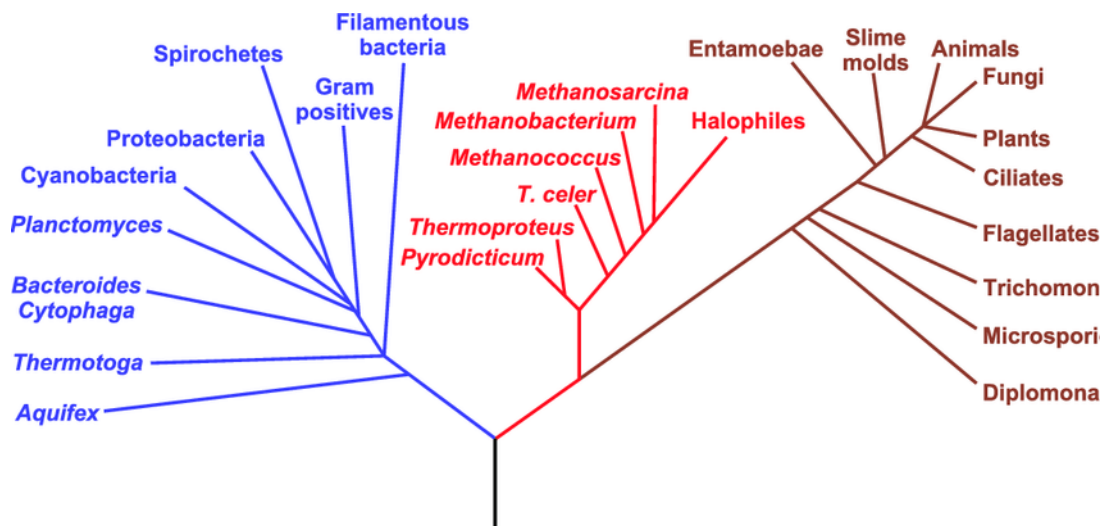


Figura 32. El "árbol de la vida" que muestra los tres dominios principales de los organismos vivos: las bacterias, las arqueas y la eucariota. Basado en sus secuencias de rRNA 16S y 18S (Rivera y Lake, 2004).

La mayor frecuencia de la longitud interna de circRNAs computados fue de 141 bp (ver figura 12). Este es un punto interesante debido a que organismos de especies de más reciente ramificación presentan mayoritariamente circRNAs en diferentes tejidos con una longitud de alrededor de 500 bp (T. Zhou y cols., 2018). Es probable que dado que *E. histolytica* es un organismo protista unicelular, de ramificación temprana (figura 27) mantenga su estructura génica relativamente simple (longitud media de un gen 1.26 kb vs. 3 kb en humanos), aunque aparentemente comparta con organismos “superiores” los mecanismos de la biogénesis de los circRNAs, como el *backsplicing*.

La mayoría de los estudios de circRNAs se ha enfocado a analizar principalmente aquellos relacionados con enfermedades en humano, plantas y otros organismos modelo. Por lo tanto, un análisis más detallado de circRNAs de organismos que divergieron de manera temprana del árbol de la vida podría dar un mayor entendimiento del origen de estas moléculas, así como de su biogénesis sus posibles funciones en la biología de los organismos como en la evolución.

En cuanto a la estructura de los circRNAs, nuestros análisis mostraron que la mayoría de ellos están conformados por un solo elemento y, en comparación con los circRNAs murinos que generalmente tienen más de 2 elementos exónicos (incluso hasta 19 exones adyacentes del mismo gen parental) (Zhou y cols., 2018), nosotros no detectamos circRNAs con más de dos elementos. Esto puede reflejar la estructura génica relativamente simple de *E. histolytica* y a que ésta posee pocos genes multiexónicos. No es de sorprenderse que no haya circRNAs con más exones puesto que *E. histolytica* tiene un total de 358 genes con 3 exones, 70 genes con 4 exones, 11 genes con 5 exones y 5 genes con 6 exones.

Es probable que el segundo exón de los genes tri-exónicos puedan formar circRNAs dado que existen evidencias para la biogénesis de dichos RNA circulares (Barrett, Wang, y Salzman, 2015). Los dos mecanismos para biogénesis de los circRNAs del exón 2 requieren de las partes intrónicas adyacentes del exón 2, ya sea para formar los circRNAs

a través de un intermediario lariat, o formar intermediarios en forma de “Y” donde las partes intrónicas adyacentes del exón 2 se juntan por complementariedad, facilitando la biogénesis de los circRNAs. Nosotros encontramos y validamos por RT-PCR divergente un circRNA del exón 2 cuyo locus es EHI_169670 cuyo gen parental está conformado de 3 exones. Cualquiera de los dos mecanismos descritos antes puede dar lugar a la biogénesis de este circRNA.

En el análisis de la correlación entre los promedios de los TPM lineales y TPM circulares en *E. histolytica* (Figura 15) observamos que hubo una correlación de la expresión de los circRNAs con sus contrapartes lineales, teniendo un $\rho = 0.35$ y un $p\text{-value} = 1.35 \times 10^{-18}$. En rata (Zhou y cols., 2018) se encontró una correlación de $\rho = 0.498$ por lo que la correlación de nuestros es ligeramente menor. Pero, aun así, existe una correlación con un incremento monótonico positivo de la expresión de los circRNAs respecto a la expresión transcripcional de sus correspondientes genes parentales.

Por otra parte, también se tenía la interrogante de si los circRNAs expresados diferencialmente tenían una correlación con sus contrapartes lineales. Efectivamente la expresión de los 39 circRNAs presentan una correlación positiva, observando un incremento monótonico en la expresión de los circRNAs y sus contrapartes lineales en todas las muestras de cepas virulentas y no virulentas (un $\rho = 0.56433$ y un $p\text{-value} = 0.00022$) (Figura 16). Debe hacerse notar que la correlación entre la expresión de circRNAs con sus contrapartes lineales es mejor en los expresados diferencialmente que cuando se considera el total de los circRNAs. Además, encontramos mayor correlación de circRNAs expresados diferencialmente con sus contrapartes lineales que los reportados en rata. Sugiriendo que los circRNAs expresados diferencialmente se encuentran regulados por los niveles de expresión transcripcional de sus correspondientes genes parentales.

La comparación de los TPM de circRNAs entre cepas amebianas (Figura 17) mostró que en la cepa avirulenta Rahman globalmente expresa tres órdenes de magnitud más

transcritos que la virulenta HM1:IMSS. Estos hallazgos posteriormente se confirmaron en un análisis estadístico comparando la fracción de circRNAs respecto a sus contrapartes lineales entre cepas amebianas. De esta manera encontramos que la cepa avirulenta en el 25 % de las fracciones circulares por arriba del 75% (tercer cuartil) hay más fracciones circulares en avirulencia que en virulencia. Se puede observar que en avirulencia hay mayores fracciones circulares atípicas. No se observan aumentos significativos en el 25% de las fracciones circulares por debajo del 75% (primer cuartil), concluyendo que en cultivos *steady-state* de cepas amebianas no virulentas hay mayor expresión de circRNAs exónicos.

El análisis de los PCA arrojó un *clúster* definido (para la cepa avirulenta. Esto puede reflejar que la ausencia de un *clúster* definido en amebas virulentas la expresión de los circRNAs depende de factores divergentes propios de la plasticidad genómica de *E. histolytica* HM1:IMSS (Weedall y cols., 2012). Es decir, las clonas derivadas de HM1:IMSS poseen amplia variabilidad genética.

Hasta el momento se ha reportado todo un compendio de miRNAs tanto para *E. histolytica* y *E. invadens*. Probablemente alguno de los circRNAs detectados en nuestro trabajo puedan funcionar como esponjas de algunos de esos miRNAs identificados computacionalmente. Sin embargo, por su complejidad no se pudo realizar ese análisis en este momento. También cabe la posibilidad de que los circRNAs identificados pueden tener alguna función codificante.

En cáncer, algunos de estos circRNAs funcionan como esponjas de miRNAs cuyo objetivo son miRNAs relacionados a oncoprotección, por lo tanto, en ese caso las esponjas de miRNAs oncoprotectores son marcadores moleculares. También se ha descubierto que hay presencia de circRNAs y miRNAs en exosomas, los circRNAs podrían ser utilizados como marcadores moleculares de infección de amebas de vida libre. Y no sólo *E. histolytica* sino también si se realizan más experimentos en ameba de vida libre la que comúnmente se llama *Naegleria fowleri* de infección rápida la cual su detección temprana es vital para la supervivencia del paciente infectado.

Los resultados del análisis STRING de los genes parentales de los circRNA identificados de los transcriptomas de *E. histolytica* utilizados aquí permite observar que las proteínas correspondientes a los circRNAs detectadas en el análisis tienen más interacciones entre ellas de lo que se esperaría en un conjunto aleatorio de proteínas/genes con características similares. Tal enriquecimiento de vías bioquímicas indica que las proteínas/genes están biológicamente conectadas como un grupo. Por otro lado en el enriquecimiento funcional por KEGG se encontraron genes que participan en la amebiasis: EHI_183900 (snoRNA binding protein, putative), EHI_091060 (ATPase, AAA family protein), EHI_196410 (Ribosome biogenesis protein BMS1, putative), EHI_151370 (WD domain containing protein), EHI_174940 (nucleolar GTP-binding protein 1, putative), EHI_153780 (hypotetical protein) y EHI_148190 (Ran family GTPase), implicando nuevamente a las vías de la biogénesis y del metabolismo ribosomal en eventos relacionados con el establecimiento de la infección amebiana.

Por el contrario, en *E. invadens* ocurrió que los resultados del análisis STRING de los genes parentales ortólogos de *E. histolytica* de los circRNAs identificados en *E. invadens* no arrojaron un enriquecimiento funcional distinto al aleatorio. Aun así, esto no implica necesariamente que no sea una selección de proteínas biológicamente significativa, cabe la posibilidad que estos genes ortólogos no se hayan estudiado a fondo, y/o que sus interacciones tal vez no están actualizadas en la base de datos de STRING, o tal vez que la ortología de los genes no proporciona información adecuada para realizar una red con muchas interacciones. A pesar de ello, el análisis funcional KEGG arrojó información valiosa de un *clúster* en común entre estas especies del género Entamoeba el cual correspondió a los genes ribosomales.

Además, hay dos *clústeres* en común que se determinaron utilizando el análisis de enriquecimiento funcional de UniProt Keywords (apéndices D y E con un nivel de confianza 0.150). Estos son: proteínas ribosomales y proteínas de unión a Actina. Así, los *clusters* que tienen en común ambas especies ordenados por el conteo de genes parentales de los circRNAs son: Ribosomales (KEGG pathway ehi03010), proteínas

ribosomales y proteínas de unión a Actina (UniProt Keywords KW-0689 y KW-0009 respectivamente).

Finalmente, en los mismos análisis realizados con genes parentales de los circRNAs expresados diferencialmente se observa que las vías en las que participan son: ribosomas (KEGG), ribonucleoproteínas y proteínas ribosomales (UniProt Keywords), N-terminal C2 en EEIG1 y proteínas EHBP1 (PFAM Protein Domains) y el dominio DIL que no tiene función conocida (SMART Protein Domains). Interesantemente, el gen EHI_048630 que codifica una actina putativa pertenece al *clúster* de proteínas ribosomales y ribonucleoproteínas. Este gen también conecta con genes que codifican factores de virulencia reportados previamente:

- EHI_014170 (factor de virulencia, proteína hipotética) determinado por (Weber y cols., 2016)
- EHI_169670 (factor de virulencia, proteína hipotética) determinado por (Meyer y cols., 2016)
- EHI_146110 (proteína hipotética) (Loftus y cols., 2005)
- EHI_104630 (flamina putativa 2) (Loftus y cols., 2005)
- EHI_194220 (proteína que contiene el dominio de la proteína fosfatasa) (Loftus y cols., 2005)
- EHI_000240 (subunidad beta de la proteína de unión a nucleótidos de guanina, putativo) (Loftus y cols., 2005)

Ya se sabe que los circRNAs exónicos son moléculas que están evolutivamente conservadas, probablemente en la expresión génica de eucariotas (Wang y cols., 2014). Las proteínas ribosomales se encuentran entre las proteínas altamente conservadas en todas las formas de vida (Ban y cols., 2014). No es de sorprenderse encontrar más circRNAs provenientes de genes parentales de genes ribosomales y esperar circRNAs conservados evolutivamente. Los genes ribosomales están evolutivamente conservados, además se saben que los circRNAs ribosomales interactúan con la membrana asociada a ribosomas. También es probable que los circRNAs pertenecientes a genes ribosomales

sean más abundantes debido a que son parte de una maquinaria antigua propias de las proteínas ribosomales para la función y regulación de esta en eucariotas. (Pamudurti y cols., 2017). Estudios en rata indican que la vía más conspicua de los genes parentales de los circRNAs identificados es la ribosomal (T. Zhou y cols., 2018) al igual que nuestros datos. Sin embargo, no hay presencia de circRNAs exónicos en procariontes siendo este un mecanismo característico de eucariontes (L. Chen y cols., 2015). Por lo tanto, entender la evolución de los circRNAs exónicos puede dar pauta a un mejor entendimiento de la biogénesis de esto a lo largo del árbol de la vida de estas moléculas circulares.

10. Conclusiones

Se logró exitosamente caracterizar, cuantificar y ensamblar *in silico* circRNAs en *Entamoeba histolytica*, HM-1:IMSS y Rahman utilizando datos de NGS creadas a partir de genotecas no óptimas. Por otro lado, también se logró caracterizar circRNAs de *E. invadens* utilizando datos de NGS creadas a partir de genotecas no óptimas. Se observó la participación de los genes parentales de los circRNAs de diferentes rutas bioquímicas, aunque el *cluster* más conspicuo correspondió a las proteínas ribosomales y a las ribonucleoproteínas. Encontramos que hubo una correlación monotónica positiva de la expresión de los transcritos circulares con sus contrapartes lineales, la correlación aumentó cuando se comparó la expresión de los transcritos de los circRNAs expresados diferencialmente con sus contrapartes lineales.

El presente trabajo se enfocó a detectar por minería de datos secuenciación circRNAs en cepas del género *Entamoeba*. Nuestro trabajo está validado porque inicialmente corrimos los mismos análisis en bancos de datos de rata, encontrando resultados similares a los reportados (datos no mostrados). En segundo lugar, encontramos los identificadores de circRNAs esperados (BSJs) tanto cuando se minaron bases de datos de *E. histolytica* como de *E. invadens*. En tercer lugar, la comparación de circRNAs entre especies y cepas amebianas se detectaron circRNAs de genes ortólogos. En cuarto lugar, identificamos computacionalmente circRNAs idénticos en otros transcriptomas de la misma especie. Finalmente, uno de los circRNAs fue validado por RT-PCR. Así cumpliendo con todos los objetivos de este proyecto.

11. Perspectivas

Hay mucho por hacer, tanto en el aspecto bioinformático como experimental para continuar con este trabajo, que se aspiran a realizar en el doctorado, entre ellas:

- Realizar el análisis de circRNAs con funciones de esponjas de miRNAs amebianos *in silico*.
- Encontrar qué tipo proteínas interactúan con nuestro compendio de circRNAs *in silico*, se pretende realizar minería de datos a partir de bases de datos de experimentos de CLIP-seq de *E. histolytica* con el fin de encontrar proteínas que interactúen con secuencias en circRNAs.
- Realizar la búsqueda de patrones o secuencias consenso que den respuesta a la biogénesis de los circRNAs amebianos y otros organismos unicelulares, a partir de secuencias consenso que tengan en común en las zonas adyacentes genómicas de los circRNAs y también intrónicas si los circRNAs son originados a partir de genes multiexónicos, utilizando diferentes estrategias computacionales. Probablemente realizar este tipo de análisis nos ayudaría a comprender mejor la evolución de la circularización de RNAs amebianos, en comparación con la de los organismos que bifurcaron tardíamente en el árbol de la vida.
- Realizar el análisis de la proporción de los tipos de señales de *backsplicing* de los circRNA identificados en este trabajo y por identificar, debido a que es una tarea masiva y repetitiva con probables errores, sobre todo si se realiza manualmente, se pretende diseñar de un *script* en algún lenguaje de programación como Python que resuelva esta tarea.

- Realizar nuestros propios experimentos de RNA-seq a partir de RNA enriquecido con moléculas de circRNA a partir de RNA de amebas inducidas a virulencia, para obtener un mejor repertorio de circRNAs y ver la implicación de estos en virulencia.
- A partir de RNA enriquecido con circRNAs, se realizará una inmunoprecipitación dirigido a secuencias que contengan secuencias con m⁶A, que posteriormente se secuenciará masivamente para luego ser corridos por *pipelines* bioinformáticos especializados (C. Zhou y cols., 2017), con el fin de observar en los circRNAs secuencias consenso que contengan metilaciones, recordando que éstas metilaciones son indicativos de que estos circRNAs pueden dar origen a pequeños péptidos o proteínas más grandes.

12. Referencias

- Ashwal-Fluss, Reut, Markus Meyer, Nagarjuna Reddy Pamudurti, Andranik Ivanov, Osnat Bartok, Mor Hanan, Naveh Evantal, Sebastian Memczak, Nikolaus Rajewsky, and Sebastian Kadener. 2014. "CircRNA Biogenesis Competes with Pre-mRNA Splicing." *Molecular Cell*. <https://doi.org/10.1016/j.molcel.2014.08.019>.
- Ban, Nenad, Roland Beckmann, Jamie H.D. Cate, Jonathan D. Dinman, François Dragon, Steven R. Ellis, Denis L.J. Lafontaine, et al. 2014. "A New System for Naming Ribosomal Proteins." *Current Opinion in Structural Biology*. <https://doi.org/10.1016/j.sbi.2014.01.002>.
- Barrett, Steven P., Peter L. Wang, and Julia Salzman. 2015. "Circular RNA Biogenesis Can Proceed through an Exon-Containing Lariat Precursor." *ELife*. <https://doi.org/10.7554/eLife.07540>.
- Chao, Cindy Wang, David C. Chan, Ann Kuo, and Philip Leder. 1998. "The Mouse Formin (Fmn) Gene: Abundant Circular RNA Transcripts and Gene-Targeted Deletion Analysis." *Molecular Medicine*. <https://doi.org/10.1007/bf03401761>.
- Chen, Liang, Chuan Huang, Xiaolin Wang, and Ge Shan. 2015. "Circular RNAs in Eukaryotic Cells." *Current Genomics*. <https://doi.org/10.2174/1389202916666150707161554>.
- Chen, Ling Ling, and Li Yang. 2015. "Regulation of CircRNA Biogenesis." *RNA Biology*. <https://doi.org/10.1080/15476286.2015.1020271>.
- Chu, Qinjie, Panpan Bai, Xintian Zhu, Xingchen Zhang, Lingfeng Mao, Qian-Hao Zhu, Longjiang Fan, and Chu-Yu Ye. 2018. "Characteristics of Plant Circular RNAs." *Briefings in Bioinformatics*. <https://doi.org/10.1093/bib/bby111>.
- Chuang, Trees Juen, Chan Shuo Wu, Chia Ying Chen, Li Yuan Hung, Tai Wei Chiang, and Min Yu Yang. 2016. "NCLscan: Accurate Identification of Non-Co-Linear Transcripts (Fusion, Trans-Splicing and Circular RNA) with a Good Balance between Sensitivity and Precision." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkv1013>.
- Conn, Vanessa M., Véronique Hugouvieux, Aditya Nayak, Stephanie A. Conos, Giovanna Capovilla, Gökhan Cildir, Agnès Jourdain, et al. 2017. "A CircRNA from SEPALLATA3 Regulates Splicing of Its Cognate mRNA through R-Loop Formation."

- Nature Plants*. <https://doi.org/10.1038/nplants.2017.53>.
- De, Subhajyoti, Dibyarupa Pal, and Sudip K. Ghosh. 2006. "Entamoeba Histolytica: Computational Identification of Putative MicroRNA Candidates." *Experimental Parasitology*. <https://doi.org/10.1016/j.exppara.2006.01.009>.
- Du, William W., Weining Yang, Elizabeth Liu, Zhenguo Yang, Preet Dhaliwal, and Burton B. Yang. 2016. "Foxo3 Circular RNA Retards Cell Cycle Progression via Forming Ternary Complexes with P21 and CDK2." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkw027>.
- Ehrenkaufer, Gretchen M., Gareth D. Weedall, Daryl Williams, Hernan A. Lorenzi, Elisabet Caler, Neil Hall, and Upinder Singh. 2013. "The Genome and Transcriptome of the Enteric Parasite Entamoeba Invadens, a Model for Encystation." *Genome Biology*. <https://doi.org/10.1186/gb-2013-14-7-r77>.
- Ehrenkaufer, Gretchen M, Upinder Singh, Gareth D Weedall, Daryl Williams, Neil Hall, Hernan A Lorenzi, and Elisabet Caler. 2013. "The Genome and Transcriptome of the Enteric Parasite Entamoeba Invadens, a Model for Encystation." *Genome Biology*. <https://doi.org/10.1186/gb-2013-14-7-r77>.
- Enuka, Yehoshua, Mattia Lauriola, Morris E. Feldman, Aldema Sas-Chen, Igor Ulitsky, and Yosef Yarden. 2016. "Circular RNAs Are Long-Lived and Display Only Minimal Early Alterations in Response to a Growth Factor." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkv1367>.
- Faust, Daniela M., and Nancy Guillen. 2012. "Virulence and Virulence Factors in Entamoeba Histolytica, the Agent of Human Amoebiasis." *Microbes and Infection*. <https://doi.org/10.1016/j.micinf.2012.05.013>.
- Gao, Yuan, Jinfeng Wang, and Fangqing Zhao. 2015. "CIRI: An Efficient and Unbiased Algorithm for de Novo Circular RNA Identification." *Genome Biology*. <https://doi.org/10.1186/s13059-014-0571-3>.
- Gao, Yuan, and Fangqing Zhao. 2018. "Computational Strategies for Exploring Circular RNAs." *Trends in Genetics*. <https://doi.org/10.1016/j.tig.2017.12.016>.
- Gupta, Abhishek Kumar, Sunil Kumar Panigrahi, Alok Bhattacharya, and Sudha Bhattacharya. 2012. "Self-Circularizing 5'-ETS RNAs Accumulate along with Unprocessed Pre Ribosomal RNAs in Growth-Stressed Entamoeba Histolytica."

- Scientific Reports*. <https://doi.org/10.1038/srep00303>.
- Hon, Chung Chau, Christian Weber, Odile Sismeiro, Caroline Proux, Mikael Koutero, Marc Deloger, Sarbashis Das, et al. 2013. "Quantification of Stochastic Noise of Splicing and Polyadenylation in *Entamoeba Histolytica*." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gks1271>.
- Huang, Guanli, Hua Zhu, Yixiong Shi, Wenzhi Wu, Huajie Cai, and Xiangjian Chen. 2015. "Cir-ITCH Plays an Inhibitory Role in Colorectal Cancer by Regulating the Wnt/ β -Catenin Pathway." *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0131225>.
- Huang, Guanqun, Shuaihu Li, Nuo Yang, Yongdong Zou, Duo Zheng, and Tian Xiao. 2017. "Recent Progress in Circular RNAs in Human Cancers." *Cancer Letters*. <https://doi.org/10.1016/j.canlet.2017.07.002>.
- Kean, B. H. 2017. "Clinical Parasitology." *The American Journal of Tropical Medicine and Hygiene*. <https://doi.org/10.4269/ajtmh.1984.33.515>.
- Kelly, Steven, Chris Greenman, Peter R. Cook, and Argyris Papantonis. 2015. "Exon Skipping Is Correlated with Exon Circularization." *Journal of Molecular Biology*. <https://doi.org/10.1016/j.jmb.2015.02.018>.
- Lamm, Ayelet T., Michael R. Stadler, Huibin Zhang, Jonathan I. Gent, and Andrew Z. Fire. 2011. "Multimodal RNA-Seq Using Single-Strand, Double-Strand, and CircLigase-Based Capture Yields a Refined and Extended Description of the *C. Elegans* Transcriptome." *Genome Research*. <https://doi.org/10.1101/gr.108845.110>.
- Lasda, Erika, and Roy Parker. 2016. "Circular RNAs Co-Precipitate with Extracellular Vesicles: A Possible Mechanism for Circrna Clearance." *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0148407>.
- Li, Zhaoyong, Chuan Huang, Chun Bao, Liang Chen, Mei Lin, Xiaolin Wang, Guolin Zhong, et al. 2015. "Exon-Intron Circular RNAs Regulate Transcription in the Nucleus." *Nature Structural and Molecular Biology*. <https://doi.org/10.1038/nsmb.2959>.
- Liu, Degao, Ritesh Mewalal, Rongbin Hu, Gerald A. Tuskan, and Xiaohan Yang. 2017. "New Technologies Accelerate the Exploration of Non-Coding RNAs in Horticultural Plants." *Horticulture Research*. <https://doi.org/10.1038/hortres.2017.31>.
- Loftus, Brendan, Iain Anderson, Rob Davies, U. Cecilia M. Alsmark, John Samuelson,

- Paolo Amedeo, Paola Roncaglia, et al. 2005. "The Genome of the Protist Parasite *Entamoeba Histolytica*." *Nature*. <https://doi.org/10.1038/nature03291>.
- Long, Yicheng, Xueyin Wang, Daniel T Youmans, and Thomas R Cech. 2017. "G E N E E X P R E S S I O N How Do LncRNAs Regulate Transcription?"
- Mar-Aguilar, Fermín, Victor Trevino, Jannet E. Salinas-Hernández, Marcela M. Taméz-Guerrero, María P. Barrón-González, Eufemia Morales-Rubio, Jaime Treviño-Neávez, Jorge A. Verduzco-Martínez, Mario R. Morales-Vallarta, and Diana Reséndez-Pérez. 2013. "Identification and Characterization of MicroRNAs from *Entamoeba Histolytica* HM1-IMSS." *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0068202>.
- Martinez-Palomo, A., A. Gonzalez-Robles, and M. De La Torre. 1973. "Selective Agglutination of Pathogenic Strains of *Entamoeba Histolytica* Induced Con A." *Nature New Biology*. <https://doi.org/10.1038/newbio245186a0>.
- Memczak, Sebastian, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, et al. 2013. "Circular RNAs Are a Large Class of Animal RNAs with Regulatory Potency." *Nature*. <https://doi.org/10.1038/nature11928>.
- Mendoza-Figueroa, María S., Eddy E. Alfonso-Maqueira, Cristina Vélez, Elisa I. Azuara-Liceaga, Selene Zárate, Nicolás Villegas-Sepúlveda, Odila Saucedo-Cárdenas, and Jesús Valdés. 2018. "Postsplicing-Derived Full-Length Intron Circles in the Protozoan Parasite *Entamoeba Histolytica*." *Frontiers in Cellular and Infection Microbiology*. <https://doi.org/10.3389/fcimb.2018.00255>.
- Meyer, Martin, Helena Fehling, Jenny Matthiesen, Stephan Lorenzen, Kathrin Schuldt, Hannah Bernin, Mareen Zaruba, et al. 2016. "Overexpression of Differentially Expressed Genes Identified in Non-Pathogenic and Pathogenic *Entamoeba Histolytica* Clones Allow Identification of New Pathogenicity Factors Involved in Amoebic Liver Abscess Formation." *PLoS Pathogens*. <https://doi.org/10.1371/journal.ppat.1005853>.
- Nozaki, Tomoyoshi, and Alok Bhattacharya. 2015. *Amebiasis: Biology and Pathogenesis of Entamoeba*. *Amebiasis: Biology and Pathogenesis of Entamoeba*. <https://doi.org/10.1007/978-4-431-55200-0>.

- Pamudurti, Nagarjuna Reddy, Osnat Bartok, Marvin Jens, Reut Ashwal-Fluss, Christin Stottmeister, Larissa Ruhe, Mor Hanan, et al. 2017. "Translation of CircRNAs." *Molecular Cell*. <https://doi.org/10.1016/j.molcel.2017.02.021>.
- Pearson, Richard J., and Upinder Singh. 2010. "Approaches to Characterizing Entamoeba Histolytica Transcriptional Regulation." *Cellular Microbiology*. <https://doi.org/10.1111/j.1462-5822.2010.01524.x>.
- Ren, Xiaoxia, Yongxing Du, Lei You, and Yupei Zhao. 2017. "Potential Functions and Implications of Circular RNA in Gastrointestinal Cancer." *Oncology Letters*. <https://doi.org/10.3892/ol.2017.7118>.
- Rivera, Maria C., and James A. Lake. 2004. "The Ring of Life Provides Evidence for a Genome Fusion Origin of Eukaryotes." *Nature*. <https://doi.org/10.1038/nature02848>.
- Sanchez, Lidya, Vincenzo Enea, and Daniel Eichinger. 1994. "Identification of a Developmentally Regulated Transcript Expressed during Encystation of Entamoeba Invadens." *Molecular and Biochemical Parasitology*. [https://doi.org/10.1016/0166-6851\(94\)90102-3](https://doi.org/10.1016/0166-6851(94)90102-3).
- Sargeant, P. G., and J. E. Williams. 1978. "Electrophoretic Isoenzyme Patterns of Entamoeba Histolytica and Entamoeba Coli." *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [https://doi.org/10.1016/0035-9203\(78\)90053-6](https://doi.org/10.1016/0035-9203(78)90053-6).
- Sargeant, P. G., J. E. Williams, and J. D. Grene. 1978. "The Differentiation of Invasive and Non-Invasive Entamoeba Histolytica by Isoenzyme Electrophoresis." *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [https://doi.org/10.1016/0035-9203\(78\)90174-8](https://doi.org/10.1016/0035-9203(78)90174-8).
- Shi, Yigong. 2017. "Mechanistic Insights into Precursor Messenger RNA Splicing by the Spliceosome." *Nature Reviews Molecular Cell Biology*. <https://doi.org/10.1038/nrm.2017.86>.
- Song, Xiaofeng, Naibo Zhang, Ping Han, Byoung San Moon, Rose K. Lai, Kai Wang, and Wange Lu. 2016. "Circular RNA Profile in Gliomas Revealed by Identification Tool UROBORUS." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkw075>.
- Strachan, W. D., W. M. Spice, P. L. Chiodini, A. H. Moody, and J. P. Ackers. 1988. "IMMUNOLOGICAL DIFFERENTIATION OF PATHOGENIC AND NON-PATHOGENIC ISOLATES OF ENTAMOEBIA HISTOLYTICA." *The Lancet*.

- [https://doi.org/10.1016/S0140-6736\(88\)91355-4](https://doi.org/10.1016/S0140-6736(88)91355-4).
- Szabo, Linda, Robert Morey, Nathan J. Palpant, Peter L. Wang, Nastaran Afari, Chuan Jiang, Mana M. Parast, Charles E. Murry, Louise C. Laurent, and Julia Salzman. 2015. "Statistically Based Splicing Detection Reveals Neural Enrichment and Tissue-Specific Induction of Circular RNA during Human Fetal Development." *Genome Biology*. <https://doi.org/10.1186/s13059-015-0690-5>.
- Szabo, Linda, and Julia Salzman. 2016. "Detecting Circular RNAs: Bioinformatic and Experimental Challenges." *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg.2016.114>.
- Tovy, Ayala, Rivka Hertz, Rama Siman-Tov, Sylvie Syan, Daniela Faust, Nancy Guillen, and Serge Ankri. 2011. "Glucose Starvation Boosts *Entamoeba histolytica* Virulence." *PLoS Neglected Tropical Diseases*. <https://doi.org/10.1371/journal.pntd.0001247>.
- Venkataraman, Krithika, Kip E. Guja, Miguel Garcia-Diaz, and A. Wali Karzai. 2014. "Non-Stop mRNA Decay: A Special Attribute of Trans-Translation Mediated Ribosome Rescue." *Frontiers in Microbiology*. <https://doi.org/10.3389/fmicb.2014.00093>.
- Wang, Peter L., Yun Bao, Muh Ching Yee, Steven P. Barrett, Gregory J. Hogan, Mari N. Olsen, José R. Dinneny, Patrick O. Brown, and Julia Salzman. 2014. "Circular RNA Is Expressed across the Eukaryotic Tree of Life." *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0090859>.
- Weber, Christian, Mikael Koutero, Marie Agnes Dillies, Hugo Varet, Cesar Lopez-Camarillo, Jean Yves Coppée, Chung Chau Hon, and Nancy Guillén. 2016. "Extensive Transcriptome Analysis Correlates the Plasticity of *Entamoeba histolytica* Pathogenesis to Rapid Phenotype Changes Depending on the Environment." *Scientific Reports*. <https://doi.org/10.1038/srep35852>.
- Weedall, Gareth D, C Graham Clark, Pia Koldkjaer, Suzanne Kay, Iris Bruchhaus, Egbert Tannich, Steve Paterson, and Neil Hall. 2012. "Genomic Diversity of the Human Intestinal Parasite *Entamoeba histolytica*." *Genome Biology*. <https://doi.org/10.1186/gb-2012-13-5-r38>.
- "Weekly Epidemiological Record." 2017. *Japanese Journal of Leprosy* JAPANESE

- JOURNAL OF LEPROSY*. <https://doi.org/10.5025/hansen.85.157>.
- Xie, Binhui, Zhenxian Zhao, Qingquan Liu, Xiaonong Wang, Zhenjiang Ma, and Heping Li. 2019. "CircRNA Has_circ_0078710 Acts as the Sponge of MicroRNA-31 Involved in Hepatocellular Carcinoma Progression." *Gene*. <https://doi.org/10.1016/j.gene.2018.10.043>.
- Yang, Yun, Xiaojuan Fan, Miaowei Mao, Xiaowei Song, Ping Wu, Yang Zhang, Yongfeng Jin, et al. 2017. "Extensive Translation of Circular RNAs Driven by N⁶-Methyladenosine." *Cell Research*. <https://doi.org/10.1038/cr.2017.31>.
- Ye, Chu Yu, Xingchen Zhang, Qinjie Chu, Chen Liu, Yongyi Yu, Weiqin Jiang, Qian Hao Zhu, Longjiang Fan, and Longbiao Guo. 2017. "Full-Length Sequence Assembly Reveals Circular RNAs with Diverse Non-GT/AG Splicing Signals in Rice." *RNA Biology*. <https://doi.org/10.1080/15476286.2016.1245268>.
- Zhang, Hanbang, Gretchen M. Ehrenkaufner, Neil Hall, and Upinder Singh. 2013. "Small RNA Pyrosequencing in the Protozoan Parasite *Entamoeba Histolytica* Reveals Strain-Specific Small RNAs That Target Virulence Genes." *BMC Genomics*. <https://doi.org/10.1186/1471-2164-14-53>.
- Zhang, Hanbang, Gretchen M. Ehrenkaufner, Dipak Manna, Neil Hall, and Upinder Singh. 2015. "High Throughput Sequencing of *Entamoeba* 27nt Small RNA Population Reveals Role in Permanent Gene Silencing but No Effect on Regulating Gene Expression Changes during Stage Conversion, Oxidative, or Heat Shock Stress." *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0134481>.
- Zhang, Xiao Ou, Hai Bin Wang, Yang Zhang, Xuhua Lu, Ling Ling Chen, and Li Yang. 2014. "Complementary Sequence-Mediated Exon Circularization." *Cell*. <https://doi.org/10.1016/j.cell.2014.09.001>.
- Zhang, Yang, Xiao Ou Zhang, Tian Chen, Jian Feng Xiang, Qing Fei Yin, Yu Hang Xing, Shanshan Zhu, Li Yang, and Ling Ling Chen. 2013. "Circular Intronic Long Noncoding RNAs." *Molecular Cell*. <https://doi.org/10.1016/j.molcel.2013.08.017>.
- Zheng, Yi, Peifeng Ji, Shuai Chen, Lingling Hou, and Fangqing Zhao. 2019. "Reconstruction of Full-Length Circular RNAs Enables Isoform-Level Quantification." *Genome Medicine*. <https://doi.org/10.1186/s13073-019-0614-1>.
- Zhou, Chan, Benoit Molinie, Kaveh Daneshvar, Joshua V. Pondick, Jinkai Wang,

Nicholas Van Wittenberghe, Yi Xing, Cosmas C. Giallourakis, and Alan C. Mullen. 2017. "Genome-Wide Maps of M6A CircRNAs Identify Widespread and Cell-Type-Specific Methylation Patterns That Are Distinct from MRNAs." *Cell Reports*. <https://doi.org/10.1016/j.celrep.2017.08.027>.

Zhou, Tong, Xueying Xie, Musheng Li, Junchao Shi, Jin J. Zhou, Kenneth S. Knox, Ting Wang, Qi Chen, and Wanjun Gu. 2018. "Rat BodyMap Transcriptomes Reveal Unique Circular RNA Features across Tissue Types and Developmental Stages." *RNA*. <https://doi.org/10.1261/rna.067132.118>.

13. Apéndice

Apéndice A. Lista de circRNAs con más de un elemento interno (Los amarillos son circRNAs que están expresados diferencialmente)

CircID contig:start end	Gene Parental	Locus de los elementos internos del circRNA	CircID contig:start end	Gene Parental	Locus de los elementos internos del circRNA
DS571186:1600 1788	EHI_000240	1600-1657,1717-1788,	DS571208:14489 35573	EHI_030760	14489-14494,35477-35573,
DS571186:1639 1788	EHI_000240	1639-1644,1717-1788,	DS571208:35400 35575	EHI_030800	35400-35443,35570-35575,
DS571186:1623 1809	EHI_000240	1623-1657,1717-1809,	DS571252:36071 36248	EHI_033710	36071-36127,36243-36248,
DS571186:1621 1788	EHI_000240	1621-1657,1717-1788,	DS571190:19758 19957	EHI_038680	19758-19858,19912-19957,
DS571197:44686 44880	EHI_006880	44686-44778,44830-44880,	DS571224:19773 19953	EHI_040310	19773-19863,19948-19953,
DS571200:45997 46195	EHI_010020	45997-46042,46107-46195,	DS571578:2205 2368	EHI_040820	2205-2243,2363-2368,
DS571168:92541 92882	EHI_013220	92541-92599,92834-92882,	DS571370:6373 6567	EHI_053450	6373-6408,6463-6567,
DS571286:30182 30380	EHI_026410	30182-30267,30320-30380,	DS571313:26919 27096	EHI_056490	26919-26924,27033-27096,
DS571286:30182 30393	EHI_026410	30182-30267,30320-30393,	DS571301:28271 28420	EHI_060420	28271-28291,28415-28420,
DS571155:128498 128673	EHI_092580	128498-128619,128668-128673,	DS571374:3123 3361	EHI_146920	3123-3299,3356-3361,
DS571185:37588 37725	EHI_104490	37588-37593,37654-37725,	DS571145:250028 250099	EHI_152360	250028-250073,250079-250099,
DS571976:1246 1560	EHI_116360	1246-1251,1414-1560,	DS571145:321292 321434	EHI_152760	321292-321328,321405-321434,
DS571976:1144 1464	EHI_116360	1144-1149,1414-1464,	DS571145:321292 321538	EHI_152760	321292-321328,321405-321538,
DS571281:25995 26177	EHI_122870	25995-26060,26133-26177,	DS571145:359827 399020	EHI_152970	359827-359931,399015-399020,

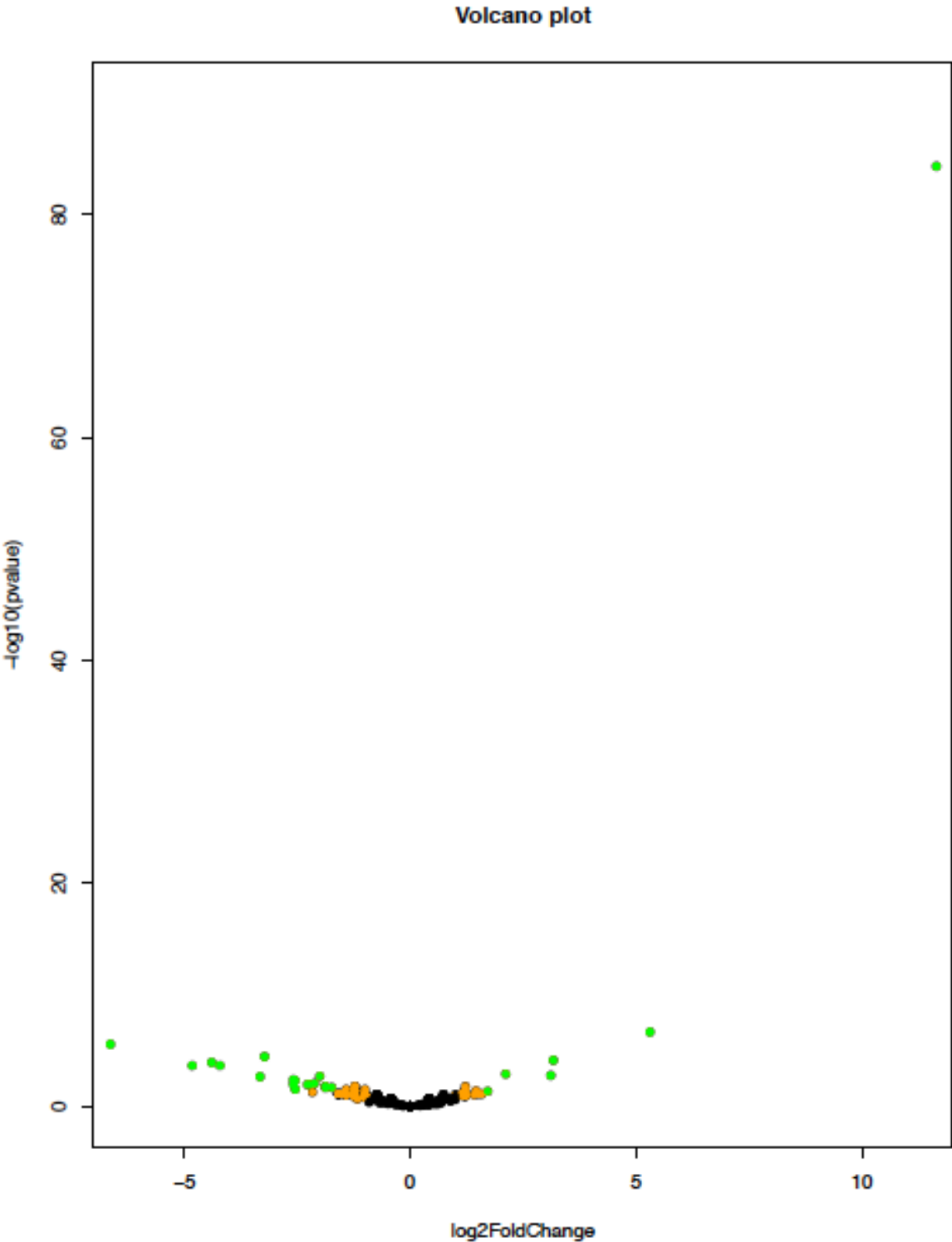
CircID contig:start end	Gene Parental	Locus de los componentes internos del circRNA
DS571145:383172 462091	EHI_153080	383172-383244,462086-462091,
DS571145:404776 404972	EHI_153190	404776-404807,404871-404972,
DS571145:452405 452566	EHI_153480	452405-452424,452476-452566,
DS571377:17778 18245	EHI_169670	17778-17836,18195-18245,
DS571377:17778 18245	EHI_169670	17778-17827,18195-18245,
DS571377:17778 18241	EHI_169670	17778-17836,18236-18241,
DS571175:48265 48427	EHI_177380	48265-48351,48405-48427,
DS571175:50415 50584	EHI_177400	50415-50454,50510-50584,
DS571431:15130 15298	EHI_179390	15130-15193,15249-15298,
DS571436:10774 11054	EHI_184250	10774-10858,10946-11054,
DS571159:12325 12529	EHI_186340	12325-12427,12505-12529,
DS571424:12489 12740	EHI_190450	12489-12510,12569-12740,
DS571317:3869 4090	EHI_191900	3869-3964,4019-4090,
DS571191:16879 17058	EHI_194220	16879-16884,16988-17058,
DS571250:37477 37630	EHI_195060	37477-37546,37607-37630,

Apéndice B. Lista de los 39 circRNAs expresados diferencialmente

CircID	log2FoldChange	pvalue	padj
DS571170:16194 16361	-6.6345524902929	0.0000027533328	0.0000915483172
DS571557:235 392	-4.8248974562402	0.0002243343332	0.0033151629245
DS571358:9721 9921	-4.3926591532429	0.0001173150708	0.0022289863445
DS571244:23861 24043	-4.2100844385459	0.0002224878424	0.0033151629245
DS571170:15984 16124	-3.3222620544095	0.0022350710210	0.0228664958304
DS571398:16736 16881	-3.2228081646782	0.0000349883709	0.0009306906653
DS571215:8591 8752	-2.5853062112851	0.0042099022226	0.0399940711149
DS571231:33971 34111	-2.5849944549704	0.0111306767532	0.0786980686251
DS571244:24172 24328	-2.5505485857624	0.0048461626641	0.0429693089549
DS571265:30984 31142	-2.5442417293395	0.0298242181944	0.1486474226062
DS571160:76317 76466	-2.2733938861738	0.0112425812322	0.0786980686251
DS571185:66039 66198	-2.2228032517662	0.0110003582495	0.0786980686251
DS571235:30353 30515	-2.1617565468121	0.0537015364665	0.1879553776326
DS571152:69791 69970	-2.1157385232258	0.0088757166556	0.0737793947000
DS571314:18860 19004	-2.0003808339792	0.0020164194551	0.0223486489606
DS571222:36290 36451	-1.8748944526808	0.0160193647245	0.0967797346201
DS571145:91445 91585	-1.8747688399879	0.0215101933308	0.1144342285200
DS571324:14068 14208	-1.7371314528354	0.0204020062983	0.1130611182367
DS571145:517901 518030	-1.4153116427040	0.0303136634759	0.1486474226062
DS571186:1621 1788	-1.2230398222097	0.0442493300026	0.1623073757975
DS571213:10608 10760	-1.2228620196020	0.0167147133281	0.0967797346201
DS571387:5087 5245	-1.2228339447027	0.0362253387083	0.1609866945823
DS571190:19758 19957	-1.2227825177493	0.0167198340593	0.0967797346201
DS571145:250019 250114	-1.2226848748435	0.0443215444399	0.1623073757975
DS571191:16879 17058	-1.2226848748435	0.0443215444399	0.1623073757975
DS571154:19722 19913	-1.2226088797518	0.0443208228563	0.1623073757975
DS571145:445852 445989	-1.0003770949183	0.0312941942329	0.1486474226062
DS571145:381681 381748	1.1630803183242	0.0597443640009	0.2037436003106
DS571377:17837 17993	1.2224281435579	0.0443430431296	0.1623073757975
DS571159:48214 48408	1.2225501784495	0.0363127882516	0.1609866945823
DS571146:38576 38725	1.2225794343900	0.0167363450847	0.0967797346201
DS571145:350423 350484	1.4592672299268	0.0410779874134	0.1623073757975
DS571145:151614 151702	1.7159503755121	0.0451531797331	0.1623073757975
DS571244:24189 24328	2.1156136405741	0.0012862196576	0.0171067214463
DS571286:30182 30393	3.1154393871798	0.0017181267006	0.0207737137435

DS571172:69095 69235	3.1700653408583	0.0000739936869	0.0016401933919
DS571377:17778 18245_C	5.3099842341243	0.0000002176756	0.0000096502857
DS571377:17778 18245_B	11.6486705514678	0.0000000000000	0.0000000000000
DS571377:17778 18245_A	13.2197013635409	0.0000000000000	0.0000000000000

Apéndice C. Gráfica de volcán abarcando los 39 circRNAs expresados diferencialmente





Network Stats

number of nodes: 336	expected number of edges: 3400
number of edges: 3595	PPI enrichment p-value: 0.000467
average node degree: 21.4	<i>your network has significantly more interactions than expected (what does that mean?)</i>
avg. local clustering coefficient: 0.322	

Functional enrichments in your network

Molecular Function (GO)




GO-term	description	count in gene set	false discovery rate	
GO:0005488	binding	10 of 59	0.0178	
GO:0005515	protein binding	4 of 9	0.0324	

KEGG Pathways

pathway	description	count in gene set	false discovery rate
ehi03010	Ribosome	21 of 135	4.05e-05
ehi00010	Glycolysis / Gluconeogenesis	8 of 24	0.00063
ehi00500	Starch and sucrose metabolism	6 of 17	0.0034
ehi01110	Biosynthesis of secondary metabolites	12 of 86	0.0058
ehi01100	Metabolic pathways	24 of 266	0.0058
ehi01130	Biosynthesis of antibiotics	10 of 71	0.0105
ehi05146	Amoebiasis	8 of 54	0.0205
ehi03008	Ribosome biogenesis in eukaryotes	9 of 72	0.0281
ehi00040	Pentose and glucuronate interconversions	3 of 7	0.0324

(less ...)

UniProt Keywords

keyword	description	count in gene set	false discovery rate	
KW-0689	Ribosomal protein	21 of 143	0.00019	
KW-0009	Actin-binding	6 of 17	0.0070	
KW-0479	Metal-binding	24 of 257	0.0080	
KW-0411	Iron-sulfur	4 of 6	0.0084	
KW-0521	NADP	3 of 5	0.0421	
KW-0251	Elongation factor	4 of 12	0.0421	
KW-0547	Nucleotide-binding	30 of 425	0.0465	

(less ...)

Apéndice E. Red e información de otros tipos de enriquecimiento funcional diferentes a KEGG en la red de interacciones entre proteínas/genes parentales ortólogos de *E. histolytica* de los circRNAs identificados en *E. invadens* con una red de interacciones con una confianza baja del 0.150



Network Stats

number of nodes: 88
number of edges: 331
average node degree: 7.52
avg. local clustering coefficient: 0.419

expected number of edges: 323
PPI enrichment p-value: 0.338
your network does not have significantly more interactions than expected (what does that mean?)

Functional enrichments in your network

KEGG Pathways

pathway	description	count in gene set	false discovery rate
ehi03010	Ribosome	7 of 135	0.0167

UniProt Keywords

keyword	description	count in gene set	false discovery rate
KW-0689	Ribosomal protein	8 of 143	0.0072
KW-0687	Ribonucleoprotein	9 of 170	0.0072
KW-0175	Coiled coil	37 of 2063	0.0106
KW-0670	Pyruvate	2 of 3	0.0167
KW-0009	Actin-binding	3 of 17	0.0167

(more ...)

PFAM Protein Domains

domain	description	count in gene set	false discovery rate
PF00164	Ribosomal protein S12/S23	3 of 4	0.0040
PF14429	C2 domain in Dock180 and Zizimin proteins	3 of 10	0.0155
PF06920	Dock homology region 2	3 of 13	0.0197

Apéndice F. Lista de circRNAs detectados en el trabajo de Weber y cols. (2016).
En amarillo son los circRNAs en común con los experimentos de RNA-seq de Hon y cols. (2013)

circRNA_ID
DS571265:30984 31142
DS571318:8455 8661
DS571151:142149 142301
DS571153:94757 94905
DS571151:141726 141938
DS571151:142162 142301
DS571152:69653 132393
DS571153:94906 139081
DS571159:22414 101760
DS571166:4029 43908
DS571167:48830 80209
DS571170:15434 75812
DS571177:56242 56380
DS571178:17905 18100
DS571183:9038 19890
DS571209:23199 23370
DS571209:23199 23370
DS571318:8557 23235

Apéndice G. IDs de CircRNAs detectados en las genotecas de *Entamoeba histolytica* HM1:IMSS

	<u>_05</u>	<u>_06</u>	<u>_07</u>
1	DS571145:112984 113118	DS571145:151614 151667	DS571145:107599 107719
2	DS571145:151614 151702	DS571145:151614 151702	DS571145:121260 121370
3	DS571145:16567 16641	DS571145:151836 151934	DS571145:151836 151934
4	DS571145:175932 176039	DS571145:174848 174904	DS571145:16554 16629
5	DS571145:18277 18354	DS571145:216972 217082	DS571145:186938 187052
6	DS571145:186821 186891	DS571145:226073 226133	DS571145:189761 189858
7	DS571145:186938 187018	DS571145:230398 230460	DS571145:215665 215767
8	DS571145:186938 187052	DS571145:244220 244297	DS571145:216972 217082
9	DS571145:186950 187018	DS571145:250019 250114	DS571145:307058 307153
10	DS571145:213515 213611	DS571145:250028 250099	DS571145:313386 313467
11	DS571145:215665 215767	DS571145:250341 250454	DS571145:320536 320619
12	DS571145:221937 222046	DS571145:265559 265648	DS571145:350423 350517
13	DS571145:230398 230460	DS571145:285533 285737	DS571145:381650 381744
14	DS571145:244220 244297	DS571145:313386 313467	DS571145:381681 381748
15	DS571145:249977 250078	DS571145:320536 320619	DS571145:381706 381778
16	DS571145:249977 250099	DS571145:350423 350484	DS571145:383103 383225
17	DS571145:250019 250114	DS571145:350423 350535	DS571145:421932 422027
18	DS571145:27299 27409	DS571145:351829 351936	DS571145:422028 422106
19	DS571145:27830 27901	DS571145:381650 381744	DS571145:445852 445989
20	DS571145:285369 285532	DS571145:381681 381748	DS571145:448780 448854
21	DS571145:285482 285580	DS571145:381706 381778	DS571145:452252 452368
22	DS571145:285533 285737	DS571145:383172 383280	DS571145:45500 45586
23	DS571145:307058 307154	DS571145:408394 408534	DS571145:48775 48844
24	DS571145:313386 313467	DS571145:421932 422029	DS571145:487765 487890
25	DS571145:321292 321434	DS571145:422028 422106	DS571145:49298 49441
26	DS571145:338522 338597	DS571145:445852 445989	DS571145:49322 49441
27	DS571145:350423 350517	DS571145:445900 445989	DS571145:497094 497231
28	DS571145:361747 361908	DS571145:449166 449327	DS571145:517901 518030
29	DS571145:381650 381744	DS571145:449255 449327	DS571145:64214 64316
30	DS571145:381681 381748	DS571145:452230 452283	DS571145:64268 64375
31	DS571145:381706 381778	DS571145:452252 452368	DS571145:69619 69714
32	DS571145:404020 404093	DS571145:497475 497564	DS571145:73919 74027
33	DS571145:421932 422028	DS571145:503546 503682	DS571145:91445 91585
34	DS571145:422028 422106	DS571145:503713 503853	DS571146:38569 38725
35	DS571145:445852 445989	DS571145:517974 518030	DS571146:38572 38725
36	DS571145:445900 445989	DS571145:72433 72540	DS571148:153027 153166
37	DS571145:44885 44957	DS571146:60569 60739	DS571149:174825 174967

38	DS571145:452252 452368	DS571148:153027 153166	DS571151:141040 141186
39	DS571145:452405 452566	DS571149:146665 146818	DS571151:22538 22696
40	DS571145:458351 458453	DS571150:131087 131248	DS571152:69791 69970
41	DS571145:488235 488314	DS571151:145001 145153	DS571153:94757 94905
42	DS571145:49298 49441	DS571151:157461 157600	DS571154:115353 115525
43	DS571145:49322 49441	DS571151:39225 39392	DS571158:136019 136153
44	DS571145:504073 504213	DS571151:39393 39533	DS571158:40942 41109
45	DS571145:517901 518030	DS571152:8851 8989	DS571159:111043 111221
46	DS571145:72433 72540	DS571153:94757 94905	DS571160:48956 49099
47	DS571145:72835 72957	DS571154:11264 11453	DS571160:76317 76466
48	DS571145:91445 91585	DS571154:11300 11453	DS571163:126893 127031
49	DS571146:31945 32121	DS571154:11318 11453	DS571163:34232 34375
50	DS571147:19647 19823	DS571154:18666 18893	DS571166:13433 13588
51	DS571147:36435 36593	DS571154:19722 19913	DS571170:15984 16124
52	DS571149:103536 103697	DS571155:7230 7386	DS571172:40198 40341
53	DS571149:106861 107007	DS571155:7480 7637	DS571172:5905 6048
54	DS571149:175158 175298	DS571156:83064 83203	DS571172:8982 9124
55	DS571150:81862 82089	DS571157:17936 18106	DS571172:8982 9153
56	DS571151:141586 141758	DS571157:54474 54620	DS571175:5384 5527
57	DS571151:141804 141938	DS571159:70600 70742	DS571175:89542 89703
58	DS571151:142302 142544	DS571160:76317 76466	DS571179:19218 19378
59	DS571151:145001 145153	DS571162:107301 107468	DS571179:51556 51702
60	DS571151:39296 39455	DS571163:126868 127031	DS571179:92870 93025
61	DS571151:48094 48232	DS571163:126893 127031	DS571180:51221 51364
62	DS571152:103507 103665	DS571163:36940 37115	DS571181:10793 10932
63	DS571152:69791 69970	DS571163:47158 47334	DS571182:21784 21922
64	DS571152:93773 93919	DS571164:100317 100489	DS571182:8883 9059
65	DS571153:40248 40391	DS571164:71475 71622	DS571183:37865 38000
66	DS571153:67334 67477	DS571164:9684 9819	DS571185:66039 66179
67	DS571153:94657 94806	DS571168:92541 92882	DS571186:1600 1788
68	DS571153:94757 94905	DS571169:101705 101857	DS571186:1621 1788
69	DS571154:11273 11453	DS571170:15984 16124	DS571186:1623 1809
70	DS571154:11291 11453	DS571170:16194 16361	DS571187:21500 21714
71	DS571154:11300 11453	DS571170:96096 96242	DS571190:19758 19957
72	DS571154:11318 11453	DS571171:18812 18961	DS571191:40634 40789
73	DS571154:18666 18848	DS571171:91609 91761	DS571191:42370 42506
74	DS571154:18894 19094	DS571172:8982 9132	DS571192:23679 23837
75	DS571154:19722 19913	DS571174:68364 68505	DS571192:61786 61929
76	DS571155:106607 106768	DS571179:17276 17427	DS571193:10432 10572
77	DS571155:145772 145930	DS571179:92870 93025	DS571198:20194 20337

78	DS571155:43479 43622	DS571181:10793 10932	DS571200:45997 46195
79	DS571155:71008 71156	DS571181:15063 15202	DS571200:5252 5391
80	DS571155:7480 7637	DS571181:37519 37686	DS571202:24801 24962
81	DS571156:104612 104763	DS571182:43016 43156	DS571206:41345 41503
82	DS571157:18859 19026	DS571183:38688 38891	DS571208:55922 56056
83	DS571158:133826 133966	DS571184:86213 86386	DS571209:59621 59763
84	DS571158:37319 37480	DS571185:11890 12114	DS571213:10608 10760
85	DS571158:40942 41109	DS571185:37588 37725	DS571215:8591 8752
86	DS571160:48887 49099	DS571185:66039 66198	DS571219:57650 57793
87	DS571160:48956 49099	DS571185:66127 66290	DS571222:44119 44256
88	DS571160:76317 76466	DS571185:70691 70902	DS571222:7402 7540
89	DS571160:84858 85010	DS571186:1621 1788	DS571226:36554 36691
90	DS571163:126893 127031	DS571186:1639 1788	DS571231:33971 34111
91	DS571163:36927 37103	DS571190:19758 19957	DS571232:46216 46398
92	DS571163:39840 40002	DS571191:16879 17058	DS571232:54127 54279
93	DS571163:47158 47334	DS571191:63017 63158	DS571233:37450 37593
94	DS571163:87278 87434	DS571193:49352 49534	DS571235:30353 30515
95	DS571164:71475 71622	DS571203:35344 35498	DS571236:50043 50186
96	DS571165:113186 113321	DS571206:40960 41097	DS571244:23861 24043
97	DS571166:14148 14305	DS571206:41345 41503	DS571244:24172 24328
98	DS571167:59790 59927	DS571207:43494 43655	DS571245:36344 36492
99	DS571168:92541 92882	DS571208:14489 35573	DS571246:32004 32154
100	DS571169:22785 22925	DS571209:61454 61591	DS571250:37477 37630
101	DS571169:46889 47028	DS571213:10608 10760	DS571251:29076 29246
102	DS571170:15984 16124	DS571218:46928 47110	DS571252:35757 35912
103	DS571170:16194 16361	DS571221:58566 58717	DS571252:35809 36017
104	DS571171:18812 18961	DS571222:36290 36451	DS571256:47804 47942
105	DS571173:77741 77920	DS571223:39635 39790	DS571260:36482 36637
106	DS571175:50415 50584	DS571224:19773 19953	DS571276:17000 17152
107	DS571175:78002 78173	DS571228:38011 38193	DS571289:23677 23817
108	DS571176:31613 31757	DS571231:13042 13209	DS571305:11742 11899
109	DS571179:39637 39780	DS571231:7088 7256	DS571312:19054 19202
110	DS571181:10793 10932	DS571235:30353 30515	DS571313:27109 27249
111	DS571183:38745 38897	DS571242:29052 29189	DS571314:11274 11453
112	DS571183:60631 60770	DS571243:12058 12204	DS571314:11289 11438
113	DS571184:51408 51547	DS571244:23861 24043	DS571314:18860 19004
114	DS571184:86213 86386	DS571244:24172 24328	DS571328:12668 12868
115	DS571185:19229 19426	DS571252:35757 35912	DS571328:12714 12868
116	DS571185:37588 37725	DS571252:36071 36248	DS571331:17521 17664
117	DS571185:66039 66198	DS571254:3561 3737	DS571342:14071 14232

118	DS571185:66039 66237	DS571258:31252 31386	DS571342:21713 21851
119	DS571185:66669 66866	DS571261:43348 43514	DS571347:17763 17918
120	DS571185:66792 66941	DS571265:30984 31142	DS571357:9333 9486
121	DS571186:17135 17314	DS571273:17741 17884	DS571358:4786 4936
122	DS571186:73502 73696	DS571274:23481 23648	DS571365:14233 14388
123	DS571187:19660 19848	DS571286:38638 38772	DS571376:8082 8234
124	DS571189:48581 48727	DS571287:28567 28704	DS571377:11587 11742
125	DS571190:19758 19957	DS571299:11915 12063	DS571377:17894 18058
126	DS571190:26912 27049	DS571299:27640 27805	DS571377:17952 18101
127	DS571190:37028 37172	DS571302:25503 25714	DS571387:5087 5245
128	DS571191:16879 17058	DS571313:7091 7237	DS571388:4166 4324
129	DS571191:62825 62962	DS571314:18860 19004	DS571398:16736 16881
130	DS571193:11071 11220	DS571317:3869 4090	DS571398:20761 20928
131	DS571194:37665 37808	DS571318:8455 8653	DS571404:18682 18855
132	DS571196:45037 45189	DS571319:27871 28048	DS571405:13124 13280
133	DS571196:54536 54688	DS571324:13488 13626	DS571405:13127 13280
134	DS571197:44686 44880	DS571324:14068 14208	DS571405:13140 13280
135	DS571199:38438 38625	DS571347:17763 17918	DS571418:16689 16845
136	DS571202:43288 43455	DS571353:7700 7839	DS571424:12489 12740
137	DS571205:25696 25883	DS571358:9721 9921	DS571431:14235 14377
138	DS571205:37961 38095	DS571387:17267 17430	DS571431:15130 15298
139	DS571206:40960 41097	DS571387:5087 5245	DS571451:2960 3117
140	DS571206:41194 41346	DS571394:16355 16526	DS571473:5957 6108
141	DS571206:41345 41503	DS571398:16600 16809	DS571547:2124 2276
142	DS571207:54439 54585	DS571398:16736 16881	DS571547:2133 2279
143	DS571208:35400 35575	DS571398:16882 17027	DS571547:2133 2339
144	DS571209:61454 61591	DS571402:7652 7798	DS571547:2299 2447
145	DS571213:10608 10760	DS571402:8885 9075	DS571557:235 392
146	DS571214:57616 57831	DS571404:18682 18855	DS571579:2752 2893
147	DS571215:8591 8752	DS571405:13124 13280	DS571609:8185 8346
148	DS571216:20037 20174	DS571405:13127 13280	DS571641:3153 3319
149	DS571222:36290 36451	DS571436:10774 11054	DS571891:2567 2752
150	DS571224:19773 19953	DS571438:13519 13700	DS571976:1193 1356
151	DS571225:9277 9441	DS571438:13525 13671	
152	DS571227:19549 19713	DS571484:2194 2393	
153	DS571227:37232 37376	DS571494:9970 10120	
154	DS571227:47860 47996	DS571506:4201 4356	
155	DS571231:13042 13209	DS571518:3788 3928	
156	DS571231:33971 34111	DS571519:10609 10826	
157	DS571232:54113 54279	DS571545:3758 3919	

158	DS571235:30353 30515	DS571547:2133 2339
159	DS571235:31320 31493	DS571561:5382 5537
160	DS571240:36752 36913	DS571574:7906 8046
161	DS571241:12424 12568	DS571579:2752 2893
162	DS571243:8495 8641	DS571699:817 995
163	DS571243:8550 8705	DS571976:1144 1464
164	DS571244:23861 24043	DS571976:1193 1332
165	DS571244:24172 24328	DS571976:1193 1368
166	DS571249:36977 37125	DS571976:1193 1392
167	DS571251:39399 39542	DS571976:1193 1404
168	DS571252:35838 36017	DS571976:1193 1428
169	DS571253:31443 31610	DS571976:1246 1560
170	DS571253:48611 48809	
171	DS571258:28249 28396	
172	DS571258:31001 31221	
173	DS571264:7494 7643	
174	DS571265:30984 31142	
175	DS571265:31443 31700	
176	DS571273:25241 25483	
177	DS571273:25323 25483	
178	DS571274:34576 34713	
179	DS571276:14560 14694	
180	DS571279:12469 12606	
181	DS571282:12501 12645	
182	DS571289:23677 23817	
183	DS571294:11156 11311	
184	DS571299:11915 12063	
185	DS571299:11930 12100	
186	DS571299:17226 17360	
187	DS571300:33577 33714	
188	DS571303:22093 22278	
189	DS571307:4869 5027	
190	DS571313:27109 27249	
191	DS571314:18860 19004	
192	DS571324:14068 14208	
193	DS571329:4047 4232	
194	DS571338:9135 9285	
195	DS571344:22966 23102	
196	DS571358:19617 19774	
197	DS571358:9721 9921	

198 DS571359:22738|22886
199 DS571384:22608|22775
200 DS571387:17368|17517
201 DS571388:4166|4324
202 DS571392:15088|15313
203 DS571396:11297|11441
204 DS571398:16736|16881
205 DS571402:7652|7798
206 DS571405:13124|13280
207 DS571408:14951|15091
208 DS571414:12731|12872
209 DS571416:17030|17168
210 DS571423:14207|14344
211 DS571453:11740|11882
212 DS571462:7553|7747
213 DS571525:8810|8965
214 DS571525:8822|8979
215 DS571557:235|392
216 DS571561:5382|5537
217 DS571579:2752|2893
218 DS571581:4163|4369
219 DS571691:5639|5782
220 DS571976:1193|1368
221 DS571976:1193|1392
222 DS571976:1193|1428
223 DS571976:1193|1440
224 DS571976:1193|1464
225 DS571976:1193|1560

Apéndice H. IDs de CircRNAs detectados en las genotecas de *Entamoeba Histolytica* Rahman.

_08	_09	_10	
DS571145:112984 113118	DS571145:107570 107704	DS571145:151614 151667	1
DS571145:151591 151667	DS571145:130925 131028	DS571145:151614 151702	2
DS571145:151614 151667	DS571145:151614 151667	DS571145:151635 151702	3
DS571145:151614 151702	DS571145:151614 151702	DS571145:151836 151934	4
DS571145:151836 151934	DS571145:151836 151934	DS571145:16554 16629	5
DS571145:16567 16641	DS571145:180858 180962	DS571145:16567 16641	6
DS571145:186938 187018	DS571145:186821 186891	DS571145:186821 186891	7
DS571145:215665 215767	DS571145:187064 187190	DS571145:186938 187018	8
DS571145:216972 217082	DS571145:215665 215767	DS571145:209377 209452	9
DS571145:221937 222046	DS571145:239942 240032	DS571145:216906 216971	10
DS571145:313386 313467	DS571145:249977 250063	DS571145:230398 230460	11
DS571145:339893 339964	DS571145:273110 273202	DS571145:239942 240032	12
DS571145:350423 350484	DS571145:284926 285001	DS571145:244220 244297	13
DS571145:350423 350517	DS571145:307058 307153	DS571145:285533 285737	14
DS571145:381650 381744	DS571145:313386 313467	DS571145:313386 313467	15
DS571145:381681 381748	DS571145:333026 333130	DS571145:320536 320619	16
DS571145:381681 381778	DS571145:338084 338200	DS571145:321292 321434	17
DS571145:381706 381778	DS571145:350423 350484	DS571145:321292 321538	18
DS571145:398147 398253	DS571145:367745 367827	DS571145:326005 326085	19
DS571145:408394 408534	DS571145:381650 381744	DS571145:338084 338200	20
DS571145:418909 419004	DS571145:381681 381748	DS571145:347795 347858	21
DS571145:421932 422027	DS571145:381691 381778	DS571145:350423 350484	22
DS571145:432156 432377	DS571145:381698 381817	DS571145:359827 399020	23
DS571145:449166 449327	DS571145:381706 381778	DS571145:361806 361908	24
DS571145:452252 452368	DS571145:383172 462091	DS571145:381650 381744	25
DS571145:49322 49441	DS571145:390595 390667	DS571145:381681 381748	26
DS571145:497475 497564	DS571145:404866 404978	DS571145:381681 381778	27
DS571145:49957 50036	DS571145:421932 422027	DS571145:381706 381778	28
DS571145:64268 64375	DS571145:422028 422106	DS571145:383172 383280	29
DS571146:38569 38725	DS571145:427950 428041	DS571145:404776 404972	30
DS571146:38572 38725	DS571145:445900 445989	DS571145:421932 422029	31
DS571146:38576 38725	DS571145:452252 452368	DS571145:421932 422106	32
DS571149:103536 103697	DS571145:452265 452368	DS571145:422028 422106	33
DS571150:120794 120930	DS571145:48775 48844	DS571145:445900 445989	34
DS571151:142149 142301	DS571145:488235 488314	DS571145:449255 449327	35
DS571151:39296 39455	DS571145:5763 5857	DS571145:452265 452368	36
DS571152:58968 59123	DS571145:6020 6086	DS571145:45428 45499	37

DS571153:71567 71725	DS571145:64268 64375	DS571145:487309 487410	38
DS571153:94757 94905	DS571145:9872 9961	DS571145:48775 48844	39
DS571154:11300 11453	DS571146:38569 38725	DS571145:488069 488127	40
DS571154:115571 115768	DS571146:38572 38725	DS571145:49242 49321	41
DS571155:128498 128673	DS571146:38576 38725	DS571145:49957 50036	42
DS571155:7480 7637	DS571149:152953 153141	DS571145:503785 503853	43
DS571156:123066 123207	DS571150:26441 26586	DS571145:52942 53054	44
DS571157:138976 139198	DS571151:11728 11877	DS571145:64214 64316	45
DS571158:39786 39938	DS571153:94616 94815	DS571145:64268 64375	46
DS571159:48214 48408	DS571153:94757 94905	DS571145:76095 76160	47
DS571160:76317 76466	DS571154:11300 11453	DS571145:8859 8948	48
DS571161:7377 7553	DS571154:61756 61897	DS571146:38569 38725	49
DS571162:47882 48069	DS571157:54474 54620	DS571146:38576 38725	50
DS571163:122225 122368	DS571158:45605 45744	DS571149:138731 138874	51
DS571163:126857 127024	DS571159:48804 48952	DS571149:174472 174658	52
DS571163:126893 127031	DS571159:59212 59350	DS571151:141586 141758	53
DS571163:34232 34375	DS571160:48890 49099	DS571151:141804 141938	54
DS571163:47158 47334	DS571160:48956 49099	DS571151:157629 157808	55
DS571165:113186 113321	DS571164:100329 100463	DS571152:8851 8989	56
DS571170:15984 16124	DS571164:71475 71622	DS571153:94636 94788	57
DS571171:18812 18961	DS571165:37965 38102	DS571153:94757 94905	58
DS571171:18818 18961	DS571168:92541 92882	DS571155:7480 7637	59
DS571171:91609 91761	DS571169:22785 22925	DS571159:12325 12529	60
DS571172:69095 69235	DS571169:98451 98602	DS571159:48214 48408	61
DS571179:51556 51702	DS571172:69095 69235	DS571160:48956 49099	62
DS571181:10793 10932	DS571172:8982 9132	DS571163:47158 47334	63
DS571186:74510 74662	DS571175:48265 48427	DS571163:69950 70090	64
DS571188:77541 77679	DS571181:10793 10932	DS571164:9684 9819	65
DS571189:55220 55363	DS571190:37025 37172	DS571168:92541 92882	66
DS571191:63017 63158	DS571192:51471 51650	DS571169:22785 22925	67
DS571194:41987 42165	DS571192:52563 52730	DS571172:69095 69235	68
DS571206:40960 41097	DS571199:58210 58350	DS571175:70990 71205	69
DS571206:41345 41503	DS571203:9820 9981	DS571179:8303 8457	70
DS571207:54439 54585	DS571206:41345 41503	DS571180:51221 51364	71
DS571209:59915 60064	DS571206:41496 41641	DS571181:10793 10932	72
DS571218:37251 37397	DS571207:54439 54585	DS571181:80164 80304	73
DS571218:46928 47110	DS571208:35384 35533	DS571181:81420 81569	74
DS571224:19773 19953	DS571209:36572 36715	DS571184:4933 5101	75
DS571225:9277 9441	DS571217:3476 3658	DS571185:37588 37725	76
DS571232:54127 54279	DS571218:46928 47110	DS571187:19660 19848	77

DS571235:30470 30655	DS571219:57624 57796	DS571206:40960 41097	78
DS571237:26221 26358	DS571219:57650 57793	DS571206:41194 41346	79
DS571238:31583 31720	DS571224:19773 19953	DS571206:41345 41503	80
DS571239:19157 19297	DS571226:36458 36634	DS571207:54439 54585	81
DS571243:12058 12204	DS571228:50580 50768	DS571208:35384 35542	82
DS571244:24172 24328	DS571231:7088 7256	DS571209:59764 59926	83
DS571244:24189 24328	DS571234:5749 5901	DS571209:61454 61591	84
DS571245:36794 36936	DS571235:12182 12325	DS571214:36016 36238	85
DS571251:39399 39542	DS571235:30353 30515	DS571219:57650 57793	86
DS571252:35757 35912	DS571237:17066 17284	DS571224:19773 19953	87
DS571252:35809 36017	DS571243:12058 12204	DS571231:33971 34111	88
DS571273:13881 14045	DS571244:23861 24043	DS571235:30353 30515	89
DS571276:17000 17152	DS571244:24189 24328	DS571236:49668 49808	90
DS571304:23033 23176	DS571246:2673 2831	DS571244:11497 11644	91
DS571312:19054 19202	DS571252:35685 35828	DS571244:24189 24328	92
DS571313:27109 27249	DS571252:35757 35912	DS571244:3852 3995	93
DS571359:22847 22985	DS571252:35809 36017	DS571248:16442 16612	94
DS571374:3123 3361	DS571286:30182 30393	DS571252:35757 35912	95
DS571377:17778 18245	DS571291:37470 37609	DS571252:35838 36017	96
DS571377:17778 18245	DS571301:28271 28420	DS571265:30984 31142	97
DS571377:17778 18245	DS571313:27109 27249	DS571278:21540 21692	98
DS571380:11056 11229	DS571316:11031 11185	DS571281:25995 26177	99
DS571388:4166 4324	DS571316:12747 12956	DS571286:30104 30255	100
DS571402:7652 7798	DS571324:13488 13626	DS571286:30182 30380	101
DS571405:13124 13280	DS571330:30522 30689	DS571286:30182 30393	102
DS571405:13127 13280	DS571344:23365 23514	DS571289:23677 23817	103
DS571405:784 933	DS571344:8262 8421	DS571295:23597 23738	104
DS571416:17030 17168	DS571350:19688 19877	DS571305:11742 11899	105
DS571431:9755 9900	DS571352:15570 15719	DS571313:26857 27009	106
DS571559:6680 6820	DS571372:3786 3983	DS571313:26919 27096	107
DS571561:5382 5537	DS571377:17778 18241	DS571318:8455 8661	108
DS571579:2752 2893	DS571377:17778 18245	DS571335:16401 16594	109
DS571976:1450 1632	DS571377:17778 18245	DS571341:17368 17538	110
	DS571377:17778 18245	DS571353:7700 7887	111
	DS571377:17837 17993	DS571355:24039 24247	112
	DS571377:17894 18052	DS571370:6373 6567	113
	DS571377:17952 18097	DS571377:17778 18245	114
	DS571377:17952 18115	DS571377:17778 18245	115
	DS571377:17994 18241	DS571377:17778 18245	116
	DS571387:19371 19537	DS571377:17837 17993	117

DS571402:7652 7798	DS571377:17952 18097	118
DS571404:17073 17219	DS571377:17952 18101	119
DS571405:13124 13280	DS571377:17994 18184	120
DS571429:17967 18163	DS571398:16600 16809	121
DS571431:5200 5373	DS571402:7652 7798	122
DS571547:2124 2276	DS571404:18682 18855	123
DS571547:2124 2339	DS571405:13127 13280	124
DS571561:5382 5537	DS571438:13355 13518	125
DS571579:2752 2893	DS571547:2124 2276	126
DS571976:1193 1332	DS571547:2124 2339	127
DS571976:1193 1368	DS571547:2133 2276	128
DS571976:1193 1392	DS571547:2184 2330	129
DS571976:1193 1560	DS571561:5382 5537	130
	DS571574:8047 8187	131
	DS571578:2205 2368	132
	DS571615:5710 5847	133
	DS571627:2199 2351	134
	DS571976:1193 1356	135
	DS571976:1486 1632	136

Apéndice I. *Script* en R de la expresión diferencial utilizando DESeq2.

```
#####  
#Programa: Análisis de expresión diferencial con DESeq2  
#Lugar: CINVESTAV  
#Fecha: 19 de abril 2019  
#Desarrollador: César Padrón  
#####  
library('DESeq2')  
directory <- "/Users/CristianPadron/Desktop/ED/ed1"  
setwd (directory)  
#usar grep para buscar los archivos  
sampleFiles <-grep('_conteo',list.files(directory),value=TRUE)  
sampleFiles  
#Asignar nombre de la condicion, cuadrar  
#con el orden de la linea 22 muy IMPORTANTE  
sampleCondition <-c('Vir','Vir','Vir','NoV','NoV','NoV')  
sampleTable <-data.frame (sampleName=sampleFiles,  
fileName=sampleFiles,  
condition=sampleCondition)  
sampleTable  
ddsHTSeq<-DESeqDataSetFromHTSeqCount(  
sampleTable=sampleTable,  
directory=directory, design=~condition)  
#La primera etiqueta corresponde al control  
colData(ddsHTSeq)$condition<-factor(colData(ddsHTSeq)$condition,  
levels=c('Vir','NoV'))  
dds<-DESeq(ddsHTSeq)  
res<-results(dds)  
res<-res[order(res$padj),]  
head(res)  
#Resumen de resultados  
summary(dds)
```

```

#Guardadando
write.csv(res, file="Tabla_ED_crudos.csv")
resOrdered <- res[order(res$pvalue),]
summary(res)
#####
# Algunos graficos
#####
###MA plot
plotMA(dds,ylim=c(-1.0,1.0),main='DESeq2')
dev.copy (png,'deseq2_MAprt.png')
dev.off()
###PCA
rld<- rlogTransformation(dds, blind=TRUE)
plotPCA(rld, intgroup=c('condition'))
# basic volcanoplot
head(res)
# Make a basic volcano plot
with(res, plot(log2FoldChange, -log10(pvalue), pch=20, main="Volcano plot",
xlim=c(-15,15)))
# Add colored points: red if padj<0.05, orange if log2FC>1, green if both
with(subset(res, padj<.05 ), points(log2FoldChange, -log10(pvalue), pch=20,
col="red"))
with(subset(res, abs(log2FoldChange)>1.2), points(log2FoldChange, -
log10(pvalue), pch=20, col="orange"))
with(subset(res, padj<.05 & abs(log2FoldChange)>1.2), points(log2FoldChange, -
log10(pvalue), pch=20, col="green"))

```

Apéndice J . Generación de gráfico con línea de regresión local.

R code:

```
data1 <- read.csv(file.choose(), header=T)
```

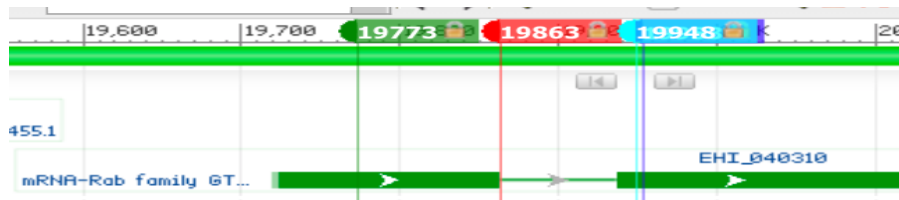
```
names(data1) <- c("Xvar", "Yvar")
```

```
plot(data1, log="xy", col= "blue")
```

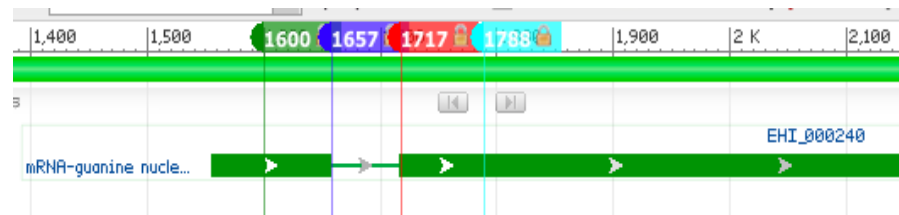
```
scatter.smooth(x=data1$Xvar, y=data1$Yvar, col= "blue", pch=20,,  
log="xy", xlim=c(0.00001,100000), ylim=c(0.00001,100000), lpars =list (col = "red", lwd =  
2, lty = 2))
```

Apéndice J. Mapa genómico de algunos circRNAs conformados por dos elementos internos

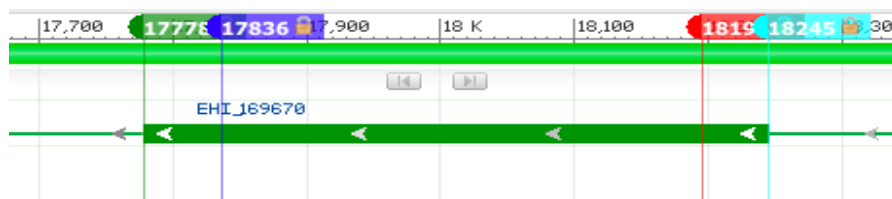
DS571224:19773|19953



DS571186:1621|1788 EHI_000240 1621-1657,1717-1788, ED.



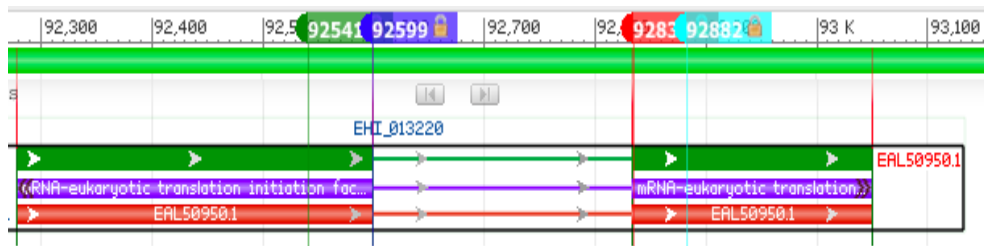
DS571377:17778|18245_B EHI_169670 17778-17836,18195-18245,



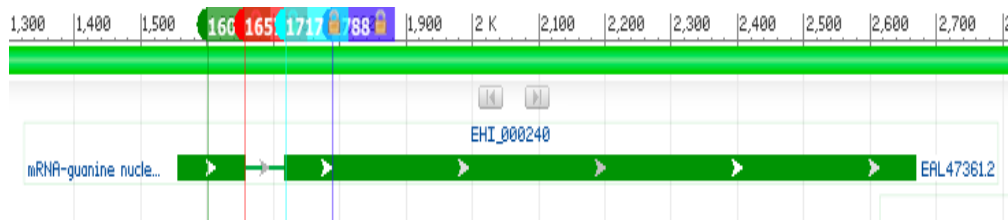
DS571377:17778|18245_C EHI_169670 17778-17827,18195-18245,



DS571168:92541|92882 EHI_013220 92541-92599,92834-92882,



DS571186:1600|1788



Apéndice K. Genes ortólogos de *E. histolytica* de los genes parentales de los circRNAs detectados en *E. invadens*.

	<i>E. invadens</i>	<i>E. histolytica</i>	<i>E. invadens</i>	<i>E. histolytica</i>
1	EIN_005670	EHI_053830	EIN_173710	EHI_093580
2	EIN_031580	EHI_030890	EIN_187940	EHI_118840
3	EIN_032050	EHI_097980	EIN_215120	EHI_040380
4	EIN_043960	EHI_058330	EIN_222920	EHI_110740
5	EIN_044390	EHI_054800	EIN_222960	EHI_150470
6	EIN_047670	EHI_010670	EIN_223290	EHI_110740
7	EIN_065890	EHI_166920	EIN_223290	EHI_110740
8	EIN_065940	EHI_178990	EIN_223290	EHI_110740
9	EIN_066080	EHI_015750	EIN_223290	EHI_110740
10	EIN_092220	EHI_198930	EIN_224250	EHI_023230
11	EIN_093520	EHI_052790	EIN_226130	EHI_148220
12	EIN_093540	EHI_049640	EIN_228310	EHI_058110
13	EIN_095510	EHI_188730	EIN_229670	EHI_044770
14	EIN_095510	EHI_188730	EIN_230150	EHI_135040
15	EIN_095760	EHI_147440	EIN_231040	EHI_188170
16	EIN_114020	EHI_155160	EIN_248100	EHI_178460
17	EIN_117910	EHI_200800	EIN_251130	EHI_025090
18	EIN_129260	EHI_111550	EIN_269080	EHI_108500
19	EIN_129500	EHI_069110	EIN_269080	EHI_020310
20	EIN_135020	EHI_006810	EIN_269080	EHI_040830
21	EIN_135020	EHI_006810	EIN_269080	EHI_081320
22	EIN_145900	EHI_120640	EIN_273630	EHI_163260
23	EIN_152450	EHI_143650	EIN_274510	EHI_051060
24	EIN_153430	EHI_045080	EIN_281250	EHI_201820
25	EIN_153430	EHI_045080	EIN_283750	EHI_170470
26	EIN_153430	EHI_045080	EIN_284550	EHI_141090
27	EIN_155450	EHI_164070	EIN_284550	EHI_141090
28	EIN_162170	EHI_140240	EIN_284890	EHI_044760
29	EIN_162220	EHI_114400	EIN_284940	EHI_092690
30	EIN_171040	EHI_147010	EIN_284970	EHI_167320

<i>E. invadens</i>	<i>E. histolytica</i>	<i>E. invadens</i>	<i>E. histolytica</i>	
EIN_306920	EHI 109900	EIN_391640	EHI 169670	1
EIN_308160	EHI 085970	EIN_398130	EHI 147010	2
EIN_309600	EHI 125150	EIN_398310	EHI 024650	3
EIN_327660	EHI 182620	EIN_403300	EHI 110180	4
EIN_327940	EHI 147540	EIN_403630	EHI 022980	5
EIN_327940	EHI 147540	EIN_403770	EHI 065630	6
EIN_327940	EHI 147540	EIN_405260	EHI 005150	7
EIN_327940	EHI 147540	EIN_407990	EHI 136410	8
EIN_327940	EHI 147540	EIN_409300	EHI 064710	9
EIN_335410	EHI 082560	EIN_409670	EHI 064710	10
EIN_335470	EHI 001780	EIN_411070	EHI 021190	11
EIN_335470	EHI 001780	EIN_411070	EHI 021190	12
EIN_337640	EHI 120630	EIN_416970	EHI 104630	13
EIN_359870	EHI 111060	EIN_424210	EHI 174070	14
EIN_369450	EHI 152970	EIN_428080	EHI 013180	15
EIN_369710	EHI 151850	EIN_428970	EHI 096340	16
EIN_371050	EHI 092290	EIN_453940	EHI 104710	17
EIN_371160	EHI 138510	EIN_461630	EHI 153670	18
EIN_371410	EHI 148900	EIN_468220	EHI 006860	19
EIN_379990	EHI 201710	EIN_468220	EHI 006860	20
EIN_380070	EHI 201510	EIN_468500	EHI 051060	21
EIN_380970	EHI 201510	EIN_476210	EHI 077500	22
EIN_386100	EHI 151800	EIN_486020	EHI 196570	23
EIN_390170	EHI 017610	EIN_492940	EHI 177640	24
EIN_391640	EHI 169670	EIN_496850	EHI 166850	25
EIN_391640	EHI 169670	EIN_496860	EHI 003930	26
EIN_391640	EHI 169670	EIN_498400	EHI 020280	27
EIN_391640	EHI 169670	EIN_498410	EHI 068050	28
EIN_391640	EHI 169670	EIN_498890	EHI 050330	29
EIN_391640	EHI 169670			30

